MEGA

Molecular Evolutionary Genetics Analysis

Version 4.0 (Beta release)

Koichiro Tamura, Joel Dudley

Masatoshi Nei, Sudhir Kumar,

Draft

Center for Evolutionary Functional Genomics

Biodesign Institute

Arizona State University

Table of Contents

1 Preface

1.1 Preface

Genome sequencing is generating vast amounts of DNA sequence data from a wide range of organisms. As a result, gene sequence databases are growing rapidly. In order to conduct efficient analyses of these data, there is a need for easy-to-use computer programs, containing fast computational algorithms and useful statistical methods.

The objective of the *MEGA* software has been to provide tools for exploring, discovering, and analyzing DNA and protein sequences from an evolutionary perspective. The first version was developed for the limited computational resources that were available on the average personal computer in early 1990s. *MEGA1* made many methods of evolutionary analysis easily accessible to the scientific community for research and education. *MEGA2* was designed to harness the exponentially greater computing power and graphical interfaces of the late 1990's, fulfilling the fast-growing need for more extensive biological sequence analysis and exploration software. It expanded the scope of its predecessor from single gene to genome wide analyses. Two versions were developed (2.0 and 2.1), each supporting the analyses of molecular sequence (DNA and protein sequences) and pairwise distance data. Both could specify domains and genes for multi-gene comparative sequence analysis and could create groups of sequences that would facilitate the estimation of within- and among- group diversities and infer the higher-level evolutionary relationships of genes and species. *MEGA2* implemented many methods for the estimation of evolutionary distances, the calculation of molecular sequence and genetic diversities within and among groups, and the inference of phylogenetic trees under minimum evolution and maximum parsimony criteria. It included the bootstrap and the confidence probability tests of reliability of the inferred phylogenies, and the disparity index test for examining the heterogeneity of substitution pattern between lineages.

MEGA 4 continues where *MEGA2* left off, emphasizing the integration of sequence acquisition with evolutionary analysis. It contains an array of input data and multiple results explorers for visual representation; the handling and editing of sequence data, sequence alignments, inferred phylogenetic trees; and estimated evolutionary distances. The results explorers allow users to browse, edit, summarize, export, and generate publication-quality captions for their results. *MEGA 4* also includes distance matrix and phylogeny explorers as well as advanced graphical modules for the visual representation of input data and output results. These features, which we discuss below, set MEGA apart from other comparative sequence analysis programs

As with previous versions, *MEGA 4* is specifically designed to reduce the time needed for mundane tasks in data analysis and to provide statistical methods of molecular evolutionary genetic analysis in an easy-to-use computing workbench. While *MEGA 4* is distinct from previous versions, we have made a special effort to retain the user-friendly interface that researchers have come to identify with

MEGA. This interface is obtains information from the user only on a need-to-know basis. Furthermore, the data subsets and output results are stored in files for viewing only if the user specifically needs to do so.

1.2 Copyright

1.3 Disclaimer

1.4 Acknowledgements

Many friends and colleagues have provided encouragement and assistance in the development of *MEGA*. Beta Test versions of MEGA have been used in the research laboratories of the authors, in the classrooms of Sudhir Kumar at the Arizona State University and Masatoshi Nei at the Pennsylvania State University, and by the thousands of users that signed up for the MEGA 4 Beta program. The feedback and bug reports provided by these groups of users were invaluable to the development team. Almost all facets of design and implementation benefited from their comments and suggestions.

MEGA software development is supported by research grants from the NIH, NSF, and Burroughs-Wellcome Fund.

1.5 MEGA 4 Software Development Team

Project Director

Sudhir Kumar

Programming Efforts

Principal Programmers

Koichiro Tamura & Sudhir Kumar

Associate Programmer

Joel Dudley

User Interface Design

Sudhir Kumar & Koichiro Tamura

Documentation

Sudhir Kumar

Koichiro Tamura

Joel Dudley

Website Designs and Implementation

Wayne Parkhurst

Joel Dudley

Quality Assurance

Graziela Valente

See also Acknowledgements.

