MacVector 12.6

MacVector, Inc.
Software for Scientists

Copyright statement

Trademarks

Gateway®, TOPO®, Vector NTI® and Zero Blunt® are regiestered trademarks of Life Technologies, Carlsbad, California, USA.
Vector NTI Advance™ is a trademark of Life Technologies, Carlsbad, California, USA.

Part one

Introduction

The following chapters introduce you to the structure of this user guide and provide an overview of the MacVector application.

Chapter 1, *“Introduction to the User Guide”*, describes the conventions used in this user guide and outlines the content of each user guide part.

Chapter 2, *“Introduction to MacVector”*, introduces the MacVector Sequence Analysis Software and its principal features.

Introduction to the User Guide

Overview

This chapter describes the information provided in this user guide, where to find it, and the conventions used throughout.

Contents

The MacVector documentation set

Please refer to the release notes included with your copy of MacVector for new and updated information about a particular release. Also please check our website for news of updates, recent issues, and also for freely downloadable updaters for your version of MacVector.

http://www.macvector.com

About this user guide

This user guide describes the MacVector sequence analysis software program and how to use it. The user guide is divided into the following parts:

Introduction	An outline describing the MacVector documentation set and the major functional areas of MacVector.
Using MacVector	A description of using the features of MacVector on a task by task basis.
Understanding MacVector	A description of the theory behind the analyses performed by MacVector, with full references.
Appendices	Additional information, e.g. the file formats that can be used with MacVector, and reference tables of code formats.
Index	A comprehensive index to the MacVector User Guide.

Conventions in this user guide

This user guide assumes that you are familiar with the following:

- the graphical user interface of your operating system
- principles of basic file management.

Refer to Chapter 4, *“Working with MacVector Files”*, for instructions on file management and an explanation of how MacVector manages files.

Interface conventions

The following styles are used in this user guide to describe actions and components of the interface:

- Choose **File | Open** means: choose the **File** menu, then choose **Open** from that menu
- Items on the interface that you need to select, such as buttons or check boxes, are shown in bold, for example:

  the **Offset** check box

  the **and** radio button
- Text that you need to type in, or text that is displayed in a window, is shown as follows:

  `2.0`

  example1.seq

Navigation aids

There is a full table of contents for this user guide on page 1.

As explained earlier, this user guide is divided into five parts, where each part contains one or more chapters.

Each chapter has a short table of contents and an overview that explains exactly what information is contained within the chapter.

Where possible, information on any particular topic is grouped together, ensuring that you can easily locate the information that you need. There are also cross references to other relevant information about a topic.

There is a comprehensive index at the end of the user guide, so that you can easily locate all the information about a particular topic.

2 Introduction to MacVector

Overview

MacVector is a sequence analysis application for Macintosh computers.

This chapter provides an overview of MacVector, introducing:

- MacVector sequence files
- the various data views that can be used to analyze sequence data.
- the search methods and other analyses that can be applied to your sequence data

Data displayed in the various views can be printed and saved in a quality suitable for publication.

Contents

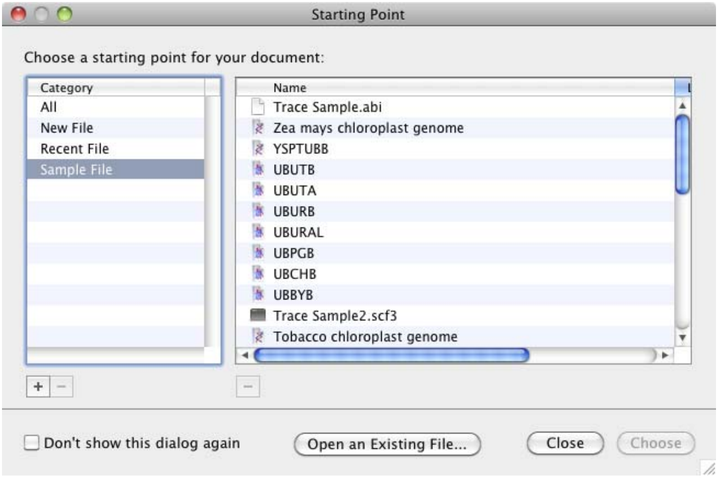
MacVector sequence files

MacVector uses its own binary file format to store sequences, along with features and annotations of the sequence. It can also read many binary file formats.

All major ASCII sequence formats are supported. Refer to Appendix G, *“Supported File Formats and File Extensions”* for a summary.

Starting Point dialog

The **Starting Point** dialog is displayed every time you start MacVector. It provides quick and easy access to sample files and the other folders and files you use most.



Starting Point
Choose a starting point for your document:
+
−
−
□ Don't show this dialog again
Open an Existing File...
Close
Choose

Category
All
New File
Recent File
Sample File

Name
Trace Sample.abi
Zea mays chloroplast genome
YSPTUBB
UBUTB
UBUTA
UBURB
UBURAL
UBPGB
UBCHB
UBBYB
Trace Sample2.scf3
Tobacco chloroplast genome

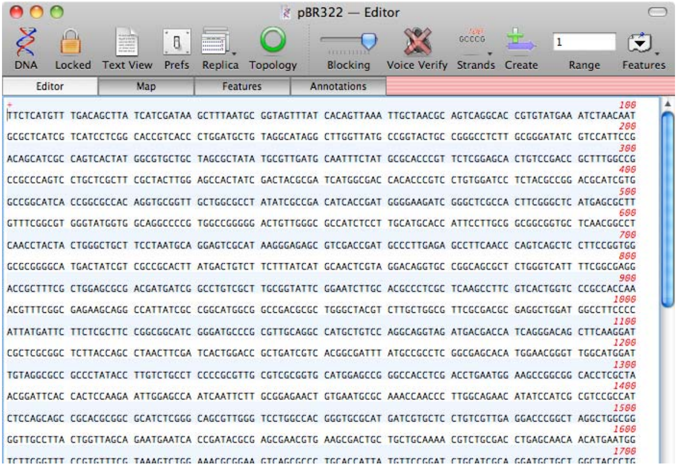
You can add new folders to the list on the left using the **+** icon and remove them using the **-** icon. Alternatively, you can drag and drop new folders onto the dialog. You can also remove recently used files from the list on the right using the **-** icon.

Sequence window

Any sequence file opened in MacVector is displayed in a Sequence window. This comprises a context dependent OS X style toolbar and four tabbed and linked views of the sequence: Editor, Map, Features and Annotations. Refer to *“Sequence window”* on page 76, for further details.

Editing sequence data

The Editor view allows you to enter and edit sequence data.You can use the readback facility to check that the data has been entered correctly. See *“Editor view”* on page 82.



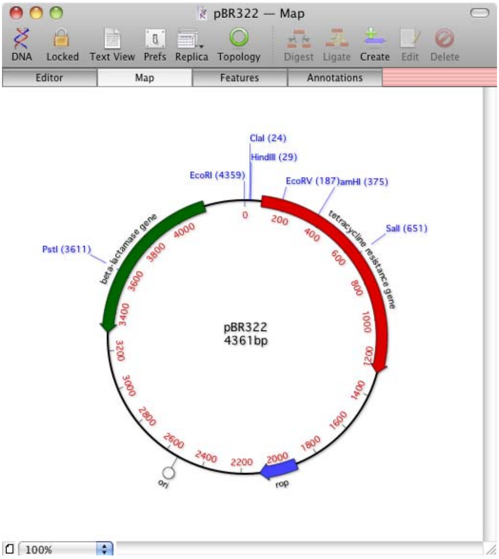
```
                                                                                                          100
TTCTCATGTT TGACAGCTTA TCATCGATAA GCTTTAATGC GGTAGTTTAT CACAGTTAAA TTGCTAACGC AGTCAGGCAC CGTGTATGAA ATCTAACAAT
                                                                                                          200
GCGCTCATCG TCATCCTCGG CACCGTCACC CTGGATGCTG TAGGCATAGG CTTGGTTATG CCGGTACTGC CGGGCCTCTT GCGGGATATC GTCCATTCCG
                                                                                                          300
ACAGCATCGC CAGTCACTAT GGCGTGCTGC TAGCGCTATA TGCGTTGATG CAATTTCTAT GCGCACCCGT TCTCGGAGCA CTGTCCGACC GCTTTGGCCG
                                                                                                          400
CCGCCCAGTC CTGCTCGCTT CGCTACTTGG AGCCACTATC GACTACGCGA TCATGGCGAC CACACCCGTC CTGTGGATCC TCTACGCCGG ACGCATCGTG
                                                                                                          500
GCCGGCATCA CCGGCGCCAC AGGTGCGGTT GCTGGCGCCT ATATCGCCGA CATCACCGAT GGGGAAGATC GGGCTCGCCA CTTCGGGCTC ATGAGCGCTT
                                                                                                          600
GTTTCGGCGT GGGTATGGTG GCAGGCCCCG TGGCCGGGGG ACTGTTGGGC GCCATCTCCT TGCATGCACC ATTCCTTGCG GCGGCGGTGC TCAACGGCCT
                                                                                                          700
CAACCTACTA CTGGGCTGCT TCCTAATGCA GGAGTCGCAT AAGGGAGAGC GTCGACCGAT GCCCTTGAGA GCCTTCAACC CAGTCAGCTC CTTCCGGTGG
                                                                                                          800
GCGCGGGGCA TGACTATCGT CGCCGCACTT ATGACTGTCT TCTTTATCAT GCAACTCGTA GGACAGGTGC CGGCAGCGCT CTGGGTCATT TTCGGCGAGG
                                                                                                          900
ACCGCTTTCG CTGGAGCGCG ACGATGATCG GCCTGTCGCT TGCGGTATTC GGAATCTTGC ACGCCCTCGC TCAAGCCTTC GTCACTGGTC CCGCCACCAA
                                                                                                         1000
ACGTTTCGGC GAGAAGCAGG CCATTATCGC CGGCATGGCG GCCGACGCGC TGGGCTACGT CTTGCTGGCG TTCGCGACGC GAGGCTGGAT GGCCTTCCCC
                                                                                                         1100
ATTATGATTC TTCTCGCTTC CGGCGGCATC GGGATGCCCG CGTTGCAGGC CATGCTGTCC AGGCAGGTAG ATGACGACCA TCAGGGACAG CTTCAAGGAT
                                                                                                         1200
CGCTCGCGGC TCTTACCAGC CTAACTTCGA TCACTGGACC GCTGATCGTC ACGGCGATTT ATGCCGCCTC GGCGAGCACA TGGAACGGGT TGGCATGGAT
                                                                                                         1300
TGTAGGCGCC GCCCTATACC TTGTCTGCCT CCCCGCGTTG CGTCGCGGTG CATGGAGCCG GGCCACCTCG ACCTGAATGG AAGCCGGCGG CACCTCGCTA
                                                                                                         1400
ACGGATTCAC CACTCCAAGA ATTGGAGCCA ATCAATTCTT GCGGAGAACT GTGAATGCGC AAACCAACCC TTGGCAGAAC ATATCCATCG CGTCCGCCAT
                                                                                                         1500
CTCCAGCAGC CGCACGCGGC GCATCTCGGG CAGCGTTGGG TCCTGGCCAC GGGTGCGCAT GATCGTGCTC CTGTCGTTGA GGACCCGGCT AGGCTGGCGG
                                                                                                         1600
GGTTGCCTTA CTGGTTAGCA GAATGAATCA CCGATACGCG AGCGAACGTG AAGCGACTGC TGCTGCAAAA CGTCTGCGAC CTGAGCAACA ACATGAATGG
                                                                                                         1700
TCTTCGGTTT CCGTGTTTCG TAAAGTCTGG AAACGCGGAA GTCAGCGCCC TGCACCATTA TGTTCCGGAT CTGCATCGCA GGATGCTGCT GGCTACCCTG
```

Standard sequence conversions allow you to reverse, complement, and reverse complement regions within a sequence and to create new sequences using translate, and reverse translate functions.

Sequence maps

MacVector gives you complete control over the appearance of the Map view. You choose the way that the features, ruler, and sequence are displayed, and can choose to display some or all of the sequence features.

You can optimize the maps either for on-screen appearance or for high-quality printed output.

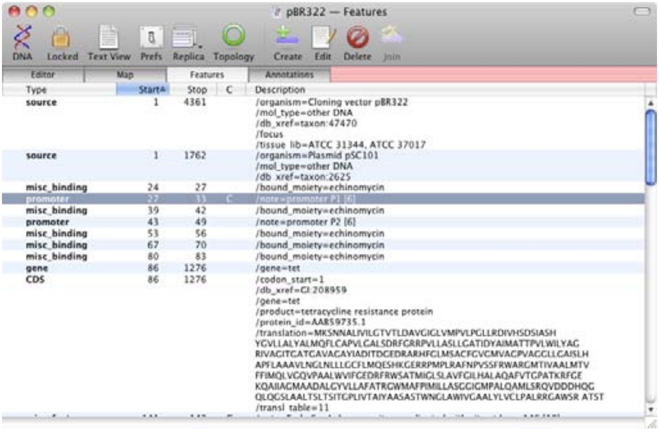


Features and annotations

Sequence features and other regions of interest can be marked in a number of ways, following GenBank formatting conventions.

They can be displayed graphically in the Map view and as a list in the Features view. See *“Features view”* on page 113.

When a sequence is unlocked, features are edited easily from the table, using the toolbar at the top of the Sequence window (or by double clicking). Clicking once on a feature automatically selects the residues corresponding to that feature in the Editor view.

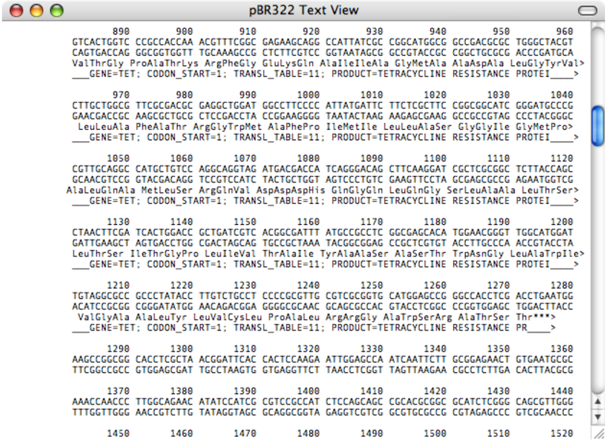


Type	Start	Stop	C	Description
source	1	4361		/organism=Cloning vector pBR322 /mol_type=other DNA /db_xref=taxon:47470 /focus /tissue_lib=ATCC 31344, ATCC 37017
source	1	1762		/organism=Plasmid pSC101 /mol_type=other DNA /db_xref=taxon:2625
misc_binding	24	27		/bound_moiety=echinomycin
promoter	27	33	C	/note=promoter P1 [6]
misc_binding	39	42		/bound_moiety=echinomycin
promoter	43	49		/note=promoter P2 [6]
misc_binding	53	56		/bound_moiety=echinomycin
misc_binding	67	70		/bound_moiety=echinomycin
misc_binding	80	83		/bound_moiety=echinomycin
gene	86	1276		/gene=tet
CDS	86	1276		/codon_start=1 /db_xref=GI:208959 /gene=tet /product=tetracycline resistance protein /protein_id=AAB59735.1 /translation=MKSNNALIVILGTVTLDAVGIGLVMPVLPGLLRDIVHSDSIASH YGVLLALYALMQFLCAPVLGALSDRFGRRPVLLASLLGATIDYAIMATTPVLWILYAG RIVAGITGATGAVAGAYIADITDGEDRARHFGLMSACFGVGMVAGPVAGGLLGAISLH APFLAAAVLNGLNLLLGCFLMQESHKGERRPMPLRAFNPVSSFRWARGMTIVAALMTV FFIMQLVGQVPAALWVIFGEDRFRWSATMIGLSLAVFGILHALAQAFVTGPATKRFGE KQAIIAGMAADALGYVLLAFATRGWMAFPIMILLASGGIGMPALQAMLSRQVDDDHQG QLQGSLAALTSLTSITGPLIVTAIYAASASTWNGLAWIVGAALYLVCLPALRRGAWSR ATST /transl_table=11

Text annotations associated with a sequence are displayed in the Annotations view. These provide extra information about the sequence, such as a description, academic references, or identifying numbers assigned by the databases from which the sequences are taken.

Text view

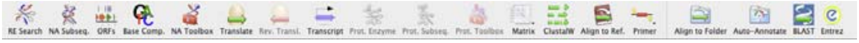
A Text view can be customized to show additional information including defined sequence features, and, in the case of nucleic acid sequences, the complement strand and translation of major features.



```
        890        900        910        920        930        940        950        960
GTCACTGGTC CCGCCACCAA ACGTTTCGGC GAGAAGCAGG CCATTATCGC CGGCATGGCG GCCGACGCGC TGGGCTACGT
CAGTGACCAG GGCGGTGGTT TGCAAAGCCG CTCTTCGTCC GGTAATAGCG GCCGTACCGC CGGCTGCGCG ACCCGATGCA
ValThrGly ProAlaThrLys ArgPheGly GluLysGln AlaIleIleAla GlyMetAla AlaAspAla LeuGlyTyrVal>
___GENE=TET; CODON_START=1; TRANSL_TABLE=11; PRODUCT=TETRACYCLINE RESISTANCE PROTEI____>

        970        980        990       1000       1010       1020       1030       1040
CTTGCTGGCG TTCGCGACGC GAGGCTGGAT GGCCTTCCCC ATTATGATTC TTCTCGCTTC CGGCGGCATC GGGATGCCCG
GAACGACCGC AAGCGCTGCG CTCCGACCTA CCGGAAGGGG TAATACTAAG AAGAGCGAAG GCCGCCGTAG CCCTACGGGC
 LeuLeuAla PheAlaThr ArgGlyTrpMet AlaPhePro IleMetIle LeuLeuAlaSer GlyGlyIle GlyMetPro>
___GENE=TET; CODON_START=1; TRANSL_TABLE=11; PRODUCT=TETRACYCLINE RESISTANCE PROTEI____>

       1050       1060       1070       1080       1090       1100       1110       1120
CGTTGCAGGC CATGCTGTCC AGGCAGGTAG ATGACGACCA TCAGGGACAG CTTCAAGGAT CGCTCGCGGC TCTTACCAGC
GCAACGTCCG GTACGACAGG TCCGTCCATC TACTGCTGGT AGTCCCTGTC GAAGTTCCTA GCGAGCGCCG AGAATGGTCG
AlaLeuGlnAla MetLeuSer ArgGlnVal AspAspAspHis GlnGlyGln LeuGlnGly SerLeuAlaAla LeuThrSer>
___GENE=TET; CODON_START=1; TRANSL_TABLE=11; PRODUCT=TETRACYCLINE RESISTANCE PROTEI____>

       1130       1140       1150       1160       1170       1180       1190       1200
CTAACTTCGA TCACTGGACC GCTGATCGTC ACGGCGATTT ATGCCGCCTC GGCGAGCACA TGGAACGGGT TGGCATGGAT
GATTGAAGCT AGTGACCTGG CGACTAGCAG TGCCGCTAAA TACGGCGGAG CCGCTCGTGT ACCTTGCCCA ACCGTACCTA
LeuThrSer IleThrGlyPro LeuIleVal ThrAlaIle TyrAlaAlaSer AlaSerThr TrpAsnGly LeuAlaTrpIle>
___GENE=TET; CODON_START=1; TRANSL_TABLE=11; PRODUCT=TETRACYCLINE RESISTANCE PROTEI____>

       1210       1220       1230       1240       1250       1260       1270       1280
TGTAGGCGCC GCCCTATACC TTGTCTGCCT CCCCGCGTTG CGTCGCGGTG CATGGAGCCG GGCCACCTCG ACCTGAATGG
ACATCCGCGG CGGGATATGG AACAGACGGA GGGGCGCAAC GCAGCGCCAC GTACCTCGGC CCGGTGGAGC TGGACTTACC
 ValGlyAla AlaLeuTyr LeuValCysLeu ProAlaLeu ArgArgGly AlaTrpSerArg AlaThrSer Thr***>
___GENE=TET; CODON_START=1; TRANSL_TABLE=11; PRODUCT=TETRACYCLINE RESISTANCE PR____>

       1290       1300       1310       1320       1330       1340       1350       1360
AAGCCGGCGG CACCTCGCTA ACGGATTCAC CACTCCAAGA ATTGGAGCCA ATCAATTCTT GCGGAGAACT GTGAATGCGC
TTCGGCCGCC GTGGAGCGAT TGCCTAAGTG GTGAGGTTCT TAACCTCGGT TAGTTAAGAA CGCCTCTTGA CACTTACGCG

       1370       1380       1390       1400       1410       1420       1430       1440
AAACCAACCC TTGGCAGAAC ATATCCATCG CGTCCGCCAT CTCCAGCAGC CGCACGCGGC GCATCTCGGG CAGCGTTGGG
TTTGGTTGGG AACCGTCTTG TATAGGTAGC GCAGGCGGTA GAGGTCGTCG GCGTGCGCCG CGTAGAGCCC GTCGCAACCC

       1450       1460       1470       1480       1490       1500       1510       1520
```

Analysis toolbar

The Analysis toolbar appears across the top of the screen whenever MacVector is active. It provides shortcuts to most of the analysis functions available in MacVector. The toolbar icons are enabled whenever a suitable Sequence window is active.



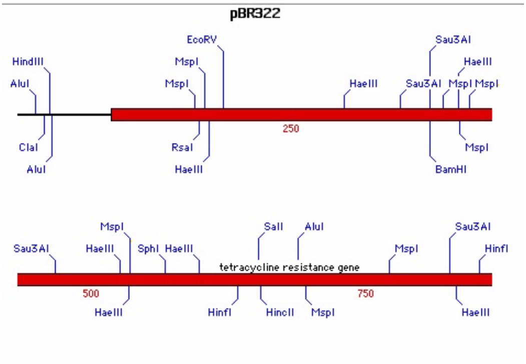
The toolbar is fully customizable. You can turn the toolbar off and on using **Window | Show/Hide Analyses Tools** or modify its contents and appearance using the **Customize Toolbar** option accessed by right-clicking on the toolbar when no other windows are open.

All of the Analysis toolbar icons can be added to the standard Sequence window toolbars too.

Site and motif searches

MacVector includes protein motif, nucleic acid motif, restriction site and proteolytic site searching. Motifs can consist of 1, 2 or 3 parts, and you can define the gap between each part. ORF detection can be used to assist gene detection. Search results can be displayed as annotated sequences, subsequence maps, fragment predictions and sizes, and tabular data summaries, allowing flexible processing of data.

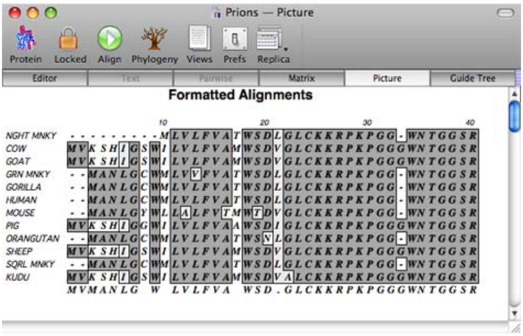
All motif searches are based upon subsequence files, which contain a set of motifs. You can edit and create new files to contain your own search analyses.



Sequence comparison, alignment and phylogeny

MacVector uses Pustell matrix analysis to compare two sequences, generating a dot plot that provides an excellent overview of the sequence similarity. You can also compare a query sequence *via* the Internet to databases available on the NCBI server, using the BLAST algorithm.

You can compare and align a set of sequences that are stored in a folder, or use different algorithms to align multiple sequences. The MSA Editor lets you enter, import and edit multiple sequences, and perform phylogenetic analyses, generating tree diagrams of their evolutionary relationships.

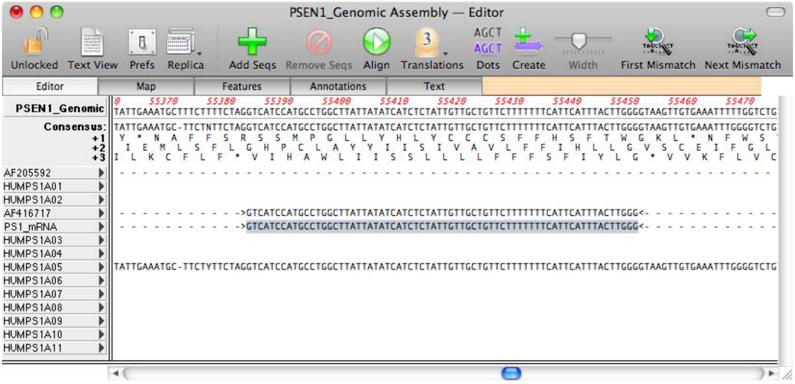


Formatted Alignments

```
                            10        20        30        40
NGHT MNKY   ---------MLVLFVATWSDLGLCKKRPKPGG-WNTGGSR
COW         MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSR
GOAT        MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSR
GRN MNKY    --MANLGCWMLVVFVATWSDLGLCKKRPKPGG-WNTGGSR
GORILLA     --MANLGCWMLVLFVATWSDLGLCKKRPKPGG-WNTGGSR
HUMAN       --MANLGCWMLVLFVATWSDLGLCKKRPKPGG-WNTGGSR
MOUSE       --MANLGYWLLALFVTMWTDVGLCKKRPKPGG-WNTGGSR
PIG         MVKSHIGGWILVLFVAAWSDIGLCKKRPKPGGGWNTGGSR
ORANGUTAN   --MANLGCWMLVLFVATWSNLGLCKKRPKPGG-WNTGGSR
SHEEP       MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSR
SQRL MNKY   --MANLGCWMLVLFVATWSDLGLCKKRPKPGG-WNTGGSR
KUDU        MVKSHIGSWILVLFVAMWSDVALCKKRPKPGGGWNTGGSR
            MVMANLG W LVLFVA WSD.GLCKKRPKPGGGWNTGGSR
```

Aligning sequences to a reference

The Align to Reference tool in MacVector enables you to align one or more sequences against a reference sequence so that you can quickly identify differences or similarities. Two different options are available:



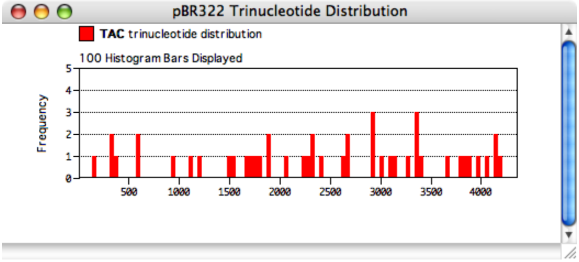
```
               0      55370     55380     55390     55400     55410     55420     55430     55440     55450     55460     55470
PSEN1_Genomic  TATTGAAATGCTTTCTTTTCTAGGTCATCCATGCCTGGCTTATTATATCATCTCTATTGTTGCTGTTCTTTTTTTCATTCATTTACTTGGGGTAAGTTGTGAAATTTTTGGTCTG
Consensus:     TATTGAAATGC-TTCTNTTCTAGGTCATCCATGCCTGGCTTATTATATCATCTCTATTGTTGCTGTTCTTTTTTTCATTCATTTACTTGGGGTAAGTTGTGAAATTTGGGGTCTG
+1              Y  *  N  A  F  F  S  R  S  S  M  P  G  L  L  Y  H  L  Y  C  C  C  S  F  F  H  S  F  T  W  G  K  L  *  N  F  W  S
+2               I  E  M  L  S  F  L  G  H  P  C  L  A  Y  Y  I  I  S  I  V  A  V  L  F  F  I  H  L  L  G  V  S  C  E  I  F  G  L
+3             I  L  K  C  F  L  F  *  V  I  H  A  W  L  I  I  S  S  L  L  L  L  F  F  F  S  F  I  Y  L  G  *  V  V  K  F  L  V  C
AF205592       - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HUMPS1A01
HUMPS1A02
AF416717       - - - - - - - - - - - ->GTCATCCATGCCTGGCTTATTATATCATCTCTATTGTTGCTGTTCTTTTTTTCATTCATTTACTTGGG<- - - - - - - - - - - -
PS1_mRNA       - - - - - - - - - - - ->GTCATCCATGCCTGGCTTATTATATCATCTCTATTGTTGCTGTTCTTTTTTTCATTCATTTACTTGGG<- - - - - - - - - - - -
HUMPS1A03
HUMPS1A04
HUMPS1A05      TATTGAAATGC-TTCTYTTCTAGGTCATCCATGCCTGGCTTATTATATCATCTCTATTGTTGCTGTTCTTTTTTTCATTCATTTACTTGGGGTAAGTTGTGAAATTTGGGGTCTG
HUMPS1A06
HUMPS1A07
HUMPS1A08
HUMPS1A09
HUMPS1A10
HUMPS1A11
```

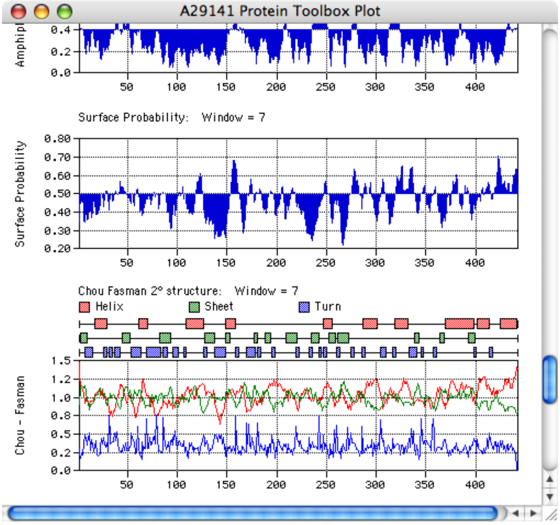
The Sequence Confirmation tool lets you align one or more sequences or chromatogram sample files against a reference sequence. While the cDNA Alignment tool lets you align one or more cDNA clones against a reference genomic sequence.

Protein and nucleic acid toolbox analysis

Property profiles for both nucleic acid and protein sequences can be plotted.

MacVector can be customized to display various combinations of properties. It is easy to magnify regions of interest.


pBR322 Trinucleotide Distribution
TAC trinucleotide distribution
100 Histogram Bars Displayed
Frequency
0
1
2
3
4
5
500
1000
1500
2000
2500
3000
3500
4000



Part two

Using MacVector

This part describes how to use MacVector and includes the following chapters:

Chapter 3, *“General Procedures”*, describes how you can customize the appearance of the sequences and graphics displayed by MacVector.

Chapter 4, *“Working with MacVector Files”*, introduces all the file types that MacVector uses, and how to modify non-sequence files.

Chapter 5, *“Working with Sequences and Features”*, describes how to create and modify sequences and their associated features and annotations.

Chapter 6, *“Searching and Downloading Sequences in the Entrez Database”*, describes how to locate and extract complete sequences from the Entrez database released by NCBI.

Chapter 7, *“Calculating Sequence Properties”*, describes how to calculate property profiles and how to search a nucleic acid sequence for coding regions.

Chapter 8, *“Searching for Sites and Motifs”*, provides guidelines on working with the sequence search methods in MacVector.

Chapter 9, *“Click Cloning”* describes the click cloning functionality available in MacVector and how it can be used to simulate commercial cloning technologies.

Chapter 10, *"Primer and Probe Design"*, describes the methods used to design primers and screen for primers and probes.

Chapter 11, *"Using Transcription and Translation Functions"*, describes the facilities for controlling the transcription and translation of DNA to protein using any genetic code, and also reverse translation.

Chapter 12, *"Aligning Sequences"*, describes the range of methods available in MacVector for sequence comparison and alignment.

Chapter 13, *"Aligning sequences using a folder search"* describes how to use the Align to Folder functionality provided in MacVector.

Chapter 14, *"Aligning and Downloading Sequences with BLAST"* describes how to align sequences using internet BLAST.

Chapter 15, *"Comparing Sequences using Pustell Matrix Analysis (Dot Plot)"*, describes the MacVector implementation of Pustell matrix analysis for sequence similarity.

Chapter 16, *"Aligning Sequences to a Reference"*, describes how to align one or more sequences against a reference sequence so that you can quickly identify differences or similarities.

Chapter 17, *"Aligning Multiple Sequences"*, describes how to handle, display and edit multiple sequences.

Chapter 18, *"Reconstructing Phylogeny"*, describes methods for generating phylogenetic trees from multiple alignments of nucleic acid or protein sequences.

Chapter 19, *"Sequence Assembly"*, describes the Assembler add-on module for Contig Assembly.

3 General Procedures

Overview

This chapter describes several generally useful functions available in MacVector for visualizing, formatting and interpreting results. It also describes how you can use:

- the numeric keypad to make it easier to enter nucleotide sequences
- the proofreader to have the residue sequenece read back to you
- the Find tool to search for and replace specific subsequences or features

Contents

Performing sequence analysis

In general performing sequence analysis with MacVector involves three steps:

- First you must open a sequence, then select an analysis tool from a toolbar or menu.
- Secondly, you must setup the analysis. Most analyses require you to supply input parameters or supporting files.
- Finally, you can filter the results of the analysis. Most analyses produce summary results and at this stage it is possible to select which results are shown in detail. You can also choose how you want the results to be displayed by selecting which type of results window to use.

This is illustrated in the example given below.

Example - searching for restriction enzyme sites

To search for restriction enzyme sites manually, then filter and display the results

1. Make the required nucleic acid Sequence window active and choose **Analyze | Restriction Enzyme** from the main menu.
2. Select your search criteria by choosing from the **Search Using** drop-down menu as follows:

- choose **all enzymes** to use all the enzymes in the selected enzyme file
- choose **selected enzymes** to use only the enzymes that you have selected within the enzyme file
- choose **other criteria** to display further options

And, if you have chosen **other criteria**, do one or more of the following:

- to restrict the search to enzymes whose recognition sites are a certain size, type the size limits into the **site size** text boxes.
- to limit the search to enzymes that leave 5' ends, 3' ends, or blunt ends when they cut the sequence, select the appropriate option from the **end structure** drop-down menu.

Finally, to restrict the search to enzymes that cut the sequence a limited number of times, type the upper limit in the **number of cuts** text box. Then select **OK** to perform the analysis. When this is complete, a result filtering dialog is displayed.

3. Select from the following filters for the results, as required:
   - by the number of cuts an enzyme makes
   - by the end structure of cut fragments
   - by restriction site size
   - by the residue range in which the cut site occurs

And select **Display Options** to generate separate windows displaying the results. Click **OK** to display the results windows.

Viewing results

The results of a MacVector sequence analysis are displayed in one or more windows. Many graphical results are displayed as a map, whose appearance is controlled by two dialog boxes:

- the Symbol Editor, see *“Editing the global symbol set”* on page 94
- the Graphics Palette, see *“Editing the general map appearance”* on page 109.

Non-graphical results appear as text listings.

Magnifying a graphic area

You can enlarge a region of any displayed graphic result.

To magnify an area of the graphic

1. Move the mouse cursor into the graphic area so that the mouse cursor changes to a magnifying glass.
2. Hold the mouse button down at one side of the area you want to enlarge and drag the cursor to the other side of the required area.
3. When you have selected the required area for enlargement, release the mouse button.

The selected area of the graphic is enlarged and redisplayed.

To return a magnified area to the original size

1. Double-click the mouse button in the graphic area.

The graphic is redisplayed with the entire residue range.

You can also resize any graphical result so that it fits entirely within the current window using the **Fit to Window** option on the Graphics Palette (see *“Editing the general map appearance”* on page 109).

Outputting results

You can output the results of an analysis in two ways:

- Print the results
- Save the results to a file.

To print results

1. Make the plot or list you want to print active by clicking on it.
2. Choose **File | Print**.

A dialog box is displayed, enabling you to choose your printer and its settings.

3. Select **Print** to print the results.

A dialog box is displayed, informing you of the progress of the printing.

Tip. All or part of a list result can be copied to the clipboard by highlighting the required text.

To save results to a file

1. Make the view you want to save active by clicking on the corresponding tab.
2. Choose **File | Export**.

The **Export** dialog box is displayed.

3. If required, navigate to an alternative folder, or create a **New Folder**.
4. If required, change the file name by typing in the **Save As** text box.
5. Select **Save** to export the file.

The representation of the underlying data displayed in the current view is exported to a file in an appropriate format, for example a PDF file for graphics or a text file for annotations.

Saving graphics

In line with most modern OS X applications, MacVector uses the PDF format for exporting most graphics. However, some parts of MacVector still use the older PICT format, for example, Phylogenetic trees and ClustalW guide trees still use the PICT format.

You can export these graphics files directly from the corresponding graphics view using the **File | Export** option.

Alternatively, you can copy the graphics from a graphics view and place them on the clipboard for pasting directly into another application.

PDF format retains high resolution and can be pasted into many applications (Adobe Illustrator, TextEdit, Pages, KeyNote etc.) with no loss of resolution. However, some older applications, particularly Microsoft Office 2004 programs such as PowerPoint or Word, cannot import clipboard information in PDF format. So, MacVector also places graphical information on the clipboard in bitmap format.

You can control the resolution used by MacVector for bitmap copies using the **Map View** preferences dialog accessible using **MacVector | Preferences** from the menu, then clicking the **Map View** icon on the preferences dialog, or by clicking the **Prefs** icon on the toolbar when a Map view is selected.

PDF files

Most graphics in MacVector are exported as PDF files. This format is best as long as the intended application supports PDF. PDF is a vector format, and if your intended application supports both PDF file input and vector image editing (e.g. Adobe Illustrator CS3, or Adobe Photoshop CS3), then the MacVector graphic can be edited further in that application.

To Save a PDF image of a graphic

1. Select the required graphics view and choose **File | Export** from the menu.

The **Export** dialog box is displayed.

2. If required, navigate to an alternative folder, or create a **New Folder**.
3. If required, change the file name by typing in the **Save As** text box.
4. Select **Save** to export the file.

Or, alternatively

1. Select the required graphics view and choose **File | Print** from the menu.

The **Print** dialog box is displayed.

2. Click the **PDF** button and choose **Save as PDF...** from the menu.
3. Navigate to the folder where you want to save the graphic data and choose **Save**.

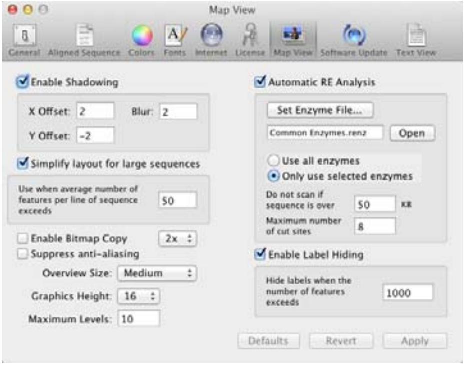
To Copy a PDF image of a graphic to another application

1. Choose the required graphics view and select **Edit | Copy** from the menu.
2. Navigate to the application you want to copy the graphic to, and select **Edit | Paste** in that application.

Note. It is not necessary to select anything within the graphics view in order to copy its contents to the clipboard for pasting.

Bitmap files

MacVector places graphical information on the clipboard in bitmap format. To increase the quality of the exported bitmap image you can use oversampling to increase the resolution of the image placed in the clipboard. By default, MacVector has bitmapped copy turned on with 1x oversampling. This copies an image with the same resolution as you can see on the screen.



☑ Enable Shadowing

X Offset: 2 Blur: 2

Y Offset: -2

☑ Simplify layout for large sequences

Use when average number of features per line of sequence exceeds 50

□ Enable Bitmap Copy 2x

□ Suppress anti-aliasing

Overview Size: Medium

Graphics Height: 16

Maximum Levels: 10

☑ Automatic RE Analysis

Common Enzymes.renz

○ Use all enzymes

● Only use selected enzymes

Do not scan if sequence is over 50 KB

Maximum number of cut sites 8

☑ Enable Label Hiding

Hide labels when the number of features exceeds 1000

When you choose **Edit | Copy**, the entire graphical image is copied and placed on the clipboard, not just the visible region. For large sequences this can be a tremendous amount of information, particularly if you use 2x or 4x oversampling to increase the bitmap resolution. You can limit the amount of memory MacVector allocates for the bitmap copy by changing the appropriate value in the **Map View** preferences dialog. By default this is set to 16 MB. If the required memory would exceed this value, the bitmap copy is simply skipped.

PICT files

Some parts of MacVector still use the older PICT format, for example, Phylogenetic trees and ClustalW guide trees are exported in PICT format. PICT files are specific to Macintosh computers. They include extra high-resolution information, which is not visible on the screen, but increases the printed quality. If you are always working and printing on a Mac, PICT files are standard and convenient.

Note. This feature is included for legacy purposes only. Most graphics in MacVector are now exported in PDF and bitmap format.

Formatting the Aligned Sequence view

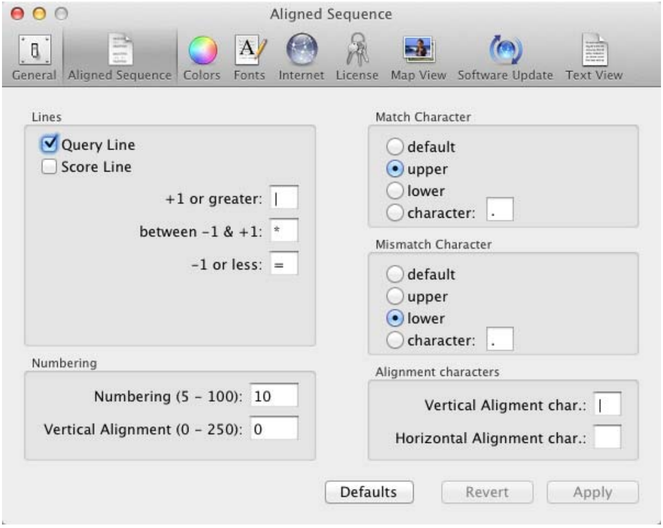
You can customize the appearance of the alignment portion of the Aligned Sequence view, both as it appears on the screen and as it will be printed. This formatting does not apply to multiple sequence alignments, which have their own formatting options.

The aligned sequence formatting enables you to control:

- whether or not a query sequence line will appear
- the style and display of a score line
- how often the aligned sequences will be numbered and how often vertical alignment characters will appear
- the display of vertical and horizontal alignment characters
- the representation of matching and mismatching residues.

Characteristics of the query sequence are customized using the **Text Display** options. Examples of formatting can be found in Appendix D, *“Formatting Examples”*.

This functionality can be accessed at any time, whether an aligned sequence is present or not. If an aligned sequence is present, it will be updated to reflect changes when you click **Apply**, switch to a different view or close the dialog.



Lines

☑ Query Line

☐ Score Line

+1 or greater: |

between -1 & +1: *

-1 or less: =

Match Character

○ default
● upper
○ lower
○ character: .

Mismatch Character

○ default
○ upper
● lower
○ character: .

Numbering

Numbering (5 – 100): 10

Vertical Alignment (0 – 250): 0

Alignment characters

Vertical Aligment char.: |

Horizontal Alignment char.:

There are three possible components to an Aligned Sequence view:

- the aligned sequence
- the score line
- the query sequence.

If all three are present, they appear beneath each other in the order shown above.

Formatting the aligned sequence

The format of an aligned sequence affects the numbering of residues, representation of matched and mismatched residues, and visual alignment aids.

To format an aligned sequence

1. Choose **MacVector | Preferences** from the menu, then click the **Aligned Sequence** icon on the preferences dialog.

The **Aligned Sequence** preferences dialog box is displayed.

Note. You can also access the **Aligned Sequence** preferences dialog using **Options | Aligned Sequence** from the menu.

2. Type a number in the **Numbering** text box to number the sequence residues at that interval.

3. Type a number in the **Vertical Alignment** text box to insert vertical alignment characters at that interval through the display.

This is an aid to keeping your place as you scan down through the aligned sequences. If you enter a zero in this box, no vertical alignment characters will appear.

4. Type a character in the **Vertical Alignment char.** text box to use as a vertical alignment character.

If you do not want any vertical alignment characters to appear, type a space in the box or leave it empty.

5. Type a character in the **Horizontal Alignment char.** text box to use as a horizontal alignment character.

This is an aid to keeping your place on the line as you scan left to see the sequence name. If you do not want any horizontal alignment characters to appear, type a space in the box or leave it empty.

Note. Do not use a hyphen, because MacVector uses hyphens to represent deletions and gaps in the aligned sequence.

6. From the **Match Character** panel, select the radio button corresponding to how you want to represent residues in the aligned sequences that match residues in the query sequence.

The **character** radio button enables you to enter a character of your choice in the text box.

7. From the **Mismatch Character** panel, select the radio button corresponding to how you want to represent residues in the aligned sequences that do not match residues in the query sequence.

The **character** radio button enables you to enter a character of your choice in the text box.

8. Select **OK** to apply the formatting.

Each Aligned Sequence view is updated with the formatting changes.

Displaying a score line

Each character in the score line represents the score assigned to the comparison between the residue in the aligned sequence and the corresponding residue in the query sequence. This gives you a semi-graphical indication of which regions along the length of the sequence match the query sequence well and which match poorly.

To display a score line

1. Choose **MacVector | Preferences** from the menu, then click the **Aligned Sequence** icon on the preferences dialog.

The **Aligned Sequence** preferences dialog box is displayed.

Note. You can also access the **Aligned Sequence** preferences dialog using **Options | Aligned Sequence** from the menu.

2. Select the **Score Line** check box to display a score line beneath each of the sequences in the alignment.
3. Type a character of your choice in each of the boxes labeled **+1 or greater**, **between -1 and +1**, and **-1 or less**.

These characters can identify matching residues along the aligned sequence. Most scoring matrices give a positive score to an exact or equivalent match, and a negative score to a mismatch. Any or all of these boxes may contain a space character or be left blank.

4. Select **OK** to apply the score line.

Each Aligned Sequence view is updated with the formatting changes.

Displaying a query line

The query sequence can be displayed beneath each of the sequences in the alignment. The format of the query sequence is controlled using the **Text Display** options. See *“Formatting the Aligned Sequence view”* on page 41.

To display a query line

1. Choose **MacVector | Preferences** from the menu, then click the **Aligned Sequence** icon on the preferences dialog.

The **Aligned Sequence** preferences dialog box is displayed.

Note. You can also access the **Aligned Sequence** preferences dialog using **Options | Aligned Sequence** from the menu.

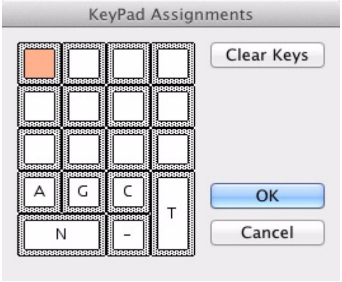
2. Select the **Query Line** check box to display a query line beneath each of the sequences in the alignment.
3. Select **OK** to apply the query line.

Each Aligned Sequence view is updated with the formatting changes.

Configuring the numeric keypad

This option enables you to assign the one-letter codes for nucleotides to keys of the numeric keypad (if your Macintosh has one) to make it easier to enter sequence data with one hand.

Assignments can be made whenever the menu bar is accessible, but they will be in effect only when a nucleic acid sequence tab is active. At other times, the keypad keys will retain their normal assignments as numbers or symbols. The assignment is not retroactive; if you type in some bases from the keypad, then change the keypad assignment in mid-sequence, the bases you previously typed in will not be changed.



To configure the numeric keypad

1. Choose **Options | Set Keypad**.

The **KeyPad Assignments** dialog box appears.

2. To cancel all keypad assignments that are currently in effect, select the **Clear Keys** button.
3. Click on the box corresponding to the key that you want to assign.
4. Type the letter to assign to that key.

Only valid IUPAC one-letter codes will be accepted.

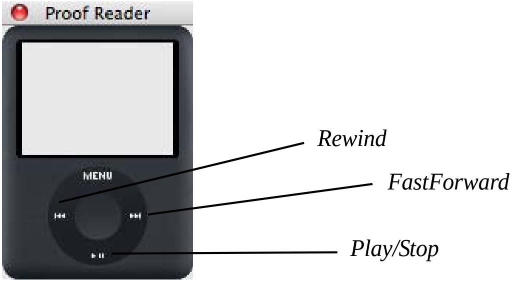
Note. When assigning IUPAC codes to the keypad, the letter **U** is not used. Instead, the **T** key enters **T** into a DNA sequence and **U** into an RNA sequence.

5. To make changes permanent, do one of the following:
   - select the **Cancel** button to discard all changes made. MacVector ignores any changes you may have made to the keypad assignments and retains existing ones.
   - Select **OK** to accept the keypad assignments that are currently displayed.

The assignments are saved when you exit MacVector.

Using the proofreader

MacVector provides a proofreading tool that enables you to have the residue sequence read back to you. This facility is available both as you type a sequence in, and as a proofreading tool on a selected sequence.



Reading a typed sequence

You can have a sequence repeated to you as you type.

To read back a sequence as it is typed

1. Choose **Windows | Show Proofreader** to display the proofreader.
2. Click on **MENU** to select a male or female voice and the volume level as required.
3. Position the cursor in the Sequence window at the insertion point where you want to enter the new or additional residue sequence.
4. Type in the sequence.

As you type the sequence, it is read back to you.

Proofreading a sequence

After you have entered a sequence, you can proofread it.

To proofread a sequence

1. Select the required portion of the sequence to be proofread.
2. Choose **Windows | Show Proofreader** to display the proofreader.
3. Select a male or female voice and the volume level as required.
4. Select the **start** button on the proofreader to begin.

Each residue is unhighlighted in turn as the sequence is read.

5. If required, you can pause the playback at any time with the **pause** button on the proofreader.

6. If required, you can move backwards or forwards along the sequence by using the **fast forward** and **rewind** buttons of the proofreader.

Selecting the sequence using the proofreader

The sequence for proofreading can be selected by using the **fast forward** and **rewind** buttons of the proofreader.

To select the sequence using the proofreader

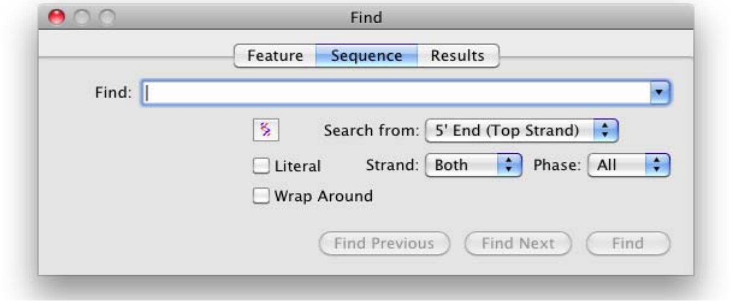
1. Position the insertion point at the beginning or end of the required selection.
2. Use either the **fast forward** button or the **rewind** button to highlight residues from the insertion point.
3. Use the **stop** button when the required residue block has been highlighted.

The sequence can now be proofread as described previously.

Searching

The Find tool in MacVector provides powerful and flexible search and replace capabilities for finding specific subsequences in proteins and DNA in the current sequence or multiple sequence alignment. It also enables you to search features associated with a sequence and analysis results, such as BLAST search results.

Choose **Edit | Find** to display the **Find** dialog box. The dialog box has three tabs for the three different types of search: Feature, Sequence and Results.



Find:

Search from: 5' End (Top Strand)

☐ Literal

Strand: Both

Phase: All

☐ Wrap Around

Searching sequences

The sequence search enables you to find specific subsequences in proteins and DNA in the current sequence or multiple sequence alignment.

Tip. If you only need to find a specific residue number, use **Edit | Jump to Location**. The number you specify can be relative to the 5' end of the sequence, the 3' end of the sequence, the current position of the insertion point, or the plus or minus origin.

To perform a subsequence search

1. Choose **Edit | Find** from the main menu.

The **Find** dialog box is displayed.

2. Ensure that the **Sequence** tab is selected.

This tab is selected by default if the **Find** dialog box is opened when a Sequence Editor or MSA Editor view is active.

3. Type the subsequence that you want to find in the **Find** text box.
4. Choose where to start the search from using the **Search from** drop-down menu.
5. Choose which strand to search from using the **Strand** drop-down menu. By default, both strands are searched.
6. Choose which phase to search from the **Phase** drop-down menu.
7. By default, the **Find** tool uses IUPAC codes to search for matching residues, so that the search string AGY will also locate the ambiguous residues AGT and AGC. If you just want to locate the sequence AGY, then check the **Literal** box.
8. By default, the **Find** tool will match any queries in circular seqeuences that cross the origin, searching from the 5' end of a forward sequence (and 3' end of the reverse sequence) until it finds the first match, then looking for the subsequent matches until it reaches the end of a sequence. If you want to locate matches indefinitely around a sequence, then check the **Wrap Around** option.
9. Do one of the following:
   - to find the first occurrence of the subsequence, select **Find**
   - to find the next occurrence of the subsequence, select **Find Next**
   - to find the previous occurrence of the subsequence, select **Find Previous**

If the subsequence is present in a sequence, it will be highlighted, if not, the alert will sound.

You can also use the **Find** tool to search a protein sequence using a DNA subsequence and *vice versa*. This is useful if, for example, you want to find all subsequences in a DNA sequence that could code for the tripeptide leucine-serine-arginine.

MacVector automatically performs the translation (or reverse-translation) using the current genetic code.

To search a DNA sequence using a protein subsequence

1. Choose **Edit | Find** from the main menu.

The **Find** dialog box is displayed.

2. Ensure that the **Sequence** tab is selected.

This tab is selected by default if the **Find** dialog box is opened when a Sequence Editor view is active.

3. Click the **molecule** icon to toggle from DNA to protein.
4. Type the protein subsequence in the **Find** text box.
5. Select other options if required, as described in steps 2 through 8 in the procedure *"To perform a subsequence search"* on page 48.
6. Select **Find** to perform the search.

Tip. If you select the **molecule** button again after entering the subsequence, you can see the translated DNA subsequence.

To search a protein sequence using a DNA subsequence

1. Choose **Edit | Find** from the main menu.

The **Find** dialog box is displayed.

2. Ensure that the **Sequence** tab is selected.

This tab is selected by default if the **Find** dialog box is opened when a Sequence Editor view is active.

3. Click the **molecule** icon to toggle from protein to DNA.
4. Type the DNA subsequence in the **Find** text box.
5. Select other options if required, as described in steps 2 through 8 in the procedure *"To perform a subsequence search"* on page 48.
6. Select **Find** to perform the search.

The **Find** tool also enables you to specify a replacement string.

To replace one subsequence with another in a sequence

1. Choose **Edit | Find** from the main menu.

The **Find** dialog box is displayed.

2. Ensure that the **Sequence** tab is selected.

This tab is selected by default if the **Find** dialog box is opened when a Sequence Editor view is active.

3. Type the subsequence you want to find in the **Find** text box.
4. Type a replacement subsequence in the **Replace** text box.
5. Select other options if required, as described in the procedure *"To perform a subsequence search"* on page 48.
6. Do one of the following:
    - to substitute the replacement string for the first occurrence of the subsequence, select **Find** then **Replace**. After replacement, the **Replace** button is unavailable.
    - to substitute the replacement string for the selected occurrence of the subsequence and find the next occurrence of the subsequence, select **Replace & Find**
    - to substitute the replacement string for all occurrences of the subsequence, select **Replace All**

Searching features

The feature search enables you to find specified text within the features associated with a sequence.

Search results are selected automatically in the Sequence Map and Sequence Features views, either all at once using the **All** button or one at a time using the **Next** and **Previous** buttons.

You can limit your search to a subset of the features by including a **Feature** type or a **Qualifier** type in your search.

You can also build more complex queries, by searching only those features that are already selected.

To perform a feature search

1. Choose **Edit | Find** from the main menu.

The **Find** dialog box is displayed.

2. Ensure that the **Feature** tab is selected.

This tab is selected by default if the **Find** dialog box is opened when a Sequence Feature view or a Sequence Map view is active.

3. Type the term that you want to find in the **Find** text box.
4. Optionally, restrict your search to a particular **Feature** type by selecting it from the list.

5. Optionally, restrict your search to a particular **Qualifier** type by selecting it from the list.
6. Do one of the following:
   - to select all features that match the search parameters, use **All**
   - to scroll through the features that match the search parameters one at a time, use **Next** and **Previous**

Searching results

The results search enables you to find specified text within the results windows from analyses such as Blast and Align to Folder.

To perform a results search

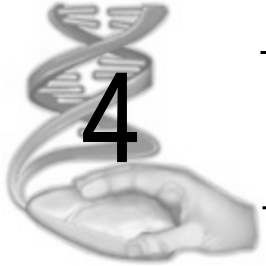
1. Choose **Edit | Find** from the main menu.

The **Find** dialog box is displayed.

2. Ensure that the **Results** tab is selected.

This tab is selected by default if the **Find** dialog box is opened when Results window is active.

3. Type the term that you want to find in the **Find** text box.
4. Check **Ignore Case** to make your search insensitive to the case of any matching text.
5. Check **Wrap Around** to make your search locate matching text which straddles more than one line.
6. Scroll through the search results one at a time, using **Next** and **Previous**.



4 Working with MacVector Files

Overview

This chapter introduces all the file types that are used by MacVector.

It also describes how the following file types can be modified:

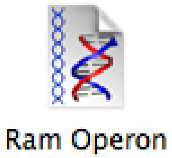
- enzyme files
- subsequence files
- scoring matrix files

Chapter 5, *“Working with Sequences and Features”* describes how to manage and edit sequence files.

Contents

MacVector file types

Sequence files

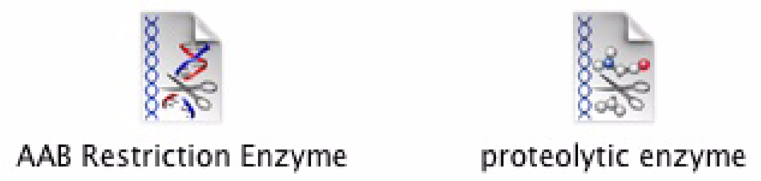


These two icons are used for nucleic acid and protein sequence files. They are in a binary file format and can only be opened and edited within MacVector.

MacVector can also open nucleic acid and protein sequences in all major ASCII formats. See Appendix G, *“Supported File Formats and File Extensions”* for a complete list of the sequence formats MacVector supports.

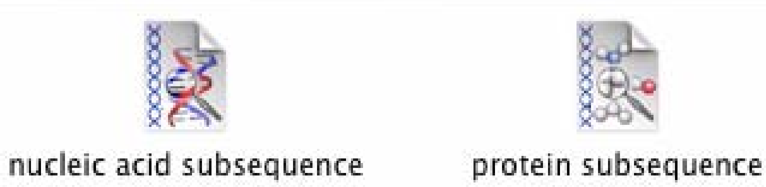
For more information about sequence files, see *“Sequence files”* on page 72, and Chapter 5, *“Working with Sequences and Features”*.

Enzyme files



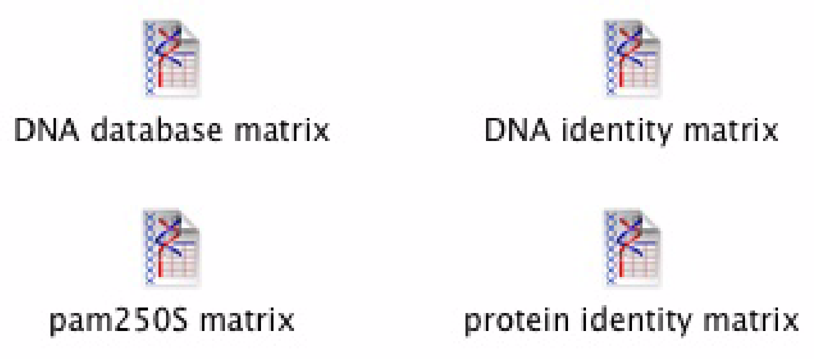
These icons are used to represent restriction enzyme and proteolytic enzyme files. They are in a binary file format and can only be opened and edited within MacVector.

Subsequence files



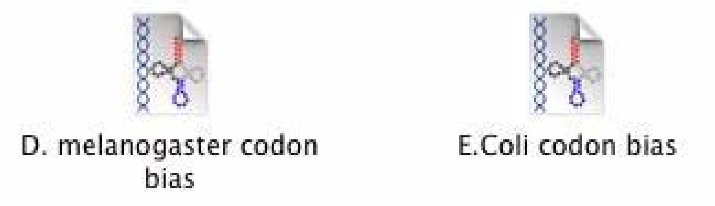
These file icons are used for nucleic acid and protein subsequence files. They are in a binary file format and can only be opened and edited within MacVector.

Scoring matrix files



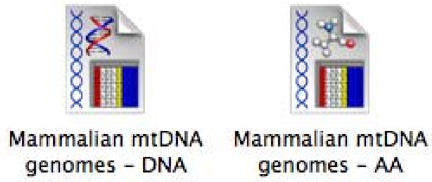
These icons are used to represent the nucleic acid and protein scoring matrix files that are used in the matrix comparisons and database searches. They are in a binary file format and can only be opened and edited within MacVector.

Codon bias files



This icon is used for files that contain codon bias tables. You cannot open and edit a codon bias file directly.

Multiple alignment files



These icons are used for files generated by MacVector that contain multiple alignments of DNA or protein sequences. They can be opened and edited within MacVector.

Managing files

This section describes some general procedures for managing files:

- creating a file
- opening a file
- saving a file
- closing a file

Creating new files

You can create new files of the following MacVector file types:

- restriction enzyme or proteolytic enzyme files
- nucleic acid or protein sequence files
- nucleic acid or protein subsequence files
- nucleic acid or protein multiple alignment files
- nucleic acid or protein scoring matrix files

Codon bias files are created by a different process. See Chapter 10, *“Primer and Probe Design”* for further details.

To create a new file

1. Choose **File | New** from the menu.
2. Choose a file type from the submenu.

An untitled window of the appropriate type is displayed.

Alternatively, press **<⌘>** + **N** on the keyboard to create a new nucleic acid sequence file.

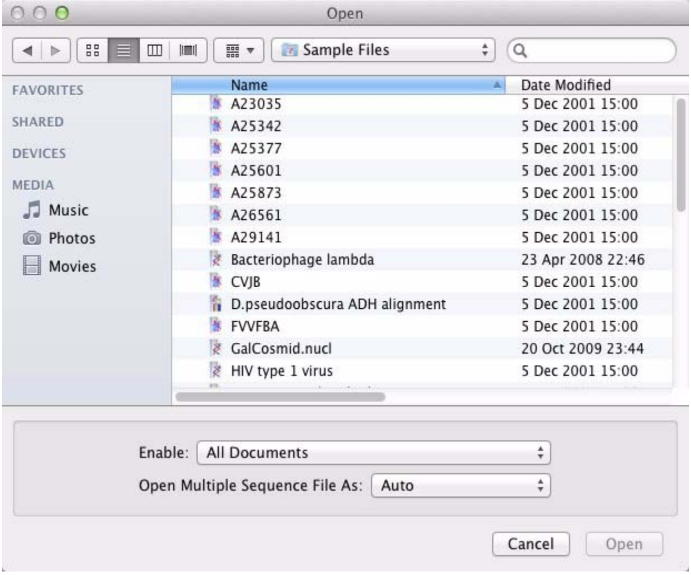
Opening files

You can open any of the following types of files for editing:

- restriction enzyme or proteolytic enzyme files
- nucleic acid or protein sequence files
- nucleic acid or protein subsequence files
- nucleic acid or protein multiple alignment files
- nucleic acid or protein scoring matrix files

Codon bias files can only be accessed during codon preference analysis. They cannot be opened for editing.

You can open any file that MacVector recognizes using the **Open** dialog box.



Name	Date Modified
A23035	5 Dec 2001 15:00
A25342	5 Dec 2001 15:00
A25377	5 Dec 2001 15:00
A25601	5 Dec 2001 15:00
A25873	5 Dec 2001 15:00
A26561	5 Dec 2001 15:00
A29141	5 Dec 2001 15:00
Bacteriophage lambda	23 Apr 2008 22:46
CVJB	5 Dec 2001 15:00
D.pseudoobscura ADH alignment	5 Dec 2001 15:00
FVVFBA	5 Dec 2001 15:00
GalCosmid.nucl	20 Oct 2009 23:44
HIV type 1 virus	5 Dec 2001 15:00

Enable: All Documents

Open Multiple Sequence File As: Auto

To open a file

1. Choose **File | Open**.

The **Open** dialog box is displayed.

2. To restrict the types of files displayed, click the **Enable** arrow button and choose the required file type from the drop-down list.
3. Browse to the appropriate directory and choose the required file from the list. You can also use the search box at the top right of the dialog box to help you locate your files.
4. To open more than one file, hold down the **<shift>** key and continue to choose files form the list. (A second **<shift>**-click on the same file will deselect it.)
5. If you have selected a file containing multiple sequences, the **Open Multiple Sequence Files As:** menu enables you to choose whether to open each sequence in a separate window, or all the sequences

aligned, in a single window. See *"Opening multiple sequences"* on page 312 for details.

6. Select **Open** to open the file(s) and close the dialog box.
7. The files are displayed in windows of the appropriate type.

Saving files

When files are modified, the entries take immediate effect on the copy in memory. The changes are only made permanent when you use either **File | Save** or **File | Save As**.

To save a file

1. Ensure the required window is active.
2. Do one of the following:
   - choose **File | Save** to save changes to the file.
   - choose **File** | **Save As** to save the changes as a new file. If you do this, you must provide a new name for the file in the **Save As** edit box, before saving the file using the **Save** button.

Closing files

To close a file

1. Choose **File | Close**.

The file and its associated window are closed.

Enzyme files

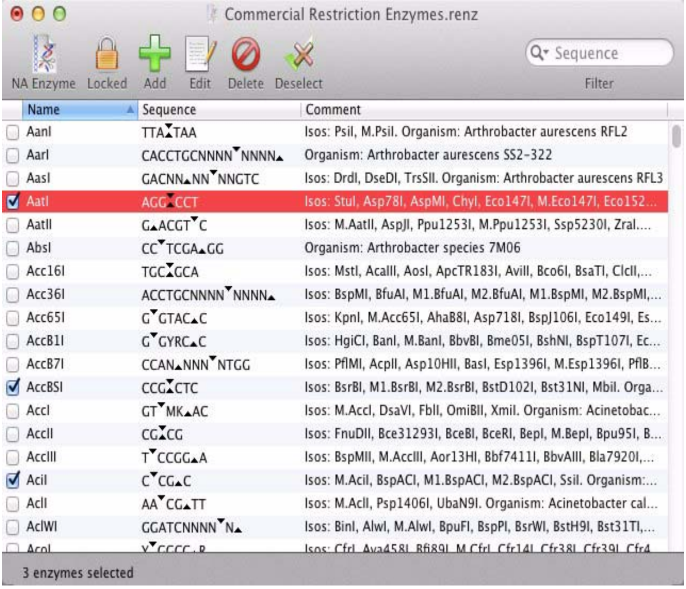
Enzyme files contain a list of enzymes and details of the recognition site they require for cleaving a sequence.

Restriction enzymes (for cleaving nucleic acid sequences) and proteolytic enzymes (for cleaving amino acid sequences) are stored in separate files. MacVector is supplied with several enzyme files. The restriction enzyme files are a mirror of the REBASE database of restriction enzymes.

You can edit restriction enzyme and proteolytic enzyme files. MacVector enables you to:

- select a subset of enzyme entries and save the selection for use with the restriction enzyme or proteolytic enzyme analyses
- delete existing enzyme entries from the file
- change the data for existing enzyme entries

- add new enzyme entries to the file



	Name	Sequence	Comment
□	AanI	TTA⧗TAA	Isos: PsiI, M.PsiI. Organism: Arthrobacter aurescens RFL2
□	AarI	CACCTGCNNNN▼NNNN▲	Organism: Arthrobacter aurescens SS2-322
□	AasI	GACNN▲NN▼NNGTC	Isos: DrdI, DseDI, TrsSII. Organism: Arthrobacter aurescens RFL3
☑	AatI	AGG⧗CCT	Isos: StuI, Asp78I, AspMI, ChyI, Eco147I, M.Eco147I, Eco152...
□	AatII	G▲ACGT▼C	Isos: M.AatII, AspJI, Ppu1253I, M.Ppu1253I, Ssp5230I, ZraI....
□	AbsI	CC▼TCGA▲GG	Organism: Arthrobacter species 7M06
□	Acc16I	TGC⧗GCA	Isos: MstI, AcaIII, AosI, ApcTR183I, AviII, Bco6I, BsaTI, ClcII,...
□	Acc36I	ACCTGCNNNN▼NNNN▲	Isos: BspMI, BfuAI, M1.BfuAI, M2.BfuAI, M1.BspMI, M2.BspMI,...
□	Acc65I	G▼GTAC▲C	Isos: KpnI, M.Acc65I, AhaB8I, Asp718I, BspJ106I, Eco149I, Es...
□	AccB1I	G▼GYRC▲C	Isos: HgiCI, BanI, M.BanI, BbvBI, Bme05I, BshNI, BspT107I, Ec...
□	AccB7I	CCAN▲NNN▼NTGG	Isos: PflMI, AcpII, Asp10HII, BasI, Esp1396I, M.Esp1396I, PflB...
☑	AccBSI	CCG⧗CTC	Isos: BsrBI, M1.BsrBI, M2.BsrBI, BstD102I, Bst31NI, MbiI. Orga...
□	AccI	GT▼MK▲AC	Isos: M.AccI, DsaVI, FblI, OmiBII, XmiI. Organism: Acinetobac...
□	AccII	CG⧗CG	Isos: FnuDII, Bce31293I, BceBI, BceRI, BepI, M.BepI, Bpu95I, B...
□	AccIII	T▼CCGG▲A	Isos: BspMII, M.AccIII, Aor13HI, Bbf7411I, BbvAIII, Bla7920I,...
☑	AciI	C▼CG▲C	Isos: M.AciI, BspACI, M1.BspACI, M2.BspACI, SsiI. Organism:...
□	AclI	AA▼CG▲TT	Isos: M.AclI, Psp1406I, UbaN9I. Organism: Acinetobacter cal...
□	AclWI	GGATCNNNN▼N▲	Isos: BinI, AlwI, M.AlwI, BpuFI, BspPI, BsrWI, BstH9I, Bst31TI,...
□	AcoI	Y▼CCCC▲R	Isos: CfrI, Ava458I, Bfi89I, M.CfrI, Cfr14I, Cfr38I, Cfr39I, Cfr4

3 enzymes selected

Selecting enzymes for site analysis

You can choose a set of enzymes to use for a site analysis. This may be useful, for example, when you want to restrict your analysis to those enzymes readily available in your laboratory.

To select a subset of enzymes in an enzyme file

1. Open the required enzyme file, and ensure that the file window is active.

A selection of enzyme files is available in the Restriction Enzymes folder.

2. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
3. Scroll through the enzyme list to find the name of each enzyme you want to select.

Tip. You can navigate through the list by typing the first character of the restriction enzyme, or by using the up and down arrows on the keyboard.

4. Click on each required enzyme.

   A check mark appears, and a line at the bottom of the window indicates how many enzymes have been selected.

Tip. Use the **clear selection** button to deselect all entries.

5. Save the file to disk, using **File | Save**.

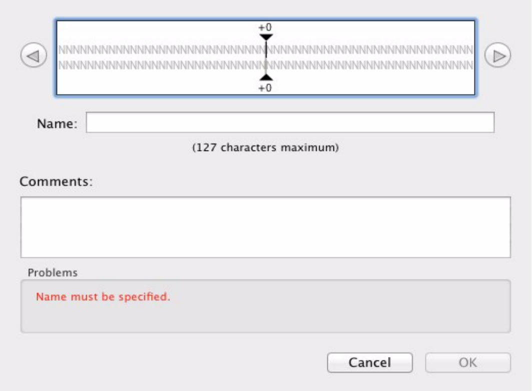
   Your selections are saved and you will not have to repeat the selection process every time you open the file.

Note. When you perform a restriction or proteolytic site search, you may use either all entries in the enzyme file, or only those enzymes selected.

Tip. If you save the changes using **File | Save As**, you can save different selection subsets.

Adding entries to an enzyme file

Entries can be added to an enzyme file at any time that the window for that file is active.



Name:

(127 characters maximum)

Comments:

Problems

Name must be specified.

To add a new entry to an enzyme file

1. Open the required enzyme file and ensure that the file window is active.
2. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
3. Click the **Add** button on the toolbar.

The **Enzyme Editor** dialog box is displayed.

4. Type a name for the enzyme in the **Name** text box.

5. Click the sequence representation in the center of the dialog box to position the insertion point, then type the recognition site using the IUPAC one-letter codes.

For proteins, enter the recognition sequence in the amino to carboxy direction; for nucleic acids, in the 5' to 3' direction.

Note. Use parentheses to enclose a combination of amino acids for which there is no one-letter assignment. For example, if the agent cleaves after an aspartic (D) or glutamic (E) acid residue, denote it as (DE).

6. Drag the arrow-shaped control to the cut position.

For restriction enzymes, there are two arrows: the upper one to indicate the cut position on the plus strand, the lower one to indicate the cut position on the minus strand.

7. If required, you can enter a comment up to 254 characters long.
8. Select **OK** to add the enzyme to the file.
9. Choose **File | Save** to make the changes permanent.

Copying entries between enzyme files

Copying existing entries between files can be useful, for example to create a customized enzyme subset from more than one file. This requires that one or more enzyme files are open at the same time, and that the file receiving the enzyme entries is unlocked.

To copy entries between files

1. Make the file from which you are copying entries the active window.
2. Select the enzymes that you want to copy using one of the following methods:
   - click to select a single entry
   - hold down the mouse button and drag to select a continuous block of entries
   - hold the <**shift>** key in combination with either of the above to retain already selected entries.
3. Choose **Edit | Copy**.
4. Make the file that is receiving the entries the active window.
5. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
6. Choose **Edit | Paste**.

The entries are added to the file in alphabetical order.

7. Choose **File | Save** to make the changes permanent.

Note. If any of the entries were marked as selected, the selection will be retained.

Editing entries in an enzyme file

Entries in an enzyme file can be modified. This is generally useful only when a mistake has been noticed in the existing entry.

To edit an enzyme entry

1. Open the required enzyme file and ensure that the file window is active.
2. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
3. Select the enzyme that you want to edit.
4. Click the **Edit** button on the toolbar.

The **Enzyme Editor** dialog box is displayed.

5. Modify the entry as required.

See *"To add a new entry to an enzyme file"* on page 60, for further details.

6. Select **OK** to add the changes to the enzyme file.
7. Choose **File | Save** to make the changes permanent.

Deleting entries from an enzyme file

Entries in an enzyme file can be deleted when required.

To delete an enzyme entry

1. Open the required enzyme file, and ensure that the file window is active.
2. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
3. Select the enzymes that you want to delete using one of the following methods:
   - click to select a single entry
   - hold down the mouse button and drag to select a continuous block of entries
   - hold the <**shift**> key in combination with either of the above to retain already selected entries

4. Click the **Delete** button on the toolbar.

The selected entries are deleted from the file.

5. Choose **File | Save** to make the changes permanent.

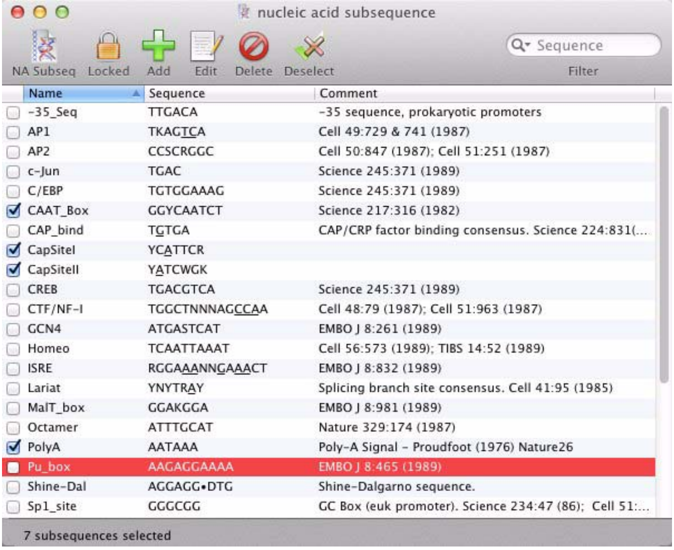
Note. This is equivalent to using **Edit | Clear**. If you want to paste the deleted entries into another file, use **Edit | Cut** instead of the **Delete** button.

Subsequence files

Subsequence files contain a list of motifs that can be used to look for structural and functional patterns within sequences. A subsequence may consist of up to three parts, with the permitted gap between each part defined, as well as the allowed mismatch for each part in a search. The logic operators OR and AND can be used to specify whether any or all of the parts need to be found in a sequence for it to be accepted.

MacVector is supplied with several subsequence files for both proteins and nucleic acids. You can edit these to:

- select a subset of entries and save the selection for use with the subsequence analyses
- delete existing entries from the file
- change the data for existing entries
- add new entries to the file



	Name	Sequence	Comment
☐	-35_Seq	TTGACA	-35 sequence, prokaryotic promoters
☐	AP1	TKAGTCA	Cell 49:729 & 741 (1987)
☐	AP2	CCSCRGGC	Cell 50:847 (1987); Cell 51:251 (1987)
☐	c-Jun	TGAC	Science 245:371 (1989)
☐	C/EBP	TGTGGAAAG	Science 245:371 (1989)
☑	CAAT_Box	GGYCAATCT	Science 217:316 (1982)
☐	CAP_bind	TGTGA	CAP/CRP factor binding consensus. Science 224:831(...
☑	CapSiteI	YCATTCR	
☑	CapSiteII	YATCWGK	
☐	CREB	TGACGTCA	Science 245:371 (1989)
☐	CTF/NF-I	TGGCTNNNAGCCAA	Cell 48:79 (1987); Cell 51:963 (1987)
☐	GCN4	ATGASTCAT	EMBO J 8:261 (1989)
☐	Homeo	TCAATTAAAT	Cell 56:573 (1989); TIBS 14:52 (1989)
☐	ISRE	RGGAAANNGAAACT	EMBO J 8:832 (1989)
☐	Lariat	YNYTRAY	Splicing branch site consensus. Cell 41:95 (1985)
☐	MalT_box	GGAKGGA	EMBO J 8:981 (1989)
☐	Octamer	ATTTGCAT	Nature 329:174 (1987)
☑	PolyA	AATAAA	Poly-A Signal - Proudfoot (1976) Nature26
☐	Pu_box	AAGAGGAAAA	EMBO J 8:465 (1989)
☐	Shine-Dal	AGGAGG•DTG	Shine-Dalgarno sequence.
☐	Sp1_site	GGGCGG	GC Box (euk promoter). Science 234:47 (86); Cell 51:...

7 subsequences selected

Selecting subsequences for searches

You can choose a set of subsequences to use for a search. This may be useful, for example, when you want to restrict your search to particular functional patterns.

To select a subset of subsequences

1. Open the required subsequence file and ensure that the file window is active.
2. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
3. Scroll through the subsequence list to find the name of each subsequence you want to select.

Tip. You can navigate through the list by typing the first character of the subsequence, or by using the up and down arrows on the keyboard.

4. Click on each required subsequence.

   A check mark appears and a line at the bottom of the window indicates how many subsequences have been selected.

Tip. Use the **clear selection** button to deselect all selected entries.

5. Save the file to disk, using **File | Save**.

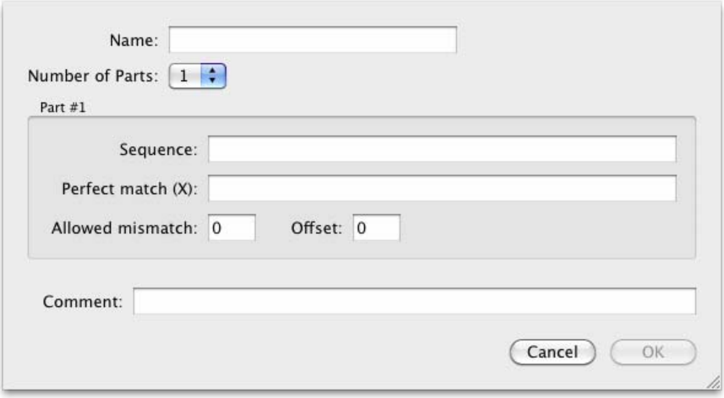
   Your selections are saved and you will not have to repeat the selection process every time you open this file.

Note. When you perform a search, you may use either all entries in the subsequence file, or only those subsequences selected.

Tip. If you save the changes using **File | Save As**, you can save different selection subsets.

Adding entries to a subsequence file

Entries can be added to a subsequence file at any time that the window for the file is active.



Number of Parts: 1

Allowed mismatch: 0

Offset: 0

To add a new entry to a subsequence file

1. Open the required subsequence file and ensure that the file window is active.
2. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
3. Click the **Add** button on the toolbar.

The **Subsequence Editor** dialog box is displayed.

4. Type a name for the subsequence in the **Name** text box.
5. Select the **No. of parts** radio button that corresponds to how many parts the subsequence will have.
6. In the **Part #** text boxes, type the sequence for each part of the subsequence, using standard IUPAC one-letter codes.

Note. Use parentheses to enclose a combination of amino acids for an ambiguous residue. For example, if either an aspartic (D) or glutamic (E) acid residue can occur at a given position, denote it as (DE). However, for an ambiguous nucleic acid residue, use standard IUPAC symbols, not parentheses.

7. Use the **Perfect Match** and **Allowed mismatch** text boxes to adjust the tolerance of the match as follows:
   - to specify an exact match, type `0` in **Allowed mismatch** and leave **Perfect match** empty

- to allow some degree of mismatching, type the maximum number of mismatching residues that you will accept in the **Allowed mismatch** text box
- to require that specific residues match exactly, type X beneath those residues in the **Perfect match** box, using spaces to position the Xs. For amino acid motifs, only one X can be placed beneath each group of amino acids enclosed in parentheses. The allowed mismatch cannot exceed the length of the sequence for that part of the subsequence.

8. If you want to add an offset to the reported position of the subsequence when a search is performed, type a value in the **Offset** text box.

Usually, the number of the first residue in the subsequence is reported.

9. If required, you can enter a comment up to 254 characters long.
10. For subsequences with more than one part, you can treat each part of the subsequence as if it were a separate motif:
    - select the **Logic: or** radio button to report a subsequence found if any combination of the subsequence parts is found
    - select the **Logic: and** radio button to report a subsequence found only if all subsequence parts are found.
11. For subsequences with more than one part, select one of the **Offset part:** radio buttons to report the position of the subsequence relative to any of the parts.

This option appears only if you are using AND logic. It is used in conjunction with the **Offset** value. If you select **1** as the offset part and type 0 as the offset, MacVector will report the position of the first residue of the first part of the subsequence. If you select **3** as the offset part and type 4 as the offset, MacVector will report the location of the subsequence to be the fifth residue of the third part of the subsequence.

12. For subsequences with more than one part, use the **Gap** text boxes to specify the limits on the number of residues that may occur between each part of the subsequence.

This option appears only if you are using AND logic. For example, if your subsequence consists of two parts which may be separated by 10 to 20 residues, you would type 10 and 20 in the **#1 - #2 Gap** text boxes. The gap values can be any number from zero upwards, and the second number must be equal to or larger than the first.

13. Select **OK** to add the subsequence to the file.
14. Choose **File | Save** to make the changes permanent.

Copying entries between subsequence files

Copying existing entries between files can be useful, for example to create a customized subsequence subset from more than one file. One or more subsequence files must be open at the same time, and the file receiving the subsequence entries must be unlocked.

To copy entries between files

1. Make the file from which you are copying entries the active window.
2. Select the subsequences that you want to copy:
   - click to select a single entry
   - hold down the mouse button and drag to select a continuous block of entries
   - hold the <**shift>** key in combination with either of the above to retain already selected entries.
3. Choose **Edit | Copy**.
4. Make the file that is receiving the entries the active window.
5. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
6. Choose **Edit | Paste**.

The entries are added to the file in alphabetical order.

7. Choose **File | Save** to make the changes permanent.

Note. If any of the entries were marked as selected, the selection will be retained.

Editing entries in a subsequence file

Entries in a subsequence file can be modified. This is generally useful only when a mistake has been noticed in the existing entry.

To edit a subsequence entry

1. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
2. Click the **Edit** button on the toolbar.

The **Subsequence Editor** dialog box is displayed.

3. Modify the entry as required.

See *"To add a new entry to a subsequence file"* on page 65, for further details.

4. Select **OK** to add the changes to the subsequence file.
5. Choose **File | Save** to make the changes permanent.

Deleting entries from a subsequence file

Entries in a subsequence file can be deleted when required.

To delete a subsequence entry

1. If the file is locked, unlock it by clicking the **Locked** icon in the toolbar.
2. Select the subsequences that you want to delete:
   - click to select a single entry
   - hold down the mouse button and drag to select a continuous block of entries
   - hold the <**shift>** key in combination with either of the above to retain already selected entries.
3. Click the **Delete** button on the toolbar.

The selected entries are deleted from the file.

4. Choose **File | Save** to make the changes permanent.

Note. This is equivalent to using **Edit | Clear**. If you want to paste the deleted entries into another file, use **Edit | Cut** instead of the **Delete** button.

Scoring matrix files

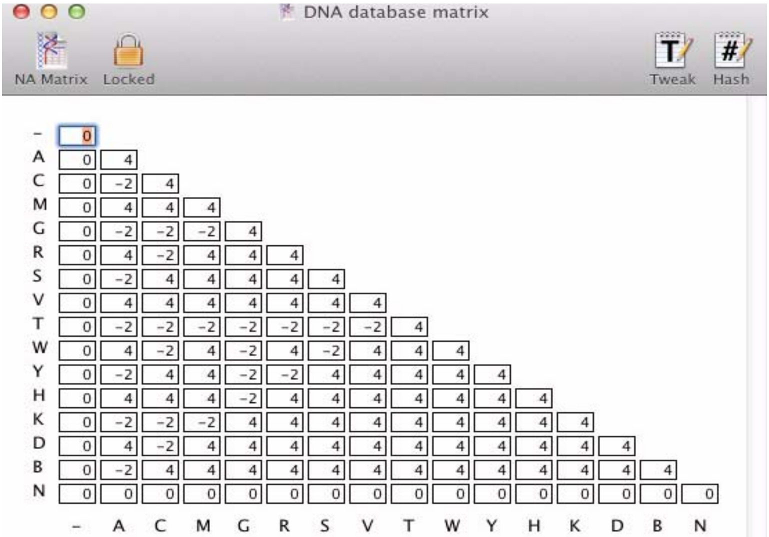
You can make the following modifications to the scoring matrix files to customize both the matrix comparisons and database searches:

- change the match/mismatch scores in the scoring matrix
- change the hash codes used during the hashing step
- change the deletion penalty
- change the gap penalty
- change the parameters used to calculate the cut-off score for performing an optimized alignment

The default values for the above parameters in the scoring matrix files supplied with MacVector are listed in Appendix C, *"Reference Tables"*. Modifications should be made with caution.

Editing match and mismatch scores

The scoring matrix is displayed as a lower-triangular matrix. The number located at coordinate (H,V) in the matrix is the score that is assigned for an alignment between the two residues H and V.



-	0															
A	0	4														
C	0	-2	4													
M	0	4	4	4												
G	0	-2	-2	-2	4											
R	0	4	-2	4	4	4										
S	0	-2	4	4	4	4	4									
V	0	4	4	4	4	4	4	4								
T	0	-2	-2	-2	-2	-2	-2	-2	4							
W	0	4	-2	4	-2	4	-2	4	4	4						
Y	0	-2	4	4	-2	-2	4	4	4	4	4					
H	0	4	4	4	-2	4	4	4	4	4	4	4				
K	0	-2	-2	-2	4	4	4	4	4	4	4	4	4			
D	0	4	-2	4	4	4	4	4	4	4	4	4	4	4		
B	0	-2	4	4	4	4	4	4	4	4	4	4	4	4	4	
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-	A	C	M	G	R	S	V	T	W	Y	H	K	D	B	N

From left to right across the horizontal axis, and from top to bottom down the vertical axis, the order for nucleotide codes is:

- A C M G R S V T W Y H K D B N

and for amino acid codes is:

- A C D E F G H I K L M N P Q R S T V W Y B Z X *

The match and mismatch scoring is symmetric, so any change to the score for a pair (H,V) is automatically applied to (V,H) in the matrix.

To edit the match/mismatch scores

1. Open the scoring matrix that you want to edit.

A window opens, displaying the lower-triangular scoring matrix.

2. Select the value to be edited using one of the following methods:

   - click directly on a value
   - type the first letter of a pair, then hold down the **<shift>** key and type the second letter.

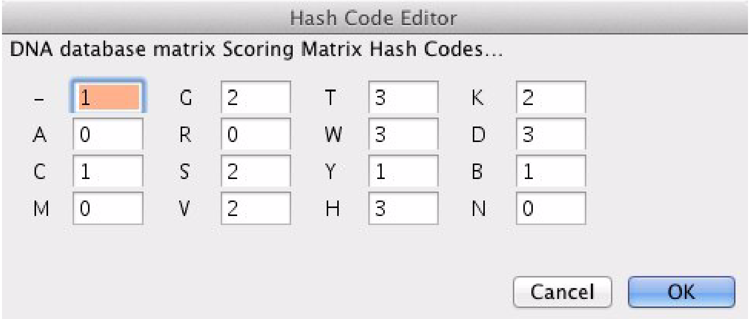
3. Type the new score (it may be positive or negative) and press the **Enter** key.

The scores are symmetrical, so if you change the score for the aligned pair A by C, the score for the pair C by A is changed simultaneously. Values assigned to the scoring matrix must lie between -99 and 99.

4. Select **OK** to save the changes.
5. Choose **File | Save** to make the changes permanent.

Editing hash codes

All residues that have the same hash code are treated as identical residues by the program. As an example, look at the hash codes for the nucleic acid scoring matrix DNA database matrix. The ambiguous bases W (which stands for A or T) and D (not C) are arbitrarily assigned the same hash code as T. There are 21 possible hash codes for amino acids (0 to 20) and four possible for nucleic acids (0 to 3).



Hash Code Editor

DNA database matrix Scoring Matrix Hash Codes...

–	1	G	2	T	3	K	2
A	0	R	0	W	3	D	3
C	1	S	2	Y	1	B	1
M	0	V	2	H	3	N	0

To edit the hash codes

1. Open the scoring matrix that you want to edit.

A window opens, displaying the lower-triangular scoring matrix.

2. Click the **Hash** button on the toolbar.

The **Hash Code Editor** dialog box is displayed.

3. Type values in the text boxes as required to make the residue types equivalent.
4. Select **OK** to save the changes.
5. Choose **File | Save** to make the changes permanent.

Editing tweak values

The tweak values are used to evaluate alignment scores, and can be edited if required.

The cut-off score parameters are substituted into an equation that is used to calculate the minimum score that an initial alignment must have before an optimized alignment is performed. Ordinarily, these parameters should be left at their default values. Accepted values are from 1 to 200.

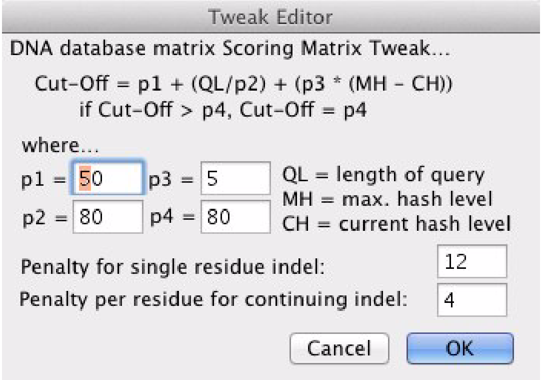
The deletion penalty (single residue indel) is the value subtracted from the score of an aligned region for every one-residue insertion or deletion that was introduced in order to improve the alignment.

If an insertion or deletion longer than one residue was introduced into an alignment, the score is reduced by the gap penalty (continuing indel) times the number of bases in the insertion or deletion after the first one.

Accepted values for the deletion and gap penalties are from 0 to 100. By altering the values for these penalties, you can control the number and size of the insertions or deletions that you will permit when aligning two sequences.

A large deletion penalty coupled with a small gap penalty will make it difficult to introduce an insertion or deletion into an alignment, but will make it fairly easy to extend one when it is introduced. Alignments performed using this combination will not contain many insertions or deletions, but those that do occur may be fairly long. This type of setting should be used if you think that sequences in the database may contain insertions or deletions of blocks of residues when compared to your query sequence.

A small deletion penalty coupled with a large gap penalty will tend to produce alignments containing many short insertions or deletions. You would use this type of setting if you expect that differences between the sequences will involve single-residue insertions or deletions.



Tweak Editor

DNA database matrix Scoring Matrix Tweak...

Cut-Off = p1 + (QL/p2) + (p3 * (MH - CH))
if Cut-Off > p4, Cut-Off = p4

where...

p1 = 50 p3 = 5 QL = length of query
p2 = 80 p4 = 80 MH = max. hash level
CH = current hash level

Penalty for single residue indel: 12

Penalty per residue for continuing indel: 4

To edit tweak values

1. Open the scoring matrix that you want to edit.

A window opens, displaying the lower-triangular scoring matrix.

2. Click the **Tweak** button on the toolbar.

The **Tweak Editor** dialog box is displayed.

3. Type values in the text boxes as required to modify the tweak values.
4. Select **OK** to save the changes.
5. Choose **File | Save** to make the changes permanent.

Sequence files

MacVector sequence files are binary files, rather than ASCII text files. This means that you cannot read them using a word processor or text editor. MacVector stores sequence file annotation and features table data, as well as the sequence data itself. By using binary files that can be edited only with MacVector, we can ensure that the files will be properly formatted.

MacVector provides features table support for files in GenBank format (or GenBank variants). As new features are added to the GenBank format over time, support for them is included in upgrades of MacVector. See Appendix E, *“GenBank Feature Tables”* for details of the feature keywords supported in this version of MacVector.

When non-GenBank files are imported, MacVector creates a GenBank-style annotation section. Any nonsequence information in the original files is placed in the annotation section under Comment, so no information is lost.

When directed to open a file of type TEXT, MacVector parses it to see if it matches one of the supported file formats. If there is no match, it assumes that the file is sequence only.

MacVector reads the molecule type (nucleic acid or protein) directly from the file, if possible. When it encounters a file format that does not contain this information (such as line files, Staden files, and FastA files) MacVector displays an alert box for you to indicate whether the file contains a nucleic acid or a protein sequence.

See Appendix G, *"Supported File Formats and File Extensions"* for a complete list of the sequence formats MacVector supports.

For details of how to manage and edit sequence files, see Chapter 5, *"Working with Sequences and Features"*.

5 Working with Sequences and Features

Overview

This chapter describes the Sequence window and how to work with sequence files, including:

- opening, creating, saving and printing sequence files
- editing sequences, features and annotations

Multiple sequences are described in the series of chapters beginning with Chapter 12, *"Aligning Sequences"*.

Contents

Sequence window

The Sequence window is displayed whenever a sequence is opened or created. It comprises a context dependent toolbar and four tabbed and linked views of the sequence: Editor, Map, Features and Annotations. The functionality available in each view is described in the sections below.

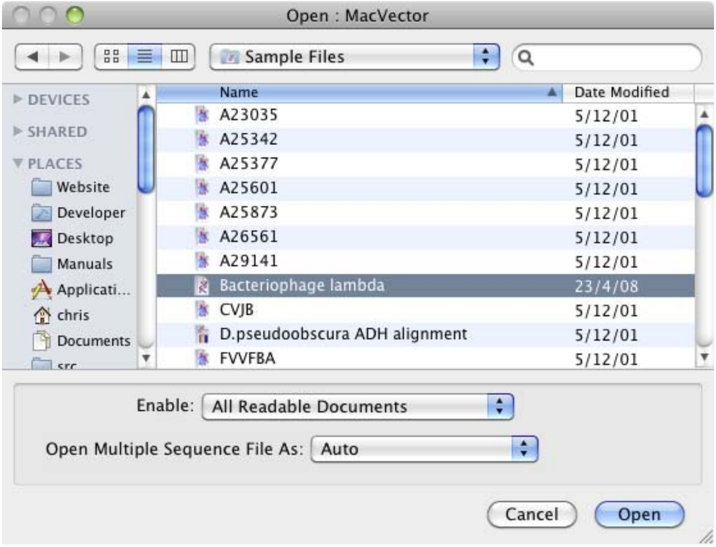
Managing sequences

This section describes some general procedures for managing sequences:

- opening a sequence
- creating a sequence
- saving a sequence
- exporting a sequence view
- printing a sequence
- closing a sequence

Opening a sequence

You can use this procedure to open any sequence file that MacVector recognizes.



Name	Date Modified
A23035	5/12/01
A25342	5/12/01
A25377	5/12/01
A25601	5/12/01
A25873	5/12/01
A26561	5/12/01
A29141	5/12/01
Bacteriophage lambda	23/4/08
CVJB	5/12/01
D.pseudoobscura ADH alignment	5/12/01
FVVFBA	5/12/01

Enable: All Readable Documents

Open Multiple Sequence File As: Auto

To open a sequence

1. Choose **File | Open** from the menu.

The **Open** dialog box is displayed.

2. To choose a file type to view, select the **Show** arrow button and choose the required file type from the drop-down list.
3. Choose the required file from the list.
4. To open more than one file, hold down the **<shift>** key and continue to choose files from the list. (A second **<shift>** click on the same file will deselect it.)
5. If you have selected a file containing multiple sequences, the **Open Multiple Sequence Files As:** menu lets you choose whether to open each sequence in a separate Sequence window, or all the sequences aligned in a single Multiple Sequence Alignment (MSA) window. See *"Opening multiple sequences"* on page 312 for details.
6. Select **Open** to open the file(s) and close the dialog box.

Note. Regardless of the file type opened, it is converted internally into MacVector binary format. This is the default format for saving the file.

Creating a new sequence

You can create a new sequence at any time.

To create a new sequence file from scratch

1. Choose **File | New** from the menu.
2. Select the required sequence type from the submenu.

A new empty Sequence window of the selected type is displayed.

Tip. You can also create new empty sequence files for Nucleic Acids and proteins using the command key shortcuts ⌘N and **<option>**+⌘N respectively.

To create a new sequence file from the contents of the clipboard

- Copy the required sequence fragment to the clipboard by selecting it and choosing **Edit | Copy** from the menu.
- Or choose **File | New from Clipboard** from the menu.

A new Sequence window containing the copied fragment is displayed.

Saving a sequence

You can save a sequence using its existing file name or a new file name. You can save using any of the supported sequence file formats. See

Appendix G, *"Supported File Formats and File Extensions"* for a complete list of the sequence formats MacVector supports.

Note. When you save a sequence you are saving the underlying sequence data not the representation displayed in the current sequence view. See *"Exporting view-specific data"* on page 80 for information about exporting these representations.

Save

When a new sequence file is created or an existing file is opened, **File | Save** is disabled until the first change is made.

If you choose **File | Save** for a sequence that does not exist as a MacVector file, the Save As dialog box is displayed. This makes it unlikely that you will accidentally overwrite the original version of a file in another format.

To save an existing MacVector sequence

1. Ensure the required Sequence window is active.
2. Choose **File | Save** from the menu.

The file is saved as a MacVector binary file, overwriting the previous version of the file, regardless of the previous format.

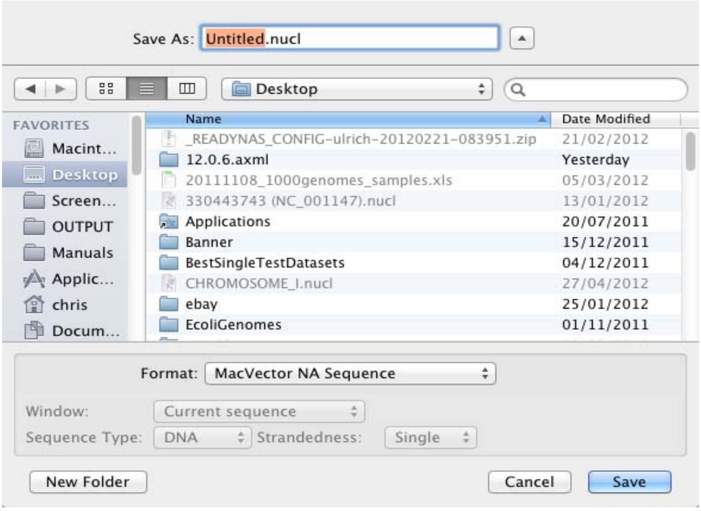
Note. If the file is newly created, **File | Save** acts like **File | Save As** and a dialog box is displayed, asking for a filename.

Save As

Use the **Save As** command when you want to:

- change the file name
- save the file to a different folder or to a different disk
- save the file in a format other than MacVector binary format
- change the molecule type or strand type of the sequence if the file is a nucleic acid sequence file

The **Save As** command is enabled whenever a Sequence window is active. If a file with the same name already exists on disk, it will be overwritten.



Save As: Untitled.nucl

Desktop

Name	Date Modified
_READYNAS_CONFIG-ulrich-20120221-083951.zip	21/02/2012
12.0.6.axml	Yesterday
20111108_1000genomes_samples.xls	05/03/2012
330443743 (NC_001147).nucl	13/01/2012
Applications	20/07/2011
Banner	15/12/2011
BestSingleTestDatasets	04/12/2011
CHROMOSOME_I.nucl	27/04/2012
ebay	25/01/2012
EcoliGenomes	01/11/2011

Format: MacVector NA Sequence

Window: Current sequence

Sequence Type: DNA Strandedness: Single

To save a sequence

1. Ensure the required Sequence window is active.
2. Choose **File | Save As** from the menu.

The **Save As** dialog box is displayed.

3. If required, navigate to the appropriate folder, or create a **New** folder.
4. If required, change the file name by typing in the **Save As** text box.
5. Choose a format from the **Format** drop-down menu.
6. If necessary, select the required **Sequence Type**.

The **Sequence Type** option is only applicable to the IG_Suite and GenBank file formats.

7. Select **Save** to save the file.

See also *“Saving multiple sequence alignments”* on page 315

The MacVector interface allows you to write the contents of all currently open sequence windows to a single file in GCG RSF, Fasta or Genbank format. (Sequences in MSA windows cannot be saved in this way.)

To save several sequences to a single file

1. Ensure that the only open sequence windows are the ones to be saved, and that a sequence window is active.
2. Choose **File | Save As** from the menu.

The **Save As** dialog box is displayed.

3. If required, navigate to a new folder.
4. If required, change the file name by typing in the **Save As** text box.
5. Choose the GCG(RSF) format from the **File Format** drop-down menu.
6. Choose **All open sequences** from the **Window** drop-down menu.
7. Select **Save** to save the file.

All currently open sequences are saved to the file.

File extensions for sequence files

When you save a sequence using the **Save As** dialog box, MacVector will by default, add a filename extension appropriate to the format you select. See Appendix G, *"Supported File Formats and File Extensions"* for a complete list of the file extensions used by MacVector.

These file extensions may be useful when files are used with other operating systems which use extensions to identify the file type.

Exporting view-specific data

You can now export the representation of the sequence displayed in the current sequence view to a file in an appropriate format, for example a PDF file for graphics displayed in the Map view or a text file for comments displayed in the Annotations view.

To export view-specific data

1. Ensure the required Sequence view is displayed.
2. Choose **File | Export** from the menu.

The **Export** dialog box is displayed.

3. If required, navigate to an alternative folder, or create a **New Folder**.
4. If required, change the file name by typing in the **Save As** text box.
5. Select **Save** to export the file.

Printing a sequence

The information in sequence files can be printed in two different ways:

- **Annotated** - an annotated sequence file in a user-customizable format
- **Separate...** - an unannotated form, containing one or more of the three parts of the sequence file (annotations, features table, and sequence).

If you print an annotated sequence, you must format the appearance first. See Chapter 3, *“General Procedures”*, for details of formatting the annotated sequence display.

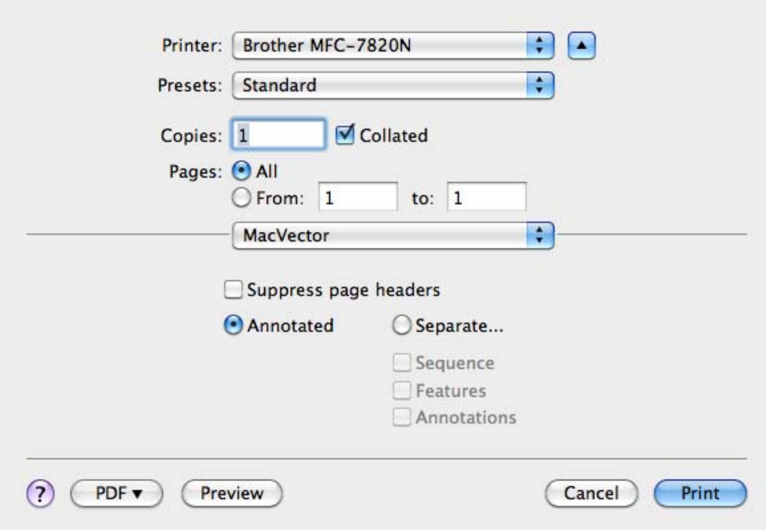
To print a sequence

1. Choose **File | Page Setup** from the menu.

The **Page Setup** dialog box is displayed

2. Adjust the paper size and orientation as required.
3. Select **OK**.
4. Choose **File | Print** from the menu.

The **Print** dialog is displayed.



Printer: Brother MFC-7820N
Presets: Standard
Copies: 1 ☑ Collated
Pages: ● All
○ From: 1 to: 1
MacVector
□ Suppress page headers
● Annotated ○ Separate...
□ Sequence
□ Features
□ Annotations

5. Optionally, check **Suppress page headers** to remove page headings from the printed output.
6. Select the required output option, from the two described above.

Note. If you select the **Separate...** option, then you must also choose which of the components **Sequence**, **Features** and **Annotations** you want to include in the printed output.

7. Click **Print**.

The specified sequence information is printed to the chosen device.

Closing a sequence

When you close a sequence that has not been saved, a message dialog box is displayed, asking whether you want to save the sequence before closing it.

To close a sequence

1. Choose **File | Close** from the menu.

The sequence closes, and the Sequence window is closed.

Editor view

The Editor view is displayed by default whenever a sequence is opened or created. To display it at any other time, click on the **Editor** tab in the Sequence window.

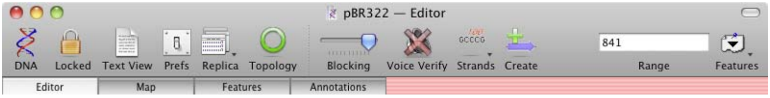
It enables you to perform the following operations on sequence data that appears in the Sequence window:

- insert or delete individual residues or groups of residues using the keyboard
- transfer blocks of residues within a sequence or between sequences of the same type using cut, copy, and paste operations
- reverse, complement, or reverse complement a sequence or a portion of a sequence
- display the complementary strand and the 3 and 6 frame translations underneath a sequence
- verify your entries by having MacVector say the name of each residue as you enter it
- proofread the sequence by directing MacVector to read it back to you after you have finished entering it

Note. Sequences can be modified only if they are unlocked. This is done by clicking on the **Locked** indicator icon in the toolbar.

If the sequence file has entries in the features table, any features that are affected by changes made to the sequence are updated automatically. In addition, cut, copy, and paste operations are automatically documented in the features table by the addition of a frag feature, which describes the operation performed.

The Editor view has a toolbar containing tools that are used to perform the particular functions, as described below.



Common tools

These tools are included by default on all of the sequence view toolbars:

For proteins, the sequence type indicator icon is **Protein**. For nucleic acids, the sequence type indicator icon is either **DNA** or **RNA**. Clicking the icon changes how the sequences are interpreted from DNA to RNA and back again.

The **Locked/Unlocked** indicator icon displays the current lock status of the Sequence file. Clicking the icon changes the lock status from locked to unlocked and back again. Sequences can only be altered when the sequence file is unlocked.

The **Text View** icon provides access to the Text view, which contains the sequence text with features marked along its length. The format of this window controlled by the **Text Display** preferences dialog box, accessible via **MacVector | Preferences** on the main menu. See *“Formatting the Text view”* on page 124 for more information about this dialog box. The information in the window cannot be changed, but it can be highlighted and copied to the clipboard.

The **Prefs** icon provides access to the preferences dialog box, which enables you to specify general MacVector preferences, including the font used in the Editor view.

The **Replica** icon is used to create a linked copy of the current sequence file in a new Sequence window.

Note. You can choose which view is displayed by default in the replica window by selecting the view name from the drop-down menu that appears when you click and hold down the **Replica** icon.

The **Topology** indicator icon shows which topological representation is currently being used in analyses of the sequence. Clicking the icon changes the topological representation from linear to circular and back again. This icon is displayed only if the Sequence window contains a nucleic acid sequence.

Editor view specific tools

These tools only appear by default in the Editor view toolbar:

The **Blocking** icon is a horizontal slider used to control the number of residues in a block. Drag the control with the mouse to change the number of residues in a block from 1 to 10.

The **Voice Verify** indicator icon displays the current voice verification status of the Sequence file. Clicking the icon changes the voice verification status from off to on and back again. When voice verification is on, each residue you enter is spoken by a computer voice. Refer to *"Using the proofreader"* on page 46, for further information.

The **Strands** icon enables you to display the complementary strand and the 3 and 6 frame translations of the current sequence. Click and hold down the icon to select the translations you want to display from the drop-down menu and click and release the icon to toggle the display of the complementary strand on and off. The translations, which are displayed underneath the original sequence, use the currently selected genetic code (see *"Selecting a different genetic code"* on page 242) and the amino acid letter code option specified in the **Text Display** preferences dialog.

Note. You can display the codons for the currently selected genetic code in a popup reference table at any time using the **Window | Genetic Code Key** menu option.

The color of the complementary strand can be set using the **Colors** preferences dialog. This icon is displayed only if the Sequence window contains a nucleic acid sequence.

The **Create** icon provides access to the **Feature Editor** dialog box which you can use to add new features to the sequence. Refer to *"Adding a feature"* on page 115, for further information.

The **Range** text box displays the number or range of the current selection within the displayed sequence. You can edit the values in the box to change the selection. Refer to *"Selection range"* on page 85, for further information.

You can also alter the selected residues by using the **Features** drop-down menu. Choose from the list of sequence features or select the entire sequence.

To change the font used in the Editor view

1. Choose **MacVector | Preferences** from the menu, then click the **Fonts** icon on the preferences dialog, or click the **Prefs** icon on the toolbar when the **Editor** tab is selected.

The **Fonts** preferences dialog box is displayed.

2. In the **Editor window font** panel, use the drop down menus to change the font and size.
3. Select **Apply** to save your changes.

To change the default colors used in the Editor view

1. Choose **MacVector | Preferences** from the menu, then click the **Colors** icon on the preferences dialog.

The **Colors** preferences dialog box is displayed.

2. Specify custom colors for the display of:
   - The **Reference** and **Complement** strands
   - The sequence **Numbering** scheme
   - The outlines used to highlight the **Left** and **Right End** selections on the click cloning ligation pop-up window (see *"Click Cloning"* on page 188).
   - The **Traces** used in chromatogram displays
3. Select one of the following options from the **Features** drop-down list to specify how features are colored in the Editor view:
   - **No Color** - no color is used to highlight features in the Editor view.
   - **Color Residues** - feature residues are colored with the same color used to identify the feature in the Map view.
   - **Color Background** - the background of feature residues is colored with the same color used to identify the feature in the Map view.
4. Select **Apply** to save your changes or **Defaults** to restore the default colors.

You can also specify a background color for a selected range in the Editor view. See *"To specify a new background color for a selected range"* on page 89.

Selection range

The **Range** text box shows one of the following:

- where the insertion point is located in the sequence

- the selected residue range

When you click within the content portion of the sequence window, a blinking vertical bar will appear within the sequence to mark the insertion point. The number of the residue immediately after the insertion point is displayed in the selection state window.

When one or more residues in the sequence are highlighted, the numbers of the beginning and ending residues of the block are displayed.

Clicking inside the **Range** box at any time enables you to alter the residue selection, by typing a new residue range or number.

Tip. If you intend to copy the updated residue selection to the clipboard, ensure that you click **<enter>**, **<tab>** or **<esc>** after you have typed in new range values to make the Sequence window, rather than the **Range** box, the active window.

Setting the origin

The plus origin of the sequence is indicated by a small red plus sign above the sequence in the content part of the window, the minus origin by a small red minus sign.

MacVector allows you to set the origin of the sequence to be a location other than the beginning or end of the sequence, so you can conform to the negative numbering system often used when referring to sequence data upstream from some significant feature.

Note. It is only possible to set the origin for linear seqeuences.

To set the plus origin, place the pointer over the plus sign, hold down the mouse button, and drag the plus sign to the residue that you want to be number 1. The sequence numbering is adjusted accordingly, and the numbers of features in the features table are also adjusted to reflect the new origin.

You can also specify a positive number other than one as the plus origin. This might be useful if, for example, you are working on a smaller, more manageable, region of a larger chromosome and want to retain the original numbering. To set an alternative start value for the plus origin, double-click on the red plus origin indicator and type the new start value in the pop-up window.

Note. MacVector automatically adjusts the start value of the plus origin to retain the original numbering when you copy and paste a sequence segment into a new sequence window.

To reset the plus origin to 1, double-click on the red plus origin indicator and type 1 in the pop-up window.

Using the Editor view context menu

When you right-click or **<ctrl>**-click within the Editor view a context-sensitive menu is displayed. It provides access to some or all of the following options depending on the current selection in the Editor view:

Create Feature provides access to the **Feature Editor** dialog box which you can use to add new features to the sequence. Refer to *"Adding a feature"* on page 115, for further information.

Add to Feature provides access to the **Features** drop-down menu. Selecting a feature from this menu adds the residues currently selected in the Editor view to that feature as a continuation segment. This enables you to build multi-segmented features, such as a CDS interrupted by introns.

Set Origin provides access to the origin pop-up window which enables you to specify an alternative plus origin (see *"Setting the origin"* on page 86).

Note. It is only possible to set the origin for linear seqeuences.

Reset Origin to 1 resets the plus origin to 1.

Selecting residues

For many editing operations, you need to select a region of the sequence first. This may be done by using the mouse, the selection range tools, the Map view or the features table. The features table must contain entries for selection.

Mouse selections

To select a small block of residues using the mouse

1. Position the cursor to the left of the first residue in the block.
2. Hold down the mouse button and move the cursor until all the required residues are highlighted.

The residue block is highlighted.

To select a large block of residues using the mouse

1. Position the cursor to the left of the first residue in the block, then click the mouse button.
2. Scroll through the sequence to the required end point of the block of residues.
3. Hold down the **<shift>** key and click the mouse button after the last required residue.

The residue block is highlighted.

Selection range selections

The **Range** box and **Features** drop-down menu provide a convenient method of selecting residues when either the residue number range or feature name are known.

To select residues using the selection range box

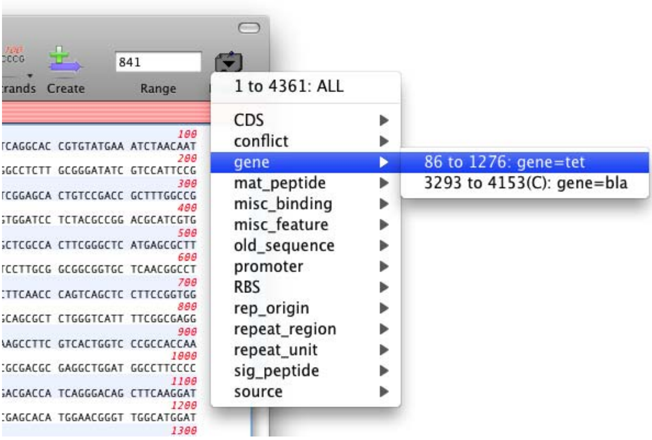
1. Click inside the **Range** text box.
2. Type a residue number or range of numbers. Use a space or a colon (:) to separate the range values.

Either the residue range is highlighted, or the cursor is positioned immediately before the residue number typed.

To select residues using the features drop-down menu

1. Select a feature from the **Features** drop-down menu.

The list contains the defined features of the sequence, plus **All 1 to n** for selecting the entire sequence.



Map view selections

To select a block of residues from the Map view

1. Click on the **Map** tab

The Map view is displayed.

2. Click on a feature in this view to select it.

The residue block is highlighted.

Features table selections

To select a block of residues using the features table

1. Click on the **Features** tab.

The Sequence Features view is displayed.

2. Scroll to the feature that you want to highlight.
3. Click on the feature.

The residue block is highlighted.

Tip. To select the entire sequence, choose **Edit | Select All**.

To specify a new background color for a selected range

1. Select the range of residues you want to highlight in the Editor view.
2. Choose **Edit | Transformations** from the menu bar.
3. Select the desired highlight **Color** from the options provided.

Adding residues

You can add residues manually to a sequence displayed in the Editor view. You can use both upper and lowercase letters, mixing and matching them to highlight regions of interest. The case you specify will be used in any printed output. However, you must use the standard IUPAC code for either nucleic acids or amino acids.

Refer to Appendix C, *"Reference Tables"*, for details of these codes.

Tip. If you are unfamiliar with the one-letter codes for the nucleotides or amino acids, you can display a small movable window that contains the information by choosing **Window | IUPAC Key** from the main menu.

When you add a residue, the residues that follow it in the sequence are moved along one position.

In addition to entering bases using the alphanumeric portion of the keyboard, you can assign nucleotide codes to keys of the numeric keypad (if your Macintosh has one) to make it easier to enter nucleic acid sequence data with one hand. Choose **Options | Set Keypad** to use this facility.

To add residues to a sequence

1. Use the mouse cursor to select an insertion point in the sequence.

Residues will be added immediately after this insertion point.

2. Enter the residues that you want to add.

Alternatively, you can choose **Edit | Paste** to add residues that have been copied to the clipboard.

To overwrite residues in a sequence

You can overwrite residues by selecting a region of the sequence before adding the new residues. When you enter new residues, the selected region is overwritten.

To change the case of all the residues in a sequence

1. Choose **Edit | Transformations** from the menu bar.
2. Select either **Make Upper Case** to make all residues upper case or **Make Lower Case** to make all residues lower case.

To change the case of a range of residues in a sequence

1. Choose **Edit | Transformations** from the menu bar.
2. Select **Enable Mixed Case Entry.**
3. Select the range of residues you want to modify.
4. Choose **Edit | Transformations** from the menu bar.
5. Select either **Make Upper Case** to make the selected residues upper case or **Make Lower Case** to make the selected residues lower case.

Note. The case used to identify residues in a sequence has no impact on any of the analysis functions in MacVector.

Deleting residues

You can delete residues manually from a sequence displayed in the Sequence Editor view. When you delete residues, the residues that follow the deletion in the sequence are moved back the appropriate number of positions.

Any deletions made from a double-stranded nucleic acid sequence result in residues being deleted from both strands.

Deleting a single residue

You can delete single residues from a sequence.

To delete single residues from a sequence

1. Use the mouse cursor to select an insertion point in the sequence.

Residues will be deleted immediately before this insertion point.

2. Use the **<backspace>** key to delete the unwanted residues.

Deleting a range of residues

You can delete a range of residues from a sequence.

To delete a range of residues from a sequence

1. Use the mouse cursor to select a region of the sequence that you want to delete.
2. Press the **<delete>** key to delete the selected residues.

Alternatively, choose **Edit** | **Cut** to remove the residues from the sequence and copy them to the clipboard for later use.

Note. Adding or deleting residues that are part of a feature will corrupt that feature. MacVector automatically adds a *frag* feature to the features table whenever a cut, edit or paste operation is performed, so you can keep track of changes.

Reversing a sequence

You can reverse the order of the residues in part or all of a sequence. This option applies to both protein and nucleic acid sequences in the sequence window.

To reverse a region or the entire sequence

1. Select either a region of the sequence, or the entire sequence.
2. Choose **Edit | Reverse**.

The region is reversed and updated in the sequence window.

Complementing a sequence

You can complement the residues in an entire sequence, or in a selected region of it. This option is available only if a nucleic acid sequence is displayed in the Sequence window.

To complement a region or the entire sequence

1. Select either a region of the sequence, or the entire sequence.
2. Choose **Edit | Complement**.

The region is complemented and updated in the sequence window.

Reversing and complementing a sequence

You can reverse then complement the order of the residues in an entire sequence, or in a selected region of it. This option is available only if a nucleic acid sequence is displayed in the Sequence window.

To reverse and complement a region or the entire sequence

1. Select either a region of the sequence, or the entire sequence.
2. Choose **Edit | Reverse & Complement**.

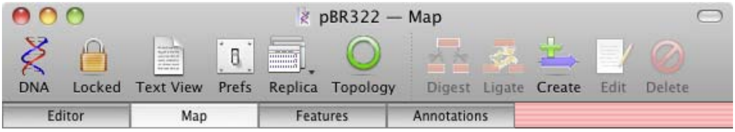
The region is reversed and complemented, and updated in the sequence window.

Map view

The Map view displays an annotated graphic showing the features of the sequence. For DNA sequences, either a linear or circular map can be displayed. At the bottom of the Map view, a collapsible overview pane displays a full-length miniature version of the sequence to help you to orient yourself when you are working with a long sequence.

The map can be edited in many ways to give a tailored map for on-screen or printed presentation. See *“Formatting the Map view”* on page 93, for further details.

The Map view has a toolbar containing tools that are used to perform particular functions. In addition to the common tools that appear by default on all of the sequence toolbars (see *“Common tools”* on page 83), the Map view toolbar includes the following map specific tools by default:



The **Digest** icon is enabled only when two restriction enzyme sites are selected. It is used to copy the fragment between the two sites, together with information about the structures of the ends produced by the enzymes and any overlapping annotations or feature information, to the clipboard for cloning (see *“Click Cloning”* on page 188).

The **Ligate** icon is enabled only when a fragment is present on clipboard and one or more restriction enzyme sites are selected. It provides access to the ligation pop-up window which enables you to manipulate the fragment on the clipboard before inserting it into the sequence (see *“Click Cloning”* on page 188).

The **Create** icon provides access to the **Feature Editor** dialog box which you can use to add new features to the sequence. Refer to *“Adding a feature”* on page 115, for further information.

The **Edit** icon is enabled only when a feature is selected. It provides access to the **Feature Editor** dialog box which you can use to edit the details of the selected feature. Refer to *"Editing a feature"* on page 118, for further information.

Tip. The **Feature Editor** does not allow you to modify the appearance of the feature. Instead this is done using the **Symbol Editor** which can be accessed by double-clicking on the feature in the Map view.

The **Delete** icon is enabled only when a feature is selected. It is used to delete the selected feature.

The **Range** text box displays the number or range of the current selection within the displayed sequence. You can edit the values in the box to change the selection. Refer to *"Selection range"* on page 85, for further information.

You can also alter the selected residues by using the **Features** drop-down menu. Choose from the list of sequence features or select the entire sequence.

Note. The Sequence Map toolbar, like all toolbars in MacVector, can be customized. Right-click on the toolbar to access this functionality. The tools described above appear in the default Sequence Map toolbar. Some of these tools may be absent and other tools may be present depending on your settings.

To display a sequence map

1. Make the sequence window active by clicking inside it.
2. Click on the **Map** tab.

The Map view is displayed, showing the sequence.

The appearance of the map can be controlled in a variety of ways, and can also be optimized for printing. See *"Configuring the numeric keypad"* on page 45.

Formatting the Map view

You can customize the Map view by manipulating the appearance of the following elements:

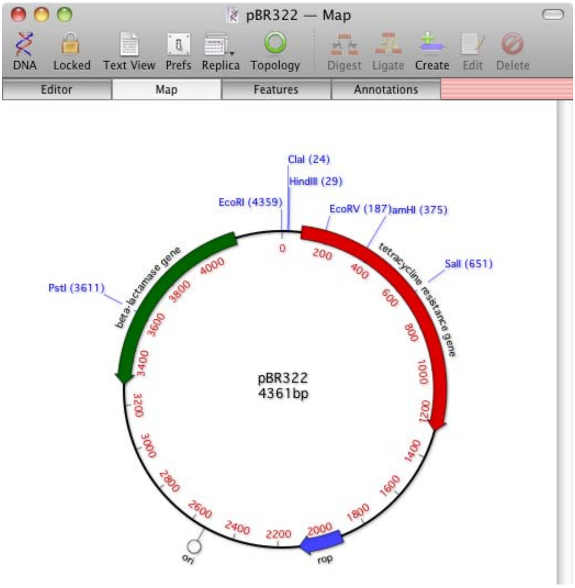
- title
- sequence line, either showing residue letters or a single line
- ruler indicating the scale of the sequence line
- feature sites
- result sites, if there are any

- segment map

This is done by editing symbol sets, of which there are three types:

- sequence default symbol set, which can be specified and saved for each sequence
- global default symbol set, which can be specified and saved as the default set used when a new sequence is created
- MacVector default symbol set, which cannot be edited, only reapplied to the editable symbol sets

There are also a number of options that can be used to control the overall appearance of the map and to optimize the map for printing.



Editing the global symbol set

The global default symbol set is edited using the Symbol Editor dialog box. Editing the global default symbol set affects only new sequences or features created after modifications are made; existing sequences or features are not affected.

When no windows are present, the dialog box is displayed by choosing **Options | Default Symbols**. If an editor view, map view or results view is active, then the menu says **Options | Symbols for *name***, where *name* is a description of the active window. In this case, you need to hold down the option key before making the menu selection, and the item **Default Symbols** will be present again. If any other view type is active, the item will be grayed out.

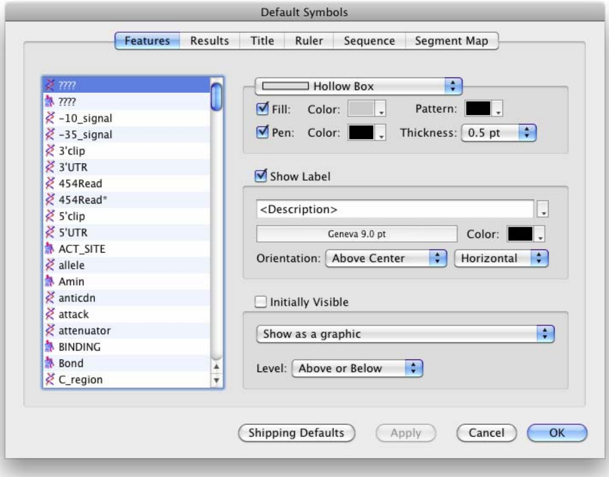
The dialog box consists of six panels, which are used to modify the appearance of the following elements of the graphical displays:

- features
- results
- title
- ruler
- sequence
- segment map
- coverage map

The use of each panel is described in the following sections.

Features panel

Using this panel of the **Symbol Editor** dialog box, you can select each feature type and alter its settings.



Hollow Box

☑ Fill: Color: Pattern:

☑ Pen: Color: Thickness: 0.5 pt

☑ Show Label

<Description>

Geneva 9.0 pt Color:

Orientation: Above Center Horizontal

□ Initially Visible

Show as a graphic

Level: Above or Below

To modify the global feature symbols

1. Choose **Options | Default Symbols** from the menu.

Note. If an editor view, map view or results view is active, you must hold down the option key before making this menu selection.

The Symbol Editor dialog box is displayed.

2. Select **Features** from the list of tabs at the top of the dialog box.

A list of all available features is displayed in the scrolling panel. The symbols to the left of the feature name identify the feature as either a protein or nucleic acid feature.

3. From the scrolling list, select a feature name whose appearance you want to modify.

Note. You can select more than one feature type to modify by holding down the **<shift>** or ⌘**<command>** key while selecting features. Each selected feature will be given the modified settings when **OK** is selected.

4. Choose a style for the feature symbol from the **Style** drop-down menu.
5. If you want the symbol to be filled, select the **Fill** check box, then choose the fill color and pattern from the associated drop-down menus.
6. If you want a border around the symbol, select the **Pen** check box, then choose the color and border thickness from the associated drop-down menus.
7. If you want to label the symbol, select the **Show Label** check box, then do the following in the Label panel:

- enter the text for the label in the text box. The following Metatags are available for the label text, click the small triangle to the left of the Label text box to display a pop up menu which lists them all:
  <Description> or **<Desc>** inserts the feature comment text
  <Length> inserts the feature residue length
  <Size> inserts the feature size (same as **<Length>** above)
  <Start> and **<Stop>** insert the start and stop residue numbers
  <Type> inserts the feature type name, e.g., CDS
  <Count>
  <Index> inserts the index number of this particular instance of the feature type
  <Segment>
  <Total> inserts the total number of this particular feature type present
  For example, “**<Start>:<Stop> <Type>**” would display “123:456 CDS” for a CDS feature that spanned residues 123 to 456. Alternatively, **“<Desc> (<Index>/<Count>)”** would display “BamHI (1/

3)", "BamHI (2/3)", and so on, useful for describing restriction enzyme cut sites.

- choose the label color from the **Color** drop-down menu
- click in the box to choose an alternative font style, size and weight to the one displayed.
- choose the position and orientation of the label with respect to the feature symbol from the **Orientation** drop-down menus.

Note. When plotting circular maps, vertical label text runs along a radius of the circle, and horizontal label text runs parallel to the circumference. Align to the left means that the text begins at the left edge of the symbol.

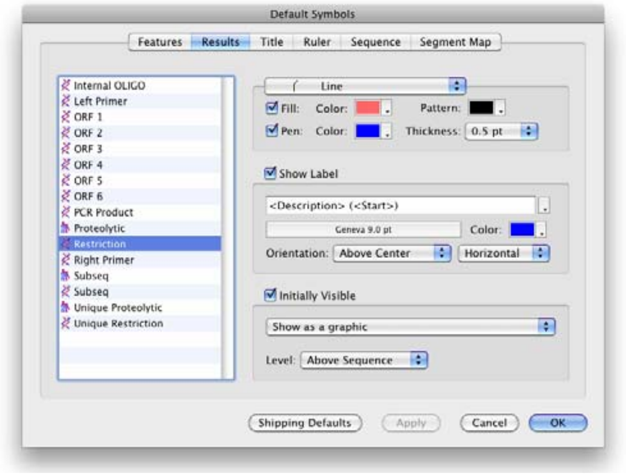
8. If you want the symbol visible by default, select the **Initially Visible** check box.
9. If you want to display selected features as residues when there is sufficient space on the map (for example, when you are zoomed in), then select **Show residue letters if room** from the drop-down menu. The features that can be displayed in this way are CDS, RNA and primer_bind features. Alternatively, select **Show as a graphic** to disable this feature.
10. Choose the position of the symbol relative to the sequence line from the **Level** drop-down menu.

There are six levels above or below the sequence line. They can be thought of as lines on which the feature symbol is displayed. You would normally leave this as "above or below", in which case MacVector assigns a level automatically.

11. To make changes permanent, do one of the following:
    - select **OK** to save the changes you have made and to close the dialog box
    - select **Cancel** to discard all changes made
    - select **Shipping Defaults** to revert to the MacVector default symbols.

Results panel

Using this panel of the Symbol Editor dialog box, you can select each result type and alter its settings.



Line

☑ Fill: Color: 🟥 Pattern: ■

☑ Pen: Color: 🟦 Thickness: 0.5 pt

☑ Show Label

<Description> (<Start>)

Geneva 9.0 pt Color: 🟦

Orientation: Above Center Horizontal

☑ Initially Visible

Show as a graphic

Level: Above Sequence

To modify the global results symbols

1. Choose **Options | Default Symbols** from the menu.

Note. If an editor view, map view or results view is active, you must hold down the option key before making this menu selection.

The Symbol Editor dialog box is displayed.

2. Select **Results** from the list of tabs at the top of the dialog box.

A list of all available result types is displayed in the scrolling panel. The symbols to the left of the result type identify the result as either a protein or nucleic acid result.

3. From the scrolling list, select a result type whose appearance you want to modify.
4. Choose a style for the symbol from the **Style** drop-down menu.
5. If you want the symbol to be filled, select the **Fill** check box, then choose the fill color and pattern from the associated drop-down menus.
6. If you want a border around the symbol, select the **Pen** check box, then choose the color and border thickness from the associated drop-down menus.

7. If you want to label the result symbol, select the **Show Label** check box, then do the following in the Label panel:

- enter the text for the label in the text box. The following Metatags are available for the label text, click the small triangle to the left of the Label text box to display a pop up menu which lists them all:
  <Description> or **<Desc>** inserts the result type name
  <Length> inserts the
  <Size> inserts the
  <Start> and **<Stop>** insert the start and stop residue numbers
  <Type> inserts the result type name, e.g., CDS
  <Count>
  <Index>
  <Segment>
  <Total>
  For example, **<Start> <Type>** would display 123 BamHI on a restriction enzyme map that had a BamHI cut site at residue 123.
- choose the label color from the **Color** drop-down menu
- click in the box to choose an alternative font style, size and weight to the one displayed.
- choose the position and orientation of the label with respect to the result symbol from the **Orientation** drop-down menus.

Note. Align to the left means that the text begins at the left edge of the symbol.

8. If you want the result type visible by default, select the **Initially Visible** check box.
9. Choose the position of the result symbol relative to the sequence line from the **Level** drop-down menu.

There are six levels above or below the sequence line. They can be thought of as lines on which the result symbol is displayed. You would normally leave this as “above or below”, in which case MacVector assigns a level automatically.

10. To make changes permanent, do one of the following:

- select **OK** to save the changes you have made and to close the dialog box
- select **Cancel** to discard all changes made
- select **Shipping Defaults** to revert to the MacVector default symbols.

Title panel

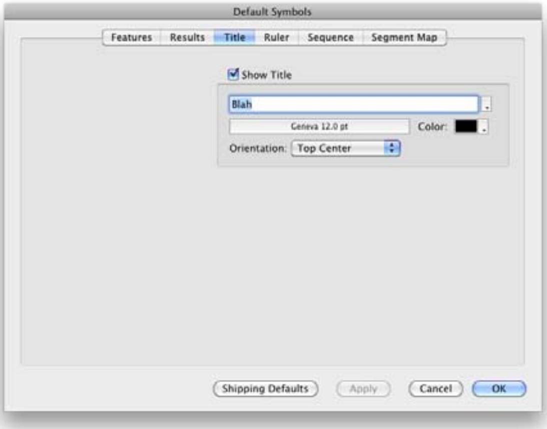
The Title panel is used to define the title text.

To modify the global map tab title

1. Choose **Options | Default Symbols** from the menu.

Note. If an editor view, map view or results view is active, you must hold down the option key before making this menu selection

The Symbol Editor dialog box is displayed.



Default Symbols

☑ Show Title

Blah

Geneva 12.0 pt

Color: ■

Orientation: Top Center

2. Select **Title** from the list of tabs at the top of the dialog box.
3. If you want the title visible by default, select the **Show Title** check box.
4. Enter a title in the text box.

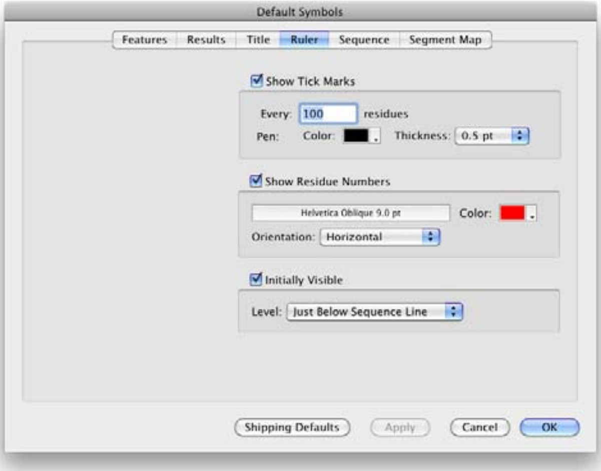
The text **<Sequence Name>** is special text that is replaced by the name of the sequence whose map is displayed.

5. Click in the box to choose an alternative font style, size and weight to the one displayed.
6. Choose the text color from the **Color** drop-down menu.
7. Choose the title position from the **Orientation** drop-down menu.
8. To make changes permanent, do one of the following:
   - select **OK** to save the changes you have made and to close the dialog box
   - select **Cancel** to discard all changes made

- select **Shipping Defaults** to revert to the MacVector default symbols.

Ruler panel

The ruler panel controls the appearance of the sequence numbering.



☑ Show Tick Marks

Every: 100 residues

Pen: Color: ■ Thickness: 0.5 pt

☑ Show Residue Numbers

Helvetica Oblique 9.0 pt Color: 🟥

Orientation: Horizontal

☑ Initially Visible

Level: Just Below Sequence Line

To modify the global ruler markings

1. Choose **Options | Default Symbols** from the menu.

Note. If an editor view, map view or results view is active, you must hold down the option key before making this menu selection.

The Symbol Editor dialog box is displayed.

2. Select **Ruler** from the list of tabs at the top of the dialog box.
3. If you want tick marks, select the **Show Tick Marks** check box, then do the following:
   - enter the tick interval in the text box
   - choose the tick color from the **Color** drop-down menu
   - choose the tick thickness from the **Thickness** drop-down menu.
4. If you want to label the ruler ticks, select the **Show Residue Numbers** check box, then do the following:
   - click in the box to choose an alternative font style, size and weight to the one displayed.
   - choose the label color from the **Color** drop-down menu

- choose the orientation of the label with respect to the ruler from the **Orientation** drop-down menus

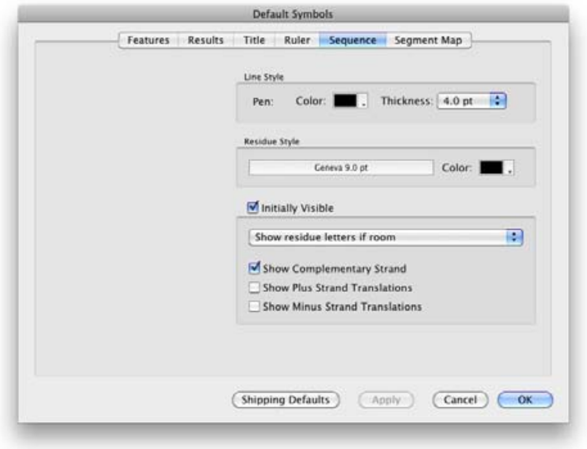
5. If you want the ruler visible by default, select the **Initially Visible** check box.
6. Choose the level of the ruler from the **Level** drop-down menu.

The levels above or below the sequence line can be thought of as lines on which the graphic features are displayed.

7. To make changes permanent, do one of the following:
   - select **OK** to save the changes you have made and dismiss the dialog box
   - select **Cancel** to discard all changes made
   - select **Shipping Defaults** to revert to the MacVector default symbols.

Sequence panel

The sequence panel defines the default appearance of any new sequence maps that are created.



Line Style

Pen: Color: ■ Thickness: 4.0 pt

Residue Style

Geneva 9.0 pt Color: ■

☑ Initially Visible

Show residue letters if room

☑ Show Complementary Strand

□ Show Plus Strand Translations

□ Show Minus Strand Translations

To modify the global sequence appearance

1. Choose **Options | Default Symbols** from the menu.

Note. If an editor view, map view or results view is active, you must hold down the **option** key before making this menu selection.

The Symbol Editor dialog box is displayed.

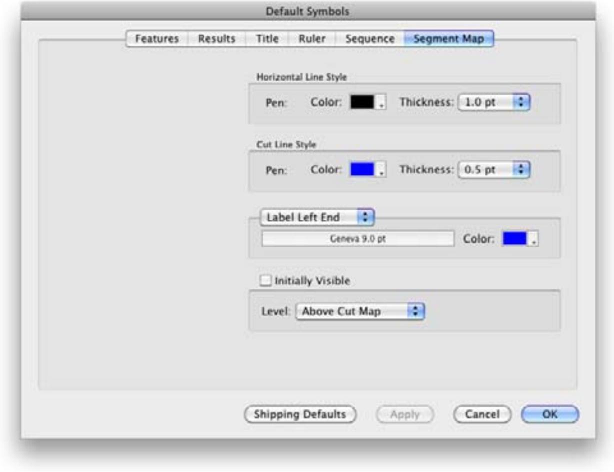
2. Select **Sequence** from the list of tabs at the top of the dialog box.
3. Choose the line color and thickness by selecting from the **Color** and **Thickness** drop-down menus in the **Line Style** panel.
4. Choose an alternative font style size and weight for residue labels by clicking in the box displaying the current settings in the **Residue Style** panel.
5. Choose the residue color by selecting from the **Color** drop-down menu.
6. If you want sequences visible by default, select the **Initially Visible** check box.
7. If you want to view the sequence only as a line, select the **Show as a Line** from the drop-down menu.
8. If you want to display the sequence letters, select the **Show residue letters if room** from the drop-down menu.

The sequence letters will be displayed if there is room. This will depend on the settings on the Graphic Palette dialog box, see *“Editing the general map appearance”* on page 109.

9. To make changes permanent, do one of the following:
   - select **OK** to save the changes you have made and to close the dialog box
   - select **Cancel** to discard all changes made
   - select **Shipping Defaults** to revert to the MacVector default symbols.

Segment Map panel

The segment map panel controls the appearance of the segment map. It is only available when MacVector is displaying a linear map showing the results of an enzyme digest or subsequence search.



Horizontal Line Style

Pen: Color: ■ Thickness: 1.0 pt

Cut Line Style

Pen: Color: ■ Thickness: 0.5 pt

Label Left End

Geneva 9.0 pt Color: ■

□ Initially Visible

Level: Above Cut Map

To modify the global segment map appearance

1. Choose **Options | Default Symbols** from the menu.

Note. If an editor view, map view or results view is active, you must hold down the **option** key before making this menu selection.

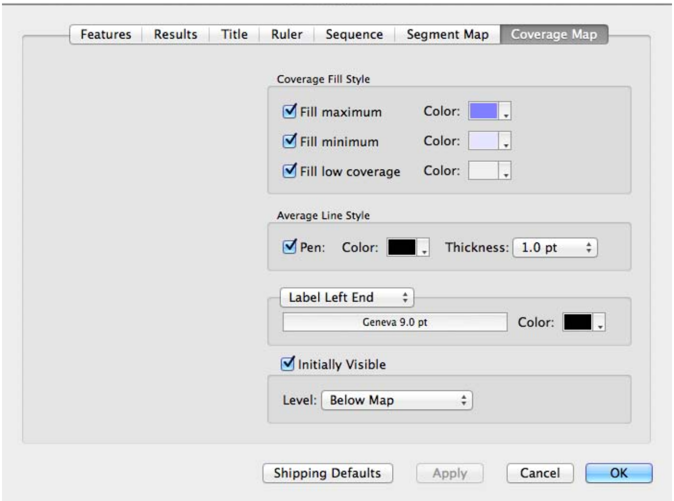
The Symbol Editor dialog box is displayed.

2. Select **Segment Map** from the list of tabs at the top of the dialog box.
3. Choose the color and thickness of the horizontal lines, by selecting from the **Color** and **Thickness** drop-down menus in the **Horizontal Line Style** panel.
4. Choose the color and thickness of the vertical cut lines, by selecting from the **Color** and **Thickness** drop-down menus in the **Cut Line Style** panel.
5. To alter the appearance and position of the map labels, do the following in the **Label** panel:
   - choose the label position from the drop-down menu
   - click in the box to choose an alternative font style, size and weight to the one displayed.
   - choose the label color from the **Color** drop-down menu

6. If you want the segment map visible by default, select the **Initially Visible** check box.
7. Choose the position of the segment map (above or below the cut map) from the **Level** drop-down menu.
8. To make changes permanent, do one of the following:
   - select **OK** to save the changes you have made and to close the dialog box
   - select **Cancel** to discard all changes made
   - Select **Shipping Defaults** to revert to the MacVector default settings.

Coverage Map panel

The coverage map panel controls the appearance of the coverage map. It is only available when MacVector is displaying a Reference Contig or a Child Contig from a Sequence Assembly project



Coverage Fill Style

☑ Fill maximum Color:

☑ Fill minimum Color:

☑ Fill low coverage Color:

Average Line Style

☑ Pen: Color: Thickness: 1.0 pt

Label Left End

Geneva 9.0 pt Color:

☑ Initially Visible

Level: Below Map

To modify the global coverage map appearance

1. Choose **Options | Default Symbols** from the menu.

Note. If an editor view, map view or results view is active, you must hold down the **option** key before making this menu selection.

The Symbol Editor dialog box is displayed.

2. Select **Coverage Map** from the list of tabs at the top of the dialog box.

3. Choose the display options and fill styles for the maximum, minimum and low coverage regions by checking the boxes and selecting from the **Color** drop-down menus in the **Coverage Fill Style** panel.
4. Choose the color and thickness of the average lines, by selecting from the **Color** and **Thickness** drop-down menus in the **Average Line Style** panel.
5. To alter the appearance and position of the map labels, do the following in the label panel:
   - choose the label position from the drop-down menu.
   - click in the box to choose an alternative font style, size and weight to the one displayed.
   - choose the label color from the **Color** drop-down menu.
6. If you want the coverage map visible by default, select the **Initially Visible** check box.
7. Choose the position of the coverage map (above or below the map) from the **Level** drop-down menu.
8. To make changes permanent, do one of the following:
   - select **OK** to save the changes you have made and to close the dialog box
   - select **Cancel** to discard all changes made
   - Select **Shipping Defaults** to revert to the MacVector default settings.

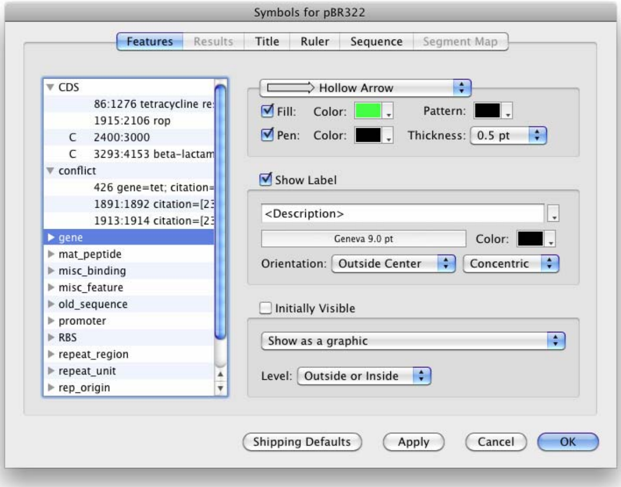
Editing the sequence symbol set

When a new sequence is created, its symbol set is defined to be the same as the global symbol set, see *“Editing the global symbol set”* on page 94. The symbol set for each sequence can be individually edited and saved.

When a change is made to the sequence symbol set, the active maps for the sequence are updated immediately. If you are refining the appearance of a large map, you can interrupt redrawing at any time by one of the following methods:

- pressing the ⌘ and **period** (.) keys together
- pressing the **esc** key

- changing any setting in the Graphics Palette dialog box.



▼ CDS
86:1276 tetracycline re
1915:2106 rop
C 2400:3000
C 3293:4153 beta-lactam
▼ conflict
426 gene=tet; citation=
1891:1892 citation=[2
1913:1914 citation=[2
▶ gene
▶ mat_peptide
▶ misc_binding
▶ misc_feature
▶ old_sequence
▶ promoter
▶ RBS
▶ repeat_region
▶ repeat_unit
▶ rep_origin

Hollow Arrow
☑ Fill: Color: 🟩 Pattern: ■
☑ Pen: Color: ■ Thickness: 0.5 pt
☑ Show Label
<Description>
Geneva 9.0 pt Color: ■
Orientation: Outside Center Concentric
☐ Initially Visible
Show as a graphic
Level: Outside or Inside

To edit the sequence symbol set

1. Do one of the following:
   - if a sequence editor, map or results view is active, choose **Options | Symbols for *name***, where *name* is a description of the active window.
   - select the item that you want to edit from the scrolling list of the Graphics Palette, then select **Edit**.

Note. The Graphics Palette is displayed by selecting **Window | Show Graphics Palette**. This menu item is a toggle, and once set, the Graphics Palette will be displayed whenever a map view is selected, until you choose **Window | Hide Graphics Palette**.

The Symbol Editor dialog box is displayed. If you used the Edit button on the Graphics Palette dialog box, the appropriate Symbol Editor panel is displayed.

Note. The scrolling list shows a tree view of the map items on the Features and Results panels. These can be expanded or contracted by clicking on the triangles to the left of the items.

2. For details of how to edit the symbols, see the corresponding sections in *"Editing the global symbol set"* on page 94, (omit step 1 in each case):

- *"Features panel"* on page 95
- *"Results panel"* on page 98
- *"Title panel"* on page 100
- *"Ruler panel"* on page 101
- *"Sequence panel"* on page 102
- *"Segment Map panel"* on page 104

Note. When you access the dialog boxes in this way, they each have one additional button, next to the **Cancel** button. It is the **Apply** button. This lets you update the open maps with the changes you make, without closing the dialog box. This means that you can make a number of edits to the map display conveniently.

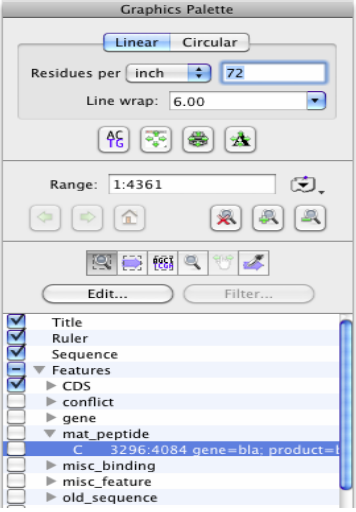
The tree hierarchy of map items for the Features and Results panels enables you to modify the symbol for each item individually, for a complete group of features, or any combination of features.

Note. Modifying the symbol of an item on the Results or Features panel affects all other items subordinate to it in the tree hierarchy.

3. The edited symbol set is saved with the sequence, so you can save different map views of the same sequence.

Editing the general map appearance

The Graphics Palette is used to control the overall appearance of the Map view. Any changes made directly from this dialog box are applied immediately to the Map view.

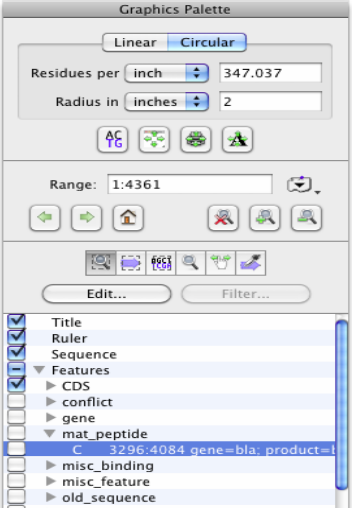


Residues per inch 72

Line wrap: 6.00

Range: 1:4361

- ☑ Title
- ☑ Ruler
- ☑ Sequence
- ⊟ ▼ Features
  - ☑ ▶ CDS
  - □ ▶ conflict
  - □ ▶ gene
  - □ ▼ mat_peptide
    - □ C 3296:4084 gene=bla; product=
  - □ ▶ misc_binding
  - □ ▶ misc_feature
  - □ ▶ old_sequence



Residues per inch 347.037

Radius in inches 2

Range: 1:4361

- ☑ Title
- ☑ Ruler
- ☑ Sequence
- ⊟ ▼ Features
  - ☑ ▶ CDS
  - □ ▶ conflict
  - □ ▶ gene
  - □ ▼ mat_peptide
    - □ C 3296:4084 gene=bla; product=
  - □ ▶ misc_binding
  - □ ▶ misc_feature
  - □ ▶ old_sequence

To modify the general map appearance

1. When a Map view is displayed, choose **Windows | Show Graphics Palette** if the Graphics Palette is not visible.

Note. This menu item is a toggle. Once set, the Graphics Palette will be displayed each time a Map view is displayed, until you choose **Windows | Hide Graphics Palette**.

The Graphics Palette is displayed. The scrolling list shows a tree directory of all the map elements whose symbols can be edited. The list can be expanded or contracted by clicking on the triangles to the left of the items. The items on the dialog box differ, depending on whether the sequence is displayed as a linear or circular map.

2. If the graphic map is for a DNA sequence, you can toggle between linear and circular display by selecting either the **Linear** or the **Circular** option at the top of the Graphics Palette.
3. To adjust the display density of residues, select **inch**, **cm** or **line** from the **Residues per** drop-down menu and enter the number of residues to appear in that interval in the associated text box.
4. If your map is linear, you can adjust the line length by choosing whether to make one long unbroken line of residues, or re-size to the

width of the window or the printable page, by selecting from the **Line wrap** drop-down menu.

5. If your map is circular, you can control the size of the circle by selecting a value for the radius units from the **Radius in** drop-down menu, then entering a value for the radius in the associated text box.
6. In addition, the following Quick Fit buttons are provided to enable you to adjust the range and scaling of the sequence display quickly and easily:

- **Zoom to Sequence** adjusts the residues per inch automatically, such that the residues are just visible.

Note. The **Zoom to Sequence** button does not change the current displayed sequence range.

- **Fit to Window** adjusts the residues per inch automatically, such that the entire sequence range can be displayed in the current window.

Note. The **Fit to Window** button re-sizes a circular map to the longest dimension of your window. It will not shrink the map below a minimum legible size, so some scrolling will be needed to view the entire map if your window is small.

- **Fit to Page** the residues per inch automatically, such that entire sequence range will fit on the currently selected printable page.

Note. The printable page size is determined by the current settings of the page setup dialog.

- **Fit Residues** resets the displayed range to the entire sequence and adjusts the residues per inch automatically, such that the residues are just visible.

To display a region of a sequence in a Map view

You can restrict the display to a certain region of the sequence, using either the mouse or the Graphics Palette:

- Using the mouse, click and drag to select the region on the map that you want to display. If your map is linear, the selection will be fitted to the window automatically.
- Using the Graphics Palette: type, in the **Range** panel, the sequence numbers of the first and last residues of the selected region, separated by a colon (:). Alternatively, you can select a region from the features table drop-down menu at the right of the text box. When you make a selection from the Graphics Palette, your current settings for residue density and line wrap are retained.

Note. For results maps, the range cannot be set outside the range that was specified when the results calculation was performed. For example, if restriction enzyme cut sites were only computed between bases 1000 and 2000, the range cannot be set to show bases 1 to 4000, because result data is not available for the wider range. If the user types in range values outside the permitted limit, they will be changed to be within the limit (i.e. no less than 1000, no more than 2000 in the example just given).

Navigation

The followng Navigation buttons are provided to enable you to scroll around the Map view quickly and easily:

- **Slide Left** if a linear map is displayed and a range is selected, then this nudges the range to the left. If a circular map is displayed, then this rotates the sequence slightly to the left. The left arrow key also performs this function.
- **Slide Right** if a linear map is displayed and a range is selected, then this nudges the range to the right. If a circular map is displayed, then this rotates the sequence slightly to the right. The right arrow key also performs this function.
- **Home** if a linear map is displayed and a range is selected, then this centers the current range in the top of the window. If a circular map is displayed then this resets the rotation.
- **Zoom In** this zooms in such that the displayed range is reduced two-fold.
- **Zoom Out** this zooms out such that the displayed range is increased two-fold.
- **Reset Zoom** this resets the zoom such that the entire sequence is displayed in the window.

Tip. You can also navigate around the Map view using the overview pane. If only part of the main sequence is visible, click on a different part of the sequence in the overview pane and the Map view will move to display that section.

To select objects in a Map view

The following Selection Mode buttons are provided to enable you to select different objects in the Map view quickly and easily:

- **Zoom** this mode allows you to click on objects such as features or sites to select them, or to click, hold and drag to zoom in on the objects you selected. This is the default Selection Mode.

- **Feature Selection** this mode allows you to click, hold and drag to select all of the features or sites within the selection rectangle.
- **Sequence Selection** this mode allows you to click, hold and drag to select the portion of the sequence within the selection rectangle.
- **Magnify** this mode allows you to click to increase the magnification of the display two-fold or <option>-click to reduce the magnification of the display two-fold.
- **Slide** this mode allows you to drag the current zoomed region to the left or the right. It is also used to rotate circular maps, such that an arbitary location appears at the 12 o'clock position.
- **Copy Feature Appearance** this mode allows you to click on a feature to copy its appearance and paste it on to any currently selected features. This mode is enabled only if one or more features are selected.

Printing maps

After you have drawn a map to your requirements, you can print it. MacVector generates high-quality print output. The map tab can show you how the map will be printed if it is large enough to cover several pages, or alternatively you can use **Page Setup...** to scale a map to fit onto a single page.

Maps are always printed at 100% view; the view scale controls on the map tab have no effect on the printed output and are there only to enable you to see the split of the map over several pages.

When choosing a font for labels and titles, be aware of the differences between fonts that look good on the screen and fonts that look good on the printer. Geneva, Monaco and Chicago fonts are designed to look clear and legible on the screen. However, when you print them to a standard PostScript laser printer, their letter spacing can look a little uneven. (In fact, when you try to print Geneva you actually get Helvetica letters, and when you print Monaco you may get Courier letters, but with letter spacing appropriate to their original screen font. The mismatch between which font's letters are being drawn and which font is being used for the inter-letter spacing accounts for the uneven appearance.)

Geneva is the default label font supplied with MacVector, to give best legibility on screen. If your primary concern is to get the best quality printed map, edit the symbols to set them to use a true PostScript printer font like Helvetica. You will notice the difference, particularly in the quality of rotated text on circular maps.

To print a map view

1. If you want to preview how the map will split across printed pages, click the **page mode** button in the lower left corner of the view.

This toggles the page mode. In page mode, the tab has dotted lines that represent page boundaries, so you can see where your map will split.

2. If necessary, adjust the view so that the whole map appears in the window, by selecting **Fit In Window** from the **View scale** drop-down menu next to the **page mode** button.

Note. This does not affect the printed output size.

3. If you want to scale the map, either to fit on a single sheet of paper, or to control where the map splits across a page, do the following:
   - choose **File | Page Setup**. The Page Setup dialog box appears
   - select **Page Attributes** from the drop-down menu at the top left of the dialog box
   - enter a value in the **Scale** text box, then select **OK**.

The map tab is redrawn. If you need to adjust the size further, repeat this step as necessary.

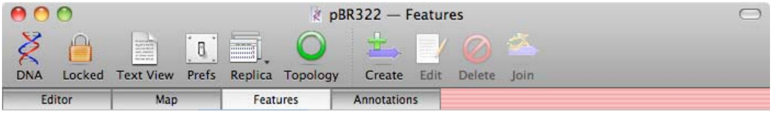
4. Choose **File | Print.**
5. Adjust the print options as required.
6. Select **Print**.

The map is printed to the chosen device.

Features view

The Features view displays the features table associated with the sequence. A feature is an item of information about a sequence that is associated with a specific position within that sequence. The position may be a single residue, the gap between two residues, or a contiguous series of residues.

The Features view has a toolbar containing tools that are used to perform particular functions. In addition to the common tools that appear by default on all of the sequence toolbars (see *“Common tools”* on page 83), the Features view toolbar includes the following feature specific tools by default:



The **Create** icon provides access to the **Feature Editor** dialog box which you can use to add new features to the sequence. Refer to *"Adding a feature"* on page 115, for further information.

The **Edit** icon is enabled only when a feature is selected. It provides access to the **Feature Editor** dialog box which you can use to edit the details of the selected feature. Refer to *"Editing a feature"* on page 118, for further information.

Tip. The **Feature Editor** does not allow you to modify the appearance of the feature. Instead this is done using the **Symbol Editor** which can be accessed by double-clicking on the feature in the Map view.

The **Delete** icon is enabled only when a feature is selected. It is used to delete the selected feature.

The **Join** icon is enabled only when two or more features are selected. It is used to join the selected features to create a single segmented feature.

Note. The Sequence Features toolbar, like all toolbars in MacVector, can be customized. Right-click on the toolbar to access this functionality. The tools described above appear in the default Sequence Features toolbar. Some of these tools may be absent and other tools may be present depending on your settings.

To display sequence features

1. Make the Sequence window active by clicking inside it.
2. Click on the **Features** tab.

The Features view is displayed.

You can modify sequence features as follows:

- add a new feature
- edit an existing feature
- delete an existing feature.

Features can be modified only when the sequence is unlocked. This is done by clicking on the **Locked** indicator icon in the Sequence window toolbar.

Feature keywords

The MacVector feature keyword sets for proteins and nucleic acids match the SwissProt and GenBank standard sets, respectively, see Appendix E, *“GenBank Feature Tables”* for further details. Any standard feature can be selected from the keyword list when editing the features table of protein or nucleic acid sequences.

Adding a feature

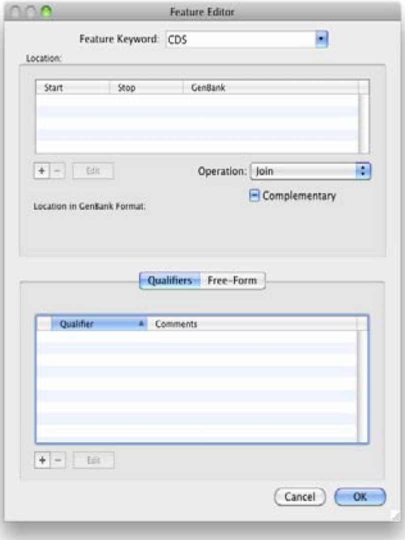
You can create a feature directly from a selected region in either the Editor view or the Map view or create one from scratch using the Features view.

You can also use the auto-annotation tool to automatically add features that match those on annotated sequences in a specified folder (see *“Using auto-annotation to add features and feature appearance information”* on page 127).

To add a protein feature

1. Click the **Create** icon on the toolbar.

The **Feature Editor** dialog box is displayed.



Feature Keyword: CDS

Operation: Join

By default, any region selected in the Sequence Editor or Sequence Map view is used to populate the **Start Location** and **Stop Location** boxes.

2. If necessary, supply or modify residue range of the feature using the **Start Location** and **Stop Location** text boxes and the associated **At** and **Before (<)** or **After (>)** drop-down menus.
3. Scroll through the **Feature Keyword** list and select the feature type that you want to add.

A complete list of the protein feature keywords supported in MacVector can be found in Appendix E, *"GenBank Feature Tables"*.

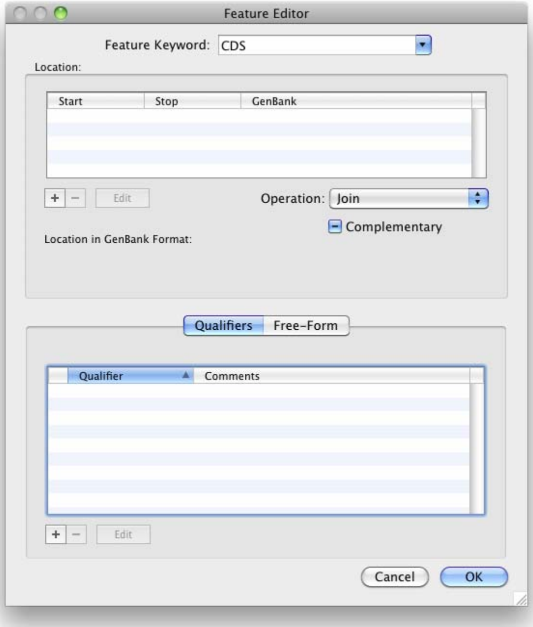
4. Optionally, type a comment in the **Comments** text box.

Select **OK** to add the new feature.

To add a nucleic acid feature

1. Click the **Create** icon on the toolbar.

The **Feature Editor** dialog box is displayed.



Feature Keyword: CDS

Operation: Join

By default, any region selected in the Sequence Editor or Sequence Map view appears as the first entry in the **Location** list.

2. Click on the + button to add additional segments to the feature.

A new sheet is displayed in the dialog box, which you can use to define additional segments.

3. Select a segment in the **Location** list and click **Edit** to modify it.
4. Select a segment in the **Location** list and click - to remove it from the feature.
5. Scroll through the **Feature Keyword** list and select the feature type that you want to add, or type the name of the feature type you want to add in the text box.

Note. If you type an invalid keyword into the **Feature Keyword** text box, then the **OK** button on the **Feature Editor** dialog box will be disabled. Select or type a valid feature keyword to re-enable the **OK** button.

A complete list of the nucleic acid feature keywords supported in MacVector can be found in Appendix E, *"GenBank Feature Tables"*.

6. Specify how the segments in the **Location** list should be combined to construct the feature using the **Operation** drop-down menu.

Select **Join** to indicate that the segments should be placed end-to-end to form one contiguous sequence. Alternatively, select **Order** to indicate that the segments can be found in the specified order (5' to 3' direction) but are not necessarily joined.

7. If the feature you are creating is located on the minus strand (the strand complementary to the one that is actually present in the sequence file), check the **Complementary** box. If the feature is located on the plus strand, ensure the this box is not checked.
8. To add GenBank qualifiers to the feature:
   - Ensure the **Qualifiers** tab is selected.
   - Click on the + button.

A new sheet is displayed in the dialog box, which you can use to add qualifiers.

- Select the **Qualifier** you want to add from the drop-down list.

Note. Only qualifiers that are allowed by the GenBank specification for the selected feature type are available in the list.

Note. MacVector now honors the standard GenBank `/codon_start` qualifier for defining the start frame of a coding region. This is useful when annotating exons in eukaryotic sequences where triplet codons may not be in phase due to the presence of introns. Valid values are 1, 2 and 3.

- Optionally, type **Comments** in the text box.
- Click **OK** to add the qualifier to the feature or **Cancel** to dismiss the sheet without adding the qualifier.

- Repeat these steps to add as many qualifiers as you want.

Note. Some feature types have mandatory qualifiers. If one of these is selected, then MacVector will automatically add the appropriate qualifiers to the Qualifier list.

9. Alternatively, if you are familiar with GenBank format, you can type qualifiers in by hand using the **Free-Form** tab.
10. Select **OK** to add the new feature.

You can also use the auto-annotate tool to add matching features on annotated sequences in a specified folder automatically

Editing a feature

To edit a feature

1. Select the feature that you want to edit by clicking on it with the mouse button.
2. Click the **Edit** icon on the toolbar to display the current feature information in full.
3. Edit the information as required.

Deleting a feature

To delete a feature

1. Select the feature that you want to delete by clicking on it with the mouse.
2. Click the **Delete** button on the toolbar to delete the selected feature.

Sorting features

The list of features displayed in the Features view can be sorted in several ways:

- in alphabetical order by feature keyword
- in ascending order by starting position on the 5' strand
- in ascending order by stopping position on the 5' strand
- in reverse-alphabetical order by feature keyword
- in descending order by starting position on the 5' strand
- in descending order by stopping position on the 5' strand
- by strand

To sort features

1. On the Features table, click the column heading label of the column you want to sort.

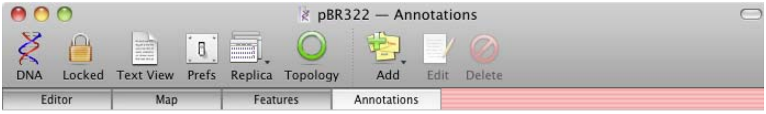
The feature list will be sorted by that column.

Note. Sorting is an additive process. So, if you sort by Type and then by Strand you will obtain a list in which, for example, all the CDS features on the upper strand are grouped together. This makes it easy to select related features and modify their properties, collectively.

Annotations view

The Annotations view displays text annotations associated with the sequence. Annotations provide extra information about the sequence, such as a description, academic references, or identifying numbers assigned by the databases from which the sequences are taken.

The Annotations view has a toolbar containing tools that are used to perform particular functions. In addition to the common tools that appear by default on all of the sequence toolbars (see *“Common tools”* on page 83), the Annotations view toolbar includes the following annotation specific tools by default:



The **Add** icon provides access to the **Annotation Editor** dialog box which you can use to add new annotations to the sequence. Refer to *“Adding an annotation”* on page 123, for further information.

The **Edit** icon is enabled only when an annotation is selected. It provides access to the **Annotation Editor** dialog box which you can use to edit the selected annotation. Refer to *“Editing an annotation”* on page 123, for further information.

The **Delete** icon is enabled only when an annotation is selected. It is used to delete the selected annotation. Refer to *“Deleting an annotation”* on page 123, for further information.

Note. The Annotations view toolbar, like all toolbars in MacVector, can be customized. Right-click on the toolbar to access this functionality. The tools described above appear in the default Annotations view toolbar. Some of these tools may be absent and other tools may be present depending on your settings.

To display sequence annotations

1. Make the Sequence window active by clicking inside it.
2. Click on the **Annotations** tab.

The Annotation view is displayed.

PUBMED abstracts can be retrieved directly from the Annotation view, by double-clicking on any reference that contain a PUBMED ID.

You can also modify sequence annotations as follows:

- add a new annotation
- edit an existing annotation
- delete an existing annotation.

Annotations can only be modified when the sequence is unlocked. This is done by clicking on the **Locked** indicator icon in the Sequence window toolbar.

Types of annotation

The LOCUS line in the GenBank annotation field is not fully editable in MacVector. Instead, MacVector fills in the LOCUS line information automatically whenever you save a file. The first ten characters of the name that you gave to the sequence file are used as the locus name. Although, you can change this. The current number of bases in the sequence file is computed by MacVector. The molecule representations (DNA or RNA, circular or linear) are taken from the current settings of the molecule and linear / circular icons unless you override the settings in the Save As dialog box. The entry date is the date the modifications were made.

All other GenBank annotation types can be edited directly.

Field	Description
definition	This is a short description of the sequence entry. It starts with the common name of the source organism, followed by the criteria that distinguish this sequence from the other parts of the source genome (gene name and what the gene codes for, the protein name and mRNA, or a description of the sequence’s function if it is a noncoding region). The definition line of coding regions can end with a completeness qualifier such as complete sequence, complete genome, or cds (complete coding sequence). MacVector limits this field to 254 characters.

accession This contains one or more accession numbers that apply to the sequence entry. Accession numbers are assigned by GenBank and you would usually change them only if you knew there was an error in an existing number or if it was a new sequence and you had just received the accession number assignment from GenBank. Each accession number consists of an alphabetic character, followed by five digits. A space character is used to separate multiple accession numbers. The first accession number listed is unique to this particular sequence entry. MacVector limits this field to 254 characters.

version This field contains a compound identifier, consisting of the primary accession number and a numeric version number associated with the current version of the sequence data in the record. This is followed by an integer key (a "GI") assigned to the sequence by NCBI. Mandatory keyword/exactly one record.

DBLink The DBLink line (DR line in SWISSPROT) is used as a pointer to information related to SWISSPROT and GENBANK entries and found in other data collections.

keywords This field consists of short phrases that provide information about the sequence entry. Use semicolons to separate the keyword phrases and a period after the last keyword phrase. MacVector limits this field to 254 characters.

segment This field is found only in segmented entries. It is used if two or more sequence entries of known relative orientation are separated by a short (less than 10 kb) segment of DNA. The format for the annotation is: n of total, where n is the segment number of the current entry and total is the total number of segments. Usually, you would not use this annotation.

source

This field has up to three subfields:

Source may contain an abbreviated form of the organism name and a molecule type.

Organism contains the scientific name for the source organism (genus and species, where applicable). Sequence files also contain in the Organism subfield a list of all the taxonomic classification levels for the organism, separated by semicolons and ending with a period.

If the sequence is from a parasitic organism, you can enter the name of the parasite’s host organism in the optional **Host** subfield. MacVector limits each subfield to 254 characters.

reference

This field has six subfields:

The optional **Title** subfield contains the title of the cited reference.

Type a number in the **Reference** subfield (references are numbered sequentially in a GenBank file, starting with 1).

The **Author** subfield is a list of authors in the order that they appear in the cited reference. Each name is listed in the form “lastname, A.A.”. The names are separated by a comma followed by a space. There is no comma after the penultimate name and the final name is preceded by the word “and”.

The **Journal** subfield contains the name of the journal, book, or thesis where the citation was published, or unpublished if the sequence has not been published. MacVector limits each subfield to 254 characters.

The **PUBMED ID** subfield contains the PUBMED references.

The **REMARK** subfield contains accredited comments that have been added by the database managers.

comment

This field is an optional, free-form text section. MacVector limits this field to 32,767 characters

base count

The contents of this section cannot be changed by the user - the base count line can only be added to or removed from the file. If the base count field is present, MacVector automatically updates the base count whenever you edit a sequence.

origin — This field specifies how the first base is located within the genome. The Origin field for pBR322, for example, reads EcoR1 site. This field may be left blank or the word `Unreported` may be entered. MacVector limits this field to 254 characters.

project — This field is an optional, free-form text section.

contig — This field is an optional, free-form text section.

Adding an annotation

To add an annotation

1. Click the **Add Annotation** icon on the toolbar and select the type of annotation that you want to add from the submenu that appears.

A template for that annotation type is displayed.

2. Type the information as required.

Editing an annotation

To edit an annotation

1. Select the annotation that you want to edit by clicking on it with the mouse.
2. Click the **Edit** icon on the toolbar to display an editable version of the annotation.
3. Edit the information as required.

Deleting an annotation

To delete an annotation

1. Select the annotation that you want to delete by clicking on it with the mouse.
2. Click the **Delete** icon on the toolbar to delete the selected annotation.

Text view

A Text view can be created and customized to show additional information including defined sequence features, and, in the case of nucleic acid sequences, the complement strand and translation of major features.

To create a Text view for the current sequence click the **Text View** icon on the toolbar.

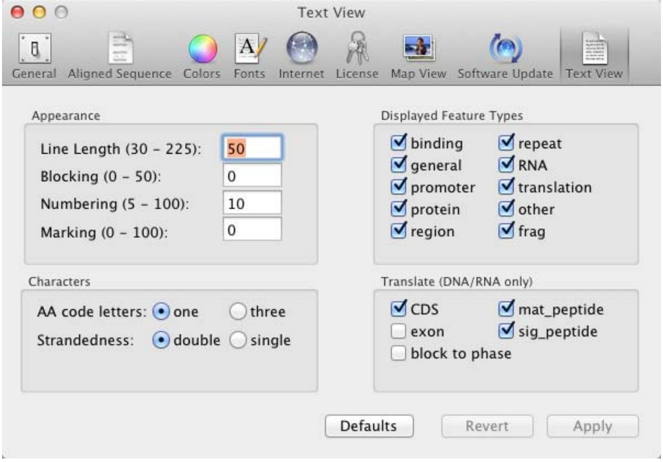
Formatting the Text view

You can customize the appearance of the Text view, both as it appears on the screen and as it will be printed. The formatting enables you to control:

- the number of residues that occur per sequence block, between sequence numbers, and between sequence marks
- how the sequence itself is displayed: uppercase or lowercase letters, single- or double-stranded format, and the fixed width font that is used.
- which feature types will be included among the annotations to the sequence
- which translated feature types will have their amino acid sequences printed beneath the nucleic acid sequence.

Text sequence formatting also affects some of the format properties of the aligned sequence display. Examples of formatting can be found in Appendix D, *“Formatting Examples”*.

This functionality may be accessed at any time, whether a text sequence is present or not. If a text sequence is present, it will be updated immediately to reflect changes.



Appearance

Line Length (30 – 225): 50
Blocking (0 – 50): 0
Numbering (5 – 100): 10
Marking (0 – 100): 0

Characters

AA code letters: ● one ○ three
Strandedness: ● double ○ single

Displayed Feature Types

☑ binding ☑ repeat
☑ general ☑ RNA
☑ promoter ☑ translation
☑ protein ☑ other
☑ region ☑ frag

Translate (DNA/RNA only)

☑ CDS ☑ mat_peptide
□ exon ☑ sig_peptide
□ block to phase

Formatting residue sequences

The sequence layout can affect the clarity of the displayed information, particularly if a lot of feature information is also visible.

To format a residue sequence

1. Choose **MacVector | Preferences** from the menu, then click the **Text Display** icon on the preferences dialog, or click the **Prefs** icon on the toolbar when the **Features** tab is selected.

The **Text Display** preferences dialog box is displayed.

Note. You can also access the **Text Display** preferences dialog using **Options | Format Annotated Display** from the menu.

2. Type a number in the **Line Length** text box to specify the number of residues per line.
3. Type a number in the **Blocking** text box to draw residues in groups of that number.
4. Type a number in the **Numbering** text box to number the sequence residues at that interval.
5. Type a number in the **Marking** text box to place an asterisk (*) at that interval along the sequence.
6. This is often set at half the interval of the **Numbering** value.
7. Select the **Letter case** as **upper** or **lower** as required.
8. Select the display of amino acid codes as **one** or **three** letters when a nucleic acid sequence translation is displayed.
9. Select either **single** or **double** to display a nucleic acid sequence as either a single- or double-stranded molecule.
10. Select **OK** to apply the formatting.

Formatting features

When the sequence is displayed, you can control the type of features that are shown beneath the sequence. The features must be present in the features table of the sequence.

To display features as sequence annotations

1. Choose **MacVector | Preferences** from the menu, then click the **Text Display** icon on the preferences dialog, or click the **Prefs** icon on the toolbar when the **Features** tab is selected.

The **Text Display** preferences dialog box is displayed.

Note. You can also access the **Text Display** preferences dialog using **Options | Format Annotated Display** from the menu.

2. In the **Displayed Feature Types** panel, select and deselect check boxes as required.

Selected features that occur in the sequence features table will appear as annotations to the sequence in the display.

3. Select **OK** to apply the formatting.

The features are displayed below the appropriate residues.

Note. If two or more features overlap, one appears below the other.

You can also view the translation of certain features, if they occur in the sequence's features table. These features are:

- mature protein product (mat_peptide)
- protein coding region (CDS)
- signal peptide (sig_peptide)
- exon regions (exon)

To display feature translations as sequence annotations

1. Choose **MacVector | Preferences** from the menu, then click the **Text Display** icon on the preferences dialog, or click the **Prefs** icon on the toolbar when the **Features** tab is selected.

The **Text Display** preferences dialog box is displayed.

Note. You can also access the **Text Display** preferences dialog using **Options | Format Annotated Display** from the menu.

2. In the **Translate** panel, select and deselect check boxes as required.
3. To display the translated region in codon groups, select **block to phase**.

This display style overrides the **Blocking** value set for the remainder of the sequence.

4. Select **OK** to apply the formatting.

Translatable features of the selected types are displayed in translated form beneath the main sequence.

To change the font used

1. Choose **MacVector | Preferences** from the menu, then click the **Fonts** icon on the preferences dialog.

The **Fonts** preferences dialog box is displayed.

2. In the **Result window font** panel, use the drop down menu to change the font.
3. Check the **Enable anti-aliasing** option to enable anti-aliasing for displaying the font. This makes fonts appear smoother at the expense of

reduced sharpness. High resolution monitors may display better without this option.

4. Select **OK** to apply the formatting.

Note. Certain printer drivers do not correctly align Monaco text when printed out, although the alignments look perfect on the screen. Andale Mono is a good substitute that is correctly aligned on all printers we have tested to date.

Using auto-annotation to add features and feature appearance information

The auto-annotation tool allows you to compare a bare sequence with annotated sequences in a specified folder and automatically annotate the bare sequence with any matching features found on sequences in the folder. It simplifies the process of annotating sequences downloaded from GenBank and other sources consistently.

Additionally, the feature appearance information associated with the any matching features is also copied to the bare sequence. This means that the auto-annotation tool can also be used to automatically adjust the appearance of existing features to suit your personal preferences, so that, for example, Ampicillin resistance genes always appear as green arrows in your sequence files.

How auto-annotation works

The auto-annotation algorithm reads each file in the specified sequence folder and determines if it encodes a DNA sequence containing annotated features. It then works through each feature in the DNA, copying the sequence corresponding to the feature and comparing it to the new sequence. If a match is found, then the feature is copied to the new sequence.

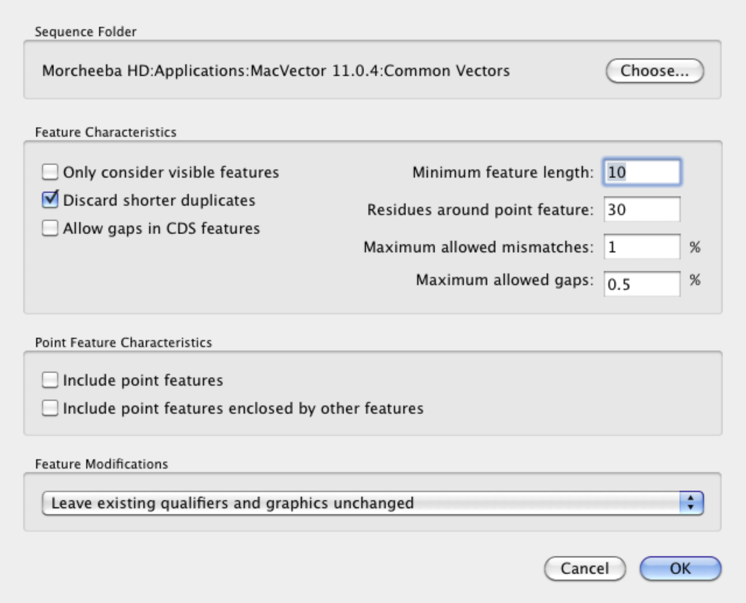
There are a couple of things to note about the auto-annotation algorithm:

- It examines all sub-folders within the sequence folder you specify. However, any aliases or shortcuts to other folders are ignored.
- It discards duplicate features, which are those of identical type that start and stop at the same location on the same strand of the sequence. It does not consider the description or qualifiers associated with a feature when determining which are duplicates.
- If a matching feature is found in a MacVector format file, then the graphical appearance information associated with that feature is also copied to the new sequence.

To use auto-annotation to add features to a new sequence

1. Ensure the required Sequence window is active.
2. Click the **Auto-Annotate** icon on the toolbar or select **Database | Auto-Annotate Sequence...** from the menu.

The **Auto-Annotation** parameters are displayed.



Sequence Folder

Morcheeba HD:Applications:MacVector 11.0.4:Common Vectors

Feature Characteristics

☐ Only consider visible features
☑ Discard shorter duplicates
☐ Allow gaps in CDS features

Minimum feature length: 10
Residues around point feature: 30
Maximum allowed mismatches: 1 %
Maximum allowed gaps: 0.5 %

Point Feature Characteristics

☐ Include point features
☐ Include point features enclosed by other features

Feature Modifications

Leave existing qualifiers and graphics unchanged

The **Sequence Folder** section displays the name of the folder containing the annotated sequences that will be compared with the new sequence.

3. Optionally, click the **Choose** button to specify a new folder.

The **Feature Characteristics** section enables you to control which types of features will be added.

4. Check **Only consider visible features** to limit the features added to only those that are visible.

This option is useful if you want to avoid cluttering your sequences with large numbers of hidden features. For example, the standard GenBank pBR322 vector sequence has around 60 features assigned to it, but only four are displayed in most schematic diagrams of the plasmid.

5. Check **Discard shorter duplicates** to discard any features that lie entirely within another feature of the same type, so that only the longest feature is retained.

Normally, if features have different start or stop locations, they are considered to be different. However, many vectors have slight differences in the extent of the replication origin or in features such as T7 or SP6 promoters. So, when comparisons are made with the numerous annotated sequences in the sequence folder, it is possible for the same feature to be annotated several times, with each annotation differing by one or two residues at either end. Checking this option removes such duplicates.

6. Check **Allow gaps in CDS features** to permit gaps in CDS features when searching for matches between features in the sequence folder and the new sequence.

In general, MacVector incorporates some fuzziness in the identification of matching features, allowing a limited number of gaps and mismatches in the alignment. However, it does not usually permit gaps in CDS features because these give rise to frameshifts in the encoded protein, potentially leading to something completely different to that which was encoded by the original annotated feature. Checking this option removes this restriction, which could be useful if you suspect that the new sequence may have sequencing errors, since allowing gaps in CDS features should ensure that they are annotated as expected.

7. Specify the **Minimum feature length** you want the auto-annotation tool to identify.

The auto-annotation algorithm uses sequence similarity to determine if a feature is present in the new sequence. This matching method can, sometimes, lead to very short features being incorrectly added. Consider, for example, a 4 base pair misc_feature being used to label an important MboI site in a sequence in the sequence folder. If no minimum feature length is specified, then that feature will be added at every MboI (GATC) site in the new sequence – every 256bp on average.

8. Specify the number of **Residues around point feature** to be examined when looking for a point feature match in the new sequence.

A point feature is one where the start and stop location are the same. These might be SNP locations, a replication start site, or just a particular point of interest. MacVector treats these as a special case.When the algorithm encounters a point feature, it examines the region containing the specified number of residues centered on that point and uses it to determine if the point feature is present in the new sequence.

Note. This parameter is used only when the **Include point features** option in the **Point Feature Characteristics** section (described below) is selected.

9. Specify the percentage of mismatched residues permitted in the **Maximum allowed mismatches** field.

MacVector incorporates some fuzziness in the identification of matching features, allowing a limited number of gaps and mismatches in the alignment. This parameters controls the number of mismatches permitted in matching features. The default value of 1 means that only 1 residue in 100 can be mismatched.

10. Specify the percentage of gaps permitted in the **Maximum allowed gaps** field.

This parameters controls the number of gaps permitted in matching features. The default value of 0.5 means that only 1 gap in 200 residues is permitted.

The **Point Feature Characteristics** section enables you to control whether and how point features are included in the search.

11. Check **Include point features** to include point features in the search.

Note. If this option is selected, then the **Residues around point feature** setting provided in the **Feature Characteristics** section (described above) will be used to identify matches.

12. Alternatively, check **Include point features enclosed by other features** to include point features only if the feature that encloses them is added.

For example, suppose you have SNPs annotated within a CDS feature. If the CDS feature gets added to the new sequence, then so do all the SNPs that lie within it.

The **Feature Modifications** section enables you to adjust the way the auto-annotate algorithm handles any existing features in the new sequence.

13. Select **Leave existing qualifiers and graphics unchanged** to discard matching features if the new sequence already has features of the same type at the same location.

Use this option to make sure that no existing features are changed while still allowing new features to be added.

14. Alternatively, select **Replace qualifiers and graphics for existing features** to replace existing features in the new sequence with matching features of the same type at the same location.

Note. Features are replaced but not removed. If there are existing features that do not match any features in the sequence folder, then they are retained unchanged.

Tip. To auto-annotate a sequence “from scratch”, you should first delete all existing features, then run the auto-annotation algorithm.

15. Or, select **Replace only graphics for existing features** to retain all of the qualifier and description information associated with any existing features but replace the graphical appearance information with that of the matching feature.

This option is particularly useful if you have downloaded a sequence from Entrez or imported a GenBank or EMBL format file, since it ensures that the feature takes on the graphical appearance you prefer, without losing any textual annotations.

16. Click **OK** to run the auto-annotation algorithm or **Cancel** to dismiss the dialog without running the algorithm.

A **Summary** dialog providing statistics about the numbers of features considered, matched, and ignored is displayed when the analysis is complete.

Limitations of auto-annotation

A limitation of the auto-annotation function is that the entire feature from the scanned folder must be present in the target sequence uninterrupted before it will be added. If even a single residue is missing from one end, it will not be considered a full-length match and will not be annotated. Similarly, if a feature has been interrupted, e.g. because you inserted a fragment of DNA into a gene, neither part of the feature will be annotated.

Importing features from a BED/GFF/GTF/GFF3 file

Import Features allows you to annotate an unannotated or partially annotated sequence with annotations (or features) contained in a BED/GFF/GTF/GFF3 file.

To annotate a sequence with a BED/GFF/GTF/GFF3 file

1. Open the sequence file you want to annotate.
2. Select **File | Import Features** from the menu bar.

3. Locate and select the BED/GFF/GTF/GFF3 file you want to import features from and click **Open**.

The Sequence IDs (SeqIDs) for all the features in the selected BED/GFF/GTF/GFF3 file are shown, along with the number of features for each SeqID and the region of the sequence that the annotations span. A warning is displayed if any of the features are outside the range of the selected sequence.

4. Select the SeqID you want to import features from and click **OK**.

The number of annotations that have been added is displayed.

If you annotate a blank sequence, for example a Fastq file, then the resulting features may be hidden. Use the **Graphics Palette** tree view to show any hidden features.

You can choose to annotate your sequence with all the features contained in the selected BED/GFF/GTF/GFF3 file or you can use only the unique features.

Note. Due to the lack of strict standards across many of these file formats it is possible for potential duplicates to be missed and some degree of manual curation of the annotated sequence will be required. However, potentially interesting or important information is never discarded. Only entires which are 100% identical are identified as duplicates. A feature will not be classed as a duplicate if the **Start**, **Stop** or **Feature Type** are different in any way.

Searching and Downloading Sequences in the *Entrez* Database

Overview

This chapter describes how to locate and extract complete sequences from the *Entrez* database maintained by NCBI. You can do the following:

- find sequences by querying any of the *Entrez* database fields
- find PUBMED abstracts in the *Entrez* database using any of the *Entrez* database fields
- extract sequences and abstracts from the database to the desktop, or save them to disk.

Refer to*"Subsequence analysis"* on page 181, for details of subsequence searching.

Refer to Chapter 12, *"Aligning Sequences"*, for a description of the sequence comparison methods used by MacVector.

Contents

The *Entrez* database

The *Entrez* database, produced by the National Center for Biotechnology Information (NCBI), provides access to DNA and protein sequences and related bibliographic information. The sequence data include the complete nucleotide and protein sequence data from the GenBank, EMBL, DDBJ, PIR, PRF, PDB, SWISS-PROT, dbEST and dbSTS databases, as well as data from U.S. and European patents. *Entrez* also contains a subset of the PUBMED database, including references and abstracts which are cited in the sequence databases and other related PUBMED records.

MacVector can search the *Entrez* database on the Internet. The data is updated on a daily basis, and the service is available free of charge.

Refer to Appendix A, *“Setting up NCBI's Entrez and BLAST Services”*, for details of accessing the database on the Internet.

Searching the *Entrez* Database

Choosing the database

The MacVector interface enables you to search the different databases available through the NCBI separately, providing you with access to the full range of information available.

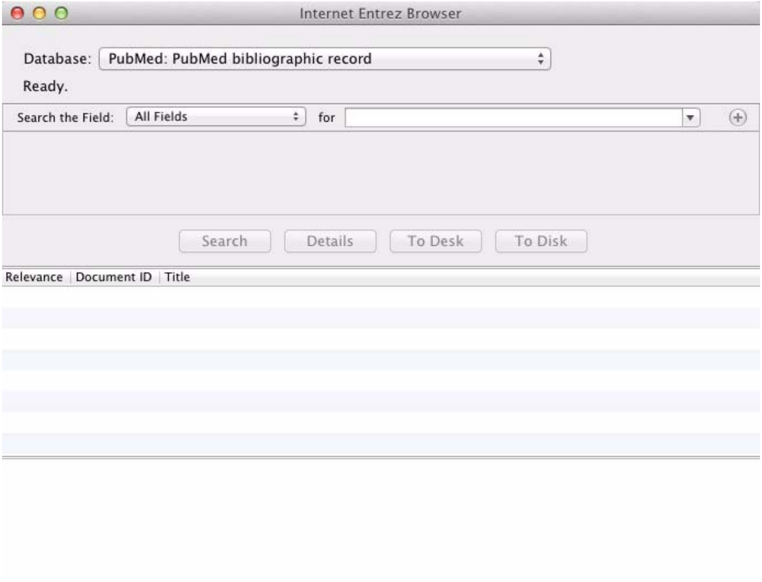
Note. Not all of the databases available through the NCBI are sequence databases (e.g. Index of NCBI web pages).

The precise list of databases you can search is updated from the NCBI's servers each time you launch an *Entrez* search.

To choose the database

1. Choose **Database | Internet Entrez Search**.

The **Internet Entrez Browser** window is displayed.



Database: PubMed: PubMed bibliographic record

Search the Field: All Fields

2. Select the database you want to search using the **Database** drop-down menu.

Among the most useful databases available are:

Protein sequence record

The protein entries in the *Entrez* search and retrieval system have been compiled from a variety of sources, including SwissProt, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq.

Core nucleotide db

The core nucleotide database contains records for all *Entrez* nucleotide sequences that are not found within the EST or GSS divisions of GenBank. These include sequences from all the remaining divisions of GenBank, NCBI Reference Sequences (RefSeqs), Whole Genome Shotgun (WGS) sequences, Third Party Annotation (TPA) sequences, and sequences imported from the Entrez Structure database.

GSS db

The GSS database contains all records found within the Genome Survey Sequence division of GenBank. GSS records contain first-pass single-read genomic sequences and rarely include annotated biological features.

The GSS division contains (but is not limited to) the following types of data:

- random "single pass read" genome survey sequences.
- cosmid/BAC/YAC end sequences
- exon trapped genomic sequences
- Alu PCR sequences
- transposon-tagged sequences

EST db

The EST database contains all records found within the Expressed Sequence Tags division of GenBank. EST records contain first-pass single-read cDNA sequences and include no annotated biological features.

Full nucleotide db

The full nucleotide database is a superset of the Core nucleotide db, GSS db and EST db.

Performing the search

The set of three drop-down menus enables you to define a single or combined (Boolean) annotation search of the selected database. The two **All Fields** drop-down menus contain a list of all possible search categories for the database you are searching.

The other drop-down menu contains logical operators that enable you to perform more complicated searches using two criteria. For example, you can extract all ribosomal protein sequences from the organism "*Canis*" using the following settings:

- "canis*" in the **organism** field
- the **And** operator in the central drop-down menu
- "ribosomal" in the **indexed words** field.

To perform a single category search

1. Choose a search category in the first **All Fields** drop-down menu.
2. In the adjacent text box, type the required search text.

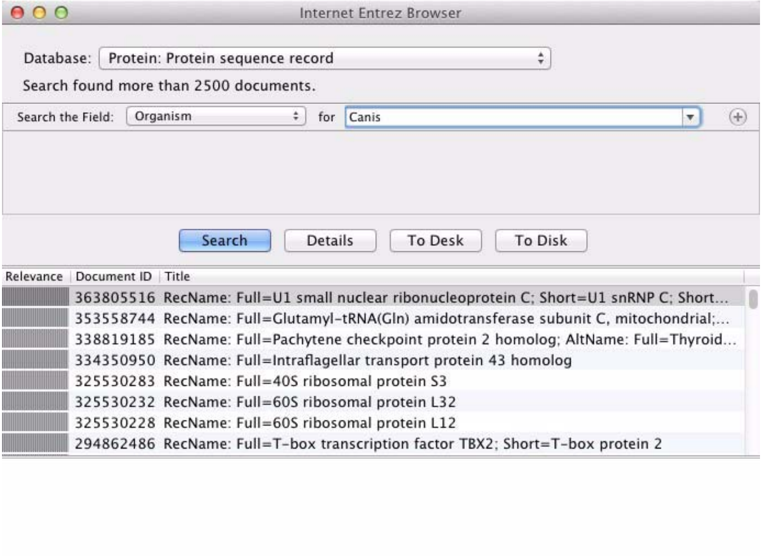
Only one string is allowed as a search query for each field, except for the entry definition field, when separate search strings can be separated by a single space.

3. Ensure that the logical operator drop-down menu is set to **None**.
4. Select **Search** to perform the search.

During the search, the **Search** button label changes to **Stop**, allowing you to stop the search.

If MacVector finds any matches to your query, it will list them in the top text panel. The **Title**, **Document ID** and the **Relevance** of the match to your search text are listed for each match.

Note. If you save the sequence as a file the **Document ID** number will be used to name the file (see *“Saving sequences to file”* on page 139).



Database: Protein: Protein sequence record

Search found more than 2500 documents.

Search the Field: Organism for Canis

Relevance	Document ID	Title
▇	363805516	RecName: Full=U1 small nuclear ribonucleoprotein C; Short=U1 snRNP C; Short...
▇	353558744	RecName: Full=Glutamyl-tRNA(Gln) amidotransferase subunit C, mitochondrial;...
▇	338819185	RecName: Full=Pachytene checkpoint protein 2 homolog; AltName: Full=Thyroid...
▇	334350950	RecName: Full=Intraflagellar transport protein 43 homolog
▇	325530283	RecName: Full=40S ribosomal protein S3
▇	325530232	RecName: Full=60S ribosomal protein L32
▇	325530228	RecName: Full=60S ribosomal protein L12
▇	294862486	RecName: Full=T-box transcription factor TBX2; Short=T-box protein 2

To perform a combined category search

1. Choose a search category in the first **All Fields** drop-down menu.
2. In the adjacent text box, type the required search text.

Only one string is allowed as a search query for each field, except for the entry definition field, when separate search strings may be separated by a single space.

3. Choose a logical operator for combining the two fields from the logical operator drop-down menu.
4. Choose a search category in the second **All Fields** drop-down menu.
5. In the adjacent text box, type the required search text.
6. Select **Search** to perform the search.

During the search, the **Search** button label changes to **Stop**.

If MacVector finds any matches to your query, it will list them in the top text panel. The **Title**, **Document ID** and the **Relevance** of the match to your search text are listed for each match.

Note. If you save the sequence as a file the **Document ID** number will be used to name the file (see *"Saving sequences to file"* on page 139).

Viewing details of search results

The top text panel of the **Internet Entrez Browser** window contains a scrollable list of the matches to your query.

To view details of search results

1. Select the required results in the results panel, by clicking or shift-clicking.
2. Select the **Details** button.

The relevant NCBI web page is displayed in the bottom text panel of the **Internet Entrez Browser** window.

Extracting information from the *Entrez* database

When you have a set of results from your search, you can do the following:

- extract one or more sequences into individual Sequence windows
- save one or more sequences into individual sequence files
- extract abstracts into individual Text windows
- save abstracts as individual text files
- extract abstracts cited in sequence annotations

Extracting sequences to the desktop

To extract sequence search results into sequence windows

1. Select the required results in the results panel, by clicking or shift-clicking.

2. Select the **To Desk** button.

The selected sequences will appear on the desktop in individual Sequence windows.

Tip. Instead of using the **To Desk** command, you can extract a sequence just by double-clicking on the entry in the results panel.

Saving sequences to file

To save sequence search results into sequence files

1. Select the required results in the results panel, by clicking or shift-clicking.
2. Select the **To Disk** button.

A dialog box is displayed, enabling you to choose a folder in which to store the sequence files.

3. To create a new folder, select the **New** button and type in a name.
4. Select **Choose**.

Each sequence is saved to its own file. The file is named using the **Document ID** number of the sequence.

Extracting abstracts to the desk top

To extract PUBMED search results into text windows

1. Select the required results in the results panel, by clicking or shift-clicking.
2. Select the **To Desk** button.

The selected abstracts will appear on the desktop in individual text windows.

Tip. Instead of using the **To Desk** command, you can extract an abstract just by double-clicking on the entry in the results panel.

Saving abstracts to file

To save PUBMED search results into files

1. Select the required results in the results panel, by clicking or shift-clicking.
2. Select the **To Disk** button.

A dialog box is displayed, enabling you to choose a folder in which to store the abstract files.

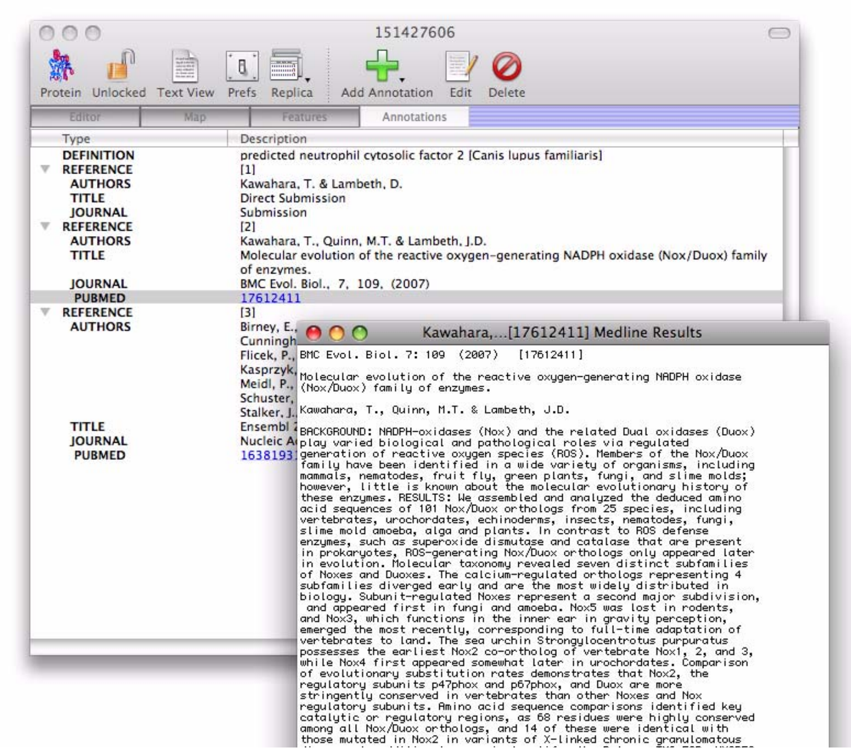
3. To create a new folder, select the **New** button and type in a name.

4. Select **Save**.

Each abstract is saved to its own file. The file is named using the **PUBMED ID** number of the abstract.

Extracting PUBMED abstracts cited in sequence annotations

If you have extracted a sequence that cites a PUBMED reference, you can retrieve the PUBMED abstract directly from the Annotations view of the sequence.



Type	Description
DEFINITION	predicted neutrophil cytosolic factor 2 [Canis lupus familiaris]
REFERENCE	[1]
AUTHORS	Kawahara, T. & Lambeth, D.
TITLE	Direct Submission
JOURNAL	Submission
REFERENCE	[2]
AUTHORS	Kawahara, T., Quinn, M.T. & Lambeth, J.D.
TITLE	Molecular evolution of the reactive oxygen-generating NADPH oxidase (Nox/Duox) family of enzymes.
JOURNAL	BMC Evol. Biol., 7, 109, (2007)
PUBMED	17612411
REFERENCE	[3]
AUTHORS	Birney, E., Cunningh Flicek, P., Kasprzyk, Meidl, P., Schuster, Stalker, J.
TITLE	Ensembl 2
JOURNAL	Nucleic A
PUBMED	1638193

Kawahara,...[17612411] Medline Results

```
BMC Evol. Biol. 7: 109  (2007)   [17612411]

Molecular evolution of the reactive oxygen-generating NADPH oxidase
(Nox/Duox) family of enzymes.

Kawahara, T., Quinn, M.T. & Lambeth, J.D.

BACKGROUND: NADPH-oxidases (Nox) and the related Dual oxidases (Duox)
play varied biological and pathological roles via regulated
generation of reactive oxygen species (ROS). Members of the Nox/Duox
family have been identified in a wide variety of organisms, including
mammals, nematodes, fruit fly, green plants, fungi, and slime molds;
however, little is known about the molecular evolutionary history of
these enzymes. RESULTS: We assembled and analyzed the deduced amino
acid sequences of 101 Nox/Duox orthologs from 25 species, including
vertebrates, urochordates, echinoderms, insects, nematodes, fungi,
slime mold amoeba, alga and plants. In contrast to ROS defense
enzymes, such as superoxide dismutase and catalase that are present
in prokaryotes, ROS-generating Nox/Duox orthologs only appeared later
in evolution. Molecular taxonomy revealed seven distinct subfamilies
of Noxes and Duoxes. The calcium-regulated orthologs representing 4
subfamilies diverged early and are the most widely distributed in
biology. Subunit-regulated Noxes represent a second major subdivision,
 and appeared first in fungi and amoeba. Nox5 was lost in rodents,
and Nox3, which functions in the inner ear in gravity perception,
emerged the most recently, corresponding to full-time adaptation of
vertebrates to land. The sea urchin Strongylocentrotus purpuratus
possesses the earliest Nox2 co-ortholog of vertebrate Nox1, 2, and 3,
while Nox4 first appeared somewhat later in urochordates. Comparison
of evolutionary substitution rates demonstrates that Nox2, the
regulatory subunits p47phox and p67phox, and Duox are more
stringently conserved in vertebrates than other Noxes and Nox
regulatory subunits. Amino acid sequence comparisons identified key
catalytic or regulatory regions, as 68 residues were highly conserved
among all Nox/Duox orthologs, and 14 of these were identical with
those mutated in Nox2 in variants of X-linked chronic granulomatous
```

To extract PUBMED abstracts from a sequence window

1. Ensure that the required sequence is displayed in the active window.
2. Select the **Annotations** tab.
3. Scroll through the annotations and identify the PUBMED reference.
4. Double-click on the reference.

MacVector retrieves the PUBMED reference from the *Entrez* database, and the abstract is displayed in a new text window.

7 Calculating Sequence Properties

Overview

This chapter describes how to:

- calculate sequence property analyses
- perform base composition analysis
- display and manage the results of a sequence property profile calculation
- find open reading frames using user-designated start and stop codons
- find regions of a nucleic acid that are likely to code for protein according to Fickett’s TESTCODE algorithm
- create codon preference plots to locate regions that have a high probability of coding for a highly expressed protein

Refer to Chapter 5, *“Working with Sequences and Features”*, for details about opening and editing sequence files before analysis.

Contents

Sequence Properties

MacVector contains many methods for analyzing the properties and composition of proteins and nucleic acid sequences. Detailed information about these analyses is available in Chapter 20, *“Understanding Protein and DNA Analysis”*.

Protein sequence properties

The available methods for predicting sequence properties for protein sequences are summarized in the following sections.

Antigenicity

Hopp-Woods

The Hopp-Woods scale was designed to predict the locations of antigenic determinants in a protein, assuming that the antigenic determinants would be exposed on the surface of the protein and thus would be located in hydrophilic regions. Its values are derived from the transfer free energies for amino acid side chains between ethanol and water.

Parker

Predicts the location of antigenic determinants by finding the area of greatest local hydrophilicity. It is based on the Hopp-Woods method, but differs in that it uses a modified hydrophilicity scale, based on the HPLC retention times of model peptides.

Protrusion

Uses the Protrusion Index, which is an antigenic scale based on a study of proteins with known 3D structure. The tendency of each residue to be located in a protruding region of the protein can be calculated using this method.

Welling

Calculates a statistical score, where the antigenicity value for each residue is calculated as the log of the quotient between its percentage in a sample of known antigenic regions and its percentage in average proteins.

Antigenic Index

Uses a modified version of the Jameson and Wolf algorithm, which combines hydrophilicity, surface, flexibility and secondary structure predictions to produce a composite surface contour of a protein.

Flexibility

Protein flexibility The Karplus and Schulz method measures the flexibility of each residue and assigns each residue a class (rigid, intermediate, or mobile) based on the flexibility of the residue and its two neighbors. Flexibility is a useful predictive parameter of antigenicity.

Secondary Structure

Chou-Fasman Uses the tendency of an amino acid to appear in a given secondary structure in known X-ray structures to predict unknown structures.

Robson-Garnier Predicts secondary structure based upon the effect that each amino acid has on the conformational state of its neighbors.

Hydrophobicity

Fauchere-Pliska Uses a hydrophobicity scale based on the experimental octanol/water partition of the N-acetyl-amino-acid amides of each residue at neutral pH. Each hydrophobicity value is expressed with respect to glycine (which scores 0), so positive values indicate a greater hydrophobicity than glycine.

Janin A method based on the accessibility of residues, with a molar fraction (%) of the occurrences of buried (< 20 Angstroms exposed surface area) and exposed (> 20 Angstroms exposed surface area) residues being derived.

Kyte-Doolittle The Kyte-Doolittle scale is the most commonly used hydropathy scale. Its values are assigned using a combination of the water-vapor transfer free energies for amino acid side chains and the preference of amino acid side chains for interior or exterior environments. Small adjustments are made to the final values based on the experience of the authors.

Manalavan A method based on a 'bulk hydrophobicity character', as the hydrophobicity of an individual amino acid residue is modified by the presence of other residues within an 8 Angstrom radius.

Sweet-Eisenberg Uses a consensus scale comprising four other hydrophobicity scales, including the Janin scale and the von Heijne scale. The method averages the four scales to reduce the effect of outlier values on the predictions.

von Heijne Uses a scale that reflects the estimated free energies of transfer of residues when moving from an alpha-helix in water to an alpha-helix in a non-polar phase (with no hydrogen bonding capacity).

Hydrophilicity

Hopp-Woods The Hopp-Woods scale was designed to predict the locations of antigenic determinants in a protein, assuming that the antigenic determinants would be exposed on the surface of the protein and thus would be located in hydrophilic regions. Its values are derived from the transfer free energies for amino acid side chains between ethanol and water.

Amphiphilicity

Amphiphilicity Helix Uses the periodicity in the protein's hydrophobicity in alpha-helical regions to calculate amphiphilicity.

Amphiphilicity Sheet Uses the periodicity in the protein's hydrophobicity in beta sheet regions to calculate amphiphilicity.

Transmembrane

Argos Helix Predicts membrane-bound, helical sequence regions, based on data from proteins which are known to interact with membranes.

von Heijne Helix Weights the contribution from each residue, so that the residues in the central apolar region of the bilayer dominate the prediction. However, good noise reduction is also obtained.

Goldman-Engelman-Steitz (GES)

The GES scale was developed in order to identify possible transmembrane helices in a protein. The scale values are the sums of hydrophobic and hydrophilic components for each amino acid. Hydrophobic components are derived from the free energy of water-oil transfer for the side chains; hydrophilic components take into consideration the free energy for inserting charged groups into a bilayer and the free-energy contributions from hydrogen-bonding with water and with backbone carbonyl groups (if the residues can form such bonds when participating in a helical structure).

Surface

Surface probability

This profile was designed to predict which regions of a protein are most likely to lie on the protein's surface, based on knowledge of which amino acids are more likely to be found on the surface of proteins of known structure. It is based on the work of Janin *et al* (1978) and Emini *et al* (1965).

General

Composition

This profile generates a listing of the amino acid composition.

MW

This profile calculates the protein molecular weight.

pI

This calculates the pH at which the protein has a net charge of zero.

Nucleic acid sequence properties

The following types of sequence property profile calculation are available for nucleic acid sequences:

- percentage base composition with respect to any combination of nucleotides
- mono-, di-, and trinucleotide frequencies
- melting or dissociation temperature of a nucleic acid sequence or oligo, including a running average for the sequence
- Coding preference plots

Window size

The window size determines the number of residues that are grouped together for analysis. For example, a window size of 3 will result in the sequence being analyzed in blocks of 3 residues, throughout the length of the sequence.

The percentage base composition is calculated on the basis of the window size that you specify in the **Base Composition Analysis** dialog box.

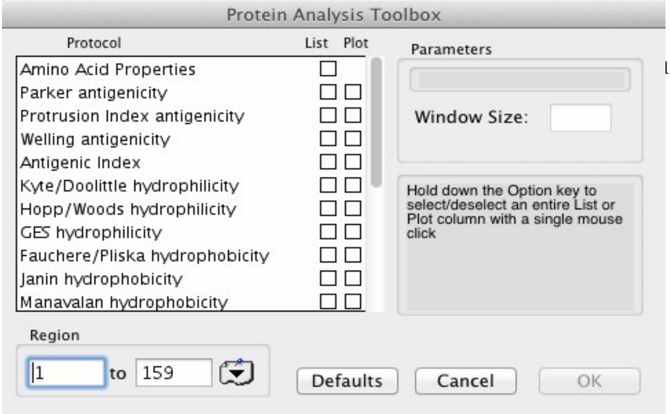
You can select a different window size for each profile algorithm.

Note. You cannot change the window size for the following protein algorithms: Protein flexibility, von Heijne transmembrane flexibility, and Argos transmembrane flexibility. These methods use a fixed window size.

Performing protein sequence analyses

Protein analyses are accessible by choosing **Analyze | Protein Analysis Toolbox**. This item is enabled when a protein sequence file is the active window.

When a protein sequence is the active window, one or more analyses can be performed for the sequence.



Protein Analysis Toolbox

Protocol	List	Plot
Amino Acid Properties	□	
Parker antigenicity	□	□
Protrusion Index antigenicity	□	□
Welling antigenicity	□	□
Antigenic Index	□	□
Kyte/Doolittle hydrophilicity	□	□
Hopp/Woods hydrophilicity	□	□
GES hydrophilicity	□	□
Fauchere/Pliska hydrophobicity	□	□
Janin hydrophobicity	□	□
Manavalan hydrophobicity	□	□

Parameters

Window Size:

Hold down the Option key to select/deselect an entire List or Plot column with a single mouse click

Region

1 to 159

ıe

To perform a protein sequence analysis

1. Choose **Analyze | Protein Analysis Toolbox**.

The **Protein Analysis Toolbox** dialog box is displayed.

2. Scroll through the **Protocol** list and select the profile you want to calculate.

When a profile is selected, a brief description of it is displayed in the text panel to the right of the **Protocol** list.

3. Select the **List** and **Plot** check boxes as necessary for a text listing or graphical view of the results.

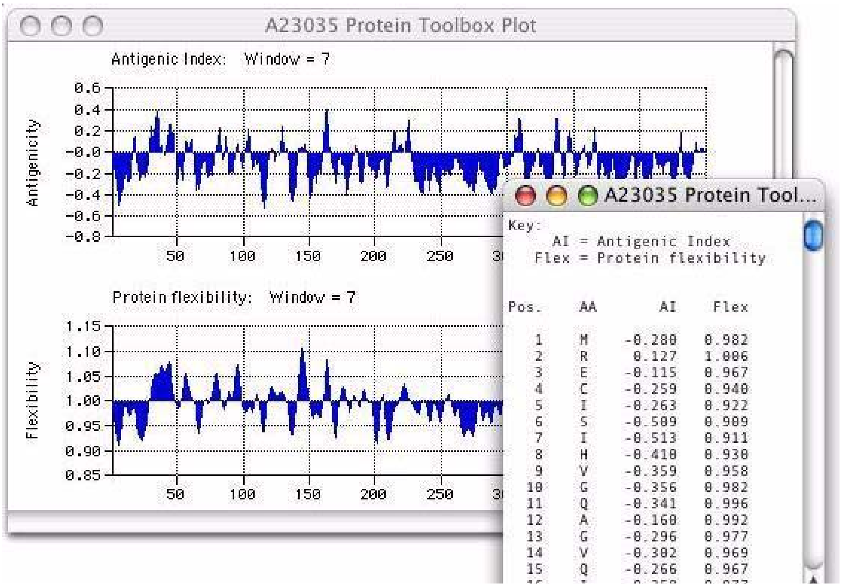
To select or deselect all checkboxes in a column, hold down the **option** key while clicking on any checkbox in the column.

4. Repeat steps 2 and 3 for each profile you want to calculate.
5. Choose the portion of the sequence that you want to analyze by typing in the sequence numbers that bracket the region in the **Region** text boxes. Alternatively, select a region from the feature selector drop-down menu to the right of the text boxes.

Note. Some profiles require a minimum sequence length, or window, for their calculation. A selection smaller than this value is the most common cause of the calculation failing.