1.6 Citing MEGA in Publications

WARNING: This version of MEGA is for testing purposes only. Please do not publish results obtained with this version.

If you wish to cite *MEGA* in your publications, we suggest the following:

(1) When referring to *MEGA* in the main text of your publication, you may choose a format such as:

Phylogenetic and molecular evolutionary analyses were conducted using *MEGA* version 4 (Tamura, Dudley, Nei, and Kumar 2006).

(For later versions of *MEGA*, replace 4 with the appropriate version number [e.g., 4.5], which is displayed at the top of the main MEGA application window.

(2) When including a *MEGA* citation in the Literature Cited/Bibliography section, you may use the following:

S Kumar, K Tamura, and M Nei (2004) *MEGA3*: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics* 5:150-163.

2 Part I: Getting Started

2.1 Installing MEGA

2.11 System Requirements

MEGA 4 was developed for use on Microsoft Windows® operating systems: Windows 95/98, NT, ME, 2000, XP, or later. We recommend a computer with at least 64 MB of RAM, 20 MB of available hard disk space, and an entry- level Pentium® processor or equivalent. Our tests show that *MEGA 4* runs well on computers with an entry-level Pentium® CPU. However, for quick computation of large datasets, you should have a faster processor and larger amount of physical memory (RAM). *MEGA 4* also can be run on other operating systems for which Windows emulators are available.

Platform	Software
Macintosh	Windows using *VirtualPC*
Sun Workstation	SoftWindows95
Linux	Windows using *VMWare*

2.12 Installing MEGA

The preferred way to install *MEGA 4* is directly from the website (www.megasoftware.net). A specially designed installation program automatically downloads *MEGA 4* and installs it in the location (directory) you specify.

If you are unable to install *MEGA 4* directly from the website, you can download it as a single compressed ZIP file. Then you must use a program, such as WinZip, to uncompress this ZIP file in a temporary directory. Click on the MEGA**SETUP.EXE** file to install *MEGA 4* on your computer automatically.

Finally, you may install *MEGA 4* from a CD obtained from the authors. In this case, insert the media into the computer and then click on MEGA**SETUP.EXE**.

We recommend that you install *MEGA 4* in one of the three ways described above. Please do not simply copy *MEGA 4*-related files from one computer to another, as *MEGA 4* may not work properly if installed in this manner.

2.13 Uninstalling MEGA

The preferred way to uninstall programs in *Windows* is to use *Add/Remove Programs* option in the control panel, which is accessible from the *Start* button on

the lower left corner of your computer desktop. A dialog box (usually named *Add/Remove programs*) will display a list of programs. To remove *MEGA 4*, scroll down to *MEGA 4* so that it is highlighted, then click *Add/Remove*.

2.2 Features & Support

2.21 What's New in Version 4

Version 4 contains a number of enhancements over *MEGA 3.1*. They include

Maximum Composite Likelihood method (MCL):

- Estimation of evolutionary distances.
- Construction of phylogenetic trees from MCL distances.
- Inference of patterns of nucleotide substitution including estimation of transition/transversion rate ratios (kappa1, kappa2), and the transition/transversion rate bias (R).

Real-time Caption Expert Engine:

- Generates publication-quality captions from analysis results.
- Captions are generated dynamically based on the analysis options used, and manual adjustments made within result explorers.
- Captions for all result explorers available in MEGA.
- Captions can be saved to a file, copied to an external program (Word, Excel, etc), or printed directly from MEGA.