6. If you want to use MacVector default settings, select **Defaults**.
7. Select **OK** to perform the analysis.

Note. The **OK** button will only be enabled if at least one **List** or **Plot** checkbox is selected.



```
Key:
   AI = Antigenic Index
 Flex = Protein flexibility


Pos.    AA        AI     Flex

   1    M     -0.280    0.982
   2    R      0.127    1.006
   3    E     -0.115    0.967
   4    C     -0.259    0.940
   5    I     -0.263    0.922
   6    S     -0.509    0.909
   7    I     -0.513    0.911
   8    H     -0.410    0.930
   9    V     -0.359    0.958
  10    G     -0.356    0.982
  11    Q     -0.341    0.996
  12    A     -0.160    0.992
  13    G     -0.296    0.977
  14    V     -0.302    0.969
  15    Q     -0.266    0.967
```

Viewing Protein Analysis Toolbox results

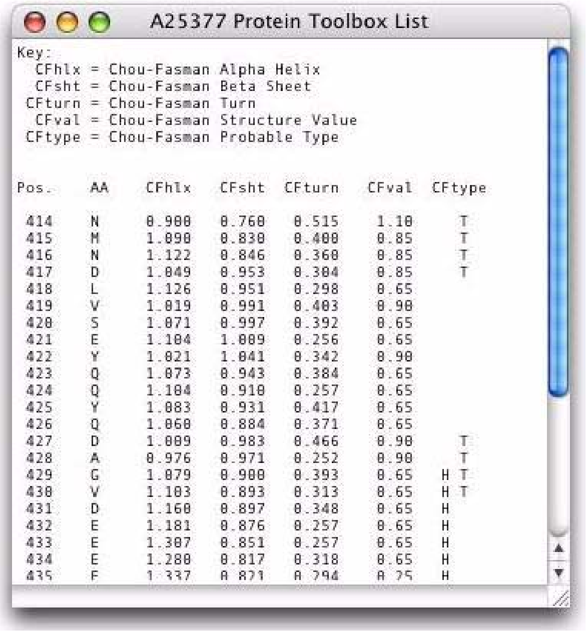
The results of the analysis are displayed automatically. Plots generated by selecting the **Plot** check box are displayed in a graphical Toolbox Plot window. Numerical results generated by selecting the **List** check box are displayed as tables in a scrollable Toolbox List text window.

Further viewing options, and saving and printing results are described in Chapter 3, *“General Procedures”*.

Protein Toolbox Lists

There are a few points to note about protein Toolbox Lists:

- each table column is in units specific to the analysis
- column abbreviations are explained at the top of the list
- the table may print incorrectly with large numbers of profiles. This will depend on your printer page width settings.

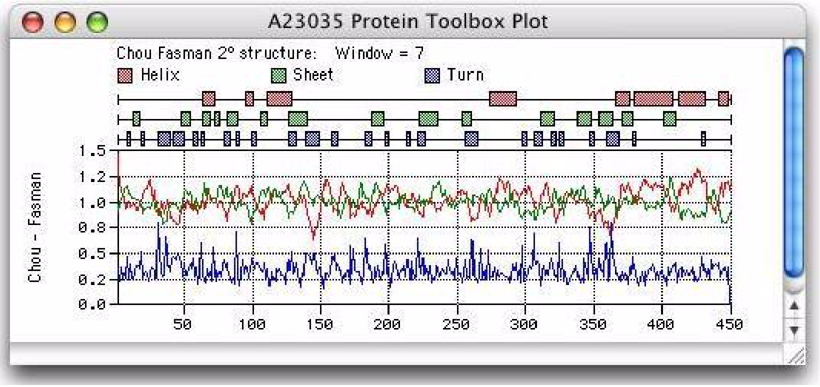


```
Key:
  CFhlx = Chou-Fasman Alpha Helix
  CFsht = Chou-Fasman Beta Sheet
 CFturn = Chou-Fasman Turn
  CFval = Chou-Fasman Structure Value
 CFtype = Chou-Fasman Probable Type


Pos.    AA    CFhlx   CFsht   CFturn    CFval   CFtype

 414    N     0.900   0.760   0.515     1.10       T
 415    M     1.090   0.830   0.400     0.85       T
 416    N     1.122   0.846   0.360     0.85       T
 417    D     1.049   0.953   0.304     0.85       T
 418    L     1.126   0.951   0.298     0.65
 419    V     1.019   0.991   0.403     0.90
 420    S     1.071   0.997   0.392     0.65
 421    E     1.104   1.009   0.256     0.65
 422    Y     1.021   1.041   0.342     0.90
 423    Q     1.073   0.943   0.384     0.65
 424    Q     1.104   0.910   0.257     0.65
 425    Y     1.083   0.931   0.417     0.65
 426    Q     1.060   0.884   0.371     0.65
 427    D     1.009   0.983   0.466     0.90       T
 428    A     0.976   0.971   0.252     0.90       T
 429    G     1.079   0.900   0.393     0.65     H T
 430    V     1.103   0.893   0.313     0.65     H T
 431    D     1.160   0.897   0.348     0.65     H
 432    E     1.181   0.876   0.257     0.65     H
 433    E     1.307   0.851   0.257     0.65     H
 434    E     1.280   0.817   0.318     0.65     H
 435    F     1.337   0.821   0.294     0.75     H
```

Protein Toolbox Plots

There are a number of points to note about protein Toolbox Plots:

- when using the zoom facility (see *"To magnify an area of the graphic"* on page 37), scales are dynamically recalculated to maximize the range
- the secondary structure plots have two parts: running average plots for the helix, coil and turn structures; and structure prediction bars across the top of the plot
- when both Robson-Garnier and Chou-Fasman plots are calculated, a consensus plot is also generated. The running average values are the product of the normalized values from each individual plot, and structure prediction bars are only drawn where both methods agree.



To navigate in the Protein Toolbox Plot window

1. **To magnify a selected region**, click and drag in the window
2. **To return to the whole plot**, double-click anywhere in the window
3. **To zoom in,** press the **Up-arrow** key (or ⌘ and **+** keys) to zoom by a factor of 2, or press **Shift+Up-arrow** to zoom to individual residues
4. **To zoom out,** press the **Down-arrow** key (or ⌘ and **-** keys) to zoom by a factor of 2, or press **Shift+Down-arrow** to zoom all the way out
5. **To shift to the right,** press the **Right-arrow** key to shift 10% to the right, or press **Shift+Right-arrow** to shift 80% to the right
6. **To shift to the left,** press the **Left-arrow** key to shift 10% to the left, or press **Shift+Left-arrow** to shift 80% to the left.

Nucleic acid analyses are accessible by choosing **Analyze | Base Composition**, **Analyze | ORF Analysis** or **Analyze | Nucleic Acid Analysis Toolbox**.

These items are enabled when a nucleic acid sequence file is the active window.

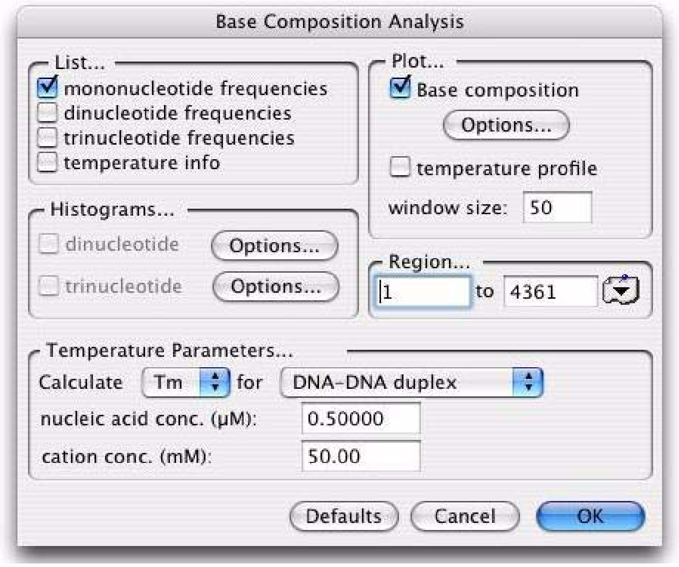
Nucleic acid analyses are accessible by choosing **Analyze | Base Composition**, **Analyze | ORF Analysis** or **Analyze | Nucleic Acid Analysis Toolbox**. These items are enabled when a nucleic acid sequence file is the active window.

Performing nucleic acid sequence analyses

Nucleic acid analyses are accessible by choosing **Analyze | Base Composition**, **Analyze | ORF Analysis** or **Analyze | Nucleic Acid Analysis Toolbox**. These items are enabled when a nucleic acid sequence file is the active window.

Performing base composition analysis

When a DNA sequence is the active window, one or more profiles can be calculated for the sequence's base composition.



Base Composition Analysis

List...
☑ mononucleotide frequencies
☐ dinucleotide frequencies
☐ trinucleotide frequencies
☐ temperature info

Plot...
☑ Base composition
Options...
☐ temperature profile
window size: 50

Histograms...
☐ dinucleotide Options...
☐ trinucleotide Options...

Region...
1 to 4361

Temperature Parameters...
Calculate Tm for DNA-DNA duplex
nucleic acid conc. (µM): 0.50000
cation conc. (mM): 50.00

Defaults Cancel OK

To perform a base composition analysis

1. Choose **Analyze | Base Composition** from the menu.

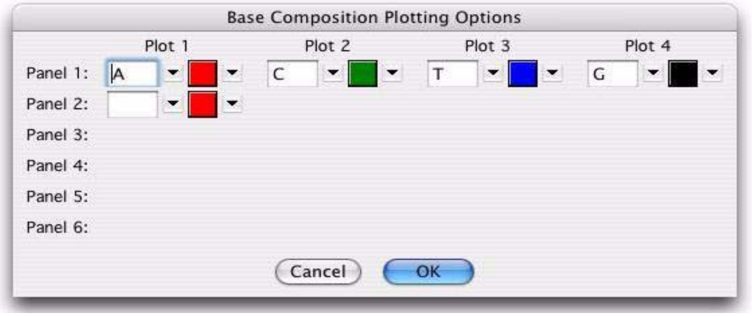
The **Base Composition Analysis** dialog box is displayed.

2. Select one or more check boxes in the **List** panel to generate a text listing of the required mono-, di- and trinucleotide frequencies and temperature information.
3. Select the **Base composition** check box to generate percentage composition plots.
4. Specify the required composition plots by selecting **Options** in the **Plot** panel. See*"Specifying base composition plots"* on page 152 for details of this procedure.
5. Select the **temperature profile** check box to generate a running average of the sequence Tm or Td.
6. Specify temperature profile parameters in the **Temperature Parameters** panel as follows:
   - choose **Tm** or **Td** from the **Calculate** drop-down menu
   - choose the duplex type from the adjacent drop-down menu
   - enter a value in the **nucleic acid conc**. text box
   - enter a value in the **cation conc**. text box.
7. Type a value in the **window size** text box for the number of residues over which properties are calculated.
8. Select the required check boxes in the **Histograms** panel to generate histograms for di- and trinucleotide repeats along the sequence.
9. Specify the required histograms by selecting **Options** to the right of each check box. See *"Specifying nucleotide frequency histograms", p 8-15,* for details of this procedure.
10. If required, you can restrict the analysis to a portion of the sequence, by typing in the sequence numbers that bracket the region in the **Region** text boxes. Alternatively, select a region from the feature selector drop-down menu to the right of the text boxes.
11. Select **OK** to perform the analysis.

Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the analysis.

Specifying base composition plots

When base composition plots are generated, MacVector can generate up to six separate graphs, each with a maximum of four plots.



Base Composition Plotting Options

	Plot 1	Plot 2	Plot 3	Plot 4
Panel 1:	A 🟥	C 🟩	T 🟦	G ⬛
Panel 2:	🟥			
Panel 3:				
Panel 4:				
Panel 5:				
Panel 6:				

To specify base composition plots

1. Choose **Analyze | Base Composition** from the menu.

The **Base Composition Analysis** dialog box is displayed.

2. Select **Options** in the **Plot** panel.

The **Base Composition Plotting Options** dialog box is displayed.

3. Enter the required composition plot for the first graph, either by typing one or more base characters in the **Panel 1 / Plot 1** text box, or by selecting the arrow button to the right of the text box and choosing from the drop-down menu.

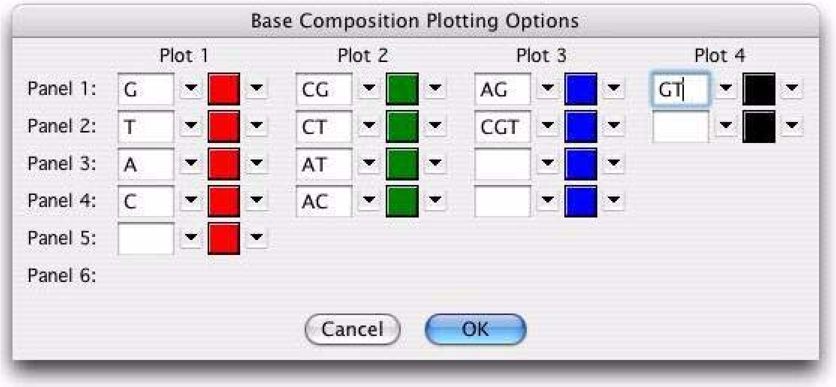
When a valid entry is made, the following changes occur:

- a second text box appears, under the **Plot 2** position
- a new text box appears in the **Panel 2** row, at the **Plot 1** position.

The **Panel** rows refer to separate graphs, the **Plot** columns to plots on each graph.

4. Enter additional composition plots as required.

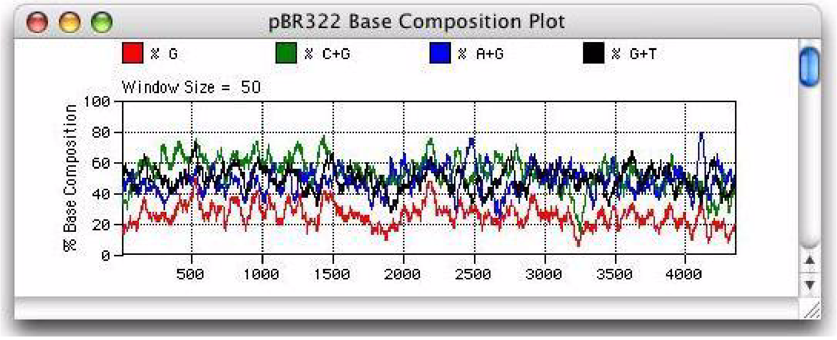
Each time an extra plot is specified for a given panel, a new text box appears under the next plot column, up to the maximum 4 plots. When the first plot for a panel is specified, a new row appears, up to the maximum 6 panels. The following picture shows a typical specification of the composition plots.



Base Composition Plotting Options

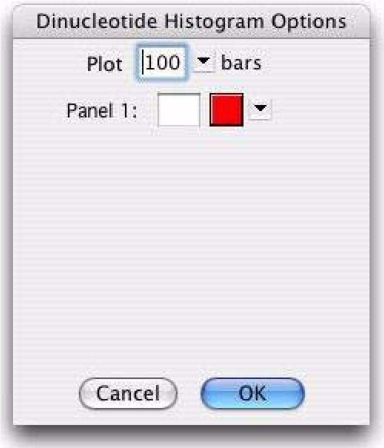
	Plot 1	Plot 2	Plot 3	Plot 4
Panel 1:	G 🟥	CG 🟩	AG 🟦	GT ⬛
Panel 2:	T 🟥	CT 🟩	CGT 🟦	⬛
Panel 3:	A 🟥	AT 🟩	🟦	
Panel 4:	C 🟥	AC 🟩	🟦	
Panel 5:	🟥			
Panel 6:				

5. To change the color of each plot, do one of the following:
   - Click on the arrow button to the right of the color swatch, and choose from the displayed color palette
   - Click on the color swatch itself, and use the system Color Picker; this is useful for fine control.
6. Select **OK** to dismiss the **Base Composition Plotting Options** dialog box.
7. Select **OK** to generate the plots.



Specifying nucleotide frequency histograms

The frequency of occurrence of di- and trinucleotides is visualized in MacVector using a histogram plot, where each histogram bar represents the number of occurrences of the selected set of nucleotides, within a region of the sequence whose length is determined by the number of bars you choose to plot on the histogram.



To specify nucleotide frequency histograms

1. Choose **Analyze | Base Composition**.

The **Base Composition Analysis** dialog box is displayed.

2. Select **Options** in the **Histograms** panel next to the dinucleotide or trinucleotide check box.

Depending on which **Options** was selected, either the **Dinucleotide** or **Trinucleotide Histogram Options** dialog box is displayed.

3. Enter either the 2 or 3 letters that represent the di- or trinucleotide sequence in the **Panel** text boxes.

Up to 6 histograms can be generated for dinucleotides and trinucleotides. As each histogram is specified, a new panel row appears, up to the maximum allowed.

4. Enter the number of bars required for each histogram in the **Plot** text box.

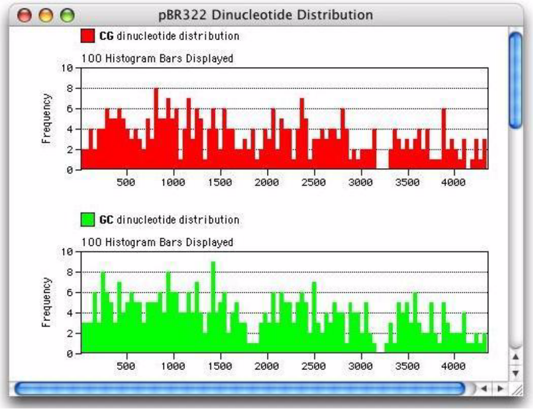
The sequence being analyzed is split into equal parts, and the number of occurrences of the nucleotide sequence in each part is represented by a bar on the histogram.

5. If you need to change the bar color, do one of the following:

   - Click on the arrow button to the right of the color swatch, and choose from the displayed color palette

- Click on the color swatch itself, and use the system Color Picker; this is useful for fine control.

6. Select **OK** to dismiss the **Histogram Options** dialog box.
7. Select **OK** to generate the histograms.


pBR322 Dinucleotide Distribution
🟥 CG dinucleotide distribution
100 Histogram Bars Displayed
Frequency
10
8
6
4
2
0
500
1000
1500
2000
2500
3000
3500
4000
🟩 GC dinucleotide distribution
100 Histogram Bars Displayed
Frequency
10
8
6
4
2
0
500
1000
1500
2000
2500
3000
3500
4000

Searching nucleic acids for coding regions

ORF analysis

Open reading frame (ORF) analysis enables you to find:

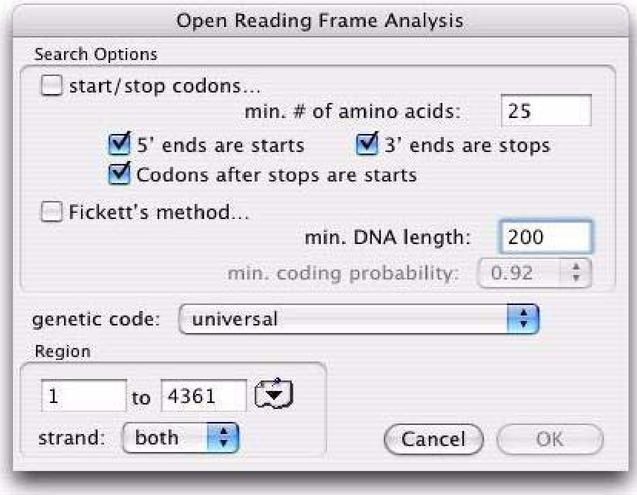
- open reading frames using user-designated start and stop codons
- regions of a nucleic acid that are likely to code for protein according to Fickett’s TESTCODE algorithm.

To run the analysis, a nucleic acid sequence window must be the active window. If you want to define your own start and stop codons, you must edit the genetic code that you are using. See Chapter 11, *“Using Transcription and Translation Functions”*, for further details of this procedure.

Finding open reading frames and coding regions

The ORF analysis and Fickett’s coding region algorithm are initiated on the same dialog box, and can be performed as a single analysis. Refer to

Chapter 20, *"Understanding Protein and DNA Analysis"*, for further details of Fickett's alogorithm.



Open Reading Frame Analysis

Search Options

☐ start/stop codons...

min. # of amino acids: 25

☑ 5' ends are starts ☑ 3' ends are stops

☑ Codons after stops are starts

☐ Fickett's method...

min. DNA length: 200

min. coding probability: 0.92

genetic code: universal

Region

1 to 4361

strand: both

To find an open reading frame

1. Choose **Analyze | Open Reading Frames** from the menu.

The **Open Reading Frame Analysis** dialog box is displayed.

2. Select the **start/stop codons** check box.

Tip. You may want to define your own start and stop codons. Refer to Chapter 11, *"Using Transcription and Translation Functions"*, for details of this.

3. Specify the minimum amino acid length that you want to consider to be a valid open reading frame in the **min. # of amino acids** text box.

Tip. The majority of proteins are larger than 75 amino acids. However, eukaryotic exons can be as small as 20 to 25 amino acids.

4. Select check boxes for treating the sequence ends as start and stop codons. The options are:

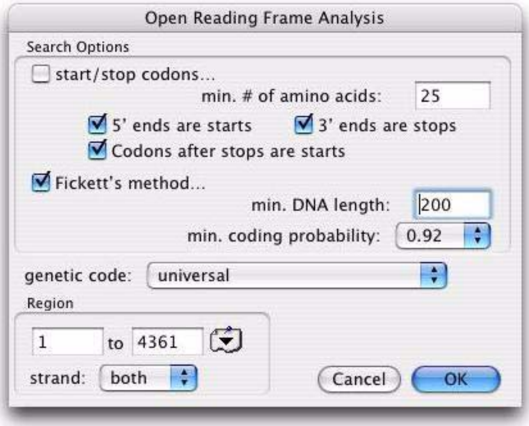
- **5' ends are starts**
- **3' ends are stops**
- **Codons after stops are starts**

Normally all three boxes should be checked because MacVector defines an open reading frame as a region flanked by a start codon and a stop codon. If your sequence fragment lies in the middle of a coding region and contains no start or stop codons, for example, MacVector will not report any open reading frames unless you have checked both **5' ends are starts codon** and **3' ends are stops**. The **Codons after stops are starts** option is recommended for eukaryote sequences where introns are common. Where there are internal coding exons, it finds the longest possible ORFs, ensuring that the entire exon is captured.

5. Select the genetic code that contains the start and stop codons you want to use from the **genetic code** drop-down menu
6. You can restrict the search to a certain region of the sequence by typing in the sequence numbers that bracket the region in the **Region** panel, or by selecting a region from the features table drop-down menu at the right of the text boxes.
7. You can restrict the analysis to one or both strands by selecting the appropriate item from the **strand** drop-down menu.
8. Select **OK** to perform the analysis.

Tip. You can perform a coding region analysis at the same time (see *"To find a coding region using Fickett's method"* on page 157). The results will include only those ORFs that are found by both methods.

After the analysis is complete, the **Open Reading Frame Analysis Display Options** dialog box is displayed. Its use is described in *"Displaying ORF and coding region search results"* on page 158.



Open Reading Frame Analysis

Search Options

□ start/stop codons...

min. # of amino acids: 25

☑ 5' ends are starts ☑ 3' ends are stops

☑ Codons after stops are starts

☑ Fickett's method...

min. DNA length: 200

min. coding probability: 0.92

genetic code: universal

Region

1 to 4361

strand: both

To find a coding region using Fickett's method

1. Choose **Analyze | Open Reading Frames** from the menu.

The **Open Reading Frame Analysis** dialog box is displayed.

2. Select the **Fickett's method** check box.
3. Type in the **min. DNA length**.

This should be at least 200 for best results.

4. Select a value for the coding probability from the **min. coding probability** drop-down menu.

Fickett's method assigns any region that has a coding probability below 0.29 as "non-coding" and any region with a coding probability greater than 0.92 as "coding." Regions with intermediate probabilities are assigned "no opinion." You can include these borderline regions with the "coding" regions by setting the probability of coding at successively lower numbers.

Note. Fickett's method is independent of start and stop codons and it cannot distinguish between frames. It simply reports regions where the cutoff value is exceeded.

5. You can restrict the search to a certain region of the sequence by typing in the sequence numbers that bracket the region in the **Region** panel, or by selecting a region from the features selector drop-down menu to the right of the text boxes.
6. You can restrict the analysis to one or both strands by selecting the appropriate item from the **strand** drop-down menu.
7. Select **OK** to perform the analysis.

Tip. You can perform an open reading frame analysis at the same time (see *"To find an open reading frame"* on page 156). The results will include only those ORFs that are found by both methods.

After the analysis is complete, the **Open Reading Frame Analysis Display Options** dialog box is displayed. The use of this dialog box is described in the next section.

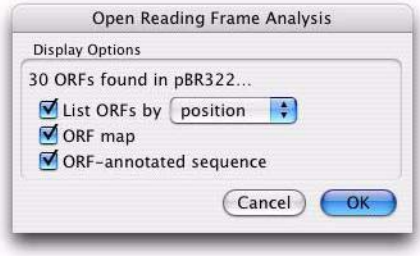
Displaying ORF and coding region search results

The results of a reading frame analysis or coding region search can be displayed in several ways:

- as a text listing
- as a linear map
- as an annotated sequence.

If you have performed both types of analysis at the same time, the combined results are displayed.

Each display type can be saved to disk or printed when the corresponding view is active. Refer to Chapter 3, *“General Procedures”*, for further details.



Open Reading Frame Analysis

Display Options

30 ORFs found in pBR322...

☑ List ORFs by position

☑ ORF map

☑ ORF-annotated sequence

To display ORF and coding region search results

1. The **Open Reading Frame Analysis Display Options** dialog box is displayed on completion of each analysis. To display this dialog box at other times, choose **Analyze | Open Reading Frames** when any open reading frame display result window is active.
2. Select **List ORFs by** to display a text window that contains a list of the open reading frames / coding regions found. The associated drop-down menu enables you to specify whether the list should be ordered according to **position** in the sequence or according to the **length** of the open reading frames / coding regions.
3. Select **ORF map** to display a map window that shows a linear map of the positions of the ORFs / coding regions.

Tip. Right-click on an ORF or coding region in the ORF map and select **Create CSD Feature** from the context menu to add that feature to the original seqeunce in a Genbank compatible format.

Note. If you select a single ORF in the ORF map the sequence will be highlighted in the original Sequence window. This provides another way of creating a feature for an ORF.

4. Select **ORF-annotated sequence** to display a Text view that shows the nucleic acid sequence and the translated sequences of any ORFs or coding regions found by the analysis.
5. Select **OK** to display the requested results.

Performing Nucleic Acid Analysis Toolbox analyses

The coding preference plots in the Nucleic Acid Analysis Toolbox enable you to locate regions in a sequence that have a high probability of coding for a highly expressed protein. Where such regions are found, the plots also help you to determine which reading frame is the coding frame.

MacVector offers a range of algorithms for calculating coding preferences and locating protein-coding regions in a sequence:

- Open reading frames
- G+C % composition
- Fickett's TestCode
- Uneven positional base frequencies
- Positional base preference
- MacVector codon preference
- Gribskov codon preference
- Staden codon preference
- Transfac profile

Refer to *"Base composition methods"* on page 441, *"Codon preference and codon bias tables"* on page 445, and *"Transfac profile scans"* on page 160 for details of the calculations.

Most algorithms generate three plots, one for each possible starting codon position. These can be displayed separately, or combined in a single panel to accentuate the differences between plots. You can display the results of several different analyses in aligned panels, and control the colors of the different plots.

To run this analysis, a nucleic acid sequence window must be the active window, and a codon bias file for the organism under investigation must be available. Several codon bias files are provided with MacVector.

Transfac profile scans

The Nucleic Acid Analysis Toolbox allows you to scan a sequence with up to two profile files. MacVector reads profile in transfac profile format. The default profiles are of Donor and Acceptor sites for RNA splicing. There is also a set of profile files for the Jaspar database of transcription factors in the `MacVector 12.6/Profiles` folder.

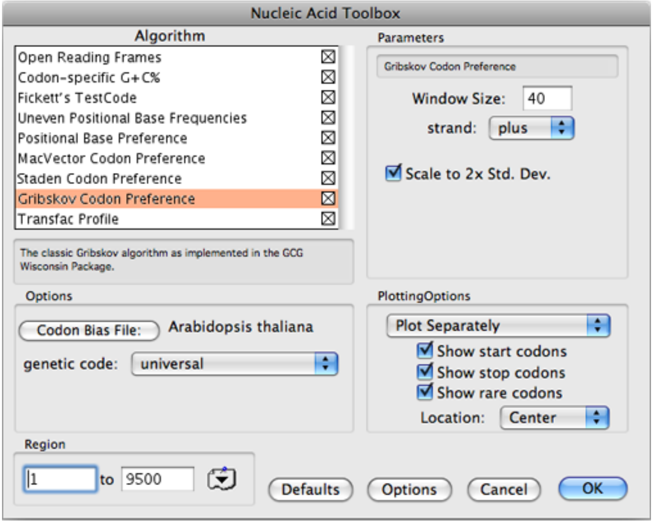
The Multiple Sequene Alignment window **Profile** view displays the transfac profile for the current alignment. You can export or copy and paste this text into text edit and use the resulting file as the source file for a Profile search in the Nuclic Acid Analysis Toolbox.

To generate the plot the profile is moved along the sequence and the score at each position is calculated by adding together the count of the corresponding residues for each position in the profile. The codon bias is taken into account and then the cumulative log odds probabilities for each position having a match to the profile are plotted along the sequence. The scale is chosen so that by default the base is set to twice the standard deviation of the results.

Additionally, whether the target site actually matches the consensus splice site or not is displayed on the same plot. There is a threshold % setting in the Profile pane, default 85%. When MacVector loads a profile it generates a consensus using the least degenerate IUPAC code that matches or exceeds 85%. So if 90% of the residues are G, then the consensus is G. If 40% are C, 45% T and the rest A or T, then it will assign Y, the IUPAC code for C or T. It then scans the target sequence for a match to the consensus. Where it finds them, it draws a box on the plot covering the length of the profile with a vertical line through at the "cut" point. If a region matches both profiles, then the box is drawn in red rather than blue or green.

If a peak is present in the profile plot corresponding with the vertical line through a consensus box then that indicates you have a high probability profile "hit" that also matches the consensus sequence for the profile.

Generating plots with the Nucleic Acid Analysis Toolbox



Nucleic Acid Toolbox

Algorithm

- Open Reading Frames ☒
- Codon-specific G+C% ☒
- Fickett's TestCode ☒
- Uneven Positional Base Frequencies ☒
- Positional Base Preference ☒
- MacVector Codon Preference ☒
- Staden Codon Preference ☒
- Gribskov Codon Preference ☒
- Transfac Profile ☒

The classic Gribskov algorithm as implemented in the GCG Wisconsin Package.

Parameters

Gribskov Codon Preference

Window Size: 40

strand: plus

☑ Scale to 2x Std. Dev.

Options

Codon Bias File: Arabidopsis thaliana

genetic code: universal

PlottingOptions

Plot Separately

☑ Show start codons

☑ Show stop codons

☑ Show rare codons

Location: Center

Region

1 to 9500

To generate nucleic acid analysis plots

1. Choose **Analyze | Nucleic Acid Analysis Toolbox** from the menu

The **Nucleic Acid Toolbox** dialog box is displayed.

2. Scroll through the **Algorithm** list and select the checkbox for an algorithm you want to use.

When an algorithm is highlighted, a brief description is displayed below the **Algorithm** list, and the options for it are displayed in the **Parameters** panel.

Tip. You can browse through descriptions and parameters for the algorithms without selecting their check boxes. Simply click on a name in the algorithm list to display this information.

3. In the **Parameters** panel, adjust the settings as required.

Window size is set independently for each algorithm. Increasing the window size gives smoother, clearer graphs, but reduces the definition of start and stop locations and can obscure short coding regions. For more details, refer to the individual algorithm entries in *"Coding regions"* on page 440.

4. Repeat steps 2 and 3 to add more algorithms, as required.

To remove an algorithm from analysis, reselect its check box.To select or deselect all the algorithms, hold down the **option** key while clicking on any checkbox.

5. To choose the correct codon bias table, select the **Codon Bias File:** button in the **Options** panel. A file selection dialog box is displayed; select the required file, normally from the `Codon Bias Tables` folder.
6. To select the genetic code, choose from the **genetic code** drop-down menu in the **Options** panel.

Tip. To modify which codons are used as starts and stops, you can create your own genetic code. See *“Modifying genetic codes”* on page 242.

7. From the drop-down menu in the **Plotting options** panel, choose either:
   - **Combined plots** - to combine the plots into a single panel
   - **Plot separately** - to plot each frame in a separate panel
8. To adjust the display of codons in the plots, select as required the checkboxes for:
   - **Show start codons**
   - **Show stop codons**
   - **Show rare codons**

To specify where the codons should appear in the display, choose **Above**, **Center** or **Below** from the **Location** drop-down menu. They will only be displayed where separate plots have been selected and where frame-specific data has been generated.

9. To restrict the search to a certain region of the sequence, type in the sequence numbers that bracket the region in the **Region** panel.

Alternatively, choose a region from the features table drop-down menu at the right of the text boxes.

10. Select **Options** if you want to alter the plot colors, codon symbols, or the definition of rare codons. See *“To set the nucleic acid analysis toolbox plot options”* on page 164.
11. Select **OK** to generate the plot.

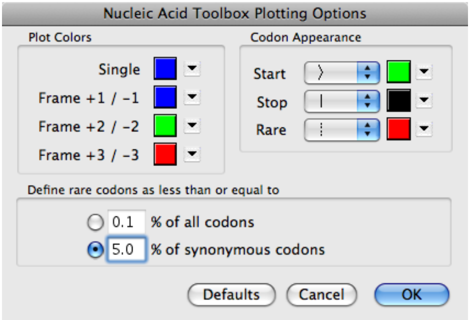
Alternatively, select the **Defaults** button to restore the default settings, or select **Cancel** to close the dialog box without performing any analyses.

Plot colors, codon symbols, and the definition of rare codons are specified using a separate dialog box, **Nucleic Acid Toolbox Plotting Options**.

To set the nucleic acid analysis toolbox plot options

1. In the **Nucleic Acid Toolbox** dialog box, select the **Options** button.

The **Nucleic Acid Toolbox Plotting Options** dialog box is displayed.



Nucleic Acid Toolbox Plotting Options

Plot Colors

Single 🟦
Frame +1 / -1 🟦
Frame +2 / -2 🟩
Frame +3 / -3 🟥

Codon Appearance

Start 〉 🟩
Stop | ⬛
Rare ¦ 🟥

Define rare codons as less than or equal to

○ 0.1 % of all codons
● 5.0 % of synonymous codons

2. In the **Plot Colors** panel, set colors as required. You can do this either by selecting an arrow button and choosing from the drop-down color palette, or by clicking on the color swatch and using the system Color Picker.

   - **Single** sets the color for all separate plots
   - **Frame +1 / -1** sets the color for the first frame in a combined plot, **Frame +2 / -2** the second, and **Frame +3 / -3** the third.

3. In the **Codon Appearance** panel, set the symbols and colors for **Start**, **Stop** and **Rare** codons, using the appropriate drop-down menus and arrow buttons.

4. In the **Define rare codons as less than or equal to** panel, select a radio button to choose the way rare codons will be defined:

   - **% of all codons**
   - **% of synonymous codons -** i.e. all codons that generate the same amino acid as that codon.

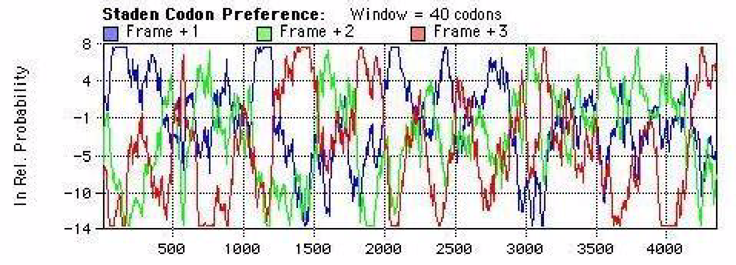
Then type the required threshold value in the selected text box. (The default setting is **5.0% of synonymous codons**. If **% of all codons** is selected, the default value is **0.1**%.)

5. Select **OK** to apply the settings.

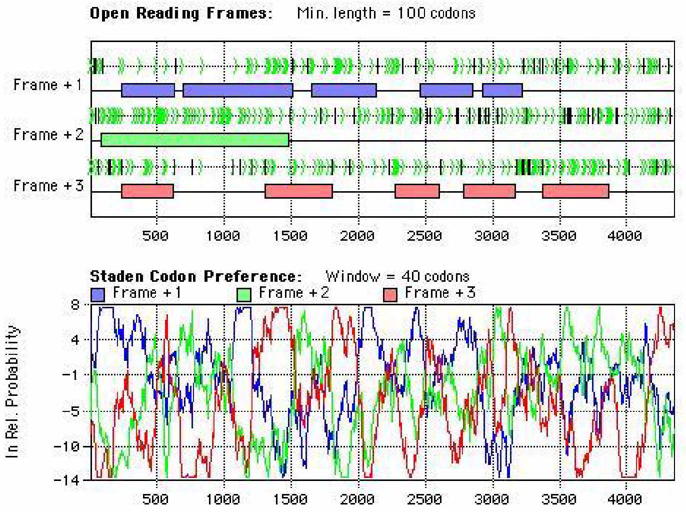
Alternatively, select the **Defaults** button to restore the default settings, or **Cancel** to close the dialog box without saving the settings.

Viewing Nucleic Acid Analysis Toolbox results

The results window is a scrollable window containing the plots from the selected analyses in aligned panels. In combined-plots mode, the results of each analysis are displayed in a single panel. In separate-plots mode, results for each reading frame is displayed on a separate graph, so there will be either three or six panels, depending on whether you specified one or both strands.



The open reading frames plot always combines the three frames in a single color-coded panel, showing markers for starts, stops and rare codons, and bars representing open reading frames that exceed the minimum length. When you select separate-plots mode, the starts, stops and rare codons are also displayed in the frame results panels of your other analyses.



Prominent peaks in the graph indicate regions that exhibit a codon preference. Such regions are very likely to code for highly expressed proteins. For more details, see *"Interpreting coding preference plots"* on page 448.

To inspect potential coding regions, you will usually need to select a region of the plot and display it at greater magnification. When you select a region, all the results plots are re-scaled, keeping them aligned.

To navigate in the Nucleic Acid Analysis Toolbox results window

- **To magnify a selected region**, click and drag in the window
- **To return to the whole plot**, double-click anywhere in the window
- **To zoom in,** press the **Up-arrow** key (or ⌘ and **+** keys) to zoom by a factor of 2, or press **Shift+Up-arrow** to zoom to individual residues
- **To zoom out,** press the **Down-arrow** key (or ⌘ and **-** keys) to zoom by a factor of 2, or press **Shift+Down-arrow** to zoom all the way out
- **To shift to the right,** press the **Right-arrow** key to shift 10% to the right, or press **Shift+Right-arrow** to shift 80% to the right
- **To shift to the left,** press the **Left-arrow** key to shift 10% to the left, or press **Shift+Left-arrow** to shift 80% to the left.

Tip. You can check for keyboard shortcuts at any time during a session. Choose **Help | MacVector Keyboard Shortcuts** to display a listing of all MacVector shortcuts.

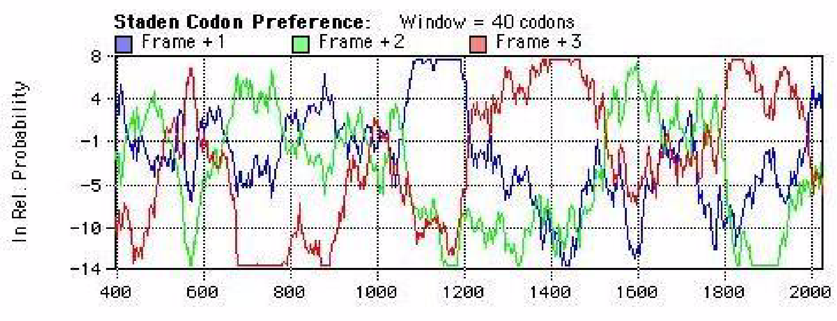
Selecting ORFs from Nucleic Acid Analysis Toolbox plots

If you click any open reading frame in the coding preference plot results window, the corresponding residues are selected in the sequence editor window.

Tip. Click on an ORF, go to the sequence editor window, choose **Analyze | Translation**, and you can quickly create a protein sequence from the selected region.

If you want more control in the selection of the ORF, click on any line of stop and start codons, instead of on the ORF bar. If there are several possible start codons, you do not have to choose the longest ORF - click just to the right of whichever start codon you want to use. The selected ORF is outlined, and its residues highlighted in the sequence editor window.

The following figure shows two possible selections for an ORF:



Note.If you choose separate-plots mode, you can select ORFs from any graph where codons are displayed.

8 Searching for Sites and Motifs

Overview

This chapter provides guidelines on working with the sequence analysis methods in MacVector, including:

- automatic and manual restriction site searching
- proteolytic site searching
- nucleic acid motif searching
- protein motif searching.

Refer to *"Searching nucleic acids for coding regions"* on page 155, for information about searching nucleic acid sequences for coding regions.

Refer to Chapter 12, *"Aligning Sequences"*, for a description of sequence comparison methods used by MacVector.

Contents

Restriction enzyme sites

You can use restriction enzyme files to:

- locate restriction enzyme cut sites in a nucleic acid sequence
- predict the fragments that would result from single or multiple digests using up to six enzymes.

To use this functionality, a nucleic acid sequence window must be the active window, and a restriction enzyme file must be available

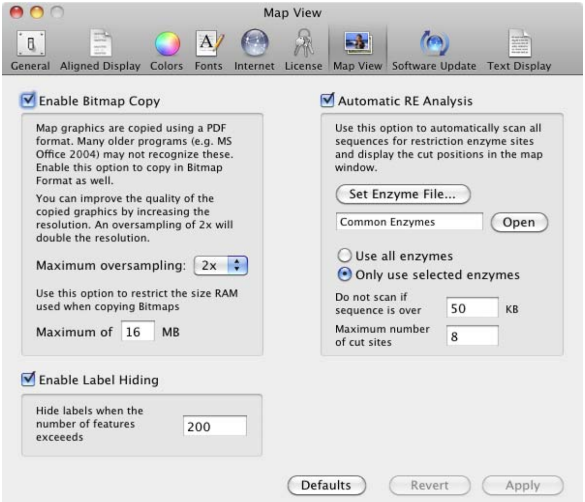
Searching for restriction enzyme sites

Two types of restriction enzyme site searching are available: automatic and manual.

With both types of search you can, if required, select a subset of the available enzymes from the file you are using. Before you perform the analysis, open the restriction enzyme file and click in front of the name of any enzyme you want to select. A check mark appears to mark your choice. If you save the restriction enzyme file, the selections are saved with the file. See *"Selecting enzymes for site analysis"* on page 59.

Automatic searching

Sequences in the active Sequence window can now be scanned automatically for restriction enzymes, with the cut sites found displayed in the Sequence Map view. The settings for this feature are specified in **Map View** preferences dialog box.



☑ Enable Bitmap Copy

Map graphics are copied using a PDF format. Many older programs (e.g. MS Office 2004) may not recognize these. Enable this option to copy in Bitmap Format as well.

You can improve the quality of the copied graphics by increasing the resolution. An oversampling of 2x will double the resolution.

Maximum oversampling: 2x

Use this option to restrict the size RAM used when copying Bitmaps

Maximum of 16 MB

☑ Enable Label Hiding

Hide labels when the number of features exceeeds 200

☑ Automatic RE Analysis

Use this option to automatically scan all sequences for restriction enzyme sites and display the cut positions in the map window.

Common Enzymes

○ Use all enzymes

● Only use selected enzymes

Do not scan if sequence is over 50 KB

Maximum number of cut sites 8

To set up automatic restriction enzyme site searching

1. Select **MacVector | Preferences** from the main menu, then click the **Map View** icon on the preferences dialog, or click the **Prefs** icon on the Sequence window toolbar when a Map view is selected.

The **Map View** preferences dialog box is displayed.

Note. You can also access the **Map View** preferences dialog using **Options | Map View Options** from the menu.

2. Ensure that the **Automatic RE Analysis** box is checked.

By default, MacVector automatically analyzes nucleic acid sequences for restriction enzyme sites using the specified enzyme file and displays the cut sites in the Sequence Map view. To switch off this feature, uncheck the **Automatic RE Analysis** box.

3. Click the **Set Enzyme File** button to choose an alternative restriction enzyme file for the automatic analysis.

By default, the file `/MacVector 12.6/Restriction Enzymes/Common Enzymes` is used.

4. Optionally, click the **Open** button to open the selected enzyme file in the background.

This shortcut enables you to access the enzyme file and modify the enzymes selected within it more quickly.

5. Select one of the following options

- Choose **Use all enzymes** to perform the automatic search using all of the enzymes in the specified enzyme file.
- Choose **Only use selected enzymes** to perform the automatic search using only those enzymes that are selected in the specified enzyme file.

6. Use the **Do not scan if sequence is over** setting to turn off automatic searching for larger sequences.

The restriction enzyme search is relatively fast but the graphical display of the results can take a long time for very large sequences. The default value is 50kb.

7. Use the **Maximum number of cut sites** setting to screen out restriction enzymes that cut your sequence too frequently.

This option can reduce search times and screen clutter. The default is six sites, but you might want to reduce this if you routinely use relatively large sequences.

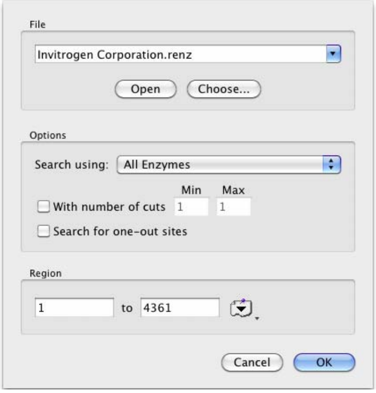
8. Click **OK** to save your settings and close the **Map View** dialog box or **Cancel** to close the dialog box without saving your changes.

When automatic restriction enzyme analysis is switched on and the active Sequence window contains a nucleic acid sequence, then the sequence is searched for restriction enzyme cut sites automatically and any that are located are displayed in the sequence Map view.

Restriction enzyme recognition sites that appear only once in a sequence are displayed in red in the Map view, whereas sites that are present more than once are shown in blue. Double-clicking on a restriction site selects all the sites of the same type, for example *Eco*RI, *Bam*HI, etc., making it easy to see where sites of a particuar type lie on the sequence.

In addition, if the **Only use selected enzymes** option was specified, all open sequence Map views are updated dynamically when changes are made to the enzyme selection in the specified enzyme file. This makes it easy to identify compatible restriction sites between two sequences (e.g. a gene and a vector), for click cloning (see *"Click Cloning"* on page 188).

Manual searching



File

Invitrogen Corporation.renz

Options

Search using: All Enzymes

	Min	Max
□ With number of cuts	1	1

□ Search for one-out sites

Region

1 to 4361

To search for restriction enzyme sites manually

1. Make the required nucleic acid Sequence window active.
2. Specify whether to analyze the sequence as a circular or linear molecule by clicking the **Topology** icon in the Sequence window toolbar until it is set correctly for your sequence.
3. Choose **Analyze | Restriction Enzyme** from the main menu or click the **RE Search** icon on the Analysis toolbar.

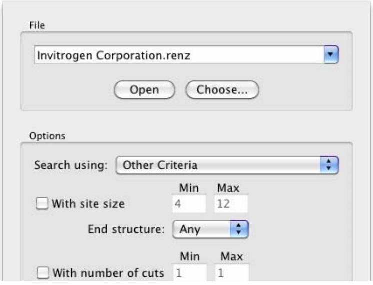
The **Restriction Enzyme Analysis** dialog box is displayed.

4. Click the **Choose...** button to choose an alternative restriction enzyme file to the one already selected (`Common Enzymes.renz`, by default) for use in the analysis.

A file selection dialog box is displayed, enabling you to choose the file.

5. Choose the required file from the scrolling list and select **Choose**.
6. Alternatively, click **Open** to display the contents of the currently selected restriction enzyme file.
7. Select your search criteria by choosing from the **Search Using** drop-down menu as follows:
   - choose **all enzymes** to use all the enzymes in the selected enzyme file
   - choose **selected enzymes** to use only the enzymes that you have selected within the enzyme file
   - choose **other criteria** to display further options

Note. When the **other criteria** search is performed, all enzymes in the file are used, not only the selected ones.



File: Invitrogen Corporation.renz

Search using: Other Criteria

□ With site size — Min 4, Max 12

End structure: Any

□ With number of cuts — Min 1, Max 1

8. If you have chosen **other criteria**, do one or more of the following:
   - to restrict the search to enzymes whose recognition sites are a certain size, type the size limits into the **site size** text boxes. MacVector accepts values from 1 to 100. Leave both boxes empty if you do not want to restrict the search by site size.
   - to limit the search to enzymes that leave 5' ends, 3' ends, or blunt ends when they cut the sequence, select the appropriate option from the **end structure** drop-down menu.
9. You can restrict the search to enzymes that cut the sequence a limited number of times by checking the **With number of cuts** box and

typing the desired **Min** and **Max** cut values in the associated boxes. MacVector accepts values from 1 to 20.

10. You can include one-out enzyme sites in the search by checking the **Search for one-out sites** box.

Note. These one-out sites are displayed in paler shades in graphical results views and with an asterisk next to them in text results views, to distinguish them from standard restriction enzyme sites.

11. You can restrict the search to a certain region of the sequence by typing in the base numbers that bracket the region in the **Region** panel, or by selecting a region from the feature selector drop-down menu to the right of the text boxes.

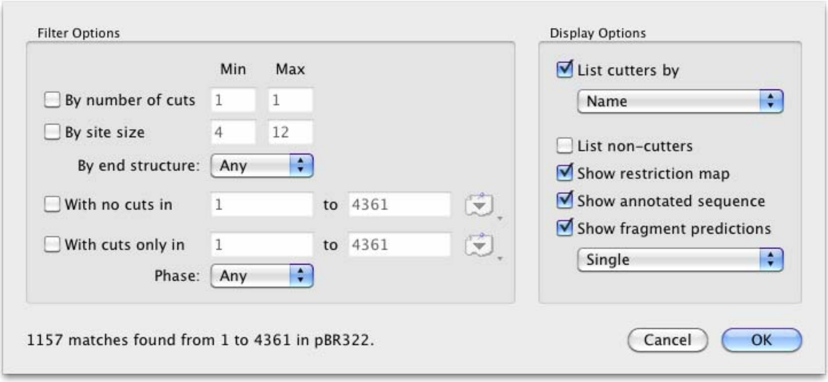
12. Select **OK** to perform the analysis.

When the analysis is completed, the **Restriction Enzyme Analysis Display** dialog box is displayed. The use of this dialog box is described in the following section.

Filtering manual restriction site search results

Manual restriction site search results can be filtered as follows:

- by the number of cuts an enzyme makes
- by restriction site size
- by the end structure of cut fragments
- by the residue range in which the cut site occurs.



Filter Options

	Min	Max
□ By number of cuts	1	1
□ By site size	4	12

By end structure: Any

□ With no cuts in 1 to 4361

□ With cuts only in 1 to 4361

Phase: Any

Display Options

☑ List cutters by Name

□ List non-cutters

☑ Show restriction map

☑ Show annotated sequence

☑ Show fragment predictions Single

1157 matches found from 1 to 4361 in pBR322.

To filter restriction site search results

1. The **Restriction Enzyme Analysis Display** dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example, to change the display parameters, choose **Analyze | Restriction Enzyme** when any restriction analysis display result window is active.
2. To restrict the displays to enzymes that cut the sequence a limited number of times, select the **By number of cuts** checkbox and type the desired **Min** and **Max** cut values in the associated boxes. MacVector accepts values from 1 to 20.
3. To restrict the search to enzymes whose recognition sites are a certain size, select the **By site size** checkbox and type the size limits into the **Min** and **Max** boxes. MacVector accepts values from 1 to 100.
4. To restrict the displays to enzymes that leave 5' ends, 3' ends, or blunt ends when they cut the sequence, select the appropriate option from the **By end structure** drop-down menu.

If you restricted the search stage to one end type, this drop-down menu will be disabled.

5. To restrict the displays to enzymes that do not cut within a given region, select the **With no cuts in** checkbox, then either type in the numbers of the bases that bracket the region, or choose a region from the features table drop-down menu at the right of the text boxes.
6. To restrict the displays to enzymes that cut only within a given region, select the **With cuts only in** checkbox, then either type in the numbers of the bases that bracket the region, or choose a region from the features table drop-down menu at the right of the text boxes.
7. To restrict the displays to enzymes that cut only within a given reading frame, select the **With cuts only in** checkbox, then choose a frame number from the **Phase** drop-down menu.

Tip. If you want to specify a reading frame but not limit the displays to a region, select **ALL 1- ...** from the features table drop-down menu.

8. Select the **Display Options** as required.

These options are described in the next section.

9. Select **OK** to display the results.

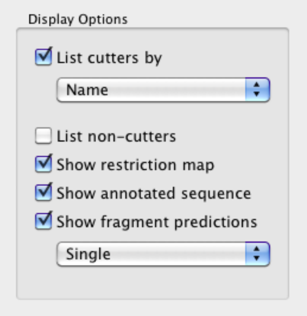
Displaying manual restriction site search results

You can display manual restriction site search results in a number of ways:

- as a list of enzymes that cut the sequence
- as a list of enzymes that did not cut the sequence
- as a restriction map
- as an annotated sequence
- as a list of the predicted fragments for each digest

One or more of the displays can be generated at once.

Each display type can be saved to disk or printed when its window is active. Refer to *“Saving graphics”* on page 38, for further details.



Display Options

☑ List cutters by

Name

□ List non-cutters

☑ Show restriction map

☑ Show annotated sequence

☑ Show fragment predictions

Single

To display restriction site search results

1. The **Restriction Enzyme Analysis Display** dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example, to change the display parameters, choose **Analyze | Restriction Enzyme** when any restriction analysis display result window is active.
2. Select the filter options as required. These are described in the previous section.
3. Select the **List cutters by** check box to display a list of the enzymes that cut the sequence and the locations of the cuts.

Use the drop-down menu to choose how to order the list: by name, by position of cuts, or by number of cuts.

4. Select the **List non-cutters** check box to display a list of the enzymes that were considered in the search but were found not to cut the sequence.
5. Select the **Show restriction map** check box to display the locations of the cut sites in a graphical form.The appearance of this display can

be changed interactively using the Graphics Palette. Refer to *"Editing the general map appearance"* on page 109, for further details.

Tip. Right-click on a restriction enzyme site in the restriction map and select **Create misc_feature Feature** from the context menu to add that feature to the original seqeunce in a Genbank compatible format.

6. Select the **Show annotated sequence** check box to display a Text view with the cut sites displayed at their proper locations along the sequence. The appearance of this view can be changed interactively using the **Text Display** preferences dialog box. Refer to *"Formatting the Text view"* on page 124, for further details.
7. Select the **Show fragment predictions** check box to display fragment predictions for the digests. The enzymes used are those that have satisfied the **Filter Options**.

Use the drop-down menu to choose from the following:

- **Single** for single digests
- **Double** for double digests
- **Single & Double** for single and double digests
- **All combined** for digests using up to six selected enzymes.

8. Select **OK** to display the results.

Proteolytic enzyme sites

You can use proteolytic enzyme files to:

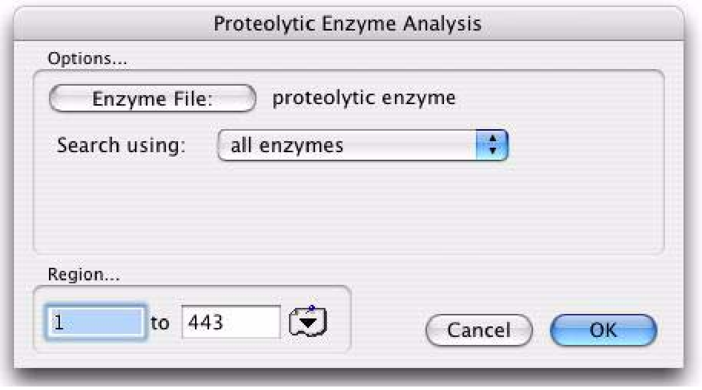
- locate cleavage sites for proteolytic agents in a protein sequence
- predict the fragments that would result from single or multiple digests using up to six proteolytic agents.

To use this functionality, a protein sequence window must be the active window, and a proteolytic enzyme file must be available.

Although we use the term proteolytic enzyme throughout this section of the manual, the analysis can locate cleavage sites for any proteolytic agents—whether enzymes or chemical agents—that cleave a protein in a sequence-specific manner.

Searching for proteolytic enzyme sites

If required, you can select a subset of the available enzymes from the file you are using. Before you perform the analysis, open the proteolytic enzyme file and click in front of the name of any enzyme you wish to select. A check mark appears to mark your choice. Save the proteolytic enzyme file to save the selections. See *"Selecting enzymes for site analysis"* on page 59, for further details.



Proteolytic Enzyme Analysis

Options...

Enzyme File: proteolytic enzyme

Search using: all enzymes

Region...

1 to 443

To search for proteolytic enzyme sites

1. Make the required amino acid sequence window active.
2. Choose **Analyze | Proteolytic Enzyme**.

The **Proteolytic Enzyme Analysis** dialog box is displayed.

3. Select **Enzyme File** to choose the restriction enzyme file that will be used for the analysis.

A file selection dialog box is displayed, enabling you to choose the file.

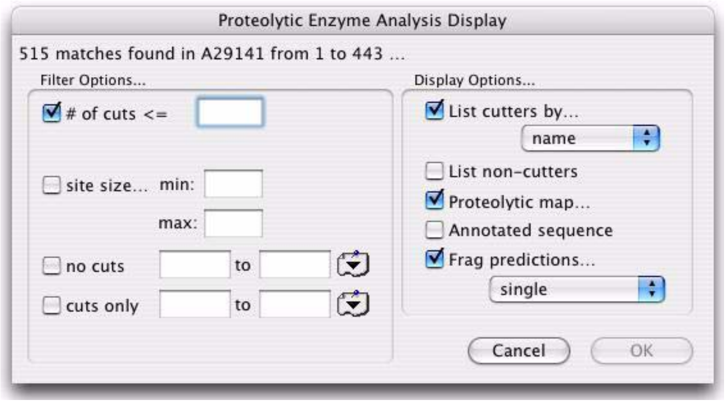
4. Choose the required file from the scrolling list and select **Choose**.
5. You can restrict the search to a certain region of the sequence by typing in the sequence numbers that bracket the region in the **Region** panel, or by selecting a region from the feature selector drop-down menu to the right of the text boxes.
6. Select your search criteria by choosing from the **Search Using** drop-down menu as follows:
   - choose **all enzymes** to use all the enzymes in the selected enzyme file
   - choose **selected enzymes** to use only the enzymes that you have selected within the enzyme file.
7. Select **OK** to perform the analysis.

When the analysis is completed, the **Proteolytic Enzyme Analysis Display** dialog box is displayed. The use of this dialog box is described in the following section.

Filtering proteolytic site search results

The proteolytic site search results can be filtered as follows:

- by the number of cuts an enzyme makes
- by cut site size
- by the residue range in which the cut site occurs.



Proteolytic Enzyme Analysis Display

515 matches found in A29141 from 1 to 443 ...

Filter Options...

☑ # of cuts <=

□ site size... min: max:

□ no cuts to

□ cuts only to

Display Options...

☑ List cutters by... name

□ List non-cutters

☑ Proteolytic map...

□ Annotated sequence

☑ Frag predictions... single

To filter proteolytic site search results

1. The **Proteolytic Enzyme Analysis Display** dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example, to change the display parameters, choose **Analyze | Proteolytic Enzyme** when any proteolytic analysis display result window is active.
2. To restrict the displays to enzymes that cut the sequence a limited number of times, select the **# of cuts** checkbox and type a number from 1 to 20 to set the upper limit for number of cuts.
3. To restrict the search to enzymes whose recognition sites are a certain size, select the **site size** checkbox and type the size limits into the **min** and **max** text boxes. MacVector accepts values from 1 to 100.
4. To restrict the displays to enzymes that do not cut within a given region, select the **no cuts** checkbox, then either type in the numbers of the bases that bracket the region, or choose a region from the features table drop-down menu at the right of the text boxes.

5. To restrict the displays to enzymes that cut only within a given region, select the **cuts only** checkbox, then either type in the numbers of the bases that bracket the region, or choose a region from the features table drop-down menu at the right of the text boxes.
6. Select the **Display Options** as required.

These options are described in the next section.

7. Select **OK** to display the results.

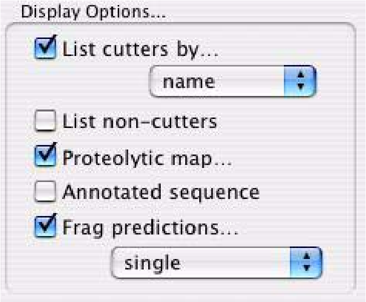
Displaying proteolytic site search results

You can display the proteolytic enzyme site search results in a number of ways:

- as a list of enzymes that cut the sequence
- as a list of enzymes that did not cut the sequence
- as a proteolytic map
- as an annotated sequence
- as a list of the predicted fragments for each digest.

One or more of the displays can be generated at once.

Each display type can be saved to disk or printed when its window is active. Refer to Chapter 3, *"General Procedures"*, for further details.



Display Options...

☑ List cutters by...

name

□ List non-cutters

☑ Proteolytic map...

□ Annotated sequence

☑ Frag predictions...

single

To display proteolytic site search results

1. The **Proteolytic Enzyme Analysis Display** dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example, to change the display parameters, choose **Analyze | Proteolytic Enzyme** when any restriction analysis display result window is active.

2. Select the **List cutters by** check box to display a list of the enzymes that cut the sequence and the locations of the cuts.

   Use the drop-down menu to choose how to order the list: by name, by position of cuts, or by number of cuts.

3. Select the **List non-cutters** check box to display a list of the enzymes that were considered in the search but were found not to cut the sequence.
4. Select the **Proteolytic map** check box to display the locations of the cut sites in a graphical form.The appearance of this display can be changed interactively using the Graphics Palette. Refer to *"Editing the general map appearance"* on page 109, for further details.

Tip. Right-click on a proteolytic enzyme site in the proteolytic map and select either **Create Cleavage Feature** or **Create SITE Feature** from the context menu to add that feature to the original seqeunce in a Genbank compatible format.

5. Select the **Annotated sequence** check box to display a Text view with the cut sites displayed at their proper locations along the sequence. The appearance of this view can be changed interactively using the **Text Display** preferences dialog box. Refer to *"Formatting the Text view"* on page 124, for further details.
6. Select the **Frag predictions** check box to display fragment predictions for the digests. The enzymes used are those that have satisfied the **Filter Options**.

   Use the drop-down menu to choose from the following:

   - **single** for single digests
   - **double** for double digests
   - **both** for single and double digests
   - **all combined** for digests using up to six selected enzymes.

7. Select **OK** to display the results.

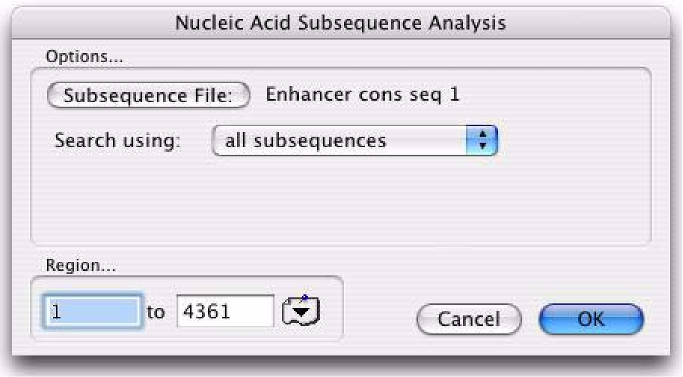
Subsequence analysis

You can scan a nucleic acid or protein sequence for known consensus sequences, primers and motifs that are defined in a subsequence file. A subsequence may consist of one, two or three parts, with the permitted gap between each part defined, as well as the allowed mismatch for each part in a search. Refer to *"Subsequence files"* on page 63, for further information.

To use this functionality, a Sequence window must be the active window, and a subsequence file must be available.

Searching for subsequences

If required, you can select a subset of subsequences from the list of all subsequences in the file. Before the subsequence search, open the subsequence file and click in front of the name of any subsequence you want to select. A check mark appears to mark your choice. Save the subsequence file to save the selections. See *"Selecting subsequences for searches"* on page 64, for further details.


Nucleic Acid Subsequence Analysis

Options...

Subsequence File: Enhancer cons seq 1

Search using: all subsequences

Region...

1 to 4361

To search for subsequences

1. Make the required Sequence window active.
2. If you are analyzing a nucleic acid sequence, specify whether to analyze the sequence as a circular or linear molecule by clicking the **Topology** icon in the Sequence window toolbar until it is set correctly for your sequence.
3. Do one of the following:
   - choose **Analyze | Nucleic Acid Subsequence** for a nucleic acid sequence
   - choose **Analyze | Protein Subsequence** for a protein sequence.

The appropriate Analysis dialog box is displayed.

4. Select the **Subsequence File** button to choose the subsequence file that will be used for the analysis.

A file selection dialog box is displayed, enabling you to choose the file.

5. Choose the required file from the scrolling list and select **Choose**.

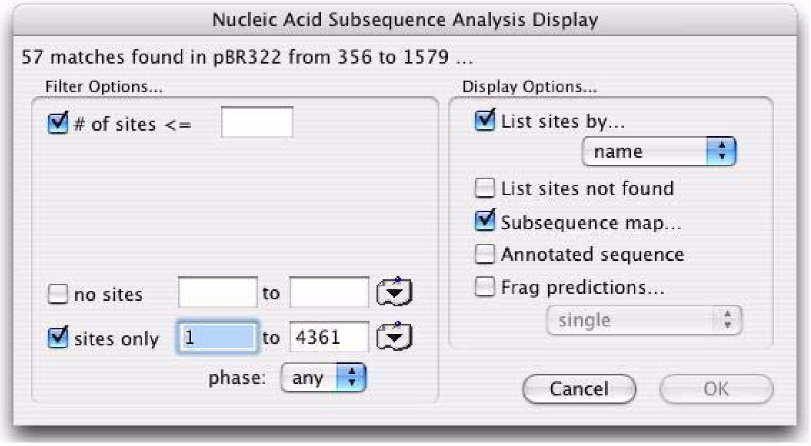
6. You can restrict the search to a certain region of the sequence by typing in the base numbers that bracket the region in the **Region** panel, or by selecting a region from the feature selector drop-down menu at the right of the text boxes.
7. Select your search criteria by choosing from the **Search Using** drop-down menu as follows:
   - choose **all subsequences** to use all the subsequences in the selected subsequence file
   - choose **selected subsequences** to use only the subsequences that you have selected in the subsequence file.
8. Select **OK** to perform the analysis.

When the analysis is completed, the appropriate Display dialog box is displayed. The use of this dialog box is described in the following section.

Filtering subsequence search results

The subsequence search results can be filtered as follows:

- by the frequency of the motif occurrence
- by the residue range in which the motif occurs.



Nucleic Acid Subsequence Analysis Display

57 matches found in pBR322 from 356 to 1579 ...

Filter Options...

☑ # of sites <=

□ no sites to

☑ sites only 1 to 4361

phase: any

Display Options...

☑ List sites by... name

□ List sites not found

☑ Subsequence map...

□ Annotated sequence

□ Frag predictions... single

To filter subsequence search results

1. The Display dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example, to change the display parameters, do one of the following:

- choose **Analyze | Nucleic Acid Subsequence** when any nucleic acid subsequence display result window is active
- choose **Analyze | Protein Subsequence** when any protein subsequence display result window is active.

2. To restrict the displays to subsequences that occur a limited number of times, select the **# of sites** checkbox and type a number to set the maximum number of occurrences.
3. To restrict the displays to subsequences that do not occur within a given region, select the **no sites** checkbox, then either type in the numbers of the bases that bracket the region, or choose a region from the features table drop-down menu at the right of the text boxes.
4. To restrict the displays to subsequences that occur only within a given region, select the **sites only** checkbox, then either type in the numbers of the bases that bracket the region, or choose a region from the features table drop-down menu at the right of the text boxes.
5. For nucleic acids, to restrict displays to subsequences that cut only within a given reading frame, select the **sites only** checkbox, then choose the frame from the **phase** drop-down menu.

Tip. If you want to specify a reading frame but not limit the displays to a region, select **ALL 1- ...** from the features table drop-down menu.

6. Select the **Display Options** as required.

These options are described in the next section.

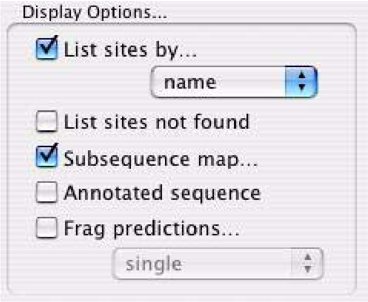
7. Select **OK** to display the results.

Displaying subsequence search results

You can display the nucleic acid subsequence search results in a number of ways:

- as a list of the subsequences found
- as a list of the subsequences not found
- as a subsequence map
- as an annotated sequence
- as a list of the sequence sizes found between subsequence occurrences.

One or more of the displays can be generated at the same time. Each display type can be saved to disk or printed when its window is active. Refer to Chapter 3, *“General Procedures”*, for further details.



Display Options...

☑ List sites by...

name

□ List sites not found

☑ Subsequence map...

□ Annotated sequence

□ Frag predictions...

single

To display subsequence search results

1. The Display dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example, to change the display parameters, do one of the following:
   - choose **Analyze | Nucleic Acid Subsequence** when any nucleic acid subsequence display result window is active
   - choose **Analyze | Protein Subsequence** when any protein subsequence display result window is active.
2. Select the **List sites by** check box to display a list of the subsequences found and the locations of the sites.

Use the drop-down menu to choose how to order the list: by name, by position of sites, or by number of sites.

3. Select the **List sites not found** check box to display a list of the subsequences that were considered in the search but were not found in the sequence.
4. Select the **Subsequence map** check box to display the locations of the cut sites in a graphical form.The appearance of this display can be changed interactively using the Graphics Palette. Refer to *“Editing the general map appearance”* on page 109, for further details.

Tip. Right-click on a cut site in the subsequence map and select either **Create misc_binding Feature** (for nucleaic acid subsequences) or **Create BINDING Feature/ Create MOTIF Feature** (for protein subsequences) from the context menu to add that feature to the original seqeunce in a Genbank compatible format.

5. Select the **Annotated sequence** check box to display a Text view with the sites displayed at their proper locations along the sequence. The appearance of this view can be changed interactively using the **Text Display** preferences dialog box. Refer to *"Formatting the Text view"* on page 124, for further details.
6. Select the **Frag predictions** check box to display sequence sizes between specified motifs. The motifs used are those that have satisfied the **Filter Options**.

Use the drop-down menu to choose from the following:

- **single** for sizes between a single motif
- **double** for sizes between each pair of motifs
- **both** for sizes of both the **single** and **double** options
- **all combined** for sizes using up to six selected motifs

9 Click Cloning

Overview

This chapter describes the click cloning functionality available in MacVector and how it can be used to design and document cloning strategies that might be performed in a practical laboratory, including:

- Point and click digestion and ligation of target sequences into vectors
- Blunting using Klenow, and other modification of digested fragments
- Simulation of the TOPO and Gateway cloning technologies from Invitrogen

Refer to Chapter 8, *"Searching for Sites and Motifs"* for information about how to find restriction sites and generate restriction maps of your sequences.

Contents

Click Cloning

The click cloning feature in MacVector allows you to replicate or design cloning experiments that you might perform in the laboratory. It provides a way of constructing new DNA molecules from existing clones and vectors by selecting restriction enzyme sites, digesting the intervening fragment and then ligating it into a target molecule.

Click cloning is usually performed in conjunction with either automatic or manual restriction enzyme searches (see *"Searching for restriction enzyme sites"* on page 170) which identify the locations of common restriction enzyme cut sites in nucleic acid sequences.

The click-cloning tool recognizes sticky or blunt ends and only allows automatic cloning of fragments if these are compatible, preserving sites and feature information, where needed. It also provides tools to manually flip the fragment and manipulate any incompatible sticky ends prior to cloning, should this be necessary.

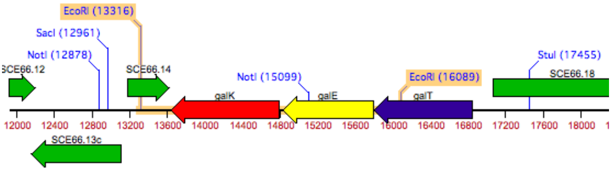
You can access the click-cloning functionality using either the **Digest** and **Ligate** tools on the Sequence Map view toolbar (see *"Map view"* on page 92) or the **Edit | Digest** and **Edit | Ligate** options from the menu.

To perform basic click cloning with an automatic restriction enzyme search

1. Open the sequence files containing the gene or fragment you want to clone and the vector sequence you want to clone into.
2. Ensure that **Automatic RE Analysis** is checked on and the appropriate enzymes are selected in the specified enzyme file. See *"To set up automatic restriction enzyme site searching"* on page 171

The sequence is searched for restriction enzyme cut sites automatically and any that are located are displayed in the sequence Map view.

3. Check that there are suitable restriction sites surrounding the gene or fragment you want to clone.

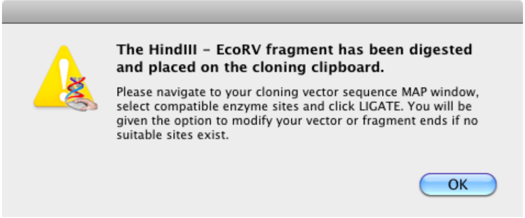


Restriction enzyme recognition sites that appear only once in a sequence are displayed in red in the Map view, whereas sites that are present more than once are shown in blue. Double-clicking on a restriction site selects all the sites of the same type, for example EcoRI, BamHI, etc., making it easy to see where sites of a particuar type lie on the sequence. Moreover, when the sequence is viewed at the residue level, the cut sites are shown. Select a restriction fragment to highlight the structure of any staggered ends present.

If the enzyme file being used for the automatic restriction enzyme search is open, then you can also select different enzymes in the file to see if they are present in both sequences.

4. Select suitable flanking sites around the gene or fragment, by clicking on one, then holding down **<shift>** and selecting the other.
5. Click the **Digest** icon on the Map view toolbar or choose **Edit | Digest** from the menu.

The selected gene or fragment, together with sticky end information and any overlapping annotations or feature information, is copied to the cloning clipboard. A dialog is displayed confirming that the fragment has been digested and which enzymes were used.



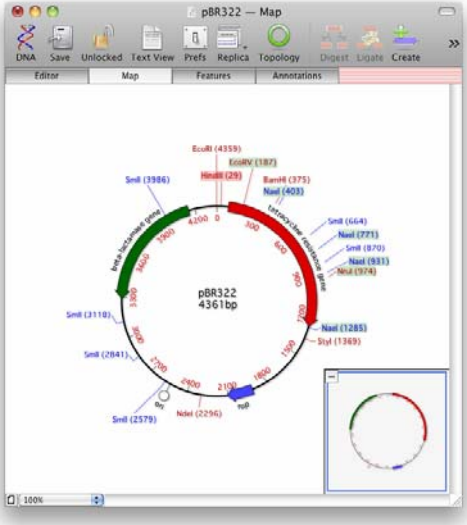
The HindIII – EcoRV fragment has been digested and placed on the cloning clipboard.

Please navigate to your cloning vector sequence MAP window, select compatible enzyme sites and click LIGATE. You will be given the option to modify your vector or fragment ends if no suitable sites exist.

Note. You can also use the **Edit | Copy** menu item or **<cmd+c>** on the keyboard to copy the selected fragment to the cloning clipboard.

6. Make the vector sequence window active and select the enzyme site you want to paste the gene or fragment into.

Sites in the vector sequence are highlighted in pale red and pale green, to indicate which end of the digested fragment they are compatible with.



7. Click the **Ligate** icon on the Map view toolbar or choose **Edit | Ligate** from the menu.
8. If prompted, click the **Unlock...** button to confirm that you want to make changes to the selected vector sequence.

The ligation pop-up window is displayed. It shows the structures of the gene or fragment and vector ends, with color-coded outlines to indicate whether they are compatible.

If the colors of the outlines around pairs of gene or fragment and vector ends match, then they are compatible and the **Ligate** button is enabled.

Tip. You can configure the colors used to identify compatible vector ends using MacVector | Preferences | Colors (see *“To change the default colors used in the Editor view”* on page 85 for more details.

9. Optionally, click the **Flip** button to reverse and complement the gene or fragment prior to ligation.

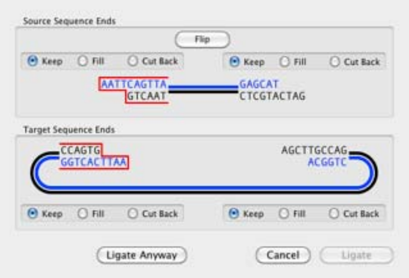
Note. MacVector detects when the gene or fragment *must* be flipped in order to make the sticky ends compatible and performs the operation automatically as part of the ligation step. In these circumstances the **Ligate** button is renamed **Flip & Ligate**.

10. Click the **Ligate** or **Flip & Ligate** button.

The gene or fragment on the cloning clipboard is inserted into the selected enzyme site on the vector sequence.

Note. You can also use the **Edit | Paste** menu item or **<cmd+v>** on the keyboard to paste the fragment on the cloning clipboard into the current sequence at the selected site. As long as the fragment and vector ends are compatible, the fragment will be cloned into the vector sequence at the specified site automatically. It will also be flipped, if required. If the ends are not compatible, then the ligation pop-up window will be displayed, allowing you to manipulate the incompatible sticky ends prior to ligation (see below).

Tip. Instead of using the **Automatic RE Analysis** tool for click cloning, you may wish to use Filtering restriction site results (see *"To filter restriction site search results"* on page 174). This, more powerful search tool, allows you to chose only enzymes that cut in a particular site (i.e. the MCS) or only enzymes that do not cut in the gene you want to clone.



Source Sequence Ends

● Keep ○ Fill ○ Cut Back

● Keep ○ Fill ○ Cut Back

Target Sequence Ends

● Keep ○ Fill ○ Cut Back

● Keep ○ Fill ○ Cut Back

Manipulating Sticky Ends

Sometimes during a cloning procedure it is necessary to modify the ends of a fragment by either filling in or cutting back the sticky ends produced by a restriction digest.

Some restriction enzymes (e.g. EcoRV) will leave no overhangs and most PCR products have blunt ends (see *"Gateway and TOPO Cloning"* on page 193). Therefore you may be left with one or more blunt ends on one or both of your fragments. If you have blunt ends on both fragments, then these can be ligated together. However, if you intend to ligate a blunt end to a sticky end, then you must modify the sticky end so that it is compatible with the blunt end. Additionally, it can be impossible to obtain a compatible site for some cloning strategies. In these circumstances, an option is to digest with whatever enzyme is suitable and then to blunt both ends to make them compatible.

Note. If both ends of a fragment are blunt, then it is impossible to determine the orientation of the fragment after ligation into the vector. In such situations MacVector will show only one of the two possible constructs.

The most commonly used techniques for these types of modifications are to use either DNA polymerase I large fragment (Klenow fragment) or T4 DNA polymerase to fill in a overhanging 5' end. In addition, T4 DNA polymerase may be used to digest away an overhanging 3' end.

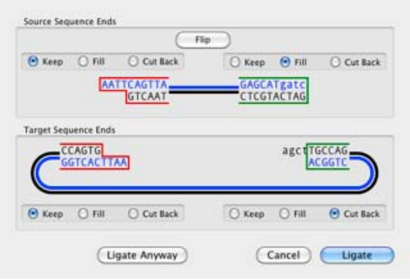
Both of these enzymes have 5'-3' polymerase and 3'-5' exonuclease activities, though the exonuclease activity of Klenow is much weaker and so T4 DNA polymerase is preferred. Neither of the enzymes has a 5'-3' exonuclease activity.

The ligation pop-up window provides the tools to perform these manipulations before joining the fragment and vector ends together.

To manipulate sticky ends prior to ligation

1. Select the **Fill** option associated with a particular end to fill in the unwanted 5’ overhang on that end.
2. Select the **Cut Back** option associated with a particular end to cut back the unwanted 3’ or 5’ overhang on that end.

Tip. Any portions of the sequence which have been filled in or cut back are now shown in lowercase.



Source Sequence Ends

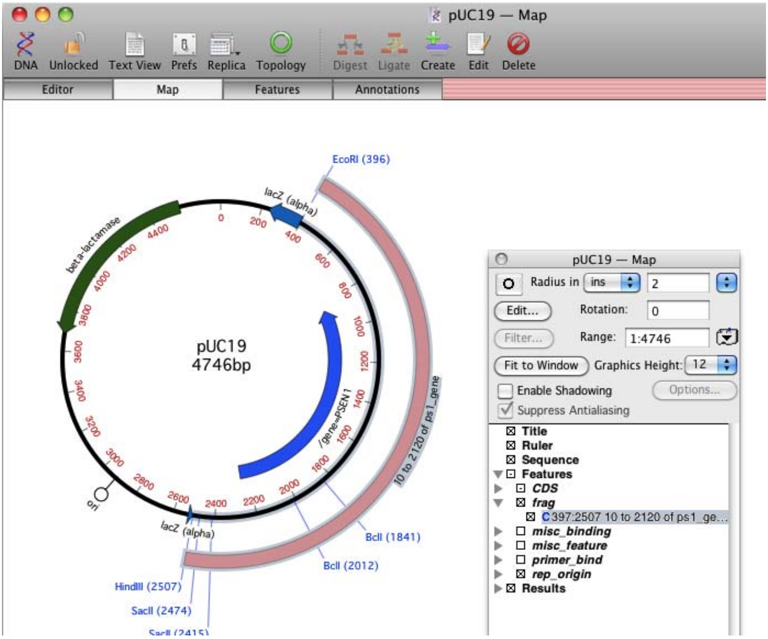
● Keep ○ Fill ○ Cut Back   ○ Keep ● Fill ○ Cut Back

Target Sequence Ends

● Keep ○ Fill ○ Cut Back   ○ Keep ○ Fill ● Cut Back

Maintaining a cloning history

Whenever you paste a fragment into a target molecule a special “frag” feature is created in the target with a note describing the source of the fragment of DNA. You can use this to keep track of the history of your constructs.



○ Radius in: ins 2

Rotation: 0

Range: 1:4746

Graphics Height: 12

☐ Enable Shadowing

☑ Suppress Antialiasing

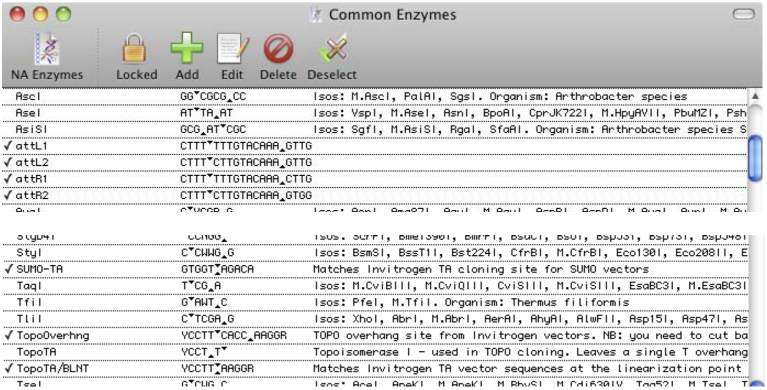
- ☒ Title
- ☒ Ruler
- ☒ Sequence
- ☐ Features
  - ☐ *CDS*
  - ☒ *frag*
    - ☒ C 397:2507 10 to 2120 of ps1_ge...
  - ☐ *misc_binding*
  - ☐ *misc_feature*
  - ☐ *primer_bind*
  - ☒ *rep_origin*
- ☒ Results

Gateway and TOPO Cloning

Invitrogen market a number of kits and cloning vectors designed to simplify the cloning and subcloning of DNA fragments, such as those generated by PCR amplification. Among the most popular are the Gateway and TOPO kits. For detailed information on the concepts, availability and pricing of these kits, please visit the Invitrogen web site at www.invitrogen.com.

MacVector now allows you to simulate these popular cloning technologies, so that you can accurately document the sequences of the constructed molecules.

Note. In order to perform the Gateway and TOPO procedures described here you must use the Common Enzymes restriction enzyme file and the default selection must include the TOPO and Gateway sites shown below.



AscI	GG▾CGCG▴CC	Isos: M.AscI, PalAI, SgsI. Organism: Arthrobacter species
AseI	AT▾TA▴AT	Isos: VspI, M.AseI, AsnI, BpoAI, CprJK722I, M.HpyAVII, PbuMZI, Psh
AsiSI	GCG▴AT▾CGC	Isos: SgfI, M.AsiSI, RgaI, SfaAI. Organism: Arthrobacter species S
✓ attL1	CTTT▾TTTGTACAAA▴GTTG	
✓ attL2	CTTT▾CTTGTACAAA▴GTTG	
✓ attR1	CTTT▾TTTGTACAAA▴CTTG	
✓ attR2	CTTT▾CTTGTACAAA▴GTGG	
StyI	C▾CWWG▴G	Isos: BsmSI, BssT1I, Bst224I, CfrBI, M.CfrBI, Eco130I, Eco208II, E
✓ SUMO-TA	GTGGT⧗AGACA	Matches Invitrogen TA cloning site for SUMO vectors
TaqI	T▾CG▴A	Isos: M.CviBIII, M.CviQIII, CviSIII, M.CviSIII, EsaBC3I, M.EsaBC3I
TfiI	G▾AWT▴C	Isos: PfeI, M.TfiI. Organism: Thermus filiformis
TliI	C▾TCGA▴G	Isos: XhoI, AbrI, M.AbrI, AerAI, AhyAI, AlwFII, Asp15I, Asp47I, As
✓ TopoOverhng	YCCTT▾CACC▴AAGGR	TOPO overhang site from Invitrogen vectors. NB: you need to cut ba
TopoTA	YCCT▴T▾	Topoisomerase I - used in TOPO cloning. Leaves a single T overhang
✓ TopoTA/BLNT	YCCTT⧗AAGGR	Matches Invitrogen TA vector sequences at the linearization point

Vector Sequences

There are a wide variety of vectors available both from Invitrogen and other sources that can be used in TOPO and Gateway cloning. The sequences and maps of the Invitrogen vectors can be downloaded from their website. MacVector includes a selection of these vectors in the */Applications/MacVector 11/Common Vectors/Invitrogen* folder.

TOPO Cloning

This technique relies on the single-stranded DNA nicking and religation properties of Topoisomerase I from Vaccinia virus. This enzyme binds to duplex DNA at specific sites and cleaves the phosphodiester backbone after 5'-YCCTT on one strand. The energy from the broken phosphodiester backbone is conserved by formation of a covalent bond between the 3' phosphate of the cleaved strand and a tyrosyl residue (Tyr-274) of topoisomerase I. The phospho-tyrosyl bond between the DNA and enzyme can subsequently be attacked by the 5' hydroxyl of the original cleaved strand, reversing the reaction and releasing topoisomerase. TOPO Cloning exploits this reaction to efficiently clone PCR products.

Invitrogen provide linearized vectors that have Topoisomerase covalently bound to each end – when mixed with PCR fragments, the fragments are joined to the vector with high efficiency, making it a far more effective method for cloning PCR fragments than standard blunt ended DNA ligase approaches.

There are two main types of TOPO cloning:

- TOPO-TA Cloning
- Zero Blunt TOPO Cloning

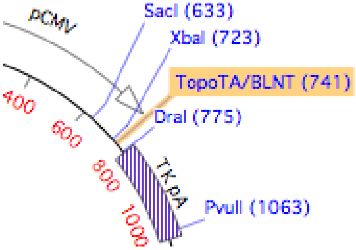
There is a third approach that uses 4 residue single-stranded overhangs at one end of the vector to allow directional cloning of fragments. You can simulate this approach using MacVector although it is not documented here. Contact MacVector support if you need more information.

TOPO-TA Cloning

Taq polymerase has a non template-dependent activity that adds a single deoxyadenosine (A) to the 3´ ends of PCR products. This reduces the efficiency of ligation into normal blunt-ended vectors. However, TOPO-TA vectors are supplied with single 3´ deoxythymidine (T) residues attached to the ends, providing a complementary overhang that increases the efficiency of ligation.

To simulate TOPO-TA cloning with MacVector

1. Select the region in a source DNA molecule corresponding to the PCR fragment.
2. Click the **Digest** icon on the Map view toolbar or choose **Edit | Digest** from the menu to copy the blunt ended fragment to the cloning clipboard.
3. Open the target vector molecule and view the TopoTA/BLNT cloning site in the Map view using the Common Enzymes set of sites.
4. Select the TopoTA/BLNT site and choose **Edit | Ligate** from the menu.



5. Click the **Ligate** button in the ligation pop-up window to insert the fragment.

Zero Blunt TOPO Cloning

This technique is very similar to the TOPO-TA approach, except that the linear vector molecules are completely blunt ended and do not have 3’T overhangs. These vectors should be used whenever a “proof-reading” polymerase is used in the PCR reaction. These polymerases, like Pfx DNA Polymerase (also available from Invitrogen) do not leave 3’A overhanging residues and instead produce a completely flush end.

To simulate Zero Blunt TOPO cloning with MacVector

1. Select the region in a source DNA molecule corresponding to the PCR fragment.
2. Click the **Digest** icon on the Map view toolbar or choose **Edit | Digest** from the menu to copy the blunt ended fragment to the cloning clipboard.
3. Open the target vector molecule and view the TopoTA/BLNT cloning site in the Map view using the Common Enzymes set of sites.
4. Select the TopoTA/BLNT site and choose **Edit | Ligate** from the menu.
5. Click the **Ligate** button in the ligation pop-up window to insert the fragment.

Gateway Cloning

Gateway cloning is a technology that takes advantage of *in vitro* recombination using att sites derived from the bacteriophage lambda chromosomal integration system. Once a fragment of interest has been cloned into an “Entry” vector, flanked by att sites, it can be rapidly transferred into a wide variety of engineered “Destination” vectors by incubating the entry clone with destination vector and recombination proteins, then transforming a suitable host and selecting the destination vector. Most Invitrogen Entry vectors use Kanamycin resistance as the primary selectable marker, whereas the Destination vectors use Ampicillin resistance. In addition, the Destination vectors typically contain a strain-specific lethal gene (usually ccdB) flanked by the att sites that gets replaced when the recombination event occurs, allowing direct selection for successful recombinant clones.

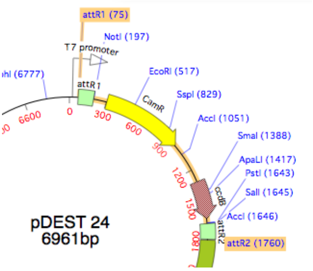
Another important feature of the system is that the recombination proteins have a slight flexibility in recognizing the sequence of the recombination sites, although the source and destination sites must be identical for recombination to occur. This means that by varying the sequence at the recombination site by a single residue, the system ensures that recombination always occurs in the desired orientation. There are 8 primary att sites used in Gateway cloning that are included in the Common Enzymes file and and separately in the Gateway file:

- `attR1 CTTT TTTGTACAAA CTTG`
- `attR2 CTTT CTTGTACAAA GTGG`
- `attR3 CTTT ATTATACATA GTTG`
- `attR4 CTTT TCTATACAAA GTTG`
- `attL1 CTTT TTTGTACAAA GTTG`
- `attL2 CTTT CTTGTACAAA GTTG`
- `attL3 CAAC TTTGTATAAT AAAGT`
- `attL4 CAAC TTTTGTATAC AAAGT`

attR1 and attL1 share the normal core sequence of the lambda att site, but attR2 and attL2 contain a C instead of a T within the core sequence. This difference ensures that attR1 can only recombine with attL1 and attR2 can only recombine with attL2. All the Invitrogen vectors contain matched flanking L1/L2 or R1/R2 att site pairs, so the recombination event always occurs in the same orientation, allowing you to exactly predict the product that will result from the recombination. There are other related L3/L4/L5 etc. variations that can be used, but these are far less common.

To simulate Gateway cloning with MacVector

1. Select the att sites in the source vector by holding down the **<shift>** key and clicking on the sites.



2. Click the **Digest** icon on the Map view toolbar or choose **Edit | Digest** from the menu to copy the fragment to the cloning clipboard.
3. Open the destination vector and select the corresponding target att sites.
4. Choose **Edit | Ligate** from the menu.
5. Click the **Ligate** button in the ligation pop-up to insert the source fragment into the destination vector.

10 Primer and Probe Design

Overview

MacVector provides the following tools to design primers and screen for likely primers and probes:

- Primer3 for searching for primer pairs or finding a matching primer for one designed using Quicktest Primer.
- Quicktest Primer for interactive design and evaluation of primers, including the introduction of mismatches or addition of tails and viewing the effect of such changes on protein open reading frames.
- screening a nucleic acid sequence for likely sequencing primers or hybridization probes.
- screening a protein sequence for the least degenerate oligos that can serve as hybridization probes.
- screening sequencing and PCR primer sequences for specific characteristics.

Contents

Scanning for suitable primers

MacVector includes an automated primer searching module based on Primer3. The module has been designed to be easy to use - you can create a new primer in as little as three mouse clicks using the default settings - whilst retaining all of the powerful features of Primer3.

As well as scanning for primer pairs, this module can also be used to:

- test pairs of primers
- design matching primers for a known primer
- design hybridization primers for use in realtime PCR analyzes and similar techniques.

Scanning for primers using Primer3

Amplify Feature/Region

The simplest way to scan for primers with MacVector is to select a particular feature or region of a sequence and design primers either side of it, to amplify the selection.

To scan for a primer using the Amplify Feature/Region method

1. Open the sequence file you want to analyze.
2. Select a feature of the sequence in the **Map** tab or the **Features** tab. Alternatively, select a region of the sequence in the **Editor** tab.

By default, primers are chosen from a 200 bp long region on either side of the feature or region selected.

3. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.

Note. When the **Primer Design (Primer3)** dialog box opens it is populated with information about the feature or region you selected in the active sequence window. You can populate the dialog box with information about an alternative feature in the active sequence using the feature selector drop-down menu, or you can specify an alternative region by providing residue numbers manually.

4. Ensure that **Amplify Feature/Region** is selected from the design method drop-down list.
5. Click **OK**.

The default primer design settings will be used to design the new primer. These will produce a ranked list and graphical map of the top five primer pairs that will amplify the selected feature or region, using the optimum values for the primer Tm and %GC. However, by modifying the advanced parameters (see *"Advanced Primer3 settings"* on page 206) it is possible to tune these, and other values, to obtain the best primers for non-standard situations.

Region to Scan

Alternatively, you can specify a region within which the amplified product should lie. Simply choose the size of product you want and Primer3 will design primers to amplify products of that size, from anywhere in that region.

To scan for a primer using the Region to Scan method

1. Open the sequence file you want to analyze.
2. Select a feature of the sequence in the **Map** tab or the **Features** tab. Alternatively, select a region of the sequence in the **Editor** tab.
3. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.

Note. When the **Primer Design (Primer3)** dialog box opens it is populated with information about the feature or region you selected in the active sequence window. You can populate the dialog box with information about an alternative feature in the active sequence using the feature selector drop-down menu, or you can specify an alternative region by providing residue numbers manually.

4. Select **Region to Scan** from the design method drop-down list.
5. Click **OK**.

By default this method will design primers to produce products between 100 and 300 bases in length. However, you can change the minimum and maximum **Product size** values to design primers to produce shorter or longer products.

Flanking Regions

Finally, you can design primers by specifying two flanking regions and selecting left and right primers from these. This is similar to the Amplify Feature/Region method but it is much more flexible, as you can specify more precisely the region that each primer is selected from.

To scan for a primer using the Flanking Regions method

1. Open the sequence file you want to analyze.
2. Select the feature of the sequence that defines the flanking regions you are interested in on the **Map** tab or the **Features** tab. Alternatively, select the region of the sequence that defines the flanking regions you are interested in on the **Editor** tab.

3. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.
4. Select **Flanking Regions** from the design method drop-down list.

The values for the left region and the right region are populated based on the sequence selection you made above. These initial values define flanking regions of zero length.

5. Modify the start point of the left region and the end point of the right region until they encompass the regions you want to choose primers from.
6. Click **OK**.

Note. There are no limits on the size of the regions that can be used, however, if they are very short then it may not be possible to find suitable primers with the default settings.

Testing primers using Primer3

You can also use the primer design module to test primer sequences. You can test existing primer pairs to ensure that they match and produce the expected product. Alternatively, you can provide one primer and use the primer design module to design a suitable matching primer for a specified region or sequence, or one that produces a product of a specified size.

To test an existing primer pair

1. Open the sequence file you want to analyze.
2. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.
3. Check the **Use this primer** option for both primers, then paste or type the existing primer sequences into the appropriate **Primer Sequences** edit boxes.
4. Select **Region to Scan** from the design method drop-down list
5. Ensure that the **Region to Scan** values entirely encompass the area that contains the expected product. If in doubt, select the entire sequence.
6. Set the **Product size** values generously, e.g. between 0.5kb below and 0.5kb above the product size you would expect.
7. Click **OK**.

To find a new primer to match an existing primer for a specified feature or region

1. Open the sequence file you want to analyze.

2. Select the feature of the sequence that you want to amplify in the **Map** tab or the **Features** tab. Alternatively, select the region of the sequence that you want to amplify in the **Editor** tab.
3. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.
4. Check the **Use this primer** option for the existing primer, then paste or type the existing primer sequence into the corresponding **Primer Sequences** edit box.
5. Ensure that the **Find primer** option is selected for the primer you want to design.
6. Ensure that **Amplify Feature/Region** is selected from the design method drop-down list.
7. Click **OK**.

MacVector will test the existing primer, then design suitable matching primers. Alternatively, to design a matching primer to produce a product of a specified size, use the following procedure.

To find a new primer to match an existing primer for any product size

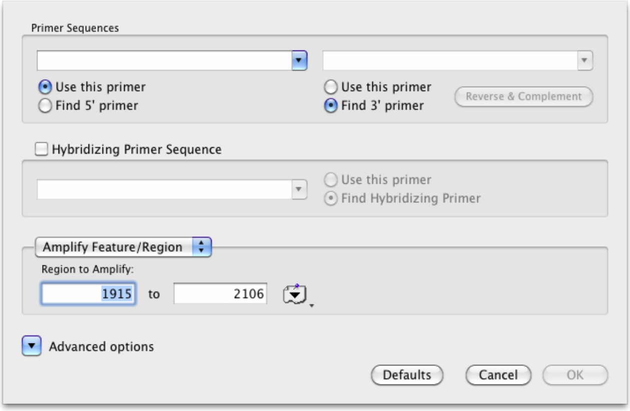
1. Open the sequence file you want to analyze.
2. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.
3. Check the **Use this primer** option for the existing primer, then paste or type the existing primer sequence into the corresponding **Primer Sequences** edit box.
4. Ensure that the **Find primer** option is selected for the primer you want to design.
5. Select **Region to Scan** from the design method drop-down list
6. Click **OK**.

By default this method will design a matching primer to produce products between 100 and 300 bases in length. However, you can change the minimum and maximum **Product size** values to design primers to produce shorter or longer products.

Real-time primer design

Many real time PCR techniques use a third oligonucleotide that will anneal inside the desired product. This makes is possible to track the progress of the PCR amplification using one of the many proprietary technologies available. The primer design module enables you to design these.

You can design an internal primer suitable for use with a pair of existing primers, design external primers to suit an existing internal primer, or design all three primers from scratch.



Primer Sequences

● Use this primer
○ Find 5' primer

○ Use this primer
● Find 3' primer

□ Hybridizing Primer Sequence

○ Use this primer
● Find Hybridizing Primer

Amplify Feature/Region

Region to Amplify: 1915 to 2106

To design a pair of external primers and a suitable internal hybridization primer from scratch

1. Open the sequence file you want to analyze.
2. Select the feature of the sequence you want to monitor in the **Map** tab or the **Features** tab. Alternatively, select the region of the sequence you want to monitor in the **Editor** tab.

By default, primers are chosen from a 200 bp long region on either side of the feature or region selected.

3. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.

Note. When the **Primer Design (Primer3)** panel opens it is populated with information about the feature or region you selected in the active sequence window. You can populate it with information about an alternative feature in the active sequence using the feature selector drop-down menu, or you can specify an alternative region by providing residue numbers manually.

4. Ensure that **Amplify Feature/Region** is selected from the design method drop-down list.

Tip. Alternatively, use the **Flanking Regions** design method to gain more control over the regions from which the primers are chosen.

5. Ensure that the **Find primer** option is selected for both external primers.
6. Check the **Hybridizing Primer Sequence** box.
7. Ensure that the **Find Hybridizing Primer** option is selected.
8. Click **OK**.

To scan for a suitable internal hybridization primer for a pair of existing external primers

1. Open the sequence file you want to analyze.
2. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.
3. Check the **Use this primer** option for both primers, then paste or type the existing primer sequences into the appropriate **Primer Sequences** edit boxes.
4. Select **Region to Scan** from the design method drop-down list
5. Ensure that the **Region to Scan** values entirely encompass the area that contains the expected product. If in doubt, select the entire sequence.
6. Set the **Product size** values generously, e.g. between 0.5kb below and 0.5kb above the product size you would expect.
7. Check the **Hybridizing Primer Sequence** box.
8. Ensure that the **Find Hybridizing Primer** option is selected.

Note. You should ensure that the advanced primer design settings on the **Hybridization Primer** tab are set to values appropriate for the primer technology you are using. The documentation supplied with your primer technology should contain this information.

9. Click **OK**.

To scan for a suitable pair of external primers for an existing internal primer

1. Open the sequence file you want to analyze.
2. Select the feature of the sequence you want to monitor in the **Map** tab or the **Features** tab. Alternatively, select the region of the sequence you want to monitor in the **Editor** tab.

By default, primers are chosen from a 200 bp long region on either side of the feature or region selected.

3. Choose **Analyze | Primers | Primer Design (Primer3)** from the menu.

Note. When the **Primer Design (Primer3)** dialog box opens it is populated with information about the feature or region you selected in the active sequence window. You can populate the dialog box with information about an alternative feature in the active sequence using the feature selector drop-down menu, or you can specify an alternative region by providing residue numbers manually.

4. Ensure that **Amplify Feature/Region** is selected from the design method drop-down list.
5. Check the **Hybridizing Primer Sequence** box.
6. Check the **Test this primer** option for the hybridizing primer, then paste or type the existing internal primer sequence into the **Hybridizing Primer Sequence** edit box.
7. Click **OK**.

Advanced Primer3 settings

The default primer design parameters are set such that most primer design experiments can be run successfully without the need to make any changes. However, all of the parameters can be accessed and modified using the **Advanced options** button on the **Primer Design (Primer3)** dialog box.

The following tabs, containing groups of related primer design parameters, are displayed.

Tip. Users familiar with Primer3 can hover the mouse over parameters on these tabs to display tooltips with information about which underlying Primer3 parameters they correspond to. See *“Primer design”* on page 421 for further information.

Characteristics

On this tab you can select minimum, maximum and optimum values for **Length**, **Percent G+C**, and **Tm**. Primer3 will always try to design primers that match the optimum values specified as closely as possible.

Note. The optimum value for these parameters does not need to be midway between the minimum and maximum values.

The **GC Clamp** is the number of G and/or C residues you want to include at the 3' end of each primer. **Maximum Poly-X** limits the number of consecutive residues of the same type that can appear in a primer sequence. The **Maximum difference in Tm between primers** setting enables you to ensure that primers are closely matched. By default, the allowed difference is set to a quite high value, however, Primer3 will still usually find well-matched primers.

Primer Binding

On this tab you can set values for **Primer vs Primer**, **3' end vs 3' end** and **Primer vs Product** which control the maximum acceptable binding between the primers and the product. The units of these parameters are defined internally by Primer3. Increase the values to make the primer design experiment less sensitive to binding. This may be necessary if you are scanning a short sequence and cannot find suitable primers.

To remove **Primer vs Product** filtering altogether, set the value to -1.

Check the **Allow ambiguous residues** box if the sequence you are analyzing contains ambiguous residues. Optionally, specify the maximum number of ambiguous residues you are willing to accept in the primers using the **Allowed Ns in primers** edit box.

Reaction Conditions

On this tab you specify the reaction conditions you use in the lab. These parameters are shared by all the Primers tools, so making a change to any of the values on this tab will affect the values displayed and used by PCR Primer Pairs, Sequenceing Primers/Probes, etc. The reaction conditions values specified, affect the reported Tm of the primer binding to the template.

Hybridization Primer

On this tab you can set alternative basic parameters for hybridization primer scans. See *“Characteristics”* on page 206 for further details about the parameters on this tab.

Note. These parameters are completely independent from the basic parameters used for external primers. So, if you modify the basic parameters used for external primers on the **Characteristics** tab, then those changes will not be reflected on the **Hybridization Primer** tab.

Misc.

On this tab you can control the number of primer pairs that are reported. The default is 5 but you can choose up to 100.

Note. The Graphical map output can only display the top 25 primers (see *“Graphical map”* on page 209).

Analyzing Primer3 results

Initially, the results of primer design experiments are presented in the **Primer Design Display** dialog box. This summarizes the number of primers found, the number that were rejected, the number of suitable pairs of primers found and the number that were rejected. This information is divided into four sections.

- A summary of the number of individual primers that were considered and how many were accepted.
- A summary of the reasons why the rejected primers were discarded. This section is useful for tuning the analysis if the default values do not produce acceptable primers.
- A summary of the number of pairs of primers that were considered and how many were rejected,
- A summary of the reasons why any rejected primer pairs were discarded.

Note. If your primer design experiment included the design of an internal hybridization primer, then statistics for this will also be included in these summaries.

The **Primer Design Display** dialog also provides a choice of three display options, which enable you to view the detailed results of the primer design experiment in different ways.

To select display options

1. Check the boxes adjacent to the display options you want to see.

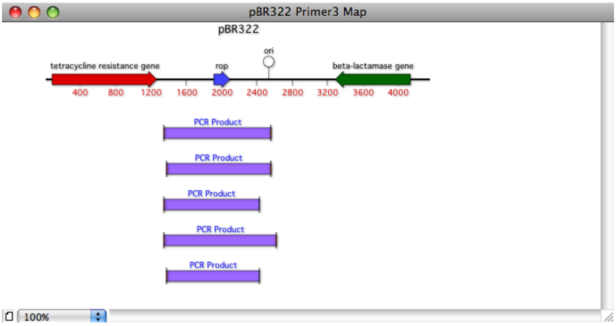
You can select as many of the different options as you like.

2. Click **OK**.

Note. The selection you make the first time the **Primer Design Display** dialog box appears is retained throughout your MacVector session.

Primer3 output

The **Primer3 output** option displays the raw output from Primer3. See *"Primer3 output file format"* on page 432 for a description of the syntax of this output.



Graphical map

The **Graphical map** option displays the primer pairs identified, the internal hybridization primers identified (if this option was selected) and the predicted product or amplicon in a new Primer3 map window, along with any existing features of the sequence.

Tip. Right-click on a primer pair or product in the graphical map and select either **Create primer_bind Feature** or **Create misc_feature Feature** from the context menu to add that feature to the original sequence in a Genbank compatible format.

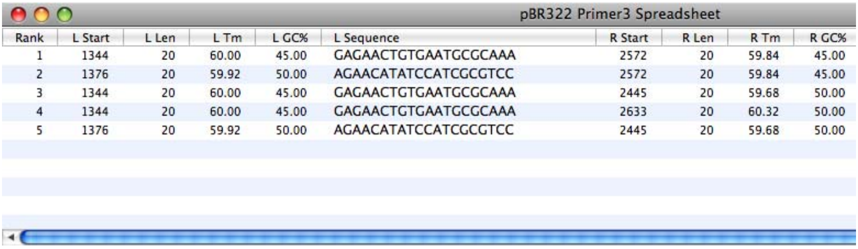
Like all other map views in MacVector, you can control the appearance of the Primer3 map using the Graphics Palette.