2.22 Feature List

MEGA Version	1.0	2.x	3.x	4.x
Platform	DOS	Win	Win	Win
Input Data				
DNA, Protein, Pairwise distance matrix	•	•	•	•
Sequence Alignment Construction				
Alignment Editor				
Manual editing of DNA and Protein sequences			•	•
Motif searching/highlighting			•	•

Synchronous alignment editing of original and translated cDNA	•	•
Copy/Paste sequences To/From Clipboard	•	•
Save alignment session for future display	•	•
Ability to read sequencer, MEGA, NEXUS, FASTA, and other formats	•	•
Apply color/highlight schemes to sequence data	•	•
Write alignment to MEGA file for direct analysis in MEGA	•	•
BLAST sequences from alignment directly	•	•
Multiple Sequence Alignment		
Complete native implementation of ClustalW	•	•
Ability to select all options on the fly	•	•
Ability to align any user-selected region	•	•
Ability to align translated cDNA sequences and automatic adjustment	•	•
Sequencer (Trace) File editor/viewer		
View ABI (*.abi, .ab1) and Studfen (*.std?)	•	•
Edit trace file	•	•
Mask vector (or any other region)	•	•
Launch direct BLAST search for whole or selected sequence	•	•
Send data directly to Alignment Editor	•	•

Integrated Web Browser and Sequence Fetching				
Direct "usual" web and GenBank browsing from MEGA			•	•
One-click sequence fetching from databanks queries			•	•
Send sequence data from BLAST search directly into alignment			•	•
Bookmark favorite sequence databank sites			•	•
Data Handling				
Handling ambiguous states (R,Y,T, etc.)		•	•	•
Extended MEGA format to save all data attributes		•	•	•
Importing Data from other formats (Clustal/Nexus/etc.)		•	•	•
Data Explorers				
Sequence	•	•	•	•
Distance matrix		•	•	•
Attributes supported				
Groups of Sequences/Taxa		•	•	•
Domains	•	•	•	•
Genes and Mixed Domain attributes		•	•	•
Explicit labels for sites		•	•	•
Automatic codon translation	•	•	•	•
Selection of codon positions	•	•	•	•
Selection of different site categories		•	•	•
Visual Specification of Domains/Groups		•	•	•
Center Analysis Preferences Dialog			•	•

Unlimited Data size for Analysis		•	•	•
Genetic Code Table Selection				
Choose a desired table	•	•	•	•
Ability to add/edit user defined tables		•	•	•
Computation of statistical attributes of a code table				
Degeneracy of codon positions		•	•	•
Numbers of potential synonymous sites		•	•	•
Inclusion of all known code tables		•	•	•
Real-Time Caption Expert Engine				
Generate Captions for Distance Matrices				•
Generate Captions for Phylogenies				•
Generate Captions for Tests				•
Generate Captions for Alignments				•
Copy Captions to External Programs				•
Save/Print Captions				•
Integrated Text File Editor				
Unlimited Text File Size		•	•	•
Multi-file Tabbed Display		•	•	•
Columnar Block selection/Editing		•	•	•
Undo/Redo operations		•	•	•
Line numbers		•	•	•
Utilities to Format Sequences/Reverse complement etc.		•	•	•
Copy Screenshots to EMF/WMF/Bitmap for presentation		•	•	•
Sequence Data Viewer				

Two dimensional display of molecular sequences	•	•	•	•
Display with identity symbol		•	•	•
Drag-drop sorting of sequences		•	•	•
Mixing coding and non-coding sequence display		•	•	•
One-click translation	•	•	•	•
Display with all or only selected taxa		•	•	•
Data Export				
PAUP3, PHYLIP	•	•	•	•
PAUP4, PHYLIP Interleaved		•	•	•
Highlighting				
0,2,4-fold degenerate sites	•	•	•	•
Variable, parsimony informative sites	•	•	•	•
Constant Sites		•	•	•
Statistical Quantities estimation				
DNA and protein sequence compositions	•	•	•	•
Estimation by genes/domains/groups		•	•	•
Codon Usage	•	•	•	•
Estimation by genes/domains/groups		•	•	•
Use only highlighted sites		•	•	•
MCL-based Estimation of Nucleotide Substitution Patterns				
4x4 Rate Matrix			•	•
Transition/Transversion Rate Ratios (k1, k2)			•	•
Transition/Transversion Rate Bias (R)			•	•