The Primer3 map window and the parent sequence window are linked so, if you click on any graphical object in the Primer 3 map window, then the corresponding region is selected in the parent sequence window.

Spreadsheet

The **Spreadsheet** option displays the primer sequences and detailed statistics about them in a new Primer3 spreadsheet window.

You can copy and paste individual sequences, entire lines, or the whole spreadsheet into other documents, including Excel spreadsheets and others that can import tab separated values. This makes it easy to distribute the primer sequence information to your oligonucleotide synthesis service. You can also save the Primer3 spreadsheet directly, as a comma separated value (CSV) file.



pBR322 Primer3 Spreadsheet

Rank	L Start	L Len	L Tm	L GC%	L Sequence	R Start	R Len	R Tm	R GC%
1	1344	20	60.00	45.00	GAGAACTGTGAATGCGCAAA	2572	20	59.84	45.00
2	1376	20	59.92	50.00	AGAACATATCCATCGCGTCC	2572	20	59.84	45.00
3	1344	20	60.00	45.00	GAGAACTGTGAATGCGCAAA	2445	20	59.68	50.00
4	1344	20	60.00	45.00	GAGAACTGTGAATGCGCAAA	2633	20	60.32	50.00
5	1376	20	59.92	50.00	AGAACATATCCATCGCGTCC	2445	20	59.68	50.00

You can also sort any of the columns in the Primer3 spreadsheet. By default the results are sorted by rank each pair has been given. So, the first primer pair listed is the most suitable and so on. However, you can also sort the results by Tm or %GC values. To do this click on the heading cell or the appropriate column.

The Primer3 map and spreadsheet windows are linked so, if you select a primer in the Primer3 spreadsheet window, then that primer feature is selected in the Primer3 map window and in the parent sequence window. Similarly, if you select a primer feature in the Primer3 map window, then it is also selected in the Primer3 spreadsheet window and the parent sequence window.

Quicktest Primer

Quicktest Primer allows you to quickly evaluate the suitability of any short (~20 residues) DNA sequence for use in PCR or sequencing experiments. Simply open the **Quicktest Primer** dialog using **Analyze | Primers | Quicktest Primer...** and type or paste in a primer sequence to display of the primer properties instantly. Alternatively, you can evaluate any short (up to 80 nucleotides) portion of an existing sequence quickly by selecting that portion in the **Editor**, **Map** or **Features** view before choosing **Analyze | Primers | Quicktest Primer...** In these circumstances, the **Quicktest Primer** dialog opens with properties for the selected primer already loaded.

The **Quicktest Primer** dialog displays the following information about each “primer”:

- Tm
- Presence of “bad” residues like non-G/C 3’ end and runs of >3 homopolymers (esp G)

- “Score” of hairpin loops and primer dimers
- Graphical display of hairpin loops and primer dimers
- Location of any other matches on the target sequence

In addition, the **Quicktest Primer** dialog enables you to:

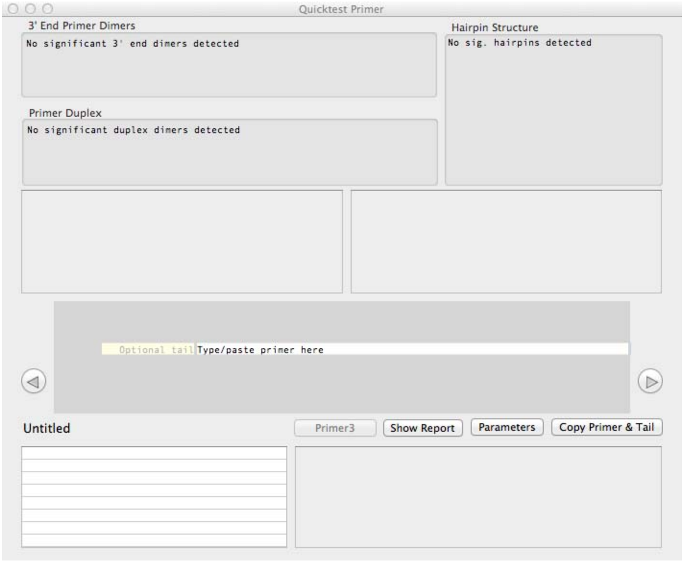
- Add a user defined 5’ tail
- Nudge the primer along the target sequence
- Find a balanced second primer using Primer3
- Create a mismatch at any location (esp corresponding to a “one out” site)

Quicktest Primer can be used in conjunction with Primer3 to rapidly design pairs of primers with mismatches and/or tails and to generate the predicted product of the reaction, complete with tails and/or mismatches.

To assess the suitability of a sequence as a primer

1. Choose **Analyze | Primers | Quicktest Primer...** from the menu.

The **Quicktest Primer** dialog box is displayed



No significant 3' end dimers detected

No sig. hairpins detected

No significant duplex dimers detected

2. Type or paste the primer sequence you want to evaluate into the central primer editing box, with the white background.

Information panes above the primer editing box show graphical representations of the highest scoring primer dimer and/or primer duplex that can be formed by the primer with itself (if any) and any hairpin loop structures that may be present in the sequence. Scrollable text boxes also display primer characteristics including molecular weight, Tm, thermodynamic properties, etc. Any other suitable binding sites in the current sequence are listed in the information pane below the primer editing box, together with a graphical representation of the alignment and other basic summary statistics.

If the primer binds to the active nucleic acid sequence, then the matching sequence is also shown aligned below the primer, on a gray background (see *"To assess the suitability of a portion of an existing sequence as a primer"* on page 212 for more details).

No binding information is displayed if the primer does not bind to the active nucleic acid sequence or if no nucleic acid sequence is active.

3. Click in the **Optional tail** region of the primer editing box and type in a sequence to add a tail to the primer.

When you make these changes, the **Quicktest Primer** display updates dynamically.

4. Click **Copy Primer & Tail** to copy the full primer to the clipboard for pasting into other applications.
5. Click **Show Report** to generate a report summarizing all the information displayed in the **Quicktest Primer** dialog.
6. Click **Primer3** to open the **Primer Design (Primer3)** dialog box, pre-populated with the primer you have been assessing. See *"Scanning for suitable primers"* on page 200 for details of how to use Primer3.

To assess the suitability of a portion of an existing sequence as a primer

1. Open the nucleic acid sequence file you want to analyze.
2. Select a primer-sized portion (less than 80 residues) of the sequence in the **Editor**, **Map** or **Features** view.
3. Choose **Analyze | Primers | Quicktest Primer...** from the menu.

The **Quicktest Primer** dialog box is displayed

The selected primer is shown in the central primer editing box, on a white background. If the selected primer binds to the active nucleic acid sequence, then below it the matching sequence is shown, aligned to the primer, on a gray background. In addition, the translations of any CDS features that overlap the binding region are also shown.

Note. The plus and minus strands of the binding region are displayed in the colors currently specified (by default these are black for the plus strand and blue for the minus strand). If the primer binds preferentially to the minus strand, then this will be displayed above the plus strand.

Information panes above the primer editing box show graphical representations of the highest scoring primer dimer and/or primer duplex that can be formed by the primer with itself (if any) and any hairpin loop structures that may be present in the sequence. Scrollable text boxes also display primer characteristics including molecular weight, Tm, thermodynamic properties, etc. Any other suitable binding sites in the current sequence are listed in the information pane below the primer editing box, together with a graphical representation of the alignment and other basic summary statistics.

The **Quicktest Primer** dialog box is linked to the active nucleic acid sequence window. So, if you select a different nucleic acid sequence window, then MacVector will recalculate the **Quicktest Primer** data.

Tip. Since no binding information is displayed if the selected primer does not bind to the current sequence, you can use this feature to establish quickly if (and where) the primer binds to any open nucleic acid sequences.

4. Click the left and right arrow icons to nudge the primer selection left and right along the sequence.
5. Click in the **Optional tail** region of the primer editing box and type in a sequence to add a tail to the primer.

When you make these changes, the **Quicktest Primer** display updates dynamically, allowing you to establish the optimum primer location quickly.

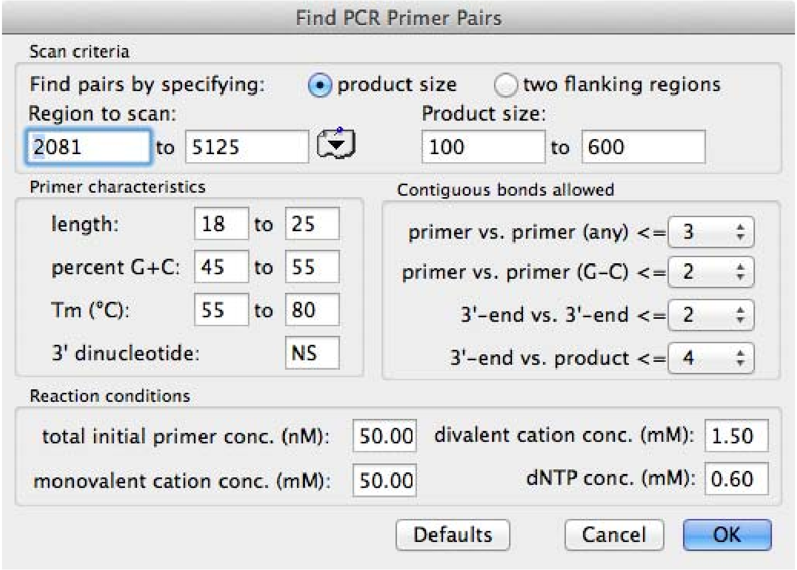
6. Click **Copy Primer & Tail** to copy the full primer to the clipboard for pasting into other applications.
7. Click **Show Report** to generate a report summarizing all the information displayed in the **Quicktest Primer** dialog.
8. Click **Primer3** to open the **Primer Design (Primer3)** dialog box, prepopulated with the primer you have been assessing. See *“Scanning for suitable primers”* on page 200 for details of how to use Primer3.

PCR primer pairs

MacVector provides an automated means of screening a nucleic acid sequence for PCR primer pairs. Refer to *“Primers and probes”* on page 421, for details of the method used.

Screening for PCR primer pairs

To use this functionality, a nucleic acid sequence must be active.



Find PCR Primer Pairs

Scan criteria

Find pairs by specifying: ● product size ○ two flanking regions

Region to scan: 2081 to 5125

Product size: 100 to 600

Primer characteristics

length: 18 to 25

percent G+C: 45 to 55

Tm (°C): 55 to 80

3' dinucleotide: NS

Contiguous bonds allowed

primer vs. primer (any) <= 3

primer vs. primer (G-C) <= 2

3'-end vs. 3'-end <= 2

3'-end vs. product <= 4

Reaction conditions

total initial primer conc. (nM): 50.00

divalent cation conc. (mM): 1.50

monovalent cation conc. (mM): 50.00

dNTP conc. (mM): 0.60

To screen for PCR primer pairs

1. Choose **Analyze | Primers | PCR Primer Pairs**.

The **Find PCR Primer Pairs** dialog box is displayed.

2. In the **Scan criteria** panel, do one of the following:

   - select **product size** if the entire nucleic acid sequence is known and you want to find primer pairs that amplify a product of a certain size in a given region
   - select **two flanking regions** to scan if each primer must lie in a certain region, or if you do not know the sequence of the product you want to amplify but do know the sequences of the regions flanking the required product.

The choice made affects the other text boxes in this panel.

3. If you selected **product size**, then do the following:
   - enter the base numbers that bracket the region you want to scan by typing them in the **Region to scan** text boxes, or by selecting a region from the feature selector drop-down menu, to the right of the boxes.
   - enter the range of product sizes that you want to amplify in the **Product size** text boxes.
4. If you selected **two flanking regions**, then do the following:
   - enter the base numbers that bracket the region where the forward primer should be found by typing them in the **Forward primer (5') region** text boxes, or by selecting a region from the features table drop-down menu at the right of the boxes.
   - enter the base numbers that bracket the region where the backward primer should be found by typing them in the **Backward primer (3') region** text boxes, or by selecting a region from the features table drop-down menu at the right of the boxes.
5. Type the required range of primer lengths in the **length** text boxes.
6. Type the required percentage range of primer G+C content in the **percent G+C** text boxes.
7. Type the required range of primer Tm values in the **Tm (°C)** text boxes.
8. Type the IUPAC code for the two nucleotides that you want to appear at the 3' end of the primers in the **3' dinucleotide** text box.
9. Use the **primer vs. primer (any)** drop-down menu to specify the maximum number of consecutive bonds of any type that you will allow the primer to form with itself (hairpin formation) or with another primer (dimer formation).
10. Use the **primer vs. primer (G-C)** drop-down menu to specify the maximum number of consecutive G-C bonds that you will allow the primer to form with itself (hairpin formation) or with another primer (dimer formation).
11. Use the **3'-end vs. 3'-end** drop-down menu to specify the maximum number of consecutive bonds that you will allow for the formation of “primer dimers.”

Primer dimers can lead to amplification of the primers alone instead of the required product.

12. Use the **3'-end vs. product** drop-down menu to specify the maximum number of consecutive bonds you will allow between the 3' end of a primer and the product it amplifies.

This is used to eliminate primers that may bind to alternate sites tightly enough to result in false priming and amplification of the wrong product(s).

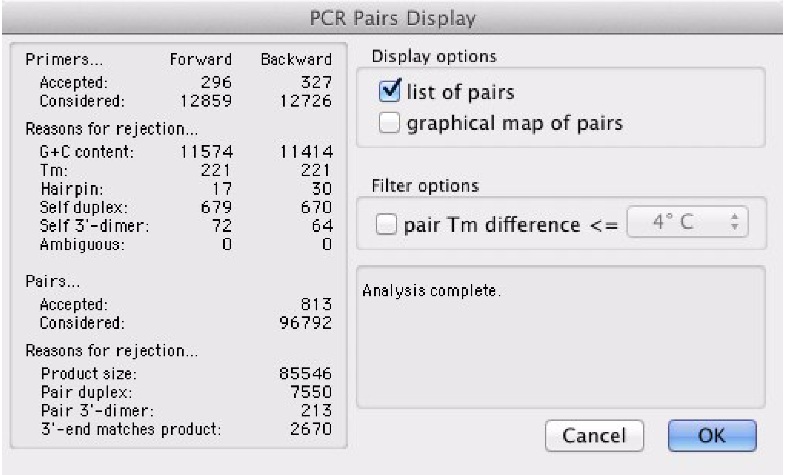
13. Type in the sum of the concentrations of the two primers at the start of the amplification reaction in the **total initial primer conc. (nM)** text box.
14. Type in the monovalent cation (Na^+ and K^+) concentration of the reaction mixture in the **monovalent cation conc. (mM)** text box.
15. Type in the divalent cation (Mg^{++} and Mn^{++}) concentration of the reaction mixture in the **divalent cation conc. (mM)** text box.
16. Type in the deoxyribonucleotide triphosphate concentration for the reaction mixture in the **dNTP conc. (mM)** text box.

Note. These **Reactions conditions** parameters are shared between all the Primers tools, so making a change to any of the values here will affect the values displayed and used by Primer3, Sequenceing Primers/Probes, etc.

17. Select **OK** to perform the screen.

Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the screen.

At the end of the screen, the **PCR Pairs Display** dialog box is displayed.



PCR Pairs Display

```
Primers...              Forward    Backward
  Accepted:                 296         327
  Considered:             12859       12726
Reasons for rejection...
  G+C content:            11574       11414
  Tm:                       221         221
  Hairpin:                   17          30
  Self duplex:              679         670
  Self 3'-dimer:             72          64
  Ambiguous:                  0           0

Pairs...
  Accepted:                             813
  Considered:                         96792
Reasons for rejection...
  Product size:                       85546
  Pair duplex:                         7550
  Pair 3'-dimer:                        213
  3'-end matches product:              2670
```

Display options

☑ list of pairs

□ graphical map of pairs

Filter options

□ pair Tm difference <= 4° C

Analysis complete.

The **PCR Pairs Display** dialog box contains statistical information about the PCR primer pair screening: how many were considered, how many eliminated, and the reasons for rejection. If no pairs were found, look at these statistics to determine the most common cause of rejection. This can help you decide which setup parameters to modify.

Displaying PCR primer screen results

The **PCR Pairs Display** dialog box provides a number of options for displaying the results of the screen.

To display the results of a PCR primer pairs screen

1. The **PCR Pairs Display** dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example to change the display parameters, choose **Analyze | Primers | PCR Primer Pairs** when any PCR pairs display result window is active.
2. Select the **pair Tm difference** checkbox to restrict the displays to those primer pairs whose Tm's differ by less than the amount chosen in the associated drop-down menu.
3. Select the **list of pairs** checkbox to see a list of the primer pairs, and information about them and the products they amplify: Tm, G+C content, length, location, and the optimum annealing temperature.
4. Select the **graphical map of pairs** checkbox to see a graphical representation of the pairs and the products they amplify.

5. Select **OK** to display the results.

PCR primer pair characteristics

For any input pair of PCR primers, MacVector can generate useful information about the primers and the product. This enables you to assess primer pairs from sources other than the MacVector primer design functions. The calculated characteristics include:

- length and G-C percentage of each sequence
- average Tm, derived from the reverse complement of the primer
- primer pair Tm difference
- self-dimer and dimer formation
- hairpin formation
- self-duplex and duplex formation
- binding site information
- product formation.

To use this functionality, a DNA sequence must be the active window.

To calculate PCR primer pair characteristics

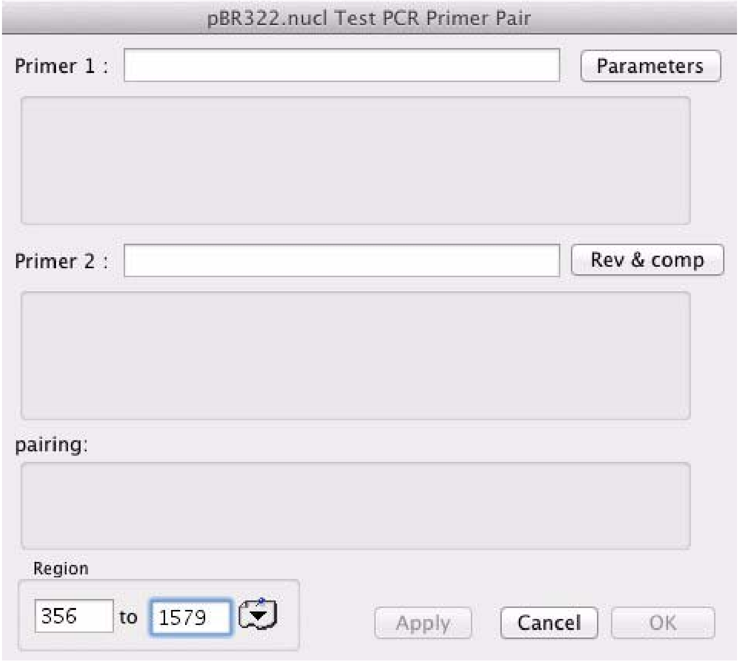
1. Choose **Analyze | Primers | Test PCR Primer Pai**r.

The **PCR Characteristics** dialog box is displayed.

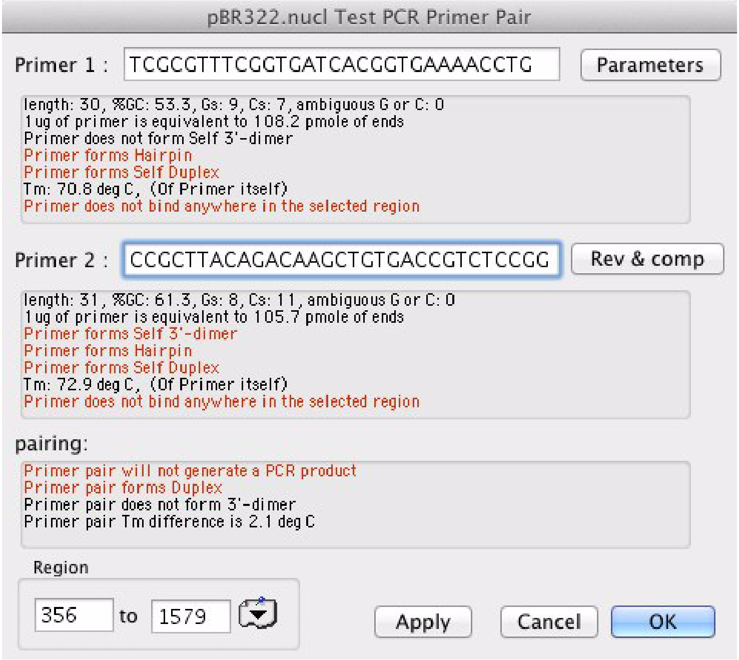
2. Enter the sequence of the first primer in the pair into the **Primer 1** text box, using the standard IUPAC one-letter codes. The sequence order must be 5’ to 3’.
3. Enter the sequence of the second primer in the pair into the **Primer 2** text box, using the standard IUPAC one-letter codes. The sequence order must be 5’ to 3’.

Tip. You might want to cut and paste text into both Primer 1 and Primer 2 boxes. You can do this by storing text in the Text Editor. You can then switch to the Text Editor to copy new text to the clipboard, before pasting into MacVector.

4. If required, select the **Parameters** button to adjust the parameters used to control the tests that generate the characteristics. See *“To modify the PCR primer test parameters”* on page 221, for further details.



5. You can restrict the search for unwanted binding sites to a certain region of the active sequence you are testing primers against, by typing in the base numbers that bracket the region in the **Region** text boxes, or by selecting a region from the features table drop-down menu at the right of the text boxes.
6. Do one of the following:
   - select **Apply** to run the analysis and display results in the information text boxes. This leaves the **PCR Characteristics** dialog box displayed



pBR322.nucl Test PCR Primer Pair

Primer 1 : TCGCGTTTCGGTGATCACGGTGAAAACCTG

```
length: 30, %GC: 53.3, Gs: 9, Cs: 7, ambiguous G or C: 0
1ug of primer is equivalent to 108.2 pmole of ends
Primer does not form Self 3'-dimer
Primer forms Hairpin
Primer forms Self Duplex
Tm: 70.8 deg C, (Of Primer itself)
Primer does not bind anywhere in the selected region
```

Primer 2 : CCGCTTACAGACAAGCTGTGACCGTCTCCGG

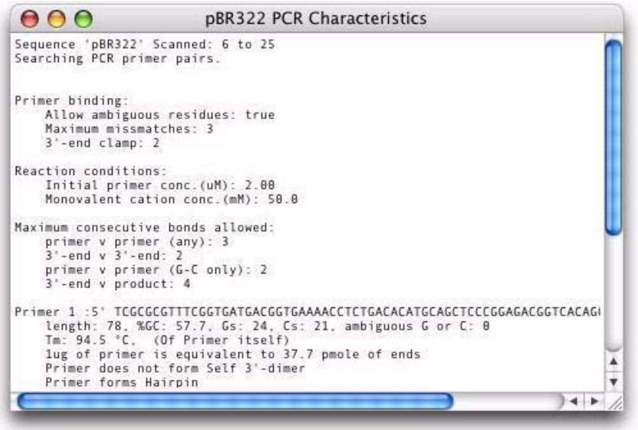
```
length: 31, %GC: 61.3, Gs: 8, Cs: 11, ambiguous G or C: 0
1ug of primer is equivalent to 105.7 pmole of ends
Primer forms Self 3'-dimer
Primer forms Hairpin
Primer forms Self Duplex
Tm: 72.9 deg C, (Of Primer itself)
Primer does not bind anywhere in the selected region
```

pairing:

```
Primer pair will not generate a PCR product
Primer pair forms Duplex
Primer pair does not form 3'-dimer
Primer pair Tm difference is 2.1 deg C
```

Region: 356 to 1579

- select **OK** to close the **PCR Characteristics** dialog box and display a text window of the same name that duplicates the results shown in the information text boxes. It also shows aligned sequences and product sequences.



pBR322 PCR Characteristics

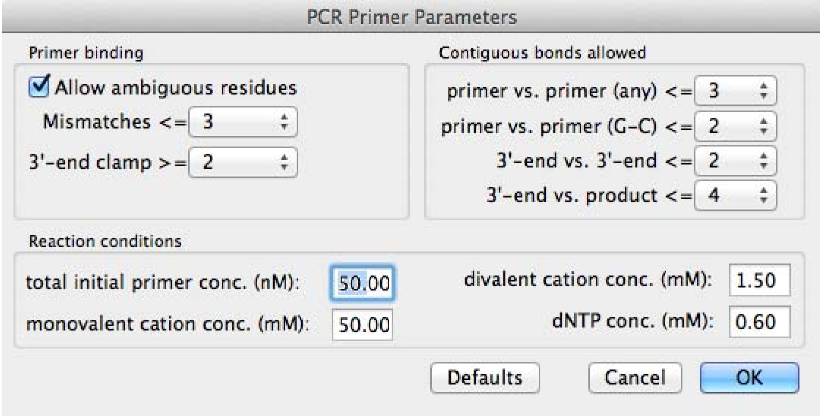
```
Sequence 'pBR322' Scanned: 6 to 25
Searching PCR primer pairs.


Primer binding:
    Allow ambiguous residues: true
    Maximum missmatches: 3
    3'-end clamp: 2

Reaction conditions:
    Initial primer conc.(uM): 2.00
    Monovalent cation conc.(mM): 50.0

Maximum consecutive bonds allowed:
    primer v primer (any): 3
    3'-end v 3'-end: 2
    primer v primer (G-C only): 2
    3'-end v product: 4

Primer 1 :5' TCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAG
    length: 78, %GC: 57.7, Gs: 24, Cs: 21, ambiguous G or C: 0
    Tm: 94.5 °C,  (Of Primer itself)
    1ug of primer is equivalent to 37.7 pmole of ends
    Primer does not form Self 3'-dimer
    Primer forms Hairpin
```

Note. If a text window of the correct name is already open, then the contents will simply be updated. Because the contents are only updated when the **OK** button is clicked, it is possible to display information in the text window that is out of step with the PCR Primer Pair dialog information boxes.

PCR primer test parameters

The characteristics are calculated from a series of tests, and you can modify the test parameters. The parameters are accessible from the PCR Characteristics dialog box, and include several parameters used on the Find PCR Primer Pairs dialog box (see *"Screening for PCR primer pairs"* on page 214).

Note. The shared parameters are linked, so changes in one dialog box will be reflected in the other.



☑ Allow ambiguous residues

Mismatches <= 3

3'-end clamp >= 2

primer vs. primer (any) <= 3

primer vs. primer (G-C) <= 2

3'-end vs. 3'-end <= 2

3'-end vs. product <= 4

total initial primer conc. (nM): 50.00

monovalent cation conc. (mM): 50.00

divalent cation conc. (mM): 1.50

dNTP conc. (mM): 0.60

To modify the PCR primer test parameters

1. Choose **Analyze | Primers | Test PCR Primer Pai**r.

The **PCR Characteristics** dialog box is displayed.

2. Select the **Parameters** button.

The **PCR Primer Parameters** dialog box is displayed.

Note. The **Primer binding** parameters are used to identify regions of the target sequence that the primer can potentially bind to.

3. Select **Allow Ambiguous residues** to let primer sequences contain characters other than A, C, T, and G, and to accept matches between ambiguous bases in the test sequence.

4. Use the **Mismatches** drop-down menu to specify the maximum number of residues by which the primer can differ from the target sequence.
5. Use the **3'-end clamp** drop-down menu to specify the number of contiguous residues at the 3' end that must bind with the target.

Note. The **Contiguous bonds allowed** parameters are used to screen the primer for self-annealing artifacts that might interfere with the reaction.

6. Use the **primer vs. primer (any)** drop-down menu to specify the maximum number of consecutive bonds of any type that you will allow the primer to form with itself (hairpin formation) or with another primer (dimer formation).
7. Use the **primer vs. primer (G-C)** drop-down menu to specify the maximum number of consecutive G-C bonds that you will allow the primer to form with itself (hairpin formation) or with another primer (dimer formation).
8. Use the **3'-end vs. 3- end** drop-down menu to specify the maximum number of consecutive bonds that you will allow for the formation of primer dimers.

Primer dimers can lead to amplification of the primers alone instead of the required product.

9. Use the **3'-end vs. product** drop-down menu to specify the maximum number of consecutive bonds you will allow between the 3' end of a primer and the product it amplifies.

This is used to check for primers that may bind to alternate sites tightly enough to result in false priming and amplification of the wrong product(s).

10. Type in the sum of the concentrations of the two primers at the start of the amplification reaction in the **total initial primer conc. (nM)** text box.
11. Type in the monovalent cation (Na^+ and K^+) concentration of the reaction mixture in the **monovalent cation conc. (mM)** text box.
12. Type in the divalent cation (Mg^{++} and Mn^{++}) concentration of the reaction mixture in the **divalent cation conc. (mM)** text box.
13. Type in the deoxyribonucleotide triphosphate concentration for the reaction mixture in the **dNTP conc. (mM)** text box.

Note. These **Reactions conditions** parameters are shared between all the Primers tools, so making a change to any of the values here will affect the values displayed and used by Primer3, Sequenceing Primers/Probes, etc.

Note. Steps 10 to 13 only affect the Tm calculation.

14. Select **OK** to return to the **PCR Characteristics** dialog box.

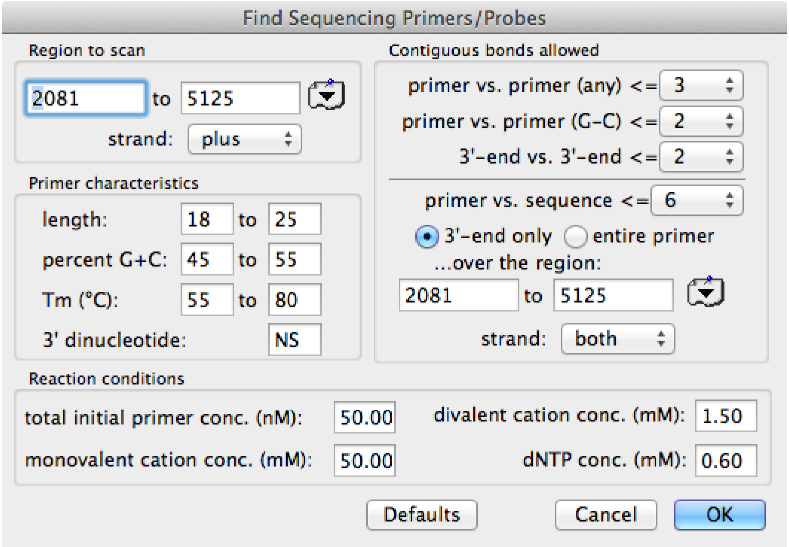
Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without saving the parameters.

Sequencing primers and hybridization probes

MacVector provides functionality for screening a nucleic acid sequence for likely sequencing primers or hybridization probes of up to 80 nucleotides.

Screening for sequencing primers or probes

To use this functionality, a nucleic acid sequence window must be the active window.



Find Sequencing Primers/Probes

Region to scan

2081 to 5125

strand: plus

Primer characteristics

length: 18 to 25

percent G+C: 45 to 55

Tm (°C): 55 to 80

3' dinucleotide: NS

Contiguous bonds allowed

primer vs. primer (any) <= 3

primer vs. primer (G-C) <= 2

3'-end vs. 3'-end <= 2

primer vs. sequence <= 6

● 3'-end only ○ entire primer

...over the region:

2081 to 5125

strand: both

Reaction conditions

total initial primer conc. (nM): 50.00

divalent cation conc. (mM): 1.50

monovalent cation conc. (mM): 50.00

dNTP conc. (mM): 0.60

To screen for sequencing primers or probes

1. Choose **Analyze | Primers | Sequencing Primers/Probes**.

The **Find Sequencing Primers/Probes** dialog box is displayed.

2. In the **Region to scan** panel, enter the base numbers that bracket the region you want to scan by typing them in the text boxes, or by selecting a region from the features table drop-down menu at the right of the boxes.
3. Choose the strand to scan from the **strand** drop-down menu.
4. Type the required range of primer/probe lengths in the **length** text boxes.
5. Type the required percentage range of primer/probe G+C content in the **percent G+C** text boxes.
6. Type the required range of primer/probe Tm values in the **Tm (°C)** text boxes.
7. Type the IUPAC code for the two nucleotides that you want to appear at the 3' end of the primer/probe in the **3' dinucleotide** text box.
8. Use the **primer vs. primer (any)** drop-down menu to specify the maximum number of consecutive bonds of any type that you will allow the primer/probe to form with itself (hairpin formation) or with another primer/probe molecule (dimer formation).
9. Use the **primer vs. primer (G-C)** drop-down menu to specify the maximum number of consecutive G-C bonds that you will allow the primer/probe to form with itself (hairpin formation) or with another primer/probe molecule (dimer formation).

Use the **3'-end vs. 3'-end** drop-down menu to specify the maximum number of consecutive bonds that you will allow for the formation of dimers.

10. This is important for sequencing primers; for hybridization probes, choose the largest value.
11. Use the **primer vs. sequence** drop-down menu to specify the maximum number of consecutive bonds you will allow between the primer/probe secondary binding sites and the target sequence. There are two options associated with this:
    - for a sequencing primer, you would usually select the radio button that specifies **3'-end only.**
    - for a hybridization probe, you would usually select the radio button that specifies the **entire primer**.

12. In the ...**over the region** text boxes, type the base numbers that bracket a comparison region, or select a region from the features table drop-down menu to the right of the boxes.

This is the region of the sequence that you want to compare with the potential primer/probe to eliminate those that bind to alternate sites.

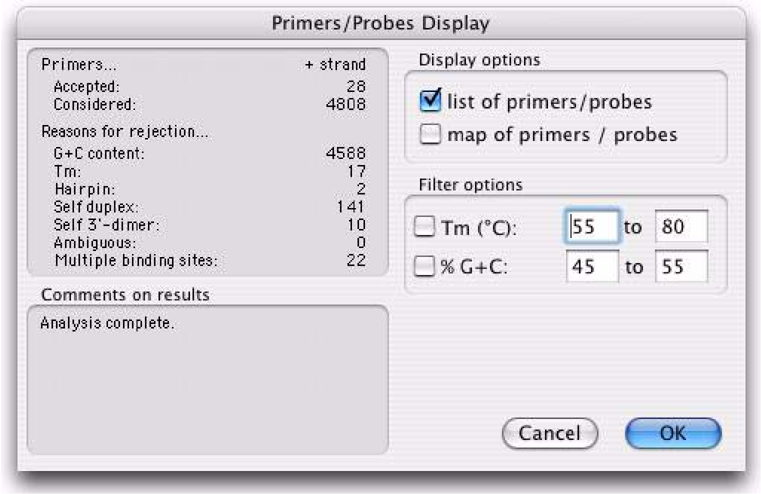
13. Choose the strand to use in the comparison from the **strand** drop-down menu.
14. Type in the sum of the concentrations of the two primers at the start of the amplification reaction in the **total initial primer conc. (nM)** text box.
15. Type in the monovalent cation (Na^+ and K^+) concentration of the reaction mixture in the **monovalent cation conc. (mM)** text box.
16. Type in the divalent cation (Mg^{++} and Mn^{++}) concentration of the reaction mixture in the **divalent cation conc. (mM)** text box.
17. Type in the deoxyribonucleotide triphosphate concentration for the reaction mixture in the **dNTP conc. (mM)** text box.

Note. These **Reactions conditions** parameters are shared between all the Primers tools, so making a change to any of the values here will affect the values displayed and used by Primer3, PCR Primer Pairs, etc.

18. Select **OK** to perform the screen.

Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the screen.

At the end of the screen, the **Primers/Probes Display** dialog box is displayed.



Primers...	+ strand
Accepted:	28
Considered:	4808
Reasons for rejection...	
G+C content:	4588
Tm:	17
Hairpin:	2
Self duplex:	141
Self 3'-dimer:	10
Ambiguous:	0
Multiple binding sites:	22

Comments on results

Analysis complete.

Display options

☑ list of primers/probes

□ map of primers / probes

Filter options

□ Tm (°C): 55 to 80

□ % G+C: 45 to 55

The **Primers/Probes Display** dialog box contains statistical information about the scan for primers/probes: how many were considered, how many eliminated and the reasons for rejection. If none were found, look at these statistics to determine the most common cause of rejection. This can help you decide which setup parameters to modify.

Displaying sequencing primer or probe screen results

The **Primers/Probes Display** dialog box provides a number of options for displaying the results of the screen.

To display the results of a sequencing primer or probe screen

1. The **Primers/Probes Display** dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example to change the display parameters, choose **Analyze | Primers | Sequencing Primers/Probes** when any Primers/Probes display result window is active.
2. Select the **Tm (°C)** checkbox to restrict the displays to those primers/ probes whose Tm’s lie within the range specified in the text boxes.
3. Select the **percent G+C** checkbox to restrict the displays to those primers/probes whose G+C percentage lies within the range specified in the text boxes.

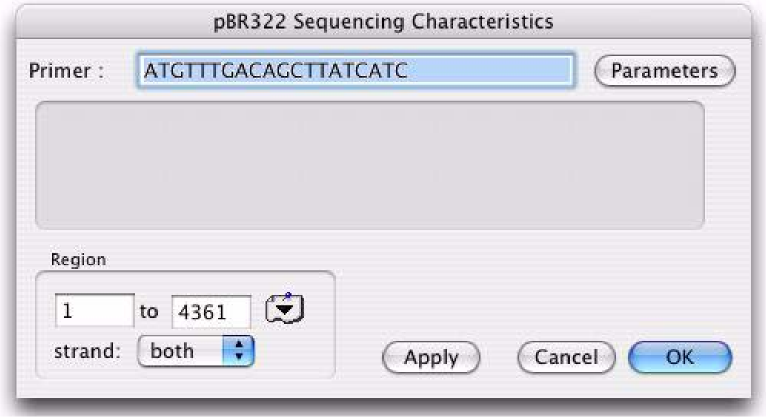
4. Select the **list of primers/probes** checkbox to see a list of the primers or probes and information about them: Tm, G+C content, length, and location.
5. Select the **map of primers/probes** checkbox to see a graphical representation of the sequencing primers or hybridization probes.
6. Select **OK** to display the results.

Sequencing primer characteristics

For any sequencing primer, MacVector can generate useful information about its properties and binding characteristics. This enables you to assess sequencing primers from sources other than the MacVector screening function. The calculated characteristics include:

- length and G-C percentage of each sequence
- average Tm, derived from the reverse complement of the primer
- self-dimer formation
- hairpin formation
- self-duplex formation
- binding site information.

To use this option, a DNA sequence must be the active window.



pBR322 Sequencing Characteristics

Primer : ATGTTTGACAGCTTATCATC

Region: 1 to 4361

strand: both

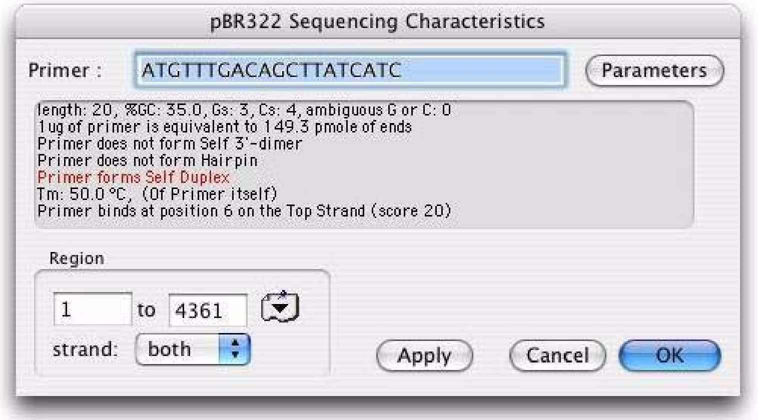
To calculate sequencing primer characteristics

1. Choose **Analyze | Primers | Test Sequencing Primers/Probes**.

The **Sequencing Characteristics** dialog box is displayed.

2. Enter the primer sequence into the **Primer** text box, using the standard IUPAC one-letter codes. The sequence order must be 5’ to 3’.
3. If required, select the **Parameters** button to adjust the parameters used to control the tests that generate the characteristics. See *“To modify the sequencing primer test parameters”* on page 230, for further details.
4. You can restrict the search for unwanted binding sites to a certain region of the active sequence you are testing primers against, by typing in the base numbers that bracket the region in the **Region** text boxes, or by selecting a region from the feature selector drop-down menu to the right of the text boxes.
5. Do one of the following:

- select **Apply** to run the analysis and display results in the information text box. This leaves the **Sequencing Characteristics** dialog box displayed.



pBR322 Sequencing Characteristics

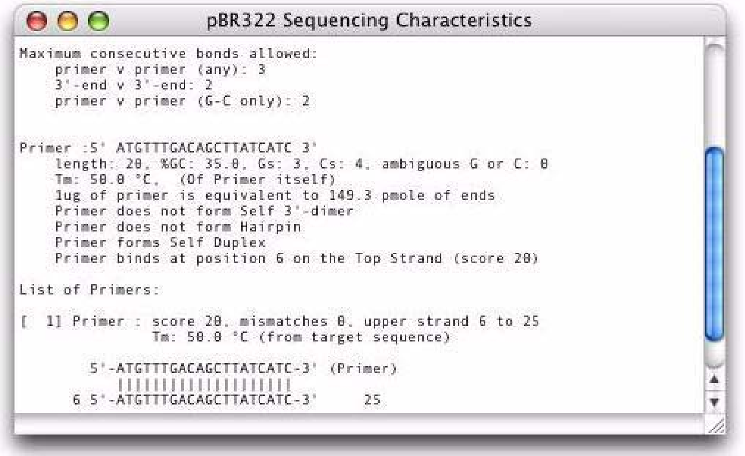
Primer : ATGTTTGACAGCTTATCATC

```
length: 20, %GC: 35.0, Gs: 3, Cs: 4, ambiguous G or C: 0
1ug of primer is equivalent to 149.3 pmole of ends
Primer does not form Self 3'-dimer
Primer does not form Hairpin
Primer forms Self Duplex
Tm: 50.0 °C, (Of Primer itself)
Primer binds at position 6 on the Top Strand (score 20)
```

Region: 1 to 4361

strand: both

- select **OK** to close the **Sequencing Characteristics** dialog box and display a text window of the same name that duplicates the results shown in the information text box. It also shows aligned sequences and product sequences.



pBR322 Sequencing Characteristics

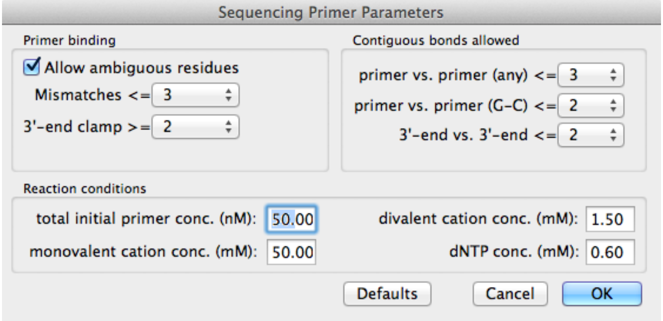
```
Maximum consecutive bonds allowed:
    primer v primer (any): 3
    3'-end v 3'-end: 2
    primer v primer (G-C only): 2


Primer :5' ATGTTTGACAGCTTATCATC 3'
    length: 20, %GC: 35.0, Gs: 3, Cs: 4, ambiguous G or C: 0
    Tm: 50.0 °C,  (Of Primer itself)
    1ug of primer is equivalent to 149.3 pmole of ends
    Primer does not form Self 3'-dimer
    Primer does not form Hairpin
    Primer forms Self Duplex
    Primer binds at position 6 on the Top Strand (score 20)

List of Primers:

[  1] Primer : score 20, mismatches 0, upper strand 6 to 25
               Tm: 50.0 °C (from target sequence)

        5'-ATGTTTGACAGCTTATCATC-3' (Primer)
           ||||||||||||||||||||
      6 5'-ATGTTTGACAGCTTATCATC-3'     25
```

Note. If a text window of the correct name is already open, then the contents will simply be updated. Because the contents are only updated when the **OK** button is clicked, it is possible to display information in the text window that is out of step with the **PCR Primer Pair** dialog boxes.

Sequencing primer test parameters

The characteristics are calculated from a series of tests, and you can modify the test parameters. The parameters are accessible from the **Sequencing Characteristics** dialog box, and include several parameters used on the **Find Sequencing Primers/Probes** dialog box (see *"Screening for sequencing primers or probes"* on page 223).

Note. The shared parameters are linked, so changes in one dialog box will be reflected in the other.



Sequencing Primer Parameters

Primer binding

☑ Allow ambiguous residues

Mismatches <= 3

3'-end clamp >= 2

Contiguous bonds allowed

primer vs. primer (any) <= 3

primer vs. primer (G-C) <= 2

3'-end vs. 3'-end <= 2

Reaction conditions

total initial primer conc. (nM): 50.00

divalent cation conc. (mM): 1.50

monovalent cation conc. (mM): 50.00

dNTP conc. (mM): 0.60

To modify the sequencing primer test parameters

1. Choose **Analyze | Primers | Test Sequencing Primer/Probe**.

The **Sequencing Characteristics** dialog box is displayed.

2. Select the **Parameters** button.

The **Sequencing Primer Parameters** dialog box is displayed.

3. Select **Allow Ambiguous residues** to let primer sequences contain characters other than A, C, T, and G.
4. Use the **Mismatches** drop-down menu to specify the maximum number of residues by which the primer can diverge from the target sequence.
5. Use the **3'-end clamp** drop-down menu to specify the number of residues at the 3' end that must bind with the target.

This is used to reduce the number of potential binding sites.

6. Use the **primer vs. primer (any)** drop-down menu to specify the maximum number of consecutive bonds of any type that you will allow the primer to form with itself (hairpin formation) or with another primer (dimer formation).
7. Use the **primer vs. primer (G-C)** drop-down menu to specify the maximum number of consecutive G-C bonds that you will allow the primer to form with itself (hairpin formation) or with another primer (dimer formation).
8. Use the **3'-end vs. 3'-end** drop-down menu to specify the maximum number of consecutive bonds that you will allow for the formation of “primer dimers.”

Primer dimers can lead to amplification of the primers alone instead of the required product.

9. Type in the sum of the concentrations of the two primers at the start of the amplification reaction in the **total initial primer conc. (nM)** text box.
10. Type in the monovalent cation (Na^+ and K^+) concentration of the reaction mixture in the **monovalent cation conc. (mM)** text box.
11. Type in the divalent cation (Mg^{++} and Mn^{++}) concentration of the reaction mixture in the **divalent cation conc. (mM)** text box.
12. Type in the deoxyribonucleotide triphosphate concentration for the reaction mixture in the **dNTP conc. (mM)** text box.

Note. These **Reactions conditions** parameters are shared between all the Primers tools, so making a change to any of the values here will affect the values displayed and used by Primer3, PCR Primer Pairs, etc.

Note. Steps 9 to 12 only affect the Tm calculation.

13. Select **OK** to return to the **Sequencing Characteristics** dialog box.

Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without saving the parameters.

Least degenerate hybridization probes

MacVector provides a way to find an oligo to use as a hybridization probe for a gene whose exact DNA sequence is not known, provided the amino acid sequence is known. The process used is as follows:

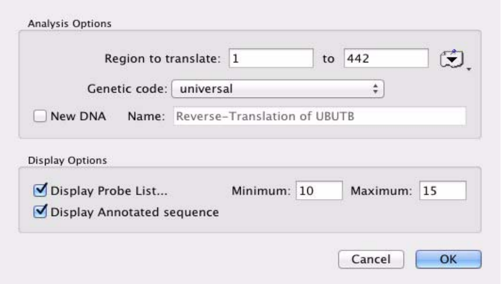
- reverse-translate the amino acid sequence;
- scan the resulting DNA sequence to find a region that is not very degenerate;
- use this region to make probes, thus minimizing the number of oligonucleotide probes that would have to be synthesized.

Refer to *"Screening a protein to find a hybridization probe"* on page 439, for more details of this process.

Screening for least degenerate hybridization probes

To use this functionality, a protein sequence window must be the active window.

Before you run the analysis, you may want to reduce the degeneracy of the resultant DNA sequence by eliminating certain codons from the genetic code used for the reverse translation. Your knowledge of the codon preferences of the organism of interest will be your best guide in eliminating codons. Refer to *"Reducing the degeneracy of a genetic code"* on page 245 for details of this procedure.



Analysis Options

Region to translate: 1 to 442

Genetic code: universal

☐ New DNA Name: Reverse-Translation of UBUTB

Display Options

☑ Display Probe List... Minimum: 10 Maximum: 15

☑ Display Annotated sequence

To find least degenerate hybridization probes

1. Choose **Analyze | Reverse Translation**.
2. Type the residue numbers that bracket the region you want to scan in the **Region** text boxes, or select a region from the features table drop-down menu at the right of the boxes.
3. Select the genetic code to be used from the **genetic code** drop-down menu.
4. Select the **Probe list** check box to generate a list of least-ambiguous oligonucleotide probes for the specified region.

The output lists the sequence of each probe, the corresponding amino acid sequence, the percent G+C content of the probe, and the dissociation temperature of the probe-DNA complex.

5. If you are generating a probe list, limit the range of probe sizes by entering values in the **minimum** and **maximum** text boxes.
6. Select the **Annotated sequence** check box to display the degenerate nucleic acid sequence that results from reverse translating the protein. The appearance of this display can be changed interactively by choosing **Customize | Format Annotated Display.** Refer to *"Formatting the Aligned Sequence view"* on page 41, for further details.
7. Select the **New DNA** check box to create a new DNA sequence window for the degenerate sequence created by reverse translating the protein sequence. Type a name for the new sequence in the text box.
8. Select **OK** to perform the screen.

Screen results

If you have displayed an annotated sequence, the amino acid that results from translation is shown beneath the degenerate nucleic acid sequence. Note that this amino acid sequence may not match the original protein because certain degenerate codons could code for more than one amino acid. For example, serine has six codons, four of the form TCN and two of the form AGY. These two forms reduce to the single degenerate codon WSN. If this degenerate codon is then retranslated, it will be assigned the "unknown" amino acid X, because WSN can expand to any of the following:

- TCN = Ser
- ACN = Thr
- AGY = Ser

- AGR = Arg
- TGY = Cys
- TGA = End
- TGG = Trp.

11 Using Transcription and Translation Functions

Overview

This chapter describes how to transcribe and translate DNA sequences to protein.You can assemble the required sequence for transcription by choosing particular features such as introns and exons. MacVector provides flexible facilities for translating and reverse translating sequence data, according to the chosen genetic code. In addition to using the genetic codes supplied with MacVector, you can also create your own codes.

The reverse translation functionality is described in *“Least degenerate hybridization probes”* on page 231

Contents

Transcribing and translating DNA to protein

MacVector enables you to transcribe a DNA sequence, and translate the resulting mRNA into an amino acid sequence using a selected genetic code. MacVector can also display a codon usage table for the translation.

Using the **Translation Analysis** dialog box, you can transcribe and translate a DNA sequence in a single operation. Alternatively, using the **Generate Transcript** dialog box, you can first perform a transcription analysis and then translate. This option is particularly useful if you require the mRNA sequence itself, or if you want to assemble the transcription sequence from discontinuous segments of DNA.

Both the **Transcription Analysis** and the **Generate Transcript** dialog boxes allow you to specify the segments to be transcribed by choosing exon, intron, RNA or coding sequence (CDS) features.

All segments are assumed to be on the same strand, so you cannot merge segments from the plus and minus strands in the same transcription. To run a transcription or translation analysis, a nucleic acid sequence window must be the active window.

Generating a transcript

An entire sequence of DNA in the sequence window can be transcribed automatically by clicking on the sequence type indicator icon on the toolbar. The indicator icon will change from **DNA** to **RNA**. Click on the icon again to reverse the transcription.

Alternatively, to transcribe one or more features in the sequence, use the **Generate Transcript** dialog box.

Note. MacVector detects features by their feature table entries. It does not predict exons, so you must first ensure that the appropriate features have been defined.

Specifying a CDS feature

This is the initial default method. When you choose the **Single CDS** method and select a range of the current DNA sequence, MacVector lists all CDSs found within the range. The CDS features for both strands are listed together; those on the minus (complementary) strand are marked with a blue ‘C’. Each CDS may contain several exons; these are listed on successive lines within the CDS feature entry. Only one CDS feature can be selected for transcription, because only a single CDS can contribute to a mRNA molecule.

Note. A CDS feature may or may not include the terminal stop codon, depending on the preference of the submitting author.

Specifying exons

When you choose **Specify exons**, and select a strand and a range of the current DNA sequence, MacVector lists all exons found on that strand within the range. Initially all exons are checked, so they will be included in the transcription. However, you can deselect exons to prevent them from being transcribed. This is useful where a gene has alternate splice sites.

The default assembly order is from 5’ to 3’ on the selected strand. Therefore the exons are listed in ascending order of residue numbering on the plus strand, and in descending order on the minus strand. You can change the order of assembly by moving items in the list.

Tip. Rearranging the order of exons may be useful if the gene has been cloned into circular DNA, and the origin of the circular DNA is inside the gene.

Specifying introns

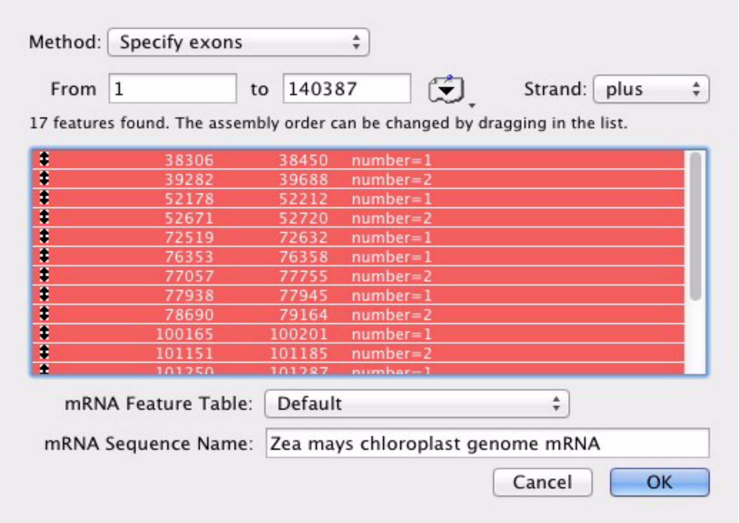
When you choose the **Specify introns** method, and select a strand and a range of the current DNA sequence, MacVector lists all introns found on that strand within the range. In contrast to MacVector’s treatment of exons, all introns are initially *un*checked, so they will be excluded from the transcription. If you select any introns, they will be transcribed.

Specifying RNA features

When you choose the **Specify RNA features** method, and select a strand and a range of the current DNA sequence, MacVector lists all RNA features found on that strand within the range. Initially all features are *un*checked, so they will be excluded from the transcription. If you select any RNA features, they will be transcribed. You can also change the order of assembly, by moving items in the list.

Preserving features

When the mRNA window is created, the relevant features from the DNA sequence’s feature table can be copied into it. Where a feature lies partially within a selected range, it will be split and copied as a feature fragment. If several segments have been transcribed, there can be many such feature fragments in the resulting mRNA sequence. MacVector lets you choose how much feature information to include in the transcribed sequence: all the features, just the features that define the selected segments, or no features at all. The default is to include only the defining features. If no features are copied, the mRNA window will contain only the residues.



Method: Specify exons

From 1 to 140387 Strand: plus

17 features found. The assembly order can be changed by dragging in the list.

↕	38306	38450	number=1
↕	39282	39688	number=2
↕	52178	52212	number=1
↕	52671	52720	number=2
↕	72519	72632	number=1
↕	76353	76358	number=1
↕	77057	77755	number=2
↕	77938	77945	number=1
↕	78690	79164	number=2
↕	100165	100201	number=1
↕	101151	101185	number=2
↕	101250	101287	number=1

mRNA Feature Table: Default

mRNA Sequence Name: Zea mays chloroplast genome mRNA

To transcribe DNA features into mRNA

1. Choose **Analyze | Generate Transcript** from the menu.

The **Generate Transcript** dialog box is displayed.

2. From the **Method** drop-down menu, choose one of the following methods for selecting segments to transcribe:

- **Single CDS**
- **Specify exons**
- **Specify introns**
- **Specify RNA sequences**

The appropriate features are displayed in a list.

3. To enter a range in the sequence, type the appropriate sequence numbers in the **From** and **to** text boxes.

You will be selecting features for transcription within this range.

4. From the **Strand** menu, select **plus** or **minus**.

If you have selected the **Single CDS** method, CDS features of both strands are listed and this menu is not available.

5. Do one of the following, depending on the transcription method you chose in step 2:
    - **Specify exons**: Deselect any listed features that you do not want transcribed by clicking in their check boxes.
    - **Single CDS**: Select a CDS feature for transcription from the list. The selected feature is highlighted.
    - **Specify introns** or **Specify RNA sequences**: Select any listed introns that you want transcribed by clicking in their check boxes.
6. If you chose **Specify exons** or **Specify RNA features**, you can change the order of assembly of the listed features. To move a feature, click on the double arrow at the left of the line. Holding down the mouse button, drag the feature to its new position in the list, and then release the mouse button.

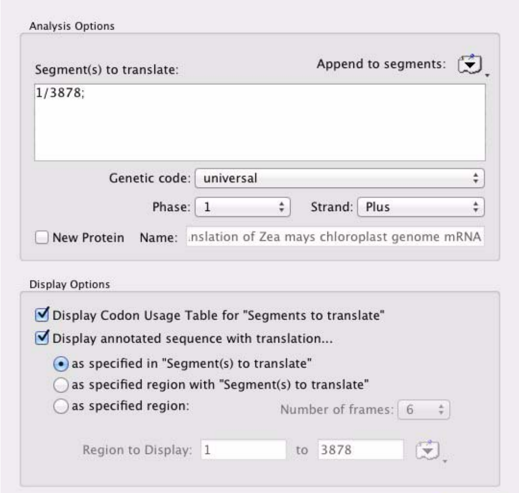
As the line is dragged, a dashed outline follows the cursor, and an arrow at the left of the list shows where the insertion position will be if the mouse is released.

7. From the **mRNA Feature Table** drop-down menu, choose which features should be copied into the transcribed sequence: **Copy all features**, **Default**, or **Do not copy features**.
8. Type a name for the transcribed sequence in the **mRNA Sequence Name** text box.
9. Select **OK** to perform the transcription and display the resulting mRNA sequence.

Translating DNA or mRNA to protein

The **Translation Analysis** dialog box is used to control the translation of a nucleic acid sequence and to set up displays of the resulting protein sequence. The nucleic acid sequence can be either DNA, or the mRNA output from a transcription analysis (see *"To transcribe DNA features into mRNA"* on page 238). If you are using a DNA sequence, select one or more segments for translation, using the **Segment(s) to translate** text box or the feature selector drop-down menu. If you have a transcribed mRNA sequence, you will normally translate the entire sequence.

For instructions on choosing and modifying genetic codes, see *"Genetic codes"* on page 242.



Analysis Options

Segment(s) to translate: Append to segments:

1/3878;

Genetic code: universal

Phase: 1 Strand: Plus

□ New Protein Name: .nslation of Zea mays chloroplast genome mRNA

Display Options

☑ Display Codon Usage Table for "Segments to translate"

☑ Display annotated sequence with translation...

● as specified in "Segment(s) to translate"

○ as specified region with "Segment(s) to translate"

○ as specified region: Number of frames: 6

Region to Display: 1 to 3878

To translate a nucleic acid sequence to protein

1. Choose **Analyze | Translation** from the menu.

The **Translation Analysis** dialog box is displayed.

2. If you do not want to translate the entire sequence, select the parts of the sequence to translate, by doing one of the following:

- type the required ranges in the **Segment(s) to translate** text box
- select the required features from the feature selector drop-down menu to the right of the box.

When you are typing ranges, there must be a slash (/) between the numbers of a range, and a semicolon (;) after each specified range. For segments on the minus (reverse complement) strand, type the base numbers with reference to the plus strand.

Note. If two or more segments are specified, they must be listed in the order that they are to be translated. If several segments on the minus strand are specified, the first segment to be translated will be the segment with the highest number.

3. Choose the genetic code to use for the translation from the **genetic code** drop-down menu.

For instructions on choosing and modifying genetic codes, see *“Genetic codes”* on page 242.

4. Choose the reading frame from the **phase** drop-down menu.

The reading frame is relative to the start of the first segment entered in the **Segment(s) to translate** box. In almost all cases, the phase should be set to 1.

5. Choose which strand to translate from the **strand** drop-down menu. If regions are entered from the features table, the strand is set automatically to the strand corresponding to the last feature that was used.
6. To create a new protein sequence window containing the protein sequence specified by the translation settings, select the **Create new protein** check box. Type a name for the protein sequence in the text box.

Note. The protein sequence will only include the translations of the specified segments.

7. To generate a table displaying the frequency of each of the 64 codons in the selected sequence, select the **Codon Usage Table** check box.
8. Select the **Annotated sequence with translation** check box to generate an annotated sequence of the DNA and translated protein. This always displays the portion of the nucleic acid sequence specified in the **Region to Display** text boxes. The displayed sequence can be further controlled as follows:
   - select the **as specified in...** radio button to see the translation only for those segments specified in the **Segment(s) to translate** box
   - select the **of entire display region** radio button to see the translation of the whole display region. With this option, you can choose the number of reading frames translated from the **no. of frames** drop-down menu.
9. You can restrict the display to a certain region of the sequence by typing in the base numbers that bracket the region in the **Region to Display** text boxes, or by selecting a region from the feature selector drop-down menu to the right of the text boxes.

This enables you to display more than just the translated portion of the sequence, for example, if you want to show regulatory regions upstream from the coding region.

Note. The annotated sequence will show all bases in the specified region, not just those which have been translated.

10. Select **OK** to perform the translation and display the requested results.

Genetic codes

Genetic codes are used for translating and reverse translating sequence data. MacVector enables you to:

- designate a default genetic code for performing auto-translations
- create a new genetic code
- “delete” codons from a genetic code to reduce the amount of degeneracy when reverse translating an amino acid sequence to a DNA sequence.

Autotranslations are controlled by the annotated sequence formatting. When enabled, they display selected features with a translation beneath the main sequence display.

Refer to *“Formatting the Aligned Sequence view”* on page 41, for further information.

You can display the codons for the currently selected genetic code in a popup reference table at any time using the **Window | Genetic Code Key** menu option.

Selecting a different genetic code

The default genetic code can be changed at any time. If an annotated sequence window is open at the time you make the change, any autotranslations that are present will be updated immediately to reflect the new genetic code settings. However, the results of previous analyses will not be changed. For example, if you translate a DNA sequence and create a new protein sequence window, the amino acids in the protein sequence will not change if you subsequently change the genetic code.

To select a genetic code

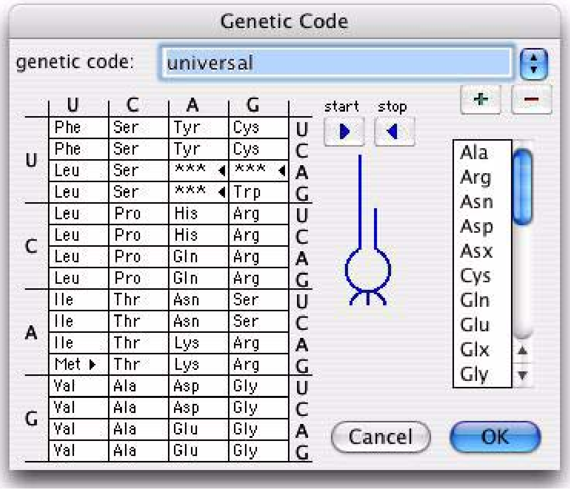
1. Choose **Options | Modify Genetic Codes** from the menu.

The **Genetic Code** dialog box is displayed.

2. Choose a code from the **genetic code** drop-down menu.
3. Select **OK** to remove the Genetic Code dialog box.

Modifying genetic codes

Creating and editing codes is done using the **Genetic Code** dialog box.



Genetic Code

genetic code: universal

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	*** ◂	*** ◂	A
	Leu	Ser	*** ◂	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met ▸	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

At the left side of the dialog box is a chart with 64 cells representing the 64 codons. The labels on the left side of the chart are for the first nucleotide of the codon, the labels on top for the second nucleotide, and the labels on the right for the third nucleotide. Each cell of the chart contains the three-letter abbreviation of the amino acid coded for by the codon controlling that cell. To the right of the chart is a graphic that looks like a stylized tRNA anticodon loop. When you click on a cell, the codon corresponding to the selected cell will be displayed beneath the loop.

Adding a new genetic code

New genetic codes can be added at any time.

To add a new genetic code

1. Choose **Options | Modify Genetic Codes**.

The **Genetic Code** dialog box is displayed.

2. Choose the code that will be used as a basis for the new one from the **genetic code** drop-down menu.
3. Type the name of the new genetic code in place of the selected code, then select the button labeled with a plus sign (+) in the upper right corner of the dialog box

Note. The new code will not be saved permanently until you select **OK**.

4. Modify the code as required.

Refer to the procedure *"To modify a genetic code"* on page 244.

5. Select **OK** to save the changes.

Note. The new code will become the default code, unless you select an alternative one before closing the **Genetic Code** dialog box.

Modifying a genetic code

The selected genetic code can be modified at any time. We recommend that any modifications to codes supplied with MacVector are done on a copy of the code. See *"To add a new genetic code"* on page 243, for details of how to copy a code.

To modify a genetic code

1. Choose **Options | Modify Genetic Codes**.

The **Genetic Code** dialog box is displayed.

2. Choose the code to be modified from the **genetic code** drop-down menu.
3. Select a cell whose assignment you want to change.
4. Scroll down the amino acid list on the right of the dialog box, and double-click on the amino acid you want to be coded for in the highlighted cell. If the codon does not code for an amino acid (for example, TAA, TAG, and TGA in the universal code) choose *** instead of an amino acid name.

The contents of the highlighted cell change to reflect the new assignment.

5. To designate any of the 64 codons to be a start or stop codon, select the appropriate cell, then select the **start** or **stop** button above the stylized tRNA anticodon loop.
6. Repeat steps 3 through 5 for each change you require.
7. Select **OK** to save the changes.

Deleting a genetic code

A genetic code can be deleted at any time. This should be done with caution, because the only way you can restore it is to re-enter it by hand or to reinstall MacVector.

To delete a genetic code

1. Choose **Options | Modify Genetic Codes**.

The **Genetic Code** dialog box is displayed.

2. Choose the code to be deleted from the **genetic code** drop-down menu.
3. Select the button labeled with a minus sign (-) in the upper right corner of the dialog box.
4. Select **OK** to delete the code permanently.

Note. If you select **Cancel**, the code will not be lost, even if you have done step 3 above.

Reducing the degeneracy of a genetic code

When you perform a reverse translation, you can eliminate codons from consideration to reduce the degeneracy of the DNA sequence that will result from the reverse translation. This can be useful, for example, when you know that your organism uses only two of the six standard serine codons in the universal genetic code.

To reduce the degeneracy of a genetic code

1. Choose **Options | Modify Genetic Codes**.

The **Genetic Code** dialog box is displayed.

2. Choose the code to be modified from the **genetic code** drop-down menu.
3. Select a codon that you want to ignore.
4. Scroll to the end of the amino acid list on the right of the dialog box, and select “- - -”.
5. Repeat steps 3 and 4 for each codon you want to ignore in the reverse translation.
6. Select **OK** to save the changes.

Each codon marked as “- - -” will not be considered when performing reverse translations.

12 Aligning Sequences

Overview

MacVector enables you to align both protein and DNA sequences from the following sources:

- sequences stored locally in any format recognized by MacVector
- sequences in the online NCBI databases accessible *via* the Internet.

Several methods are available to perform these alignments and each is outlined in this chapter:

- Multiple Sequence Alignment
- Align to Reference
- Pustell Matrix Analysis (also known as a Dot Plot)
- Internet BLAST
- Align to Folder
- Contig Assembly

Contents

Multiple Sequence Alignment

If you have two or more related sequences (DNA or Protein) and you want to examine the relationship between them, use this function.

Multiple sequence alignment allows you to align two or more sequences without requiring a reference. Alignments can be generated using the ClustalW, T-Coffee and Muscle algorithms. The generated alignments can be edited and displayed in many ways and submitted to additional phylogenetic reconstruction algorithms to help determine the relationship between sequences.

Choose **File | New | Protein Alignment** (or **File | New | Nucleic Acid Alignment**) to create an empty Multiple Sequence Alignment window.

Add sequences to the alignment by using **Edit | Add Sequences** from the main menu then click on the **Align** icon on the toolbar to align the sequences. Initially, the ClustalW algorithm is invoked by default, but subsequently the last algorithm that was used becomes the default. Click and hold the **Align** icon to choose a different alignment algorithm.

Click on the **Prefs** icon on the toolbar to control the appearance and behavior of the data in each of tabs that represent diffesrent views or analyses of the alignment.

This functionality is most suited for protein alignments, or for nucleic acid sequences where you are interested in examining phylogenetic relationships. If you wish to compare two or more DNA sequences, you should definitely consider if one of the other alignment functions may be more suitable.

To find out more see *“Performing Multiple Sequence Alignment”* on page 341.

Align to Reference

Use this tool if you have a reference sequence and you want to align one or more DNA sequences against it. A typical example would be in resequencing, e.g. sequencing a cloned PCR fragment to check no errors were introduced, sequencing across end junctions, scanning for successful mutagenesis clones, etc.

In each case, open the file that represents the parent or reference sequence, then choose **Analyze | Align to Reference**. In the Align to Reference window that opens, click on the **Add Seqs** icon on the toolbar to add sequences from disk – these can be in any format that MacVector can read, typically ABI or SCF chromatogram files but you can add plain sequences as well. Click the **Align** icon on the toolbar and choose the Sequence Confirmation algorithm, which is tuned to expect the small insertions/deletions you would expect in raw chromatogram files.

Compared to multiple sequence alignment, Align to Reference has the advantage that it will automatically "flip" sequences to guarantee optimal alignment.

Align to Reference can also be used to align cDNA clones against a genome sequence. The steps are similar: use the genomic sequence as the reference, then add one or more cDNA clones to the alignment. Again, these can be chromatogram files. Now choose the cDNA Alignment algorithm when you Align. This is tuned to expect large insertions representing the intronic regions.

See Chapter 16, *"Aligning Sequences to a Reference"* for more information.

Pustell Matrix Analysis (Dot Plot)

This tool is useful for identifying weak regions of similarity between two sequences. It is not designed to show full-length alignments between two sequences, but instead shows shorter segments of direct or inverted similarity.

You can use this to identify shorter regions of similarity, then copy those sections to new Sequence windows for more in depth analysis using Multiple Sequence Alignment or Align to Reference.

Dot Plots are also the best way of identifying sequence rearrangements – the display clearly shows insertions and deletions (the main diagonal will be broken and have an offset) and inversions (the inverted diagonal will run bottom left to top right and be colored blue).

Finally, you can use it to identify repetitive regions which appear as parallel diagonals offset from the main diagonal.

Pustell Matrices can be used not only to compare DNA:DNA and Protein:Protein, but also DNA:Protein, where the algorithm will translate the DNA in all 6 frames before aligning to the protein.

See Chapter 15, *"Comparing Sequences using Pustell Matrix Analysis (Dot Plot)"* for more information.

Internet BLAST

Use this tool to identify and align a test sequence to the databases at the NCBI using the popular BLAST algorithm. It also provides a quick and easy method for retrieving sequences from a BLAST search. Simply select a hit in the BLAST results and then choose **Database | Retrieve to Disk** or **Database | Retrieve to Desktop** to download the matching sequence from the NCBI.

See Chapter 14, *"Aligning and Downloading Sequences with BLAST"* for more information.

Align to Folder

This tool allows you to scan a local folder containing sequences (in any format MacVector can recognize) and align them using the FastA alignment algorithm. Using Align to Folder is similar to performing a local BLAST search but it is more sensitive.

Like the Pustell Matrix, you can choose to search DNA with Protein and vice versa.

Many users like this function because the text alignment output also shows the features in the test sequence. This can be very useful for demonstrating the differences between your sequence and other sequences for patent purposes.

See Chapter 13, *"Aligning sequences using a folder search"* for more information.

Contig Assembly

This requires our optional Assembler add-on. Use this tool if you want to align ten or more DNA sequences, with the intention of assembling them into a longer sequence with a consensus.

Contig Assembly allows you to create *de novo* assemblies, in which no reference or scaffold sequence is used, using `Phrap` and reference assemblies using Bowtie.

While you can use Assembler for resequencing, you should consider whether the Align to Reference function might be a better choice. In general, Align to Reference is better for small-scale sequencing projects such as confirming constructs, site directed mutagenesis, etc. Whereas, Assembler should be used for larger resequencing projects, like those with more than a few megabases of template, multiple templates or more than a few thousand reads (e.g. NGS data).

See Chapter 19, *"Sequence Assembly"* for more information.

13 Aligning sequences using a folder search

Overview

The Align to Folder feature allows you to search on sequences stored in a single folder, and enables you to:

- search folders to see if any sequences in those folders have a match to the active sequence
- search folders to rank the sequences in those folders by how closely they are related to the active sequence
- search folders to find DNA sequences that, when translated, match the active protein sequence.

Contents

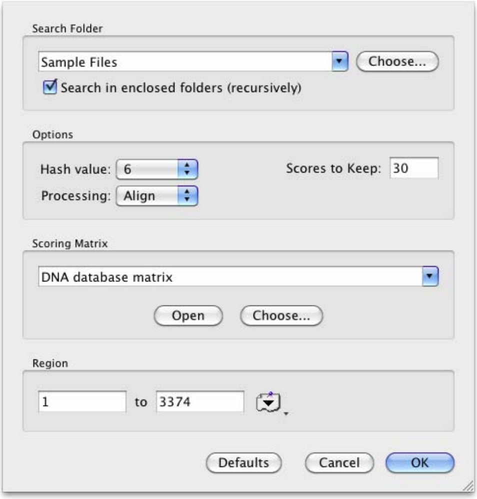
The scoring matrix pre-filter

MacVector uses a scoring matrix to pre-filter the sequences under consideration. An outline of the scoring matrix and its parameters is presented in Chapter , *“Understanding Sequence Comparisons”*. The match and mismatch scores used in the scoring can be reassigned, or new scoring matrices can be created and edited. See *“Scoring matrix files”* on page 68, for further details.

When it finds a match, MacVector computes an initial score for the matching sequence. You have the choice of using this score as a ranking to reject sequences before alignment. Alternatively, you can perform an optimal alignment on each sequence that passes a cut-off score, introducing gaps and indels if required, and calculate an optimal score to use as the basis for acceptance or rejection.

Aligning sequences in a folder

User folders may contain files in any format read by MacVector.



Search Folder

Sample Files

☑ Search in enclosed folders (recursively)

Options

Hash value: 6

Scores to Keep: 30

Processing: Align

Scoring Matrix

DNA database matrix

Region

1 to 3374

To align sequences using a scoring matrix pre-filter

1. Choose **Database | Align to Folder**.

The **Align to Folder** dialog box is displayed.

2. Click **Choose...**.

A standard dialog box appears that enables you to locate and select the folder containing sequence files.

3. Optionally, uncheck **Search in enclosed folders (recursively)** to exclude sequences in sub-folders of the selected folder from the alignment.
4. Choose a value from the **Hash value** drop-down menu.

The **Hash value** is a measure of how long an exact match must be between two sequences before MacVector will attempt to score and align that matching region. For protein sequences, a hash value of 1 is the most sensitive, and 2 is the least sensitive. For DNA sequences, a hash value of 1 is the most sensitive, and 6 is the least sensitive.

Tip. For most comparisons, start with a hash value of 2 for a protein query sequence, or 6 for a DNA query sequence. This is because it is unusual for two sequences to possess significant similarity without having regions of those sizes that match exactly.

5. Enter the number of matching sequences to retain in the **Scores to Keep** text box.

If the program finds more matches than this, the list will be trimmed by dropping sequences with the lowest scores.

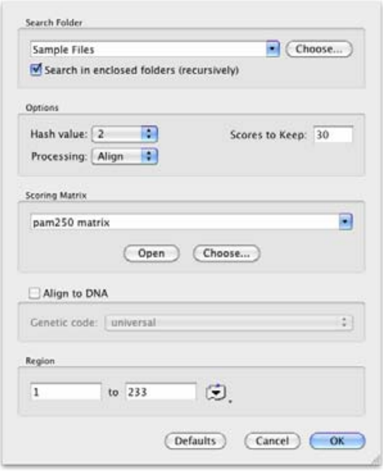
6. Choose from the **Processing** drop-down menu whether an optimal alignment should be done on-the-fly or at the end of the database search, as follows:
   - choose **None** to save the sequences in order of the initial score. If more matches are found than you wanted to keep, the sequences with the lowest initial scores are dropped. At the end of the search, MacVector performs an optimal alignment of each of the saved sequences with the query sequence. Indels and gaps are introduced if they will improve the optimized score. The matches are then listed by optimized score in the results windows.
   - choose **Align** to perform an optimal alignment for any sequence whose initial score exceeds a minimum cut-off score. The matching sequences are saved in order of the optimized score rather than

the initial score. If more matches are found than you wanted to keep, the sequences with the lowest optimized scores are dropped.

7. Click **Choose...** in the Scoring Matrix panel to choose a scoring matrix file.

A standard dialog box is displayed so that you can select the file to use. Make sure that you choose a protein scoring matrix for a protein query sequence, and a DNA scoring matrix for a DNA query sequence.

8. You can limit the query sequence to a region of the entire sequence by typing in the numbers that bracket the region in the **Region** text boxes, or by selecting a region from the feature selector drop-down menu that appears to the right of the text boxes.
9. If you have a protein query sequence, select the **Align to DNA** check box to compare the query only to the nucleic acid sequences that are in the folder.



Sample Files
☑ Search in enclosed folders (recursively)
Hash value: 2
Processing: Align
Scores to Keep: 30
pam250 matrix
☐ Align to DNA
Genetic code: universal
1 to 233

Note. The **Align to DNA** check box only appears when you have a protein query sequence.

Each nucleic acid sequence is translated on-the-fly in all six reading frames and the resulting amino acid sequences are compared with the protein query sequence. Do not select this check box if you want to compare the query with protein sequences.

10. If the **Align to DNA** check box is selected, use the **Genetic code** drop-down menu to choose the genetic code that will be used to translate the nucleic acid sequences in the folder.

11. Select **OK** to perform the alignment.

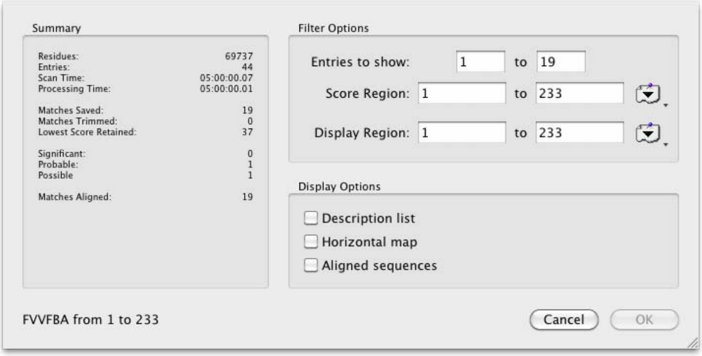
Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the analysis.

When the analysis is complete, the **Folder Sequence Query** display dialog box is displayed. This dialog box is described in the following section.

Displaying alignment results

You can use the **Folder Sequence Query** panel to filter results and display the following:

- a text file describing the analysis
- a horizontal map displaying the aligned sequences represented as horizontal bars
- an aligned sequence display.



Summary

Residues:	69737
Entries:	44
Scan Time:	05:00:00.07
Processing Time:	05:00:00.01
Matches Saved:	19
Matches Trimmed:	0
Lowest Score Retained:	37
Significant:	0
Probable:	1
Possible	1
Matches Aligned:	19

Filter Options

Entries to show: 1 to 19

Score Region: 1 to 233

Display Region: 1 to 233

Display Options

□ Description list

□ Horizontal map

□ Aligned sequences

FVVFBA from 1 to 233

To display sequence alignment results

1. The **Folder Sequence Query** panel is displayed on completion of each analysis. To display it at other times, for example to change the display parameters, choose **Database | Align to Folder** when any analysis display result window is active.
2. You can limit the displays to a single match or to a consecutive subset of the ordered matches by typing the numbers in the **Entries to show** text boxes.
3. You can limit the region that will be scored by typing in the numbers that bracket the region in the **Score Region** text boxes, or by selecting a region from the feature selector drop-down menu that appears to the right of the text boxes.

The numbers are with reference to the query sequence.

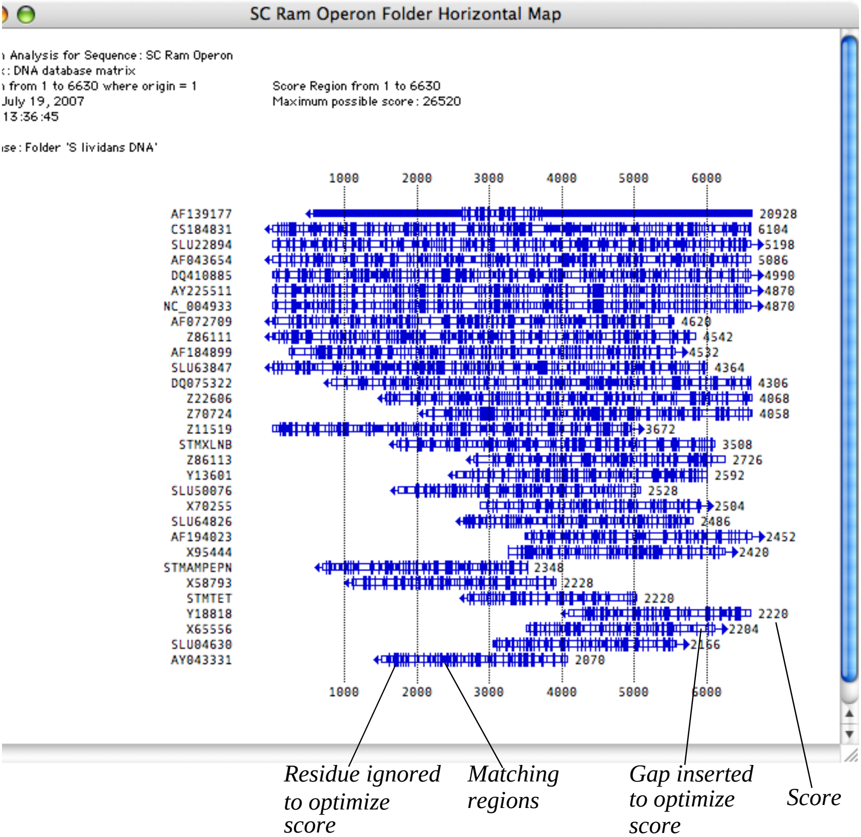
4. You can limit the region that will be displayed by typing in the numbers that bracket the region in the **Display Region** text boxes, or by selecting a region from the feature selector drop-down menu that appears to the right of the text boxes.

The numbers are with reference to the query sequence. The display region must lie within the score region.

5. Check **Description list** to display a list of the saved matches in order of optimized score.
6. Check **Horizontal map** to display a graphical representation of the alignments.
7. Check **Aligned sequences** to display the optimized alignments between the query sequence and each of the saved matching sequences.
8. Select **OK** to generate the results displays.

The results windows

An example Horizontal Map window is shown below.



SC Ram Operon Folder Horizontal Map

```
ı Analysis for Sequence : SC Ram Operon
ː: DNA database matrix
ı from 1 to 6630 where origin = 1          Score Region from 1 to 6630
July 19, 2007                              Maximum possible score : 26520
13:36:45

ıse : Folder 'S lividans DNA'

AF139177    20928
CS184831    6104
SLU22894    5198
AF043654    5086
DQ410885    4990
AY225511    4870
NC_004933   4870
AF072709    4620
Z86111      4542
AF184899    4532
SLU63847    4364
DQ075322    4306
Z22606      4068
Z70724      4058
Z11519      3672
STMXLNB     3508
Z86113      2726
Y13601      2592
SLU50076    2528
X70255      2504
SLU64826    2486
AF194023    2452
X95444      2420
STMAMPEPN   2348
X58793      2228
STMTET      2220
Y18818      2220
X65556      2204
SLU04630    2166
AY043331    2070
```

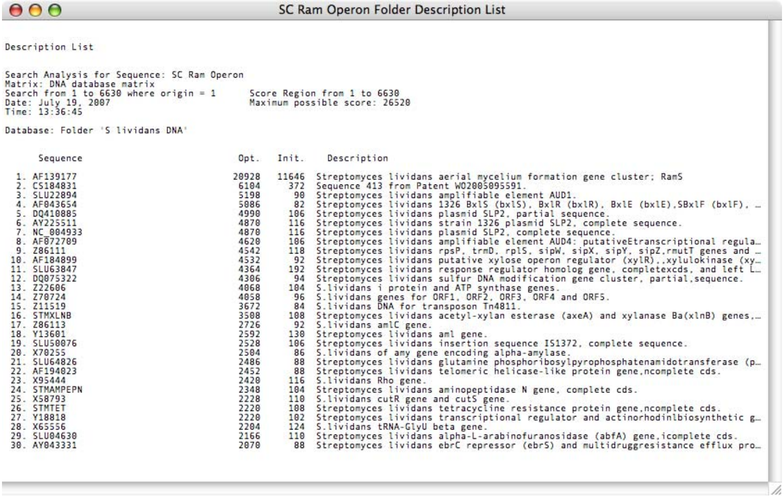
Residue ignored to optimize score

Matching regions

Gap inserted to optimize score

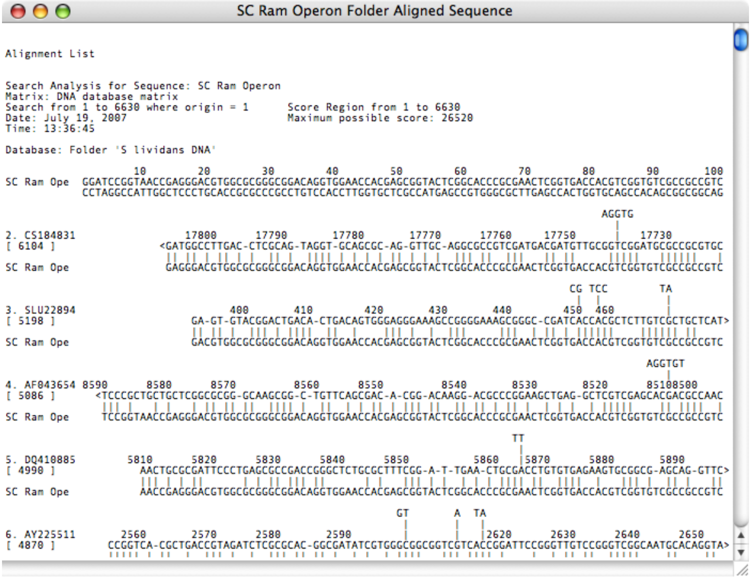
Score

The Description List window shows the locus name of the sequence, its optimized and initial scores, and the first part of the sequence entry's definition line, if one was present. An example Description List window is shown below:



```
Description List

Search Analysis for Sequence: SC Ram Operon
Matrix: DNA database matrix
Search from 1 to 6630 where origin = 1      Score Region from 1 to 6630
Date: July 19, 2007                         Maximum possible score: 26520
Time: 13:36:45

Database: Folder 'S lividans DNA'

      Sequence                          Opt.   Init.    Description

  1. AF139177                          20928   11646  Streptomyces lividans aerial mycelium formation gene cluster; RamS
  2. CS184831                           6104     372  Sequence 413 from Patent WO2005095591.
  3. SLU22894                           5198      90  Streptomyces lividans amplifiable element AUD1.
  4. AF043654                           5086      82  Streptomyces lividans 1326 BxlS (bxlS), BxlR (bxlR), BxlE (bxlE),SBxlF (bxlF), …
  5. DQ410885                           4990     106  Streptomyces lividans plasmid SLP2, partial sequence.
  6. AY225511                           4870     116  Streptomyces lividans strain 1326 plasmid SLP2, complete sequence.
  7. NC_004933                          4870     116  Streptomyces lividans plasmid SLP2, complete sequence.
  8. AF072709                           4620     106  Streptomyces lividans amplifiable element AUD4: putativeEtranscriptional regula…
  9. Z86111                             4542     118  Streptomyces lividans rpsP, trmD, rplS, sipW, sipX, sipY, sipZ,rmutT genes and …
 10. AF184899                           4532      92  Streptomyces lividans putative xylose operon regulator (xylR),,xylulokinase (xy…
 11. SLU63847                           4364     192  Streptomyces lividans response regulator homolog gene, completexcds, and left L…
 12. DQ075322                           4306      94  Streptomyces lividans sulfur DNA modification gene cluster, partial,sequence.
 13. Z22606                             4068     104  S.lividans i protein and ATP synthase genes.
 14. Z70724                             4058      96  S.lividans genes for ORF1, ORF2, ORF3, ORF4 and ORF5.
 15. Z11519                             3672      84  S.lividans DNA for transposon Tn4811.
 16. STMXLNB                            3508     108  Streptomyces lividans acetyl-xylan esterase (axeA) and xylanase Ba(xlnB) genes,…
 17. Z86113                             2726      92  S.lividans amlC gene.
 18. Y13601                             2592     130  Streptomyces lividans aml gene.
 19. SLU50076                           2528     106  Streptomyces lividans insertion sequence IS1372, complete sequence.
 20. X70255                             2504      86  S.lividans of amy gene encoding alpha-amylase.
 21. SLU64826                           2486      88  Streptomyces lividans glutamine phosphoribosylpyrophosphatenamidotransferase (p…
 22. AF194023                           2452      88  Streptomyces lividans telomeric helicase-like protein gene,ncomplete cds.
 23. X95444                             2420     116  S.lividans Rho gene.
 24. STMAMPEPN                          2348     104  Streptomyces lividans aminopeptidase N gene, complete cds.
 25. X58793                             2228     110  S.lividans cutR gene and cutS gene.
 26. STMTET                             2220     108  Streptomyces lividans tetracycline resistance protein gene,ncomplete cds.
 27. Y18818                             2220     102  Streptomyces lividans transcriptional regulator and actinorhodinlbiosynthetic g…
 28. X65556                             2204     124  S.lividans tRNA-GlyU beta gene.
 29. SLU04630                           2166     110  Streptomyces lividans alpha-L-arabinofuranosidase (abfA) gene,icomplete cds.
 30. AY043331                           2070      88  Streptomyces lividans ebrC repressor (ebrS) and multidruggresistance efflux pro…
```

The Aligned Sequence window shows the residue-by-residue comparison of each aligned sequence with the query sequence. And example of the Aligned Sequence window is shown below:



```
Alignment List

Search Analysis for Sequence: SC Ram Operon
Matrix: DNA database matrix
Search from 1 to 6630 where origin = 1      Score Region from 1 to 6630
Date: July 19, 2007                         Maximum possible score: 26520
Time: 13:36:45

Database: Folder 'S lividans DNA'

                    10        20        30        40        50        60        70        80        90      100
SC Ram Ope  GGATCCGGTAACCGAGGGACGTGGCGCGGGCGGACAGGTGGAACCACGAGCGGTACTCGGCACCCGCGAACTCGGTGACCACGTCGGTGTCGCCGCCGTC
            CCTAGGCCATTGGCTCCCTGCACCGCGCCCGCCTGTCCACCTTGGTGCTCGCCATGAGCCGTGGGCGCTTGAGCCACTGGTGCAGCCACAGCGGCGGCAG

                                                                                            AGGTG
                                                                                              |
2. CS184831                 17800      17790       17780       17770      17760      17750    |    17730
[ 6104 ]                <GATGGCCTTGAC-CTCGCAG-TAGGT-GCAGCGC-AG-GTTGC-AGGCGCCGTCGATGACGATGTTGCGGTCGGATGCGCCGCGTGC
                         || || | || | |  || |  |||| | | | | || | | |  ||| ||  |||   || ||     |||||   ||||||   |
SC Ram Ope               GAGGGACGTGGCGCGGGCGGACAGGTGGAACCACGAGCGGTACTCGGCACCCGCGAACTCGGTGACCACGTCGGTGTCGCCGCCGTC

                                                                                        CG TCC       TA
                                                                                         |  |         |
3. SLU22894                        400       410        420       430       440        450  460       |
[ 5198 ]                     GA-GT-GTACGGACTGACA-CTGACAGTGGGAGGGAAAGCCGGGGAAAGCGGGC-CGATCACCACGCTCTTGTCGCTGCTCAT>
                             || || |  ||| | ||||  ||  |    ||| |  |  |||     |||  | || | ||||||    |||||| ||
SC Ram Ope                   GACGTGGCGCGGGCGGACAGGTGGAACCACGAGCGGTACTCGGCACCCGCGAACTCGGTGACCACGTCGGTGTCGCCGCCGTC

                                                                                                   AGGTGT
                                                                                                      |
4. AF043654 8590      8580      8570         8560      8550         8540       8530       8520      85108500
[ 5086 ]      <TCCCGCTGCTGCTCGGCGCGG-GCAAGCGG-C-TGTTCAGCGAC-A-CGG-ACAAGG-ACGCCCGGAAGCTGAG-GCTCGTCGAGCACGACGCCAAC
               ||| |   | |   | || || ||  |||| |  ||  | | || | ||| ||  || || | || |  | | |  | |||||    || ||||  |
SC Ram Ope     TCCGGTAACCGAGGGACGTGGCGCGGGCGGACAGGTGGAACCACGAGCGGTACTCGGCACCCGCGAACTCGGTGACCACGTCGGTGTCGCCGCCGTC

                                                                               TT
                                                                                |
5. DQ410885        5810      5820      5830      5840      5850          5860   |5870      5880       5890
[ 4990 ]             AACTGCGCGATTCCCTGAGCGCCGACCGGGCTCTGCGCTTTCGG-A-T-TGAA-CTGCGACCTGTGTGAGAAGTGCGGCG-AGCAG-GTTC>
                     ||| | | ||      | | || ||| ||      | |   ||| | |  | | | |||| ||  ||||  |   ||| |  || |   ||
SC Ram Ope           AACCGAGGGACGTGGCGCGGGCGGACAGGTGGAACCACGAGCGGTACTCGGCACCCGCGAACTCGGTGACCACGTCGGTGTCGCCGCCGTC

                                                             GT       A  TA
                                                              |       |   |
6. AY225511       2560      2570      2580        2590        |       |   |2620     2630      2640      2650
[ 4870 ]        CCGGTCA-CGCTGACCGTAGATCTCGCGCAC-GGCGATATCGTGGGCGGCGGTCGTCACCGGATTCCGGGTTGTCCGGGTCGGCAATGCACAGGTA>
                ||||| | ||  |  ||| |  |  ||| || || |  | |  | ||||   ||| |||| |    |  | || ||  |||||    ||    ||
```

The exact appearance of this window is controlled by options on the **Aligned Display** preferences dialog box, accessed by choosing **MacVector | Preferences** from the menu, then clicking the **Aligned Display** icon on the preferences dialog. See *“Formatting the Aligned Sequence view”* on page 41, for further details.

For any sequence in the folder, only one match can be reported. If a folder sequence yields more than one match to the query sequence, only the best match is shown in the results window.

Searching Align to Folder results

The Results search available on the **Find** dialog box enables you to find specified text within the results windows from analyses such as Align to Folder. See *“Searching results”* on page 51.

Additional information

When searching a user folder, there are certain very rare situations that would cause MacVector to treat some protein sequence files in the folder as if they were nucleic acid sequences and *vice versa*. This is because of the way that MacVector differentiates sequence types. If a sequence file contains no information about the type of molecule present (this occurs with line or old-format GCG files), MacVector attempts to classify the sequence by counting the number of A, C, G, T, and U characters and non-printing ASCII characters in the first line of the sequence data. If any non-printing ASCII characters (null characters or control characters) are found in the first line of the sequence data, MacVector assumes that the file is not a sequence file and will not attempt to compare it to the query sequence. If there are no non-printing characters, it checks to see if the count of ACGTU characters exceeds 90 percent of the total characters in the line. If so, the sequence is assumed to be a nucleic acid sequence. Otherwise it is considered to be a protein sequence. If any of the protein sequences in the user folder you are searching are line or GCG files, and have an unusual amino acid composition (abnormally rich in alanine, cysteine, glycine, or threonine), they may be mis-tagged as nucleic acid sequences. Alternatively, if there are any nucleic acid sequences present that contain an unusually high percentage of ambiguous base assignments in the first part of the sequence, they may be misclassified as protein sequences and thus not compared with the query sequence. Such misclassifications will probably occur only for very short sequences with unusual compositions.

14 Aligning and Downloading Sequences with BLAST

Overview

MacVector enables you to use the BLAST (Basic Local Alignment Search Tool) heuristic search algorithm to search sequence databases available on the National Center for Biotechnology Information (NCBI) server, to find sequences that are similar to either nucleic acid or protein query sequences.

Contents

Introduction

Currently, homologs to a query sequence can be determined using the following search types:

- **BLASTN** - compares a nucleotide query sequence against a nucleotide sequence database.
- **BLASTP** - compares an amino acid query sequence against a protein sequence database.
- **BLASTX** - compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- **TBLASTN** - compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
- **TBLASTX** - compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

All of the above search types use the newest BLAST algorithms (BLAST 2; Altschul *et al.* 1997). All except TBLASTX offer gapped alignment searching. Because they yield more biologically meaningful results, gapped searches are enabled by default.

These BLAST search programs ascribe significance to their findings using the statistical methods of Karlin and Altschul (1990) or (1993). For a discussion of basic issues in similarity searching of sequence databases, see Altschul *et al.*(1994); for information about the BLAST 2 programs, see Altschul *et al.* (1997).

PSI-BLAST, PHI-BLAST, and 2-sequence BLAST searches are not currently available.

The BLAST programs have been tailored for sequence similarity searching, and are not generally useful for motif style searches.

The BLAST analysis also enables you to:

- extract matching sequences from the database to the desktop, or save them to a newly created folder on your hard drive;
- extract the annotation and features table sections of the sequence homologs that you have identified;
- extract PUBMED abstracts from your sequence homologs and save them to a newly created folder on your hard drive.

Note. NCBI request that authors should cite Altschul *et al.* (1997) when reporting results using gapped BLAST searches.

Using the BLAST searches

MacVector can directly perform BLAST searches of the sequence databases available on the NCBI server over the Internet.

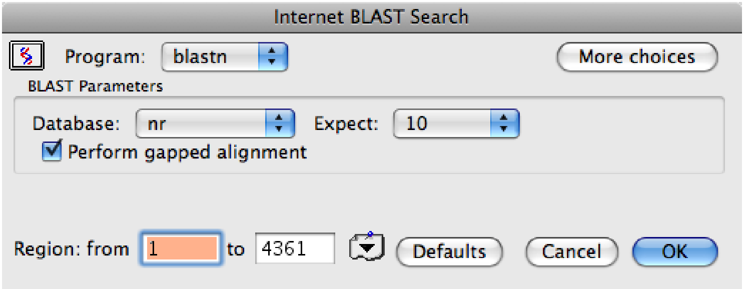
Refer to Appendix A, *“Setting up NCBI's Entrez and BLAST Services”*, for details of setting up the Internet access.

To perform a BLAST search, a nucleic acid or a protein Sequence window must be the active window.

The query sequence cannot exceed 100,000 residues. If your sequence is longer than this, you will be required to specify a region of the sequence to be used as the query.

Performing a standard BLAST search

The MacVector interface to the BLAST server has been designed to give the same results as the NCBI web page interface. To accomplish this, a standard search is provided on a dialog box with few options.



Internet BLAST Search

Program: blastn

BLAST Parameters

Database: nr

Expect: 10

☑ Perform gapped alignment

Region: from 1 to 4361

To perform a standard BLAST search

1. Choose **Database | Internet BLAST Search**.

While the connection is being made to the NCBI server, an information box is displayed. The **Internet BLAST Search** dialog box is then displayed.

Note. If any part of the connection process fails, an appropriate error message is displayed and the entire BLAST search is aborted.

2. Choose a search program from the **Program** drop-down menu.

MacVector only displays program options relevant to the current query sequence type.

3. Choose a database to search from the **Database** drop-down menu.

MacVector only displays databases relevant to the current search program you have requested.

4. Choose a threshold score for the BLAST search from the **Expect** drop-down menu.

Lower **Expect** thresholds are more stringent, leading to fewer chance matches being reported.

5. If you do not want a gapped search, click in the **Perform gapped alignment** checkbox to deselect this option.

6. You can limit the query sequence to a region of the entire sequence. One way to do this is to type, in the **Region** text boxes, the numbers of the first and last residues in the region. Another way is to select a region from the feature selector drop-down menu to the right of the text boxes.

7. Click **More choices** to expand the dialog box and modify the default settings.

The extra settings are described in *“Performing an advanced BLAST search”* on page 265.

Note. If a **Fewer Choices** button is displayed, the expanded form of the **Internet BLAST Search** dialog box is already displayed.

8. Select **OK** to perform the BLAST search.

Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the analysis.

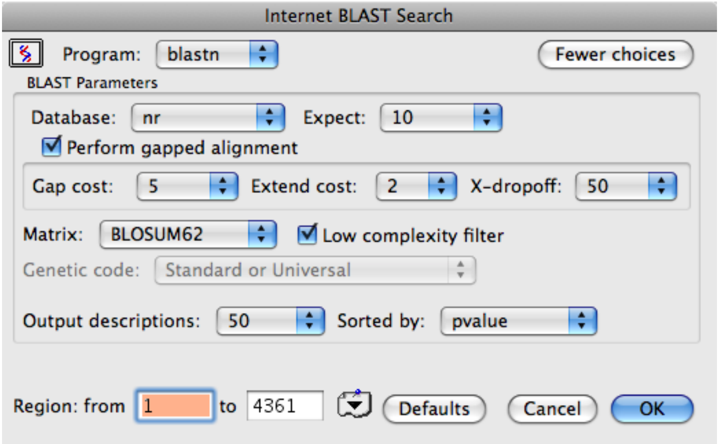
An information box indicates the elapsed time since the BLAST search was started, and the estimated time for the completion of the job. You can click **Close** to dismiss this dialog. Once dismissed the progress of the Blast job can be monitored in the Job Manager. See Appendix B, *“Using the Job Manager”* for more details about the Job Manager. There is also a **Stop** button which will cancel the BLAST Search.

When the search is complete, the **Stop** button changes to **View**.

9. Click **View** or use the Job Manager to access the results of the completed BLAST search.

Performing an advanced BLAST search

MacVector enables you to change the standard BLAST search options, by expanding the **Internet BLAST Search** dialog box.



Program: blastn

Database: nr Expect: 10

☑ Perform gapped alignment

Gap cost: 5 Extend cost: 2 X-dropoff: 50

Matrix: BLOSUM62 ☑ Low complexity filter

Genetic code: Standard or Universal

Output descriptions: 50 Sorted by: pvalue

Region: from 1 to 4361

To perform an advanced BLAST search

1. Choose **Database | Internet BLAST Search**.

While the connection is being made to the NCBI server, an information box is displayed. The **Internet BLAST Search** dialog box is then displayed.

Note. If any part of the connection process fails, an appropriate error message is displayed and the entire BLAST search is aborted.

2. Set the **Program**, **Database**, and **Expect** values, as described in *“To perform a standard BLAST search”* on page 263.
3. Click **More choices** to expand the dialog box.

The dialog box expands to show additional options.

Note. If a **Fewer Choices** button is displayed, the expanded form of the **Internet BLAST Search** dialog box is already displayed.

4. The default setting is to perform a gapped alignment. If you do not want this, click in the **Perform gapped alignment** check box to deselect this option.
5. Set the gapped alignment controls as required (these will be disabled if you deselected **Perform gapped alignment** in the previous step):

- Use the **Open cost** drop-down menu to choose a penalty value for inserting a gap
  A high value will favor alignments with as few gaps as possible (range 3 - 19, default 11)
- Use the **Extend cost** drop-down menu to choose a penalty value for extending a gap
  A high value will favor alignments with many short gaps over ones with fewer longer gaps (range 1 - 3, default 1)
- Use the **X-dropoff** drop-down menu to choose a criterion for discarding initially unpromising alignments.
  Lower values will give faster searches, but may miss some significant matches (range 15 - 50, default 50).

6. Choose a scoring matrix for the BLAST program from the **Matrix** drop-down menu.

Note. This item is disabled if the BLASTN program is selected in the program drop-down menu.

7. Select the **Low Complexity Filter** check box to enable filtering.

   This feature uses sequence similarity to mask regions that are non-specific for protein identification. It can eliminate spuriously high scores that reflect compositional bias rather than specific pairwise alignment (e.g., hits against proline-rich regions or poly-A tails). Queries searched with the BLASTN program are filtered with DUST. Other programs use SEG.

8. Choose a genetic code from the **Genetic code** drop-down menu.

   This is the code used to translate the query sequence, and defaults to the standard universal genetic code.

Note. This item is only enabled if the BLASTX program is selected under the program drop-down menu.

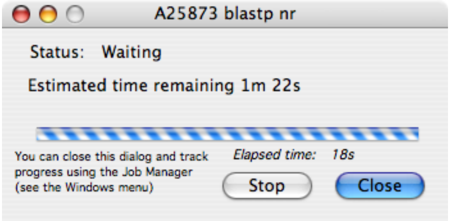
9. Choose a value from the **Output descriptions** drop-down menu to restrict the number of short descriptions of matching sequences reported.
10. Use the **Sorted by** drop-down menu to control the order in which database sequence matches are reported in the output from a BLAST search:
    - choose **pvalue** to sort from the most statistically significant to the least statistically significant

- choose **count** to sort from highest to lowest by the number of High Scoring Segment Pairs (HSP) found for each database sequence
- choose **highscore** to sort from highest to lowest by the score of the highest scoring HSP for each database sequence
- choose **total score** to sort from the highest to the lowest by the sum total score of all HSPs for each database sequence.

11. You can limit the query sequence to a region of the entire sequence. To do this, type in the numbers that bracket the region in the **Region** text boxes, or select a region from the feature selector drop-down menu that appears to the right of the text boxes.
12. Select **OK** to perform the BLAST search.

Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the analysis.

An information box indicates the elapsed time since the BLAST search was started, and the estimated time for the completion of the job. You can click **Close** to dismiss this dialog. Once dismissed the progress of the BLAST job can be monitored in the Job Manager. See Appendix B, *"Using the Job Manager"* for more details about the Job Manager. There is also a **Stop** button which will cancel the BLAST Search.



A25873 blastp nr

Status: Waiting

Estimated time remaining 1m 22s

You can close this dialog and track progress using the Job Manager (see the Windows menu)

Elapsed time: 18s

Note. At the conclusion of a BLAST search, matching sequences are not automatically retrieved from the remote NCBI server. If you want to perform additional analyses on matching sequences, you must first retrieve them. See *"Retrieving matching sequences"* on page 270, for further details.

When the search is complete, the **Stop** button changes to **View**.

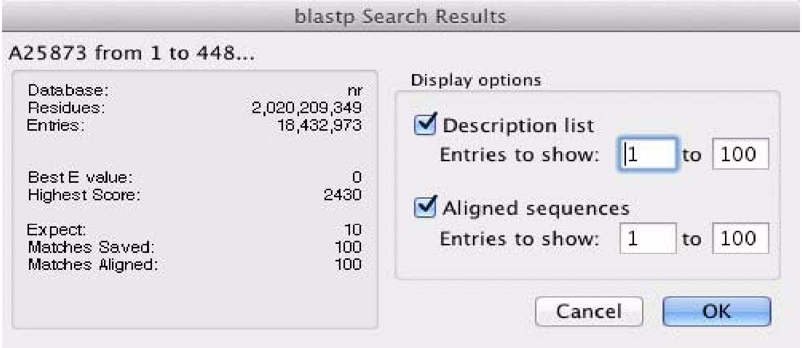
13. Click **View** or use the Job Manager to access the results of the completed BLAST search.

This is described in the following section.

Displaying BLAST search results

You can use the **BLAST Search Results** display dialog box to filter results and display the following:

- a text file listing the search results
- a textual aligned sequence display



blastp Search Results

A25873 from 1 to 448...

Database:	nr
Residues:	2,020,209,349
Entries:	18,432,973
Best E value:	0
Highest Score:	2430
Expect:	10
Matches Saved:	100
Matches Aligned:	100

Display options

☑ Description list

Entries to show: 1 to 100

☑ Aligned sequences

Entries to show: 1 to 100

A few statistics of the search are displayed in the statistics panel on the left side of the dialog. These include the name, size and build date of the database used. Below these are the best scores that were found, in particular the best (i.e. lowest) SumP(n) and the largest score obtained by a single high-scoring segment pair. Finally, the value of the expect parameter is shown, and the number of matches that were saved and aligned for viewing in the Description list and Aligned sequences windows.

To display BLAST search results

1. Ensure that the **BLAST Search Results** display dialog box is displayed.

This dialog box can be displayed by clicking **View** on the **Internet BLAST Search** dialog box after a search has finished. Alternatively, click **View** on the appropriate job in the Job Manager to display the **BLAST Search Results** display dialog box.

2. Select the **Description list** check box to generate an annotated list of the matching sequences, numbered in the order that was specified in the **Sorted by** search parameter.

Note. If you want to save matching sequences to a new folder, or open sequence windows for them, you must generate a description list.

You can restrict the description list, by typing in appropriate list numbers in the **Entries to show** text boxes.

3. Select **Aligned sequences** if you want to display the optimized alignments between the query sequence and each of the matching database sequences.

You can restrict the aligned sequences displayed, by typing in appropriate list numbers in the **Entries to show** text boxes.

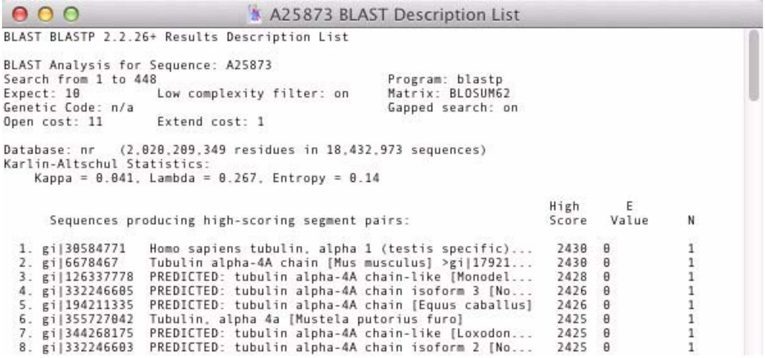
Note. The aligned sequences window will not be created unless the **Aligned sequences** checkbox is selected.

4. Select **OK** to close the BLAST Search Results dialog box and display the chosen results.

This button is only activated if at least one of the checkboxes has been selected.

5. If you need to display further results from the same search, ensure that a BLAST search results window is active, and then choose **Database | Internet BLAST Search**. The **BLAST Search Results** display dialog box will be displayed again.

The results windows



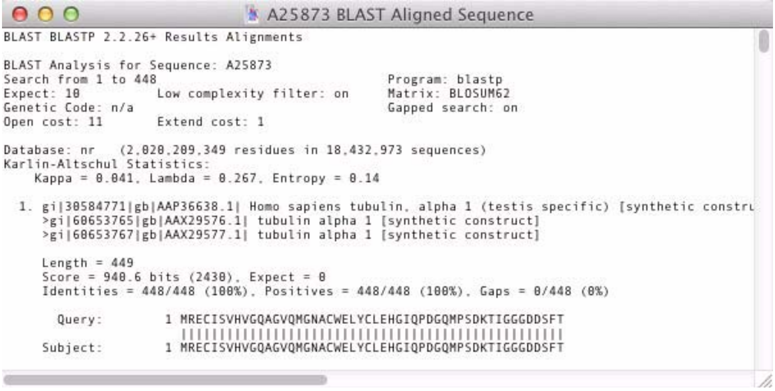
```
BLAST BLASTP 2.2.26+ Results Description List

BLAST Analysis for Sequence: A25873
Search from 1 to 448                           Program: blastp
Expect: 10          Low complexity filter: on     Matrix: BLOSUM62
Genetic Code: n/a                              Gapped search: on
Open cost: 11       Extend cost: 1

Database: nr   (2,020,209,349 residues in 18,432,973 sequences)
Karlin-Altschul Statistics:
    Kappa = 0.041, Lambda = 0.267, Entropy = 0.14

                                                                High      E
     Sequences producing high-scoring segment pairs:            Score   Value      N

  1. gi|30584771   Homo sapiens tubulin, alpha 1 (testis specific)...   2430  0          1
  2. gi|6678467    Tubulin alpha-4A chain [Mus musculus] >gi|17921...   2430  0          1
  3. gi|126337778  PREDICTED: tubulin alpha-4A chain-like [Monodel...   2428  0          1
  4. gi|332246605  PREDICTED: tubulin alpha-4A chain isoform 3 [No...   2426  0          1
  5. gi|194211335  PREDICTED: tubulin alpha-4A chain [Equus caballus]   2426  0          1
  6. gi|355727042  Tubulin, alpha 4a [Mustela putorius furo]            2425  0          1
  7. gi|344268175  PREDICTED: tubulin alpha-4A chain-like [Loxodon...   2425  0          1
  8. gi|332246603  PREDICTED: tubulin alpha-4A chain isoform 2 [No...   2425  0          1
```

The Description List window shows you the saved matches in the order determined by the **sort by** choice on the **Internet BLAST Search** dialog box. The list shows the following information:

- the document ID of the sequence, usually a Universal Identifier prefaced by the characters gi |
- a short definition of the sequence
- the score of the highest-scoring segment pair

- the smallest sum probability value of the match
- the number of segments that were aligned to the target sequence.

Note. When the Description List window is active, it can be used to extract database results. See*"Retrieving matching sequences"* on page 270, for further details.



A25873 BLAST Aligned Sequence

```
BLAST BLASTP 2.2.26+ Results Alignments

BLAST Analysis for Sequence: A25873
Search from 1 to 448                              Program: blastp
Expect: 10          Low complexity filter: on     Matrix: BLOSUM62
Genetic Code: n/a                                 Gapped search: on
Open cost: 11       Extend cost: 1

Database: nr   (2,020,209,349 residues in 18,432,973 sequences)
Karlin-Altschul Statistics:
    Kappa = 0.041, Lambda = 0.267, Entropy = 0.14

  1. gi|30584771|gb|AAP36638.1| Homo sapiens tubulin, alpha 1 (testis specific) [synthetic constru
     >gi|60653765|gb|AAX29576.1| tubulin alpha 1 [synthetic construct]
     >gi|60653767|gb|AAX29577.1| tubulin alpha 1 [synthetic construct]

     Length = 449
     Score = 940.6 bits (2430), Expect = 0
     Identities = 448/448 (100%), Positives = 448/448 (100%), Gaps = 0/448 (0%)

       Query:        1 MRECISVHVGQAGVQMGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFT
                         ||||||||||||||||||||||||||||||||||||||||||||||||||
     Subject:        1 MRECISVHVGQAGVQMGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFT
```

The Aligned Sequences window displays the alignments of the high scoring segment pairs between the query sequence and the database sequence. The appearance of the aligned sequence display can be changed interactively, see Chapter , *"General Procedures"*, for details.

The aligned sequences output closely follows the traditional BLAST format used by NCBI:

- each matching sequence has a header section describing the accession numbers and definitions of the entry in the searched database
- the individual high-scoring segment pairs are presented with scoring information for each segment
- the actual alignments with the query sequence presented on top and the database sequence presented below.

Searching BLAST results

The Results search available on the **Find** dialog box enables you to find specified text within the results windows from analyses such as BLAST. See *"Searching results"* on page 51.

Retrieving matching sequences

Two methods are available to retrieve the matching sequences:

- open a new sequence window for each match of interest
- save each match of interest to a new folder on disk.

To create new sequence windows

1. Make the Description List window active.
2. Select the sequences of interest in the description list. You just need to select at least one character on the line containing the sequence for it to be considered selected.
3. Choose **Database | Retrieve to Desktop**.

All selected files are placed in their own sequence window.

To save sequences to a folder

1. Make the Description List window active.
2. Select the sequences of interest in the description list.
3. Choose **Database | Retrieve to Disk**.

A standard dialog box is displayed, so you can save the sequences to an existing folder or to create a new folder.

Storing BLAST searches

You can save the contents of the Description List and Aligned Sequences windows to text files in the normal way, by choosing **File | Save As...** while the window you want to save is selected.

15 Comparing Sequences using Pustell Matrix Analysis (Dot Plot)

Overview

MacVector provides a number of routines for aligning and comparing sequences. For sequence similarity, a matrix comparison is the method of choice for obtaining an overall picture of how the sequences are related, and is the subject of this chapter.

The MacVector sequence alignment routines use elements of the matrix comparison method as a rapid filter before the actual alignment.

Refer to Chapter 12, *"Aligning Sequences"*, for information about scoring matrix methods for sequence comparison and alignment, and sequence comparisons using the *Entrez* database.

Contents

Pustell DNA matrix

The MacVector implementation of DNA matrix analysis uses techniques developed by Pustell. The method enables you to:

- look for regions of similarity between two nucleic acid sequences using a dot matrix plot
- look for direct or inverted repeats within a single nucleic acid sequence
- display matching regions as either diagonal lines or as diagonally arranged characters whose values indicate the degree of similarity

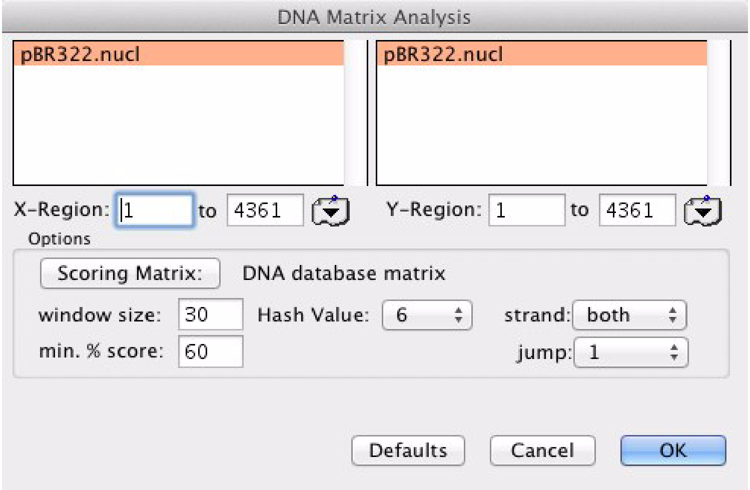
An outline of the matrix analysis and the parameters used are presented in Chapter 21, *“Understanding Sequence Comparisons”*. The match and mismatch scores used in scoring the comparison can be reassigned, or new scoring matrices can be created and edited. See *“Scoring matrix files”* on page 68, for further details.

Performing a DNA matrix analysis

To use this functionality, one or more nucleic acid sequences must be open, a window containing a nucleic acid sequence must be the active window, and a nucleic acid scoring matrix file must be open or available to the program.

The output is a two-dimensional plot, with the residues of one sequence along the X-axis, and the residues of the comparison sequence along the Y-axis.

Before running this analysis, we recommend that you read Chapter 21, *“Understanding Sequence Comparisons”*, to understand the implications of changing the scoring parameters.



pBR322.nucl

pBR322.nucl

X-Region: 1 to 4361

Y-Region: 1 to 4361

Scoring Matrix: DNA database matrix

window size: 30

Hash Value: 6

strand: both

min. % score: 60

jump: 1

To perform a Pustell DNA matrix analysis

1. Choose **Analyze | Create Dot Plot | Pustell DNA Matrix** from the menu.

The **DNA Matrix Analysis** dialog box is displayed.

2. In the left-hand scrolling list, select a sequence to display along the X-axis.
3. In the right-hand scrolling list, select a sequence to display along the Y-axis.

The name that is highlighted in each list is the sequence assigned to that axis.

4. You can limit the analysis to a region of each of the sequences by typing in the numbers that bracket the region in the **X-Region** and **Y-Region** text boxes, or by selecting a region from the feature selector drop-down menu that appears to the right of the text boxes.

Tip. If you are comparing a very long sequence with a very short one, the procedure will need less memory if you display the shorter sequence on the x-axis.

5. Click the **Scoring Matrix** button to choose a scoring matrix file.

A standard dialog box is displayed, enabling you to search for and select the file to use.

6. Enter a window size in the **window size** text box.

Whenever MacVector finds an exact match, it examines the segment (or window) of the aligned sequences that surround the matching region. The length of the segment is the value typed in for window size.

7. Enter a minimum score in the **min. % score** text box.

Whenever MacVector finds an exact match, it computes a total score for the window using the match / mismatch scores in the scoring matrix. It then determines a percent score by dividing the window's score by the score that would occur if all of the bases in the window matched. If this percent score equals or exceeds the value of **min. % score**, the window is saved.

8. Choose a value from the **Hash Value** drop-down menu.

The hash value is a measure of how long an exact match between two sequences must be before MacVector will attempt to score and align that matching region. A hash value of 1 is the most sensitive, 6 is the least sensitive.

Tip. For most comparisons, start with a hash value of 6, because it is unusual for two sequences to possess significant similarity without having regions of that size that match exactly.

9. Choose a value from the **jump** drop-down menu.

When the **jump** setting is 1, the hash value represents the number of bases in a row that must match perfectly. When the **jump** setting is 3, the hash value is the number of triplets in a row whose first base must match perfectly. When the **jump** setting is **both**, matching regions found for both **jump** settings are scored.

Tip. For the initial analysis stage, we recommend that the **jump** setting be set to **both**.

10. Use the **strand** drop-down menu to choose whether to compare the plus strands of the two sequences (++), the plus strand of the X-axis sequence with the reverse complement strand of the Y-axis sequence (+–), or both.

Tip. For the initial analysis, we recommend that you select **both**.

11. Select **OK** to perform the analysis.

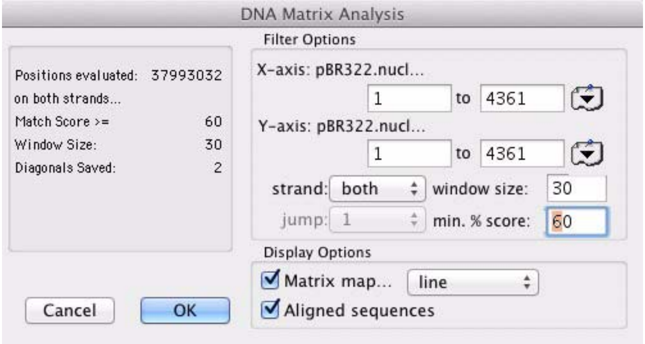
Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the analysis.

When the analysis is complete, the **DNA Matrix Analysis** display dialog box is displayed. This dialog box is described in the following section.

Displaying DNA matrix analysis results

You can use the **DNA Matrix Analysis** display dialog box to filter results and display the following:

- matrix map
- aligned sequences



DNA Matrix Analysis

Positions evaluated: 37993032
on both strands...
Match Score >= 60
Window Size: 30
Diagonals Saved: 2

Filter Options
X-axis: pBR322.nucl... 1 to 4361
Y-axis: pBR322.nucl... 1 to 4361
strand: both
window size: 30
jump: 1
min. % score: 60

Display Options
☑ Matrix map... line
☑ Aligned sequences

To display DNA matrix analysis results

1. The **DNA Matrix Analysis** display dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example to change the display parameters, choose **Analyze | Create Dot Plot | Pustell DNA Matrix** when any matrix analysis display result window is active.
2. You can limit the display to a region of each of the sequences by typing in the numbers that bracket the region in the **X-axis** and **Y-axis** text boxes, or by selecting a region from the feature selector drop-down menu that appears to the right of the text boxes.
3. If you chose **both** for the analysis **strand** setting, you can now choose to display the comparison of the plus strands only (++), the plus strand of the X-axis sequence with the reverse complement strand of the Y-axis sequence (+–), or both. If you limited the initial analysis, you will not be able to display the other types of comparisons.
4. If you set **jump** to **both** in the analysis, you can now separately display the matches found for the **jump** setting of 1 and the **jump** setting of 3, or you can look at both simultaneously. If you limited the initial analysis, the drop-down menu for the jump setting will be disabled.

5. Enter a value for the window size in the **window size** text box.

This can have any value starting from the **window size** used in the original analysis stage up to 60.

6. Enter a value for the score in the **min. % score** text box.

This can have any value starting from the **min. % score** used in the original analysis stage up to 100.

7. Select the **Matrix map** check box to display a dot matrix plot of the results of the comparison.
8. Use the drop-down menu at the end of the **Matrix map** check box to choose how you want to display the matching regions.

The **line** setting displays matching regions as diagonal lines. The **character** setting displays matching regions as diagonally arranged characters. The character used indicates the percent match at that point along the diagonal. “A” represents 100 percent match, and each succeeding letter is two percent less, so “B” is 98 percent, “C” is 96 percent, and so on. A listing of the letter codes can be found in Appendix C, *“Reference Tables”*.

9. Select the **Aligned sequences** check box to see a display of the X-axis sequence aligned with all of the regions of the Y-axis sequence that meet or exceed the **min. % score**.
10. Select **OK** to display the results.

Pustell protein matrix

The MacVector implementation of protein matrix analysis uses techniques developed by Pustell. The method enables you to:

- look for regions of similarity between two protein sequences using a dot matrix plot
- look for repeats within a single protein sequence
- display matching regions as either diagonal lines or as diagonally arranged characters whose values indicate the degree of similarity

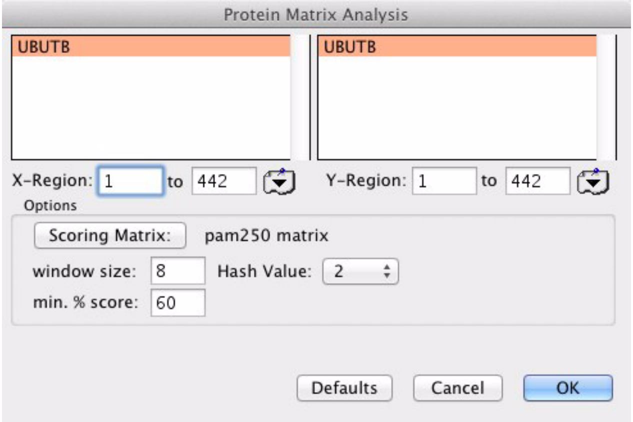
An outline of the matrix analysis and its parameters is presented in Chapter 21, *“Understanding Sequence Comparisons”*. The match and mismatch scores used in scoring the comparison can be reassigned, or new scoring matrices can be created and edited. See *“Scoring matrix files”* on page 68, for further details.

Performing a protein matrix analysis

To run this analysis, one or more protein sequences must be open, a window containing a protein sequence must be the active window, and a protein scoring matrix file must be open or available to the program.

The output is a two-dimensional plot, with the residues of one sequence along the X-axis, and the residues of the comparison sequence along the Y-axis.

Before using this functionality, we recommend that you read Chapter 21, *"Understanding Sequence Comparisons"*, to understand the implications of changing the scoring parameters.



Protein Matrix Analysis

UBUTB

UBUTB

X-Region: 1 to 442

Y-Region: 1 to 442

Options

Scoring Matrix: pam250 matrix

window size: 8

Hash Value: 2

min. % score: 60

To perform a Pustell protein matrix analysis

1. Choose **Analyze | Create Dot Plot | Pustell Protein Matrix** from the menu.

The **Protein Matrix Analysis** dialog box is displayed.

2. In the left-hand scrolling list, select a sequence to display along the X-axis.
3. In the right-hand scrolling list, select a sequence to display along the Y-axis.

The name that is highlighted in each list is the sequence assigned to that axis.

4. You can limit the analysis to a region of each of the sequences by typing in the numbers that bracket the region in the **X-Region** and **Y-Region** text boxes, or by selecting a region from the feature selector drop-down menu that appears to the right of the text boxes.

Tip. If you are comparing a very long sequence with a very short one, the procedure will need less memory if you display the shorter sequence on the x-axis.

5. Click the **Scoring Matrix** button to choose a scoring matrix file.

A standard dialog box appears, enabling you to search for and select the file to use.

6. Enter a window size in the **window size** text box.

Whenever MacVector finds an exact match, it examines the segment (or window) of the aligned sequences that surround the matching region. The length of the segment is the value typed in for window size.

7. Enter a minimum score in the **min. % score** text box.

Whenever MacVector finds an exact match, it computes a total score for the window using the match / mismatch scores in the scoring matrix. It then determines a percent score by dividing the window’s score by the score that would occur if all of the residues in the window matched. If this percent score equals or exceeds the value of **min. % score**, the window is saved.

8. Choose a value from the **Hash Value** drop-down menu.

The hash value is a measure of how long an exact match between two sequences must be before MacVector will attempt to score and align that matching region. A hash value of 1 is the most sensitive, 2 is the least sensitive.

Tip. For most comparisons, start with a hash value of 2, because it is unusual for two sequences to possess significant similarity without having regions of that size that match exactly.

9. Select **OK** to perform the analysis.

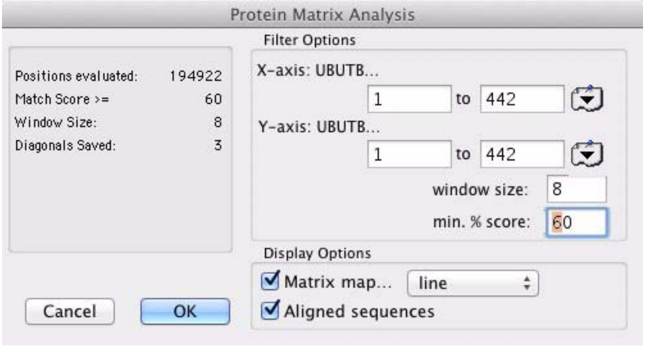
Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the analysis.

When the analysis is complete, the **Protein Matrix Analysis** display dialog box is displayed. This dialog box is described in the following section.

Displaying protein matrix analysis results

You can use the **Protein Matrix Analysis** display dialog box to filter results and display the following:

- matrix map
- aligned sequences.



Protein Matrix Analysis

Positions evaluated: 194922
Match Score >= 60
Window Size: 8
Diagonals Saved: 3

Filter Options
X-axis: UBUTB... 1 to 442
Y-axis: UBUTB... 1 to 442
window size: 8
min. % score: 60

Display Options
☑ Matrix map... line
☑ Aligned sequences

To display protein matrix analysis results

1. The **Protein Matrix Analysis** display dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example to change the display parameters, choose **Analyze | Create Dot Plot | Pustell Protein Matrix** when any matrix analysis display result window is active.
2. You can limit the display to a region of each of the sequences by typing in the numbers that bracket the region in the **X-axis** and **Y-axis** text boxes, or by selecting a region from the features table drop-down menu that appears to the right of the text boxes.
3. Enter a value for the window size in the **window size** text box.

This can have any value starting from the **window size** used in the original analysis stage up to 60.

4. Enter a value for the score in the **min. % score** text box.

This can have any value starting from the **min. % score** used in the original analysis stage up to 100.

5. Select the **Matrix map** check box to display a dot matrix plot of the results of the comparison.
6. Use the drop-down menu at the end of the **Matrix map** check box to choose how you want to display the matching regions.

The **line** setting displays matching regions as diagonal lines. The **character** setting displays matching regions as diagonally arranged characters.

The character used indicates the percent match at that point along the diagonal. “A” represents 100 percent match, and each succeeding letter is two percent less, so “B” is 98 percent, “C” is 96 percent, “a” is 50 percent, and so on. A listing of all the letter codes can be found in Appendix C, *“Reference Tables”*.

7. Select the **Aligned sequences** check box to see a display of the X-axis sequence aligned with all of the regions of the Y-axis sequence that meet or exceed the **min. % score**.
8. Select **OK** to display the results.

Pustell protein and DNA matrix

The MacVector implementation of matrix comparison enables you to compare protein and DNA sequences. You can:

- look for regions of similarity at the amino acid level between a nucleic acid sequence and a protein sequence using a dot matrix plot
- display matching regions as either diagonal lines or as diagonally arranged characters whose values indicate the degree of similarity

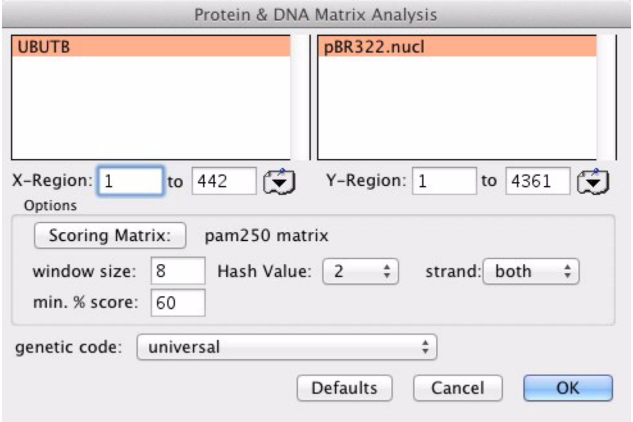
An outline of the matrix analysis and its parameters is presented in Chapter 21, *“Understanding Sequence Comparisons”*. The match and mismatch scores used in scoring the comparison can be reassigned, or new scoring matrices can be created and edited. See *“Scoring matrix files”* on page 68, for further details.

Performing a protein / DNA matrix analysis

To run this analysis, at least one nucleic acid sequence file and one protein sequence file must be open, and a window containing a sequence must be the active window. A protein scoring matrix file must be open or available on disk.

When performing this analysis, MacVector first translates the nucleic acid sequence in the three reading frames of one of the DNA strands (if the strand options of ++ or +– are chosen) or in all six reading frames (if both strands are chosen). The comparison is then performed on the resulting amino acid sequences. The output is a two-dimensional plot, with the residues of the protein sequence along the X-axis, and the residues of the translated DNA sequence along the Y-axis. The aligned sequence display shows each aligned sequence from all the reading frames.

Before running this analysis, we recommend that you read Chapter 21, *"Understanding Sequence Comparisons"*, to understand the implications of changing the scoring parameters.



Protein & DNA Matrix Analysis

UBUTB

pBR322.nucl

X-Region: 1 to 442

Y-Region: 1 to 4361

Options

Scoring Matrix: pam250 matrix

window size: 8

Hash Value: 2

strand: both

min. % score: 60

genetic code: universal

To perform a Pustell protein / DNA matrix analysis

1. Choose **Analyze | Create Dot Plot | Pustell Protein & DNA** from the menu.

The **Protein & DNA Matrix Analysis** dialog box is displayed.

2. In the left-hand scrolling list, select a protein sequence to display along the X-axis. Only protein sequences appear in this list.
3. In the right-hand scrolling list, select a DNA sequence to display along the Y-axis. Only DNA sequences appear in this list.
4. You can limit the analysis to a region of each of the sequences by typing in the numbers that bracket the region in the **X-Region** and **Y-Region** text boxes, or by selecting a region from the feature selector drop-down menu that appears to the right of the text boxes.
5. Click the **Scoring Matrix** button to choose a scoring matrix file.

A standard dialog box is displayed, enabling you to search for and select the file to use.

6. Enter a window size in the **window size** text box.

Whenever MacVector finds an exact match, it examines the segment (or window) of the aligned sequence that surrounds the matching region. The length of the segment is the value typed in for window size.

7. Enter a minimum score in the **min. % score** text box.

Whenever MacVector finds an exact match, it computes a total score for the window using the match / mismatch scores in the scoring matrix. It then determines a percent score by dividing the window's score by the score that would occur if all of the bases in the window matched. If this percent score equals or exceeds the value of **min. % score**, the window is saved.

8. Choose a value from the **Hash Value** drop-down menu.

The hash value is a measure of how long an exact match between two sequences must be before MacVector will attempt to score and align that matching region. A hash value of 1 is the most sensitive, 2 is the least sensitive.

Tip. For most comparisons, start with a hash value of 2, because it is unusual for two sequences to possess significant similarity without having regions of that size that match exactly.

9. Use the **strand** drop-down menu to choose whether to use the plus strand of the DNA sequence (++), the reverse complement strand (+–), or both.

Tip. For the initial analysis, we recommend that you choose **both**.

10. Choose the genetic code for the translation of the DNA from the **genetic code** drop-down menu.

11. Select **OK** to perform the analysis.

Alternatively, select **Defaults** to restore the default settings, or **Cancel** to close the dialog box without performing the analysis.

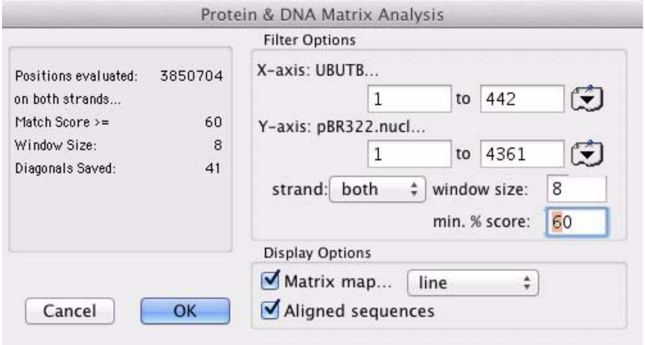
When the analysis is complete, the **Protein & DNA Matrix Analysis** display dialog box is displayed. This dialog box is described in the following section.

Displaying protein / DNA matrix analysis results

You can use the **Protein & DNA Matrix Analysis** display dialog box to filter results and display the following:

- matrix map

- aligned sequences.



Protein & DNA Matrix Analysis

Positions evaluated: 3850704
on both strands...
Match Score >= 60
Window Size: 8
Diagonals Saved: 41

Filter Options

X-axis: UBUTB... 1 to 442

Y-axis: pBR322.nucl... 1 to 4361

strand: both
window size: 8
min. % score: 60

Display Options

☑ Matrix map... line

☑ Aligned sequences

To display protein / DNA matrix analysis results

1. The **Protein & DNA Matrix Analysis** display dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example to change the display parameters, choose **Analyze | Create Dot Plot | Pustell Protein & DNA** when any matrix analysis display result window is active.
2. You can limit the display to a region of each of the sequences by typing in the numbers that bracket the region in the **X-axis** and **Y-axis** text boxes, or by selecting a region from the features table drop-down menu that appears to the right of the text boxes.
3. If you chose **both** for the analysis **strand** setting, you can now choose to display the comparison of the protein with the plus strand only (++), the reverse complement strand only (+–), or both at once. If you limited the initial analysis, you will not be able to display both types of comparisons.
4. Enter a value for the window size in the **window size** text box.

This can have any value starting from the **window size** used in the original analysis stage up to 60.

5. Enter a value for the score in the **min. % score** text box.

This can have any value starting from the **min. % score** used in the original analysis stage up to 100.

6. Select the **Matrix map** check box to display a dot matrix plot of the results of the comparison.

7. Use the drop-down menu at the end of the **Matrix map** check box to choose how you want to display the matching regions.

The **line** setting displays matching regions as diagonal lines. The **character** setting displays matching regions as diagonally arranged characters. The character used indicates the percent match at that point along the diagonal. “A” represents 100 percent match, and each succeeding letter is two percent less, so “B” is 98 percent, “C” is 96 percent, “a” is 50 percent, and so on. A listing of all the letter codes can be found in Appendix C, *“Reference Tables”*.

8. Select the **Aligned sequences** check box to see a display of the protein sequence aligned with all of the regions of the translated DNA sequence that meet or exceed the **min. % score**.
9. Select **OK** to display the results.

16 Aligning Sequences to a Reference

Overview

This chapter describes how to use MacVector to align one or more sequences against a reference sequence so that you can quickly identify differences or similarities. There are two functions available:

- **Sequence Confirmation** lets you align one or more sequence or chromatogram sample files against a reference sequence. This can be used for confirming the sequence of a cloned fragment of DNA or for identifying SNPs or other differences between clones. It is very similar to sequence assembly, except that the alignment always requires a reference sequence to act as an assembly scaffold.
- **cDNA Alignment** lets you align one or more cDNA clones against a reference genomic sequence. This can be used for identifying where exon/intron boundaries lie in cDNA/mRNA sequences.

Contents

Align to Reference

The **Align to Reference** tool in MacVector allows you to align one or more sequences against a reference sequence so that you can quickly identify differences or similarities. To use it, open the reference sequence, launch the **Align to Reference** tool by selecting **Analyze | Align to Reference...** or clicking the **Align to Ref.** icon on the analysis toolbar, and add read files, then run the alignment using either the **Sequence Confirmation** or the **cDNA Alignment** method. The **Sequence Confirmation** method expects small gaps (2 - 4 residues) and only considers the longest aligned segment, whereas the **cDNA Alignment** method expects much larger gaps.

Sequence Confirmation

The **Sequence Confirmation** function lets you align one or more sample sequences against a reference sequence. It has many similarities to sequence assembly, except that it requires and takes advantage of the use of a known reference sequence as a scaffold.

You can use this functionality to help you solve a number of typical laboratory problems:

- Confirming the sequence of a cloned fragment
- Sequencing across the ends of a cloned fragment to confirm the junction sequence
- Screening clones from a site-specific mutagenesis experiment to identify successful mutations
- Screening related clones for single nucleotide polymorphisms

The main limitation of this approach is that you must have a reference sequence to act as a scaffold against which the sample sequences can be aligned. You cannot, therefore, use this function to assemble trace files from *de novo* sequencing projects. For these types of tasks MacVector Assembler is a far better tool. See Chapter 19, *"Sequence Assembly"* for a discussion of the features of MacVector Assembler. However, for the less complex resequencing assembly tasks commonly encountered during a molecular biologist's day, the **Sequence Confirmation** function is ideal.

To align trace files against a template sequence

1. Open the MacVector sequence file you want to use as your template.

2. Choose **Analyze | Align to Reference...** or click the **Align to Ref.** icon on the analysis toolbar to open a new Align to Reference window.

The template sequence is displayed along the top of the Editor view in the new Align to Reference window.

3. Click the **Add Seqs** icon on the toolbar in the Editor view, select the sequence files you wish to assemble against your template then click the **Open** button.

Alternatively, select **Edit | Add Sequences From File** from the main menu, select the sequence files you wish to assemble against your template then click the **Open** button.

Note. You can hold down the **<shift>** key to select multiple sequences to import.

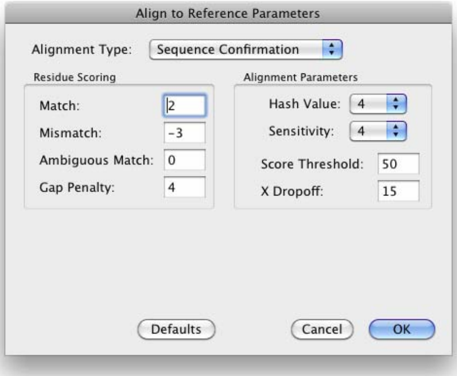
MacVector attempts to import all of the selected files into the assembly project. An error message is displayed listing the names of any files that could not be imported.

The Editor view updates to display the sequences in the order in which they were imported. If trace data is present, this is also displayed in a lower trace panel.

Note. Unaligned sequences are displayed in italics starting at residue “1” to alert you to the fact that they have yet to be aligned (or could not be aligned after automated assembly). Italicized sequences are not included in consensus calculations.

4. Click the **Align** icon on the toolbar in the Editor view.

The **Align to Reference Parameters** dialog box is displayed.



Alignment Type: Sequence Confirmation

Match: 2

Mismatch: -3

Ambiguous Match: 0

Gap Penalty: 4

Hash Value: 4

Sensitivity: 4

Score Threshold: 50

X Dropoff: 15

5. Ensure that **Sequence Confirmation** is selected from the **Alignment Type** drop-down menu.

6. Click **OK** to perform the alignment using the default parameters.

See *"Alignment parameters"* on page 293 for more information about modifying the alignment parameters.

A dialog box is displayed, informing you of the progress of the alignment. You can close this and monitor progress using the Job Manager if you prefer. See Appendix , *"Using the Job Manager"* for more details about the Job Manager. There is also a **Stop** button which will cancel the alignment. The **Align to Reference** job runs in the background, allowing you to continue to use other functions in MacVector while the job is in progress.

Note. **Align to Reference** calculations can take some time if the sequence is long and there are many trace files in the alignment.

7. Finally you can save the edited reference sequence in MacVector format by choosing **File | Save As...** from the Align to Reference window. You can save the actual sequence confirmation project itself at any time and you will be prompted to save any changes when you close the project window.

cDNA Alignment

The **cDNA Alignment** function lets you align one or more cDNA clones against a reference genomic sequence. This can be used for identifying where exon/intron boundaries lie in your cDNA/mRNA sequences.

To align cDNAs against a genomic sequence

1. Open the MacVector sequence file you want to use as your template.
2. Choose **Analyze | Align to Reference...** or click the **Align to Ref.** icon on the analysis toolbar to open a new Align to Reference window.

The template sequence is displayed along the top of the Editor view in the new Align to Reference window.

3. Click the **Add Seqs** icon on the toolbar in the Editor view, select the cDNA sequence files you wish to assemble against your template then click the **Open** button.

Alternatively, select **Edit | Add Sequences From File** from the main menu, select the cDNA sequence files you wish to assemble against your template then click the **Open** button.

Note. You can hold down the **<shift>** key to select multiple sequences to import.

MacVector attempts to import all of the selected files into the assembly project. An error message is displayed listing the names of any files that could not be imported.

The Editor view updates to display the sequences in the order in which they were imported. If trace data is present, this is also displayed in a lower trace panel.

Note. Unaligned sequences are displayed in italics starting at residue “1” to alert you to the fact that they have yet to be aligned (or could not be aligned after automated assembly). Italicized sequences are not included in consensus calculations.

4. Click the **Align** icon on the toolbar in the Editor view.

The **Align to Reference Parameters** dialog box is displayed.

5. Select **cDNA Alignment** from the **Alignment Type** drop-down menu.
6. Click **OK** to perform the alignment using the default parameters.

See *“Alignment parameters”* on page 293 for more information about modifying the alignment parameters.

A dialog box is displayed, informing you of the progress of the alignment. You can close this and monitor progress using the Job Manager if you prefer. See Appendix , *“Using the Job Manager”* for more details about the Job Manager. There is also a **Stop** button which will cancel the alignment. The **Align to Reference** job runs in the background, allowing you to continue to use other functions in MacVector while the job is in progress.

Note. Align to Reference calculations can take some time if the genomic sequence is long and there are many sequences in the alignment.

7. Finally you can save the edited reference sequence in MacVector format by choosing **File | Save As...** from the Align to Reference window. You can save the actual sequence confirmation project itself at any time and you will be prompted to save any changes when you close the project window.

Alignment algorithms

During the first step of a **Sequence Confirmation** alignment, the algorithm finds all of the short sequences of perfect *Hash Value* matches and then extends each match left and right to the end of the perfect alignment. This “diagonal” is scored based on how many residues of perfect matches are found. The highest scoring diagonal is then extended past the mismatches at each end to get the alignment generating the best possible score. The *Sensitivity* setting determines how far ahead the algorithm looks to find the optimal alignment. An extension is terminated either when the end of one of the sequences is reached, or if the current score is *X Dropoff* below the best score.

The **cDNA Alignment** algorithm takes a more complex approach. It needs to deal with multiple segments, so instead of just taking the best diagonal, it keeps *Max. Diagonals* in an array. Each of these diagonals is extended separately to find the best possible score for each diagonal. The diagonals are then assembled into a segmented alignment by starting with the highest scoring diagonal, then adding the next best scoring diagonal that will fit on either side of it. The process continues through the sorted list of diagonals until all diagonals that can be fitted in the alignment have been accounted for. MacVector will also use the existence of various splice site consensus sequences commonly found around splice site junctions (for example, the GT...AG rule) to help confirm gaps being introduced into cDNA/mRNA sequences being aligned.

Note. The most CPU intensive part of the algorithm is the extension of the perfect diagonals, accounting for mismatches, insertions/deletions etc. With a large number of diagonals, this can take a substantial amount of time, especially if a high *Sensitivity* value is used.

The *Minimum Score* for a diagonal is the score which must be exceeded before a diagonal is saved. With the current default of 25, where the residue *Match* is 2, this means that at least 13 perfect residues are needed before a diagonal is saved.

The default alignment parameter values for **cDNA Alignments** are different to those for **Sequence Confirmation**. The *Hash* value is larger (6) so the algorithm needs at least 6 perfect matches before a diagonal is extended - this speeds the initial search somewhat. The *Sensitivity* is also reduced (3), to speed up the extensions. The *Mismatch* and *Gap Penalty* are higher, to decrease the likelihood of ragged ends. However, it should be noted that this can be disadvantageous if you are aligning trace files of cDNA clones against a genome and, in those circumstances, these values should be reduced to 3 or 4.

Alignment parameters

The following parameters have an impact on the alignment algorithm:

Match

(Valid range -100 to 100, default 2). This is the value the algorithm assigns to a match between a sample residue and the reference residue. It should typically be a positive value.

Mismatch

(Valid range -100 to 100, default -4). This is the value the algorithm assigns to a mismatch between a sample residue and the reference residue. It should typically be a negative value so that it reduces the cumulative match value as the algorithm extends the match between the sequences.

Ambiguous Match

(Valid range -100 to 100, default 0). This is the value the algorithm assigns to an ambiguous match between the sample and reference residues. The default is a neutral value – you can increase this to the match value if you want ambiguous matches to be treated exactly the same as matches. Alternatively, assigning it the same value as the mismatch parameter treats ambiguous matches as full mismatches.

Gap Penalty

(Valid range -100 to 100, default 3). This is the value the algorithm subtracts from the cumulative match score whenever it has to insert a gap character. Unlike some other algorithms, the SNP Assembly algorithm does not distinguish between gap insertions and gap extensions – all gaps are treated as a gap insertion.

Hash Value

(Valid range 1 to 6, default 4). This is the number of bases MacVector uses for the hashing algorithm. A value of 4 (the default) means that MacVector initially only searches for perfect 4 base matches between the reference and sample sequences. Larger values lead to faster searches, but with reduced sensitivity.

Sensitivity

(Valid range 1 to 8, default 4). This value affects what MacVector does when it is extending a match and encounters a mismatch. The value determines how far ahead the algorithm should look to determine whether to insert a gap in either of the sequences or to accept the mismatch. A value of 4 means it looks ahead in all directions until 4 additional mismatches (or the end of the sequence) has been encountered. The larger the value, the more likely the algorithm can handle longer regions of poor quality sequence, but performance will be much slower. The default (or even a smaller value) is quite adequate for most sequencing projects. Only increase this if you have long regions of poor quality sequence that is being poorly aligned using the default parameter.

Score Threshold

(Valid range 1 to 1000, default 50). This value controls how MacVector determines that a match is significant. After finding an initial match, MacVector attempts to extend the match in each direction using the match/mismatch and gap penalty scoring parameters. It retains the extended segment that gives the highest score. If the best score exceeds the score threshold, then MacVector considers this to be a significant match and includes the sample sequence in the alignment. If no individual match segment exceeds this score, the sample is treated as “unaligned” and will appear in the alignment view in italics. If you dramatically change the values for match/mismatch, you should alter this to keep it approximately in sync.

X Dropoff

(Valid range 1 to 1000, default 15). This value is used by MacVector to tell it when to give up extending a match. When extending, it keeps track of the best possible match score. It continues to extend the match in each direction, only giving up when the cumulative score falls to less than *X Dropoff* from the best score. If you reduce this to a low value (e.g. “0”) only perfect matches will be generated in the final alignment. You should typically set this to a value that will permit a few mismatches to be incorporated into the alignment to allow for sequencing errors. A smaller value will speed up calculations, but may not correctly align longer regions of poor quality matches.

Max. Diagonals (cDNA Alignment only)

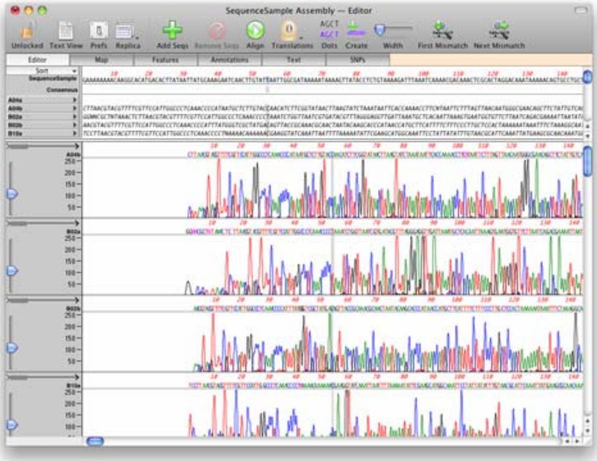
(Valid range 1 - 1000, default 50) This is the maximum number of diagonals that MacVector should keep and evaluate as potential exons during cDNA alignment. If this is set too low, MacVector may miss some short exon sequences. Setting this to a large value will lead to longer computation times.

Minimum Score (cDNA Alignment only)

(Valid range 1 - 1000, default 25) This is the minimum acceptable score for a diagonal to be saved. The default setting of 25, in conjunction with the default residue *Match* score of 2, means that a diagonal must span at least 13 perfect matches before it is saved for later evaluation.

Align to Reference window

The Align to Reference window is displayed whenever the **Align to Reference** analysis tool is launched.

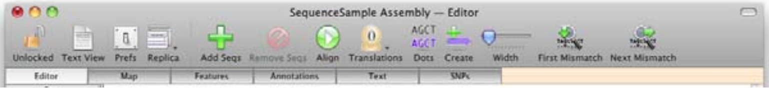


The Align to Reference window comprises a context dependent toolbar and six tabbed and linked views: Editor, Map, Features, Annotations, Text and SNPs. The functionality available in each view is described in the sections below.

Editor view

The Editor view is displayed by default when the **Align to Reference** tool is launched. To display it at any other time, click on the **Editor** tab in the Align to Reference window.

The Editor view has a toolbar containing tools that are used to perform the particular functions, as described below.



Common tools

These tools are included by default on all of the Align to Reference view toolbars:

The **Locked/Unlocked** indicator icon displays the current lock status of the Sequence file. Clicking the icon changes the lock status from locked to unlocked and back again. Sequences can only be altered when the sequence file is unlocked.

The **Text View** icon provides access to the Text view, which contains the sequence text with features marked along its length. The format of this window controlled by the **Text Display** preferences dialog box, accessible via **MacVector | Preferences** on the main menu. See *“Formatting the Text view”* on page 124 for more information about this dialog box. The information in the window cannot be changed, but it can be highlighted and copied to the clipboard.

Note. The **Text View** icon does not appear on the Align to Reference Text view or SNPs view toolbars.

The **Prefs** icon provides access to the preferences dialog box, which enables you to specify Align to Reference window preferences, including the line length used in the Text view.

The **Replica** icon is used to create a linked copy of the current sequence file in a new Align to Reference window.

Note. You can choose which view is displayed by default in the replica window by selecting the view name from the drop-down menu that appears when you click and hold down the **Replica** icon.

Editor view specific tools

These tools appear by default in the Editor view toolbar:

The **Add Seqs** icon provides access to the **Open** dialog, which you can use to add new sequences or trace files to the alignment. Newly added sequences are shown in italics to indicated that they have not yet been aligned.

The **Remove Seqs** icon removes the current sequence selection from the alignment.

The **Align** icon provides access to the Align to Reference Parameters dialog which you can use to specify alignment parameters and perform the alignment. See *“Alignment parameters”* on page 293.

The **Translations** icon toggles the display of the 3 or 6-frame translations of the reference seqeuence, shown immediately below the consensus line. The translation uses the genetic code specified on the **Genetic Code** dialog (see *"Selecting a different genetic code"* on page 242).

The **Dots** indicator icon shows which display mode is currently being used. Clicking the icon changes the display mode from normal to dots and back.

In normal mode, all sequences are displayed as expected. In dots mode, residues in the consensus and sample sequences that exactly match the reference sequence are shown as dots. This includes gap characters, but not unaligned sample sequences, spaces or "masked" regions where a sample sequence is not considered to be aligned to the reference. You can edit dots as with any normal residue – if you type a residue that matches the reference, that residue will be displayed as a dot.

The **Create** icon provides access to the **Feature Editor** dialog box which you can use to add new features to the sequence. Refer to *"Adding a feature"* on page 115, for further information.

The **Width** icon is a horizontal slider used to control the scale of the x-axis in any displayed traces. Drag the control with the mouse to increase or decrease the scale factor. The default value displays the traces at a scale of one pixel per sample point.

The **First Mismatch** icon selects and centers the display on the first mismatch between the reference and consensus.

The **Next Mismatch** icon selects and centers the display on the next mismatch between the reference and consensus.

Consensus calculation

The consensus line shown below the reference sequence in the Align to Reference **Editor** view is calculated according to the following rules:

For each consensus residue, the algorithm counts the number of overlapping A, G, C, T and gap residues. Scoring is as follows:

- A non ambiguous residue scores 12
- A two residue ambiguity scores 6
- A three residue ambiguity scores 4
- An 'N' scores 3 for each residue
- A gap scores 12 for gap

For example Y would score 6 for C and 6 for T whereas B (not A) would score 4 for C, G and T.

If there are no overlapping residues or gaps, then the consensus is set to a space. Otherwise, the required consensus threshold is calculated as the user defined threshold % multiplied by the maximum achievable score. So, if there are 10 overlapping residues and the threshold is set to 75%, then the threshold score is 0.75 * 120 = 90.

If any single residue or the gap score meets or exceeds the threshold, then that residue is chosen as the consensus.

If a single residue plus the gap score meets or exceeds the threshold, then the lower case of the residue is used as the consensus.

Otherwise, the scoring process is repeated with all the two-residue ambiguities. If a combination of two residues meets or exceeds the threshold, then that IUPAC ambiguity is used as the threshold. If two (or more) different two residue ambiguities exceed the threshold, but have the same score, then 'N' is used. Or the higher scoring ambiguity is returned.

If the highest scoring two residue ambiguity (not three residue ambiguities) plus the gap score meets or exceeds the threshold, then the lower case ambiguity is returned.

In all other circumstances, 'N' is returned.

Working with the Align to Reference Editor view

The main part of the Editor view is divided into two panels: an upper panel that displays the sequence residues of any loaded sequences and a lower panel that shows any loaded trace files.

Note. Where no trace files are present, the lower panel is hidden.

The Align to Reference Editor view is highly interactive. All of the sequence residues can be edited, although there are some restrictions that simplify the use of the editor for sequence confirmation and SNP identification purposes.

The numbering used for the reference sequence is the original numbering. Gaps are not included in the numbering, or in the locations shown in the features table, so that these are directly comparable to the original sequence.

You can edit the reference sequence at any time. To delete a base, simply type the **<space>** character. Visually, this converts it into a gap character but behind the scenes the base is physically deleted from the reference. If all of the sequences that overlap a position have gaps, the position will be closed up and the shared gaps will be deleted.

Note. All editing uses overwrite mode rather than insert mode. Hold down the **<option>** key to insert gaps or residues.

You can edit any of the sample sequences (in either the upper or lower pane) in a similar way. As you type, the consensus sequence is updated in real time. If you delete the last non-gap character at a position, the gap will close up as with the reference sequence.

You can copy any selection of more than one residue. There are important difference in the way the reference sequence is copied compared to the consensus or sample sequences:

- when the reference sequence has the primary highlight, the reference sequence and all overlapping features are copied to the clipboard. In this case, all gaps are stripped out so that you see the "real" ungapped sequence if you then paste the copied sequence into another window.
- when a sample sequence or the consensus is copied to the clipboard, only the actual sequence characters are copied, In addition, all gaps are retained in the sequence.

If you want to change the reference sequence to match the consensus sequence, select the region of the consensus that you want to copy. Choose **Edit | Copy**, carefully place the cursor on the first residue in the reference sequence that you wish to replace, then choose **Edit | Paste**. The first time you do this you will get a warning message, but when you click "Paste Anyway" the reference sequence residues will be replaced by the copied consensus sequence.

You can sort the sequences in the Editor view by position or alphabetically by sequence name using the **Sort** options above the sequence names in the top panel. Alternatively, you can re-order the sequences arbitarily, using drag and drop.

You can remove a sequence from the alignment by clicking on the sequence name button on the left to select the entire sequence and clicking the **Remove Seqs** icon on the toolbar or pressing the **<delete>** key.

Note. If you are editing very large assemblies, certain actions (e.g. deleting one or more sequences) may be slow, while a copy is made of the assembly to allow a later undo.

To delete a residue, overwrite it with either a **<space>** or a "-" character. The consensus will be updated dynamically. If all overlapping sequences have a gap at that location, the gap will be closed up.

Deletions at either end of a sequence are considered “true” deletions and the residues will actually be deleted, rather than being replaced by gaps.

The following keys may be used to edit the alignment:

- any IUPAC letter replaces the base
- any IUPAC + **<option>** inserts that base
- **<space>** or '-' replaces the selected base with a gap
- **<option>** + **<space>** or **<option>** + '-' inserts a gap. In the reference sequence, this will insert gaps into all of the reads sequences to maintain the alignment. If this is not what you desired, simply use the **<delete>** or **<backspace>** keys to delete the gaps in the read sequences.
- double-click on a read to select the entire sequence, then the left and right cursor keys and will "nudge" the read left or right. This only works for a selection of the entire sequence, and will not nudge exons or introns.
- **<delete>** or **<backspace>** physically deletes the selected bases.

Modifications to the Align to Reference Editor view

To simplify the use of the Align to Reference Editor view, the following restrictions apply:

- Although you can copy any number of residues, pasting is disabled for all sequences.
- You cannot perform a block selection of the sample sequences - only one sample sequence can have residues selected at any one time (although you can select multiple sample sequences to remove them from the assembly).
- You cannot edit the consensus sequence, it is always calculated dynamically. However, you can select and copy residues on the consensus line.

Colors and fonts in the Align to Reference Editor view

By default, MacVector displays all residues in Editor views in black Monaco 9 normal font. You can change this by choosing **MacVector | Preferences** from the main menu, then selecting the **Fonts** icon. The **Editor window font** section of this dialog box controls the font used for all of the different Editor views in MacVector.

In the Align to Reference Editor view, certain font variations are used to indicate status:

- Italics – unassembled sequences are shown in italics. They are always positioned at "1" in the alignment. All sequences that are added to an alignment are initially unassembled and will be shown in italics. After assembly, only those sequences that do not have significant alignments with the reference sequence are shown in italics.

Note. When italics are used, they always apply to the entire sequence. There are no circumstances in which only a subset of a sequence is displayed in italics.

- Gray text – This is used to indicate the regions where a sequence could not be aligned with the reference using the automated assembly algorithm. Grayed out residues are not considered in consensus calculations.
- Lower case text - This is used to indicate unaligned residues between 2 cDNA aligned exon segments. It is similar to gray text except that gray text is used at the ends of the aligned sequence, whereas lower case text is used for internal unaligned residues.

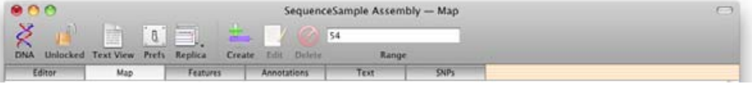
Map view

The Align to Reference Map view displays the features of the main sequence. It also shows aligned and unaligned reads. This view is connected to the Editor view so, clicking a feature on the Map view will highlight that sequence in the Editor view.

Tip. To select a single feature within a multi-segmented feature, hold down the **<option>** key whilst clicking on that feature.

The map can be edited in many ways to give a tailored map for on-screen or printed presentation. See *"Formatting the Map view"* on page 93, for further details.

The Map view has a toolbar containing tools that are used to perform particular functions. In addition to the common tools that appear by default on all of the Align to Reference toolbars (see *"Common tools"* on page 295), the Map view toolbar includes the following tools by default:



For nucleic acids, the sequence type indicator icon is either **DNA** or **RNA**. Clicking the icon changes how the sequences are interpreted from DNA to RNA and back again.

The **Create** icon provides access to the **Feature Editor** dialog box which you can use to add new features to the sequence. Refer to *“Adding a feature”* on page 115, for further information.

The **Edit** icon is enabled only when a feature is selected. It provides access to the **Feature Editor** dialog box which you can use to edit the details of the selected feature. Refer to *“Editing a feature”* on page 118, for further information.

Tip. The **Feature Editor** does not allow you to modify the appearance of the feature. Instead this is done using the **Symbol Editor** which can be accessed by double-clicking on the feature in the Map view.

The **Delete** icon is enabled only when a feature is selected. It is used to delete the selected feature.

The **Range** text box displays the number or range of the current selection within the displayed sequence. You can edit the values in the box to change the selection. Refer to *“Selection range”* on page 85, for further information.

Note. The Align to Reference Map toolbar, like all toolbars in MacVector, can be customized. Right-click on the toolbar to access this functionality. The tools described above appear in the default Align to Reference Map toolbar. Some of these tools may be absent and other tools may be present depending on your settings.

Features view

The Align to Reference Features view displays the features table associated with the reference sequence. It is exactly equivalent to the Features view in the Sequence window (see *“Features view”* on page 113). Any changes you make here will be saved in the assembly and the reference sequence, if it is saved as a MacVector single nucleic acid format file. Sample sequences added to the assembly are shown in the Features view as “Read” (assembled) or “Read*” (unassembled) features.

Annotations view

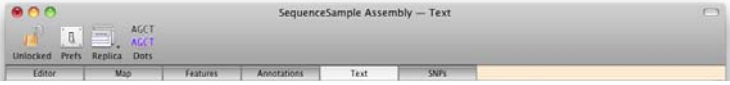
The Align to Reference Annotations view displays text annotations associated with the reference sequence. It is exactly equivalent to the Annotations view in the Sequence window (see *“Annotations view”* on page 119). Any changes you make here will be saved in the assembly and the reference sequence, if it is saved as a MacVector single nucleic acid format file.

Text view

The Align to Reference Text view shows the sequence text for the entire alignment. You can customize the appearance of the Text view, both as it appears on the screen and as it will be printed using the **Text Display** preferences dialog box (see *“Formatting the Text view”* on page 124).

Note. The line length used in the Text view is set on the **Align to Reference Preferences** dialog. accessed by clicking the **Prefs** icon on the Align to Reference window toolbar.

The Text view has a toolbar containing tools that are used to perform particular functions. In addition to the common tools that appear by default on all of the Align to Reference toolbars (see *“Common tools”* on page 295), the Text view toolbar includes the following tools by default:

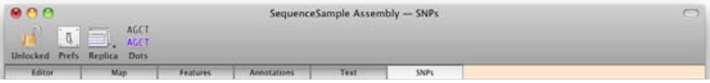


The **Dots** indicator icon shows which display mode is currently being used. Clicking the icon changes the display mode from normal to dots and back. See *“Editor view specific tools”* on page 296 for further details.

SNPs view

The SNPs view lists all of the potential SNPs that are present in the reference sequence and aligned reads.

The SNPs view has a toolbar containing tools that are used to perform particular functions. In addition to the common tools that appear by default on all of the Align to Reference toolbars (see *“Common tools”* on page 295), the SNPs view toolbar includes the following tools by default:



The **Dots** indicator icon shows which display mode is currently being used. Clicking the icon changes the display mode from normal to dots and back. See *“Editor view specific tools”* on page 296 for further details.

Probable and Possible SNPs in the Assembled Reads

Initially, any mismatches between residues in each aligned read and the reference sequence are considered as potential SNPs. If the read does not contain trace data then all mismatches are considered to be possible SNPs. However, if the read contains trace data then the putative SNP is scored by comparing the area under the curve for the called base versus the total area under the curves for all 4 bases. The thresholds are set to 90% for probable and 75% for possible. If the score is below 75% it is regarded as a low quality base call, rather than a SNP. So, if "A" accounts for 90% or more of the signal then it is a probable SNP or if it contains more than 75% it is a possible SNP.

These are listed at the bottom of the SNP tab below the SNP analysis of the reference sequence.

Probable SNPs in the Reference Sequence

For the reference sequence SNPs are grouped according to whether they are possible or probable. All SNPs identified in a read are shown, even if only a single read contains that SNP. A SNP is regarded as probable if one or more reads contains a probable SNP at that point.

Note. The presence of a SNP that has been identified in a read without trace data means it will be regarded as a probable SNP, even if any corresponding SNPs in trace files are only possible SNPs.

Navigation

Scroll Bars

There are three scroll bars in the Align to Reference window. Each will respond interactively to drags on the “thumb” and will scroll the appropriate display incrementally when clicked on the arrows or in the “page scroll” region.

- Alignment Vertical – this scroll bar lets users scroll vertically through the sample sequences that have been added to an assembly. The reference and consensus sequences always stay at the top of the alignment pane. Because each sample sequence has its own line in the display, you may see a lot of “white space” while scrolling through a large assembly.
- Multiple Trace Vertical – this scroll bar lets users scroll vertically through the trace sample sequences that overlap the current Alignment view. This is typically only a subset of the available sample sequences.

- Horizontal – this scroll bar lets users scroll horizontally through the entire alignment.

Scaling Controls

These controls affect the way the trace panes are displayed. There are two types:

- Width Control – this is a single control on the toolbar. When dragged, it interactively adjusts the horizontal scaling factor of the trace displays. The default value displays the traces at a scaling factor of one pixel per sample point.
- Vertical Gain – each individual trace pane has a vertical slider in the left hand pane that controls the vertical scaling of the traces. The default is that the tallest peak in the trace is “100%”. When dragged, it interactively adjust the vertical scaling factor of the trace so that users can “zoom in” to more closely examine areas of ambiguity.

Miscellaneous

Cut/Copy/Paste

Cut is always disabled. Copy is active only when a selection is present and acts on the currently selected sequence as follows;

- If the reference sequence is highlighted, the selected residues in the reference are copied to the clipboard as a nucleic acid sequence. Any overlapping features are NOT copied.
- If the consensus sequence is highlighted, the selected residues in the consensus are copied to the clipboard as a nucleic acid sequence. Any overlapping features are NOT copied.
- If a sample sequence is highlighted (and the reference is not), the selection in that sample sequence is copied to the clipboard as a nucleic acid sequence.

Paste is enabled only when the reference sequence is selected. When you paste into the reference, the residues on the clipboard will *overwrite* the reference sequence, NOT *insert*. You can use this feature to copy and paste the consensus sequence into the reference.

Undo

Most operations in the Align to Reference window can be undone, as with other MacVector windows. In particular, you can undo an assembly operation, and also adding/deleting sequences.

Using the Find options

Standard Find

Use the **Edit | Find** menu option to display the standard **Find** dialog. This lets you find matching residues in the reference sequence. See *"Searching sequences"* on page 47 for further details.

Note. Any gaps in the reference sequence are ignored during the search, so searching for "GATC" will find a match with the sequence "G-AT--C".

Mismatch Find

Use the Mismatch tools to find mismatches between the reference sequence and the consensus. Click the **First Mismatch** icon on the Align to Reference Editor view toolbar to select and display the first mismatch between the reference and consensus. Click the **Next Mismatch** icon to select and display the next mismatch between reference and consensus.

Saving and loading alignments

File format details

Alignment files are BSML (Biological Sequence Markup Language) XML files. The file contents can be viewed and edited with any standard text editor. The contents of the file largely follow the standard BSML format with some variations as appropriate. There have been a few changes to support sequence assembly:

- The "model-type" is "assembly" rather than "sequence"
- Each sequence in the alignment is stored in standard BSML/DS Gene format – however, the "ID" field of each sequence has specific meaning. The edited reference sequence has ID "REFERENCE_SEQ", the consensus is "CONSENSUS_SEQ" and each component is labeled "SEQ_0", "SEQ_1" etc. A copy of the original sequence, along with all features, annotations and feature appearance information is stored in "ORIGINAL_SEQ".
- Each component sequence has an "interval-loc" entry that defines its location and strand on the alignment. This is a standard BSML model, but is usually associated with features rather than sequences.

- Each component sequence has additional attributes;
  - o “IsAssembled” – “0” for no, “1” for yes
  - o “ClipLeft” – the extent of the masking at the 5’ end of the sequence.
  - o “ClipRight” – the start of the masking at the 3’ end of the sequence.
- There are additional attributes stored at the “definitions” level describing the parameters used in the assembly. Most correspond directly to parameters in the assembly dialog;
  - o “GapPenalty”
  - o “MismatchScore”
  - o “MatchScore”
  - o “AmbiguousScore”
  - o “HashValue”
  - o “ScoreThreshold”
  - o “XDropOff”
  - o “Recursion” – this is shown in the dialog as “sensitivity”
  - o “Threshold”
- There are additional attributes stored at the “definitions” level describing the current settings of the Align to Reference window;
  - o “ShowDots” – “0” for no, “1” for yes – determines if matches to the reference are shown as dots or standard residue characters.

Saving alignments

The user can save the alignment at any time. The usual rules apply:

- **Save As...** is always enabled. The user can choose to save the alignment under a different name or in a different format.
- **Save** is only enabled if the alignment has been edited since the last save. It is also enabled for new alignments, but will bring up the **Save As...** dialog as a valid file name is not initially specified.

Saving the modified reference sequence

The user can choose **Save As...** at any time and choose **MacVector NA Sequence** format from the **Format** popup menu. This saves the reference sequence (complete with any annotations and features) in standard MacVector single sequence format.

Loading alignments

The user can open an existing alignment at any time using the **File | Open...** command. Currently, alignments files can only be viewed/ selected if **All Documents** or **All Readable Documents** is selected in the **Enable** popup menu. After selection, the alignment will load and be positioned at the start of the alignment. The status of the **show dots** toolbar button is remembered, but all other display information is reset to the default.

17 Aligning Multiple Sequences

Overview

This chapter describes the features of the Multiple Sequence Alignment (MSA) window, and how to use it with nucleic acid and protein sequence alignments. It describes how to import and export multiple sequence files, how to create an empty MSA document and insert sequences into it. It also describes how to perform multiple sequence alignments using the ClustalW, T-Coffee and Muscle algorithms and how to edit and display alignment results.

For information on phylogenetic analysis, refer to *“Phylogenetic reconstruction”* on page 358.

Contents

Multiple Sequence Alignments

Multiple sequence alignments may come from one of several sources. MacVector can generate a multiple sequence alignment automatically, by means of the ClustalW, T-Coffee or Muscle alignment algorithms. Alternatively, you can open a pre-calculated alignment which has been generated by another program such as the GCG Wisconsin Package PileUp command. You can also create a completely new alignment, by importing sequence data into a new Multiple Sequence Alignment window, or typing in the residues manually.

There are very few differences between an alignment which has been generated automatically and one that has been imported or created manually. In all cases, the consensus line will be shown correctly and a phylogenetic tree can be generated.

Note. The guide tree can only be displayed if a ClustalW or T-Coffee analysis has been performed (see *"Displaying a guide tree"* on page 354).

Multiple Sequence Alignment window

The Multiple Sequence Alignment (MSA) window is displayed whenever a multiple sequence alignment document is opened or created. It comprises a context dependent toolbar and seven tabbed and linked views of the multiple sequence alignment: Editor, Text, Pairwise, Matrix, Picture, Guide Tree and Profile. The functionality available in each view is described in the sections below.

All of the MSA tabs are visible in all MSA windows, irrespective of whether the corresponding views have been calculated. However, if a view has not been calculated, then the tab is greyed out and is non-clickable.

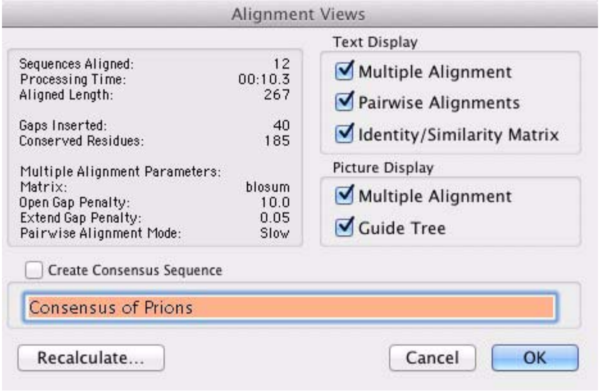
The **Alignment Views** dialog box enables you to generate any missing views, after you have performed an alignment. You can choose from the following output options:

- multiple alignment text display
- pairwise alignment text display
- identity/similarity matrix text display
- multiple alignment picture display
- guide tree (for ClustalW and T-Coffee alignments only; see *"Displaying a guide tree"* on page 354)

- a consensus sequence

In addition, this dialog box includes an information panel. If you have performed a multiple sequence alignment, a summary of the results and parameters is provided here. The **Recalculate** button lets you perform (or repeat) the alignment (see *"Performing Multiple Sequence Alignment"* on page 341).

Note. The settings and parameters reported in the information panel are only valid following a multiple sequence alignment.



Alignment Views

Sequences Aligned: 12
Processing Time: 00:10.3
Aligned Length: 267

Gaps Inserted: 40
Conserved Residues: 185

Multiple Alignment Parameters:
Matrix: blosum
Open Gap Penalty: 10.0
Extend Gap Penalty: 0.05
Pairwise Alignment Mode: Slow

Text Display
☑ Multiple Alignment
☑ Pairwise Alignments
☑ Identity/Similarity Matrix

Picture Display
☑ Multiple Alignment
☑ Guide Tree

□ Create Consensus Sequence

Consensus of Prions

To display the **Alignment Views** dialog box, click the **Views** icon on the MSA window toolbar. This dialog box is also displayed whenever a multiple sequence alignment is performed.

The data in the different views is updated in real time but it is refreshed only when a new view is displayed after the underlying alignment has been edited. To see this, open an MSA document and click the **Replica** icon on the MSA window toolbar to create a duplicate MSA window. Display the MSA Editor view in one window and the MSA Text view in the other. If you edit the alignment in the Editor view, then nothing will happen to the Text view in the other window. However, if you click on the second window, to make it active the Text view is refreshed to reflect the changes you made it the Editor view. The same will happen with the other views.

This is different to the way the views in the Sequence window are updated. The reason for the difference is that some of Text views can take a considerable time to be refreshed. So if they were updated every time you typed a character, updating the views could be very slow, preventing you from being able to edit the alignment easily. This approach ensures the views are only updated when you are ready to look at them.

Managing multiple sequence alignments

Opening multiple sequences

You can open multiple sequences from various sources:

- Several single-sequence files
- One multiple-sequence file
- Several multiple-sequence files, or a combination of single- and multiple-sequence files.

When you open multiple sequences from any of these sources, you may want to open them in individual sequence windows, or in a single MSA window.

The default behavior of MacVector is to use the following guidelines when opening one or more sequence files:

File type	**How the files are opened**
MacVector, GCG RSF, ASCII/ Plain	Individual sequence windows
GenBank, EMBL/Swiss-Prot or NBRF files, containing single or multiple sequences	Individual sequence windows
NEXUS, GCG MSF, PHYLIP or FastA (MacVector multiple sequence) files, containing single or multiple sequences	MSA windows, one for each input file

You can override the defaults and require either that all sequences are opened in individual windows, or that all sequences are opened in a single MSA window. In the **File | Open** dialog box, the **Open Multiple Sequence Files As** drop-down menu allows you to make this choice.

When **Multiple Alignment** is chosen, all sequences are opened in a single MSA window. If the sequences are unaligned and are of unequal length, gap symbols are appended to the end of the shorter sequences to make the lengths equal.

When **Single Sequences** is chosen, all sequences are opened in individual sequence windows.

To open multiple sequences from one or more files

1. Choose **File | Open**.

The **Open** dialog box is displayed.

2. Select the sequence files you want to open.
3. Select the **Open Multiple Sequences As** arrow button and choose an option from the drop-down menu:
   - **Auto**: MacVector will use the default guidelines to open the files
   - **Multiple Alignment**: All sequences will be opened in a single MSA window
   - **Single Sequences**: All sequences will be opened in individual sequence windows.
4. Select **Open** to open the sequence files.

Creating new multiple sequence alignments

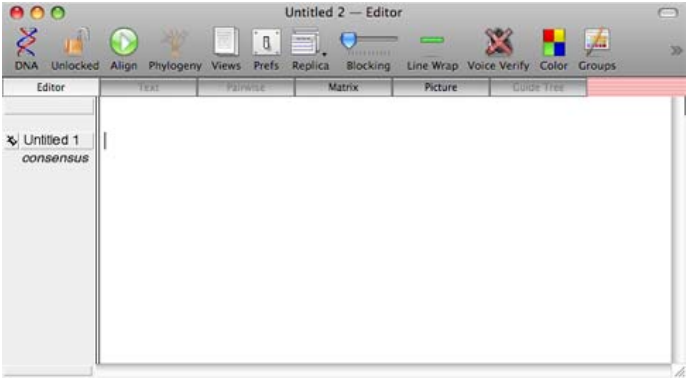
You can create a new MSA document in MacVector. This is particularly useful if you want to import aligned sequences from databases or web pages, by cutting and pasting from other applications. You may also want to type sequences into an empty MSA document.

Creating an empty MSA document

To create an empty MSA document

1. From the **File** menu, choose either **New | Nucleic Acid Alignment** or **New | Protein Alignment**.

An MSA window of the appropriate type is displayed.



Inserting sequences into an empty MSA document

You can either type a sequence into the new window, or paste residues or full sequences from the clipboard. You can also add sequences from files.

To type a sequence in an empty MSA document:

1. Ensure that the **Editor** tab is selected and type in the required sequence, using the spacebar to indicate any gaps.

When you are working with nucleic acid sequences, the **DNA/RNA** icon on the toolbar indicates whether you should type in DNA or RNA codes,

Note. It is not possible to type in both DNA and RNA codes in the same MSA window.

As you type the sequence, the sequence numbering will appear over the line at the set intervals.

Note. You can program the numeric keypad for entering sequences. See *“Configuring the numeric keypad”* on page 45.

To paste a sequence into an empty MSA document

1. Ensure that the **Editor** tab is selected and choose **Edit | Paste**.

The clipboard contents will be inserted at the cursor position, as follows:

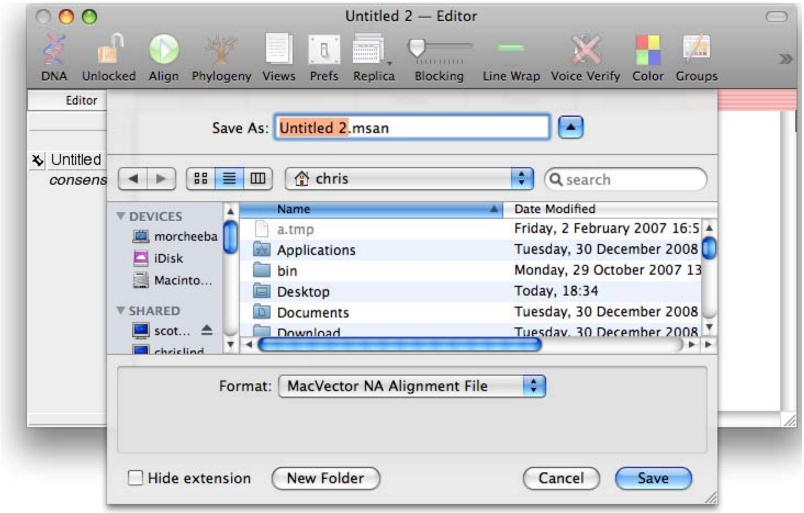
- If one or more full sequences have been selected for copying (see *“Selecting sequences in the MSA Editor”* on page 331), they will be inserted as new sequences

- If, in a single sequence, only residues have been selected (i.e. not the sequence name), they will be pasted into the untitled sequence at the insertion point
- If the selection is a block extending over several sequences, it will be inserted as new sequences

Saving multiple sequence alignments

You can save multiple sequence alignments in the MSA window in any of the supported multiple sequence formats. See Appendix G, *“Supported File Formats and File Extensions”* for a complete list of the multiple sequence formats MacVector supports.

For all formats except PHYLIP, you save MSA documents in exactly the same way as single sequences. The **File | Save As** dialog box for MSA documents looks like this:



Save As: Untitled 2.msan

Name	Date Modified
a.tmp	Friday, 2 February 2007 16:5
Applications	Tuesday, 30 December 2008
bin	Monday, 29 October 2007 13
Desktop	Today, 18:34
Documents	Tuesday, 30 December 2008
Download	Tuesday, 30 December 2008

Format: MacVector NA Alignment File

The **Options** button is only enabled when **Format** is set to **PHYLIP**.

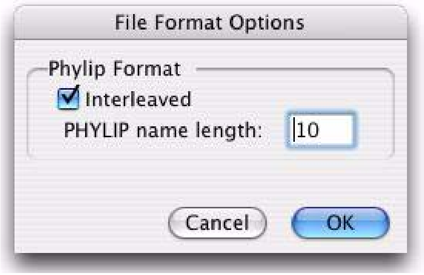
Saving sequences as PHYLIP files

When saving aligned sequences as a PHYLIP file, you need to set the name length (the length of the field containing the sequence names). If the value you enter is less than the length of any of the current sequence names, they will be truncated. In cases where truncation would result in two or more sequences being given the same name, MacVector will append numbers to the truncated name to distinguish the different sequences.

Before saving as a PHYLIP file, you also need to indicate whether the sequences should be interleaved or non-interleaved.

To save a multiple aligned sequence as a PHYLIP file

1. Ensure the required MSA window is active.
2. Choose **File** | **Save As** to display the **Save As** dialog box.
3. If required, navigate to the required folder, or create a **New** folder.
4. If required, change the file name by typing in the **Name** text box.
5. Choose **PHYLIP** from the **Format** drop-down menu. The **Options** button becomes active.
6. Select **Options**. The **File Format Options** dialog box is displayed.



File Format Options

Phylip Format

☑ Interleaved

PHYLIP name length: 10

7. Set the interleaving control:
   - If you want the sequences to be interleaved, ensure that the check box is selected.
   - If you want the sequences to be non-interleaved, ensure that the check box is not selected.
8. Type the required name length in the **PHYLIP name length** text box.
9. Select **OK** to apply the settings and return to the Save As dialog box.
10. Select **Save** to save the file.

File extensions for aligned sequence files

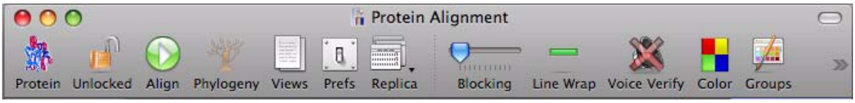
When you save an aligned sequence file using the Save As dialog box, MacVector will by default add a filename extension appropriate to the format you select. See Appendix G, *“Supported File Formats and File Extensions”* for a complete list of the file extensions used by MacVector.

These file extensions will be particularly useful when files are used with other operating systems which use extensions to identify the file type.

MSA Editor view

The MSA Editor view is displayed by default whenever a multiple sequence alignment is opened or created. To display it at any other time, click on the **Editor** tab in the MSA window.

It is used for modifying and comparing the alignment of several selected sequences. It can be used to edit individual sequences in multiple alignments, and to delete parts of sequences and import other parts.



The MSA Editor has a toolbar containing icons that are used to perform the following functions:

Note. The MSA Editor toolbar, like all toolbars in MacVector, can be customized. Right-click on the toolbar to access this functionality. The tools described below are those that appear in the default MSA Editor toolbar. Some of these tools may be absent and other tools may be present depending on your settings.

For protein alignments, the sequence type indicator icon is **Protein**. For nucleic acid alignments, the sequence type indicator icon is either **DNA** or **RNA**. Clicking the icon changes how the sequences are interpreted from DNA to RNA and back again.

The **Locked/Unlocked** indicator icon displays the current lock status of the MSA document. Clicking the icon changes the lock status from locked to unlocked and back again. Sequences can only be altered when the MSA document is unlocked.

The **Align** icon provides access to the various **Alignment** dialog boxes which enable you to align multiple sequences using the ClustalW, Muscle or T-Coffee algorithms. Click and hold the icon to select an alignment algorithm. Click the icon to access alignment parameters for the algorithm that was used last. See *“Performing Multiple Sequence Alignment”* on page 341.

The **Phylogeny** icon provides access to the **Phylogenetic Reconstruction** dialog box. This button becomes enabled when the MSA editor contains at least four sequences. See *“Phylogenetic reconstruction”* on page 358.

The **Views** icon provides access to the **Alignment Views** dialog box. See *“Multiple Sequence Alignment window”* on page 310 for more information about this dialog box.

The **Prefs** icon provides access to the **Multiple Alignment Options** dialog box. See *“Setting the MSA Editor display options”* on page 319 for more information about this dialog box.

The **Replica** icon is used to create a copy of the current MSA document in a new MSA window.

The **Blocking** icon is a horizontal slider used to control the number of residues in a block. Drag the control with the mouse to change the number of residues in a block from 1 to 10.

The **Line Wrap** indicator icon displays the current line wrap display style of the MSA document. Clicking the icon changes the display style from linear mode to page mode and back again.

In linear mode, each sequence is displayed on a single horizontal line, and sequences are aligned on top of each other. You use the horizontal scroll bar to scroll through the entire sequence.

In page mode, the linear display is wrapped to fit into the window, as in a word processor. You use the vertical scroll bar to scroll through the sequence.

The **Voice Verify** indicator icon displays the current voice verification status of the MSA document. Clicking the icon changes the voice verification status from off to on and back again. When voice verification is on, each residue you enter is spoken by a computer voice. Refer to *“Map view”* on page 92, for further information.

The **Color** indicator icon displays the current color display style of the MSA document. Clicking the icon changes the display style from color to black and white and back again.

In black and white mode, sequences are displayed as black text in a monospaced font on a white background. In color mode, sequences are displayed as black text with residue-specific color backgrounds. To help visualize similarities during manual alignments, each amino acid or nucleotide can be assigned a unique color. Furthermore, groups of amino acids can be given colors to permit visual alignments based on properties. The color groups can be edited using the **Color Group Editor** dialog box.

The **Groups** icon provides access to the **Color Group Editor** dialog box. This allows you to edit the colors and similarity groups used to calculate the consensus.

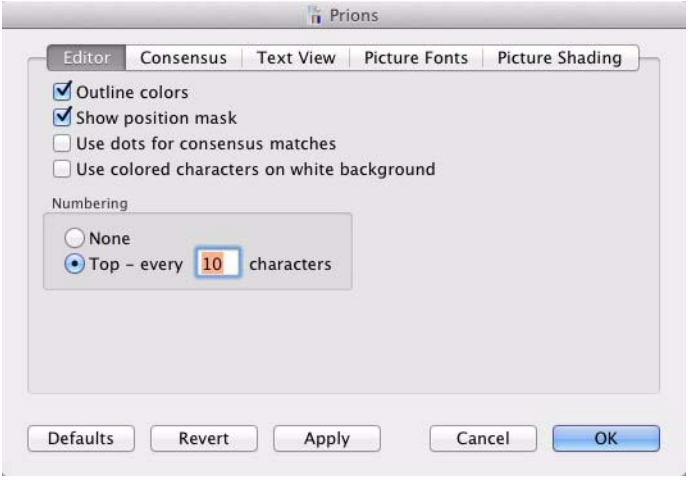
The **Width** icon is a horizontal slider used to control the width of the displayed residues when **Color** mode is enabled. Drag the slider to the left to reduce the width of the residues, and to the right to increase the width. If the width you set is less than the width of the default font, no characters are displayed and the residues are represented as color bars only.

Position mask

A position-masking bar can be displayed across the top of the aligned sequences. It is used in phylogenetic analysis, to exclude positions in the sequence from the analyses. For instructions, refer to *"Position masking"* on page 358.

Setting the MSA Editor display options

You can modify the appearance of the MSA window. Controls for all multiple alignment preferences are grouped together in the **Multiple Alignment Options** dialog box, which you can access using the **Prefs** icon on the MSA Editor toolbar.



Prions

Editor | Consensus | Text View | Picture Fonts | Picture Shading

☑ Outline colors
☑ Show position mask
□ Use dots for consensus matches
□ Use colored characters on white background

Numbering

○ None
● Top - every 10 characters

In the **Multiple Alignment Options** dialog box, settings are organized into categories, accessed by selecting tabs at the top of the dialog box. General MSA Editor display options are set using the **Editor** tab. For information about setting the options on the **Consensus** tab, see *"Calculating the consensus line"* on page 320. For information about setting the options on the **Text Display**, **Picture Fonts** and **Picture Shading** tabs, see *"Multiple Sequence Alignment window"* on page 310.

To change the Editor settings in the MSA window

1. Click the **Prefs** icon on the toolbar.

The **Multiple Alignment Options** dialog box is displayed, with the **Editor** tab selected.

2. To control sequence numbering, do one of the following:
   - select the **None** radio button if you do not want residue numbers to be displayed.
   - select the **Top** radio button to display residue numbers across the top of the alignment; then use the text box to set the interval at which numbers will be displayed (default is **10**).
3. Select the **Outline colors** check box to display each residue with a dark outline, for greater contrast.

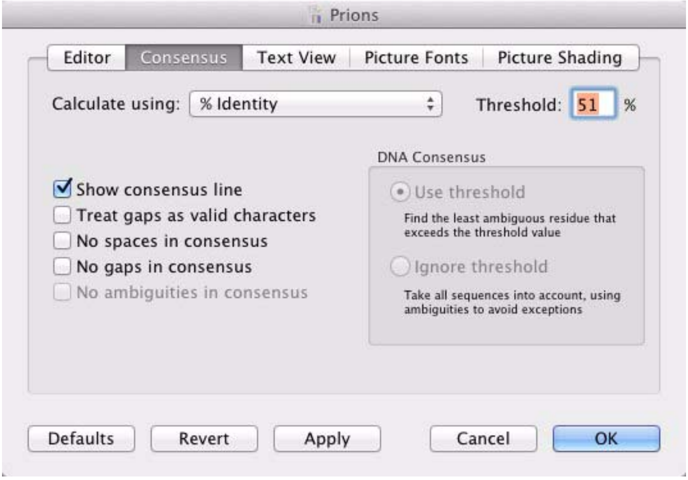
Note. The color display style must be enabled for this option to work. To toggle the color display style, click on the **Color** icon on the toolbar.

4. To allow regions of the aligned sequences to be excluded from phylogenetic reconstruction analyses, select the **Show position mask** check box.
5. Select **Apply** to see the effects of your changes on any open MSA windows.
6. Select **OK** to save the changes and close the dialog box.
7. Alternatively, select **Defaults** to reset the default settings, or **Cancel** to close the dialog box without saving.

Note. The font used in the MSA Editor view can be controlled using the MacVector preferences dialog box which is accessed by choosing **MacVector | Preferences** from the main menu, then clicking the **Fonts** icon on the preferences dialog

Calculating the consensus line

Underneath an aligned sequence, you can display a line of residue codes to show the nature and extent of their consensus. To set the consensus calculation options, use the **Consensus** tab on the **Multiple Alignment Options** dialog box.



Calculate using: % Identity

Threshold: 51 %

☑ Show consensus line

□ Treat gaps as valid characters

□ No spaces in consensus

□ No gaps in consensus

□ No ambiguities in consensus

DNA Consensus

● Use threshold

Find the least ambiguous residue that exceeds the threshold value

○ Ignore threshold

Take all sequences into account, using ambiguities to avoid exceptions

MacVector offers three basic modes for calculating the consensus line:

- Consensus Identity
- % Identity
- Property-based color groups, of which ClustalW default groups is a special case.

These modes are discussed in *“Working with color groups”* on page 324.

For protein sequence alignments, the default options for consensus calculations are:

- ClustalW default groups
- Consensus Identity
- % Identity

and the property-based color groups:

- Functionality
- Alpha-helix
- Alpha-helix + P450
- Hydrophobicity + Charge
- Acidity + Basicity

- Smooth Scaling
- Structural Position
- Steric Bulk
- Gascuel & Golmard
- Chou-Fasman Alpha Helix
- Chou-Fasman Beta Sheet
- Levitt Alpha Helix Forming
- Levitt Beta Sheet Forming
- Levitt Turn Forming
- Dayhoff Matrix
- Helix Termintors
- Chemical Type

For nucleotide alignments, the corresponding default options are:

- % Identity
- Consensus Identity
- DNA Residues

User-defined color groups will also appear in these menus.

Use threshold mode

MacVector offers a choice of two methods to deal with conflicting nucleic acids in the consensus calculation: **Use threshold** and **Ignore threshold**. The **Use threshold** option generates a consensus residue code for each position based on the highest percentage agreement that can be found between the aligned sequences. For example, if in 10 aligned sequences there are 9 Gs and one C, there is 90% agreement with a consensus code of G. The percentage agreement must exceed a minimum threshold value. Use the **Threshold** text box to set this value, within the range 51-100%.

If the consensus threshold is not met by any single-base code, MacVector then tests all 2-base ambiguities against the criterion (see *"IUPAC-IUB codes for nucleotides and amino acids"* on page 486). For example, if the separate percentages of G and C fail the criterion, but their combined percentage meets it, the consensus line will indicate S at that position. If no 2-base ambiguity meets the criterion, MacVector then tests all 3-base ambiguities. If all these fail, the consensus residue at that position will be N.

In Threshold mode, ambiguous base codes in sequences are treated as equal probabilities of the alternative bases (e.g. an S residue contributes 0.5 C and 0.5 G to the consensus base for its position).

Ignore threshold mode

When the **Ignore threshold** option is selected, the residue code on the consensus line must match every residue in that column, so that ambiguous residue codes may be required. For example, if in 10 aligned sequences there are 9 Gs and one C, the consensus code is S. The Threshold text box is disabled when you select this option, since the effective threshold is fixed at 100%.

For aligned protein sequences, the threshold method is always used to generate the consensus code.

Treatment of gaps in sequences

The **Treat gaps as valid characters** check box lets you choose whether to treat gaps as valid characters in the consensus calculation, or to ignore them. If gaps are valid characters, then the consensus line will show a gap wherever the percentage of gaps exceeds the threshold. For example, if the threshold is set at 51% and there are 11 aligned sequences, with 6 gaps, 3 Gs and 2 Cs, then the consensus code for that position will be “-” (gap), with 55% agreement.

If gaps are ignored, agreement is calculated as a percentage of the remaining total. In the example above, the consensus code will be G, with 60% agreement.

Displaying the consensus line

You can display the consensus line in the Editor, Text and Picture views in the MSA window. To set the consensus display options, use the **Consensus** tab on the **Multiple Alignment Options** dialog box.

Appearance of the consensus line

The **Show consensus line** check box lets you display the consensus line. It will appear in the MSA window in the Editor, Text and Picture views.

The **No spaces in consensus** check box lets you replace undefined characters in the consensus by N (in nucleotide windows) or X (in protein windows).

The **No gaps in consensus** check box allows you to replace the gap character "-" in the consensus line with a space.

The **No ambiguous in consensus** check box is used only with nucleotide alignments. It allows you to display only unambiguous residue codes (A, G, C, T or U) on the consensus line.

To change the Consensus settings in the MSA window

1. Click the **Prefs** icon on the toolbar.

The **Multiple Alignment Options** dialog box is displayed.

2. Choose the **Consensus** tab.
3. To display a consensus line in the MSA window, select the **Show consensus line** checkbox.
4. Set the required calculation mode for generating the consensus, by choosing from the **Calculate using** drop-down menu.
5. Set the threshold criterion, if required, by typing a value in the **Threshold** text box.
6. For nucleotide alignments, choose a **DNA Consensus** method for dealing with conflicts in the calculation:
   - select the **Ignore threshold** radio button if you want the consensus residue code to match all residues in the column
   - select the **Use threshold** radio button to calculate the consensus using percentage agreement and the **Threshold** criterion.
7. If required, select the **Treat gaps as valid characters** check box.
8. If required, select the **No spaces in consensus** check box.

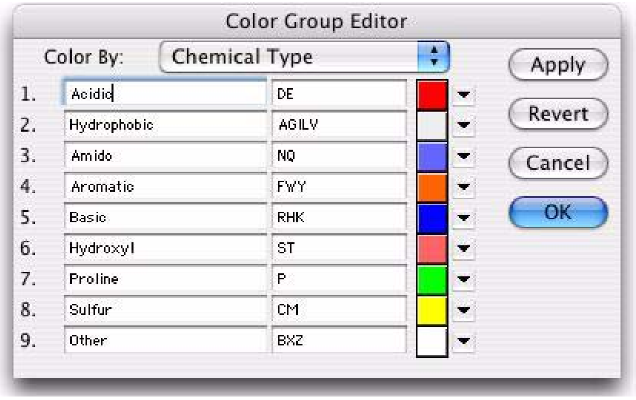
Note. This option is not applicable in **Ignore threshold** mode, because spaces are never present.

9. If required, select the **No gaps in consensus** check box.
10. Select **Apply** to see the effects of your changes on any open MSA windows.
11. Select **OK** to save the changes and close the dialog box.
12. Alternatively, select **Defaults** to reset the default settings, or **Cancel** to close the dialog box without saving.

Working with color groups

For the color display, you can specify the color for each nucleic acid base or amino acid residue. You can assign amino acid residues to color groups according to their chemical properties. Alternatively, residues can be colored according to degree of consensus among the sequences at each position. In addition, you can create and apply your own color schemes.

To choose or edit a color group use the **Color Group Editor** dialog box, which you can access using the **Groups** icon on the MSA Editor toolbar.



Color By: Chemical Type

1.	Acidic	DE	🟥
2.	Hydrophobic	AGILV	⬜
3.	Amido	NQ	🟦
4.	Aromatic	FWY	🟧
5.	Basic	RHK	🟦
6.	Hydroxyl	ST	🟥
7.	Proline	P	🟩
8.	Sulfur	CM	🟨
9.	Other	BXZ	⬜

The **Color Group Editor** dialog box differs slightly between protein and nucleotide sequence alignments, with some different **Color By** options available in each case.

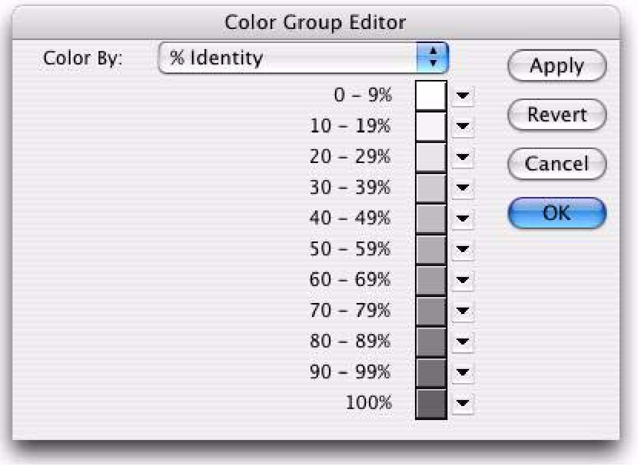
Consensus Identity and % Identity color groups

These groupings are based on the degree of consensus among the aligned sequences at each position.

The **Consensus Identity** option uses three colors:

- **100% Identity** - where all the sequence residues match the consensus, then the 100% Identity color is used for the entire column
- **Consensus Match** - where one or more residues in a column do not match the consensus, those that do match will be displayed in this color
- **Mismatch** - this color is used for all the residues that do not match the consensus.

The **% Identity** option colors residues according to the percentage of residues at a given position that share the same symbol. The color displayed depends on the residues in the other sequences, not on the calculated consensus.



Color Group Editor

Color By: % Identity

0 - 9%	□
10 - 19%	□
20 - 29%	▨
30 - 39%	▨
40 - 49%	▨
50 - 59%	▨
60 - 69%	▨
70 - 79%	▨
80 - 89%	▨
90 - 99%	▨
100%	▨

Property-based color groups

When there is a protein alignment in the MSA Editor view, you can select a range of preset color groups that are based on chemical or physical properties of the residues. The **Color By** drop-down menu includes the following property options:

- Chemical Type
- Functionality
- Alpha-helix
- Alpha-helix + P450
- Hydrophobicity + Charge
- Acidity + Basicity
- Smooth Scaling
- Structural Position
- Steric Bulk
- Gascuel & Golmard
- Chou-Fasman Alpha Helix
- Chou-Fasman Beta Sheet
- Levitt Alpha Helix Forming
- Levitt Beta Sheet Forming

- Levitt Turn Forming
- Dayhoff Matrix
- Helix Termintors

The default option for protein alignments is **Chemical Type**. When you choose a color group from the menu, the corresponding groups are displayed in the **Color Group Editor** dialog box.

ClustalW Default Groups

The default property groups used by ClustalW alignment are included as a **Color By** option for protein alignments. These groups are unusual in that a given amino acid can appear in more than one group, and therefore could be represented by more than one color. To simplify the MSA editor view, these groups are always shown in white.

To specify individual residue colors

1. With nucleotide or protein sequences displayed in the MSA window and the **Editor** tab selected, click the **Groups** icon.

The **Color Group Editor** dialog box is displayed.

2. If you are using DNA sequences, ensure that **DNA Residues** is selected in the **Color By** drop-down menu.
3. If you are using protein sequences, select **Smooth Scaling** from the **Color By** drop-down menu.
4. To alter the color scheme, choose a color for each residue from the drop-down color palette to the right of each residue symbol. Alternatively, use the system Color Picker to select a color, by clicking on the current color swatch.
5. Select **Apply** to preview the changes. If they are not satisfactory, select **Revert** to restore the previous color set.
6. Select **OK** to save the changes.

Applying color groups

To apply an existing residue color scheme

1. With nucleotide or protein sequences displayed in the MSA window and the **Editor** tab selected, click the **Groups** icon.

The **Color Group Editor** dialog box is displayed.

2. Choose a preset color scheme from the **Color By** drop-down menu.

3. Select **Apply** to preview the changes. If they are not satisfactory, select **Revert** to restore the previous color set.
4. Select **OK** to save the change to the display.

Creating, renaming and deleting color groups

To create a new residue color scheme

1. With nucleotide or protein sequences displayed in the MSA window and the **Editor** tab selected, click the **Groups** icon.

The **Color Group Editor** dialog box is displayed.

You can choose a preset color scheme to use as a template for your own scheme, as follows:

2. Choose a scheme from the **Color By** drop-down menu.
3. Choose **Copy This Group** from the **Color By** drop-down menu. A **New Color Group Name** prompt is displayed. The default name will be **Copy of *Name***, where ***Name*** is the scheme you selected.

Note. You can modify a preset scheme, but we recommend that you always make a copy of the scheme to modify, because you can delete it later if necessary.

4. Type a name for the new scheme, and select **OK** to return to the **Color Group Editor** dialog box.
5. Edit the residue groups as required. For each group you create, do the following:
   - type a name for the group
   - type the IUPAC one-letter code for each residue in the group
   - choose a color for the group from the drop-down palette.

Note. Each residue type can only appear in one group. If you enter a residue that already appears in another group, it is automatically removed from the previous group.

6. Select **Apply** to preview the changes. If they are not satisfactory, select **Revert** to restore the previous color set.
7. Select **OK** to save the changes.

To rename or delete a color scheme:

1. With nucleotide or protein sequences displayed in the MSA window and the **Editor** tab selected, click the **Groups** icon.

The **Color Group Editor** dialog box is displayed.

2. Choose the user-defined color group that you want to rename or delete from the **Color By** drop-down menu.

Note. You cannot use the **Color Group Editor** to rename or delete the pre-defined color groups.

3. Do one of the following:
   - To rename the color group, choose **Rename This Group** from the **Color By** drop-down menu. A prompt to enter a **New Color Group Name** is displayed. Type the new name in the text box, and select **OK** to return to the Color Group Editor
   - To delete the color group, choose **Delete This Group** from the **Color By** drop-down menu.
4. Select **OK** to close the **Color Group Editor** dialog box.

Editing aligned sequences

The MSA Editor allows you to copy, move, insert, rename and delete sequences. For your source material, you can select and use single or multiple sequences, or parts of sequences, from sequence windows, MSA windows, or other applications.

The MSA Editor incorporates familiar text-editor features, such as double-clicking to select a contiguous sequence of residues and moving sequences by selecting the sequence label and dragging. This section describes the available editing actions in the MSA window. Selection in single-sequence windows is described in *“Selecting residues”* on page 87.

Adding sequences to an existing alignment

You can add a new blank sequence to the MSA window, insert one or more existing sequences by pasting from the clipboard, or import sequences from files.

Before adding a sequence, ensure that the MSA document is unlocked. If necessary, click on the **Locked** icon on the toolbar to unlock the MSA document.

To create a new sequence in an existing alignment

1. Ensure that the MSA document you want to edit is active and the **Editor** tab is selected.
2. Choose **Edit | Add New Sequence** from the menu.

A newly created sequence is displayed below any existing sequences. Its length is set to the alignment length, and it consists entirely of gap characters. The cursor is positioned at the first character.

3. To input the sequence, do either of the following, as appropriate:
    - Paste in the required residues (see *"Inserting residues"* on page 332)
    - Type the required residues, using the cursor arrows to move along the sequence if necessary.

When you type, residues are inserted into the sequence. At the end of the sequence, gaps are truncated to keep the length constant.

Tip. You can program the numeric keypad for entering sequences. See *"Configuring the numeric keypad"* on page 45.

To paste sequences into an existing alignment

1. Ensure that the MSA document you want to edit is active and the **Editor** tab is selected.
2. Choose where you want the sequences to be inserted, and position the cursor somewhere in the sequence line immediately above this point.

Note. When there are multiple sequences on the Clipboard, the **Paste** command will not work if any sequences or residues are selected.

3. Choose **Edit | Paste** from the menu.

The pasted sequences will be inserted below the cursor.

Another way to add sequences is to import them from single or multiple-sequence files. You can import several files in one operation.

To add sequences from a file

1. Ensure that the MSA document you want to edit is active and the **Editor** tab is selected.
2. Choose where you want the sequences to be inserted, and position the cursor somewhere in the sequence line immediately above this point.
3. Choose **Edit | Add Sequences from File** from the menu.

The **Open** dialog box is displayed.

4. Select the sequence file(s) you want to open.
5. Select **Open** to open the sequence files.

The sequences are inserted at the cursor.

Selecting sequences in the MSA Editor

MacVector supports a full range of options to select residues and sequences.

To select all the contents of the MSA editor

1. Ensure that the MSA document you want to edit is active and the **Editor** tab is selected.
2. Choose **Edit | Select All** from the menu.

All the residues are highlighted. If you want to cancel the selection, click on white space anywhere in the sequence alignment display.

Tip. If you need to insert gaps at the start of an alignment, choose **Edit | Select All**. Then move the entire alignment to the right, by dragging with the mouse.

You can select one or more full sequences for copying or deletion. (Full sequences include both the sequence name and all its residues.)

To select full sequences in the MSA editor

1. Ensure that the MSA document you want to edit is active and the **Editor** tab is selected.
2. Select a sequence name by clicking on it.

The selected sequence is highlighted.

3. To select multiple sequences, hold down the **Shift** key while clicking on the required sequence names.

If you need to de-select a sequence, hold down the **Shift** key and click on its name again. To de-select all sequences, click anywhere in the sequence alignment display.

You can also select all or part of a residue sequence without selecting its name.

To make selections in sequences

1. Do one of the following:
   - click and drag on one or more residues. You can select a block of residues from several sequences in this way. However, discontinuous blocks cannot be selected
   - double-click a residue to select that residue and all residues contiguous with it. The selection is made in both directions, until an inserted gap or the end of the sequence is encountered

- click the cursor in front of the first residue in the range you want to select. Holding down the **Shift** key, scroll the window until you can see the last residue in the required range, and click after it.

The selected residues are highlighted. You can copy them to the clipboard using **Edit | Copy**, or move them as described in the next section.

Changing and deleting sequences

Before changing or deleting a sequence, ensure that the MSA document is unlocked. If necessary, click on the **Locked** icon on the toolbar to unlock the MSA document.

To move a selection within a sequence

1. Do one of the following:
   - place the cursor over the selection until the cursor symbol changes to a hand, then drag the selection to a new location
   - use the cursor keys to move the selection to the left or right.

Note. A selection can only be moved within the gaps between it and the rest of the sequence. If there are no gaps around the selection, it cannot be moved, and the cursor symbol will not change to a hand.

Inserting residues

When you type or paste residues, they are inserted into the sequence. At the end of the sequence, gaps are truncated to keep the alignment length constant. If there are no gaps at the end of the sequence, gaps are appended to all the other sequences in the alignment to keep them the same length.

To insert and remove gaps in a sequence

1. Position the cursor where you want to add or remove a gap.
2. Do one of the following:
   - to insert gaps, use the spacebar
   - to remove gaps, use the backspace key

To insert residues by typing

1. Position the cursor where you want to add the residues.
2. Type the residues, using the keyboard or programmed numeric keypad.

To paste residues into a sequence

1. Position the cursor where you want to add the residues.

2. Use **Edit | Paste** to insert the residues from the clipboard.

Note. If you paste a block of residues from more than one sequence into the alignment, they are always inserted as new sequences, with names based on their parent sequences. You cannot insert residues into more than one sequence at a time.

During alignment, it may be necessary to insert one or more gaps in an entire block of sequences:

To insert gaps in multiple sequences

1. Select a block of residues immediately after the position where the gap should be.
2. Press any “gap” key (space, hyphen, or period). A gap will be inserted in front of the selected block, in all the selected sequences. Repeat as required.

Deleting and replacing residues

To delete residues in a sequence

1. Select the residues you want to delete. You can select a block of residues from more than one sequence.
2. Choose **Edit | Cut** to delete the residues.

To keep all sequences the same length, gaps are appended to the affected sequences as required.

You can also use the delete key to delete residues.

To replace or clear residues in a sequence

1. Select the residues you want to replace.
2. Do one of the following:
   - Choose **Edit | Clear** to replace all the residues by gaps
   - Type or paste in the new residues.

Note. You cannot use a block of residues from more than one sequence to replace another selected block. You must insert residues into one sequence at a time.

To delete sequences

1. Select the name of a sequence you want to delete. To select more than one sequence, hold down the **Shift** key and click on the required names.
2. Choose **Edit | Delete Sequence** to delete the sequences. Alternatively, use the **Cut** function; the sequences will be placed on the Clipboard.

Ordering and naming sequences

To control the ordering and naming of sequences, use the sequence name buttons in the panel to the left of the displayed residues in the MSA editor.

Changing the sequence display order

You can change the order in which the sequences are displayed.

To change the order in which sequences are displayed

1. Select a sequence label by clicking on it.
2. Hold down the mouse button and drag the sequence to a new position

Note. Altering the sequence order does not affect the order of sequences in the results windows, until you regenerate the results.

To view the new sequence order in the results windows

1. Click the **Views** icon on the toolbar.

The **Alignment Views** dialog box is displayed.

2. Do one of the following:
   - Select **Recalculate** and use the **Alignment Views** dialog box to repeat the multiple sequence alignment calculation
   - Alternatively, select **OK** in the **Alignment Views** dialog box to regenerate the windows.

Setting reference sequences

In linear **Line Wrap** mode, you can keep reference sequences at the top of the list and compare them to others. Only the topmost of the scrollable sequences can be made into a reference sequence.

To set reference sequences

1. Ensure that the MSA editor display is in linear **Line Wrap** mode. Click the **Line Wrap** icon if necessary.

A pin icon will be visible to the right of the topmost sequence name.

2. Click on the pin icon to fix the position of the reference sequence.

The sequence is now stationary. It will remain at the top of the list while the other sequences are scrolled underneath. A new pin icon appears to the left of the sequence under the reference sequence.