Substitution Pattern Homogeneity Test

Composition Distance		•	•	•
Disparity Index		•	•	•
Monte-Carlo Test		•	•	•

Distance Estimation Methods

Nucleotide-by-Nucleotide

Models

No. of differences, p-distance, Jukes-Cantor, Kimura 2P	•	•	•	•
Tajima-Nei, Tamura 3-parameter, Tamura-Nei distance	•	•	•	•
LogDet (Tamura-Kumar)			•	•
Maximum Composite Likelihood				•

Subcomponents

Transitions (ts), tranversions (tv), ts/tv ratio	•	•	•	•
Number of common sites		•	•	•
Account for rate variation among sites	•	•	•	•
Relaxation of the homogeneity assumption			•	•

Synonymous/Nonsynonymous (Codon-by-Codon)

Models

Nei-Gojobori (1986) method	•	•	•	•
Modified Nei-Gojobori method		•	•	•
Li-Wu-Lou, PBL, Kumar method		•	•	•

Subcomponents

Synonymous (s), nonsynonymous (n) distances	•	•	•	•
Numbers of synonymous and nonsynonymous sites		•	•	•
Differences and ratios (s-n, n-s, s/n, n/s)		•	•	•
4-fold degenerate site distances		•	•	•
0-fold degenerate site distances		•	•	•
Number of 0-fold and 4-fold degenerate sites		•	•	•
Protein distance				
Number of differences, p-distance, Poisson	•	•	•	•
Dayhoff and JTT distances			•	•
Account for rate variation among sites		•	•	•
Relaxation of the homogeneity assumption			•	•
Distance Calculations				
Pairwise	•	•	•	•
Between Group Average		•	•	•
Within Group Average		•	•	•
Net between group Average		•	•	•
Overall average		•	•	•
Sequence Diversity Calculations				
Mean Diversity within Subpopulations		•	•	•
Mean Diversity for Entire Populaton		•	•	•
Mean Interpopulational Diversity		•	•	•

Coefficient of Differentiation		•	•	•
Variance Calculations				
Analytical	•	•	•	•
Bootstrap		•	•	•
Handling missing data	•	•	•	•
Automatic translation	•	•	•	•
Automatic pasting of partial codons between exons		•	•	•
Tests of Selection				
Codon-based tests				
Large sample Z-test	•	•	•	•
Between Sequences	•	•	•	•
Within groups		•	•	•
Overall sequences		•	•	•
Fisher's Exact Test		•	•	•
Tajima's Test of Neutrality		•	•	•
Molecular Clock Test				
Tajima's relative rate test		•	•	•
Tree-making Methods				
Neighbor-Joining	•	•	•	•
Randomized tie-breaking in bootstrapping		•	•	•
Minimum Evolution method		•	•	•
Branch-swapping (Close-Neighbor-Interchange; CNI)		•	•	•
Fast OLS computation method		•	•	•
UPGMA	•	•	•	•
Randomized tie-breaking in bootstrapping		•	•	•
Maximum Parsimony				

Nucleotide sequences	•	•	•	•
Protein sequences		•	•	•
Max-mini branch-and-bound and min-mini searches	•	•	•	•
Branch-swapping (CNI)		•	•	•
Average branch length estimation		•	•	•
Bootstrap Test of Phylogeny				
Neighbor-joining/UPGMA	•	•	•	•
Minimum Evolution		•	•	•
Maximum Parsimony		•	•	•
Confidence Probability Test				
Neighbor-joining	•	•	•	•
Minimum Evolution		•	•	•
Consensus tree construction	•	•	•	•
Condensed tree construction	•	•	•	•
Distance Matrix Viewer				
View pairwise distances		•	•	•
View between group distances		•	•	•
View within group distances		•	•	•
View distances and standard errors simultaneously		•	•	•
Sort the distance matrix		•	•	•
Drag-and-drop		•	•	•
Group-wise		•	•	•
By Sequence names		•	•	•
Control display precision		•	•	•
Export Data for printing or re-importing		•	•	•
Tree Explorers				

Phylogeny Display and Graphic printing	•	•	•	•
On-the-spot taxa name editing		•	•	•
Multiple phylogeny views		•	•	•
Linearized Tree		•	•	•
Estimation of divergence time by calibrating molecular clock		•	•	•
Copy to Clipboard/save to file as an EMF drawing		•	•	•
Save to Newick format			•	•
Read trees from Newick format			•	•
User specified control for				
Placement and precision of branch length		•	•	•
Scale bar addition		•	•	•
Collapsing branches or groups		•	•	•
Display only a subtree		•	•	•
Ability to view multiple trees in different viewers		•	•	•
Tree Editing				
Flipping, re-rooting		•	•	•
Add marker symbols to names		•	•	•
Multi-color display and printing		•	•	•
Change Tree Size				
Vertical separation between taxa		•	•	•
Horizontal size		•	•	•
Change Tree shape		•	•	•
Multiple tree display		•	•	•
Save tree session for future display		•	•	•

What you see is what you get printing	•	•	•
Multi- or single page printing	•	•	•
Display images on tree for groups and taxa		•	•

2.23 Using MEGA in the Classroom

Because *MEGA* includes many statistical methods for the study of molecular evolution in an interactive framework, it is instructive for classroom teaching. If you are interested in using *MEGA* in the classroom, there are no restrictions. Your students may download a copy from the website www.megasoftware.net or you may install copies on multiple computers in a common computing area. However, if you want to use *MEGA* in any other form, please contact the authors by e-mail (s.kumar@asu.edu).

If you are using *MEGA 4* in classroom teaching, please send us the following information by e-mail for our records (s.kumar@asu.edu). (1) Your name, position and institution, (2) course number and title, (3) number of students, and (4) course semester and year.

2.24 Technical Support and Updates

All minor (bug fix) and major updates of *MEGA* will be made available at the website www.megasoftware.net. We will send e-mail to all registered *MEGA* users whenever an updated version of the program or the online help manual is made available.

2.25 Reporting Bugs

If you encounter technical problems such as unexplained errors, documentation inconsistencies, or program crashes, please report them to us by e-mail at bugs@megasoftware.net. For further information on reporting problems, consult the bug report page on the MEGA website (www.megasoftware.net). Please note that telephone inquiries will not be accepted.

Please include the following information in your report: (1) your name and address, (2) the version of MEGA you are working with, (3) the version of Windows you are working in, (4) a copy of your data file (if possible), (5) a description of the problem, and (6) the sequence of events that led to that problem

[this often is crucial to understanding and remedying the problem quickly].

2.26 Guide to Notations Used

Item	Convention	Example
Directory & file names	Small Cap + Bold	*INSTALL.TXT*
File name extensions	Small Cap + Bold	*.TXT, .DOC, .MEG*
Email address/URLs	Underlined	www.megasoftware.net
Pop-up help links	Dotted Underlined + Green	`statement`
Help Jumps	Underlined + Green	set of rules
Menu/Screen Items	Italic	*Data* Menu
User-Entered Text	Monospace font	!Title

2.3 A Walk Through MEGA

Introduction to Walk Through MEGA

This section provides a MEGA tutorial. The data files for these examples can be found in the **EXAMPLES** folder, located in the MEGA installation directory (example in **C:\Program Files\MEGA 4\Examples)**. In these example files, data are deliberately written in different input formats. We recommend that you study the examples in the order presented because the techniques explained in the initial examples are used again in the subsequent ones.