Note. Further sequences can be locked as reference sequences whenever a pin icon appears. You may first need to change the sequence display order to place the required sequence below the existing reference sequences.

Renaming aligned sequences

After a sequence alignment, it is often useful to rename aligned sequences.

To rename an aligned sequence

1. Select a sequence name by double-clicking on it.

The **Rename Sequence** dialog box is displayed.

2. Type a name in the **New sequence name** text box.
3. Select **OK** to save.

MSA Text view

The MSA Text view displays the alignments in standard form. Under the alignment there is an identity line, which indicates with symbols where there are identities and similarities in the alignment. The view includes a custom header showing a score for the alignment, together with additional parameters.

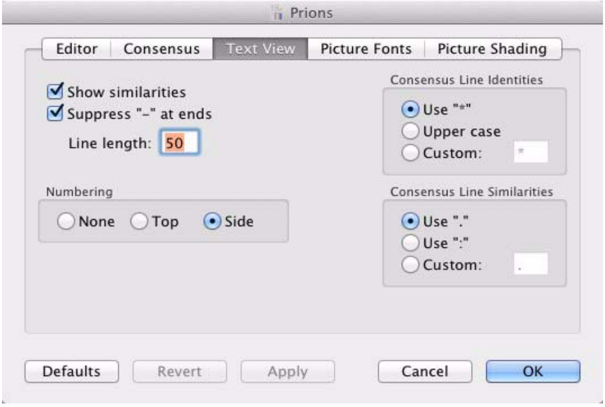
Note. The settings and parameters reported in the MSA Text view are only valid following a multiple sequence alignment.

The text multiple alignment can be saved to disk as a text file.

Tip. The MSA Text view is enabled by default. To disable it, uncheck **Multiple Alignment** in the **Text Display** panel on the **Alignment Views** dialog box.

The appearance of the MSA Text view be modified in a number of ways, using the **Text Display** tab on the **Multiple Alignment Options** dialog box.

Note. These settings also affect the MSA Pairwise view.



☑ Show similarities

☑ Suppress "-" at ends

Line length: 50

Numbering

○ None ○ Top ● Side

Consensus Line Identities

● Use "*"
○ Upper case
○ Custom: *

Consensus Line Similarities

● Use "."
○ Use ":"
○ Custom: .

To modify the appearance of the MSA Text view

1. Click the **Prefs** icon on the toolbar in the MSA window.

The **Multiple Alignment Options** dialog box is displayed.

2. Select the **Text Display** tab.
3. If you want to show identical residues in the consensus line, select the **Show identities** check box, and then choose the required **Identities** marker from the radio button and text box options.
4. If you want to show similar residues in the consensus line, select the **Show similarities** check box, and then choose the required **Similarities** marker from the radio button and text box options.
5. If you want to stop the ends of aligned sequences from being padded with the “-” symbol, select the **Suppress “-” at ends** check box.
6. Enter the number of residues per line in the **Line length** text box.
7. Choose a numbering style for the alignments, by selecting one of the **Numbering** radio buttons.
8. Select **Apply** to see the effects of your changes on any open MSA windows.
9. Select **OK** to save the changes and close the dialog box.

Alternatively, select **Defaults** to reset the default settings, or **Cancel** to close the dialog box without saving.

MSA Pairwise view

The MSA Pairwise view displays all pairwise combinations of input sequences aligned with each other. Each pairwise alignment includes a summary report of the percentage identity, percentage similarity, and number of gaps. For two sequences there is a single alignment, for three sequences there are three alignments, for four sequences there are six alignments, and so on. The alignments are derived from the complete multiple alignment, with all other sequences removed and all residual gaps closed. This approach closely matches the results of an independent pairwise sequence alignment.

To enable the MSA Pairwise view

1. Click the **Views** icon on the toolbar in the MSA window.

The **Alignment Views** dialog box is displayed.

2. Check **Pairwise Alignments** in the **Text Display** panel.
3. Click **OK**.

Note. You can enable and disable any combination of the results tabs at one time by setting the required options before selecting **OK**.

The appearance of the MSA Pairwise view can be modified in a number of ways, using the **Text Display** tab on the **Multiple Alignment Options** dialog box, as described in *"MSA Text view"* on page 335 above.

MSA Matrix view

The MSA Matrix view displays a table that shows the similarity and identity scores between pairs of sequences in an alignment. These scores are generated as part of the process of generating pairwise alignments between sequences.

To enable the MSA Matrix view

1. Click the **Views** icon on the toolbar in the MSA window.

The **Alignment Views** dialog box is displayed.

2. Check **Identity/Similarity Matrix** in the **Text Display** panel.
3. Click **OK**.

Note. You can enable and disable any combination of the results tabs at one time by setting the required options before selecting **OK**.

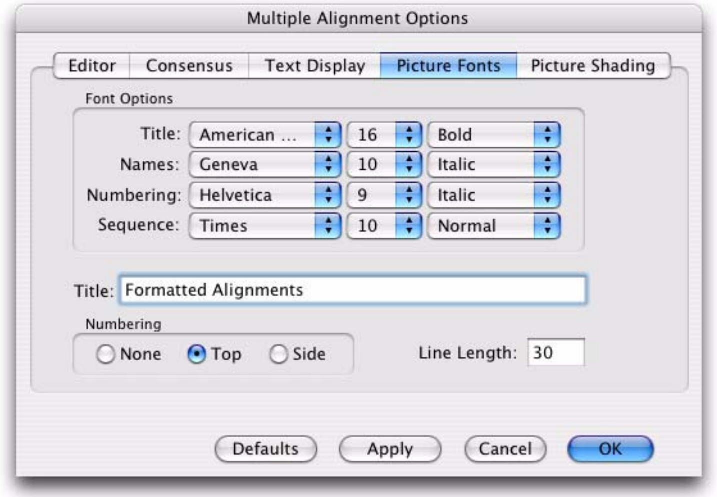
For nucleic acid alignments only the identity scores are relevant. However, for protein alignments the identity and similarity are distinct. Identity is the absolute relationship between sequences, for example 100% identity between two sequences means that they are absolutely identical. However, similarity refers to amino acids that are chemically, or rather biologically, related. For example two hydrophobic residues would score higher than a hydrophobic and a hydrophilic residue. Scoring of similarity is done according to the similarity/substitution matrices that are used to generate the alignment.

MSA Picture view

The MSA Picture view displays the multiple alignments in high resolution PDF form. When printed on a laser printer, the quality of this window is to publication standard. The aligned sequences can be copied and pasted as graphics into other applications, for example word processors or graphical editing applications.

Tip. The MSA Picture view is enabled by default. To disable it, uncheck **Multiple Alignment** in the **Picture Display** panel on the **Alignment Views** dialog box.

You can format the contents of the MSA Picture view, using the **Picture Fonts** and **Picture Shading** tabs on the **Multiple Alignment Options** dialog box.


Multiple Alignment Options

Font Options

Title:	American ...	16	Bold
Names:	Geneva	10	Italic
Numbering:	Helvetica	9	Italic
Sequence:	Times	10	Normal

Title: Formatted Alignments

Numbering: ○ None ● Top ○ Side

Line Length: 30

To modify the appearance of the MSA Picture view

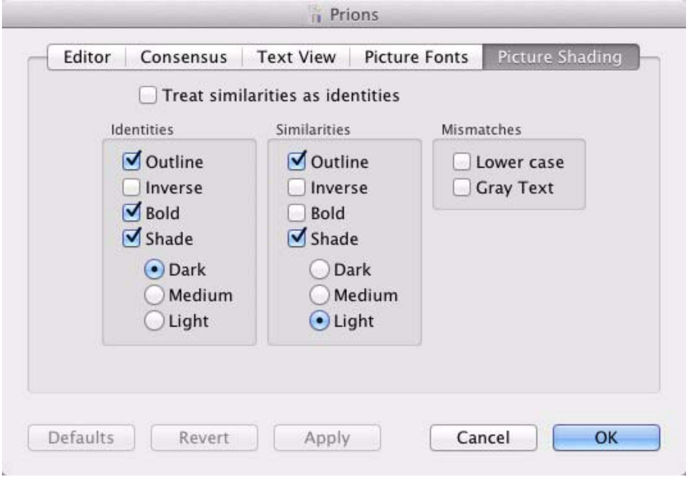
1. Click the **Prefs** icon on the toolbar in the MSA window.

The **Multiple Alignment Options** dialog box is displayed.

2. Select the **Picture Fonts** tab.
3. Use the **Set** button to choose alternative font styles, sizes and weights for the different types of text.
4. Enter a title for the window in the **Title** text box.
5. Choose a numbering style for the alignments, using one of the **Numbering** radio buttons.
6. Type in the required number of residues per line in the **Line Length** text box.
7. Select **Apply** to see the effects of your changes on any open MSA editor windows.

If you need to cancel the changes, select **Cancel**. If you need to reset the defaults, select **Defaults**.

8. Select the **Picture Shading** tab.
9. If you want similar residues to be treated as identities for the purpose of shading, select the **Treat similarities as identities** check box.
10. Choose the appearance of residue highlighting for identities, using the controls in the **Identities** panel.



□ Treat similarities as identities

Identities
☑ Outline
□ Inverse
☑ Bold
☑ Shade
● Dark
○ Medium
○ Light

Similarities
☑ Outline
□ Inverse
□ Bold
☑ Shade
○ Dark
○ Medium
● Light

Mismatches
□ Lower case
□ Gray Text

11. Choose the appearance of residue highlighting for similarities and mismatches, using the controls in the **Similarities** and **Mismatches** panels.

12. Select **Apply** to see the effects of your changes on any open MSA editor windows.
13. Select **OK** to save the changes and close the dialog box.

Alternatively, select **Defaults** to reset the default settings, or **Cancel** to close the dialog box without saving.

To modify the consensus line in the Picture view

1. To view the options for calculating and displaying the consensus line, select the **Consensus** tab in the **Multiple Alignment Options** dialog box.

The consensus line controls are the same for the MSA Editor, Text and Picture views. For detailed instructions, refer to *"Calculating the consensus line"* on page 320 and *"Displaying the consensus line"* on page 323.

MSA Guide Tree view

The MSA Guide Tree view displays a guide tree for aligned sequence results when four or more sequences are aligned together.

To enable the MSA Guide Tree view

1. Click the **Views** icon on the toolbar in the MSA window.

The **Alignment Views** dialog box is displayed.

2. Check **Guide Tree** in the **Picture Display** panel.
3. Click **OK**.

The tree shows the relationships generated when doing pairwise alignments. These are displayed as a phylogram or a cladogram, with the distances between nodes shown in phylogenetic units. As an approximate guide, a value of 0.1 corresponds to a difference of 10% between two sequences.

Note. These guide trees may not be true phylogenetic trees.

You can use the toolbar buttons and the **Tree Viewer Options** dialog box to select a tree shape, root the tree and edit its appearance. You can also print and save the guide tree.

MSA Profile view

The MSA Profile view displays the frequency of occurrence of each residue at each position in the alignment. It is designed primarily to display this information along with the consensus sequence however, the text is in Transfac format, so you can save it (using **File** | **Export**) and use it in other applications (including the MacVector Nucleic Acid toolbox, in the case of DNA sequences).

The consensus sequence window

The consensus can be used to generate a sequence in a standard MacVector sequence window. You can use this sequence in a variety of ways - for example, to perform BLAST searches.

The sequence is generated using the settings on the **Consensus** tab of the **Multiple Alignment Options** dialog. Any spaces are ignored, resulting in a continuous sequence of the matching residues.

To display a consensus sequence window

1. Click the **Views** icon on the toolbar in the MSA window.

The **Alignment Views** dialog box is displayed

2. Select the **Create Consensus Sequence** check box.
3. Type a name for the sequence in the text box.
4. Select **OK**.

Performing Multiple Sequence Alignment

There are 3 multiple sequence alignment algorithms available in MacVector:

- ClustalW
- Muscle
- T-Coffee

All three are progressive alignment algorithms. These generally build a guide tree that represents the pairwise relationships between each possible pair of sequences in the alignment. A multiple sequence alignment is then built sequentially using the tree as a guide to construction.

All three algorithms are integrated into the MSA editor. This means you can try all three algorithms on the same alignment to see the results. MacVector provides the following graphical features for displaying alignment results:

- Both multiple and pairwise alignments can be displayed in standard MacVector text windows.
- Multiple alignments can be displayed as high quality PDF images that can be exported to other applications.

- A summary of the similarity scores between each pair of sequences can be displayed in tabular form.
- A consensus sequence can be displayed and saved.
- A guide tree of the sequence comparisons can be displayed, showing the relationships between sequences (ClustalW and T-Coffee alignments only).

ClustalW Algorithm

You can use the ClustalW algorithm (Thompson *et al*, 1994) to align several nucleotide or amino acid sequences simultaneously.

To perform a ClustalW alignment, you need two or more sequences of the same type (nucleic acid or protein) open - either in individual Sequence windows or in one or more MSA windows.

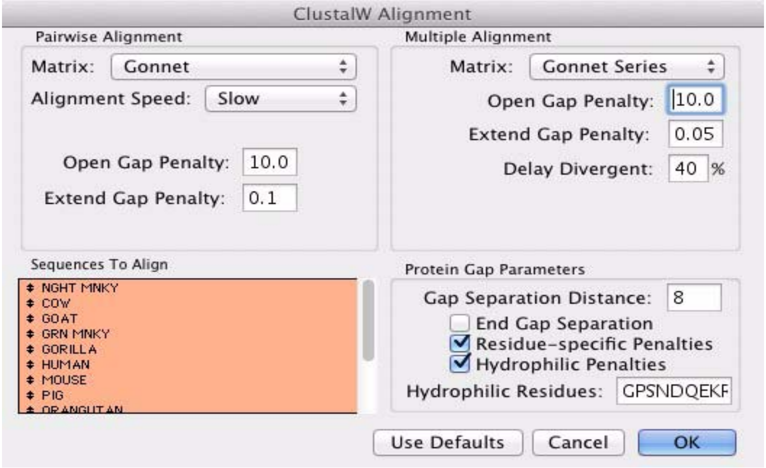
Performing a ClustalW alignment

The **ClustalW Alignment** dialog box is divided into four panels:

- **Pairwise Alignment**
- **Multiple Alignment**
- **Sequences To Align**
- **Protein Gap Parameters**

The parameters that can be set within each panel vary depending on the type of sequences being aligned and the alignment speed that is selected. To perform your alignment, set parameters in each panel according to your sequence type and alignment requirements.

ClustalW aligns sequences in two stages. The first, pairwise, stage compares each sequence individually with every other sequence. The Pairwise Alignment parameters control the speed and sensitivity of this stage. The second, multiple alignment, stage progressively merges the sequences, starting with the pair that scored highest in the pairwise stage. The Multiple Alignment parameters control this stage.



ClustalW Alignment

Pairwise Alignment

Matrix: Gonnet

Alignment Speed: Slow

Open Gap Penalty: 10.0

Extend Gap Penalty: 0.1

Multiple Alignment

Matrix: Gonnet Series

Open Gap Penalty: 10.0

Extend Gap Penalty: 0.05

Delay Divergent: 40 %

Sequences To Align

NGHT MNKY
COW
GOAT
GRN MNKY
GORILLA
HUMAN
MOUSE
PIG
ORANGUTAN

Protein Gap Parameters

Gap Separation Distance: 8

□ End Gap Separation

☑ Residue-specific Penalties

☑ Hydrophilic Penalties

Hydrophilic Residues: GPSNDQEKF

To perform a ClustalW alignment

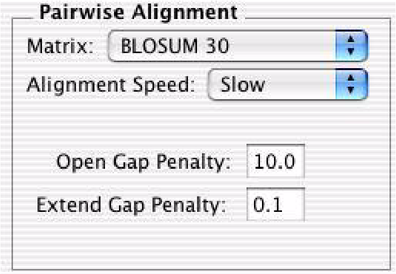
1. Choose **Analyze | Align Multiple Sequences Using | ClustalW** from the menu or click and hold the **Align** icon on the MSA Editor toolbar and select **ClustalW** from the context menu.

The **ClustalW Alignment** dialog box is displayed.

Setting options in the Pairwise Alignment panel

2. If you are aligning protein sequences, choose the protein scoring matrix from the **Matrix** drop-down menu.
3. Choose from the **Alignment Speed** drop-down menu:
   - **Slow** means that the initial pairwise alignments are performed using a full dynamic programming algorithm

- **Fast** uses the Wilbur & Lipman (1983) method.



Pairwise Alignment

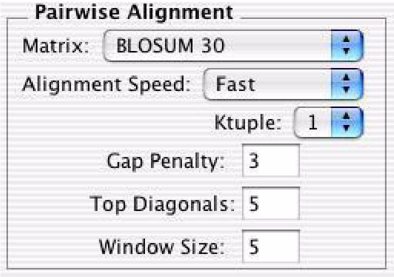
Matrix: BLOSUM 30

Alignment Speed: Slow

Open Gap Penalty: 10.0

Extend Gap Penalty: 0.1

4. For **Slow** alignments, do the following:
   - enter a value between 0 and 100 in the **Open Gap Penalty** text box. This is the score that is subtracted when a gap is inserted in an alignment. Increasing the gap opening penalty makes gaps less frequent
   - enter a value between 0 and 10 in the **Extend Gap Penalty** text box. This is the penalty for extending the gap by 1 residue. Increasing the extend gap penalty makes gaps shorter. Terminal gaps are not penalized.



Pairwise Alignment

Matrix: BLOSUM 30

Alignment Speed: Fast

Ktuple: 1

Gap Penalty: 3

Top Diagonals: 5

Window Size: 5

5. For **Fast** alignments, do the following:
   - choose a value from the **Ktuple** drop-down menu. This is the number of consecutive residues that must match the query sequence exactly before MacVector will attempt to score the matching region. Increasing the value speeds the alignment, but if the value is too large, significant homology between sequences may be missed. Throughout MacVector, Ktuple size is also referred to as *hash size* or *word size*

- enter a value between 1 and 500 in the **Gap Penalty** text box. This is the penalty for each gap. The parameter has little effect on speed or sensitivity unless extreme values are entered
- enter a value between 1 and 50 in the **Top Diagonals** text box. This value determines the number of Ktuple matches on each diagonal in an imaginary dot-matrix plot. A large value increases sensitivity, a small value makes the alignment faster
- enter a value between 1 and 50 in the **Window Size** text box. This value specifies the window size around each top diagonal. Diagonals that fall inside this window are used in the alignment. Increasing the window size results in a more sensitive, but slower alignment. Decreasing the window size increases the speed, but may result in small regions of homology being missed.

Setting options in the **Multiple Alignment** panel

6. If you are aligning protein sequences, choose the protein scoring matrix series from the **Matrix** drop-down menu.
7. Enter a value between 0 and 100 in the **Open Gap Penalty** text box. This is the score that is subtracted when a gap is inserted in an alignment. Increasing the gap opening penalty makes gaps less frequent.
8. Enter a value between 0 and 10 in the **Extend Gap Penalty** text box. This is the penalty for extending the gap by 1 residue. Increasing the extend gap penalty makes gaps shorter. Terminal gaps are not penalized.
9. Enter a value between 0% and 100% in the **Delay Divergent** text box. By delaying the alignment of the divergent sequences until after the most closely related sequences have been aligned, a more accurate alignment is achieved. The default value is 40%, which means that if a sequence is less than 40% identical to any other sequence, its alignment is delayed.
10. Choose from the **Transitions** drop-down menu as follows:
    - **weighted** to give higher weightings to transitions (A <-> G, T <-> C) compared to transversions (A <-> T, A <-> C, G <-> T, G <-> C)
    - **unweighted** to treat transitions and transversions equally.

Note. This option is only present for nucleic acid alignments.

Setting options on the **Protein Gap Parameters** panel

If you are aligning protein sequences, you must set a number of additional parameters.

11. Enter a value between 0 and 100 in the **Gap Separation Distance** text box. This allows you to discourage gaps being opened too close together by increasing the open gap penalty for a specific distance from existing gaps. For example, if the gap separation distance is set to 8, the open gap penalty is increased if the new position is within 8 residues of an existing gap.
12. Select the **End Gap Separation** check box to treat end gaps as internal gaps, to avoid gaps opening too close together, as specified by **Gap Separation Distance**. If disabled, end gaps are ignored.
13. Select the **Residue-specific Penalties** check box to modify the gap opening penalty at each position in the alignment or sequence. The **Open Gap Penalty** is multiplied by a residue-specific gap modification factor, as shown in *"Residue-specific gap modification factors for ClustalW sequence alignment"* on page 492. If there is a mixture of residues at the position, the multiplying factor is the average of all the contributions from each sequence.
14. Select the **Hydrophilic Penalties** check box to increase the chances of a gap opening within a hydrophilic stretch, as defined by **Hydrophilic Residues**.
15. In the **Hydrophilic Residues** text box, enter the IUPAC one-letter code for each residue that is to be considered hydrophilic. Any run of 5 hydrophilic residues is considered to be a hydrophilic stretch.

Choosing sequences

16. From the scrolling list in the **Sequences To Align** panel, highlight the sequences that you want to align.

Use **<shift>** click or **<⌘>** click to toggle selections. Sequences are aligned in the order they appear in the list.

17. If you want to change the order in which sequences are aligned, do the following:
    - put the mouse pointer over the arrow at the left of the sequence you want to move
    - hold the mouse button down and drag the sequence to the required position

- release the mouse button.

18. Select **OK** to perform the alignment.

A dialog box is displayed, informing you of the progress of the alignment. You can close this and monitor progress using the Job Manager if you prefer. See Appendix B, *"Using the Job Manager"* for more details about the Job Manager. There is also a **Stop** button which will cancel the ClustalW Alignment. The ClustalW job runs in the background, allowing you to continue to use other functions in MacVector while the job is in progress.

Note. While a ClustalW alignment job is in progress you cannot edit the alignments in the original window, nor can you submit additional alignment jobs from that window. However, you can submit additional alignment jobs from other windows.

When the calculation is complete, the **Alignment Views** display dialog box is displayed. This is described in *"Displaying alignment results"* on page 353. The MSA Editor view is also displayed. This is described in *"MSA Editor view"* on page 317.

Muscle Algorithm

The Muscle algorithm (Edgar, 2004) is a fast algorithm optimized for aligning protein sequences.

Muscle does not perform a pairwise alignment but instead uses an approximate method, comparing the number of short subsequences (k-mers, k-tuples or words) that each pair of sequences share. You can see immediately how this is much faster for alignments containing many sequences where the number of pairwise alignments needed to construct the tree is high.

Muscle is generally regarded as faster than Clustalw and T-Coffee at the penalty of being slightly less accurate.

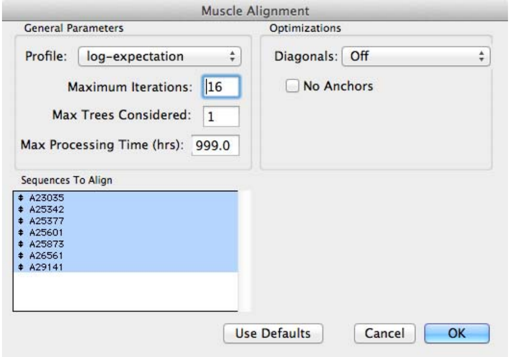
To perform a Muscle alignment, you need two or more sequences of the same type (nucleic acid or protein) open - either in individual Sequence windows or in one or more MSA windows.

Performing a Muscle alignment

The **Muscle Alignment** dialog box is divided into three panels.

- **General Parameters**
- **Optimizations**

- **Sequence to Align**



Muscle Alignment

General Parameters

Profile: log-expectation

Maximum Iterations: 16

Max Trees Considered: 1

Max Processing Time (hrs): 999.0

Optimizations

Diagonals: Off

□ No Anchors

Sequences To Align

A23035
A25342
A25377
A25601
A25873
A26561
A29141

To perform a Muscle alignment

1. Choose **Analyze | Align Multiple Sequences Using | Muscle** from the menu or click and hold the **Align** icon on the MSA Editor toolbar and select **Muscle** from the context menu.

The **Muscle Alignment** dialog box is displayed.

Setting options in the General Parameters panel

2. For protein alignments only, select a protein profile scoring function to be used from the **Profile** drop down list. The following options are available:

   - **log-expectation**
   - **PAM200**
   - **VTML240**

3. Specify the maximum number of iterations that should be performed in the **Maximum Iterations** text box. The default value is 16.

Tip. For large alignments the recommended value for Maximum Iterations is 2.

4. Specify the maximum number of trees that should be considered in the **Max Trees Considered** text box. The default value is 1. Usually, only one tree is needed.

5. Specify the maximum time in hours that should be spent processing the alignment job in the **Max Processing Time (hrs)** text box If the specified time is exceeded the alignment job will terminate even though it is incomplete. The default value is 999 hours.

Setting options in the Optimizations panel

6. Control whether or how to use 6-mer common words between pairs of sequences as seeds for diagonals to optimize the alignment by selecting one of the following options from the **Diagonals** drop down list:
   - **Off** - Optimization is not used
   - **On** - Optimization is used
   - **1st Iteration** -Optimization is used in the first iteration only.
   - **2nd Iteration** - Optimization is used in the second iteration only.
7. Check the **No Anchors** option to disable the use of ‘vertical blocks’ to anchor alignments.

Tip. This option should only be used if absolute accuracy is required. It will slow down the alignment considerably.

Choosing sequences

8. From the scrolling list in the **Sequences To Align** panel, highlight the sequences that you want to align.

Use **<shift>** click or **<⌘>** click to toggle selections. Sequences are aligned in the order they appear in the list.

9. If you want to change the order in which sequences are aligned, do the following:
   - put the mouse pointer over the arrow at the left of the sequence you want to move
   - hold the mouse button down and drag the sequence to the required position
   - release the mouse button.
10. Select **OK** to perform the alignment.

A dialog box is displayed, informing you of the progress of the alignment. You can close this and monitor progress using the Job Manager if you prefer. See Appendix B, *“Using the Job Manager”* for more details about the Job Manager. There is also a **Stop** button which will cancel the Muscle alignment. The Muscle job runs in the background, allowing you to continue to use other functions in MacVector while the job is in progress.

Note. While a Muscle alignment job is in progress you cannot edit the alignments in the original window, nor can you submit additional alignment jobs from that window. However, you can submit additional alignment jobs from other windows.

When the calculation is complete, the **Alignment Views** display dialog box is displayed. This is described in *“Displaying alignment results”* on page 353. The MSA Editor view is also displayed. This is described in *“MSA Editor view”* on page 317.

T-Coffee Algorithm

The T-Coffee algorithm (Notredame *et al*, 2000) is a fast algorithm optimized for aligning DNA and protein sequences.

T-Coffee builds a library of all pairwise alignments but also aligns each sequence in the pair with a third sequence in the sequence set before building the multiple sequence alignment.

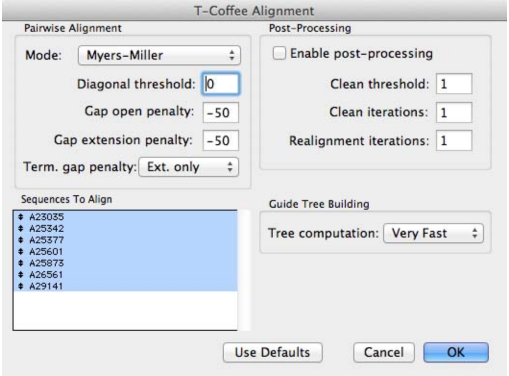
It is regarded as being slightly slower than ClustalW but will produce more accurate alignments for distantly related amino acid sequences.

To perform a T-Coffee alignment, you need two or more sequences of the same type (nucleic acid or protein) open - either in individual Sequence windows or in one or more MSA windows.

Performing a T-Coffee alignment

The **T-Coffee Alignment** dialog box is divided into four panels:

- **Pairwise Alignment**
- **Post-Processing**
- **Sequences To Align**
- **Guide Tree Building**



Mode: Myers-Miller
Diagonal threshold: 0
Gap open penalty: -50
Gap extension penalty: -50
Term. gap penalty: Ext. only

□ Enable post-processing
Clean threshold: 1
Clean iterations: 1
Realignment iterations: 1

A23035
A25342
A25377
A25601
A25873
A26561
A29141

Tree computation: Very Fast

To perform a T-Coffee alignment

1. Choose **Analyze | Align Multiple Sequences Using | T-Coffee** from the menu or click and hold the **Align** icon on the MSA Editor toolbar and select **T-Coffee** from the context menu.

The **T-Coffee Alignment** dialog box is displayed.

Setting options in the Pairwise Alignment panel

2. Select an alignment mode from the **Mode** drop down list. The following options are available:
   - **Myers-Miller** - uses an implementation of the Myers-Miller dynamic programming algorithm (quadratic in time and linear in space). This algorithm is recommended for very long sequences. It is about 2 times slower than Gotoh and only accepts tg_mode=1or 2 (i.e. gaps penalized for opening).
   - **Gotoh** - uses an implementation of the Gotoh algorithm (quadratic in memory and time)
   - **FastA** - uses an implementation of the FastA algorithm. The sequence is hashed, looking for k-tuple words. Dynamic programming is carried out only on the ndiag best scoring diagonals. This is much faster but less accurate than **Gotoh** and **Myers-Miller**.
   - **CFastA**- uses an implementation of the FastA algorithm with checking. Dynamic programming is carried out on the ndiag best diagonals, and then on the 2*ndiags, and so on until the scores converge. Complexity will depend on the level of divergence of the sequences, but will usually be L*log(L), with an accuracy comparable to that of **Gotoh** and **Myers-Miller**.
3. Specify the value of the threshold used when selection diagonals in the **Diagonal Threshold** text box. The default value is 0.
4. Specify the penalty that is applied when a gap is opened in the **Gap Open Penalty** text box. The penalty must be negative. If no value is provided when using a substitution matrix, then a value will be automatically computed. The default value is 0.

Tip. In T-Coffee matches get a score between 0 (match) and 1000 (match perfectly consistent with the library). The default cosmetic penalty is set to -50 (5% of a perfect match). If you want to tune the **Gap Open Penalty** and see a strong effect, you should use values between 0 and -1000.

5. Specify the penalty that is applied when a gap is extended in the **Gap Extension Penalty** text box. The penalty must be negative. The default value is 0.
6. Choose how terminal gaps are penalized using the **Term. Gap Penalty** drop down list. The following options are available:
   - **Open+Ext.** - Terminal gaps penalized with (**Gap Open Penalty** + G**ap Extension Penalty**) x gap length
   - **Ext. only** - Terminal gaps penalized with **Gap Extension Penalty** x gap length.
   - **None** - terminal gaps unpenalized.

Setting options in the Post-Processing panel

7. Optionally, **Enable Post Processing** by checking the box. When post-processing is enabled, residues that have a reliability score less than or equal to the specified **Clean threshold** are realigned to the rest of the alignment. Residues with a score higher than the **Clean threshold** constitute a rigid framework that cannot be altered. The cleaning algorithm starts from the top left segment of low constituency residues and works its way left to right, top to bottom along the alignment. It is computationally expensive. By default the option is unchecked.

Note. The use of this option can result in memory overflow when aligning large sequences.

8. Specify a threshold reliability score for residues in the **Clean threshold** text box. Residues with scores below this value are realigned when post processing is enabled.
9. Specify the number of iterations over which post-processing is performed in the **Clean iterations** text box.

Setting options on the **Guide Tree Building** panel

10. Select the method use for computing the distance matrix (distance between every pair of sequences) required for the computation of the dendrogram from the **Tree Computation** drop down list. The following options are available:
    - **Fast** - Used the FastA dp_mode with the extended library.
    - **Very Fast** - Uses the FastA dp_mode with blosum62mt.
    - **ktup** - Uses ktup matching (c.f. Muscle)

Choosing sequences

11. From the scrolling list in the **Sequences To Align** panel, highlight the sequences that you want to align.

Use **<shift>** click or **<⌘>** click to toggle selections. Sequences are aligned in the order they appear in the list.

12. If you want to change the order in which sequences are aligned, do the following:

- put the mouse pointer over the arrow at the left of the sequence you want to move
- hold the mouse button down and drag the sequence to the required position
- release the mouse button.

13. Select **OK** to perform the alignment.

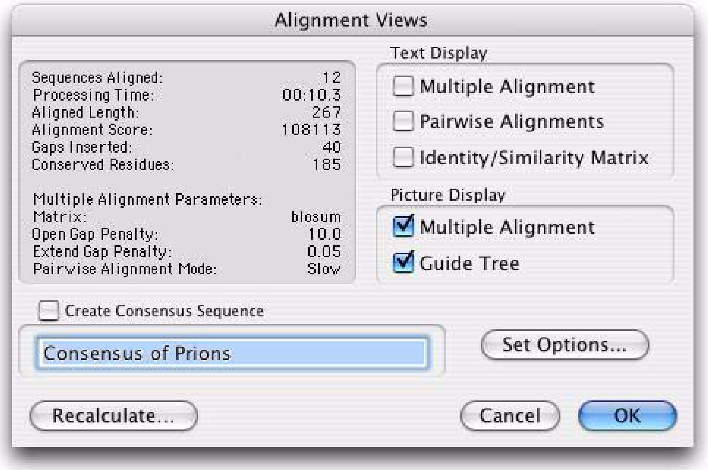
A dialog box is displayed, informing you of the progress of the alignment. You can close this and monitor progress using the Job Manager if you prefer. See Appendix B, *"Using the Job Manager"* for more details about the Job Manager. There is also a **Stop** button which will cancel the T-Coffee alignment. The T-Coffee job runs in the background, allowing you to continue to use other functions in MacVector while the job is in progress.

Note. While a T-Coffee alignment job is in progress you cannot edit the alignments in the original window, nor can you submit additional alignment jobs from that window. However, you can submit additional alignment jobs from other windows.

When the calculation is complete, the **Alignment Views** display dialog box is displayed. This is described in *"Displaying alignment results"* on page 353. The MSA Editor view is also displayed. This is described in *"MSA Editor view"* on page 317.

Displaying alignment results

The **Alignment Views** display dialog box allows you to choose one or more types of results to display.



Alignment Views

Sequences Aligned:	12
Processing Time:	00:10.3
Aligned Length:	267
Alignment Score:	108113
Gaps Inserted:	40
Conserved Residues:	185
Multiple Alignment Parameters:	
Matrix:	blosum
Open Gap Penalty:	10.0
Extend Gap Penalty:	0.05
Pairwise Alignment Mode:	Slow

Text Display

□ Multiple Alignment

□ Pairwise Alignments

□ Identity/Similarity Matrix

Picture Display

☑ Multiple Alignment

☑ Guide Tree

□ Create Consensus Sequence

Consensus of Prions

You can choose from the following output options:

- multiple alignment text display
- pairwise alignment text display
- identity/similarity matrix text display
- multiple alignment picture display
- consensus sequence display
- guide tree (for ClustalW and T-Coffee alignments only).

Only guide trees are discussed here. Refer to *“Multiple Sequence Alignment window”* on page 310 for a full discussion of the main display options.

The MSA Editor is also displayed with the results; for more information, see *“MSA Editor view”* on page 317.

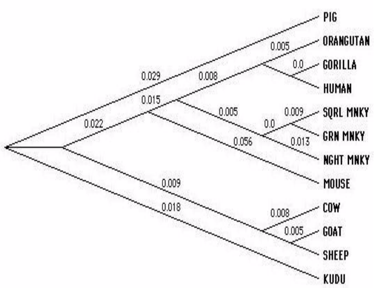
If you want to recalculate the alignment, select **Recalculate** to return to the previously used alignment parameters dialog box.

The **Alignment Views** dialog box is displayed on completion of each analysis. To display this dialog box at other times, for example to change the display parameters, click on the **Views** icon on the MSA Editor toolbar.

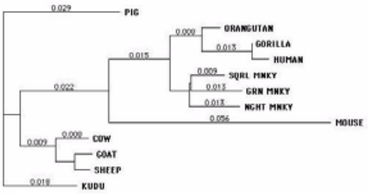
Displaying a guide tree

You can display a guide tree for aligned sequence results when four or more sequences are aligned by ClustalW and T-Coffee. The guide tree shows the relationships that these alignment algorithms generate when performing pairwise alignments, and is shown as a phylogram or a cladogram, with the distances between nodes shown in phylogenetic units. As an approximate guide, a value of 0.1 corresponds to a difference of 10% between two sequences.

Example windows:



Slanted cladogram



Phylogram

Note. Guide trees may not be true phylogenetic trees. See Chapter 18, *"Reconstructing Phylogeny"*, and *"Phylogenetic analysis"* on page 465.

To display a guide tree

1. In the **Picture Display** panel of the **Alignment Views** dialog box, select the **Guide Tree** check box.
2. Select **OK** to close the dialog box.

The guide tree is displayed in a Tree Viewer window.

You can use the toolbar buttons and the **Tree Viewer Options** dialog box to select a tree shape, root the tree and edit its appearance. You can also print and save the guide tree. For full details, see *"Tree Viewer window"* on page 365, *"Editing Trees"* on page 367, and *"Saving and printing tree diagrams"* on page 375.

18 Reconstructing Phylogeny

Overview

Phylogenetic trees indicate the evolutionary relationships among sequences. This chapter describes how to generate phylogenetic trees from multiple alignments of nucleic acid or protein sequences, and introduces the different methods for calculating evolutionary distances and building trees. Further information on the methods is presented in *“Phylogenetic analysis”* on page 465.

This chapter also describes how to use the MacVector Tree Viewer to inspect and edit phylogenetic trees, and how to print and save them.

The Tree Viewer can also be used to inspect guide trees for ClustalW multiple alignments.

Contents

Phylogenetic reconstruction

MacVector offers two modes of phylogenetic analysis, depending on the type of output required. Best Tree mode calculates the best tree using a given method. Bootstrap mode repeatedly resamples the data, generating phylogenies from each of the new data sets. A consensus tree of these phylogenies indicates how reproducible the Best Tree analysis is. The following sections describe how to generate phylogenetic trees using different methods, and how to view and save them. For further information on the methods, see *“Phylogenetic analysis”* on page 465.

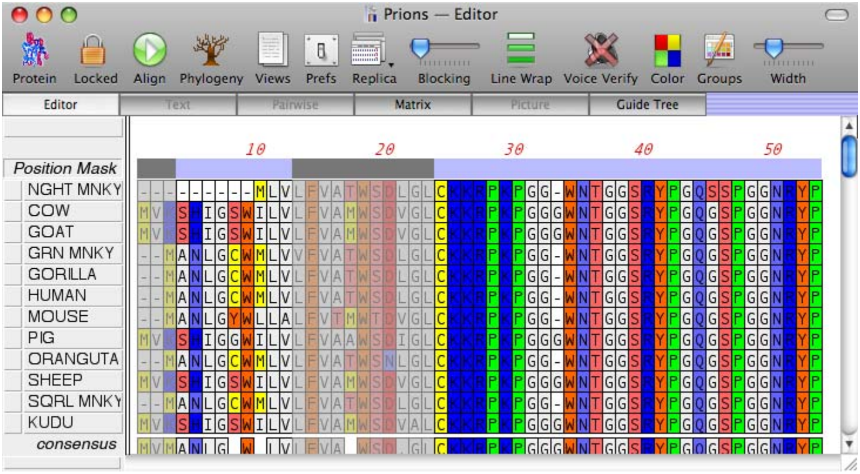
To run a phylogenetic analysis, the active window must be a Multiple Sequence Alignment (MSA) window containing four or more sequences.

Position masking

Alignments are statements of homology. We assume that amino acids or nucleotides at the same position have evolved from the same ancestral residue. In practice, however, it is often difficult to have any confidence in this homology for some regions of the alignment, and these should be omitted from the phylogenetic analysis.

MacVector allows you to mask any positions in the alignment, to exclude them from the analysis. This is done using the MSA Editor. When you enable the position masking tool, a blue bar appears above the alignment. By clicking in this bar at the appropriate locations, you can mask particular sites, which appear grayed.

Tip. If you save the multiple alignment as a NEXUS file, the position mask will be saved with it (as an exclusion set in the assumptions block of the file).



Position Mask	
NGHT MNKY	---------MLVLFVATWSDLGLCKKRPKPGG-WNTGGSRYPGQSSPGGNRYP
COW	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYP
GOAT	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYP
GRN MNKY	--MANLGCWMLVVFVATWSDLGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYP
GORILLA	--MANLGCWMLVLFVATWSDLGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYP
HUMAN	--MANLGCWMLVLFVATWSDLGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYP
MOUSE	--MANLGYWLLALFVTMWTDVGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYP
PIG	MVKSHIGGWILVLFVAAWSDIGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYP
ORANGUTA	--MANLGCWMLVLFVATWSNLGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYP
SHEEP	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYP
SQRL MNKY	--MANLGCWMLVLFVATWSDLGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYP
KUDU	MVKSHIGSWILVLFVAMWSDVALCKKRPKPGGGWNTGGSRYPGQGSPGGNRYP
consensus	MVMANLG W LVLFVA WSD.GLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYP

To exclude sites from phylogenetic analysis

By default, position masking is not enabled, and there is no blue bar displayed above the alignment in the MSA Editor view.

1. With the multiple alignment active and MSA Editor view selected, click the **Prefs** icon on the MSA Editor toolbar.

The **Multiple Alignment Options** dialog box is displayed.

2. Select the **Editor** tab, and check the **Show position mask** box to enable this feature.
3. Select **OK.**

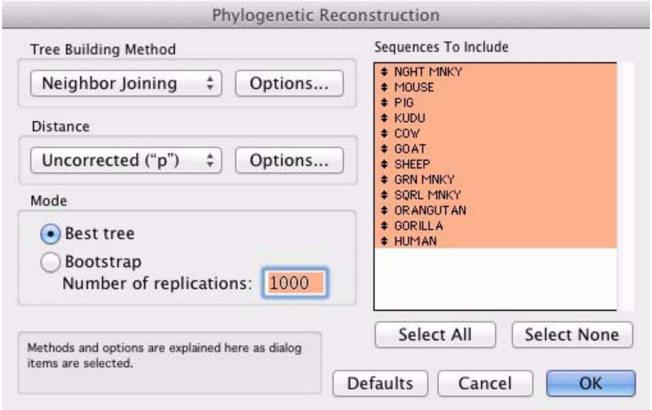
The blue position masking bar is displayed above the alignment in the MSA Editor view.

4. Select the positions to mask, using the mouse as follows:
   - to mask a single position, click in the masking bar at that position
   - to mask adjacent positions, click and drag
   - to unmask a position, click on it again
   - to mask all positions, hold down the Option key and click anywhere in the masking bar. Repeat to unmask all positions

Generating phylogenetic trees

This section describes how you can generate a phylogenetic tree from any alignment of four or more nucleotide or protein sequences. You can limit the analysis to a subset of the alignment, if required.

It also describes how you can check the reliability of this tree, by running a bootstrap analysis to display a bootstrap consensus tree for the same data. There are two stages in the generation of trees: calculation of the distances between pairs of sequences, and reconstruction of the phylogeny using the distance information. All building methods proceed by joining the two most similar sequences first, and then adding the other sequences one by one, in order of decreasing similarity.



Phylogenetic Reconstruction

Tree Building Method: Neighbor Joining

Distance: Uncorrected ("p")

Mode:
● Best tree
○ Bootstrap
Number of replications: 1000

Sequences To Include:
NGHT MNKY
MOUSE
PIG
KUDU
COW
GOAT
SHEEP
GRN MNKY
SQRL MNKY
ORANGUTAN
GORILLA
HUMAN

Methods and options are explained here as dialog items are selected.

To generate a phylogenetic tree

1. With the multiple alignment active and the MSA Editor view selected, click the **Phylogeny** icon on the MSA Editor toolbar.

Alternatively, choose **Analyze | Phylogenetic Analyses | Reconstruct Phylogeny.**

The **Phylogenetic Reconstruction** dialog box is displayed.

2. Use the **Tree Building Method** drop-down menu to choose a method of building trees.

- **Neighbor Joining** is the default. This method makes no assumptions about rates of divergence in different lineages
- The **UPGMA** method assumes that sequences have diverged at a constant rate.

3. To specify the method for resolving ties, click the **Options** button in the **Tree Building Method** panel. The **Tree Building Method Options** dialog box is displayed; select either the **Systematic** or **Random** radio button.

When there are two nodes or sequences equidistant from the current node in the tree:

- the **Systematic** method resolves the tie by adding the nodes in the order of the aligned sequences. This is the default method.
- the **Random** method chooses the order randomly.

4. Use the **Distance** drop-down menu to choose a method of calculating the pairwise distances between sequences.

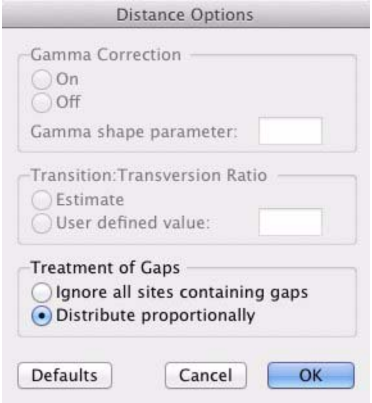
For nucleotide sequences the options are:

- **Absolute (# differences)**
- **Uncorrected (“p”)**
- **Jukes-Cantor**
- **Tajima-Nei**
- **Kimura 2-parameter**
- **Tamura-Nei** (default)
- **LogDet/Paralinear**

For protein sequences the options are:

- **Absolute (# differences)**
- **Uncorrected (“p”)** (default)
- **Poisson-correction**

5. To set the distance options, click the **Options** button in the **Distance** panel.



Gamma Correction
○ On
○ Off
Gamma shape parameter:

Transition:Transversion Ratio
○ Estimate
○ User defined value:

Treatment of Gaps
○ Ignore all sites containing gaps
● Distribute proportionally

The **Distance Options** dialog box is displayed. Options that are not relevant for your chosen **Distance** method will be grayed out and unavailable for selection.

- In the **Gamma Correction** panel, use the radio buttons to turn gamma correction **Off** or **On**, and enter a **Gamma shape parameter** value in the text box if required. Where gamma correction is available, the default value is 0.5
- In the **Transition:Transversion Ratio** panel, select the **Estimate** radio button to have the ratio estimated from your sequence data, or select **User defined value** and enter the required value in the text box. Where available, the default value is 1.0
- In the **Treatment of Gaps** panel, select the radio button either to **Ignore all sites containing gaps** or to **Distribute proportionally.** The default is to **Ignore all sites containing gaps**

Select **OK** to confirm the settings and close the dialog box. You can use the **Defaults** button to restore the default settings.

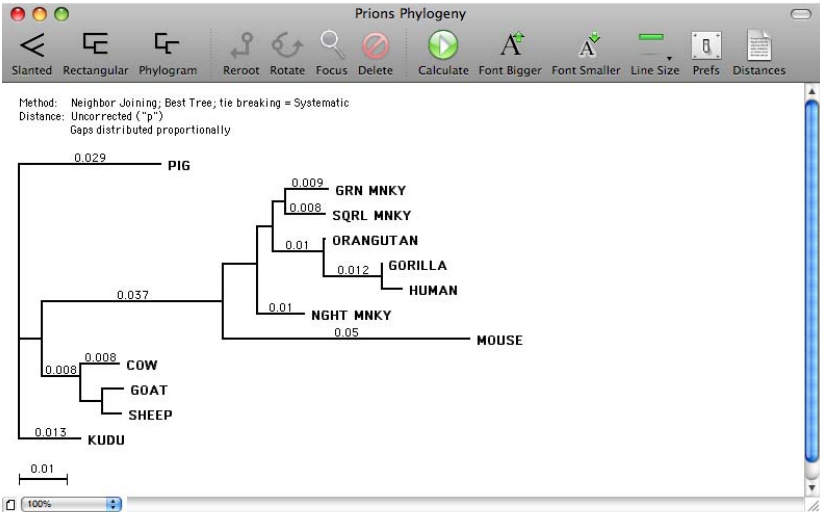
6. In the **Mode** panel, select **Best tree**.
7. By default, all sequences will be included in the tree, and they are all highlighted in the **Sequences To Include** list. If necessary, you can edit this list:

- To exclude a highlighted sequence from the tree, hold down the **<shift>** key and click on the sequence in the list
- To include a sequence that is not highlighted, hold down the **<shift>** key and click on the sequence
- To include all sequences, select the **Select All** button
- To exclude all sequences, select the **Select None** button.

8. Select **OK** to confirm the settings and perform the analysis.

Alternatively, you can use the **Defaults** button to restore the default settings, or the **Cancel** button to close the dialog without retaining the settings.

During the analysis an information box is displayed, showing a progress bar for each calculation stage. When it is complete, the resulting tree is displayed in a Tree Viewer window. You can control its appearance interactively (see *“Tree Viewer window”* on page 365).



Method: Neighbor Joining; Best Tree; tie breaking = Systematic
Distance: Uncorrected ("p")
Gaps distributed proportionally

Invalid distance warning

In some situations, no valid distances can be calculated for a pair of sequences. When this happens, a warning dialog box is displayed:

If you continue, then for each distance that cannot be determined, MacVector will assign the largest distance that was calculated in the matrix. For more details, see *“Invalid distances”* on page 467.

To generate a bootstrap consensus tree

1. With the tree displayed, click the **Calculate** icon on the Tree Viewer toolbar.

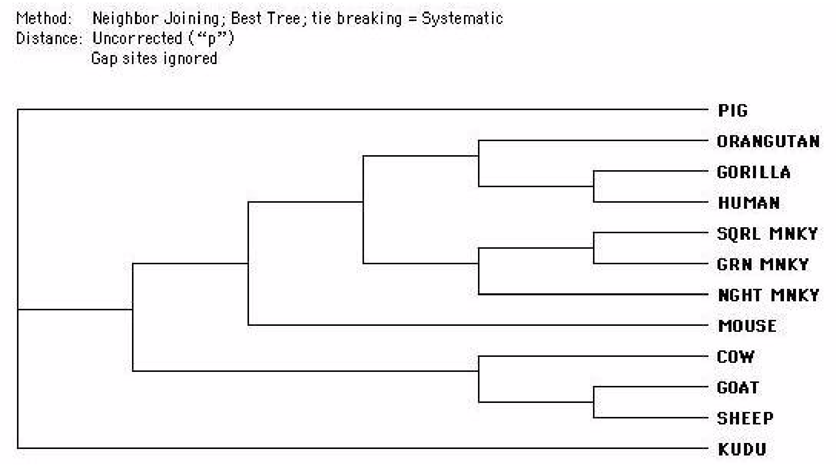
The **Phylogenetic Reconstruction** dialog box is displayed.

2. In the **Mode** panel, select the **Bootstrap** radio button.
3. Type in the required **Number of replications**.

The default value is 1000. Reducing this number gives faster but less reliable results.

4. Select **OK** to perform the analysis.

After the analysis is complete, the consensus bootstrap tree is displayed in the Tree Viewer window. The consensus bootstrap tree only includes nodes that appear consistently when the data is repeatedly resampled. Nodes that occur very frequently in the resampled data are labeled with their percentage occurrence. You can adjust the cutoff points for displaying and labeling nodes (see *“Setting the tree display options”* on page 373).



```
Method:   Neighbor Joining; Best Tree; tie breaking = Systematic
Distance: Uncorrected (“p”)
          Gap sites ignored
```

The topology of the bootstrap consensus tree will not always match that of the best tree.

For more details of the method and its interpretation, see *“Bootstrapping”* on page 470.

5. To display the original tree, click the **Calculate** icon and repeat the calculation using **Best tree** mode.

Displaying existing phylogenetic trees

When a phylogenetic tree has been calculated, it can be saved along with its associated multiple alignment file, and retrieved when the file is opened.

To display an existing phylogenetic tree

1. With the required alignment active and the MSA Editor view selected, click the **Phylogeny** icon on the MSA Editor toolbar.

The tree is displayed in a Tree Viewer window.

Displaying ClustalW guide trees

The ClustalW alignment algorithm generates a guide tree which is then used to specify the order in which sequences are aligned (see *“Performing Multiple Sequence Alignment”* on page 341). This may not be a true phylogenetic tree, because it is based on local pairwise alignments. However, the tree’s appearance may sometimes help you to judge the reliability of a ClustalW alignment.

To display a ClustalW guide tree

ClustalW guide trees are saved with their associated multiple alignment file, and you can view them when the file is retrieved.

1. With the ClustalW alignment active and the MSA Editor view selected, click the **Views** icon on the MSA Editor toolbar or choose **Analyze | ClustalW Alignment** from the main menu.

The **Alignment Views** dialog box is displayed.

2. In the **Picture Display** panel, select the **Guide Tree** checkbox, and select **OK**.

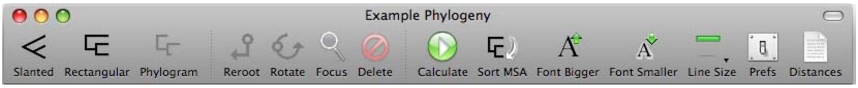
The ClustalW guide tree is displayed in the MSA Guide Tree view.

Note. You cannot recalculate a ClustalW tree. When a ClustalW tree is displayed in the MSA Guide Tree view, the controls for recalculating and deleting are not available.

Tree Viewer window

The Tree Viewer window can be used to display phylogenetic trees.

It includes tools for rooting trees to emphasize the direction of evolution, deleting parts of the tree, rotating nodes to re-order the sequences, and formatting the tree diagram for publication.



The Tree Viewer has a toolbar containing icons that are used to perform the following functions:

Note. The Tree Viewer toolbar, like all toolbars in MacVector, can be customized. Right-click on the toolbar to access this functionality. The tools described below are those that appear in the default Tree Viewer toolbar. Some of these tools may be absent and other tools may be present depending on your settings.

The **Slanted**, **Rectangular** and **Phylogram** icons enable you to control the type of tree displayed:

- slanted cladogram
- rectangular cladogram
- phylogram

The **Reroot** icon enables you to select an outgroup. See.*“Rooting”* on page 369

The **Rotate** icon enables you to change the order in which two nodes are arranged on the screen. See *“Rotating nodes”* on page 372.

The **Focus** icon enables you to display only the part of the tree you have selected. To re-display the whole tree, double-click anywhere in the Tree Viewer window.

The **Delete** icon enables you to delete the part of the tree you have selected. See *“Deleting nodes”* on page 371.

The **Calculate** icon provides access to the Phylogenetic Reconstruction dialog box, which can be used to recalculate phylogenies.

The **Sort MSA** icon enables you to sort the alignment to match the phylogenetic tree, when the tree has been calculated for a multiple alignment. See *“Sorting a multiple aligned sequence to match a tree”* on page 372.

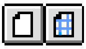
The **Font Bigger** and **Font Smaller** icons enable you to adjust the font size in the Tree Viewer window. Each click increases or decreases the font size by one increment.

The **Line Size** icon enables you to select the weight of the lines in the Tree Viewer display. Click the **Line Size** icon and choose the required line thickness from the drop-down menu.

The **Prefs** icon provides access to the **Tree Viewer Options** dialog box. See *"Setting the tree display options"* on page 373 for more information about the options available in this dialog box.

The **Distances** icon enables you to

At the bottom of the Tree Viewer window there are controls for magnification and page mode.



Use the **page mode** button to either show or hide dotted lines marking the page boundaries for printed output.

100%

The **view scale** controls let you set a magnification for viewing the tree, either by choosing from a drop-down menu or by typing in a text box. The current viewing magnification is displayed.

Tip. The width of a tree display is set to fit the current paper size. If you want to make a tree wider, use **File | Page Setup** to select a larger paper size or a landscape orientation

Editing Trees

You can edit a phylogenetic tree in a Tree Viewer window by rooting, rotating and deleting, using the icons in the toolbar. You can also format the fonts and lines used in the display, and control the display of branch and node labels. When a tree is saved along with the source MSA document, any editing you have applied to it is retained.

Note. If you have deleted part of a tree, the phylogeny will have been recalculated, so it no longer represents the full MSA document.

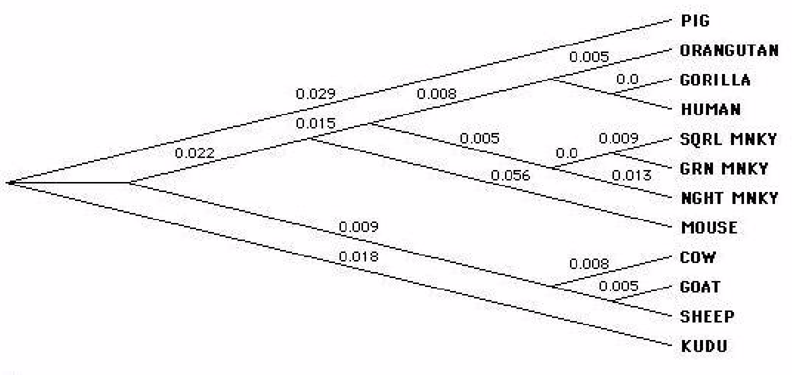
ClustalW guide trees vs. Phylogenetic trees

There are important differences in the way MacVector treats ClustalW guide trees and phylogenetic trees. Both can be edited by rooting and rotating, but ClustalW trees cannot be changed by deleting, because they cannot be recalculated. Phylogenetic trees retain all editing and formatting changes, when their associated MSA documents are saved: ClustalW guide trees, on the other hand, lose all editing changes when their MSA documents are saved.

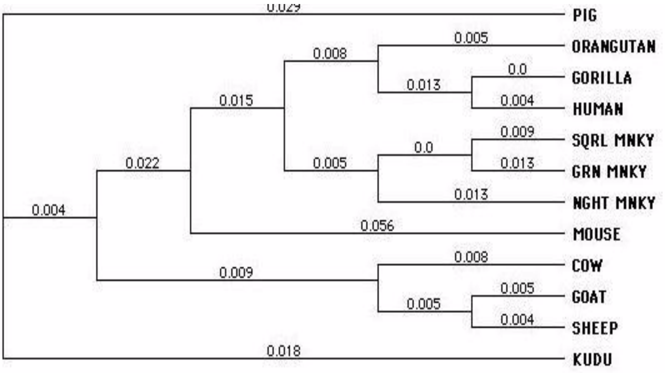
Tree shapes

The same phylogenetic data can be displayed in any of three styles:

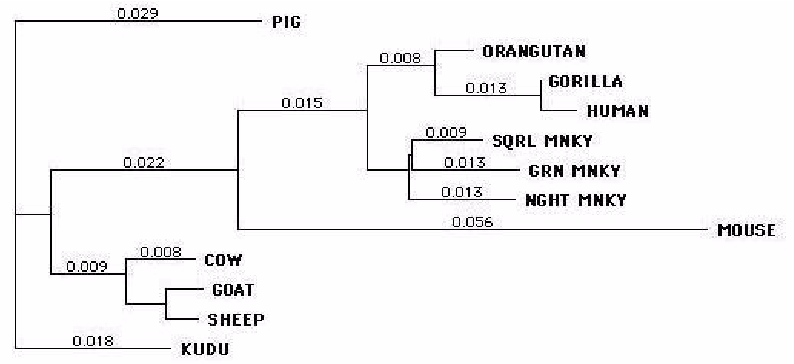
- *Slanted cladogram*



- *Rectangular cladogram*



- *Phylogram*



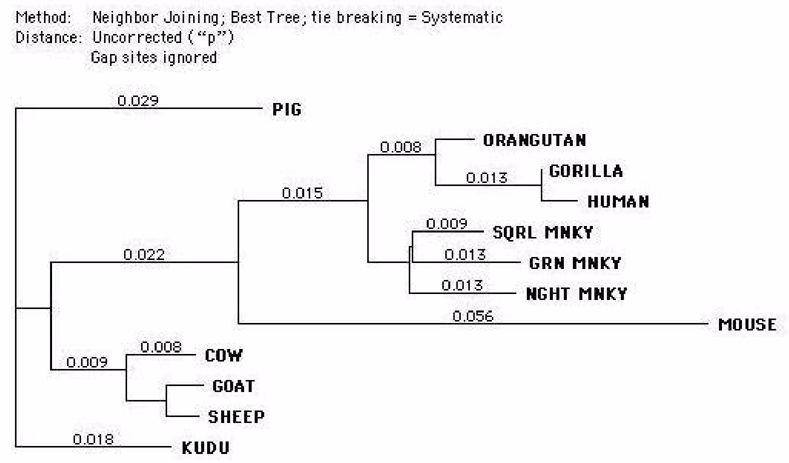
Only the phylogram display mode draws branch lengths to scale, indicating evolutionary distance. This mode is not available for bootstrap consensus trees.

Rooting

An important property of a tree is its polarity. The neighbor joining method of generating trees makes no assumption about which sequence is closest to the ancestral sequence, and in which direction evolution has proceeded. This can often be clarified by rooting the tree. Rooting affects only the appearance of the tree, not the underlying phylogenetic analysis.

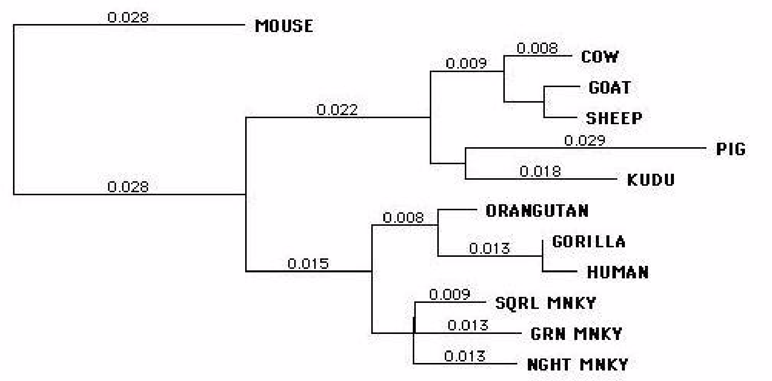
You can recognize an unrooted tree by its first (leftmost) junction, which is always trifurcated. In a rooted tree, the first junction is always bifurcated - it separates either the outgroup from the rest (outgroup rooting) or two aggregates from each other (midpoint rooting).

Consider this ***unrooted tree***, generated by neighbor joining:



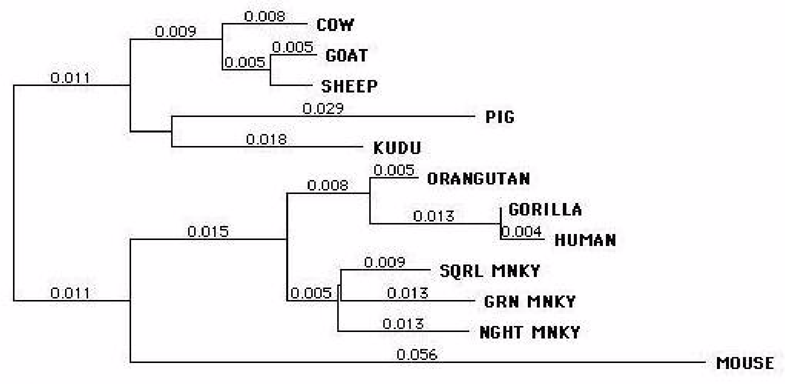
This tree does not make it clear that the primates are all closely related.

Here is the same phylogenetic analysis, but shown ***rooted on an outgroup*** (the mouse):



This tree makes it clearer that the ungulate and primate sequences form two well-defined groups.

Where there is no obvious outgroup, ***midpoint rooting*** may be useful. This method works by sampling each internal node, calculating the sum of all branch lengths under the nodes to the left and right of it. Where these two sums are most nearly equal, the midpoint is identified. By identifying its “center of gravity”, midpoint rooting helps to distinguish two groups in this tree:



You can root a tree from a selected outgroup by using the **Reroot** icon. Alternatively you can specify outgroup rooting in the Tree Viewer Options dialog box, in which case MacVector will root the tree on its longest terminal node. To specify midpoint rooting, use the **Tree Viewer Options** dialog box.

The rooting of trees generated using UPGMA cannot be changed. ClustalW guide trees can be rooted.

To root a tree on an outgroup

1. In the Tree Viewer window, select the required outgroup by clicking on its branch.

The branch becomes highlighted, and the **Reroot** icon becomes active.

Tip. Ensure that only one sequence is highlighted. You can only root the tree on a single sequence.

2. Click the **Reroot** icon or choose **Analyze | Phylogenetic Analyses | Root Tree on Node**.

The tree is shown rooted on the selected outgroup.

You can also specify outgroup rooting by using the **Tree Viewer Options** dialog box (see *“Setting the tree display options”* on page 373).

Deleting nodes

You can delete a node from a phylogenetic tree. The tree is recalculated automatically, excluding all sequences contained in the node.

To delete a node from a tree

1. In the Tree Viewer window, select a node to delete by clicking on it.

The node, together with its sub-nodes, becomes highlighted. The **Delete** icon becomes active.

2. Click the **Delete** icon or choose **Analyze | Phylogenetic Analyses | Delete Node**.

The tree is recalculated, with the node deleted.

Note. To undo a deletion, choose **Edit | Undo Delete**.

Rotating nodes

You can re-order the sequences in any tree diagram by rotating selected nodes through 180^o. Rotation flips the order of all sequences contained in the selected node. Rotating affects only the appearance of the tree, not the underlying phylogenetic analysis.

To rotate nodes on a tree

1. In the Tree Viewer window, select a node to rotate by clicking on the branch above it.

The node, together with its sub-nodes, becomes highlighted. The **Rotate** icon becomes active.

Note. The **rotate** icon only becomes active when more than one sequence has been selected.

2. Click the **Rotate** icon or choose **Analyze | Phylogenetic Analyses | Rotate Node**.

The node is rotated, flipping the order of its sequences.

3. Repeat steps 1 and 2 as required to sort the tree.

Sorting a multiple aligned sequence to match a tree

If a phylogenetic tree has been calculated for a multiple alignment, you can sort the alignment to match the tree. This can be a rapid way to sort sequences into phylogenetic groups.

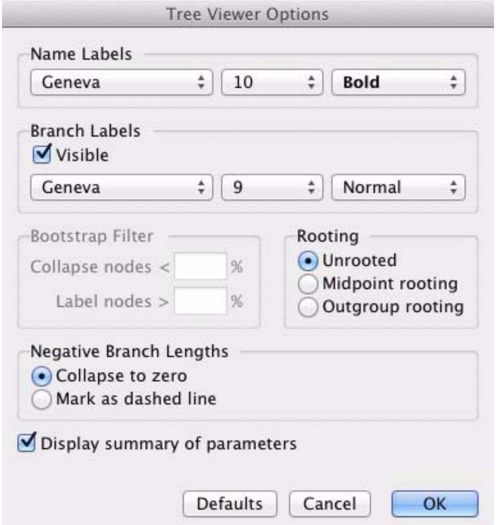
To sort the MSA by tree order

1. With the Tree Viewer window active click the **Sort MSA** icon or choose **Analyze | Phylogenetic Analyses | Sort MSA by Tree**.

The multiple alignment is sorted to match the tree sequence order.

Setting the tree display options

The Tree Viewer display can be controlled using the toolbar buttons (see *"Tree Viewer window"* on page 365). More advanced settings can be made by using the **Tree Viewer Options** dialog box.



Tree Viewer Options

Name Labels

Geneva | 10 | Bold

Branch Labels

☑ Visible

Geneva | 9 | Normal

Bootstrap Filter

Collapse nodes < %

Label nodes > %

Rooting

● Unrooted

○ Midpoint rooting

○ Outgroup rooting

Negative Branch Lengths

● Collapse to zero

○ Mark as dashed line

☑ Display summary of parameters

To set the display options for the Tree Viewer window

1. Click the **Prefs** icon on the Tree Viewer toolbar.

The **Tree Viewer Options** dialog box is displayed.

2. To adjust the appearance of the sequence name labels, use the font, size, and style drop-down menus in the **Name Labels** panel.
3. To adjust the appearance of the branch labels, use the controls in the **Branch Labels** panel:
   - The **Visible** check box controls whether or not the labels are displayed. By default the check box is selected and the labels are displayed
   - To change the appearance of the branch labels, use the font, size, and style drop-down menus.

4. When a bootstrap tree is being displayed, you can use the controls in the **Bootstrap Filter** panel to control the display threshold values:

   - Use the **Collapse nodes <** text box to set the threshold percentage occurrence for a node to be displayed in the bootstrap tree. Nodes occurring less frequently than this value will be collapsed.
     You may enter any value from 50 through 100. The default is 50%.
   - Use the **Label nodes >** text box to set the threshold for labeling a node's percentage occurrence in the bootstrap tree. Nodes occurring less frequently than this value will not be labeled.
     You may enter any value from 50 through 100, provided it is not less than the **Collapse nodes** value. The default is 70%.

   The **Bootstrap Filter** controls are disabled when there is no bootstrap tree in the window.

5. To change the rooting of a tree generated using the neighbor joining method, select the appropriate **Rooting** radio button:

   - **Unrooted**: the tree will not be rooted. This is the default setting for neighbor joining trees
   - **Midpoint rooting**: the tree will be rooted on its center of gravity
   - **Outgroup rooting**: the tree will be rooted on its longest branch

   If you have already rooted the tree manually using the **Reroot** icon, then selecting this radio button has no effect.

   Rooting controls are disabled if the current tree was calculated using UGPMA, because that method always roots the tree.

6. To adjust the rule for displaying any branches with calculated lengths that are negative, do one of the following in the **Negative Branch Lengths** panel:

   - Select **Collapse to zero** if you want to collapse such branches in the display. This is the default setting
   - Select **Mark as dashed line** to indicate negative lengths by dashed lines.

   Negative Branch Lengths controls are available only when a **phylogram** display has been selected.

7. To display details of the methods used to calculate the phylogeny, select the **Display summary of parameters** check box.

8. Select **OK** to apply the settings and close the dialog box.

Alternatively, select **Defaults** to restore the program default settings, or **Cancel** to close the dialog box without applying or saving any changes.

Saving and printing tree diagrams

Both phylogenetic trees and ClustalW guide trees can be saved as pictures or printed.

Saving trees

If a phylogenetic tree has been generated, it will be saved when you save its associated MSA document, so that when you reload the MSA file and select the **Phylogeny** button, a Tree Viewer window will open and display the tree. Any rooting, rotating, deleting and reformatting you have done is retained when you save the MSA document.

In addition, it is often useful to save an image of the tree as it appears in the Tree Viewer window.

To save a phylogenetic tree diagram

1. With the Tree Viewer window active, choose **File | Print**.

The **Print** dialog box is displayed.

2. Click the **PDF** button and choose **Save as PDF** from the menu.
3. Navigate to the folder where you want to save the graphic data and choose **Save**.

Copying trees

When the Tree Viewer window is active, you can use **Edit | Copy** to copy the tree to the clipboard. You can then paste it into another application as a picture.

Printing trees

An important function of the Tree Viewer is to provide publication-quality printed output. To keep the interface simple to use, the phylogenetic tree display is always scaled to fit the width of the selected paper size. You can set the page size with **File | Page Setup**, and view the page boundaries in the Tree Viewer window, using the **page mode** button in the bottom left corner.

Using the controls in the **Tree Viewer Options** dialog box, you can adjust the line width and font sizes so that both large and small trees are presented clearly.

To print a phylogenetic tree

1. With the Tree Viewer window active, choose **File | Print**.

The **Print** dialog box is displayed.

2. Set the printer controls, and select **Print** to print the tree.

19 Sequence Assembly

Overview

This chapter describes the Assembler plug-in module, including a Quick Start guide and an in depth Tutorial. Additional documentation on the `phred`, `cross_match` and `phrap` algorithms can be found in the `MacVector 12.6/Documentation/` folder.

Contents

Introduction

Assembler is an add-on module for MacVector. It allows you to create *de novo* assemblies using the industry standard `phred`, `cross_match` and `phrap` algorithms and reference assemblies using Bowtie.

Assembler integrates tightly into MacVector so that they appear as a single application. This means that when you create assembled contigs, you can immediately analyze the sequences using any of MacVector's DNA analysis or manipulation functions.

This chapter includes a quick start section which outlines the basics of creating assemblies (see *"Quick start"* on page 379), more detailed descriptions of both *de novo* (see *"de novo Assembly"* on page 381) and reference assembly (see *"Reference Assembly"* on page 386), and a tutorial that guides you through the process of *de novo* assembly of sequences using Assembler (see *"Tutorial"* on page 394).

Glossary

There are a few terms that are used in contig assembly that you may not be familiar with;

Term	**Description**
Base Call	The interpretation of the peaks in a trace file to identify the most likely DNA sequence.
Chromatogram	The trace information from an automated sequencing machine. The terms "chromatogram files" and "trace files" are used interchangeably in this tutorial to describe the files generated by automated sequencing machines.
Consensus	The most likely sequence of a contig, determined by taking all of the overlapping sequences into account.
Contig	An assembly of two or more overlapping sequences.
Quality Value	A value assigned to a base call to reflect the probability that the base call is in error. Uses a scale from 0 to 99.
Reads	A generic term used to describe the collection of DNA sequences that were generated during the sequencing project to be assembled into a contig. Typically these are trace files, but they can also be plain sequences.

Trace	The chromatogram data generated by an automated sequencing machine. The terms “chromatogram files” and “trace files” are used interchangeably in this chapter to describe the files generated by automated sequencing machines.

Quick start

To create a *de novo* assembly

1. Chose **File | New | Assembly Project** to create a new empty project file.
2. Click the **Add Seqs** icon on the toolbar, then select the sequence files you wish to assemble and click on the **Open** button.

Tip. You can hold down the **<shift>** key to select multiple sequences to import.

3. Click the **Phred** icon on the toolbar or choose **Analyze | Base Call (phred)** from the menu to run the `phred` algorithm on all of the sequences in the project.
4. Optionally, click the **CrossMatch** icon on the toolbar or choose **Analyze | Vector Trim (cross_match)** from the menu to mask vector sequences in the reads.

You must import the vector sequences you used into the **Vectors** tab on either the **Cross_match Parameters** or **Project Preferences** dialogs. The files must be in either MacVector or FastA format.

Note. Steps 3 and 4 above are applicable only to trace file reads. They are not applicable to Fastq reads.

5. Click the **Phrap** icon on the toolbar or choose **Analyze | Assemble (phrap)** from the menu to assemble all of the sequences of the project.
6. Double-click on a contig to open it in a Contig editor.

Tip. You can run any MacVector nucleic acid analysis function directly on the contig consensus sequence from the Contig editor.

7. Finally you can save the consensus sequence in MacVector format by choosing **File | Save As…** from the Contig editor. You can also save the assembly project itself at any time and you will be prompted to save any changes when you close the project window.

To create a *de novo* assembly of short reads

1. Chose **File | New | Assembly Project** to create a new empty project file.

2. Click the **Add Seqs** icon on the toolbar, then select the sequence files you wish to assemble and click on the **Open** button.

Tip. You can hold down the **<shift>** key to select multiple sequences to import.

3. Click the **Phrap** icon on the toolbar or choose **Analyze | Assemble (phrap)** from the menu.
4. Click **Short Read Defaults** and choose the appropriate settings for assembling short reads.
5. After assembly, contigs.will appear in the Project list. Double-click on a contig to open it in a Contig editor.
6. You can run any MacVector nucleic acid analysis function directly on the contig consensus sequence.
7. Finally you can save the consensus sequence in MacVector format by choosing **File | Save As...** from the Contig editor. You can also save the assembly project itself at any time and you will be prompted to save any changes when you close the project window.

To create a reference assembly

1. Chose **File | New | Assembly Project** to create a new empty project file.
2. Click the **Add Seqs** icon on the toolbar, then select the sequence files you wish to assemble and click on the **Open** button.

Tip. You can hold down the **<shift>** key to select multiple sequences to import.

3. Click on the **Add Ref** icon on the toolbar, then select the sequence file you wish to align the reads against and click on the **Open** button.
4. Click the **Bowtie** icon on the toolbar or choose **Analyze | Bowtie...**.
5. Click **OK** to create a reference assembly on all the sequences in the project using the default Bowtie parameters.

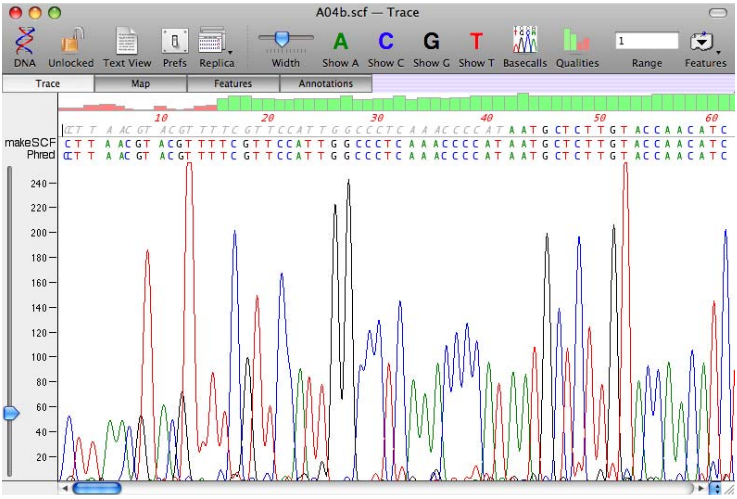
Note. If no sequences are selected, then Bowtie is run on all the sequence files in the project. To override this default behavior, select the reference sequence and one or more reads files to run Bowtie only on the selected sequences.

6. After assembly, contigs.will appear in the Project list. Double-click on a contig to open it in a Contig editor.
7. You can run any MacVector nucleic acid analysis function directly on the contig consensus sequence.
8. Finally you can save the consensus sequence in MacVector format by choosing **File | Save As...** from the Contig editor. You can also save the assembly project itself at any time and you will be prompted to save any changes when you close the project window.

de novo Assembly

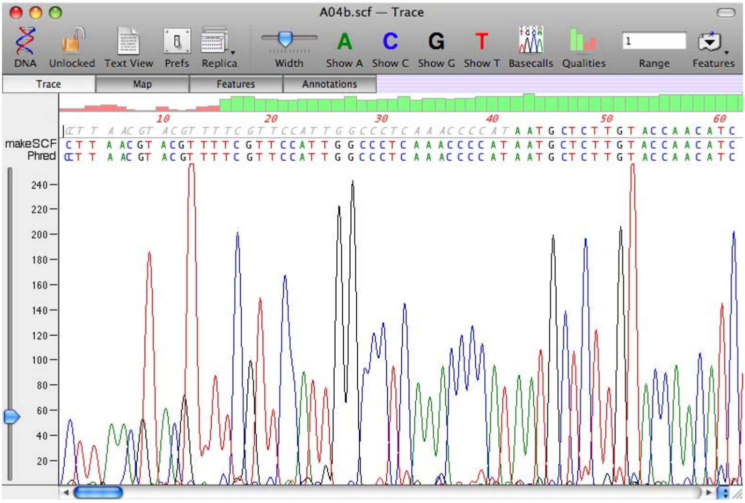
Base calling using `phred`

`Phred` is an algorithm that takes chromatogram information from an automated sequencing run and re-evaluates the peaks to produce a "base call" that is usually significantly more accurate than the original call. In addition to recalculating the residues, `phred` also adds quality score information to each residue. This is a logarithmic value from 0 to 99 where a value of 10 indicates that there is a 1 in 10 chance that the call is in error, a score of 20 indicates in 1 in 100 chance the call is in error, a score of 30 indicates a 1 in 1,000 chance of an error etc. Assembler takes advantage of multi-CPU machines (such as the Intel Core Duo machines) and splits up the `phred` jobs between the processors so that you see a speed up directly proportional to the number of processors. Once you have basecalled the imported sequences, you can view the `phred` base calls and the quality values by double-clicking on one of the sequences.



Trimming vector sequences using `cross_match`

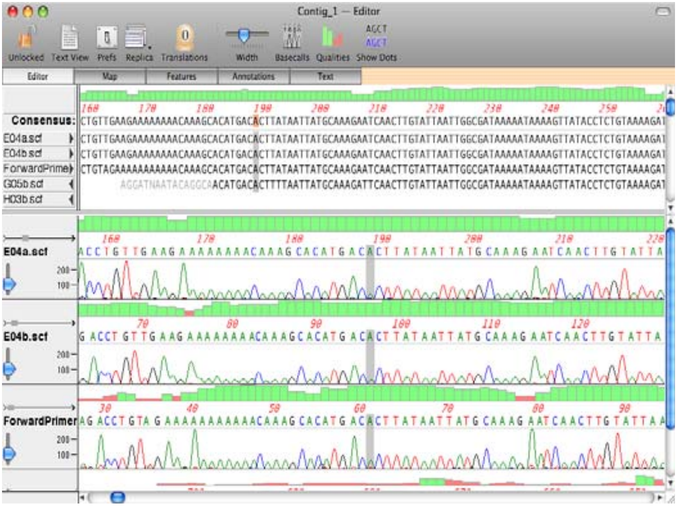
Many raw sequences from automated sequencing machines contain vector sequences at the beginning and/or end. Assembler lets you mask these out using the `cross_match` algorithm. To use this you just need to supply the sequences of the vector(s) you used for the cloning - there is no need to indicate the cloning site you used as cross_match can easily identify the exact position where the vector sequences terminate. MacVector Assembler splits up jobs between multiple CPUs if they are available. After processing by `cross_match`, you can view the masked vector sequences in the trace editor window where they appear in grayed out italics.



A trace editor window after vector sequences have been masked, indicated by gray italic text.

Assembling sequences using `phrap`

Assembler assembles sequences using the `phrap` algorithm. `phrap` does not require the sequences to have been base called by `phred`, or to have had any vector sequences masked. However, using `phred` and `cross_match` will improve the accuracy of `phrap` assemblies. `phrap` assembles sequences into contigs and creates a consensus sequence with its own set of quality values, based on the quality and strandedness of the overlapping sequences. Contigs can be viewed and edited in a Contig editor that shows the aligned sequences along with the chromatograms in a lower pane. Clicking on a residue in the consensus sequence resets the chromatogram view so that they are all aligned to that base. This allows you to easily align the chromatograms so that you can resolve ambiguities in the consensus sequence.



```
           160       170       180       190       200       210       220       230       240       250
Consensus: CTGTTGAAGAAAAAAAACAAAGCACATGACACTTATAATTATGCAAAGAATCAACTTGTATTAATTGGCGATAAAAATAAAAGTTATACCTCTGTAAAAGAT
E04a.scf   CTGTTGAAGAAAAAAAACAAAGCACATGACACTTATAATTATGCAAAGAATCAACTTGTATTAATTGGCGATAAAAATAAAAGTTATACCTCTGTAAAAGAT
E04b.scf   CTGTTGAAGAAAAAAAACAAAGCACATGACACTTATAATTATGCAAAGAATCAACTTGTATTAATTGGCGATAAAAATAAAAGTTATACCTCTGTAAAAGAT
ForwardPrime CTGTAGAAAAAAAAAAACAAAGCACATGACACTTATAATTATGCAAAGAATCAACTTGTATTAATTGGCGATAAAAATAAAAGTTATACCTCTGTAAAAGAT
G05b.scf        AGGATNAATACAGGCAACATGACACTTTTAATTATGCAAAGATTCAACTTGTATTAATTGGCGATAAAAATAAAAGTTATACCTCTGTAAAAGAT
H03b.scf
```

A complete list of the parameters used by `Phrap` Assembly and their default values can be found in Appendix I, *“Phrap Assembly Parameter Contraints”*.

de novo assembly of short reads

Assembler can assemble short read data of the type generated by next generation sequencers.

Next generation sequencers are capable of producing a very large amount of data, far greater than is seen with traditional Sanger sequencers. Assembler can generate *de novo* assemblies of up to 1.5 million short reads using `phrap`. However, such assemblies require very large amounts of RAM - much greater than is commonly found on even very recent Mac desktops and laptops. Therefore, the size of assemblies that can realistically be performed on a typical iMac or Macbook Pro is somewhat less.

`phrap` is also very CPU intensive, so Assembler will submit jobs with reduced priority. This will allow you to continue using your Mac and MacVector for normal work, as `phrap` will take less of the CPU. However, if you run any other CPU intensive jobs, then the assembly will take much longer.

Here are some examples of time taken to assemble a sample of 262,000 reads with an average length of 200bp. This has the accession number of SRR015579, taken from the NCBI's Short Read Archive

Assembler produced 408 contigs with an average length of 16,396bp

- Less than 30 minutes on a Mac Pro with two dual core CPUs and 8Gb of RAM.
- 6 hours on a MacBook Pro 3Gb of RAM
- >8 Days on an Intel iMac with 1Gb of RAM

A sample of 233,000 reads with an average length of 200bp. This has the accession number of SRR015575, taken from the NCBI's Short Read Archive

Assembler produced 48 contigs with an average length of 100,396bp

- Less than 30 minutes on a Mac Pro with two dual core CPUs and 8Gb of RAM.
- 12 hours on a MacBook Pro 3Gb of RAM

Fastq format

Assembler supports short read data in the Fastq format. This is a widely accepted format, that only contains basecalled sequence and quality data for each read. Fastq does not contain any of the image data that raw read formats (such as SRF and SFF) contain and is therefore small enough to be used practically on a desktop Mac. For example an SRF file that contained the same number of reads as a 200Mb Fastq file would likely exceed 20Gb in size.

Most sequencers' software supports the export of short read data in the Fastq format. Additionally the Short Read Archives of the EBI and NCBI only allow data to be downloaded in Fastq format.

Reads in the Fastq format may also contain a quality score for each read. There are three variations of encoding and two different quality scoring schemes commonly found in Fastq format.

- The original and most common format contains the standard `Phred` quality score. The `Phred` quality score has a range of 0 to 93. Since this is a two digit score it is encoded by using the ASCII characters codes from 33 to 126, and so it is commonly called `Phred`33.
- The second most common quality score found in Fastq files is the Illumina 1.3 format that is also a `Phred` quality score from 0 to

40. This is encoded using the ASCII characters from 64 to 104, and so is commonly called `Phred`64.

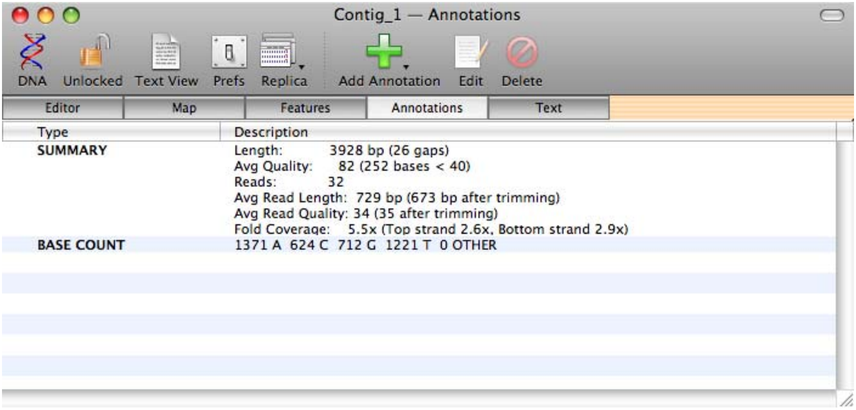
- The third format is the Solexa/Illumina 1.0. This format is now deprecated. It uses a Solexa/Illumina quality score from -5 to 40 using ASCII 59 to 104.

MacVector will prompt you to provide the format of the quality data when you import a Fastq format file.

Generally, the reads contained in Fastq format file have already been basecalled from the raw images generated initially by the sequencer. Therefore, it is not necessary nor possible to run `Phred` on assembly projects consisting wholly of Fastq format data.

Editing and analysis

You can edit the sequences in a contig and the consensus will be updated automatically. MacVector follows the `phred`/`phrap` quality value rule where edited residues are given a quality value of 99 to indicate they have been assigned by a user. These are shown in blue in the quality display. You can view a variety of statistics on the composition of the contig in the annotations window. The annotations window contains a summary of the contig composition statistics



Type	Description
SUMMARY	Length: 3928 bp (26 gaps) Avg Quality: 82 (252 bases < 40) Reads: 32 Avg Read Length: 729 bp (673 bp after trimming) Avg Read Quality: 34 (35 after trimming) Fold Coverage: 5.5x (Top strand 2.6x, Bottom strand 2.9x)
BASE COUNT	1371 A 624 C 712 G 1221 T 0 OTHER

You can invoke any MacVector DNA analysis algorithm from the contig editor window, including online NCBI blast searches. Any gaps in the consensus are removed before the analysis is performed. This lets you scan the contig for restriction enzyme sites, then edit the consensus and rescan without having to export the consensus sequence or switch to a different module.

Note. It is not possible to edit a contig assembled from file-based reads.

The consensus sequence, without gaps, can be saved to disk as a single sequence at any time. The sequence can be saved in any format supported by MacVector and retains a list of the individual reads used to generate the consensus.

de novo Assembly algorithms

`Phred` uses simple Fourier methods to examine the four base traces in the region surrounding each point in the data set in order to predict a series of evenly spaced predicted locations. That is, it determines where the peaks would be centered if there were no compressions, dropouts, or other factors shifting the peaks from their "true" locations.

Next `phred` examines each trace to find the centers of the actual, or observed, peaks and the areas of these peaks relative to their neighbors. The peaks are detected independently along each of the four traces so many peaks overlap. A dynamic programming algorithm is used to match the observed peaks detected in the second step with the predicted peak locations found in the first step.

`Phred` evaluates the trace surrounding each called base using four or five quality value parameters to quantify the trace quality. It uses a quality value lookup table to assign the corresponding quality value. The quality value is related to the base call error probability by the formula

QV = - 10 * log_10(P_e)

where P_e is the probability that the base call is an error.

`Phred` uses data from a chemistry parameter file called 'phredpar.dat' in order to identify dye primer data. For dye primer data, phred identifies loop/stem sequence motifs that tend to result in CC and GG merged peak compressions. It reduces the quality values of potential merged peaks and splits those peaks that have certain trace characteristics indicative of merged CC and GG peaks. In addition, the chemistry and dye information are passed to `phrap`.

Reference Assembly

MacVector Assembler can also perform reference assembly, in which a reference or scaffold sequence is used to assemble the reads, with Bowtie. Bowtie (Langmead *et al.*, 2009) is an ungapped aligner capable of extremely fast alignments of short sequences against a much larger reference sequence.

The interface allows easy, point and click, assembly of reads against a reference and generates reports of SNPs and other variants with Variant Calling pages. It supports Variant Calling Format (VFC) and BAM/ SAM file formats. Reads can be assembled against multiple references.

Consensus and contig sequences can be exported in Fasta and Fastq formats for further analysis.

Performing reference assembly using Bowtie

1. Chose **File | New | Assembly Project** to create a new empty assembly project.
2. Click the **Add Seqs** icon on the toolbar, then select the reads files you wish to assemble and click on the **Open** button.

Tip. You can hold down the **<shift>** key to select multiple reads files to import. You can also drag and drop files onto the Assembly Project window.

Reads in a fastq file that have less than 10 bases, have mismatched quality and sequence lines, or have missing quality lines will be ignored.

Note. This filtering step is currently disabled for paired end reads.

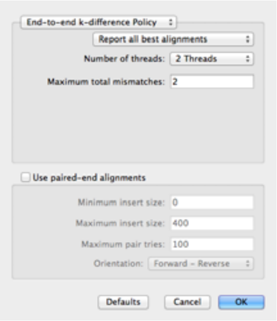
All fastq files are added as file based sequence collections, which store a reference to the original file rather than importing and storing the file. This is done to save disc space, as fastq files can be many gigabytes in size. If you move the original file you must use the **Locate** button in the Assembly Project window to restore the new filepath.

3. Click on the **Add Ref** icon on the toolbar, then select the sequence file you wish to align the reads against and click on the **Open** button.

The reference sequence is indicated by a reference icon to the left of its **Name** in the Assembly Project window and by the **Status** REF.

4. Click the **Bowtie** icon on the toolbar or choose **Analyze | Bowtie...**.

The Bowtie preferences dialog is displayed



End-to-end k-difference Policy

Report all best alignments

Number of threads: 2 Threads

Maximum total mismatches: 2

□ Use paired-end alignments

Minimum insert size: 0

Maximum insert size: 400

Maximum pair tries: 100

Orientation: Forward – Reverse

Tip. Users familiar with running Bowtie from the command line can find the equivalent command line parameter shown in brackets, along with the default parameter value, in all of the sub-sections below.

Hit Reporting

Bowtie uses a concept called strata to score alignments. A stratum is a set of reads that contain the same number of mismatches in the seed.

5. Choose the required hit reporting setting from the dropdown list. The following options are available:

- **Report all alignments** (-a) - Reports all alignments, subject to the alignment policy
- **Report the best alignment only** (--best) - Reports the best alignment in the stratum with the fewest mismatches.

Tip. Strata always trump quality so, for example a 1-mismatch alignment where the mismatched position has a Phred quality of 40 is preferred over a 2-mismatch alignment where the mismatched positions both have a Phred quality of 10.

- **Report all best alignments** (-a --best --strata, default) - Reports the best alignment in all strata.

Tip. The setting you choose depends on a number factors, for example, how many references you have, how many repeated regions you expect, whether you are using a reference sequence from the same organism or a related one, and others. Generally, it is best to start with **Report all alignments**, which is the quickest, and go from there.

6. Specify the **Number of threads** to use.

Bowtie can be run as a single or as mulitple threads. Using multiple threads will speed up the assembly process but may make your Mac less responsive for other tasks.

Type of Alignment

7. Choose the type of alignment from the dropdown list. The following options are available:

Quality Aware Policy (default)

When this option is selected the following policy settings are required to determine which alignments are valid:

- **Maximum mismatches in seed** (-n, default 2) - The maximum number of mismatches permitted in the specified **Seed length**.

- **Seed length** (-l, default 28) - The number of bases on the high-quality end of the read which constitute the seed. The lowest permitted setting is 5. Bowtie is faster when larger seed lengths are used.
- **Maximum total mismatch quality** (-e, default 70) - The maximum total quality value for all mismatched read positions throughout the alignment. Where qualities are unavailable (e.g. if the reads are from a Fasta file), the `Phred` quality defaults to 40.
- **Suppress quality rounding** (--nomaqround, default off) - Toggles on and off maximum quality rounding.

If there are many possible alignments satisfying these criteria, Bowtie gives preference to alignments with fewer mismatches and where maximum quality rounding is smaller. When hit reporting is set to either **Report all best alignments** or **Report the best alignment only** Bowtie guarantees the reported alignment(s) are the best in terms of these criteria and that the alignments are reported in best-to-worst order. Bowtie is somewhat slower when either of these hit reporting options are selected.

Bowtie is not fully sensitive when **Maximum mismatches in seed** is set to either 2 or 3. In these cases Bowtie imposes a backtracking limit to restrict the time spent trying to find valid alignments for low-quality reads that are unlikely to have any. This may cause Bowtie to miss some legal 2- and 3-mismatch alignments. The backtracking limit is set to 800 when hit reporting is set to either **Report all best alignments** or **Report the best alignment only** and 125 when it is set to **Report all alignments**.

End-to-end k-difference Policy

When this option is selected the following policy settings are required to determine which alignments are valid:

Maximum total mismatches (-v, default 2) - The maximum total number of mismatches permitted in an alignment. Valid values are 0 through 3.

Note. Quality values are ignored if this alignment mode is selected.

If there are many possible alignments satisfying these criteria, Bowtie gives preference to alignments with fewer mismatches. When hit reporting is set to either **Report all best alignments** or **Report the best alignment only** Bowtie guarantees the reported alignment(s) are the best in terms of this criterion and that the alignments are reported in best-to-worst order. Bowtie is somewhat slower when either of these hit reporting options are selected.

Paired End Alignments

8. Select whether to align paired-end reads can against a reference. The following must be specified to use this feature:

   - **Use paired-end alignments** - Toggles on and off the paired-end alignment feature. Paired-end alignment is enabled by default if two reads files are selected but disabled if three or more files are selected. However, it is possible perform paired-end alignment when more than two reads files are selected. If you do so you must ensure that he read files are sequentially numbered so that when they are submitted the pairs are together. For example “READSFILE_A_1.fastq", "READSFILE_A_2.fastq", "READSFILE_B_1.fastq" and "READSFILE_B_2.fastq" will work for two pairs called READSFILE_A and READSFILE_B.
   - **Minimum insert size** (-i, default: 0) - The minimum insert size for valid paired-end alignments. For example, if 60 is specified and a paired-end alignment consists of two 20-bp alignments in the appropriate orientation with a 20-bp gap between them, then that alignment is considered valid but a 19-bp gap would not be valid.
   - **Maximum insert size** (-X, default: 250) - The maximum insert size for valid paired-end alignments. For example, if 100 is specified and a paired-end alignment consists of two 20-bp alignments in the proper orientation with a 60-bp gap between them, that alignment is considered valid but a 61-bp gap would not be valid.
   - **Maximum pair tries** (--pairtries, default 100) - The maximum number of attempts Bowtie will make to match up an alignment for one mate in a pair with an alignment for the other mate in that pair.

Tip. Most paired-end alignments require only a few such attempts, but pairs where both mates occur in highly repetitive regions of the reference can require significantly more. Setting this value to a higher number allows Bowtie to find more paired-end alignments for repetitive pairs but the assembly will take more time.

- **Orientation** - Choose from the following orientation options for paired-ends:
  - **Forward-Reverse** (fr): The paired reads are on opposite strands and pointing towards each other.
  - **Forward-Forward** (ff): Both reads are on the same strand

- **Reverse-Forward** (rf): The paired reads are are on opposite strands and pointing away from each other.

Tip. Most, but not all, Illumina paired end reads are **Forward-Reverse**. The Mate Pair protocol in Illumina produces **Reverse-Forward** pairs and Solid reads are always **Forward-Forward**.

9. Click **OK** to create a reference assembly on all the sequences in the project using the specified Bowtie parameters.

Note. If no sequences are selected, then Bowtie is run on all the sequence files in the project. To override this default behavior, select the reference sequence and one or more reads files to run Bowtie only on the selected sequences.

Tip. The parameters used to produce the Bowtie alignment are stored in the Comment Annotations field of each Reference Contig.

Viewing, analyzing and exporting results

Viewing and analyzing contigs

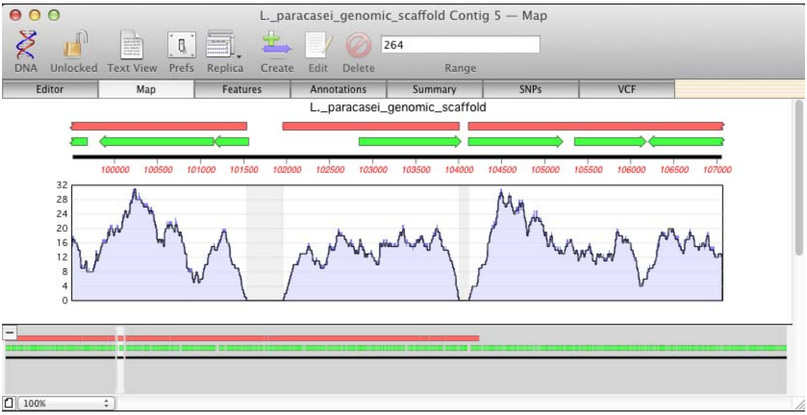
In the Assembly Project window, double-click on the reference contig to open it in the Contig Editor. To rename a reference contig option-click on the reference contig name.

Click on the disclosure triangle next to the reference contig to reveal the individual child contigs.

Note. A child contig is defined as a region of the reference contig that is bounded by two regions without overlapping reads and at least a single base with no coverage (or either end of the reference sequence).

Double-click on a child contig to open it in the Contig Editor. The #, Start and Stop columns update to display additional information. The number of reads assembled in the contig is indicated on the top line, while the orientation of each read in the contig is indicated on the other lines. The start and stop locations of each child contig are also indicated in the child contig name.

Child contigs are annotated to the reference contig as a MISC_FEATURE. These annotations are saved if you export the reference contig as a single sequence in MacVector.



Coverage map

The **Map** view of a reference or child contig shows details of the depth of reads in a coverage map. The coverage map shows four statistics. A single plot line (default color black) shows a running average of the number of reads at that point. However, since an average plot is not very sensitive when viewed at a high level, two shaded areas indicate the maximum value (default color dark blue) and the minimum value (default color light blue) of the averaged reads at that point. As the coverage map is viewed at a greater magnification these three values will become increasingly similar, to the extent that when viewed at, or close to, residue level the three plots are more or less identical.

Areas of zero coverage are shown in light grey. These areas are always displayed even when they are disproportionate to the level of magnification. For example, a region of zero coverage will always be displayed even when you are viewing a 5 megabase contig in its entirety. Also, there are no areas of zero coverage in child contigs as by definition they are bounded by either end of the reference contig and/or an area of zero coverage.

Areas with lower than average coverage are generally the result of the base composition in that region. For example, regulatory elements in a sequence (where proteins such as transcription factor bind) have lower than average coverage, perhaps due to their low GC content.

Areas with excessively high coverage can be indicative of a repeat region that may not be present in the reference sequence. In such circumstances, reads are piled up on one of the repeated regions, rather than being spread out over all of them. Paired-end reads can help to detect these and allow you to align reads correctly.

Note. For performance reasons individual reads are not shown in the **Map** view.

Tip. The coverage map makes it easy to design primers for further sequencing, such as Sanger sequencing for hybrid assembly.

Exporting contigs

The **File | Export as...** menu option allows you to export contigs and consensus sequences in Fasta or Fastq format, as follows:

- From the Assembly Project window, if the reference contig is selected it will save a Fasta/Fastq file containing all child contigs and no reference contig sequence or reference contig consensus. Selection of child contigs is ignored.
- From the Assembly Project window, if only child contigs are selected it will save a Fasta/Fastq file containing all selected child contigs and no reference contig sequence or reference contig consensus.
- With the reference contig open in the Contig Editor, it will save a multiple sequence Fasta/Fastq file containing the reference sequence and the consensus.
- With a child contig open in the Contig Editor, it will save a single sequence Fasta/Fastq file with the consensus sequence.

Adding an existing BAM/SAM alignment to an assembly project

BAM/SAM format files are becoming ubiquitous for storing reference based alignments of NGS sequencing data. Like Fastq format, for storing unaligned reads, BAM/SAM files are becoming the *de facto* standard for NGS data.

To add a BAM/SAM alignment to an assembly project

1. Chose **File | New | Assembly Project** to create a new empty assembly project.
2. Click the **Add Ref** icon on the toolbar, then select the reference sequences you wish to align the imported reads against and click on the **Open** button.

3. Click the **Add Contig** icon on the toolbar, then select the required BAM/SAM file and click on the **Open** button.
4. Select both files in the assembly project, right-click on the selection and choose **Unite reference with consensus** from the context menu. Alternatively, select **Unite reference with consensus and generate VCF report** to add the selected BAM.SAM alignment and generate a VCF report at the same time.

Note. You can also generate a VCF report from these files at a later date.

Importing large BAM/SAM files and generating VCF reports from them may take a considerable amount of time.

Tutorial

Sample files

After installing MacVector Assembler, you will find example files for this tutorial in the folder:

`MacVector/Sample Files/SequenceConfirmation`

There are 32 trace files in SCF format along with two vectors (pSG933 and Tn1000) used in the sequencing experiment.

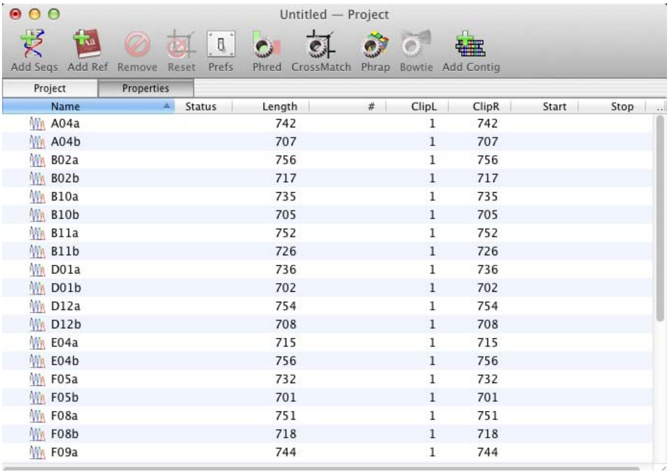
Creating and Populating a Project

The first step in the tutorial is to create a new project and add some sequences to it.

1. Select **File | New | Assembly Project**. An empty assembly project window will open.
2. Click the **Add Seqs** icon on the toolbar. In the dialog that opens, navigate to the `MacVector/Tutorial Files/Contig Assembly/Trace Files/` folder.
3. Click on the first file (`A04a.scf`) to select it, then scroll to the end of the list, hold down the **<shift>** key and click on the last file (`ReversePrimer.scf`) to select all of the files in the folder.
4. Finally, click on the **Open** button to import the selected files into the project.

Note. The data in the files is copied into the project. If you subsequently edit the original files on disk, then the data in the project will be unaffected. Similarly, any edits you make to the project data will not affect the contents of the original files.

There are no limits to the numbers or sizes of the sequences that you import. However, you may run into performance problems with projects containing large numbers of sequences. We recommend that you use a computer that has at least 0.5 MB of physical RAM for each chromatogram file you import for optimum performance i.e. use at least 512MB for a 1,000 sequence assembly. If you do see a slowdown importing, editing and saving large projects, adding more RAM to your computer is the most cost-effective way to improve performance.



Name	Status	Length	#	ClipL	ClipR	Start	Stop
A04a		742		1	742		
A04b		707		1	707		
B02a		756		1	756		
B02b		717		1	717		
B10a		735		1	735		
B10b		705		1	705		
B11a		752		1	752		
B11b		726		1	726		
D01a		736		1	736		
D01b		702		1	702		
D12a		754		1	754		
D12b		708		1	708		
E04a		715		1	715		
E04b		756		1	756		
F05a		732		1	732		
F05b		701		1	701		
F08a		751		1	751		
F08b		718		1	718		
F09a		744		1	744		

The Project window

The Project window contains two tabbed views: Project and Properties.

The Project view

The Project view has a number of columns that display information about the individual sequences and contigs. Most of the columns can be sorted by clicking on the column header.

- **Name** – the name of the sequence. All sequences and contigs in a project MUST have a unique name. If you try to import sequences with duplicate names, you will be prompted to choose how they should be handled. The icon next to the name indicates if the object is a contig, a trace or a plain sequence. You can directly edit this field to change the name.
- **Status** – initially blank, the status field indicates if a sequence has been base called with phred (”P”) or masked for vector sequences with cross_match (“X”).

- **Length** – the length of the sequence or contig.
- **#** - for contigs, this field indicates the number of reads that have been assembled. For sequences in a contig, the field indicates orientation using "->" for forward reads and "<-" for reverse reads.
- **ClipL** – the first residue from the 5' end that is not masked. Typically this will be "1", although cross_match or phrap may change this.
- **ClipR** – the last valid residue at the 3' end of a sequence. Initially, this is simply the last residue of the sequence, but cross_match and phrap may change this.
- **Start** – for sequences in a contig, the start location of the sequence within the contig.
- **Stop** – for sequences in a contig, the location of the last residue of the sequence within the contig.
- **Definition** – any descriptions associated with a sequence.

You can double-click on an item to open up the editor associated with the object, e.g. the trace editor or the contig editor. Note that in this version, you cannot directly edit plain sequences by double-clicking on them – you should complete any editing on these before adding them to the project.

The Project view also has the following toolbar buttons:

- **Add Seqs** – provides access to a dialog box which enables you to add additional sequences to the project. You can also use **Edit | Add Sequences From File**
- **Add Ref** - provides access to a dialog box which enables you to add a reference sequence to the project.
- **Remove** – removes the selected sequences from the project. You can also use this button to dissolve selected contigs. You can also use the **<delete>** key or **Edit | Clear** to accomplish the same functions.
- **Reset** – removes any clipping that has been performed on indiviual reads by Phrap.

Note. This option is not available for multiple reads files.