In the following write-up, **highlighted** words indicate the keys you must press on the keyboard. If you must press two keys simultaneously, they are shown with a + sign between them (*e.g.*, **Alt** + **F3** means that the **Alt** and **F3** keys should be pressed simultaneously). Italicized letters are used to mark the commands found in menus, submenus, and other locations as they appear on the computer screen. In every example, we discuss many procedures introducing analytical techniques. For ease of reference in later examples, these procedures are numbered in the Ex *u.v.w* format, where *u* is the example number, *v* is the procedure number, and *w* is the step number. For instance, Ex 1.3.2 refers to the 2nd step of the 3rd procedure in example 1.

A list of tutorials is as follows:

1. Aligning Sequences
2. Estimating Evolutionary Distances
3. Building Trees
4. Testing Tree Reliability
5. Marking Genes/Domains
6. Testing for Selection
7. Grouping Sequences
8. Computing Sequence Statistics
9. Trees from Distance Data

2.31 Creating Multiple Sequence Alignments

In this example, we will create an alignment from protein sequence data that will be imported into the alignment editor using different methods.

Ex 1.0.1: Start MEGA 4 by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Ex 1.0.2: Launch the Alignment Explorer by selecting the Alignment|Alignment Explorer/CLUSTAL menu command.

In order to aligning sequences contained in a Sequence Data File, do the following.

Ex 1.1.1: Add unaligned sequences from the **hsp20.fas** example file into the Alignment Explorer by selecting the Data|Open|Get Sequences From File menu command.

Ex 1.1.2: Select the Edit|Select All menu command to select every site for all sequences in the alignment.

Ex 1.1.3: Select the Alignment|Align by ClustalW menu command to align the selected sequences data using the ClustalW algorithm.

Ex 1.1.4: Save the current alignment session by selecting the Data|Save Session menu item. This will allow the current alignment session to be restored for future editing.

Ex 1.1.5: Exit the Alignment Explorer by selecting the Data|Exit Alignment Explorer menu item. A message will appear asking if you would like to save the data to a MEGA file. Choose "YES", and then a "Save As" dialog box will appear. Enter hsp20_aligned.meg as the file name and click the "Save" button. An input box will appear asking for a title for the data. Enter "HSP 20 Aligned by MEGA 4" as

the title and click the "OK" button. Another dialog will appear asking you if the sequence data is protein coding. In this case click "Yes". A final dialog box will appear asking you if you would like to open the data file in MEGA; click "Yes".

Now we examine how to send sequence data from the Internet (Web Explorer) to the Alignment Explorer

Ex 1.2.1: If the Alignment Explorer already contains sequence data select the Data|Create New menu command to create a new alignment. Choose "YES" for each message that appears.

Ex 1.2.2: Activate the Web Explorer tab by selecting Web|Query Gene Banks from the menu.

Ex 1.2.3: When the NCBI Entrez site is loaded, select either the nucleotide or protein database, enter a search term into the search box, and press the "GO" button.

Ex 1.2.4: When the search results are displayed press the "Add to Alignment" button located to the left of the address box. This will display the Web Fetch dialog window.

Ex 1.2.5: Click the box to the left of each accession number whose sequences information you would like to fetch from the web. When you are done selecting accessions press the "Fetch" button.

Ex 1.2.6: When the status column indicates that all sequences are fetched press the "Send to Alignment" button to send the fetched sequence data to the Alignment Explorer.

Ex 1.2.7: Align the fetched data using the steps detailed in Ex 1.1.2 – Ex 1.1.5.

You may also open a trace file in the Trace Data Viewer/Editor and send it directly to the Alignment Explorer.

2.32 Estimating Evolutionary Distances from Nucleotide Sequences

In this example, we will compute various distances for the *Adh* sequences from 11 *Drosophila* species. We will use the data from the previous example to study various sequence statistics. In addition, we will see how these distances can be written in a file in various formats through options for page size, precision, and relative placement of distances and their standard errors.