- **Prefs** – provides access to a dialog box which enables you to configure the default appearance of the Assembly editor and add vectors to the project.

Tip. Vectors can also be added to the project using the **Vectors** tab on the **Cross_match Parameters** dialog.

- **Phred** – runs the `phred` algorithm on all of the sequences in the project.
- **CrossMatch** – masks vector sequences in the reads.
- **Phrap** – assembles all of the sequences in the project using Phrap.
- **Bowtie** – assembles all of the sequences in the project using Bowtie.
- **Add Contig** – provides access to a dialog box which enables you to add BAM/SAM alignment files to the project.

The Properties view

The Properties view displays various useful statistics about the project.

Saving and opening assembly projects

You can save assembly projects at any time. They are saved in an xml format, meaning that the file contents are text and can (potentially) be viewed in any standard text editor such as TextEdit or Microsoft Word. The file names are not given file-type extensions.

1. Choose **File | Save**.

If this is the first time you have saved the project, you will get prompted for a filename. Otherwise, the project will be saved with its current filename.

The small tutorial project should save within a second or two. Large projects may take some time to save – approximately 15 seconds for every thousand trace sequences on an average machine. A progress dialog is displayed during the save. You can cancel this and your original file will not be affected.

2. Close the Project window.

You will be prompted to save if you have made any changes since the last save.

3. Click on the **File** menu.

A list of recent files is appended to the bottom of the menu.

4. Select the name you saved the project under.

The project will open. Again, a progress dialog is displayed during the load as large projects will take some time to open. If you have a very large project, it may take a few seconds before the progress dialog is displayed.

Base calling with **`phred`**

`Phred` is an algorithm developed by Phil Green's group at the University of Washington. `Phred` re-evaluates the chromatogram peaks in a trace file, a process known as "base calling". Not only is `phred` typically more accurate than the default base callers used by automated sequencing machines, but it also assigns "Quality Values" to each individual base call. `Phred` uses a statistically significant logarithmic scale from 0 to 99 where 10 means there is a 1 in 10 chance that the call is in error, 20 means there is a 1 in 100 chance the call is in error, 30 means there is a 1 in 1,000 chance of an error etc. The values 98 and 99 are reserved to indicate residues that have been edited by the user. A `phred` score of 20 or more is generally considered to be an acceptable score. MacVector Assembler displays `phred` scores as a histogram above the sequence using colors to indicate the quality – scores below 20 are shown in red, scores of 20 or greater in green and edited residues (score 99) in blue.

Make sure you have no selections in the Project view. To toggle a selection off, click on the selection while holding down the command (⌘) key.

You can select a subset of sequences for analysis if you wish. However, if nothing is selected, all of the chromatogram sequences in the project will be submitted for analysis.

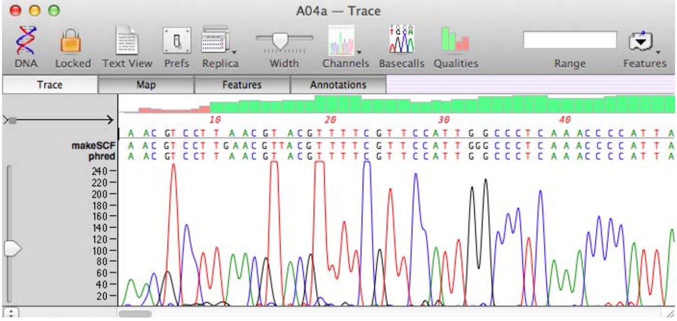
1. Click the **Phred** icon on the toolbar or choose **Analyze | Base Call (phred)** from the menu.

The **Job Manager** dialog box will appear. This allows you to follow the progress of the job.

When the job has completed, the project window will refresh to reflect the new base calls. The status of each entry changes to "P" to indicate you have run phred on the sequence.

Viewing base calls

Double-click on one of the `phred`-called sequences in the Project view to open up the Trace editor window.



```
        A ACGT CCT T AACGT ACGTTTTCGT TCCATT G GCCCTC AAACCCCATT A
makeSCF A ACGT CCT TGAACGTTACGTTTTCGT TCCATT GGGCCCTC AAACCCCATT A
phred   A ACGT CCT T AACGT ACGTTTTCGT TCCATT G GCCCTC AAACCCCATT A
```

When you open the Trace editor window from an assembly project, two additional icons are displayed in the toolbar.

- **Basecalls** – toggle this button to show or hide the basecalls displayed immediately below the main sequence line.
- **Qualities** – toggle this button to show or hide the quality histogram displayed over the top of the sequence.

You can see that the original base call for this sequence was generated using a utility called “makeSCF”. The `phred` base call is shown directly underneath this. You cannot edit the base calls – they are read-only. You can edit the upper sequence if you wish – this is the “active” editable sequence that is used in all assemblies and analyses. If you subsequently re-run `phred` on a sequence, it will replace the `phred` base call and will also replace any edits you have made to the active sequence.

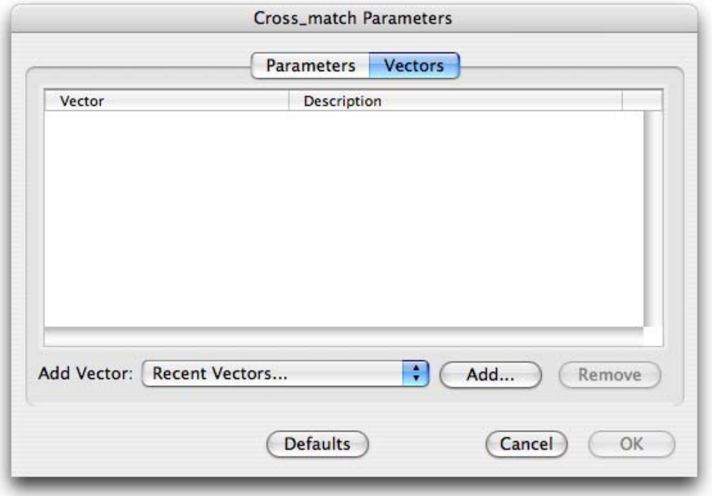
Masking vector sequences with `cross_match`

Typical sequencing projects use a directed or shotgun sub-cloning approach to generate short overlapping sequences that are then assembled into a single longer sequence. It is common for the reads to have vector sequences at the beginning and/or end which can interfere with the assembly. `cross_match` is an algorithm that can be used to mask out any vector sequences to prevent this interference. This is not an absolutely essential step as phrap (the assembly algorithm we will use) can often detect the vector sequences in a collection of similar sequences. However, using `cross_match` is highly recommended to reduce the likelihood of anomalous assemblies.

Make sure you have no selections in the Project view. To toggle a selection off, click on the selection while holding down the command (⌘) key.

1. Click the **CrossMatch** icon on the toolbar or choose **Analyze | Vector Trim (cross_match)** from the menu.

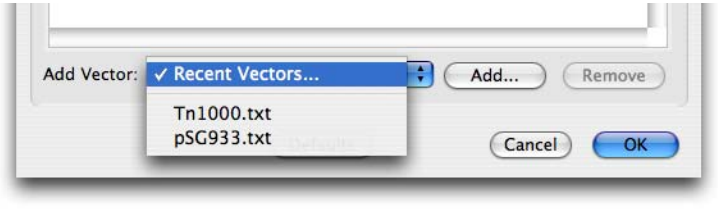
The **Cross_match Parameters** dialog box will appear. The algorithm needs to know which vectors were used in the sequencing experiments, so the dialog initially displays the empty **Vectors** tab.



2. Click **Add** to bring up the file selection dialog box.
3. Navigate to the `MacVector/Tutorial Files/Contig Assembly/` folder and select the files `pSG933.txt` and `Tn1000.txt`
4. Click **Open** to add the vector files to the **Vectors** tab.

The tab will refresh to reflect the new vectors that have been added.

5. Click on **Recent Vectors**.



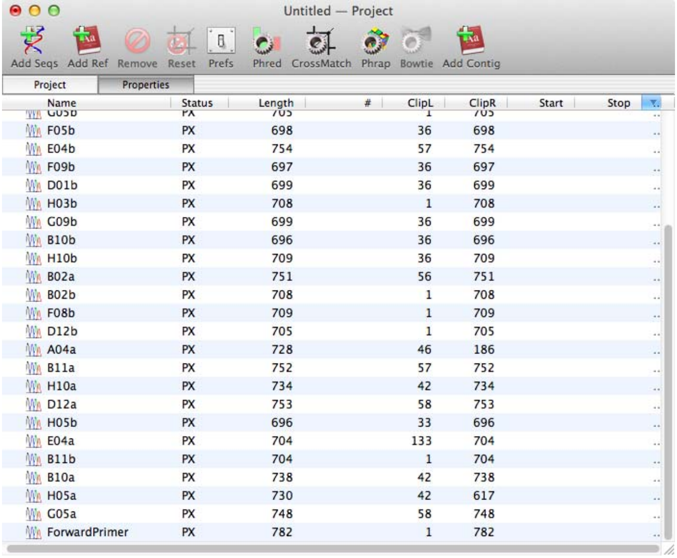
The names of the files have been added to this menu. The menu remembers the last 20 vector files you added to any project, so you can use this as a shortcut to rapidly import common vectors into any new projects you create.

6. Click on the **Parameters** tab to view the other `cross_match` parameters.

For this tutorial, we will accept the default values.

7. Click **OK** to dismiss the dialog and run the algorithm.

The algorithm should complete within a few seconds. The Project view then updates with the new data.



Name	Status	Length	#	ClipL	ClipR	Start	Stop
G05b	PX	705		1	705		
F05b	PX	698		36	698		
E04b	PX	754		57	754		
F09b	PX	697		36	697		
D01b	PX	699		36	699		
H03b	PX	708		1	708		
G09b	PX	699		36	699		
B10b	PX	696		36	696		
H10b	PX	709		36	709		
B02a	PX	751		56	751		
B02b	PX	708		1	708		
F08b	PX	709		1	709		
D12b	PX	705		1	705		
A04a	PX	728		46	186		
B11a	PX	752		57	752		
H10a	PX	734		42	734		
D12a	PX	753		58	753		
H05b	PX	696		33	696		
E04a	PX	704		133	704		
B11b	PX	704		1	704		
B10a	PX	738		42	738		
H05a	PX	730		42	617		
G05a	PX	748		58	748		
ForwardPrimer	PX	782		1	782		

In addition to the status of each sequence changing to “PX” to indicate that they have been trimmed with `cross_match`, many of the **ClipL** entries now show values other than “1” indicating that vector sequences were masked at the beginning. The sequence A04a.scf has a particularly short insert and you can see that its **ClipR** value is now only 186.

8. Double-click on the sequence A04a.scf to open up a Trace editor window.

The masked residues are shown in gray italics. If you scroll to the right, you will find additional masked residues from 186 onwards.

Assembling sequences using `phrap`

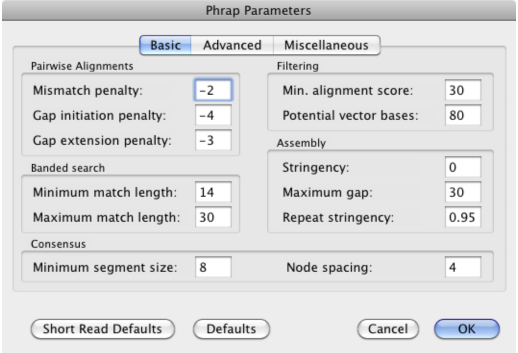
`phrap` is the assembly algorithm from the University of Washington that has been incorporated into Assembler. It is designed to work in concert with `phred` and `cross_match` – in particular it understands quality values and will use them to make better assemblies, particularly in areas with repetitive sequences. `phrap` also calculates quality values for each residue in the consensus sequence using the same scale as `phred`. However, for assemblies, a value of 40 (1 error in 10,000) is considered an acceptable value.

`phrap` is described in more detail in the phrap.pdf document that can be found in the `MacVector/Documentation` folder. This is the original documentation from the University of Washington. It is somewhat technical in places, but it describes the assembly algorithmic strategy and the effects of changing various parameters in great detail.

Make sure you have no selections in the Project view. To toggle a selection off, click on the selection while holding down the command (⌘) key.

1. Click the **Phrap** icon on the toolbar or choose **Analyze | Assemble (phrap)**.

The **Phrap Parameters** dialog box will appear. Not all of the parameters described in phrap.pdf are available in the dialog. However, it is unlikely that you will ever need to adjust any parameters other than those displayed in the **Basic** tab.



Pairwise Alignments

Mismatch penalty: -2
Gap initiation penalty: -4
Gap extension penalty: -3

Banded search

Minimum match length: 14
Maximum match length: 30

Filtering

Min. alignment score: 30
Potential vector bases: 80

Assembly

Stringency: 0
Maximum gap: 30
Repeat stringency: 0.95

Consensus

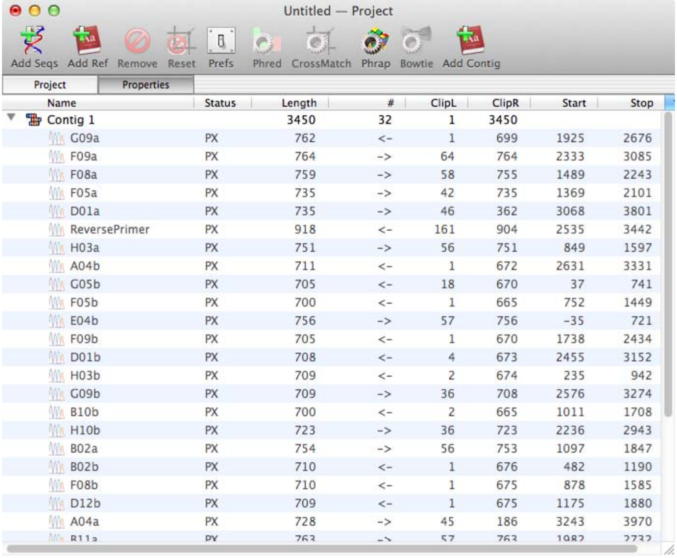
Minimum segment size: 8
Node spacing: 4

2. Click **OK** to dismiss the dialog and run the algorithm using the default values.

`phrap` is a remarkably fast algorithm and the assembly should be complete within a few seconds. Even with large (>1,000 reads) projects, assembly rarely takes more than a few minutes. As with `phred` and `cross_match`, `phrap` has been compiled for MacVector as a Universal Binary, so the algorithm will run natively on an Intel-based Macintosh.

Once assembly is complete, the project window is updated to reflect the data change. In this case, all of the reads should be assembled into a single contig.

3. Click on the disclosure triangle next to the contig to reveal the contents of the contig.



Name	Status	Length	#	ClipL	ClipR	Start	Stop
Contig 1		3450	32	1	3450		
G09a	PX	762	<-	1	699	1925	2676
F09a	PX	764	->	64	764	2333	3085
F08a	PX	759	->	58	755	1489	2243
F05a	PX	735	->	42	735	1369	2101
D01a	PX	735	->	46	362	3068	3801
ReversePrimer	PX	918	<-	161	904	2535	3442
H03a	PX	751	->	56	751	849	1597
A04b	PX	711	<-	1	672	2631	3331
G05b	PX	705	<-	18	670	37	741
F05b	PX	700	<-	1	665	752	1449
E04b	PX	756	->	57	756	-35	721
F09b	PX	705	<-	1	670	1738	2434
D01b	PX	708	<-	4	673	2455	3152
H03b	PX	709	<-	2	674	235	942
G09b	PX	709	->	36	708	2576	3274
B10b	PX	700	<-	2	665	1011	1708
H10b	PX	723	->	36	723	2236	2943
B02a	PX	754	->	56	753	1097	1847
B02b	PX	710	<-	1	676	482	1190
F08b	PX	710	<-	1	675	878	1585
D12b	PX	709	<-	1	675	1175	1880
A04a	PX	728	->	45	186	3243	3970
B11a	PX	763	->	57	763	1982	2732

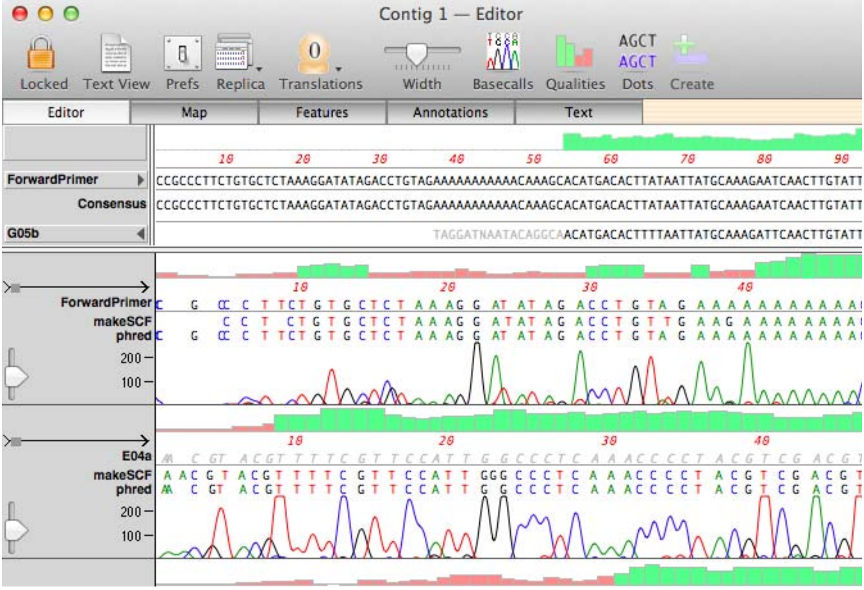
The items within the contig are grayed out to indicate that you cannot open them individually. This is to prevent you from inadvertently changing the sequence of a trace that has been carefully aligned in a contig. However, you do have full editing control from within the Contig editor (see *"Editing a contig"* on page 404).

The #, **Start** and **Stop** columns have been updated to display additional information. The number of reads assembled in the contig is indicated on the top line, while the orientation of each read in the contig is indicated on the other lines. The start and stop locations of each read within the contig are also indicated in the appropriate columns.

Editing a contig

Although `phrap` does an excellent job of assembling reads and generating an accurate consensus sequence, there are likely to be times where you need to edit the assembly, particularly if you have poor quality chromatograms, or areas of the contig that have low coverage.

1. Double-click on Contig_1 to open up the contig in the Contig editor.


Contig 1 — Editor
Locked
Text View
Prefs
Replica
Translations
Width
Basecalls
Qualities
AGCT
AGCT
Dots
Create
Editor
Map
Features
Annotations
Text
10
20
30
40
50
60
70
80
90
ForwardPrimer
CCGCCCTTCTGTGCTCTAAAGGATATAGACCTGTAGAAAAAAAAAAAACAAAGCACATGACACTTATAATTATGCAAAGAATCAACTTGTATT
Consensus
CCGCCCTTCTGTGCTCTAAAGGATATAGACCTGTAGAAAAAAAAAAAACAAAGCACATGACACTTATAATTATGCAAAGAATCAACTTGTATT
G05b
TAGGATNAATACAGGCAACATGACACTTTTAATTATGCAAAGATTCAACTTGTATT
ForwardPrimer
makeSCF
phred
200
100
E04a
makeSCF
phred
200
100

The Contig editor is based on the Assembly editor used for **Align to Reference**, with a number of important differences:

- Base calls and quality values can now be displayed, controlled by the same toolbar buttons used in the trace editor.
- There is no “reference” sequence.
- The overlapping sequences can now be displayed in “tiled” or “untiled” mode.

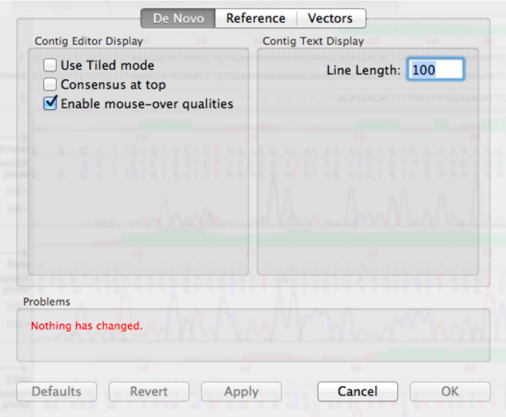
The tiled/untiled mode requires additional explanation.

In “tiled” mode, each component sequence is given its own dedicated line in the upper panel. This is fine for relatively small assemblies (< 50 reads) but for large assemblies, the layout is impractical as there is too much white space and the user spends too much time scrolling to find the right sequences to edit.

In “untiled” mode, only those sequences that actually overlap the currently visible consensus sequence are shown on the screen. This minimizes the amount of white space and reduces the need for vertical scrolling. The downside to this approach is that the reads may “move about” as you horizontally scroll through a contig.

In addition to the tiled/untiled mode, you can also choose to display the consensus sequence at the top of the panel, or in the center of the panel. This is controlled by the project preferences.

2. Click the **Prefs** icon on the toolbar to open the **Project Preferences** dialog box.



□ Use Tiled mode
□ Consensus at top
☑ Enable mouse-over qualities

Line Length: 100

Nothing has changed.

The **Project Preferences** dialog contains display preferences for **De Novo** and **Reference** assemblies and allows you to add **Vectors** to the project.

3. Select the **Use Tiled Mode** checkbox on the **De Novo** tab and click **OK**.

Note. In tiled mode, the consensus is always placed at the top of the panel.

The Contig editor display will change to use tiled mode. For the remainder of the tutorial, you should set the display to your preferred configuration. The tutorial will use the default settings of non-tiled mode with the consensus in the center of the panel. You may also want to increase the size of the window so that you can see more data at one time. Similarly, you may also want to adjust the size of the upper panel by clicking and dragging on the vertical resize control in the right hand margin.

4. Click on any residue on the consensus line.

The residue highlights, but the display also resets so that all of the traces overlapping that residue become centered in the lower multiple trace panel.

You can use this feature to click on any dubious consensus residue and immediately see the overlapping traces aligned at that position. The consensus sequence is highlighted using your primary highlight color while the reads and traces are shown in a secondary highlight color.

5. Press the right arrow button on the keyboard. Continue to hold it down.

This will scroll the contig to the right. You could potentially slowly scroll through the entire contig this way. Whenever the consensus is highlighted, the traces are always aligned and centered at that position.

6. Click on one of the read sequences, either in the upper pane or in the lower multi-trace pane.

In this case the display does not reset to align the traces at the selected position. However, the primary highlight shifts to the selected residue (in both the upper and lower panes) to indicate which character will get changed if you press a valid key.

Select any residue in one of the reads and press a different DNA character key.

Note. You cannot directly edit the consensus sequence. It is always calculated indirectly from the overlapping reads. However, because of the way quality values are handled, this is not a significant limitation.

The residue changes to the chosen character and the quality value changes to 99, represented by the blue histogram. The value 99 is very important for consensus recalculation as it always overrides all other quality values. This has two important implications;

If you edit a residue to be a valid DNA character, the consensus will always change to match that character as it overrides all other considerations.

If you edit two residues at the same position but in different reads, and they do not agree, the consensus will be given an ambiguity character.

7. Choose **Edit | Undo Typing** from the menu.

As with most MacVector functions, there is just a single level of undo.

8. Close the Contig editor.

Changes to contigs in the Contig editor are considered to be changes to the project, so you do not get prompted to save those changes until you try to close the Project window, rather than the Contig editor.

9. Choose **File | Save** from the menu.

You will be prompted for a suitable filename.

Saving the consensus sequence

You can save the consensus sequence of a contig in MacVector format at any time.

1. Double-click on a contig in the Project view to open up the Contig editor for that contig.
2. Choose **File | Save**.

You will be prompted for a suitable filename.

The file will be saved in MacVector single sequence format. Any gaps in the consensus sequence will not be present in the saved file. The locations of the reads ARE written to the file as features, but this behavior may change in future.

Analyzing contig sequences

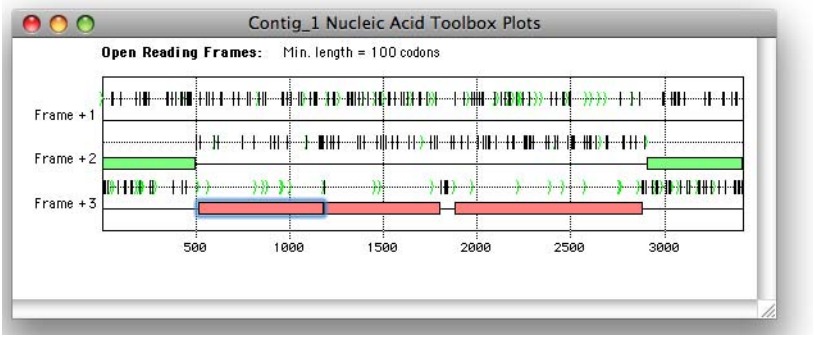
A Contig editor window is functionally equivalent to any normal MacVector single sequence window. You can run any nucleic acid sequence algorithm on the contig – in every case, it is the ungapped consensus sequence that is analyzed. Any gaps that appear in the contig editor consensus sequence are there only to maintain alignment with the overlapping reads. All analysis functions (along with save and copy functionality) strip out any gaps before analysis.

1. Double-click on the contig in the Project view to open the Contig editor, if it is not already open.
2. Choose **Analyze | Nucleic Acid Analysis Toolbox**.

The standard **Nucleic Acid Analysis** dialog box is displayed.

3. Select the **Open Reading Frames** checkbox, then click **OK** to initiate the analysis.

In the graphic window that opens, select the first open reading frame in the “red” frame as shown below (outlined in a pale blue highlight).



With the selection still in place, click on the title bar of the `Contig_1` window (or anywhere in that window). The consensus sequence is selected, along with the reads that overlap that position.

4. Choose **Analyze | Translation**.

The **Translation Analysis** dialog box is displayed.

5. Make sure the **Create new protein** checkbox is selected, then click **OK**.

A new protein Sequence window will be displayed. The protein sequence is a translation of the consensus sequence with all of the gaps removed. The same principle applies to all MacVector analysis functions – you can run any nucleic acid analysis (e.g. restriction enzyme analysis, or an online BLAST search) directly from the contig editor window and it is the ungapped consensus that gets analyzed. This allows you to get instant feedback on edits affecting the consensus sequence without requiring clumsy export to a different analysis module.

Dissolving contigs

Make sure you have saved the assembly project you are working on. Most of the functions that dissolve or significantly modify contigs cannot be undone.

1. In the assembly Project window, select a contig and then click the **Dissolve** icon on the toolbar or press the **<delete>** key.

The contig will be dissolved into its constituent reads. The project window updates to indicate the fact that all of the reads that were in the contig have now been returned to the root of the project.

Any edits you made to the individual reads will be maintained, although all gaps will be removed. It is essential that the edits be retained as this allows you to edit sequences in a “bad” contig and then reassemble them taking your edits into account. This is particularly important when assembling sequences containing closely related repeats. `phrap` will not assemble overlapping sequences that have mismatched edited bases (i.e. that have quality values of 99) – you can use this to force misassembled repeats to split by editing the mismatched bases and re-assembling.

Reassembling contigs

1. In the assembly Project window, select a contig and then click the **Phrap** icon or choose **Analyze | Assemble (phrap)** from the menu. Accept the default parameters and click on the **OK** button.

You should see that the contig is dissolved and the assembled sequences appear in the project window. After assembly is complete, a new contig will appear in the project window.

The contig is dissolved prior to reassembly because there is no guarantee that the contig will reassemble with the same sequences as before. If you have edited sequences in the contig, or chosen different phrap parameters, one or more sequences may no longer assemble.

You can select any combination of contigs and sequences in the project window to (re)assemble just those items.

Part three

Understanding MacVector

The following chapters explain some of the theories and methods used by MacVector.

Chapter 20, *“Understanding Protein and DNA Analysis”*, explains some of the theories and methods used by MacVector for protein and DNA analysis.

Chapter 21, *“Understanding Sequence Comparisons”*, explains some of the theories and methods used by MacVector for sequence comparison, alignment, and phylogenetic analysis.

20 Understanding Protein and DNA Analysis

Overview

This chapter explains some of the theories and methods used by MacVector for protein and DNA analysis. Users are referred to the academic papers from which the methods are derived. This chapter provides greater detail only where significant changes have been made to published methods.

Refer to Appendix F, *“References”*.

Contents

The protein analysis toolbox

One of the goals of molecular biology is to be able to determine the three-dimensional structure of a protein directly from its amino acid sequence—the so-called protein folding problem. As a first step toward solving this problem, researchers have developed algorithms for predicting the secondary structures of amino acid sequences and methods for predicting which regions of a protein might lie on the surface or might be buried in the protein's interior. MacVector groups these algorithms together in the protein analysis toolbox.

When applying these methods to a protein of unknown structure, the researcher must always keep in mind certain limitations. Most standard methods are based on empirical data derived from proteins whose three-dimensional structure has been determined by X-ray crystallography.

Therefore, the accuracy of the prediction will depend on how closely the protein of unknown structure resembles the set of known proteins used to derive the method. A method that performs well at predicting the structure of a cytochrome may fail when applied to an integral membrane protein; a method derived from a database of large globular proteins is probably inappropriate for deducing the structure of small peptides; and a method that is fairly accurate at predicting the structures of proteins in which alpha helix predominates may not work for proteins composed primarily of beta sheet.

Ideally, these predictive methods should be used in conjunction with biochemical and biophysical information about the protein of interest, *e.g.*, sequence similarity with proteins of known structure, local similarities (such as patterns associated with known binding sites) with proteins of similar function, circular dichroism or Raman spectroscopy data, *etc*.

Secondary structure predictions

The Chou-Fasman method

This is the most popular and widely known method of protein secondary structure prediction. From proteins with known X-ray structures, Chou and Fasman compiled statistics on the tendency of an amino acid to appear in a given secondary structure, and used these statistics to assign the 20 amino acids into four classes: helix formers, helix breakers, sheet formers, and sheet breakers. They then devised a method to predict the secondary structure of a protein, by locating clusters of helix- or sheet-forming residues in the sequence, and applying a set of heuristic rules to determine if these clusters are significant enough to nucleate a helix or beta sheet structure.

The method was originally intended to be performed by hand, which leads to its major drawback when it is adapted for computers. The same region of the protein often appears to be equally likely to nucleate a helix or sheet conformation, and the rules do not resolve this conflict by indicating the relative weight to place on each possible conformation at a given region.

By performing the method manually, the researcher resolves these conflicts by making subjective judgments (possibly biased by foreknowledge). But when the method is computerized, the programmer must explicitly assign the weights to be used to resolve such conflicts. Because each programmer may resolve them differently, one implementation of the Chou-Fasman method may yield different predictions from another implementation.

MacVector makes no attempt to resolve conflicting predictions at the same region. Instead, each prediction or conformation is treated independently of the others and each structure prediction is graphed separately. Thus, you may see a single region predicted to be in more than one conformational state. This form of presentation underscores the uncertainties of the method (or any current method for secondary structure prediction) and is a reminder that the predictions should not be over-interpreted.

The Robson-Garnier method

This is the other most popular method of protein secondary structure prediction. The Robson-Garnier method is based on information theory. Empirical studies show that an amino acid exerts a significant effect on the conformational state of residues up to eight residues distant; therefore, the information for the conformation of residue N can be based on the information contributions of the 16 nearest neighbors of N. Using a set of 25 proteins of known structure as a database, the directional information contributions of each of the 20 amino acids to a given conformational state for each of these 17 positions (N - 8 through N + 8) were derived. Using these information parameters, the likelihood of a given residue assuming each of the four possible conformations (alpha, beta, reverse turn, or coil) is calculated, and the conformation with the largest likelihood is assigned to the residue.

Unlike the Chou-Fasman method, the algorithm is clearly and unambiguously defined so that all computer implementations should yield the same result. The only choice open to the programmer involves the use of decision constants that can be assigned according to information known about the molecule's secondary structure from other sources (such as circular dichroism). MacVector follows the most commonly used procedure and automatically calculates the decision constants based on estimates of alpha-helix and beta-sheet content from a first pass of the program itself.

Combining the two methods

By graphing the secondary structure predictions only where both the Chou-Fasman and Robson-Garnier methods agree, the researcher hopes to locate those regions that are most likely to be in the predicted conformation. However, we advise you that, at best, each method has about a 60 per cent probability of being correct and that the consensus of two possibly wrong predictions does not give a correct prediction.

Protein profiles

Hydrophilicity

This profile graphs the local hydrophilicity of a protein along its amino acid sequence. Each of the 20 amino acids is assigned a hydropathy value based on some experimental or empirical measure. A window of size N is run along the length of the protein; for each window, the hydropathy values of the N amino acids are summed and divided by N to obtain the average hydrophilicity per residue for the window. The value is then plotted on the graph at the center of the window. Values above the axis denote hydrophilic regions which may be exposed on the outside of the molecule; values below the axis indicate hydrophobic regions which tend to be buried inside the molecule or inside other hydrophobic environments such as membranes.

Various amino acid hydropathy scales have been developed for hydrophilicity profiles. Three commonly used scales are Kyte and Doolittle (1982), Hopp and Woods (1981), and Engelman, Steitz and Goldman (1986) (GES). The Kyte-Doolittle and GES scales were originally used for hydrophobicity profiles. We have reversed the signs of the values so that hydrophilicity is plotted instead.

The Kyte-Doolittle scale is the most commonly used hydropathy scale. Its values are assigned using a combination of the water-vapor transfer free energies for amino acid side chains and the preference of amino acid side chains for interior or exterior environments, with small adjustments made to the final values (based on the experience of the authors).

The Hopp-Woods scale was designed to predict the locations of antigenic determinants in a protein, assuming that the antigenic determinants would be exposed on the surface of the protein and thus would be located in hydrophilic regions. Its values are derived from the transfer free energies for amino acid side chains between ethanol and water.

The GES scale was developed in order to identify possible transmembrane helices in a protein. The scale values are the sums of hydrophobic and hydrophilic components for each amino acid. Hydrophobic components are derived from the free energy of water-oil transfer for the side chains; hydrophilic components take into consideration the free energy for inserting charged groups into a bilayer and the free-energy contributions of hydrogen-bonding with water and with backbone carbonyl groups (if the residues can form such bonds when participating in a helical structure).

For most globular proteins, window sizes between seven and eleven are appropriate. (Hopp and Woods recommend a window size of six when using their scale.) In general, smaller window sizes yield noisier profiles, while larger window sizes tend to miss small hydrophilic or hydrophobic regions. Window sizes of five to seven can be useful in locating small hydrophilic regions that may protrude from the protein surface and thus be antigenic sites. Window sizes of 19 to 21 are often useful in locating possible membrane-spanning regions in integral membrane proteins.

Surface probability

This profile was designed to predict which regions of a protein are most likely to lie on the protein’s surface, based on knowledge of which amino acids are more likely to be found on the surface of proteins of known structure.

Janin *et al*. (1978) examined 28 proteins whose atomic coordinates were known and determined the solvent-accessible surface area of each residue. Residues were classified as “buried” if their accessible surface area was smaller than 20 Å and as “exposed” if the accessible surface area was larger than 60 Å. For each of the 20 amino acids, information was compiled on what percentage of the time the amino acid was found in an exposed position and what percentage of the time it was found in a buried position in these 28 proteins (Table 1 of Janin *et al.* (1978)).

Emini *et al.* (1965) used these percentages to calculate a fractional surface probability for each amino acid (%exposed / [%exposed + %buried]). The fractional surface probabilities were used to compute a surface “probability” profile for hexapeptide windows along a protein sequence by multiplying together the six fractional probabilities in the window, then multiplying this product by $(.037)^{-6}$.

Because this formula permits the “probability” to exceed 1.0, MacVector uses a different formula. MacVector sums the six fractional probabilities of the amino acids in the window and divides by six to yield a running average of the fractional surface probability along the length of the protein. Thus, a value of 1.0 at any point (which will never occur) would mean that the hexapeptide centered about that point is definitely exposed at the surface of the protein and a value of 0.0 (which also will never occur) means that the hexapeptide is definitely buried in the interior of the protein.

Flexibility

Because segmental flexibility appears to correlate with known antigenic determinants, Karplus and Schulz (1985) devised a method to predict potential antigenic sites by locating regions of the protein chain that might be relatively flexible.

Using a set of 31 proteins of known three-dimensional structure as a database, the 20 amino acids were assigned to one of two groups - rigid or flexible - depending on the normalized crystallographic temperature factors (B-values) of their alpha-carbon atoms. Three sets of B-norm values were then calculated for each amino acid, one set for each of the possible combinations of nearest neighbors: no rigid neighbors, one rigid neighbor, and two rigid neighbors. The predicted relative flexibility at a given residue is the weighted sum of the neighbor-correlated B-norm values (B) for the given residue and the six residues flanking it:

$$\begin{aligned}\text{flex[i]} = {} & 0.25(B[i-3]) + 0.5(B[i-2]) + 0.75(B[i-1]) + \\ & 1.0(B[i]) + 0.75(B[i+1]) + 0.5(B[i+2]) + 0.25(B[i+3])\end{aligned}$$

Using this method, the average flexibility of a protein is 1.0. Regions with values greater than 1.0 are predicted to be more flexible than average, while values below 1.0 indicate regions predicted to be less flexible than average.

Antigenic index

This profile is designed to locate possible exposed surface peaks of a protein, i.e., peaks that might be antigenic sites. Rather than use a single type of analysis, this method combines information from hydrophilicity, surface probability, and backbone flexibility predictions along with the secondary structure predictions of Chou-Fasman and Robson-Garnier in order to produce a composite prediction of the surface contour of a protein.

The method used in MacVector differs in one respect from that of B.A. Jameson and H.Wolf (1988): scores for each of the analyses are normalized to a value between -1.0 and +1.0 instead of using the values in Table I of the reference. The score for each analysis at each residue is multiplied by an empirically-determined weighting factor for that analysis and the weighted scores for each of the analyses are summed to yield the antigenic index, as in the reference.

$$\text{Antigenic index} = \Sigma\{0.3\text{hydro}[i] + 0.15\text{surf\_prob}[i] + 0.15\text{flex}[i] + 0.2\text{chou\_fas}[i] + 0.2\text{rob\_garn}[i]\}$$

Regions that plot above the graph axis are predicted to be exposed at the protein’s surface.

Amphiphilicity

Amphiphilicity profiles are used to detect regions of a protein that may form amphiphilic (or amphipathic) structures—structures that tend to be polar on one side and apolar on the other. Such regions are often found at protein-solvent or membrane-protein interfaces.

One way of looking for such structures is to examine the periodicity of the protein’s hydrophobicity (its hydrophobic moment). An amphiphilic alpha helix will exhibit a periodicity of about 100 degrees; for a strand of beta sheet, the periodicity theoretically would be 180 degrees, but because beta strands tend to twist, the angle varies in practice from about 160 to 180 degrees. As a compromise, MacVector uses 170 degrees when calculating beta strand amphiphilicity.

To compute the amphiphilicity profile, a window of size N is moved along the entire protein sequence. For each of these windows, the hydrophobic moment is calculated according to the formula:

$$\text{moment} = \{[\Sigma H(i)\sin(Ai)]^2 + [\Sigma H(i)\cos(Ai)]^2\}^{1/2}$$

where:

H(i) is the hydrophobicity (according to the Eisenberg *et al.* (1984) normalized consensus scale) of the ith residue of the window

A is the size of the angle (in radians) at which successive side chains emerge from the protein backbone

the index *i* takes on values from 1 to N.

The hydrophobic moment for the window is divided by N to yield an average moment per residue, and this value is plotted at the center of the window. Values greater than about 0.4 indicate amphiphilic regions.

The window size should at least be the size of one “unit” of the structure (three residues for alpha helix, two residues for beta strands) and ordinarily is chosen to equal the typical length for that structure type. Eisenberg *et al.* (1984a) suggest a window size of 11; however, von Heijne (1986) finds that the optimal window size for surface-seeking peptides ranges from 17 to 26, and for mitochondrial targeting sequences from 12 to 26.

Estimating the pI

The pI of a protein is the pH where the protein has a net charge of zero. It is calculated using the Henderson-Hasselbach equation to figure the percentage disassociation of each amino acid in the protein at various pHs. This calculation can only be an estimate, however, because it assumes that the percentage dissociation of an amino acid in a protein is the same as that of a free amino acid. This assumption is known to be false: there are effects on the ionization constants from neighboring amino acids, and these effects may extend over a considerable range of the protein.

The estimate for the pI in MacVector is based on the pKa values for the amino acid side chains and the amino and carboxy ends of the protein published in *“Biochemistry, a Problems Approach”*, by Wood, Wilson, Benbow, and Hood. The values are:

Asp	Glu	Tyr	Lys	Arg	His	NH3	COOH
3.86	4.25	10.10	9.80	12.48	6.00	8.00	3.00

These values were tested by comparing the experimentally determined values of the pIs of several proteins with estimated pIs of related proteins from the NBRF PIR database. Because the published experimental values of the pI do not state which species or exact sequence was used, the representatives chosen from the database may not correspond exactly to the proteins used in the experimental determinations. This may result in some of the discrepancy seen below between the experimental and estimated pIs:

Protein	**NBRF name**	**experimental pI**	**estimate**
Egg Albumin	OACH	4.6	5.02
Serum Albumin	ABBOS	4.9	5.83
Hemoglobin	HBBOB	6.8	7.38
Myoglobin	MYBO	7.0	7.54
Cytochrome C	CCHU	10.7	9.59
Lysozyme	LZHU	11.0	10.20

Primers and probes

Primer design

MacVector incorporates primer design using Primer3.

Note. The following is extracted from the Primer3 readme file, which is supplied in full in the MacVector application folder.

Primer3 picks primers for PCR reactions, considering as criteria:

- oligonucleotide melting temperature, size, GC content and primer-dimer possibilities
- PCR product size
- positional constraints within the source (template) sequence
- possibilities for ectopic priming (amplifying the wrong sequence)
- many other constraints

All of these criteria are user-specifiable as constraints, and some are specifiable as terms in an objective function that characterizes an optimal primer pair.

This section provides a cross-reference between the Primer3 constraints and the equivalent MacVector parameters. It is organized by MacVector user interface component.

Primer Sequences panel

- PRIMER_LEFT_INPUT (nucleotide sequence, default empty)

The sequence of a left primer to check and around which to design right primers and optional internal oligos. Must be a substring of SEQUENCE.

- PRIMER_RIGHT_INPUT (nucleotide sequence, default empty)

The sequence of a right primer to check and around which to design left primers and optional internal oligos. Must be a substring of the reverse strand of SEQUENCE.

Main function panel

- PRIMER_PRODUCT_SIZE_RANGE (size range list, default 100-300)

The associated values specify the lengths of the product that the user wants the primers to create, and is a space separated list of elements of the form <x>-<y> where an <x>-<y> pair is a legal range of lengths for the product. For example, if one wants PCR products to be between 100 to 150 bases (inclusive) then one would set this parameter to 100-150. If one desires PCR products in either the range from 100 to 150 bases or in the range from 200 to 250 bases then one would set this parameter to 100-150 200-250.

Primer3 favors ranges to the left side of the parameter string. Primer3 will return legal primers pairs in the first range regardless the value of the objective function for these pairs. Only if there are an insufficient number of primers in the first range will Primer3 return primers in a subsequent range. For those with primarily a computational background, the PCR product size is size (in base pairs) of the DNA fragment that would be produced by the PCR reaction on the given sequence template. This would, of course, include the primers themselves.

Amplify Feature/Region - Region to Scan - Flanking Regions panel

- PRIMER_TASK (string, default pick_pcr_primers)

Tell primer3 what task to perform. Legal values are pick_pcr_primers, pick_pcr_primers_and_hyb_probe, pick_left_only, pick_right_only, pick_hyb_probe_only. The tasks should be self explanatory, except that we note that pick_pcr_primers_and_hyb_probe is equivalent to the setting PRIMER_PICK_INTERNAL_OLIGO to a non-zero value and setting PRIMER_TASK to pick_pcr_primers.

Advanced options - Primer Binding

Primer vs Primer

- PRIMER_SELF_ANY (decimal,9999.99, default 8.00)

The maximum allowable local alignment score when testing a single primer for (local) self-complementarity and the maximum allowable local alignment score when testing for complementarity between left and right primers. Local self-complementarity is taken to predict the tendency of primers to anneal to each other without necessarily causing self-priming in the PCR. The scoring system gives 1.00 for complementary bases, -0.25 for a match of any base (or N) with an N, -1.00 for a mismatch, and -2.00 for a gap. Only single-base-pair gaps are allowed. For example, the alignment

```
5' ATCGNA 3'
   || | |
3' TA-CGT 5'
```

is allowed (and yields a score of 1.75), but the alignment

```
5' ATCCGNA 3'
   ||  | |
3' TA--CGT 5'
```

is not considered. Scores are non-negative, and a score of 0.00 indicates that there is no reasonable local alignment between two oligos.

3’ end vs 3’ end

- PRIMER_SELF_END (decimal 9999.99, default 3.00)

The maximum allowable 3'-anchored global alignment score when testing a single primer for self-complementarity, and the maximum allowable 3'-anchored global alignment score when testing for complementarity between left and right primers. The 3'-anchored global alignment score is taken to predict the likelihood of PCR-priming primer-dimers, for example

```
5' ATGCCCTAGCTTCCGGATG 3'
             ||| |||||
          3' AAGTCCTACATTTAGCCTAGT 5'
```

or

```
5` AGGCTATGGGCCTCGCGA 3'
         ||||||
      3' AGCGCTCCGGGTATCGGA 5'
```

The scoring system is as for the Maximum Complementarity argument. In the examples above the scores are 7.00 and 6.00 respectively. Scores are non-negative, and a score of 0.00 indicates that there is no reasonable 3'-anchored global alignment between two oligos. In order to estimate 3'-anchored global alignments for candidate primers and primer pairs, Primer assumes that the sequence from which to choose primers is presented 5'->3'. It is nonsensical to provide a larger value for this parameter than for the Maximum (local) Complementarity parameter because the score of a local alignment will always be at least as great as the score of a global alignment.

Primer vs Product

- PRIMER_MAX_TEMPLATE_MISPRIMING (decimal,9999.99, default -1.00)

The maximum allowed similarity to ectopic sites in the template. A negative value means do not check. The scoring system is the same as used for PRIMER_MAX_MISPRIMING, except that an ambiguity code in the template is never treated as a consensus (see PRIMER_LIB_AMBIGUITY_CODES_CONSENSUS).

Allowed Ns in primers

- PRIMER_NUM_NS_ACCEPTED (int, default 0)

Maximum number of unknown bases (N) allowable in any primer.

Allow ambiguous residues

- PRIMER_LIBERAL_BASE (boolean, default 0)

This parameter provides a quick-and-dirty way to get primer3 to accept IUB / IUPAC codes for ambiguous bases (i.e. by changing all unrecognized bases to N). If you wish to include an ambiguous base in an oligo, you must set PRIMER_NUM_NS_ACCEPTED to a non-0 value. Perhaps '-' and '* ' should be squeezed out rather than changed to 'N', but currently they simply get converted to N's. The authors invite user comments.

Unused Primer3 constraints

- PRIMER_PAIR_MAX_TEMPLATE_MISPRIMING (decimal,9999.99, default -1.00)

The maximum allowed summed similarity of both primers to ectopic sites in the template. A negative value means do not check. The scoring system is the same as used for PRIMER_PAIR_MAX_MISPRIMING, except that an ambiguity code in the template is never treated as a consensus (see PRIMER_LIB_AMBIGUITY_CODES_CONSENSUS). Primer3 does not check the similarity of hybridization oligos (internal oligos) to locations outside of the amplicon.

Advanced options - Characteristics

Length

- PRIMER_OPT_SIZE (int, default 20)

Optimum length (in bases) of a primer oligo. Primer3 will attempt to pick primers close to this length.

- PRIMER_MIN_SIZE (int, default 18)

Minimum acceptable length of a primer. Must be greater than 0 and less than or equal to PRIMER_MAX_SIZE.

- PRIMER_MAX_SIZE (int, default 27)

Maximum acceptable length (in bases) of a primer. Currently this parameter cannot be larger than 35. This limit is governed by maximum oligo size for which primer3's melting-temperature is valid.

Percent G+C

- PRIMER_MIN_GC (float, default 20.0%)

Minimum allowable percentage of Gs and Cs in any primer.

- PRIMER_OPT_GC_PERCENT (float, default 50.0%)

Optimum GC percent. This parameter influences primer selection only if PRIMER_WT_GC_PERCENT_GT or PRIMER_WT_GC_PERCENT_LT are non-0.

- PRIMER_MAX_GC (float, default 80.0%)

Maximum allowable percentage of Gs and Cs in any primer generated by Primer.

Tm (C)

- PRIMER_OPT_TM (float, default 60.0C)

Optimum melting temperature(Celsius) for a primer oligo. Primer3 will try to pick primers with melting temperatures are close to this temperature. The oligo melting temperature formula used can be specified by user. Please see PRIMER_TM_SANTALUCIA for more information.

- PRIMER_MIN_TM (float, default 57.0C)

Minimum acceptable melting temperature(Celsius) for a primer oligo.

- PRIMER_MAX_TM (float, default 63.0C)

Maximum acceptable melting temperature(Celsius) for a primer oligo.

GC Clamp

- PRIMER_GC_CLAMP (int, default 0)

Require the specified number of consecutive Gs and Cs at the 3' end of both the left and right primer. (This parameter has no effect on the internal oligo if one is requested.)

Maximum difference in Tm between primers

- PRIMER_MAX_DIFF_TM (float, default 100.0C)

Maximum acceptable (unsigned) difference between the melting temperatures of the left and right primers.

Maximum Poly-X

- PRIMER_MAX_POLY_X (int, default 5)

The maximum allowable length of a mononucleotide repeat, for example AAAAAA.

Unused Primer3 constraints

- PRIMER_PRODUCT_MAX_TM (float, default 1000000.0)

The maximum allowed melting temperature of the amplicon. Primer3 calculates product Tm calculated using the formula from Bolton and McCarthy, PNAS 84:1390 (1962) as presented in Sambrook, Fritsch and Maniatis, Molecular Cloning, p 11.46 (1989, CSHL Press).

Tm = 81.5 + 16.6(log10([Na+])) + .41*(%GC) - 600/length

Where [Na+] is the molar sodium concentration, (%GC) is the percent of Gs and Cs in the sequence, and length is the length of the sequence.

A similar formula is used by the prime primer selection program in GCG (http://www.gcg.com), which instead uses 675.0 / length in the last term (after F. Baldino, Jr, M.-F. Chesselet, and M.E. Lewis, Methods in Enzymology 168:766 (1989) eqn (1) on page 766 without the mismatch and formamide terms). The formulas here and in Baldino et al. assume Na+ rather than K+. According to J.G. Wetmur, Critical Reviews in BioChem. and Mol. Bio. 26:227 (1991) 50 mM K+ should be equivalent in these formulae to .2 M Na+. Primer3 uses the same salt concentration value for calculating both the primer melting temperature and the oligo melting temperature. If you are planning to use the PCR product for hybridization later this behavior will not give you the Tm under hybridization conditions.

- PRIMER_PRODUCT_MIN_TM (float, default -1000000.0)

The minimum allowed melting temperature of the amplicon. Please see the documentation on the maximum melting temperature of the product for details.

Advanced options - Reaction Conditions

Total initial primer concentration (nM)

- PRIMER_DNA_CONC (float, default 50.0 nM)

The nanomolar concentration of annealing oligos in the PCR. Primer3 uses this argument to calculate oligo melting temperatures. The default (50nM) works well with the standard protocol used at the Whitehead/ MIT Center for Genome Research - 0.5 microliters of 20 micromolar concentration for each primer oligo in a 20 microliter reaction with 10 nanograms template, 0.025 units/microliter Taq polymerase in 0.1 mM each dNTP, 1.5mM MgCl2, 50mM KCl, 10mM Tris-HCL (pH 9.3) using 35 cycles with an annealing temperature of 56 degrees Celsius. This parameter corresponds to 'c' in equation (ii) of the paper [Rychlik W, Spencer WJ and Rhoads RE (1990) "Optimization of the annealing temperature for DNA amplification in vitro", Nucleic Acids Res 18:6409-12 http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=2243783], where a suitable value (for a lower initial concentration of template) is "empirically determined". The value of this parameter is less than the actual concentration of oligos in the reaction because it is the concentration of annealing oligos, which in turn depends on the amount of template (including PCR product) in a given cycle. This concentration increases a great deal during a PCR; fortunately PCR seems quite robust for a variety of oligo melting temperatures. See *"Advice for picking primers"* on page 435.

Monovalent cation concentration (mM)

- PRIMER_DIVALENT_CONC (float, default 0.0 mM)

The millimolar concentration of divalent salt cations (usually MgCl^(2+)) in the PCR. (New in v. 1.1.0, added by Maido Remm and Triinu Koressaar) Primer3 converts concentration of divalent cations to concentration of monovalent cations using formula suggested in the paper [Ahsen von N, Wittwer CT, Schutz E (2001) "Oligonucleotide Melting Temperatures under PCR Conditions: Nearest-Neighbor Corrections for Mg^(2+), Deoxynucleotide Triphosphate, and Dimethyl Sulfoxide Concentrations with Comparision to Alternative Empirical Formulas", Clinical Chemistry 47:1956-61 http://www.clinchem.org/

cgi/ content/full/47/11/1956]. [Monovalent cations] = [Monovalent cations] + 120*(([divalent cations] - [dNTP])^0.5) According to the formula concentration of desoxynucleotide triphosphate [dNTP] must be smaller than concentration of divalent cations. If the specified concentration of dNTPs is larger than specified concentration of divalent cations then the effect of divalent cations is not considered. The concentration of dNTPs is included to the formula because of some magnesium is bound by the dNTP. Attained concentration of monovalent cations is used to calculate oligo/primer melting temperature. Use tag PRIMER_DNTP_CONC to specify the concentration of dNTPs.

Unused Primer3 constraints

- PRIMER_DNTP_CONC (float, default 0.0 mM)

The millimolar concentration of deoxyribonucleotide triphosphate. This argument is considered only if PRIMER_DIVALENT_CONC is specified. See PRIMER_DIVALENT_CONC.

- PRIMER_SALT_CORRECTIONS (int, default 0)

Specifies the salt correction formula for the melting temperature calculation. (New in v. 1.1.0, added by Maido Remm and Triinu Koressaar) A value of 1 (*RECOMMENDED*) directs primer3 to use the salt correction formula in the paper [SantaLucia JR (1998) "A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics", Proc Natl Acad Sci 95:1460-65 http://dx.doi.org/10.1073/pnas.95.4.1460] A value of 0 directs primer3 to use the salt correction formula in the paper [Schildkraut, C, and Lifson, S (1965) "Dependence of the melting temperature of DNA on salt concentration", Biopolymers 3:195-208 (not available on-line)]. This was the formula used in previous version of primer3. A value of 2 directs primer3 to use the salt correction formula in the paper [Owczarzy R, You Y, Moreira BG, Manthey JA, Huang L, Behlke MA and Walder JA (2004) "Effects of sodium ions on DNA duplex oligomers: Improved predictions of melting temperatures", Biochemistry 43:3537-54 http://dx.doi.org/10.1021/bi034621r].

Advanced options - Hybridization Primer

Because the laboratory detection step using internal oligos is independent of the PCR amplification procedure, internal oligo tags have defaults that are independent of the parameters that govern the selection of PCR primers. For example, the melting temperature of an oligo used for hybridization might be considerably lower than that used as a PCR primer.

These tags are analogous to the global input tags (those governing primer oligos) discussed above.

The exception is PRIMER_INTERNAL_OLIGO_SELF_END which is meaningless when applied to internal oligos used for hybridization-based detection, since primer-dimer will not occur. We recommend that PRIMER_INTERNAL_OLIGO_SELF_END be set at least as high as PRIMER_INTERNAL_OLIGO_SELF_ANY.

- PRIMER_INTERNAL_OLIGO_OPT_SIZE (int, default 20)
- PRIMER_INTERNAL_OLIGO_MIN_SIZE (int, default 18)
- PRIMER_INTERNAL_OLIGO_MAX_SIZE (int, default 27)
- PRIMER_INTERNAL_OLIGO_OPT_TM (float, default 60.0 degrees C)
- PRIMER_INTERNAL_OLIGO_OPT_GC_PERCENT (float, default 50.0%)
- PRIMER_INTERNAL_OLIGO_MIN_TM (float, default 57.0 degrees C)
- PRIMER_INTERNAL_OLIGO_MAX_TM (float, default 63.0 degrees C)
- PRIMER_INTERNAL_OLIGO_MIN_GC (float, default 20.0%)
- PRIMER_INTERNAL_OLIGO_MAX_GC (float, default 80.0%)
- PRIMER_INTERNAL_OLIGO_DNA_CONC (float, default 50.0 nM)
- PRIMER_INTERNAL_OLIGO_SELF_ANY (decimal 9999.99, default 12.00)
- PRIMER_INTERNAL_OLIGO_MAX_POLY_X (int, default 5)

Unused Primer3 constraints

- PRIMER_INTERNAL_OLIGO_SALT_CONC (float, default 50.0 mM)
- PRIMER_INTERNAL_OLIGO_SELF_END (decimal 9999.99, default 12.00)
- PRIMER_INTERNAL_OLIGO_DIVALENT_CONC (float, default 0.0 mM)
- PRIMER_INTERNAL_OLIGO_DNTP_CONC (float, default 0.0 mM)

Advanced options - Misc.

- PRIMER_NUM_RETURN (int, default 5)

The maximum number of primer pairs to return. Primer pairs returned are sorted by their "quality", in other words by the value of the objective function (where a lower number indicates a better primer pair). Caution: setting this parameter to a large value will increase running time.

Constraints not used by MacVector or used only at their default values

- PRIMER_PICK_ANYWAY (boolean, default 0)

If true pick a primer pair even if PRIMER_LEFT_INPUT, PRIMER_RIGHT_INPUT, or PRIMER_INTERNAL_OLIGO_INPUT violates specific constraints.

- PRIMER_PRODUCT_OPT_TM (float, default 0.0)

The optimum melting temperature for the PCR product. 0 indicates that there is no optimum temperature.

- PRIMER_PRODUCT_OPT_SIZE (int, default 0)

The optimum size for the PCR product. 0 indicates that there is no optimum product size. This parameter influences primer pair selection only if PRIMER_PAIR_WT_PRODUCT_SIZE_GT or PRIMER_PAIR_WT_PRODUCT_SIZE_LT is non-0.

- PRIMER_EXPLAIN_FLAG (boolean, default 0)

If this flag is non-0, produce PRIMER_LEFT_EXPLAIN,

- PRIMER_WT_TM_GT (float, default 1.0)

Penalty weight for primers with Tm over PRIMER_OPT_TM.

- PRIMER_WT_TM_LT (float, default 1.0)

Penalty weight for primers with Tm under PRIMER_OPT_TM.

- PRIMER_WT_SIZE_LT (float, default 1.0)

Penalty weight for primers shorter than PRIMER_OPT_SIZE.

- PRIMER_WT_SIZE_GT (float, default 1.0)

Penalty weight for primers longer than PRIMER_OPT_SIZE.

- PRIMER_WT_GC_PERCENT_LT (float, default 1.0)

Penalty weight for primers with GC percent greater than PRIMER_OPT_GC_PERCENT.

- PRIMER_WT_GC_PERCENT_GT (float, default 1.0)

Penalty weight for primers with GC percent greater than PRIMER_OPT_GC_PERCENT.

- PRIMER_WT_COMPL_ANY (float, default 0.0)

- PRIMER_WT_COMPL_END (float, default 0.0)
- PRIMER_WT_NUM_NS (float, default 0.0)
- PRIMER_WT_REP_SIM (float, default 0.0)
- PRIMER_WT_SEQ_QUAL (float, default 0.0)
- PRIMER_WT_END_QUAL (float, default 0.0)
- PRIMER_WT_POS_PENALTY (float, default 0.0)
- PRIMER_WT_END_STABILITY (float, default 0.0)
- PRIMER_WT_TEMPLATE_MISPRIMING (float, default 0.0)
- PRIMER_PAIR_WT_PR_PENALTY (float, default 1.0)
- PRIMER_PAIR_WT_IO_PENALTY (float, default 0.0)
- PRIMER_PAIR_WT_DIFF_TM (float, default 0.0)
- PRIMER_PAIR_WT_COMPL_ANY (float, default 0.0)
- PRIMER_PAIR_WT_COMPL_END (float, default 0.0)
- PRIMER_PAIR_WT_PRODUCT_TM_LT (float, default 0.0)
- PRIMER_PAIR_WT_PRODUCT_TM_GT (float, default 0.0)
- PRIMER_PAIR_WT_PRODUCT_SIZE_GT (float, default 0.0)
- PRIMER_PAIR_WT_PRODUCT_SIZE_LT (float, default 0.0)
- PRIMER_PAIR_WT_REP_SIM (float, default 0.0)
- PRIMER_PAIR_WT_TEMPLATE_MISPRIMING (float, default 0.0)
- PRIMER_IO_WT_TM_GT (float, default 1.0)
- PRIMER_IO_WT_TM_LT (float, default 1.0)
- PRIMER_IO_WT_GC_PERCENT_GT (float, default 1.0)
- PRIMER_IO_WT_GC_PERCENT_LT (float, default 1.0)
- PRIMER_IO_WT_SIZE_LT (float, default 1.0)
- PRIMER_IO_WT_SIZE_GT (float, default 1.0)
- PRIMER_IO_WT_COMPL_ANY (float, default 0.0)
- PRIMER_IO_WT_COMPL_END (float, default 0.0)
- PRIMER_IO_WT_NUM_NS (float, default 0.0)
- PRIMER_IO_WT_REP_SIM (float, default 0.0)
- PRIMER_IO_WT_SEQ_QUAL (float, default 0.0)
- PRIMER_IO_WT_END_QUAL (float, default 0.0)

Primer3 output file format

In debug mode, MacVector can produce a raw Primer3 output file This section explains the syntax of this file.

For each boulderio record passed into Primer3 via stdin, exactly one boulderio record comes out of Primer3 on stdout. These output records contain everything that the input record contains, plus a subset of the following tag/value pairs. Unless noted by (*), each tag appears for each primer pair returned.

The first version is PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO,PAIR}_<tag_name>.

Tags of additional primers chosen are of the form PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO,PAIR}_<j>_<tag_name>, where <j> is an integer from 1 to n, where n is at most the value of PRIMER_NUM_RETURN.

In the descriptions below, 'i,n' represents a start/length pair, 's' represents a string, ‘x’ represents an arbitrary integer, and ‘f’ represents a float.

PRIMER_ERROR=s (*)

s describes user-correctible errors detected in the input (separated by semicolons). This tag is absent if there are no errors.

PRIMER_LEFT=i,n

The selected left primer (the primer to the left in the input sequence). i is the 0-based index of the start base of the primer, and n is t its length.

PRIMER_RIGHT=i,n

The selected right primer (the primer to the right in the input sequence). i is the 0-based index of the last base of the primer, and n is its length.

PRIMER_INTERNAL_OLIGO=i,n

The selected internal oligo. Primer3 outputs this tag if PRIMER_PICK_INTERNAL_OLIGO was non-0. If primer3 fails to pick a middle oligo upon request, this tag will not be output. i is the 0-based index of start base of the internal oligo, and n is its length.

PRIMER_PRODUCT_SIZE=x

x is the product size of the PCR product.

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_EXPLAIN=s (*)

s is a (more or less) self-documenting string containing statistics on the possibilities that primer3 considered in selecting a single oligo. For example PRIMER_LEFT_EXPLAIN=considered 62, too many Ns 53, ok 9 PRIMER_RIGHT_EXPLAIN=considered 62, too many Ns 53, ok 9 PRIMER_INTERNAL_OLIGO_EXPLAIN=considered 87, too many Ns 39, overlap excluded region 40, ok 8. All the categories are exclusive, except the 'considered' category.

PRIMER_PAIR_EXPLAIN=s (*)

s is a self-documenting string containing statistics on picking a primer pair (plus internal oligo if requested). For example PRIMER_PAIR_EXPLAIN=considered 81, unacceptable product size 49, no internal oligo 32, ok 0

All the categories are exclusive, except the 'considered' category. In some cases primer3 will examine a primer pair before it discovers that one of the primers in the pair violates specified constraints. In this case PRIMER_PAIR_EXPLAIN might have a non-0 number 'considered', even though one or more of PRIMER_LEFT_EXPLAIN, PRIMER_RIGHT_EXPLAIN, orPRIMER_INTERNAL_OLIGO_EXPLAIN has 'ok 0'. PRIMER_PAIR_PENALTY=f

The value of the objective function for this pair (lower is better).

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_PENALTY=f

The contribution of this individual primer or oligo to the objective function.

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_SEQUENCE=s

The actual sequence of the oligo. The sequence of left primer and internal oligo is presented 5' -> 3' on the same strand as the input SEQUENCE (which must be presented 5' -> 3'). The sequence of the right primer is presented 5' -> 3' on the opposite strand from the input SEQUENCE.

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_TM=f

The melting TM for the selected oligo.

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_GC_PERCENT=f

The percent GC for the selected oligo (denominator is the number of non-ambiguous bases).

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_SELF_ANY=f

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_SELF_END=f

The self-complementarity measures for the selected oligo.

PRIMER_PAIR_COMPL_ANY=f

PRIMER_PAIR_COMPL_END=f

The inter-pair complementarity measures for the selected left and right primer

PRIMER_WARNING=s (*)

s lists warnings generated by primer (separated by semicolons); this tag is absent if there are no warnings

PRIMER_{LEFT,RIGHT,PAIR}_MISPRIMING_SCORE=f, s

f is the maximum mispriming score for the right primer against any sequence in the given PRIMER_MISPRIMING_LIBRARY; s is the id of corresponding library sequence.

PRIMER_PAIR_MISPRIMING_SCORE

the maximum sum of mispriming scores in any single library sequence (perhaps a more reasonable estimator of the likelihood of mispriming).

PRIMER_{LEFT,RIGHT,PAIR}_TEMPLATE_MISPRIMING=f

Analogous to PRIMER_{LEFT,RIGHT,PAIR}_MISPRIMING_SCORE, except that these output tags apply to mispriming within the template sequence. This often arises, for example, in genes with repeated exons. For backward compatibility, these tags only appear if the corresponding input tags have defined values.

PRIMER_PRODUCT_TM=f

f is the melting temperature of the product. Calculated using equation (iii) from the paper [Rychlik W, Spencer WJ and Rhoads RE (1990) "Optimization of the annealing temperature for DNA amplification in vitro", Nucleic Acids Res 18:6409-12 http://www.pubmedcentral.nih.gov/articlerender.fcgi? tool=pubmed&pubmedid=2243783]. Printed only if a non-default value of PRIMER_MAX_PRODUCT_TM or PRIMER_MIN_PRODUCT_TM is specified.

PRIMER_PRODUCT_TM_OLIGO_TM_DIFF=f

f is the difference between the melting temperature of the product and the melting temperature of the less stable primer. Printed only if PRIMER_MAX_PRODUCT_TM or PRIMER_MIN_PRODUCT_TM is specified.

PRIMER_PAIR_T_OPT_A=f

f is T sub a super OPT from equation (i) in [Rychlik W, Spencer WJ and Rhoads RE (1990) "Optimization of the annealing temperature for DNA amplification in vitro", Nucleic Acids Res 18:6409-12 http://www.pubmedcentral.nih.gov/articlerender.fcgi? tool=pubmed&pubmedid=2243783]. Printed only if PRIMER_MAX_PRODUCT_TM or PRIMER_MIN_PRODUCT_TM is specified.

PRIMER_INTERNAL_OLIGO_MISHYB_SCORE=f, s

f is the maximum mishybridization score for the right primer against any sequence in the given PRIMER_INTERNAL_OLIGO_MISHYB_LIBRARY; s is the id of corresponding library sequence.

PRIMER_{LEFT,RIGHT,INTERNAL_OLIGO}_MIN_SEQ_QUALITY=i

i is the minimum _sequence_ quality within the primer or oligo (not to be confused with the PRIMER_PAIR_QUALITY output tag, which is really the value of the objective function.)

PRIMER_{LEFT,RIGHT}_END_STABILITY=f

f is the delta G of disruption of the five 3' bases of the primer.

PRIMER_STOP_CODON_POSITION=i

i is the position of the first base of the stop codon, if primer3 found one, or -1 if primer3 did not. Printed only if the input tag PRIMER_START_CODON_POSITION with a non-default value is supplied.

Advice for picking primers

We suggest consulting Wojciech Rychlik (1993) "Selection of Primers for Polymerase Chain Reaction" in BA White, Ed., "Methods in Molecular Biology, Vol. 15: PCR Protocols: Current Methods and Applications", pp 31-40, Humana Press, Totowa NJ.

Some of the most important issues in primer picking can be addressed only before using Primer3. These are sequence quality (including making sure the sequence is not vector and not chimeric) and avoiding repetitive elements.

Techniques for avoiding problems include a thorough understanding of possible vector contaminants and cloning artifacts coupled with database searches using BLAST, FASTA, or other similarity searching program to screen for vector contaminants and possible repeats. Repbase (J. Jurka, A.F.A. Smit, C. Pethiyagoda, and others, 1995-1996, ftp://

ncbi.nlm.nih.gov/repository/repbase) is an excellent source of repeat sequences and pointers to the literature. (The Repbase files need to be converted to FASTA format before they can be used by Primer3.) Primer3 now allows you to screen candidate oligos against a Mispriming Library (or a Mishyb Library in the case of internal oligos).

Sequence quality can be controlled by manual trace viewing and quality clipping or automatic quality clipping programs. Low quality bases should be changed to N's or can be made part of Excluded Regions. The beginning of a sequencing read is often problematic because of primer peaks, and the end of the read often contains many low-quality or even meaningless called bases. Therefore when picking primers from single-pass sequence it is often best to use the INCLUDED_REGION parameter to ensure that Primer3 chooses primers in the high quality region of the read. In addition, Primer3 takes as input a Sequence Quality list for use with those base calling programs (e.g. Phred, Bass/Grace, Trout) that output this information.

What to do if Primer3 cannot find any primers?

Try relaxing various parameters, including the self-complementarity parameters and max and min oligo melting temperatures. For example, for very A-T-rich regions you might have to increase maximum primer size or decrease minimum melting temperature. It is usually unwise to reduce the minimum primer size if your template is complex (e.g. a mammalian genome), since small primers are more likely to be non-specific. Make sure that there are adequate stretches of non-Ns in the regions in which you wish to pick primers. If necessary you can also allow an N in your primer and use an oligo mixture containing all four bases at that position.

Primer and probe screening analysis

MacVector provides three screens for likely primers and probes:

- screening a nucleic acid sequence for likely PCR primer pairs
- screening a nucleic acid sequence for likely sequencing primers or hybridization probes
- screening a protein sequence for the least degenerate oligos that can serve as hybridization probes.

The first two work in a similar way and will be discussed together.

Screening a nucleic acid for primers or probes

There are several characteristics to consider when choosing an oligonucleotide to use as a sequencing primer, PCR primer, or hybridization probe:

- it should form a stable duplex with the target sequence under the conditions to be used in the reaction (temperature, salt concentration, and the oligo concentration)
- it should be specific for the intended target sequence, and not bind to other regions within the sequence
- it should not anneal to itself, to another copy of itself, or, in the case of PCR primer pairs, to a copy of its sister primer (hairpin or dimer formation).

Hybridization probes should be checked for possible competing binding activity; with PCR primer pairs or sequencing primers, it is more important to check the 3' end of the primer.

A computerized method can greatly reduce the amount of work needed to select a primer or probe from a nucleic acid sequence. There are two basic approaches. The first is user-oriented—the user selects a possible primer, inputs it to the computer program, and receives information about the primer that help to decide whether or not to use the primer. This can be time-consuming and demands a certain amount of knowledge from the user, but is a good method for a researcher with special needs. MacVector also offers a second approach - an automatic screening method. The user tells the program the criteria that constitute a bad primer, and the program scans the sequence, eliminating “bad” primers and displaying a list of the primers that pass this screening process. While it does not allow the degree of control that the first method provides, it works well in most cases, demands less of the user, and is faster.

MacVector’s analyses for finding PCR primer pairs and single primers for sequencing (or hybridization probes) have most of their parameters in common. You control the following:

- whether to scan an entire nucleic acid sequence or designate a shorter region (for sequencing primers / probes you can also restrict the scan to a given strand)
- the range of melting or dissociation temperatures that the oligo-sequence duplexes should have
- the required G+C content of the oligo

- the required range of oligo lengths
- the sequence of an oligo at the 3' end (a G or C “anchor,” for example).

Both analyses use the same parameters for eliminating oligos that have too much secondary structure. You control the maximum number of consecutive bonds an oligo has when it forms either a hairpin with itself or a duplex with another oligo. You can set separate limits on the following bond types:

- any bond
- G-C bonds only
- bonds only between the 3' ends of oligos.

The default values for the first two were suggested by Lowe *et al.* (1990), for the third by Rychlik and Rhoads (1989).

You can also eliminate oligos that may bind to other sites on the sequence. For PCR primer pairs, the 3' end of each primer of a pair is compared with the product that the pair amplifies. For sequencing primers / probes, the 3' end of the oligo or the entire oligo is compared against the entire sequence, or a specified region of the sequence. As with the scan region, you can limit this comparison step to a single strand.

The final set of parameters relate to the reaction conditions. The monovalent salt concentration (Na and K salts) and the oligo concentration are needed in order to compute the melting temperatures.

What if no primers or probes are found? Use the output statistics as a clue, and see what the most common reasons for oligo rejection were. Try expanding your criteria. The first two parameters that are tested by MacVector are the 3' dinucleotide and the primer length, so these are the first two that you should adjust. If the dinucleotide was set to NS, set it to NN to increase the number of oligos that will pass this first criterion. If you are using a narrow range of primer lengths or if they are skewed toward the longer sizes, widen the range and include shorter lengths. Sometimes you may have difficulty finding compatible PCR primer pairs because of widely different G+C content (and thus different Tm values) along the sequence. For example, many coding regions contain an A+T-rich region just upstream of the start of the coding region, while the coding region itself and the sequence downstream may be G+C-rich. If you try to find one primer upstream of the coding region and the other downstream, you may need to expand the range of G+C content and the range of temperatures.

Screening a protein to find a hybridization probe

If you want to find an oligo to use as a hybridization probe for a gene whose exact DNA sequence is not known, but whose amino acid sequence is known, you could reverse-translate the amino acid sequence and then scan the resulting DNA sequence to find a region that was not very degenerate to minimize the number of oligonucleotide probes that would have to be synthesized. This is easier to do with a computer than by hand, and is a feature provided by MacVector as part of the reverse translation analysis. For each probe length that you specify, MacVector scans the sequence, finds the oligos of that length that are least degenerate, and outputs a list of their sequences, dissociation temperatures, G+C content, and the number of permutations (the number of oligos that would have to be synthesized to cover all possible sequences represented by the degenerate oligo). No further analysis of the probes, such as checking for secondary structure or alternate binding sites, is performed.

You can reduce the degree of degeneracy by modifying the genetic code that is used to perform the reverse translation. For example, serine has six codons in the universal genetic code, four of the form TCN and two of the form AGY. These two forms reduce to the highly degenerate codon WSN.

If you know that your organism uses primarily the AGY codons, you can create a new genetic code that is missing the TCN codons, so that serines will be reversed translated as TCN instead of WSN. See *“Modifying genetic codes”* on page 242, for further details.

Computing the Tm and Td

While the terms Tm (melting temperature) and Td (dissociation temperature) are often used interchangeably, they are not quite the same. The melting temperature is used to describe the separation of a duplex nucleic acid molecule in solution solely because of changes in temperature of the solution. The dissociation temperature is used to describe the temperature at which a duplex bound to a substrate (as in a filter hybridization) dissociates because of temperature and washing effects. Rychlik and Rhoads (1989) estimate that Td is approximately 7.6° C lower than the Tm.

Because of its simplicity, a rule-of-thumb equation is often used to calculate the Td of an oligonucleotide duplex: 4 (number of G+C bases) + 2 (number of A+T bases). This equation provides a rough estimate for oligonucleotides between ten and twenty bases long. A more accurate determination of Tm or Td is obtained using nearest-neighbor methods, which provide valid estimates over a wider range of oligo lengths (Rychlik and Rhoads (1989), Rychlik *et al*. (1990), based on values published by Breslauer *et al*. (1986) for DNA and Freier *et al*. (1986) for RNA). In computing Tm and Td values, MacVector uses nearest-neighbor methods for oligos up to thirty bases long. For longer oligos, or when the effects of formamide are taken into account, the equation of Baldino *et al*. (1989) is used for DNA-DNA hybrids, Bodkin and Knudsen (1985) for RNA-RNA hybrids, and Casey and Davidson (1977) for DNA-RNA hybrids.

One factor that complicates Tm calculations for PCR primers is that the primers are consumed during the course of the reaction, leading to a wide variation in their actual concentration. Rychlik *et al*. (1990) examined this experimentally and suggested an adjustment factor that could be used to calculate an effective Tm which would be more applicable to real-world PCR reaction conditions. MacVector uses this factor in PCR primer calculations, yielding adjusted values that are typically 2^{O}C to 10^{O}C lower than the values obtained for sequencing primers.

Coding regions

After you determine the sequence of a piece of DNA, one of the first things you would like to know is whether it codes for a protein. If you are lucky, your sequence will contain at least one long open reading frame—a reading frame of at least 50 to 100 codons that contains no stop codons. You can translate the open reading frame and search the NBRF Protein Identification Resource database or the GenBank nucleic acid database to see if there is a match with a known protein. If you find a match, you will have answered your question.

If there is no match, you need some other method of determining the biological significance of the open reading frame. It may code for a previously unsequenced protein, or it may have no biological significance whatsoever - after all, not all open reading frames are protein coding regions.

MacVector provides a range of analyses to help you make this decision. In addition to open reading frame analysis, the program provides various methods use that base or codon composition to help you determine if your open reading frame has the characteristics of a protein coding region. In conjunction with these methods, you can use nucleic acid subsequence analysis to look for motifs in your sequence, such as ribosome binding sites or intron-exon splice sites, that may help define the exact boundaries of coding regions.

Open reading frame analysis

Unlike many other sequence analysis programs, MacVector enables you to define which codons are to be used as start and stop codons. For prokaryotic sequences, you would normally set ATG as a start codon, and possibly also GTG and TTG. The termination codons TAA, TAG, and TGA should be assigned as stop codons, unless your sequence contains suppressor mutations.

In eukaryotic sequences, there may be exons present that do not start with ATG. Therefore, in addition to using ATG as a start codon, you could also assign the stop codons (TAA, TAG and TGA) as start codons, in which case MacVector will treat the codon immediately after a stop codon as a start codon.

You can also designate the beginning and end of the sequence to act as ORF starts or stops. This can be useful if you suspect that the open reading frame for your sequence extends beyond the ends of the sequence.

Base composition methods

MacVector provides four methods that use base composition to predict protein-coding regions. These are based on bias in the overall nucleotide base composition, or on the tendencies for bases to occur at particular codon positions. These methods do not use organism-specific information such as codon bias tables.

G+C % composition

Many organisms have a skewed composition of G+C bases, which results in a strongly biased codon composition. In such cases, non-coding regions tend to reflect the overall G+C percentage, but coding regions, under the constraints of natural selection, show a different pattern. Because of redundancy in the genetic code, there is less selection on the third codon position, so that its G+C % shows a strong bias. The second position shows no bias, as there is no degeneracy in the code. The first position has some bias, due to amino acids such as Leu that can be coded by six different codons. The overall effect in an organism such as *Streptomyces coelicolor* is that coding regions tend to have ~70% G+C at the first position, ~50% at the second position and > 95% at the third. In non-coding regions, all three plots will tend to be around the average G+C% for the DNA.

The algorithm for calculating the percentages is simple. Starting at base 1 and sampling every third base in a window of *t* codons (i.e. base 1, 4, 7, 10, ...), count the number of G and C bases and plot their combined percentage. Slide the window one codon to the right, and plot again; repeat until the end of the sequence. Now repeat the procedure, this time starting at base 2 (i.e. sampling 2, 5, 8, 11, ...), and plotting the % points in a second color. Finally, repeat once more, starting at base 3 (i.e. sampling 3, 6, 9, 12, ...) and plotting the % points in a third color.

Fickett's TESTCODE algorithm for predicting coding regions

As DNA sequences accumulated in databases, researchers examined known coding and non-coding regions and discovered that there were characteristic differences in their base compositions. First of all, because of selection pressure at the amino acid level, the triplet frequencies differ. Some amino acids are more common than others and thus their codons are repeated more often within coding regions than within non-coding regions. As a consequence, nucleotides tend to be repeated with a periodicity of three in coding regions. In addition, coding regions tend to have a slightly higher G+C content than non-coding regions.

Starting from these observations, Fickett (1982) examined coding regions of sequences from the GenBank database and developed a statistical method that could be used to predict coding regions in DNA sequences of at least 200 base pairs. This method uses eight parameters that are based on the periodic properties of each of the four nucleotides within the sequence.

The first four parameters are called position parameters. A count is made of the number of times A appears in codon position 1 (A1), the number of times it appears in codon position 2 (A2), and the number of times it appears in codon position 3 (A3) in the sequence. The first parameter, A-Position, is then defined as:

$$\text{A-Position} = (\max(A1,A2,A3)/\min(A1,A2,A3))$$

The position parameters for the other three bases are computed similarly. The last four parameters are the content parameters. These are just the percentage composition of the sequence for each of the four bases (%A, %C, %G, %T).

Fickett calculated each of the eight parameters for a number of known coding and non-coding regions. He assigned weights (W) to each of the parameters according to the percentage of the time that a coding region could be correctly predicted using the value of that parameter alone. He also derived a “probability of coding” (P) for ten intervals of each of the parameter values. For example, he looked at all regions whose A-Position parameter value fell between 0.0 and 1.1. Twenty-two per cent of these were coding regions, so for his interval, P1 = 0.22. Regions with an A-Position parameter value between 1.6 and 1.7 were distributed as 93% coding and 7% non-coding, so for this interval, P1 = 0.93. The weights and probabilities for each of the eight parameters were combined to define a single TESTCODE indicator value:

$$\text{Indicator} = P1W1 + \ldots + P8W8$$

Fickett next examined the distribution of the indicator values of 321 coding and 249 non-coding sequence fragments. Only 29% of the regions with indicator values below 0.74 were coding regions, while 92% of regions with indicator values above 0.95 were coding regions.

In the borderline areas were regions scoring between 0.74 and 0.84 (40% were coding) and regions scoring between 0.84 and 0.95 (77% were coding regions).

As it was originally defined, Fickett’s method assigns a region with a coding probability below 0.29 as a “non-coding” region and a region with a probability of coding above 0.92 as a “coding” region. Regions with probabilities between these values are tagged “no opinion.”

Fickett’s method has two limitations. As we mentioned above, it may not give valid results for sequences shorter than 200 bases (67 codons). Moreover, it cannot distinguish the correct reading frame - or indeed the correct strand.

If you have more than one open reading frame in a region that is assigned as coding, you must use another method to determine which reading frame is the correct one.

As an interesting aside, Fickett notes that many of the known coding regions that this method mis-classifies as non-coding are areas of the sequence where some specialized use is made of the DNA. The mechanisms responsible for ensuring this specialized use are apparently stronger than the forces that cause the difference in coding frequencies. Some examples are viruses with overlapping genes and the variable regions of immunoglobulin genes.

Uneven positional base frequencies

In protein coding regions, the four nucleotide bases do not occur equally frequently in the three codon positions. This algorithm, developed by Staden (1984), measures the deviation from an even distribution across codon positions. It does not require prior knowledge of the organism's codon usage. A single plot is generated, showing the likelihood that a region of the sequence is protein coding. The method cannot distinguish which frame might be the actual coding frame.

In a sliding window of *t* codons, the algorithm calculates how many times each base *i* occurs in position *j* of a codon, to give an array *N*[*i*,*j*]. It then calculates the expected non-coding value for each base in each codon position for that window, i.e.:

$$E[i, 1] = E[i, 2] = E[i, 3] = \frac{(N[i, 1] + N[i, 2] + N[i, 3])}{3}$$

Finally, it computes the divergence from an even distribution across codon positions, by summing the absolute differences between observed and expected values for each of the 12 *i*,*j* values (4 bases X 3 positions):

$$D = \Sigma|E[i, j] - N[i, j]|$$

The window is moved along the sequence one base at a time, and the calculation of D repeated. The results are displayed as a single plot. Seventy-six per cent of coding regions have a *D* score higher than 0.78, and 76% of non-coding regions score lower than 0.78.

Positional base preference

Like the uneven positional base frequencies algorithm, this was also developed by Staden (1984). However, this method can show the likely protein coding frame.

It is based on the observation that, across a large sample set of protein coding regions from different species, the first two codon positions show marked base specificity. It estimates how closely a potential open reading frame matches this base composition.

A window of *t* codons is moved along the length of the sequence in increments of 3 bases. Within each window, each codon is scored by adding up three values from this table:

Frame	**1**	**2**	**3**
A	27.67	30.97	23.96
C	21.08	23.78	25.06
G	33.57	18.18	25.92
T	17.68	27.07	25.06

For instance, A at position 1 scores 27.67, and if the next base is T, then 27.07 is added. The score for each window is simply the sum for each codon in the window. Three scores are calculated for each window, starting from positions 1, 2 and 3 in the window. Finally, the user can choose to view the values as absolute or relative. In absolute mode, the score for a window is divided by the window size to get a normalized score. In relative mode, the score for each position is divided by the sum of the scores for all three positions - this accentuates differences between the frames, at the cost of some increased noise.

Codon preference and codon bias tables

There is another way that codon usage in protein coding regions may differ from that in non-coding regions. Most organisms exhibit a preferential use of codons according to the abundance of the corresponding tRNA species. In such organisms, non-coding regions exhibit no codon preference. Regions that code for proteins having a low level of expression tend to have a codon distribution that parallels the abundance of the corresponding tRNA species. Regions that code for highly-expressed proteins are biased even further—they preferentially use those codons that correspond to the most abundant tRNA species for that amino acid. For these organisms, the codon preference can be used as a basis for determining the likelihood that a given region of DNA codes for a protein. MacVector offers three algorithms for identifying coding regions using codon preference.

To use codon preference methods, you must first determine empirically the degree of codon bias that the organism possesses, by creating a codon bias table for that organism. Because codon usage in mitochondrial DNA, for example, is typically different from that of chromosomal

DNA, you should choose only sequences from the same genome source. A sequence database is searched for all sequences belonging to a given organism that contain coding regions. (MacVector uses regions of the sequence that are listed as “pept” in the features table.) The number of occurrences of each of the 64 codons in these regions is added up and divided by the total number of codons in the search to obtain a codon frequency f_{abc} for each codon *abc*. The frequency of each codon family (all codons that encode the same amino acid) is then calculated by summing the frequencies for each codon in the family. For example, the frequency of the leucine family would be:

$$F_{leu} = f_{CTT} + f_{CTC} + f_{CTA} + f_{CTG} + f_{TTA} + f_{TTG}$$

Lastly, the relative frequency of each codon within the family is obtained by dividing that codon’s frequency by the codon family frequency. The relative frequencies are stored in the codon bias table.

Codon preference methods are not always applicable. Some organisms do not exhibit a codon bias. For other organisms, there are not enough highly-expressed gene sequences in the database to derive a useful codon bias table. Even if a codon preference is present, it may only be useful for locating genes of highly-expressed or moderately-expressed proteins. Weakly-expressed genes usually have little codon bias, and plots of these sequences can be difficult to interpret.

Gribskov codon preference

This algorithm is used in the Wisconsin package. It uses a 64-codon bias table, and identifies coding regions by measuring the relative occurrence of codons that are alternatives for specifying the same amino acid. The Gribskov method normalizes for the base composition of the sequence. The algorithm calculates a preference parameter P_{abc} for each codon *abc*, as follows:

Given a target sequence of length *N*, the base composition is first calculated, generating the four probabilities N_A, N_C, N_G and N_T.

The next step is to determine f_{abc}, the absolute frequency of the codon in the supplied usage table. The frequency of its codon family (i.e. all the codons that code for the same amino acid as *abc*), F_{abc}, is then calculated as the sum of the frequencies of the family members. Next, the "random" probability of *abc* in the target sequence, r_{abc}, is calculated as:

$$r_{abc} = (N_a N_b N_c)/N^3$$

and the random probability of the *abc* codon family, R_{abc}, is obtained as the sum of the r_{abc} values in the family. The preference parameter for each codon can then be calculated as:

$$P_{abc} = (f_{abc}/F_{abc})/(r_{abc}/R_{abc})$$

For a window of length *w*, the P_{abc} values for each codon in the window are multiplied together, and the preference statistic is the *w*th root of this product. To simplify calculations, this is computed using logarithms.

Where a codon in the supplied table has a zero probability, it is assigned a probability equal to the reciprocal of the sum of codons in its codon family. If it has no family, the reciprocal of the total number of codons in the table is used.

MacVector codon preference

The MacVector codon preference analysis is a variant of the Gribskov algorithm that is not normalized for base composition. It is included primarily to allow comparisons with results from earlier versions of MacVector.

Staden codon preference

In its basic form, the algorithm of Staden & McLachlan (1982) uses the "raw" codon probabilities contained in codon usage tables such as those generated by MacVector. The table contains 64 entries, one for each codon; the sum of their probabilities is 1.0. The algorithm proceeds as follows:

A window of *t* codons is moved along the sequence in increments of 3 bases. For each window, three scores are calculated for *t* codons starting at positions 1, 2, and 3. The coding probability for each position is calculated by multiplying together the probabilities of each codon. The final score for each position is then divided by the sum of the probabilities in all three positions to produce a relative coding probability.

A variation of the method uses the coding probabilities normalized to amino acid composition. In this case the probabilities of all six Leucine codons would add up to 1.0, and the single Tryptophan codon (TGG) would always have a probability of 1.0.

Interpreting coding preference plots

The main use of the coding preference plots is to help identify potential protein coding open reading frames in a piece of DNA. Interpretation of the plots will depend to a large extent upon the source of your DNA, and in particular whether or not introns are likely to be present. It can be reasonably straightforward for sequences where introns are absent or rare, e.g. prokaryotic sequences, cDNA clones, many mitochondrial genomes, and simple eukaryotes like *S. cerevisiae*.It can also be complicated by sequencing errors, particularly insertions or deletions, which may cause frame shifts in potential coding regions.

The following tips and suggestions may help you to get the best results from your coding preference analyses.

Sequence length

The length of the analyzed sequence affects how easy it is to interpret the graphs. As most protein-coding open reading frames are less than 3kb in length, it is easier to visualize the plots if the visible on-screen range is less than 15kb. You can easily zoom in on longer sequences by selecting a region with the mouse or using the arrow keys.

Window size

The window size used for each of the algorithms is a trade-off between ease of visualization and the identification of short protein-coding regions. Longer windows give smoother plots with less noise. However, they may miss short open reading frames and may also make it difficult to determine the most likely start codon, where several alternatives exist.

Separate vs. combined mode

Algorithms that generate three plots can be viewed in separate or combined modes. The combined mode, where all three plots appear on the same graph, is very useful for accentuating differences between the frames. However, using this mode, it is not possible to display the frame-specific starts and stops with the plot.

One way to overcome this problem is to turn on just the ORF plot and the algorithm of interest. The two plots will then appear next to each other in the output window permitting color-coded visual comparison between the codons in the ORF plot and the algorithm plots.

The combined mode is also very useful for identifying the ends of open reading frames, as the three plots typically cross each other at this point. You can even zoom in to the residue level to identify the exact base at which the crossovers occur.

The separate mode is useful for a tighter integration of the codons and the frame-specific plots. For most algorithms, the start and stop of a protein-coding region often corresponds to the point at which the frame-specific plot crosses the midpoint. By turning on the start and stop codon displays with the center location, the most likely start codon and corresponding stop codon can be clearly seen.

Open reading frames

In searching for coding regions, ORF plots are often a good starting point. In many genomes, the G+C% is less than or equal to 50, and so the standard stop codons (TAG, TAA, TGA) appear quite frequently. As a result of this, ORFs longer than 50 amino acids rarely occur, and are a good indication of protein coding. However, in DNA with a high G+C%, longer ORFs occur frequently by chance, since the A/T-containing stop codons are rare. In such cases the ORF plots are less informative.

Another drawback of using ORF plots is that they are very susceptible to frameshift sequencing errors.

If you are dealing with bacterial sequences, you may want to use the **Universal + GUG start** genetic code to see more potential ORFs. (In *E. coli* over 10% of protein-coding ORFs start with this codon).

Fickett's TestCode and Uneven Positional Base Frequencies

These methods are both useful as a general guide for identifying regions of sequence that exhibit a biased base composition typical of protein-coding regions. Neither method can identify the likely frame. TestCode scores of greater than 0.92 are considered highly likely to represent protein coding regions. Similarly, Uneven Positional Base Frequency scores greater than 0.78 are also highly likely to encode proteins. Select the **use fixed scaling** option to view the plots in an optimal way - this will truncate scores above or below the optimal ranges. Turning this off prevents truncation, allowing you to see regions with exceptionally high or low scores.

Positional Base Preference

This plot is useful if you need to confirm the frame used by a coding region, but do not have additional codon usage information. Plots with values greater than 1.0 indicate a bias towards coding, whereas plots below 1.0 indicate a bias towards non-coding. Typical coding regions will have one plot significantly above 1.0, with the other two plots considerably below. Non-coding regions have all three plots close to 1.0. Plotting the scores in relative mode can help to accentuate the differences between the frames although it can introduce some noise. The fixed scaling mode displays the plots with an optimal range for identifying coding regions, although there will be truncation of high and low scores.

Codon Preference Plots

If you have access to codon usage tables for your organism or genome, the three Codon Preference Plots offer the most sensitive way of identifying coding regions. The MacVector and Gribskov plots are essentially the same, except that the Gribskov algorithm normalizes for the G+C% composition of the DNA. For 50% G+C DNA, they produce essentially identical results. For skewed G+C% DNA, the MacVector algorithm may be overly optimistic in assigning likely coding regions. All three plots allow you either to view the entire range, or to scale to twice the standard deviation. When most of the DNA is thought to encode protein, scaling to **2x Std. Dev.** helps to distinguish between the different open reading frames. However eukaryotic genomes, where coding regions are relatively rare, are best viewed at full scale: otherwise the background noise becomes amplified and interferes with the visualization.

G+C%

The G+C% plot is extremely useful for analyzing genomes that diverge significantly from 50% G+C, although it can still be useful for other organisms, as many species exhibit a codon-specific G+C% bias in coding regions. The plot is harder to interpret than the other algorithms.

Consider a genome with an overall G+C composition of 75%. In non-coding regions of DNA, there will be no codon-specific G+C bias, and the three plots will be poorly separated, all showing G+C percentages close to the average for the test DNA. However, where proteins are encoded, the plots will diverge so that the first plot averages 70%, the second 50% and the third 95%. This means that although the third plot lies nearest to the top of the graph, it is actually the middle line that represents the likely frame of the coding region. For high A+T% DNA, the converse is true (the third position will be particularly AT rich), although the middle line still represents the frame.

The situation becomes more complex when we consider issues of strandedness. The MacVector color defaults are: frame 1 = blue, frame 2 = green, and frame 3 = red. For coding regions on the top (or forward) strand, if the blue frame (position 1) represents the first position and is the middle line, then the green line should be the lower line and the red line should be uppermost. Similarly, if the red line represents the first position, then blue should be the lower line (position 2) and green the upper line (position 3). For coding regions on the opposite strand, however, the order of the second and third lines is swapped, so that if position 1 is blue, position 2 will be red, and position 3 green. If you are having difficulty, copy the sequence to a separate document, reverse complement it, and repeat the analysis.

Effect of introns

Intron-containing genomic DNA can be analyzed in much the same way, except that you should be aware that:

- internal exons may not have start or stop codons, and
- many coding exons may be less than 50 amino acids in length.

Ideally, you should combine the analysis with the results of searches for potential splice-site donor and acceptor sequences. As you may find large regions of non-coding DNA, it is useful to use the zoom functionality to home in on potential coding regions.

Sequencing errors

Sequencing errors can interfere with this analysis, particularly if they introduce frameshifts into the sequence. These often show up as areas where the plots appear to indicate that the frame has changed, but there is no corresponding start/stop codon. Use the zoom function to home in on the area where the plots cross, to identify the likely region containing the error. Re-evaluation of the original sequencing data will often reveal the source of the error.

You can use coding preference plots as a method for detecting these sequencing errors.

21 Understanding Sequence Comparisons

Overview

This chapter explains some of the theories and methods used by MacVector for sequence comparison. Users are referred to the academic papers from which the methods are derived. This chapter provides greater detail only where significant changes have been made to the methods.

Refer to Appendix F, *“References”*.

Contents

Pustell matrix analysis

When looking for similarities between two sequences, a matrix comparison is the method of first choice. A matrix analysis is unsurpassed for getting an overall picture of how sequences are related, and it can detect features that other methods may miss.

By displaying results graphically, a matrix method takes advantage of the superb pattern recognition capabilities of the human eye to allow detection of even weak similarities between two sequences. Computational alignment methods usually report only one “best” alignment between two sequences; the matrix method will reveal any significant alternative alignments that may exist, the presence of weaker regions of similarity, such as duplications, and the existence of rearrangements. Computational region-finding methods present the researcher with a list of matching regions; the matrix method displays matching regions in the context of the sequence as a whole, making it easy to determine if the regions are repeats or inverted repeats, for example.

The simplest form of matrix analysis (also called a dot plot or dot matrix analysis) can be done by hand for short sequences. The residues of one sequence are written along the X-axis of a two-dimensional graph and the residues of the other are written along the Y-axis. A dot is placed at each x,y coordinate where the residues of the two sequences are identical. Regions of similarity stand out as diagonal dotted lines against an amorphous cloud of dots.

The patterns of the diagonals in the plot can tell you a great deal about the sequences. A long unbroken diagonal from the upper left corner to the lower right corner indicates that the sequences are identical (this is called the identity diagonal). If the diagonal runs from the lower left corner to the upper right corner, one sequence is the reverse complement of the other. Short diagonals that parallel the identity diagonal represent repeats; diagonals paralleling the reverse diagonal represent reverse repeats. Breaks appearing within a diagonal represent mismatching regions; a gap is seen as a diagonal that stops, then continues at a different level. In the `Sample Files` supplied with MacVector are a number of sequences that demonstrate these patterns. If you compare SV4CG with itself and zoom in to view the region between bases 1-200, you will see examples of repeats and reverse repeats. Comparing the protein sequences CVJB and FVVFBA shows an example of a rearrangement.

MacVector's Pustell matrix analyses are more sophisticated variations on the simple dot-plot method. They offer a number of parameters to customize the analysis—more than most implementations do. In general, leaving the parameters set at their default values will result in a usable matrix plot. Once you know more about your sequence, for example, if it contains coding regions, a known transcriptional control region, and so on, the various parameters will become more useful.

The parameters are:

- the scoring matrix
- hash size
- window size
- minimum percent score
- jump parameter.

The values in the scoring matrix are the weights given to different kinds of matches. If you care only about perfect matches (A-A, G-G, etc.), you would use an identity matrix—one that scores exact matches as +1 and mismatches as 0. But under some situations for nucleic acids and, more frequently, for amino acids, you may wish to consider some mismatches to be partial matches according to criteria such as hydrophobicity, charge, or structure, and to consider some mismatches to be worse than others. By altering the values in the scoring matrix, you can control the matching criteria.

The hash size is also known as the k-tuple size or the word size. When the analysis initially scans the sequences for matches, it uses a very fast technique known as hashing. When the hash size is set to six, the program will only consider exact matches of six residues to be a "hit." If it is set to one, it will consider a single residue match to be a hit, and so on. Using larger hash sizes will result in a less sensitive but faster search.

The jump parameter is used in conjunction with the hash size. When the jump parameter is set to one, the "word" used in the hash step is formed from consecutive residues. So if the hash size is six, and the jump is one, a word consists of six consecutive residues. If the jump parameter is set to three, the word is composed of every third residue. Therefore if the hash size is six and the jump is three, the word consists of the first residues of six consecutive triplets.

When a hit is found, the program begins the second step of the comparison, the scoring step. In this step, the scoring matrix, window size parameter, and minimum percent score come into play. The hit is placed in the center of a “window” whose length is the window size set by the user. The window is then scored, using the values for the residue pairs that are found in the scoring matrix. If the score meets or exceeds the minimum percent score set by the user, the location of the window is saved by the program.

As with hash size, larger values for the window size result in less “noise” in the matrix, but the sensitivity of the search is reduced. Experience has shown that for two random sequences, the minimum percent score should be set no lower than 65% if an identity scoring matrix is used.

Let us now illustrate these concepts with an example, matching a short sequence against itself:

```
AGCCTCGATCGATTCGATCGGCGCTAGCTAGGCTAG
```

Hashing both copies of the sequence at a hash size of three will make the program look at the following three-letter “words” in both sequences to see if there are any matches:

```
Sequence A words:
1 AGC CTC GAT CGA TTC GAT CGG CGC TAG
CTA GGC TAG
Sequence B words:
1 AGC CTC GAT CGA TTC GAT CGG CGC TAG
CTA GGC TAG
```

Since these two sequences are identical, there is obviously a match starting with the first base of each sequence. But other shorter matches may exist. Note that there are two GAT words in each sequence, which means a region of similarity may also exist centered at residue 7 in A and 16 in B. Let’s assume the window size is 12. The 65% match criterion means that at least 7 of the 12 residues in a window must match. Obviously, the window containing the first GAT word in both sequences will be a valid window since 12 out of 12 residues match.

But what about the regions of the two sequences where the first instance of GAT in A matches the second instance of GAT in B?

```
1  AGC CTC GAT CGA TTC GAT CGG CGC TAG
CTA ...
    *   ** *** **

10 CGA TTC GAT CGG CGC TAG CTA GGC TAG
```

There are eight matches in the 12-base window surrounding the matching word GAT, so this window is also a valid match. After checking the windows that surround each of the matching words, the program shifts one base in the second sequence, rehashes the sequence in this phase, and compares the resulting set of words for sequence B with the original set of words for sequence A:

```
1  AGC CTC GAT CGA TTC GAT CGG CGC TAG
CTA ...
     *             *
2  GCC TCG ATC GAT TCG ATC GGC GCT AGC
TAG ...
```

Again, the program checks the windows surrounding each matching word to locate valid windows. Again, it shifts one base over in the second sequence and repeats the hashing and scoring process for the third time.

Now we'll consider the effect of the jump parameter. A jump setting of one with a hash size of three applied to our test sequence gives the words you have already seen:

```
AGC CTC GAT CGA TTC GAT CGG CGC TAG CTA
GGC TAG
```

But with a jump setting of three and a hash size of three, we get a different set of words:

Original sequence:

AGC**C**TC**G**AT**C**GA**T**TC**G**AT**C**GG**C**GC**T**AG**C**TA**G**GC**T**AG

After the jump is applied, the sequence is:

```
A C G C T G C C T C G T
```

The words obtained from hashing the processed sequence are:

```
ACG CTG CCT CGT
```

Notice that we are still comparing words of size three, but they are derived from different "samples" of the same sequence.

The jump parameter is useful when we compare nucleic acid sequences that code for proteins. Imagine comparing two distantly related protein-coding regions using a hash level of three and a jump setting of one. This means that no windows are scored unless the program finds three bases in a row that match. Because the triplet code's redundancy allows much variation in the third base, these sequences could code for identical proteins, yet still have no matching triplets because of different codon usage. In such a case, the matrix analysis would show no diagonals! If we used a jump setting of three, we would be comparing every third base. We would get strong signals when we coincided the first and second bases of the codons, even though there would be no signal for the third bases. Now the matching coding regions should show up as obvious diagonals in the matrix display.

Lipman-Pearson DNA and Wilbur-Lipman protein library searches

These high-speed sequence library search routines have been a boon to molecular biologists. A full local alignment of a query sequence against a whole library of sequences would take an enormous amount of time, even for a supercomputer. By using an initial quick hashing step to screen the sequences, as is done in the matrix comparison discussed in *"Pustell matrix analysis"* on page 454, the time-consuming alignment process is only performed on sequences that are shown to have a certain degree of similarity. MacVector uses these routines to compare a query sequence to a folder containing one or more sequences.

There are a number of differences between these analyses and the matrix comparisons. While the matrix comparison will display all matching regions that is finds, these library search methods will only show the one "best" alignment between the matching regions of two sequences.

Both methods use the hashing technique as the first step in comparing the sequences. However, the library search methods do not allow you to change the jump parameter: it is permanently set at one. As with the matrix analysis, when a hit is found, the second step of the analysis is a scoring step using the values in the scoring matrix. Unlike the matrix analysis, there is no fixed window size for the scoring step and no percent score is taken into account—the scoring starts with the hit and continues until a significant region of mismatch is encountered. The score obtained from this step is called the initial score.

Unlike the matrix analysis, there is a third step to the comparison: a full local alignment. Three additional parameters are involved:

- the cut-off score
- the deletion penalty (also called the gap-opening penalty)
- the gap penalty (also called the gap-extension penalty).

These parameters are accessed by opening the scoring matrix file and clicking the Tweak button (labeled with an icon that looks like a mouse) to open a dialog box in which the values of these parameters can be changed. The cut-off score is determined by the four parameters p1, p2, p3, and p4; ordinarily you would only change these if the query sequence is very short.

The third step, the time-consuming alignment step, is performed only if the initial score exceeds the cut-off score. This full local alignment inserts gaps in the alignment as necessary to improve the score. The score resulting from this step of the analysis is called the optimized score. The optimized score is the score of the best local alignment of the matched region of the library sequence and the query sequence. In the case of perfect identity, the initial and optimized scores will be identical and will be four times the length of the matching region (if you haven't changed any of the scoring matrix parameters with "Tweak" or changed the scoring matrix itself). So, for a 50-base-long perfect match, both scores will be 200.

It is necessary to balance the speed of the procedure (*via* parameter choices) with the accuracy of the result. For example, an incorrect choice of match, mismatch, gap-opening, and gap-extension parameters can produce alignments which favor gaps over mismatches, a generally unlikely occurrence. It may even be possible to create a parameter set which will miss near-optimal, but biologically significant, alignments.

As in the similarity matrix procedures, the hash level is an important parameter. The larger the hash level value, the faster the search, and the less sensitive the search. If the hash level is set to one and the other parameters are handled appropriately, the library search routine becomes a full metric alignment procedure. This is of significant advantage for evolutionary comparisons of genes within a gene cluster. It is the most rigorous search routine available in MacVector.

For your first pass, the best hash setting to use for amino acid sequences is two and for nucleotide sequences is six. As matching regions are identified, it is frequently useful to modify the search parameters through the "Tweak" option or the scoring matrix.

Customizing similarity searches

Tweak

At the beginning of the sequence library search, the program calculates a cut-off score. Matching regions whose scores fall below the cut-off score are not aligned and therefore not saved. The parameters used by MacVector to calculate the score can be changed.

The parameters p1, p2, p3, and p4 should normally be left at their default settings. The parameters that are useful to modify are the deletion penalty (penalty for a single residue insertion or deletion, called an indel) and the gap penalty (penalty for a continuing indel). The higher the value of the deletion penalty, the less frequently the program will introduce an indel in preference to allowing a mismatch. The higher the gap penalty, the more costly it is to extend that indel once it is opened. By decreasing the deletion penalty, you can make it easier to introduce an indel. This can be desirable in AT-rich regions, or up- and downstream control regions of a gene cluster. Similarly, for protein sequences, changing these penalties from their default values can increase the likelihood that insertions and deletions, such as have occurred in the cytochrome family, will be permitted in the sequence comparison.