Ex 2.0.1: Start MEGA 4 by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Activate the data file **Drosophila_Adh.meg** using the instructions given in **Ex 2.2.1 – Ex 2.2.3**.
Now, we begin by computing the proportion of nucleotide differences between each pair of *Adh* sequences.

Ex 2.1.1: Select the *Distance*|*Compute Pairwise* command (**F7**) to display the

distance analysis preferences dialog box.

Ex 2.1.2: In the *Distance Options* tab, click the *Models* pulldown and then select the *Nucleotide|p-distance* option.

Ex 2.1.3: You may look around at the other options but at this moment we will be using the defaults for the remaining options. Click "Compute" to begin the computation.

Ex 2.1.4: A progress indicator will appear briefly and then the distance computation results will be displayed in grid form in a new window.

Now we will compute distances and compare them using other methods.

Ex 2.2.1: Select the *Distance|Compute Pairwise* command. Use the *Models* pulldown to select the *Nucleotide|Jukes-Cantor* method. Now click "OK" to begin the computation.

Ex 2.2.2: Follow the steps Ex. 2.1.1- Ex 2.1.3 and compute the *Tamura Distance*.

Ex 2.2.3: You now have open results windows containing the distances estimated by three different methods, which you can now compare.

Ex 2.2.4: After you've compared the results, select the *File|Quit Viewer* option for each result window.

We have computed nucleotide distances from the nucleotide sequence data in the file Drosophila_Adh.meg. Let us now compute the proportion of amino acid differences. Note that MEGA will automatically translate the nucleotide sequences into amino acid sequences using the selected genetic code table.

Ex 2.3.1: Select the *Distance|Compute Pairwise* command (**F7**) to display the distance analysis preferences dialog box.

Ex 2.3.2: In the *Distance Options* tab, click the *Models* pulldown and then select the *Amino Acid|p-distance* option.

Ex 2.3.3: Click the "OK" button to accept the default values for the rest of the options and begin the computation.

Ex 2.3.4: A progress dialog box will appear briefly. As with the previous nucleotide estimation a results viewer window will be displayed, showing the distances in a grid format.

Ex 2.3.5: After you have inspected the results, use the *File|Quit Viewer* command to close the results viewer. To shut down MEGA, select the *File|Exit* menu command from the main MEGA application window and indicate that you would like to close the data file.

2.33 Constructing Trees and Selecting OTUs from Nucleotide Sequences

The Crab_rRNA.meg file contains nucleotide sequences for the large subunit mitochondrial rRNA gene from different crab species (Cunningham *et al.*

1992). Since the rRNA gene is transcribed but not translated, it is in the category of non-coding genes. Let us use this data file to illustrate the procedures of building trees and in-memory sequence data editing, using the commands present in the *Data and Phylogeny* menus.

Ex 3.0.1: Start MEGA 4 by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Ex 3.1.1: Activate the data file Crab_rRNA.meg using the instructions given in Ex 2.2.1 - Ex 2.2.3.

Let us start by building a neighbor-joining tree.

Ex 3.2.1: Select the *Phylogeny*|*Construct Tree*|*Neighbor-Joining* command to display the analysis preferences dialog box.

Ex 3.2.2: In the Options Summary tab, click the Model pulldown (found in the Substitution Model section) and then select the *Nucleotide*|*p-distance* option.

Ex 3.2.3: Click "OK" to accept the defaults for the rest of the options and begin the computations. A progress indicator will appear briefly, then the tree will be displayed in the Tree Explorer.

Ex 3.2.4: To select a branch, click on it with the left mouse button. If you click on a branch with the right mouse button, you will get a small options menu that will let you flip the branch and perform various other operations on it. To edit the OTU labels, double click on them.

Ex 3.2.5: Change the branch style by selecting the *View*|*Tree/Branch Style* command from the Tree Explorer menu.