Scoring matrix

In any search procedure, there are values assigned both for a match and for a mismatch. These values can be found in MacVector's scoring matrix files.

For example, in `DNA database matrix`, the score assigned for a match is +4, while the score assigned to a mismatch is -2. It is six times more costly to introduce an indel (insertion or deletion) than it is to tolerate a mismatch (the default deletion penalty is -12) and twice as costly to extend an existing indel than to tolerate a mismatch (the default gap penalty is -4 per base).

These default settings were assigned based on general experience with DNA sequences. For specific applications, it may be useful to alter the match and mismatch scores. You can edit the existing scoring matrix files, or create new scoring matrix files within MacVector. In setting these values, it is the ratio between the match and mismatch, and between the match and deletion penalty values, that are important, not the actual magnitude of the scores. If you report similarity scores for sequences in a scientific presentation, you should always include the values of all four parameters (match score, mismatch score, deletion penalty, and gap penalty).

What are the circumstances that might lead you to alter a scoring matrix? It is often useful to try different scoring schemes for comparing proteins. The scoring matrices pam250 and pam250S are based on Dayhoff's log-odds scoring matrix, whose values were assigned empirically on the basis of observed amino acid replacements. This matrix has been shown to be more effective than other scoring matrix schemes in verifying distant relationships between homologous proteins (Feng, *et al*. 1985). However, you may wish to assign matrix scores based on other criteria when you search for protein similarities.

The simplest scheme is an identity matrix which scores an exact match as +1 and a mismatch as 0 (the important cysteine-cysteine match is often assigned a score of +2). Another matrix scheme assigns scores based on the genetic code—higher scores are assigned to amino acid pairs whose codons differ by two or three nucleotides. It is also possible to assign scores by taking into account the structural similarities of the amino acid side chains, for example, assigning a higher score for a leucine-isoleucine pair than for a leucine-lycine pair.

In addition to allowing you to change the match / mismatch values themselves, MacVector makes it easy for you to treat groups of amino acids as if they were identical. This is done by assigning them the same hash (or group) number. To see how this works, compare the scoring matrix pam250 with its simplified version, pam250S, shown in Appendix C, *“Reference Tables”*. Notice that these two matrices contain identical match / mismatch values. If you look at the tweak values table, you will see that each non-ambiguous amino acid is assigned a unique hash number in pam250. The matrix pam250S, on the other hand, assigns several different amino acids to the same hash number. All amino acids that share the same hash number are considered to be identical according to the properties of their side chains.

Additional hash schemes are certainly possible. For instance, when searching protein sequences, it may be useful to use Kyte and Doolittle classification parameters. You might develop your own classification for amphipathic helices and NTP binding sites, among others. For nucleic acid sequences in extremely AT- or GC-rich segments, it may well be useful to use a pyrimidine-pyrimidine and purine-purine scoring scheme instead.

Protein scoring

Because of the way MacVector computes scores when matching protein sequences, it is possible for a region to exceed 100 percent match. Another artifact of the scoring method is that a region of the two sequences that is an exact match may have a score of less than 100 percent. How does this arise, and is it something to worry about?

The pam250 scoring matrix assigns scores from 1 to 17 for matches and scores from -1 to -8 for mismatches. The magnitude of the score reflects the likelihood that an amino acid in one sequence could have been replaced by the corresponding amino acid in the other sequence. For example, cysteines tend to be highly conserved, so a C-C match between two sequences gets the score 12. A replacement of a cysteine by a serine (or vice versa) is less likely, so a C-S pairing gets the score 0. The likelihood that a cysteine will be replaced by a tryptophan (or vice versa) by mutation is fairly low, so this pairing gets the score -8.

Before comparing the two sequences, MacVector computes the total score (using the scoring matrix) that would occur if the sequence on the X-axis were compared with itself. The total score is then divided by the number of amino acids in the sequence to yield the average contribution per amino acid to the total score. The amino acid score is multiplied by the window size to give the theoretical maximum score for a 100 percent match for a window of that size. This theoretical maximum score may be less than the score attained by a small region of the sequence, or it may exceed the score attained by a perfect match in a small region of the sequence. An example may clarify matters:

The peptide:

```
G D V E K G K K I F I M K  C  S Q  C  H T V
```

has scores of:

```
5 4 4 4 5 5 5 5 5 9 5 6 5  12 2 4  12 6 3 4
```

when it is compared with itself. The total score is 110, or an average of 5.5 per residue, which MacVector truncates to 5. If the window size is 8, the theoretical score for 100 percent match is 5*8, or 40. Let's now compare the sequence with itself window-by-window as the program does. The score for the first 8-residue window is 37. This becomes 37 / 40 = 0.93, so the algorithm has calculated a 93 percent match, even though the match is perfect. The score for the second window is also 37 (0.93), the score for the third window is 42 (42 / 40 = 1.05), and the score for the fourth window is 43 (43 / 40 =1.08).

Why don't we calculate the maximum score per window on-the-fly, using the scores for the amino acids that are actually present instead of the theoretical maximum score? This doesn't solve the problem, since the maximum score (and hence the percent match) for a window could depend on which of the two sequences was used to calculate the maximum. If the X-axis sequence within a window is DIMCYWRH and the corresponding Y-axis sequence within the window is EVICFFKQ, the maximum score would be either 66 or 52, depending on which sequence was compared with itself to calculate the maximum score. The score obtained by comparing the two windows with each other using pam250 is 34, so the match would be either 52 percent or 65 percent, depending on which maximum score was used. In addition, this method is slower because of the added computation.

The important thing to remember is that a matrix analysis is really a qualitative analysis method (even though our assigning letters every two percentage points makes it look very quantitative). Using the line display mode, you can easily locate regions of two sequences that are similar. By switching to the character display mode, you can instantly get a rough idea of the significance of matches that all look the same in the line display mode: a diagonal consisting of mostly Cs and Ds obviously represents a better match than one consisting of mostly Zs and As.

Finding subsequences and motifs

DNA and protein subsequence analyses allow you to locate small consensus sequences or motifs within a sequence. The search sequences are stored in a special subsequence file, see *“Subsequence files”* on page 63.

MacVector gives you a great deal of flexibility on how to describe the motif. Many of the subsequences you want to search for are simple, consisting of a single short motif. You can specify that you will allow a certain number of mismatches, and also indicate which residues in the motif must match exactly.

MacVector also allows you to search for motifs that are more complex. If the sequence has more than one part, you have two choices on how the parts should be combined: you can use AND logic or OR logic. Combining parts using OR logic means that only one of the specified parts needs to be present for a match to occur (either Part A OR Part B must be present). Combining parts using AND logic means that all parts must be present in order for a match to occur (both Part A AND Part B must be present).

An example using OR logic: site-directed mutagenesis

Site-directed mutagenesis is a powerful technique for introducing a mutation into a DNA sequence at a single known location. A synthetic oligonucleotide is made whose sequence matches the sequence of a region of the parent or wild type DNA except for a specific nucleotide. This oligo is then substituted into the parent DNA. To make sure this technique is in fact site-specific, you need to verify that the oligonucleotide will not match regions of the parent DNA other than the one specific region where you want to insert the mutation. You can do this by performing a subsequence analysis of the parent DNA using the oligonucleotide as the subsequence. If your oligo is truly site-specific, the program will report only one match.

You could use the whole oligo as a single simple subsequence, just as we did in the section on simple subsequences analyses. But a problem may arise if your oligo is very long.

Suppose, for example, that your oligo is 21 bases long, with the base to be mutated in the center. It might be possible for a short region at one end of the oligo (about seven bases or so) to match several sites in the DNA, which means your mutation would be inserted in more than one place. If you searched for the whole oligo as a single subsequence, the non-matching portion of the oligo would score more heavily than the small amount of match at the end, and the program would not report the extra matches, lulling you into a false sense of security as to the specificity of your oligo.

A way to deal with this situation is to analyze long oligos in parts. For this example, we could enter bases 1 through 10 and bases 12 through 21 into a subsequence file as two separate subsequences with separate names. Or we could enter them into the subsequence file as two parts of one subsequence, using OR logic.

An example using AND logic: cell division motif

A cell division motif is shared by the SV40 large-T antigen, adenovirus E1A protein, papilloma virus E7 protein, the v- and c-myc oncogene products, and the CDC25 gene product (yeast mitotic regulator). This motif has two parts which are separated by a gap which varies in size from one to eight amino acids:

Part 1. (ND)XXCX(STE) (beta turn)
Part 2. (DE)(DEST)(DE)XXX (alpha helix)

By entering these into the subsequence file as two parts of a single subsequence using AND logic and setting the gap range from 1 to 8, we can direct the program to look for this motif.

Phylogenetic analysis

The main method that molecular biologists use to elucidate gene function is to perform experiments with cloned DNA sequences. However, comparative analysis is an increasingly important alternative. For example, one of the first tasks of a researcher, when characterizing a newly cloned sequence, is to perform a BLAST search against a reference library such as one of the NCBI databases. If there are significant matches between query and library sequences, this may indicate functional similarities. Such findings often provide the starting point for further studies of the cloned sequence.

MacVector provides a suite of methods for undertaking sequence comparisons, including Pustell matrix analysis, DNA and protein library searches, and subsequence analysis. For the most part these methods are concerned with identifying regions of similarity between pairs of sequences. None of the methods make any assumptions about the underlying mechanisms by which sequences diverge from one another during evolution.

Phylogenetic analysis provides an alternative approach to comparative sequence analysis. Instead of concentrating on the relationships between a sequence and one or more reference sequences, phylogenetic analysis provides a means of simultaneously describing the relationships between all sequences in a given sample. Phylogenetic methods allow the researcher to look for patterns of similarity within an alignment, without having to restrict attention to a specific sequence. At the same time, phylogenetics provides a rigorous framework in which researchers can make quantitative statements about sequence similarities, and make better informed judgements as to whether a given match reflects true homology rather than chance similarity. Li (1997) and Swofford et al. (1996) provide overviews of phylogenetic analysis.

Methods for calculating phylogenies

Molecular phylogenetics offers a diverse array of methods for analyzing sequence alignments (Li 1997). A common way to categorize these approaches is to distinguish three groups of methods:

- *Distance matrix* methods are derived from statistical methods for performing cluster analysis. A pairwise distance matrix is constructed for all sequences in the alignment, and this is used to reconstruct a tree.

- *Parsimony* methods use a hypothetico-deductive approach to analyze sequence alignments. The alignment positions are considered one at a time, and the tree that requires the least number of changes is chosen.
- *Maximum likelihood* methods require an explicit model of sequence evolution. By applying the statistical approach of maximum likelihood to a quantitative model and a data set, the optimal tree can be identified for a given set of conditions.

The choice of method depends to some extent on what kind of sequence alignment is being analyzed. Most sequence alignments that show a consistent pattern of variation can be successfully analyzed using any method. In contrast, alignments that contain very diverged sequences may require particular methods of analysis. In choosing which method to use, researchers must usually make a trade-off between reliability and speed of computation.

MacVector offers two methods, both of them distance matrix approaches. Distance matrix methods are among the fastest reconstruction methods, and they are known to be reliable for a variety of sequence alignments. There are two steps to the process:

1. Estimating the evolutionary distance between every pair of sequences in the alignment
2. Reconstructing the phylogeny, using clustering methods to add the sequences one by one.

MacVector provides a variety of metrics for estimating the evolutionary distance between sequences, and two methods of phylogenetic reconstruction.

Estimating evolutionary distance

To understand distance methods it is necessary is to appreciate the underlying evolutionary process. If we compare a homologous nucleotide position in two DNA sequences there are two possibilities: identity or non-identity. If the two sequences are very closely related, it may be reasonable to assume that identity indicates that no substitutions have occurred since the two evolutionary lineages diverged. However, if the two sequences are less closely related, and have diverged over longer periods of time, it becomes increasingly likely that recent evolutionary events have masked earlier events.

For example, if we see an A at the same position in two sequences, it is possible that in one lineage A occurred in the ancestral sequence, but was then replaced by G, which was again replaced by A at some later time. Likewise, if the two sequences show a difference at a given position, this does not give us enough information to reconstruct the full series of evolutionary events since the two lineages diverged. A consequence of this is that calculations of evolutionary distance between very diverged sequences will tend to underestimate the true amount of evolution that has occurred.

Evolutionary biologists attempt to correct for unseen evolutionary events, using models of nucleotide and amino acid substitution. Nucleotide substitution models are generally the more sophisticated, largely because the process of DNA evolution is much simpler to describe than protein evolution. Consequently, a greater variety of methods for estimating DNA distances have been developed, and this is reflected in the choice of distances provided in MacVector.

Invalid distances

In some cases no valid distance can be calculated for a pair of sequences, because the selected algorithm generates an illegal operation, e.g. division by zero, or the logarithm of a negative number. This is likely to happen if the sequences are highly divergent, or if they have not been properly aligned.

MacVector warns the user when any invalid distances are found. If you choose to continue with the analysis, MacVector will use the largest value calculated in the matrix to substitute for any failed calculations - interpret with care!

Nucleotide distances

Absolute number of differences

This is calculated as the total number of nucleotide differences between two DNA sequences. This method should generally be avoided as the basis for phylogenetic reconstruction, but may occasionally be useful if the user wishes to visualize the total number of differences between sequences in an alignment.

p-distance

The *p*-distance is calculated as the proportion of differences between two sequences. In general, if the sequences being analyzed have *p*-distances less than 0.1, this method is suitable as the basis for phylogenetic reconstruction. However, when sequences have significantly diverged from one another, the *p*-distance is likely to underestimate the true amount of evolutionary divergence and produce misleading results.

Jukes-Cantor

Jukes & Cantor's (1969) model of nucleotide substitution makes the assumption that the rate of nucleotide substitution is the same between all pairs of nucleotides, so that *transitions* (e.g. purine replaced by purine) and *transversions* (e.g. purine replaced by pyrimidine) are equally likely. For moderately diverged sequences this method provides more robust estimates than the *p*-distance, but when sequences are more diverged it may be misleading.

Kimura 2-Parameter

The method of Kimura (1980) does not assume that transition and transversion rates are equal. Since an excess of transitional substitutions is a general feature of diverging DNA sequences, the Kimura 2-parameter distance is often more robust than the Jukes-Cantor method.

Tajima-Nei

The method of Tajima & Nei (1984) is appropriate when the base composition of nucleotides is unequal. Many sequences deviate from nucleotide frequencies of 0.25, and the Tajima-Nei method corrects for this bias. However, this method should be used with caution, since it assumes that transition and transversion rates are equal.

Tamura-Nei

The method of Tamura & Nei (1993) is the most general of the methods provided by MacVector. This method corrects not only for differences between transitions and transversions, but also for the different types of transitions (i.e. A <> G, and C <> T). This method is particularly appropriate for very diverged sequences with biased base compositions.

LogDet

The LogDet method for estimating nucleotide distance is useful in sequences where there are extreme differences in the base compositions of sequences in an analysis (Lockhart *et al.*, 1994). For example, in some bacterial genomes G+C contents may be close to 90%. If we include sequences from G+C rich genomes in an analysis, they are likely to be grouped in subsequent reconstruction, even if they are actually quite unrelated. The LogDet method is very powerful for resolving such situations, but is not appropriate as a general method.

Protein distances

Methods for estimating distances between amino acid sequences are currently not as sophisticated as those for nucleotide sequences. As a result MacVector offers only three methods:

Absolute number of differences

This is calculated as the total number of amino acid differences between two protein sequences.

p-distance

The *p*-distance is calculated as the proportion of differences between two sequences. In general, if the sequences being analyzed have *p*-distances less than 0.1, this method is suitable as the basis for phylogenetic reconstruction. However, when sequences have significantly diverged from one another, the *p*-distance is likely to underestimate the true amount of evolutionary divergence.

Poisson-correction

This method assumes that the number of substitutions occurring at each site follows a Poisson distribution. This method may be the most appropriate method to use when protein sequences are significantly diverged (e.g., average *p*-distance > 0.1).

Phylogenetic reconstruction methods

MacVector provides two distance matrix methods for reconstructing phylogenies, *neighbor joining* and *UPGMA* (Unweighted Pair-Group Method with Arithmetic mean). Both methods are similar in their implementation and are defined by well specified clustering algorithms. Both proceed by scanning the pairwise distance matrix and identifying the two most similar sequences. These two sequences are joined in the tree, and the underlying matrix is then recalculated, treating the newly joined sequences as a new phylogenetic unit. The process continues until all sequences have been added to the tree. Neighbor joining and UPGMA differ in the way that the matrix is transformed as each sequence is added. The important points to note are as follows:

- The UPGMA method is useful if we assume that the rate of evolution is constant in different lineages. If, in some group of organisms, a particular protein attains greater functional significance and starts to evolve at a faster rate, that assumption will be violated. This is believed to be a common situation for many genes. In such cases the UPGMA method is likely to generate misleading results.

- Neighbor-joining (Saitou & Nei 1987) makes no assumption about constant evolutionary rates, so the method is more robust than UPGMA.
- UPGMA generates rooted trees. This gives some indication of the polarity of evolution in a phylogeny. By contrast, neighbor-joining generates an unrooted tree, and the only way of making inferences about the direction of change is to use one of the supplied rooting methods.

For details of the algorithms, see Swofford et al. (1996). This article also describes several more sophisticated phylogenetic methods. However, when the appropriate distance and phylogenetic reconstruction methods are used in combination, MacVector will provide robust estimates of phylogeny.

Analyzing phylogenies

In addition to providing methods for generating phylogenetic trees, MacVector also provides methods for analyzing them. One approach is to use the set of tools provided in the Tree Viewer window. These enable users to format trees in different ways, to investigate the effect of rooting the tree using different methods, and to examine the effect of excluding sequences from an analysis (see *"Editing Trees"* on page 367). A second approach is to test the reliability of the phylogeny using bootstrap analysis (Felsenstein 1985).

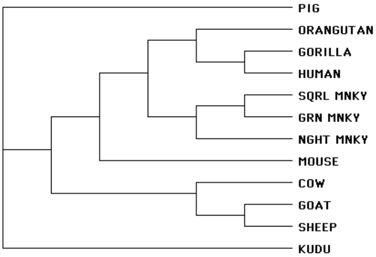
Bootstrapping

If there is variation among the sequences in an alignment, the methods provided by MacVector are guaranteed to provide a single best tree. However, in many circumstances the resulting phylogeny may be meaningless. For example, if we generate an alignment of completely unrelated DNA sequences, the relationships suggested by the phylogeny are likely to be positively misleading. The question of how much confidence can be associated with a phylogeny is a long-standing problem in phylogenetics which has attracted much attention.

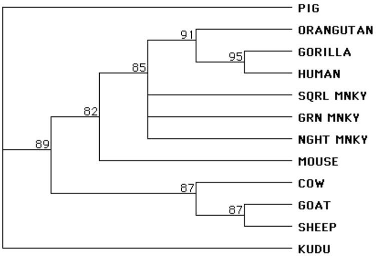
The most general method available for assessing the confidence associated with a phylogeny is *bootstrapping*. This is one of a well known group of statistical techniques called resampling methods. The essence of the method is to generate a series of random subsamples of the data set and repeat the phylogenetic reconstruction for each subsample, recording how often a particular grouping in the tree is recovered. Ideally, the same tree will always be recovered from the data set, no matter what random sample is used for the reconstruction. If, however, the data set contains no consistent phylogenetic signal, completely different trees could be recovered each time a different sample of nucleotide positions was analyzed.

To summarize the results of a bootstrap analysis, a *consensus bootstrap tree* is made of the phylogenies reconstructed from the various bootstrap data samples. The consensus bootstrap tree summarizes the shared features of the set of resampling trees. Branching points (internal nodes) are retained only if they occur in a minimum percentage of resampling trees; all other nodes are collapsed.

If the phylogeny is strongly supported, the consensus bootstrap tree will include the majority of nodes resolved as bifurcations, like this:

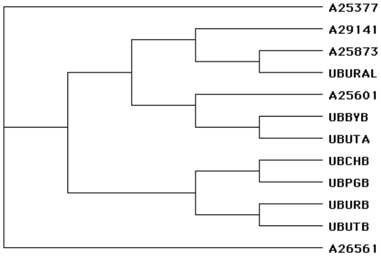


Best tree

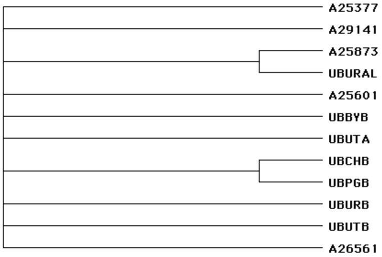


Bootstrap tree

If there is little or no resolution, the tree will show a small number of multifurcations, with multiple branches originating from the same ancestral nodes, like this:



Best tree



Bootstrap tree

Part four

Appendices

This part contains additional information about MacVector.

A

Setting up NCBI's *Entrez* and BLAST Services

Overview

This appendix contains details of how to access the *Entrez* Internet database and the BLAST search service.

NCBI’s *Entrez* database

The *Entrez* database, produced by the National Center for Biotechnology Information (NCBI), provides access to DNA and protein sequences and related bibliographic information. The sequence data include the complete nucleotide and protein sequence data from the Genbank, EMBL, DDBJ, PIR, PRF, PDB, SWISS-PROT, dbEST and dbSTS databases, as well as data from U.S. and European patents. *Entrez* also contains a subset of the PUBMED database, including references and abstracts which are cited in the sequence databases and other related PUBMED records. The data are updated on a daily basis.

NCBI’s BLAST search service

The NCBI BLAST search system provides a way of searching all the *Entrez* sequence databases by similarity. The Internet configuration for accessing BLAST is the same as for other *Entrez* searches.

Accessing the NCBI services

MacVector can directly browse the *Entrez* molecular sequence database over the Internet, and search it using BLAST. With MacVector, databases residing at NCBI are accessible by users at remote locations over the Internet.

MacVector has been designed to make accessing *Entrez* and BLAST as effortless as possible. You do not need to be an Internet expert to use MacVector. After an Internet connection is established, MacVector will handle all the communications with NCBI’s databases.

To set up your MacVector package to access *Entrez* and BLAST, you can use either a permanent network connection or a dial-up connection. Your Internet connection must be of the kind which can support a Web browser.

Firewalls

MacVector connects through http, using the same protocol as a web-browser for both Entrez and BLAST. MacVector posts requests directly to the NCBI home page using the http protocol on port 80. The data that is sent back is identical to the NCBI web page that you would get if you were using a browser. Embedded in the pages are non-displayed XML which is the "machine readable" data that we parse out using tools provided by the NCBI.

So, if you can connect with a browser to the NCBI, you should be able to connect with MacVector. If a proxy server setting is required then this will need to be done for ALL internet applications on OSX, and is done in the Network control panel for the entire machine and all applications automatically take advantage of those settings. For example you can see this in Safari - if you click on the advanced proxy settings, it opens up the system network preferences.

B

Using the Job Manager

Overview

This appendix describes how to use the Job Manager to control background jobs started by MacVector.

About the Job Manager

The MacVector Job Manager was introduced with MacVector 8.0 as a way to allow NCBI BLAST searches to be run in the background so that the user can continue to work within MacVector while waiting for the search to complete. The Job Manager also allows multiple BLAST searches to be submitted without waiting for the first job to complete, and allows submitted searches to be cancelled through a simple list-based interface.

With MacVector 9.5 this functionality was extended to include ClustalW alignment jobs. Future versions will see additional long-running analyses migrated to the control of the Job Manager. Additionally, all Assembler analysis functions are submitted as jobs to the Job Manager. The `phred` base calling algorithm and the `cross_match` vector masking algorithm are automatically split across as many cores/CPUs as the target machine possesses.

When all of the cores/CPUs are busy, the Job Manager queues additional jobs and starts them automatically when a core/CPU becomes available. You can monitor the status of all currently submitted, running and queued jobs using the Job Manager window, invoked by selecting the **Windows | Show Job Manager** menu item.

Growl Support in the Job Manager

Growl is a third party, open notification framework for OS X. It allows the basic notification system used by OS X (i.e. the bouncing icon) to be replaced by a much more flexible and arguably less intrusive system. It is widely supported by many OS X applications

From MacVector 11.1 onwards, support for Growl has been added to the Job Manager.

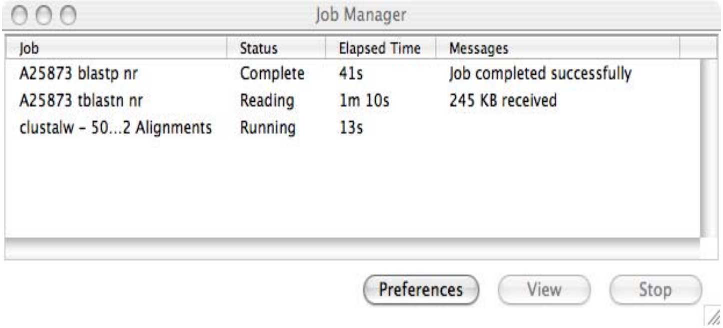
As a result, if you have Growl installed on your machine, now when a job that is running in the Job Manger finishes, you are notified in a user configurable and non-intrusive way.

Click on the dialog to display the Job Manager from which you can access the results. Alternatively, ignore the dialog and it will fade away.

Tip. When a running job finishes, even after the Growl notification dialog has faded away, the MacVector dock icon is tiled with a numeric badge. The number indicates the number of finished jobs that are available to view in the Job Manager. Open the Job Manager to find out more about these jobs.

Using the Job Manager

You can open the Job Manager window at any time by either choosing **Windows | Show Job Manager**, or pressing **< ⌘ > - B**



Job	Status	Elapsed Time	Messages
A25873 blastp nr	Complete	41s	Job completed successfully
A25873 tblastn nr	Reading	1m 10s	245 KB received
clustalw - 50...2 Alignments	Running	13s	

View

This button is only enabled if a job has completed, but the results have not yet been displayed. Currently this only applies to ClustalW and Blast jobs that completed at a time when the user was busy with other windows. If you start either of these jobs, dismiss the progress dialog and then switch to another window, the results are not automatically displayed when the job completes. This is to prevent a result dialog from being displayed and distracting you while you are busy with some other task. When you are ready, you can see the completed job results by selecting the job in the Job Manager window and then clicking **View**. Usually, `phred`, `cross_match` and `phrap` jobs are always removed from the Job Manager window when they complete because they do not need to display a completion dialog. In those cases, the Assembly Project window is simply updated with the results.

Note. If a job fails for any reason, it will remain in the Job Manager list with a status of **Failed**. You can then view the job to learn more about the failure

Stop

By highlighting a running job and then clicking the **Stop** button, the job will be stopped and removed from the Job Manager window.

Preferences

With some older multiple CPU machines, stability problems have occasionally been seen when MacVector uses both CPUs. If you believe you are seeing problems that might be related to this, you can disable the use of more than one CPU using the **Preferences** button. Please note that these problems have not been seen on newer Intel multi core machines.

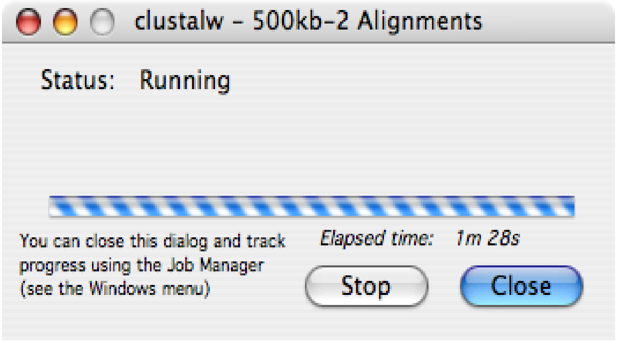
Status

The Status column can show a number of words to indicate the state of a job:

- **Complete** indicates a job that has finished and can show some results.
- **Running** indicates active jobs
- **Queued** indicates a job has been submitted and is waiting for a free processor.
- **Reading** is specific to Blasts jobs and indicates that the process is receiving data (results) from the NCBI servers.
- **Failed** indicates a job failed to complete for some reason. Additional messages will typically be displayed giving an explanation of the problem.

Job progress dialog

When a job is first submitted to the Job Manager, a dialog appears that indicates the status of the job. This dialog can be closed at any time by selecting the **Close** button. This does not stop the job and you can continue to monitor the status of the job from the main Job Manager window. You can also terminate the job at any time by selecting **Stop**. This dialog does not appear for Assembler related jobs.



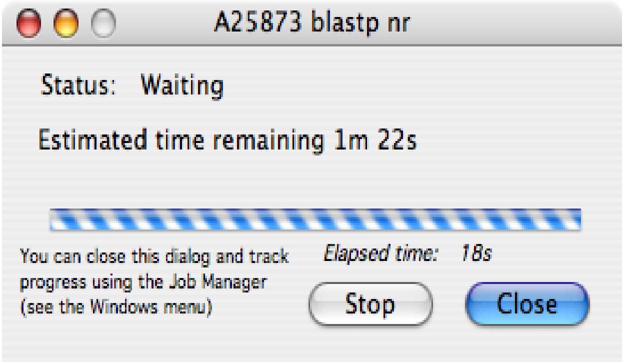
clustalw - 500kb-2 Alignments

Status: Running

You can close this dialog and track progress using the Job Manager (see the Windows menu)

Elapsed time: 1m 28s

With BLAST jobs, information related to the estimated completion time is also displayed.



A25873 blastp nr

Status: Waiting

Estimated time remaining 1m 22s

You can close this dialog and track progress using the Job Manager (see the Windows menu)

Elapsed time: 18s

C

Reference Tables

Overview

This appendix contains reference tables for the matrix analysis methods, including letter codes and scoring matrix values.

IUPAC-IUB codes for nucleotides and amino acids

Nucleic Acids

A	Adenine	
C	Cytosine	
G	Guanine	
T	Thymine	
U	Uracil	
R	puRine	(A or G)
Y	pYrimidine	(C or T/U)
K	Keto	(G or T/U)
M	aMino	(A or C)
S	Strong	(C or G)
W	Weak	(A or T)
B	not A	(C or G or T/U)
D	not C	(A or G or T/U)
H	not G	(A or C or T/U)
V	not T/U	(A or C or G)
N	aNy	(A or C or G or T/U)

Amino acids

A	Ala	Alanine
B	Asx	Asparagine or aspartic acid
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine

H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
U[a]	Sec	Selenocysteine
V	Val	Valine
W	Trp	Tryptophan
X	Xxx	Any amino acid, unknown
Y	Tyr	Tyrosine
Z	Glx	Glutamine or glutamic acid

[a] Proposed symbol for selenocysteine. You cannot use it in a MacVector sequence. When files are imported, MacVector converts U residues to X.

Letter codes for matrix analyses

Uppercase percent		**Lowercase percent**	
A	100	a	50
B	98	b	48
C	96	c	46
D	94	d	44
E	92	e	42
F	90	f	40
G	88	g	38
H	86	h	36
I	84	i	34
J	82	j	32
K	80	k	30
L	78	l	28
M	76	m	26
N	74	n	24
O	72	o	22
P	70	p	20
Q	68	q	18
R	66	r	16
S	64	s	14
T	62	t	12
U	60	u	10
V	58	v	8
W	56	w	6
X	54	x	4
Y	52	y	2
Z	50	z	0

@>100 percent (protein matrix)

Hash codes and “tweak” values for scoring matrices supplied with MacVector

pam250 protein scoring matrix

hash codes										tweak values	
--	19	Phe	5	Leu	10	Arg	15	Tyr	20	p1	27
Ala	1	Gky	6	Met	11	Ser	16	Asx	3	p2	200
Cys	2	His	7	Asn	12	Thr	17	Gkx	4	p3	50
Asp	3	Ile	8	Pro	13	Val	18	Xxx	19	p4	50
Glu	4	Lys	9	Gln	14	Trp	19	End	19		

deletion penalty (single residue indel) = 12

gap penalty (continuing indel, per residue) = 6

pam250S scoring matrix

hash codes										tweak values	
--	4	Phe	4	Leu	3	Arg	2	Tyr	4	p1	27
Ala	0	Gky	0	Met	3	Ser	0	Asx	1	p2	200
Cys	5	His	2	Asn	1	Thr	0	Gkx	1	p3	50
Asp	1	Ile	3	Pro	0	Val	3	Xxx	4	p4	50
Glu	1	Lys	2	Gln	1	Trp	4	End	4		

deletion penalty (single residue indel) = 12

gap penalty (continuing indel, per residue) = 6

DNA matrix nucleic acid scoring matrix

hash codes								tweak values	
--	1	G	2	T	3	K	2	p1	50
A	0	R	0	W	3	D	3	p2	80
C	1	S	2	Y	1	B	1	p3	5
M	0	V	2	H	3	N	0	p4	80

deletion penalty (single residue indel) = 12

gap penalty (continuing indel, per residue) = 4

DNA database matrix nucleic acid scoring matrix: match / mismatch scores

-	0															
A	0	4														
C	0	-2	4													
M	0	4	4	4												
G	0	-2	-2	-2	4											
R	0	4	-2	4	4	4										
S	0	-2	4	4	4	4	4									
V	0	4	4	4	4	4	4	4								
T	0	-2	-2	-2	-2	-2	-2	-2	4							
W	0	4	-2	4	-2	4	-2	4	4	4						
Y	0	-2	4	4	-2	-2	4	4	4	4	4					
H	0	4	4	4	-2	4	4	4	4	4	4	4				
K	0	-2	-2	-2	4	4	4	4	4	4	4	4	4			
D	0	4	-2	4	4	4	4	4	4	4	4	4	4	4		
B	0	-2	4	4	4	4	4	4	4	4	4	4	4	4	4	
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-	A	C	M	G	R	S	V	T	W	Y	H	K	D	B	N

pam250 and pam250S protein matrix: match / mismatch scores

-	0																								
A	0	2																							
C	0	-2	12																						
D	0	0	-5	4																					
E	0	0	-5	3	4																				
F	0	-4	-4	6	-5	9																			
G	0	1	-3	1	0	-5	5																		
H	0	-1	-3	1	1	-2	-2	6																	
I	0	-1	-2	-2	-2	1	-3	-2	5																
K	0	-1	-5	0	0	-5	-2	0	-2	5															
L	0	-2	-6	-4	-3	2	-4	-2	2	-3	6														
M	0	-1	-5	-3	-2	0	-3	-2	2	0	4	6													
N	0	0	-4	2	1	-4	0	2	-2	1	-3	-2	2												
P	0	1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6											
Q	0	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4										
R	0	-2	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6									
S	0	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2								
T	0	1	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3							
V	0	0	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4						
W	0	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17					
Y	0	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10				
B	0	0	-4	3	2	-5	0	1	-2	0	-3	-2	2	-1	1	0	0	0	-2	-5	-3	2			
Z	0	0	-5	2	3	-5	0	2	-2	0	-2	-1	1	0	3	0	0	0	-2	-6	-4	1	3		
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	B	Z	X	*

Residue-specific gap modification factors for ClustalW sequence alignment

Residue	Mod. Factor	Residue	Mod. Factor
A	1.13	M	1.29
C	1.13	N	0.63
D	0.96	P	0.74
E	1.31	Q	1.07
F	1.20	R	0.72
G	0.61	S	0.76
H	1.00	T	0.89
I	1.32	V	1.25
K	0.96	Y	1.00
L	1.21	W	1.23

D Formatting Examples

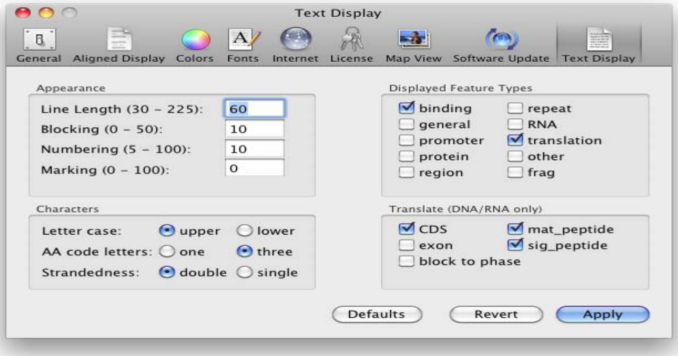
Overview

This Appendix provides some examples of formatting the MacVector output display windows.

Formatting the Text view

Example 1

To display restriction enzyme site results with translated protein coding (CDS) and binding sites from the features table, use the settings shown below from **Options | Format Annotated Display.** The Show panel labels the binding sites and translations from the features table. The Translate panel automatically translates each CDS feature and shows the sequence blocked to the phase of the translation.



Appearance

Line Length (30 – 225): 60
Blocking (0 – 50): 10
Numbering (5 – 100): 10
Marking (0 – 100): 0

Displayed Feature Types

☑ binding □ repeat
□ general □ RNA
□ promoter ☑ translation
□ protein □ other
□ region □ frag

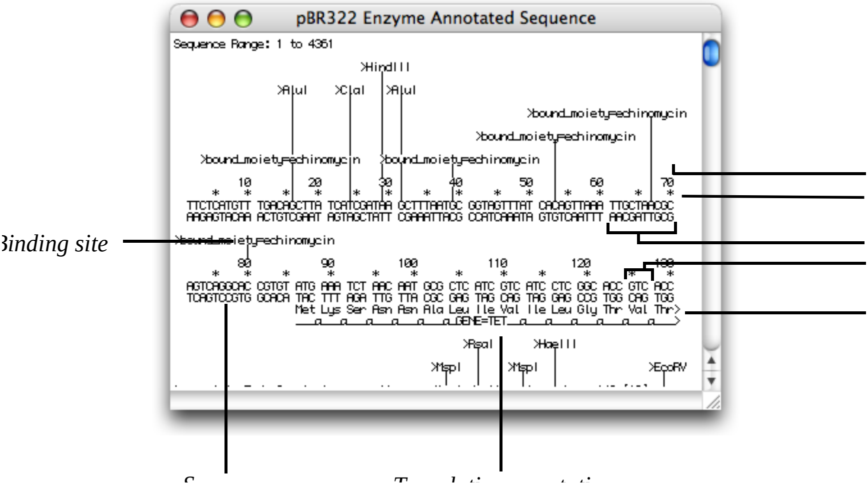
Characters

Letter case: ● upper ○ lower
AA code letters: ○ one ● three
Strandedness: ● double ○ single

Translate (DNA/RNA only)

☑ CDS ☑ mat_peptide
□ exon ☑ sig_peptide
□ block to phase

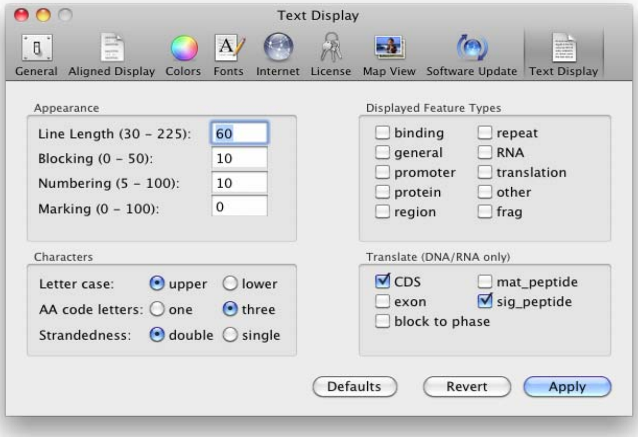
Resulting Output



```
Sequence Range: 1 to 4361

                    >HindIII
          >AluI   >ClaI  >AluI
                                        >bound_moiety=echinomycin
                                   >bound_moiety=echinomycin
  >bound_moiety=echinomycin  >bound_moiety=echinomycin
         10         20         30         40         50         60         70
     *    *    *    *    *    *    *    *    *    *    *    *    *    *
TTCTCATGTT TGACAGCTTA TCATCGATAA GCTTTAATGC GGTAGTTTAT CACAGTTAAA TTGCTAACGC
AAGAGTACAA ACTGTCGAAT AGTAGCTATT CGAAATTACG CCATCAAATA GTGTCAATTT AACGATTGCG
>bound_moiety=echinomycin
         80         90        100        110        120        130
     *    *    *    *    *    *    *    *    *    *    *    *
AGTCAGGCAC CGTGT ATG AAA TCT AAC AAT GCG CTC ATC GTC ATC CTC GGC ACC GTC ACC
TCAGTCCGTG GCACA TAC TTT AGA TTG TTA CGC GAG TAG CAG TAG GAG CCG TGG CAG TGG
                 Met Lys Ser Asn Asn Ala Leu Ile Val Ile Leu Gly Thr Val Thr>
                 ___a___a___a___a___a___a_GENE=TET_a___a___a___a___a___a___>

                                    >RsaI      >HaeIII
                                >MspI      >MspI               >EcoRV
```

Example 2

To display restriction enzyme site results with only a translated protein, use the settings shown below from **Options | Format Annotated Display.** The Translate panel automatically translates each pept entered on the features table and the protein is not labeled because the checkbox for pept in the Show panel is not chosen.



Appearance

Line Length (30 – 225): 60

Blocking (0 – 50): 10

Numbering (5 – 100): 10

Marking (0 – 100): 0

Displayed Feature Types

□ binding □ repeat

□ general □ RNA

□ promoter □ translation

□ protein □ other

□ region □ frag

Characters

Letter case: ● upper ○ lower

AA code letters: ○ one ● three

Strandedness: ● double ○ single

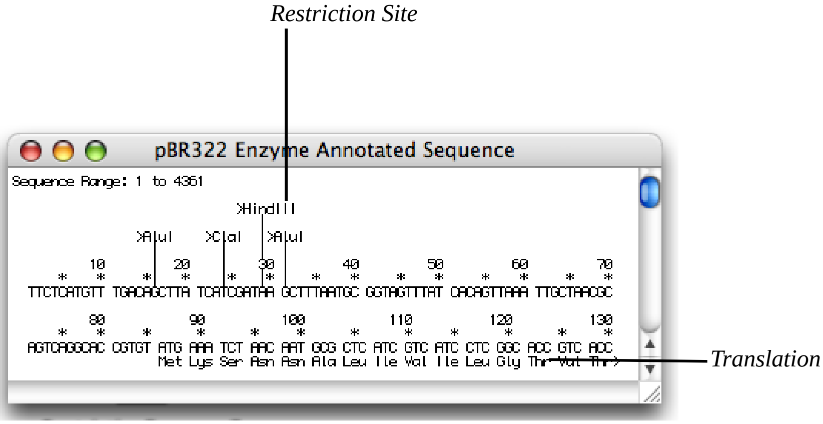
Translate (DNA/RNA only)

☑ CDS □ mat_peptide

□ exon ☑ sig_peptide

□ block to phase

Resulting output

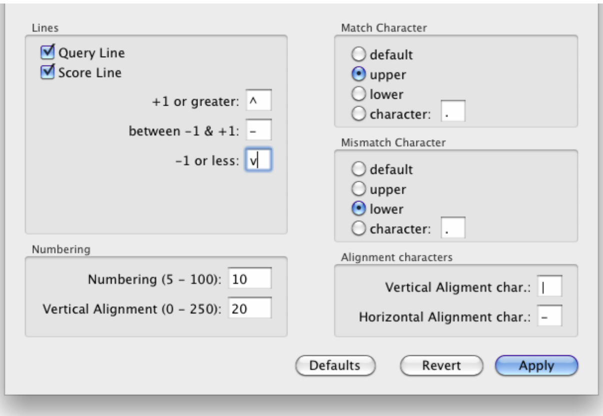


```
Sequence Range: 1 to 4361

                             >HindIII

                >AluI    >ClaI   >AluI

          10         20         30         40         50         60         70
      *    *     *    *     *    *     *    *     *    *     *    *     *    *
  TTCTCATGTT TGACAGCTTA TCATCGATAA GCTTTAATGC GGTAGTTTAT CACAGTTAAA TTGCTAACGC

          80           90          100           110          120          130
      *    *     *      *      *     *      *      *     *      *      *     *
  AGTCAGGCAC CGTGT ATG AAA TCT AAC AAT GCG CTC ATC GTC ATC CTC GGC ACC GTC ACC
                   Met Lys Ser Asn Asn Ala Leu Ile Val Ile Leu Gly Thr Val Thr>
```

Formatting the aligned display

Example 1

This aligned display has the default settings from **Options | Format Aligned Display.** The settings for the query sequence, shown double-stranded, are set on **Options | Format Annotated Display.**



Lines
- ☑ Query Line
- ☑ Score Line

+1 or greater: ^
between -1 & +1: -
-1 or less: v

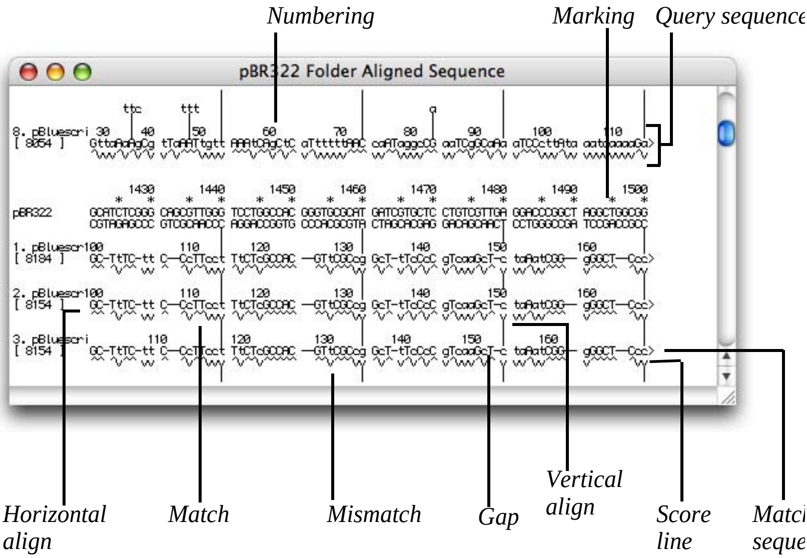
Numbering
Numbering (5 - 100): 10
Vertical Alignment (0 - 250): 20

Match Character
- ○ default
- ● upper
- ○ lower
- ○ character: .

Mismatch Character
- ○ default
- ○ upper
- ● lower
- ○ character: .

Alignment characters
Vertical Aligment char.: |
Horizontal Alignment char.: -

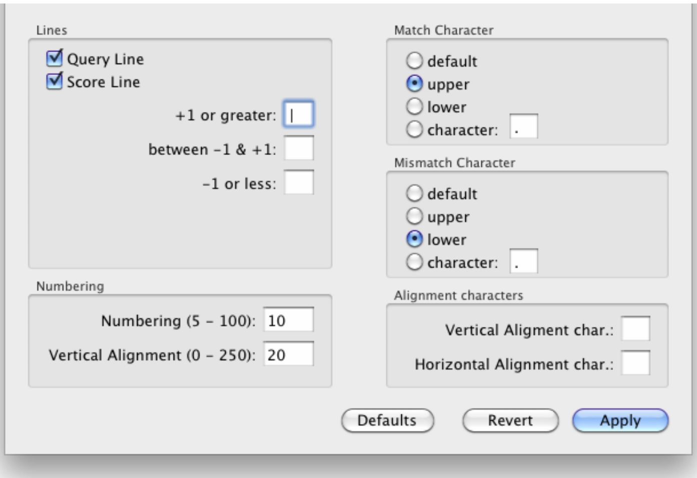
Resulting output



```
                 ttc       ttt                                 a
8. pBluescri 30       40        50          60          70          80          90         100         110
[ 8054 ]     GttaAaAgCg tTaAATtgtt AAAtCAgCtC aTtttttAAC caATaggcCG aaTCgGCaAa aTCCcttAta aatgaaaaGa>

                  1430        1440        1450        1460        1470        1480        1490        1500
                *     *     *     *     *     *     *     *     *     *     *     *     *     *     *     *
pBR322       GCATCTCGGG CAGCGTTGGG TCCTGGCCAC GGGTGCGCAT GATCGTGCTC CTGTCGTTGA GGACCCGGCT AGGCTGGCGG
             CGTAGAGCCC GTCGCAACCC AGGACCGGTG CCCACGCGTA CTAGCACGAG GACAGCAACT CCTGGGCCGA TCCGACCGCC

1. pBluescr100              110         120           130         140         150          160
[ 8184 ]     GC-TtTC-tt C—CcTTcct TtCTcGCCAC —GTtCGCcg GcT-tTcCcC gTcaaGcT-c taAatCGG— gGGCT—Ccc>

2. pBluescr100              110         120           130         140         150          160
[ 8154 ]     GC-TtTC-tt C—CcTTcct TtCTcGCCAC —GTtCGCcg GcT-tTcCcC gTcaaGcT-c taAatCGG— gGGCT—Ccc>

3. pBluescri            110         120         130          140         150          160
[ 8154 ]     GC-TtTC-tt C—CcTTcct TtCTcGCCAC —GTtCGCcg GcT-tTcCcC gTcaaGcT-c taAatCGG— gGGCT—Ccc>
```

Example 2

In this display, the score line is a "|", the Query Line is turned on, and the horizontal and vertical characters are turned off.



Lines
- ☑ Query Line
- ☑ Score Line
- +1 or greater: |
- between -1 & +1:
- -1 or less:

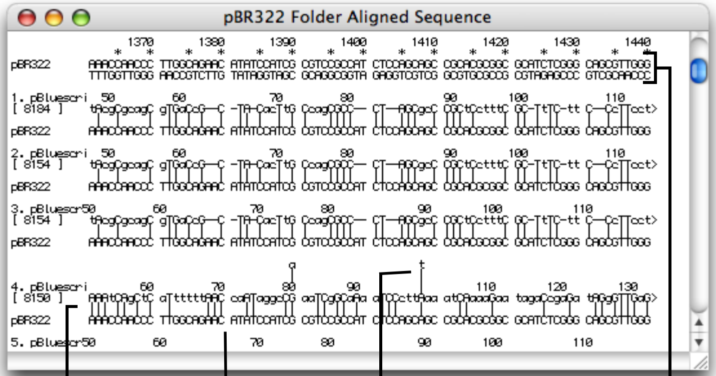
Match Character
- ○ default
- ● upper
- ○ lower
- ○ character: .

Mismatch Character
- ○ default
- ○ upper
- ● lower
- ○ character: .

Numbering
- Numbering (5 - 100): 10
- Vertical Alignment (0 - 250): 20

Alignment characters
- Vertical Aligment char.:
- Horizontal Alignment char.:

Resulting output

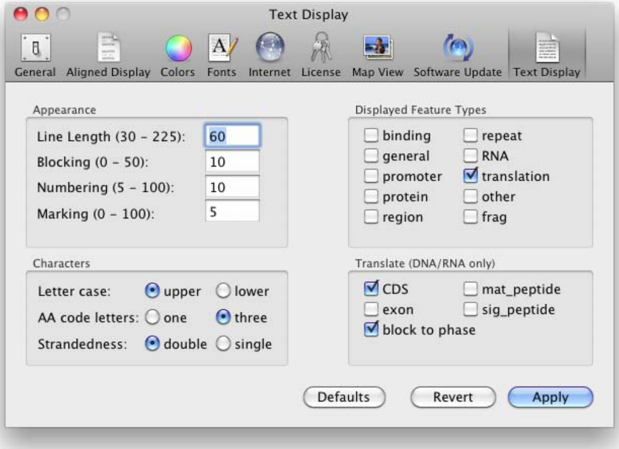


pBR322 Folder Aligned Sequence

```
                  1370       1380       1390       1400       1410       1420       1430       1440
                *    *     *    *     *    *     *    *     *    *     *    *     *    *     *    *
pBR322       AAACCAACCC TTGGCAGAAC ATATCCATCG CGTCCGCCAT CTCCAGCAGC CGCACGCGGC GCATCTCGGG CAGCGTTGGG
             TTTGGTTGGG AACCGTCTTG TATAGGTAGC GCAGGCGGTA GAGGTCGTCG GCGTGCGCCG CGTAGAGCCC GTCGCAACCC

1. pBluescri  50         60              70         80             90        100            110
[ 8184 ]     tAcgCgcagC gTGaCcG--C -TA-CacTtG CcagCGCC-- CT--AGCgcC CGCtCctttC GC-TtTC-tt C--CcTTcct>
pBR322       AAACCAACCC TTGGCAGAAC ATATCCATCG CGTCCGCCAT CTCCAGCAGC CGCACGCGGC GCATCTCGGG CAGCGTTGGG

2. pBluescri  50         60              70         80             90        100            110
[ 8154 ]     tAcgCgcagC gTGaCcG--C -TA-CacTtG CcagCGCC-- CT--AGCgcC CGCtCctttC GC-TtTC-tt C--CcTTcct>
pBR322       AAACCAACCC TTGGCAGAAC ATATCCATCG CGTCCGCCAT CTCCAGCAGC CGCACGCGGC GCATCTCGGG CAGCGTTGGG

3. pBluescr50          60              70         80             90        100            110
[ 8154 ]     tAcgCgcagC gTGaCcG--C -TA-CacTtG CcagCGCC-- CT--AGCgcC CGCtCctttC GC-TtTC-tt C--CcTTcct>
pBR322       AAACCAACCC TTGGCAGAAC ATATCCATCG CGTCCGCCAT CTCCAGCAGC CGCACGCGGC GCATCTCGGG CAGCGTTGGG

                                               g                         t
4. pBluescri         60         70         80         90                    110        120        130
[ 8150 ]     AAAtCAgCtC aTtttttAAC caATaggcCG aaTCgGCaAa aTCCcttAaa atCAaaaGaa tagaCcgaGa tAGgGTTGaG>
pBR322       AAACCAACCC TTGGCAGAAC ATATCCATCG CGTCCGCCAT CTCCAGCAGC CGCACGCGGC GCATCTCGGG CAGCGTTGGG

5. pBluescr50          60              70         80             90        100            110
```

Example 3

This shows the effect of annotated formatting on the aligned sequence display. The annotation (**Show** panel) for translation is checked, and both CDS and block to phase (**Translate** panel) are checked. The query sequence is shown double-stranded.



Appearance

Line Length (30 – 225): 60
Blocking (0 – 50): 10
Numbering (5 – 100): 10
Marking (0 – 100): 5

Displayed Feature Types

□ binding □ repeat
□ general □ RNA
□ promoter ☑ translation
□ protein □ other
□ region □ frag

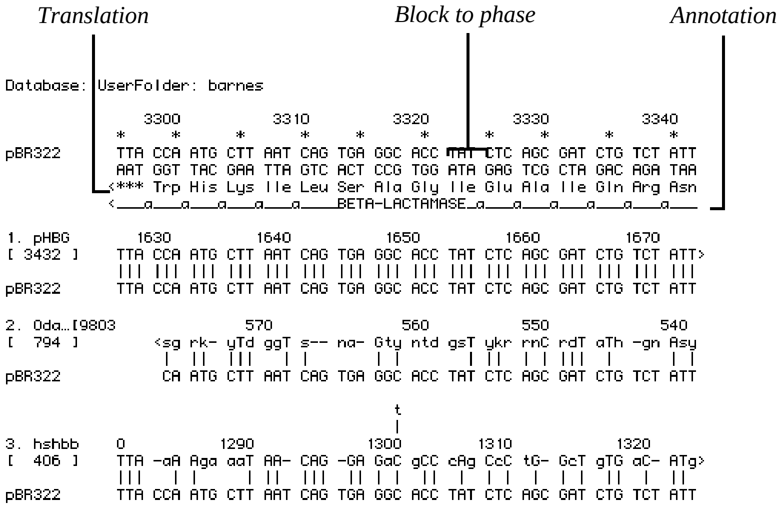
Characters

Letter case: ● upper ○ lower
AA code letters: ○ one ● three
Strandedness: ● double ○ single

Translate (DNA/RNA only)

☑ CDS □ mat_peptide
□ exon □ sig_peptide
☑ block to phase

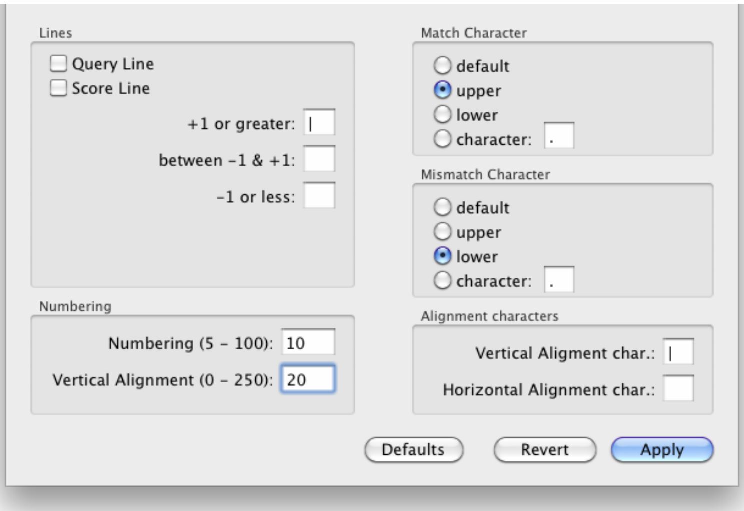
Resulting output



```
Database: UserFolder: barnes

               3300          3310         3320         3330          3340
            *     *      *      *     *      *      *     *      *      *
pBR322      TTA CCA ATG CTT AAT CAG TGA GGC ACC TAT CTC AGC GAT CTG TCT ATT
            AAT GGT TAC GAA TTA GTC ACT CCG TGG ATA GAG TCG CTA GAC AGA TAA
           <*** Trp His Lys Ile Leu Ser Ala Gly Ile Glu Ala Ile Gln Arg Asn
           <___a___a___a___a___a____BETA-LACTAMASE_a___a___a___a___a___a___

1. pHBG       1630         1640          1650         1660         1670
[ 3432 ]    TTA CCA ATG CTT AAT CAG TGA GGC ACC TAT CTC AGC GAT CTG TCT ATT>
            ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
pBR322      TTA CCA ATG CTT AAT CAG TGA GGC ACC TAT CTC AGC GAT CTG TCT ATT

2. Oda…[9803              570              560          550            540
[  794 ]        <sg rk- yTd ggT s-- na- Gty ntd gsT ykr rnC rdT aTh -gn Asy
                 |  ||  |||   | |       | |       | ||  | | |||  |      | |
pBR322           CA ATG CTT AAT CAG TGA GGC ACC TAT CTC AGC GAT CTG TCT ATT

                                          t
                                          |
3. hshbb    0          1290            1300        1310           1320
[  406 ]    TTA -aA Aga aaT AA- CAG -GA GaC gCC cAg CcC tG- GcT gTG aC- ATg>
            |||   | |     | ||  |||  || | |  ||  |  | |  |  | |  ||  |  ||
pBR322      TTA CCA ATG CTT AAT CAG TGA GGC ACC TAT CTC AGC GAT CTG TCT ATT
```

Example 4

To display results simply, turn off all settings on **Options | Format Aligned Display** and **Options | Format Annotated Displa**y.



Lines
- □ Query Line
- □ Score Line

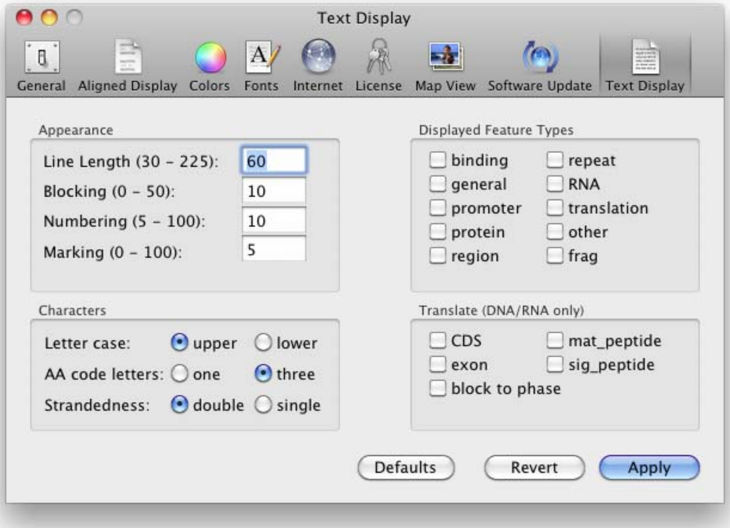
+1 or greater: |
between -1 & +1:
-1 or less:

Match Character
- ○ default
- ● upper
- ○ lower
- ○ character: .

Mismatch Character
- ○ default
- ○ upper
- ● lower
- ○ character: .

Numbering
Numbering (5 - 100): 10
Vertical Alignment (0 - 250): 20

Alignment characters
Vertical Aligment char.: |
Horizontal Alignment char.:



Text Display

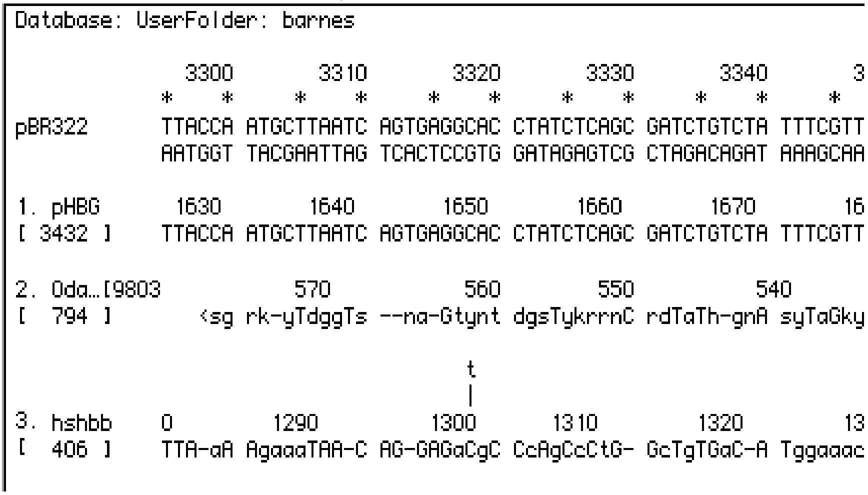
Appearance
Line Length (30 - 225): 60
Blocking (0 - 50): 10
Numbering (5 - 100): 10
Marking (0 - 100): 5

Displayed Feature Types
- □ binding
- □ general
- □ promoter
- □ protein
- □ region
- □ repeat
- □ RNA
- □ translation
- □ other
- □ frag

Characters
Letter case: ● upper ○ lower
AA code letters: ○ one ● three
Strandedness: ● double ○ single

Translate (DNA/RNA only)
- □ CDS
- □ exon
- □ block to phase
- □ mat_peptide
- □ sig_peptide

Resulting output



```
Database: UserFolder: barnes

              3300       3310       3320       3330       3340       3
            *    *     *    *     *    *     *    *     *    *     *
pBR322      TTACCA ATGCTTAATC AGTGAGGCAC CTATCTCAGC GATCTGTCTA TTTCGTT
            AATGGT TACGAATTAG TCACTCCGTG GATAGAGTCG CTAGACAGAT AAAGCAA

1. pHBG      1630       1640       1650       1660       1670       16
[ 3432 ]    TTACCA ATGCTTAATC AGTGAGGCAC CTATCTCAGC GATCTGTCTA TTTCGTT

2. Oda…[9803           570           560        550          540
[  794 ]       <sg rk-yTdggTs --na-Gtynt dgsTykrrnC rdTaTh-gnA syTaGky

                                     t
                                     |
3. hshbb    0        1290         1300      1310        1320        13
[  406 ]    TTA-aA AgaaaTAA-C AG-GAGaCgC CcAgCcCtG- GcTgTGaC-A Tggaaac
```

E

GenBank Feature Tables

Overview

This appendix contains details of the GenBank feature tables used by MacVector.

GenBank feature format

GenBank is the primary US repository of DNA and protein sequences, curated by the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) in Bethesda, Maryland.

Sequences maintained by GenBank have two main types of metadata, which we arbitarily split up into features - which we define as annotations have a defined location on the sequence (such as a gene or site) - and annotations - which are general data associated with a sequence (such as accession number, publication or authors).

GenBank features have defined types (such as CDS, mRNA, promotor, etc.) and only a limited set of types are allowed. The tables below list the feature types allowed for both protein and nucleic acid sequences in MacVector.

Further information about the GenBank file format can be found at:

`http://www.ncbi.nlm.nih.gov/collab/FT/`

Protein feature keywords

The following GenBank feature keywords are supported in MacVector for proteins:

Keyword	**Keyword**
ACT_SITE	MUTAGEN
Amin	NON_CONS
BINDING	NON_STD
CA_BIND	NON_TER
Carb	NP_BIND
CARBOHYD	PEPTIDE
CDS	PROPEP
CHAIN	REGION
Cleavage	REPEAT
COILED	SIGNAL
Comp	SIMILAR
COMPBIAS	SITE
CONFLICT	Source
CROSSLNK	STRAND
DISULFID	Tent
DNA_BIND	THIOETH

Keyword
DOMAIN
Duplic
FRAG
FrgC
FrgH
HELIX
Inhib
INIT_MET
LIPID
METAL
MOD_RES
MOTIF
THIOLEST
TOPO_DOM
TRANSIT
TRANSMEM
Trns
TURN
UNSURE
VAR_SEQ
VARIANT
VARSPLIC
ZN_FING

Nucleic acid feature keywords and qualifiers

The following GenBank keywords and associated qualifiers are supported in MacVector for set nucleic acids.

Note. Some keywords have mandatory qualifiers, as well as optional ones. These are shown in bold text in the table below.

Keyword	**Qualifiers**
-10_signal	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /standard_name
-35_signal	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /standard_name
3'clip	/note
3'UTR	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /standard_name, /trans_splicing
5'clip	/note
5'UTR	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /standard_name, /trans_splicing
allele	/note
anticdn	/note
attack	/note
attenuator	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /phenotype
C_region	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
CAAT_signal	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag
CDS	/allele, /citation, /codon, /codon_start, /db_xref, /EC_number, /exception, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /number, /old_locus_tag, /operon, /product, /protein_id, /pseudo, /ribosomal_slippage, /standard_name, /transcript_id, /translation, /transl_except, /transl_table, /trans_splicing
CDS.ps	/note
cellular	/note
conflict	/allele, **/citation, /compare,** /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /replace
connect	/note
cutds	/note
cutss	/note
D-loop	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag
D_region	/note

Keyword	Qualifiers
D_segment	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
DNA	/note
enhancer	/allele, /bound_moiety, /citation, /db_xref, /experiment, /label, /gene, /gene_synonym, /inference, /locus_tag, /map, /note, /old_locus_tag, /standard_name
exon	/allele, /citation, /db_xref, /EC_number, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /number, /old_locus_tag, /product, /pseudo, /standard_name
form	/note
frag	/frag, /note
gap	**/estimated_length**, /experiment, /inference, /map, /note
GC_signal	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag
gene	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /product, /pseudo, /phenotype, /standard_name, /trans_splicing
iDNA	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /number, /old_locus_tag, /standard_name
insertion_seq	/note
intron	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /number, /old_locus_tag, /pseudo, /standard_name
iRNA	/note
J_region	/note
J_segment	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
LTR	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /standard_name
mat_peptide	/allele, /citation, /db_xref, /EC_number, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
mat_peptide.ps	/note
methyl	/note
misc_binding	/allele, **/bound_moiety**, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag
misc_difference	/allele, /citation, /clone, /compare, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /phenotype, /replace, /standard_name
misc_feature	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /number, /old_locus_tag, /phenotype, /product, /pseudo, /standard_name

Keyword	Qualifiers
misc_recomb	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /standard_name
misc_RNA	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /product, /pseudo, /standard_name, /trans_splicing
misc_signal	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /phenotype, /standard_name
misc_structure	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /standard_name
modified_base	/allele, /citation, /db_xref, /experiment, /frequency, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, **/mod_base**, /note, /old_locus_tag
mRNA	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /product, /pseudo, /standard_name, /transcript_id, /trans_splicing
mRNA.ps	/note
mult	/note
mutation	/note
N_region	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
ncRNA	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, **/ncRNA_class**, /note, /old_locus_tag, /product, /pseudo, /standard_name, /trans_splicing, /operon
old_sequence	/allele, **/citation**, **/compare**, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /replace
operon	/allele, /citation, /db_xref, /experiment, /function, /inference, /label, /map, /note, **/operon**, /phenotype, /pseudo, /standard_name
ORF	/note
oriT	/allele, /bound_moiety, /citation, /db_xref, /direction, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /rpt_family, /rpt_type, /rpt_unit_range, /rpt_unit_seq, /standard_name
polyA_signal	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag
polyA_site	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag
precursor_RNA	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /product, /standard_name, /trans_splicing
prim_transcript	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /standard_name
primer	/note
primer_bind	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /label, /inference, /locus_tag, /map, /note, /old_locus_tag, /standard_name, /PCR_condition

Keyword	Qualifiers
promoter	/allele, /bound_moiety, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /phenotype, /pseudo, /standard_name
protein_bind	/allele, **/bound_moiety**, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /standard_name
provirus	/note
RBS	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /standard_name
Read	/note
Read*	/note
refnumbr	/note
rep_origin	/allele, /citation, /db_xref, /direction, /experiment, /gene, /gene_synonym, /label, /inference, /locus_tag, /map, /note, /old_locus_tag, /standard_name
repeat_region	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /mobile_element, /note, /old_locus_tag, /rpt_family, /rpt_type, /rpt_unit_range, /rpt_unit_seq, /satellite, /standard_name
repeat_unit	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /rpt_family, /rpt_type, /rpt_unit_range, /rpt_unit_seq
rRNA	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /product, /pseudo, /standard_name
S_region	/allele, /citation, /db_xref, /gene, /gene_synonym, /experiment, /label, /inference, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
satellite	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /rpt_type, /rpt_family, /rpt_unit_range, /rpt_unit_seq, /standard_name
scRNA	/note
sig_peptide	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
sig_peptide.ps	/note
snoRNA	/note
snRNA	/note
snRNA.ps	/note

Keyword	Qualifiers
source	/bio_material, /cell_line, /cell_type, /chromosome, /citation, /clone, /clone_lib, /collected_by, /collection_date, /country, /cultivar, /culture_collection, /db_xref, /dev_stage, /ecotype, /environmental_sample, /focus, /frequency, /germline, /haplotype, /identified_by, /isolate, /isolation_source, /label, /lab_host, /lat_lon, /macronuclear, /map, /mating_type, **/mol_type**, /note, /organelle, **/organism**, /PCR_primer, /plasmid, /pop_variant, /proviral, /rearranged, /segment, /serotype, /serovar, /sex, /specimen_voucher, /host, /strain, /sub_clone, /sub_specie, /sub_strain, /tissue_lib, /tissue_type, /transgenic, /variety
stem_loop	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /operon, /standard_name
STS	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /standard_name
TATA_signal	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag
terminator	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /operon, /old_locus_tag, /standard_name
tmRNA	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name, /tag_peptide
transit_peptide	/allele, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
transposon	/note
tRNA	/allele, /anticodon, /citation, /db_xref, /experiment, /function, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name, /trans_splicing
tRNA.ps	/note
unsure	/allele, /citation, /compare, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /replace
V_region	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
V_segment	/allele, /citation, /db_xref, /experiment, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /product, /pseudo, /standard_name
variation	/allele, /citation, /compare, /db_xref, /experiment, /frequency, /gene, /gene_synonym, /inference, /label, /locus_tag, /map, /note, /old_locus_tag, /phenotype, /product, /replace, /standard_name
virion	/note

F References

Overview

This section contains a full listing of references quoted in the User Guide, plus selected academic papers for background reading. There is also a short list of recommended reading about the use of computers in sequence analysis.

References

BLAST algorithms

Altchul et al.(1994). *Nature Genetics* **6** 119-29.
Issues in searching molecular sequence databases.

Altschul et al. (1997). *Nucleic Acids Res.* **25**:3389-3402.
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Karlin & Altschul (1990). *Proc. Natl. Acad. Sci. USA* **87** 2264-68.
Methods for assessing statistical significance of molecular sequence features by using general scoring schemes.

Karlin & Altschul (1993). *Proc. Natl. Acad. Sci. USA* **90** 5873-77.
Applications and statistics for multiple high-scoring segments in molecular sequences.

IUPAC codes for nucleotides and amino acids

A. Cornish-Bowden (1985). *Nucl. Acids Res.* **13** 3021-30.
Nomenclature for incompletely specified bases in nucleic acid sequences recommendations 1984.

IUPAC-IUB Commission on Biological Nomenclature (1968). *J. Biol. Chem.* **243** 3557-59.
A one-letter notation for amino acid sequences. Tentative rules.

Protein secondary structure prediction

Chou-Fasman method

P.Y. Chou & G.D. Fasman (1974). *Biochemistry* **13** 211-22.
Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins.

P.Y. Chou & G.D. Fasman (1974). *Biochemistry* **13** 222-45.
Prediction of protein conformation.

P.Y. Chou & G.D. Fasman (1978). *Ann. Rev. Biochem.* **47** 251-76.
Emprical predictions of protein conformations.

P.Y. Chou & G.D. Fasman (1974). *Adv. Enzymol. Relat. Areas Mol. Biol.* **47** 45-148.
Prediction of the secondary structure of proteins from their amino acid sequence.

General

W. Kabsch & C. Sander (1983). *FEBS Letters* **155** 179-82.
How good are predictions of protein secondary structure?

K. Nishikawa (1983). *Biochim. Biophys. Acta* **748** 285-99.
Assessment of secondary-structure prediction of proteins.

B.A. Wallace, M. Cascio, & D.L. Mileke (1986). *Proc. Natl. Acad. Sci. USA* **83** 9423-27.
Evaluation of methods for the prediction of membrane protein secondary structures.

Robson-Garnier method

B. Robson & E. Suzuki (1976). *J. Mol. Biol.* **107** 327-56.
Conformational properties of amino acid residues in globular proteins.

J. Garnier, D.J. Osguthorpe, & B. Robson (1978). *J. Mol. Biol.* **120** 97-120.
Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.

Protein flexibility

P.A. Karplus & G.E. Schulz (1985). *Naturwiss.* **72** 212-13.
Prediction of chain flexibility in proteins. A tool for the selection of peptide antigens.

Protein profiles

Amphiphilicity

D. Eisenberg, R.M. Weiss, & T.C. Terwilliger (1984). *Proc. Natl. Acad. Sci. USA* **81** 140-44.
The hydrophobic moment detects periodicity in protein hydrophobicity.

D. Eisenberg, E. Schwarz, M. Komaromy, & R. Wall (1984a). *J. Mol. Biol.* **179** 125-42.
Analysis of membrane and surface protein sequences with the hydrophobic moment plot.

G. von Heijne (1986). *EMBO J.* **6** 1335-42.
Mitochondrial targeting sequences may form amphiphilic helices.

Antigenicity

T.P. Hopp & K.R. Woods (1981). *Proc. Natl. Acad. Sci. USA* **78** 3824-28.
Prediction of protein antigenic determinants from amino acid sequences.

B.A. Jameson & H.Wolf (1988). *Comput. Applic. in the Biosciences* **4** 181-186.
The antigenic index A novel algorithm, for predicting antigenic determinants.

J.M.R. Parker, D. Guo & R.S. Hodges (1986). *Biochemistry*, **25** 5425.

J.M. Thornton, M.S. Edwards, W.R. Tayler & D.J. Barlow (1986). *EMBO J.*, **5** 409.

G.W. Welling, W.J. Wiejer, R. Van der Zee & S. Welling-Webster (1985). *FEBS Lett.*, **188** 215.

Hydrophilicity

D.M. Engelman, T.A. Steitz, & A. Goldman (1986). *Ann. Rev. Biophys. Biophys. Chem.* **15** 321-53.
Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins.

T.P. Hopp & K.R. Woods (1981). *Proc. Natl. Acad. Sci. USA* **78** 3824-28.
Prediction of protein antigenic determinants from amino acid sequences.

J. Kyte & R.F. Doolittle (1982). *J. Mol. Biol.* **157** 105-32.
A simple method for displaying the hydropathic character of a protein.

Hydrophobicity

J.L. Fauchere & V. Pliska (1983). *Eur. J. Med. Chem. (Chim. Ther.)*, **18** 369.

J. Janin (1979). *Nature (London)*, **277** 491.

J. Kyte & R.F. Doolittle (1982)., *J. Mol. Biol.*, **157** 105.

P. Manavalan & P.K. Ponnuswamy (1978). *Nature (London)*, **275** 673.

R.M. Sweet & D. Eisenberg (1983). *J. Mol. Biol.*, **171** 479.

G. von Heijne (1981). *Eur. J. Biochem.*, **116** 419.

Surface probability

J. Janin, S. Wodak, M. Levitt, & B. Maigret (1978). *J. Mol. Biol.* **125** 357-86.
Conformation of amino acid side-chains in proteins.

E. Emini, J.V. Hughes, D.S. Perlow, & J. Boger (1965). *J. Virol.* **55** 836-39.

Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide.

Transmembrane

P. Argos, J.K.N. Rao & P.A. Hargrave (1982). *Eur. J. Biochem.,* **128** 565.

G. von Heijne (1992). *J. Mol. Biol.,* **225** 487.

Primer and probe screening analysis

S. Rozen & H. J. Skaletsky (2000). Primer3 on the WWW for general users and for biologist programmers. In S. Krawetz & S. Misener editors, Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.

F. Baldino, Jr., M.-F. Chesselet, & M.E. Lewis (1989). *Methods in Enzymol.* **168** 761-777.

D.K. Bodkin & D.L. Knudsen (1985). J . Virol. *Methods* **10** 45.

K.J. Breslauer, R. Frank, H. Blocker, & L.A. Marky (1986). *Proc. Natl. Acad. Sci. USA* **83** 3746-50.
Predicting DNA duplex stability from the base sequence.

J. Casey & N. Davidson (1977). *Nucl. Acids Res.* **4** 1539.

S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, & D.H. Turner (1986). *Proc. Natl. Acad. Sci. USA* **83** 9373-77.
Improved free-energy parameters for predictions of RNA duplex stability.

L. Hillier & Philip Green (1991). *PCR Meth. and Applic.* **1** 124-128
OSP a computer program for choosing PCR and DNA sequencing primers.

T. Lowe, J. Sharefkin, S.Q. Yang, & C.W. Dieffenbach (1990). *Nucl. Acids Res.* **18** 1757-61.
A computer program for selection of oligonucleotide primers for polymerase chain reactions.

W. Rychlik & R.E. Rhoads (1989). *Nucl. Acids Res.* **17** 8543-51.
A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA.

W. Rychlik, W.J. Spencer & R.E. Rhoads (1990). *Nucl. Acids Res.* **18** 6409-12.
Optimization of the annealing temperature for DNA amplification *in vitro.*

Coding regions

J.W. Fickett (1982). *Nucl. Acids Res.* **10** 5303-18.
Recognition of protein coding regions in DNA sequences.

M. Gribskov, J. Devereux, & R.R. Burgess (1984). *Nucl. Acids Res.* **12** 539-549.
The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression.

R. Staden (1984). *Nucl. Acids Res.* **12** 551-567.
Measurement of the effects that coding for a protein has on a DNA sequence and their use in finding genes.

R. Staden, & A.D. McLachlan (1982). *Nucl. Acids Res.* **10** 141-156.
Codon preference and its use in identifying protein coding regions in long DNA sequences.

Sequence comparisons

D. Davison (1985). *Bull. Math. Biol.* **47** 437-60.
Sequence similarity (‘homology’) searching for molecular biologists.

D. Davison & K.H. Thompson (1984). *Bull. Math. Biol.* **46** 579-90.
A non-metric sequence alignment program.

Robert C. Edgar (2004) *Nucl. Acids Res.* **32**(5) 1792-1797
MUSCLE: multiple sequence alignment with high accuracy and high throughput.

D.F. Feng, M.S. Johnson & R.F. Doolittle (1985). *J. Mol. Evol.* **21** 112-25.
Aligning amino acid sequences comparison of commonly used methods.

S. Karlin & G. Ghandour (1985). *Proc. Natl. Acad. Sci. USA* **82** 8597-8601.
Multiple-alphabet amino acid sequence comparisons of the immunoglobin kappa-chain constant domain.

B. Langmead, C. Trapnell, M. Pop & S.L. Salzberg (2009) *Genome Biology* **10**:R25.
Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.

D.J. Lipman & W.R. Pearson (1985). *Science* **227** 1435-1440.
Rapid and sensitive protein similarity searches.

C. Notredame, D.G. Higgins & J. Heringa (2000). *J Mol Biol.* **302**(1) 205 -217.
T-Coffee: A novel method for fast and accurate multiple sequence alignment.

W.R. Pearson & D.J. Lipman (1988). *Proc. Natl. Acad. Sci. USA* **85** 2444-48.
Improved tools for biological sequence comparisons.

W.R. Pearson (1990). *Methods in Enzymol.* **183** 63-98. ed. R. F. Doolittle.
Rapid and sensitive sequence comparison with FASTP and FASTA.

J.M. Pustell (1988). *Nucl. Acids Res.* **16** 1813-1820.
Interactive molecular biology computing.

J. Pustell & F.C. Kafatos (1982). *Nucl. Acids Res.* **10** 4765-82.
A high speed, high capacity homology matrix zooming through SV40 and polyoma.

J. Pustell & F.C. Kafatos (1984). *Nucl. Acids Res.* **12** 643-655.
A convenient and adaptable package of computer programs for DNA and protein sequence management, analysis, and homology determination.

E. Sobel & H.M. Martinez (1986). *Nucl. Acids Res.* **14** 363-374.
A multiple sequence alignment program.

J. D. Thompson, D. G. Higgins & T. J. Gibson (1994). *Nucl. Acids Res.* **22** 4673-80.
Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice.

M.S. Waterman (1984). *Bull. Math. Biol.* **46** 473-500.
General methods of sequence comparison.

W.J. Wilbur & D.J. Lipman (1983). *Proc. Natl. Acad. Sci. USA* **80** 726-30.
Rapid similarity searches of nucleic acid and protein databanks.

Phylogenetic trees

General

W. H. Li (1997). Molecular Evolution. Sinauer, Sunderland, MA.

D. L. Swofford, G. J. Olsen, P. J. Waddell, & D. M. Hillis (1996). Phylogenetic inference. In Molecular Systematics, (D.M. Hillis, C. Moritz & B. K. Mable, eds.), pp. 407-514. Sinauer, Sunderland, MA.

Distance methods

T. H. Jukes & C. R. Cantor (1969). Evolution of protein molecules. In Mammalian protein metabolism, (H. N. Munro, ed.), pp. 21-132. Academic Press, New York.

M. Kimura (1980). *J. Mol. Evol.* **16** 111-120.
A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences.

P.J. Lockhart, M.A. Steel, M.D. Hendy, & D. Penny (1994). *Mol. Biol. Evol.* **11** 605-612.
Recovering evolutionary trees under a more realistic model of sequence evolution.

F. Tajima & M. Nei (1984). *Mol. Biol. Evol.* **1** 269-285.
Estimation of evolutionary distance between nucleotide sequences.

K. Tamura & M. Nei (1993). *Mol. Biol. .Evol.* **10** 512-526.
Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.

Phylogenetic reconstruction methods

N. Saitou & M. Nei (1987). *Mol. Biol. Evol.* **4** 406-425.
The neighbour-joining method: A new method for reconstructing phylogenetic trees.

Bootstrapping

J. Felsenstein (1985). *Evolution* **39** 783-791.
Confidence limits on phylogenies: an approach using the bootstrap.

Sequence Assembly

Phred

B.Ewing, L.Hillier, M.C.Wendl, and Phil Green (1998) Genome Research **8**:175-185. Base-calling of automated sequencer traces using phred. I. Accuracy assessment.

B.Ewing and P.Green 1998. Genome Research 8:186-194. Base-calling of automated sequencer traces using phred. II. Error probabilities.

Further reading

If you are not familiar with using computers to analyze DNA and protein sequences, the following books are good general references:

M.J. Bishop & C.J. Rawlings, editors (1987). Nucleic Acid and Protein Sequence Analysis A Practical Approach. IRL Press, Oxford & Washington D.C.

Russell F. Doolittle (1986). Of URFs and ORFs. A Primer on How to Analyze Derived Amino Acid Sequences. University Science Books, 20 Edgehill Road, Mill Valley CA 94941.

Russell F. Doolittle editor (1990). Methods in Enzymology, volume 183. Academic Press, San Diego CA.

Michael Gribskov & John Devereux, editors (1991). Sequence Analysis Primer. Stockton Press, New York.

Gunnar von Heijne (1987). Sequence Analysis in Molecular Biology Treasure Trove or Trivial Pursuit? Academic Press, San Diego CA.

G Supported File Formats and File Extensions

Overview

This section contains a full listing of all the sequence file formats supported by MacVector. It also provides details of the file extensions that are used by MacVector for output files.

MacVector has the ability to read and write sequence files in a wide variety of formats. In addition to the standard file **Open** and **Save**/**Save As** commands, MacVector reads sequences files through a number of other functions:

- Multiple Sequence Alignment file import. This lets you add sequences from a file(s) to a nucleic acid or protein alignment.
- Align to Reference file import. This lets you add sequences to the Align to Reference window ready for a sequence confirmation or cDNA assembly.
- Align to Folder. When you select a folder to align a sample sequence to, MacVector will automatically read all sequences of any supported format in the folder.
- Assembler file import. If you are using Assembler, you can import files in any format into the Assembler Project window.
- cross_match vector trimming. (Assembler only). cross_match requires a list of sequences to mask out any vector from input Reads.

Prior to MacVector 10.5, these other functions could only read a subset of the full range of MacVector supported file formats. However, now all of the functions can read any supported file format from the list below.

Note. MacVector automatically detects the format of the file(s) you have selected, so there is never a need to define that ahead of time. Accordingly, unlike many other sequence analysis applications, MacVector has no general import functions, the **File | Open** menu item is all that is required.

Supported file formats

Single nucleic acid sequences

Format	Description
MacVector	MacVector is forwards and backwards compatible with all versions back to 7.0. MacVector does save some additional information to the files that will be ignored by earlier versions, but all the sequence, features, annotations and feature appearance information will be maintained.
BSML	This is an XML format promoted by LabBook and a collaboration of various companies. The original site at www.bsml.org is now defunct, but the format lives on with the Accelrys Gene product from Accelrys. See http://xml.coverpages.org/bsml.html.
GenBank	This is the preferred format for US based researchers wishing to share annotated nucleic acid sequence information with the general community. MacVector fully supports all of the GenBank feature types and qualifiers, so this is the preferred format if you wish to send sequence files to collaborators who do not use MacVector. The disadvantage is that the GenBank format cannot handle "meta data" such as colors and fonts that you may have used to graphically enhance the map of your sequence. The GenBank format is maintained by the National Center for Biotechnology Information (NCBI), a US Government agency that also provides the popular *Entrez* and BLAST online database search capabilities. GenBank format files can contain multiple sequences. See ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt.
EMBL	The EMBL format is particularly popular in Europe where it is maintained by the European Bioinformatics Institute. Like GenBank, it has excellent support for features and annotations, but cannot be used to remember graphical appearance information. EMBL format files can contain multiple sequences. See http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html.
GCG RSF	This is the "Rich Sequence Format" from the Genetics Computer Group "Wisconsin Package" set of programs. GCG was aquired by Accelrys and the product has been discontinued, although many scientists worldwide continue to use the product. RSF files can contain multiple sequences.
GCG SSF	This is the older "Single Sequence Format" from GCG.

Format	**Description**
FastA	FastA files have a very simple format - each sequence entry starts with a ">" character, followed by the name of the sequence. Any text on that line after the first <space> character is consider to be the definition of the sequence. The actual sequence characters (upper or lower case) start on the next line and can continue on subsequent lines until the end of the file or the next ">" is encountered. For readability purposes, the sequence lines are typically wrapped every 80 characters or so.
ASCII/Plain	MacVector can read and write plain sequence files where the only the sequence characters are present in the file. These are similar to FastA files but without the sequence name/definition line and are also typically wrapped at 80 characters for readibility purposes.
DNAStar	The standard single sequence ".seq" LaserGene format is basically the GCG SSF format with the characteristic ".." characters introducing the sequence replaced with "^^". These are the only DNAStar format files MacVector can read, it cannot read the binary LaserGene "Map" files.
Vector NTI	To read sequences directly from Vector NTI use the **Database | Vector NTI Import...** menu item (see Appendix , *“Importing Sequences in Other Formats”*). Alternatively, export the sequences from the Vector NTI database, choosing GenBank format. MacVector can read these files, skipping over any non-GenBank-standard entries that Vector NTI may write to the file.
GeneWorks	GeneWorks is an old Macintosh application last developed by Oxford Molecular. It was discontinued in ~2000, but MacVector can read the single sequence files created by GeneWorks.
DNAStrider	DNA Strider was a popular low-end sequence analysis program from the late 1990s and early 2000s. MacVector can read the DNA, NA and Protein files produced by versions 1.1 and 1.2
IG_Suite	Intelligenetics was one of the first bioinformatics companies. It was acquired by Oxford Molecular in the mid 1990s. The file format was popular at the time but is rarely encountered nowadays.
Staden	This is the format of the Staden suite of biological analysis software. See http://staden.sourceforge.net/overview.html.

Format	**Description**
ABI	These are the chromatogram files produced by the ABI series of automated sequencing machines. MacVector can read files produced by the 310, 373, 377, 3130, 3700 and 3730 machines.
SCF	This is the "Staden Chromatogram Format", a standard format for chromatogram data. This is the only chromatogram format that MacVector can write.
ALF	The file format produced by the Pharmacia Automated Laser Fluorescence (ALF) DNA sequencer

Single protein sequences

Format	Description
MacVector	MacVector is forwards and backwards compatible with all versions back to 7.0. MacVector does save some additional information to the files that will be ignored by earlier versions, but all the sequence, features, annotations and feature appearance information will be maintained.
BSML	This is an XML format promoted by LabBook and a collaboration of various companies. The original site at www.bsml.org is now defunct, but the format lives on with the Accelrys Gene product from Accelrys.See http://xml.coverpages.org/bsml.html.
GenPept	This is the preferred format for US based researchers wishing to share annotated protein sequence information with the general community. MacVector fully supports all of the GenPept feature types and qualifiers, so this is the preferred format if you wish to send sequence files to collaborators who do not use MacVector. The disadvantage is that the GenPept format cannot handle "meta data" such as colors and fonts that you may have used to graphically enhance the map of your sequence. The GenPept format is maintained by the National Center for Biotechnology Information (NCBI), a US Government agency that also provides the popular *Entrez* and BLAST online database search capabilities. GenPept format files can contain multiple sequences.
GCG RSF	This is the "Rich Sequence Format" from the Genetics Computer Group "Wisconsin Package" set of programs. GCG was aquired by Accelrys and the product has been discontinued, although many scientists worldwide continue to use the product. RSF files can contain multiple sequences.
GCG SSF	This is the older "Single Sequence Format" from GCG.
UniProt	UniProt is the collaborative amino acid database of EBI, SIB and PIR.

Format	Description
FastA	FastA files have a very simple format - each sequence entry starts with a ">" character, followed by the name of the sequence. Any text on that line after the first <space> character is consider to be the definition of the sequence. The actual sequence characters (upper or lower case) start on the next line and can continue on subsequent lines until the end of the file or the next ">" is encountered. For readability purposes, the sequence lines are typically wrapped every 80 characters or so.
ASCII/Plain	MacVector can read and write plain sequence files where the only the sequence characters are present in the file. These are similar to FastA files but without the sequence name/definition line and are also typically wrapped at 80 characters for readibility purposes.
PIR	Protein information resource (PIR) is an annotated, non-redundant and cross-referenced database of protein sequences at the NBRF.
trEMBL	trEMBL is a large protein database, in Swiss-Prot format, generated by computer translation of the genetic information from the EMBL Nucleotide Sequence Database database. It is the protein equivalent of the EMBL DNA format.
Swiss-Prot	Swiss-Prot is a manually curated biological database of protein sequences. Swiss-Prot was developed by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. It provides reliable protein sequences associated with a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
DNAStrider	DNA Strider was a popular sequence analysis program from the late 1990s and early 2000s. MacVector can read the DNA, NA and Protein files produced by versions 1.1 and 1.2

Multiple sequences

Format	Description
MacVector	MacVector is forwards and backwards compatible with all versions back to 7.0.

Format	Description
GenBank/ GenPept	This is the preferred format for US based researchers wishing to share annotated sequence information with the general community. MacVector fully supports all of the GenBank feature types and qualifiers, so this is the preferred format if you wish to send sequence files to collaborators who do not use MacVector. The disadvantage is that the GenBank format cannot handle "meta data" such as colors and fonts that you may have used to graphically enhance the map of your sequence. The GenBank format is maintained by the National Center for Biotechnology Information (NCBI), a US Government agency that also provides the popular *Entrez* and BLAST online database search capabilities. See ftp://ftp.ncbi.nih.gov/genbank/ gbrel.txt.
EMBL/ trEMBL	The EMBL format is particularly popular in Europe where it is maintained by the European Bioinformatics Institute. Like GenBank, it has excellent support for features and annotations, but cannot be used to remember graphical appearance information. See http://www.ebi.ac.uk/embl/Documentation/User_manual/ usrman.html.
GCG RSF	This is the "Rich Sequence Format" from the Genetics Computer Group "Wisconsin Package" set of programs. GCG was acquired by Accelrys and the product has been discontinued, although many scientists worldwide continue to use the product. RSF files can contain multiple sequences.
GCG MSF	This is the older "Multiple Sequence Format" from GCG. No feature information is kept in these files.
FastA	FastA files have a very simple format - each sequence entry starts with a ">" character, followed by the name of the sequence. Any text on that line after the first <space> character is consider to be the definition of the sequence. The actual sequence characters (upper or lower case) start on the next line and can continue on subsequent lines until the end of the file or the next ">" is encountered. For readability purposes, the sequence lines are typically wrapped every 80 characters or so.

Format	Description
NEXUS	The NEXUS file format was introduced by Maddison *et.al.* (1997) and is used a number of popular phylogeny programs. See http://www.ncbi.nlm.nih.gov/pubmed/11975335?dopt=Abstract. This is a popular format for use in dedicated phylogenetic reconstruction applications such as PAUP.
PHYLIP	This is the format of Joe Felsenstein's Phylogeny Inference Package. See http://cmgm.stanford.edu/phylip/.MacVector defaults to a sequence title width of 10 characters, the PHYLIP default.

MacVector file extensions

MacVector now uses Uniform Type Identifiers (UTIs) to uniquely identify file formats that the application either reads or creates. See Apple's description of UTIs on their documentation pages:

- http://developer.apple.com/documentation/Carbon/Conceptual/understanding_utis/understand_utis_intro/chapter_1_section_1.html#//apple_ref/doc/uid/TP40001319-CH201-SW1
- http://developer.apple.com/documentation/Carbon/Conceptual/understanding_utis/understand_utis_conc/chapter_2_section_2.html#//apple_ref/doc/uid/TP40001319-CH202-SW1

The table below is a list of the file types MacVector will read and create. MacVector exports those UTIs that are unique to MacVector: matrix, bias, binary sequence files, etc. In addition MacVector imports a number of chemical MIME type or text sequence files. These UTIs are all registered with Mac OS X and Launch Services uses the information to associate files with MacVector.

The UTI brings together type identification from multiple methodologies. The files saved on Mac OS, where OSType has been a common means of identifying files with earlier systems, can now work with other systems (Windows, Linux) that might have used file extensions. When downloading files MIME types are employed to identify data formats so that browsers and other tools can locate appropriate applications to handle the data. You should be able to open all supported file formats whether created on a Macintosh or transferred from another system.

Third party formats

Name	UTI	OSType	Ext	Ext Syns	MIME Types
ASCII Sequence Document	biosequence.raw	TEXT	asc	seq txt	biosequence/plain
BSML Document	biosequence.bsml	TEXT	bsml	bsm	
DNA Chromatogram Document	biosequence.abi	ABIF	abi	ab1 abf	application/x-dna
DNA Sequence Chromatogram Document	biosequence.scf	SCF3	scf		biosequence/scf
DNA Strider DNA Degenerate Document	net.sourceforge.projects.dnastrider.xdgn	xDGN	xdgn		
DNA Strider DNA Document	net.sourceforge.projects.dnastrider.xdna	xDNA	xdna	strider	biosequence/strider
DNA Strider Protein Document	net.sourceforge.projects.dnastrider.xpr	xPRT	xprt	strider	biosequence/strider
DNA Strider RNA Document	net.sourceforge.projects.dnastrider.xrna	xRNA	xrna	strider	biosequence/strider
EMBL Nucleotide Format Document	biosequence.embl	TEXT	embl	dat ebl emb	biosequence/embl chemical/x-embl-dl-nucleotide
FASTA Formatted Sequence Document	biosequence.fasta	TEXT	fa	fna fsa fas fast fasta mpfa nt seq	biosequence/fasta chemical/x-fasta chemical/seq-na-fasta
FASTA Formatted Sequence Document	biosequence.fastp	TEXT	faa	fa faa fsa fas fast fasta fastp mpfa seq	biosequence/fasta chemical/x-fasta chemical/seq-aa-fasta
FASTQ Formatted Sequence Document	biosequence.fastq	TEXT	fq	fastq	chemical/x-fastq chemical/seq-aa-fastq chemical/seq-na-fastq

Name	UTI	OSType	Ext	Ext Syns	MIME Types
GenBank Flat Document	biosequence.genbank	TEXT	gb	gbank gbk gbs gbwithparts gen genbank	biosequence/genbank chemical/x-genbank chemical/seq-na-genbank
GeneWorks AA Document	com.intelligenetics.geneworks.aa				
GeneWorks NA Document	com.intelligenetics.geneworks.na				
GenPept Flat Document	biosequence.genpept	TEXT	gp	gbank gbk gbs gen genbank	biosequence/genbank chemical/seq-aa-genpept
IG Suite Document	biosequence.ig	TEXT	ig		biosequence/ig
Multiple Sequence Format Document	biosequence.msf	TEXT	msf		biosequence/msf
National Biomedical Research Foundation Document	biosequence.nbrf	TEXT	nbr	nbrf pir	biosequence/nbrf
NCBI ASN.1 Sequence Document	biosequence.asn1	TEXT	asn	asn1	biosequence/asn1 chemical/x-ncbi-asn1 chemical/x-ncbi-asn1-ascii
Nexus Document	biosequence.nexus	TEXT	nex	nexus nxs	biosequence/nexus
Phylip Document	biosequence.phylip	TEXT	phy	ph phylip phylip2	biosequence/phylip biosequence/phylip2
Rich Sequence Format Document	biosequence.rsf	TEXT	rsf	gcg	biosequence/gcg
Rich Text Document	public.rtf	RTF	rtf		text/rtf
Sequin FASTA Document	biosequence.sqn	TEXT	sqn		
Sequin Feature Table Document	biosequence.tbl	TEXT	tbl		
Staden Suite Document	biosequence.sta	TEXT	sta		
Swiss Protein Document	biosequence.swissprot	TEXT	sw	sp swi swiss swissprot uniprot	chemical/x-swissprot
Text Document	public.plain-text	TEXT	text	txt	text/plain
Transfac Profile Document	biosequence.transfac	TEXT	transfac		

Name	UTI	OSType	Ext	Ext Syns	MIME Types
VectorNTI DNA Archive Document	com.invitrogen.vectornti.na		ma4		
VectorNTI Protein Archive Document	com.invitrogen.vectornti.aa		pa4		

MacVector formats

Name	UTI	OSType	Ext	MIME Type
MacVector Assembly File	com.macvector.assembly	AXML	axml	application/vnd.macvector.assembly
MacVector Codon Bias File	com.macvector.basis	BIAS	bias	application/vnd.macvector.bias
MacVector NA Alignment File	com.macvector.msan	MSAN	msan	application/vnd.macvector.msan
MacVector AA Alignment File	com.macvector.msap	MSAP	msap	application/vnd.macvector.msap
MacVector NA Matrix File	com.macvector.nmat	NMAT	nmat	application/vnd.macvector.nmat
MacVector NA Subsequence File	com.macvector.nsub	NSUB	nsub	application/vnd.macvector.nsub
MacVector NA Sequence File	com.macvector.nucleotide	NUCL	nucl	application/vnd.macvector.nucleotide
MacVector AA Enzyme File	com.macvector.penz	PENZ	penz	application/vnd.macvector.penz
MacVector AA Matrix File	com.macvector.pmat	PMAT	pmat	application/vnd.macvector.pmat
MacVector AA Sequence File	com.macvector.protein	PROT	prot	application/vnd.macvector.protein
MacVector AA Subsequence File	com.macvector.psub	PSUB	psub	application/vnd.macvector.psub
MacVector NA Enzyme File	com.macvector.renz	RENZ	renz	application/vnd.macvector.renz

H

Importing Sequences in Other Formats

Overview

This section describes how to import sequences in other formats directly into MacVector.

Importing Data from Vector NTI Databases

You can now import sequence data from databases created by Vector NTI Advance v10 or earlier directly into MacVector.

To import sequence data from a Vector NTI database

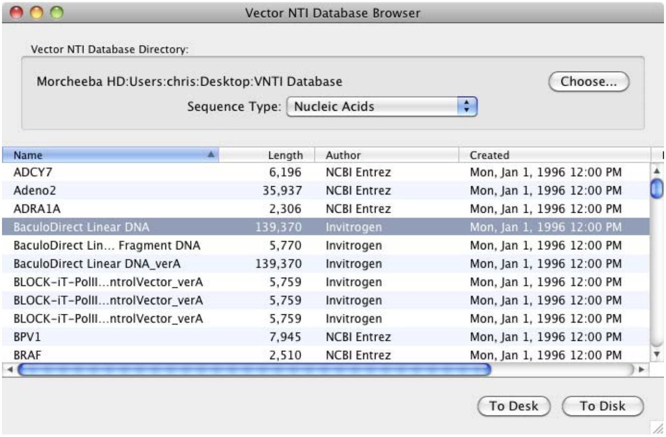
1. Select **Database | Vector NTI Import...** from the menu.

The Vector NTI Database Browser dialog is displayed.

2. Use the **Choose...** button to locate and select the Vector NTI database folder you want to import data from.

Note. The database you select can reside on a remote Windows machine. To access a database on such a machine, ensure that the parent directory, containing the database folder, is shared then connect to the remote machine using the Finder **Go | Connect to Server...** menu option.

A list of all the sequences available in the selected database is displayed.



Vector NTI Database Directory:

Morcheeba HD:Users:chris:Desktop:VNTI Database

Sequence Type: Nucleic Acids

Name	Length	Author	Created
ADCY7	6,196	NCBI Entrez	Mon, Jan 1, 1996 12:00 PM
Adeno2	35,937	NCBI Entrez	Mon, Jan 1, 1996 12:00 PM
ADRA1A	2,306	NCBI Entrez	Mon, Jan 1, 1996 12:00 PM
BaculoDirect Linear DNA	139,370	Invitrogen	Mon, Jan 1, 1996 12:00 PM
BaculoDirect Lin... Fragment DNA	5,770	Invitrogen	Mon, Jan 1, 1996 12:00 PM
BaculoDirect Linear DNA_verA	139,370	Invitrogen	Mon, Jan 1, 1996 12:00 PM
BLOCK-iT-PolII...ntrolVector_verA	5,759	Invitrogen	Mon, Jan 1, 1996 12:00 PM
BLOCK-iT-PolII...ntrolVector_verA	5,759	Invitrogen	Mon, Jan 1, 1996 12:00 PM
BLOCK-iT-PolII...ntrolVector_verA	5,759	Invitrogen	Mon, Jan 1, 1996 12:00 PM
BPV1	7,945	NCBI Entrez	Mon, Jan 1, 1996 12:00 PM
BRAF	2,510	NCBI Entrez	Mon, Jan 1, 1996 12:00 PM

3. Use the **Sequence Type** drop-down menu to toggle between Nucleic Acid and Protein sequences.
4. Click on a column heading to sort the list based on the data in that column.
5. Select the sequence or sequences in the list that you want to import.
6. Click **To Desk** to open the selected sequences in separate Sequence windows within MacVector.
7. Alternatively, click **To Disk** to save the selected sequences as MacVector format sequence files in a specified folder.

The standard features and annotations associated with the imported sequences in the Vector NTI database are retained. However, any feature appearance information is discarded and the default MacVector feature representations are used instead. MacVector also replaces any restriction enzyme sites annotated in the database with the default dynamic set of sites used in the Map view.

`Phrap` Assembly Parameter Contraints

Overview

This appendix contains reference tables for the `Phrap` assembly parameters used in MacVector.

The original in-depth `phrap` and `cross_match` documentation is installed in the `MacVector 12.6/Documentation/` folder as `phrap.pdf`. You can use the table in this appendix to cross-reference the command line parameters described in that document with the human-readable parameter names used in MacVector with Assembler.

Cross_match basic parameters and defaults

Assembler Name	**Min**	**Max**	**Command line parameter**	**Default**
Pairwise alignments				
Mismatch penalty	-100	0	Penalty	-2
Gap initiation penalty	-100	0	GapInit	Penalty-2
Gap extension penalty	-100	0	GapExt	Penalty-1
Insertion gap penalty	-100	0	InsGapExt	-3
Deletion gap penalty	-100	0	DelGapExt	-3
Bonded search				
Minimum match length	0	1000	MinMatch	14
Maximum match length	0	1000	MaxMatch	30
Minimum alignment score	0	1000	Raw	false
Maximum group size	0	1000	MaxGroupSize	20
Half bandwidth	0	1000	Bandwidth	14
Use Raw Word Lengths			WordRaw	false
Globality				
Filtering				
Mask Level (%)	0	101	MaskLevel	80
Min. Alignment score			MinScore	30

Phrap basic parameters and defaults

Assembler Name	**Min**	**Max**	**Command line parameter**	**Default**
Pairwise alignments				
Mismatch penalty	-100	0	Penalty	-2
Gap initiation penalty	-100	0	GapInit	Penalty-2

Gap extension penalty	-100	0	GapExt	Penalty-1
Bonded search				
Minimum match length	0	1000	MinMatch	14
Maximum match length	0	1000	MaxMatch	30
Filtering				
Minimum alignment score	0	1000	MinScore	30
Potential vector bases	0	1000	VectorBound	80
Assembly				
Stringency	0	10	Forcelevel	0
Maximum gap	0	1000	Maxgap	30
Repeat stringency	0.01	0.99	RepeatStringency	0.95
Consensus				
Minimum segment size	0	1000	Node_Seg	8
Node spacing	0	1000	Node_Space	4

Phrap advanced parameters and defaults

Assembler Name	**Min**	**Max**	**Command line parameter**	**Default**
Pairwise alignments				
Insertion gap penalty	-100	0	InsGapExt	-3
Deletion gap penalty	-100	0	DelGapExt	-3
Use Raw SWAT scores			Raw	false
Bonded search				
Maximum group size	0	1000	MaxGroupSize	20
Half bandwidth	0	1000	Bandwidth	14
Input				
Default quality values	0	99	DefaultQual	15

Trim start	0	1000	TrimStart	0
Assembly				
Bypass level			Bypasslevel	1
Revise Greedy			ReviseGreedy	false
Shatter Greedy			ShatterGreedy	false
Pre-assemble groups			PreAssemble	false
Ignore high Qual discrepancies			ForceHigh	false

Phrap miscellaneous parameters and defaults

Assembler Name	**Min**	**Max**	**Command line parameter**	**Default**
Miscellaneous				
Maximum subclone size	1000	10	MaxSubcloneSize	5000
Index word size	1	20	Indexwordsize	10
Retain Duplicate reads			RetainDuplicates	false
Trimming				
Trim penalty	-100	0	TrimPenalty	-2
Minimum trim score	0	1000	TrimScore	20
High quality threshold	0	99	TrimQual	13
Confirmation				
Minimum segment size	1	1000	ConfirmLength	8
Trim segment ends	0	100	ConfirmTrim	1
Penalty	-100	0	ConfirmPenalty	-5
Minimum score	1	1000	ConfirmScore	30