Ex 3.2.6: At this time the cursor assumes a triangular shape instead of the diamond (?) shape. Press M and the mirror image of the original tree is displayed instantly. Press M again and the tree reverts to its original shape.

Ex 3.2.7: Press the Up arrow key ( ) just once and the cursor moves upwards to the next branch. Press F, the Flip command, and a mirror-like effect is produced on the sub-tree anchored on the currently focused branch.

Ex 3.2.8: Select the *View*|*Topology Only* command from the Tree Explorer menu and the branching pattern (without actual branch lengths) is displayed on the screen. Press T again and the actual NJ tree reappears.

Ex 3.2.9: Press F1 to examine the help for tree editor. Use the help to become familiar with the many operations that Tree Explorer is capable of performing.

Ex 3.2.10: DO NOT remove the tree from the screen. We shall use it for illustrating how a tree can be printed.

Now you will print the NJ tree that you have on your screen in MEGA.

Ex 3.3.1: Select the *File*|*Print* command from the Tree Explorer menu to bring up a standard Windows print dialog.

Ex 3.3.2: To restrict the size of the printed tree to a single sheet of paper, choose the *File*|*Print in a Sheet* command from the Tree Explorer menu.

Ex 3.3.3: Do not change anything in this dialog box. Select the Preview command using the Tab key and a graphic image of the tree will be displayed on the screen. Press Enter to return to the option box. Now go to the Write information option, and select the Branch lengths. Again, select the Preview command (you may press Alt + V).\ to show the tree drawn with branch lengths. Press Enter to come out of the graphics image.

Ex 3.3.4: Select the *File*|*Exit* Tree Explorer (Ctrl-Q) command to exit the Tree Explorer. A warning box will inform you that your tree data has not been saved. Click the "OK" button to close Tree Explorer without saving the tree session.

In MEGA, you can also construct maximum parsimony (MP) trees. Let us construct a maximum parsimony tree(s) by using the *branch-&-bound search* option.

Ex 3.4.1: Select the *Phylogeny*|*Construct Tree*|*Maximum Parsimony* command. In the resultant preferences window, choose the Max-Mini Branch-&-Bound Search option in the MP Tree Search Options tab.

Ex 3.4.2: Click the "OK" button to accept the defaults for the other options and begin the calculation. A progress window will appear briefly and the tree will be displayed in Tree Explorer.

Ex 3.4.3: Now print this tree (See Ex 3.3.1 - 3.3.2). You do not have to specify the printer name again because *MEGA* remembers your selection.

Ex 3.4.4: Select the *File*|*Exit* Tree Explorer (Ctrl-Q) command to exit the Tree Explorer. A warning box will inform you that your tree data has not been saved. Click "OK" to close Tree Explorer without saving the tree session.

Ex 3.4.5: Compare the NJ and MP trees. For this data set, the branching pattern of these two trees is identical.

As an exercise, use the Heuristic Search for finding the MP tree. In this example, you will find the same tree as that obtained by the branch-and-bound method if you use the default option (search factor equal to 2 for all steps of OTU addition). However, the computational time will be much shorter. Actually, in this example, even a search factor equal to 0 will recover the MP tree.

We will now examine how some data editing features work in MEGA. For noncoding sequence data, OTUs as well as sites can be selected for analysis. Let us remove the first OTU from the current data set.

Ex 4.1: Select the Data|Setup/Select Taxa & Groups command. A dialog box is displayed.

Ex 4.2: All the OTU labels are checked in the left box. This indicates that all OTUs are included in the current active data subset. To remove the first OTU from the data, uncheck the checkbox next to the OTU name in the left pane.

Ex 4.3: Now from this data set construct a neighbor-joining tree (Ex 3.2.3) that contains 12 OTUs instead of 13. To inactivate the operational data set and end the current session of MEGA, press the hot-key Alt + X.

2.34 Tests of the Reliability of a Tree Obtained

In this example, we will conduct two different tests using protein-coding genes from the chloroplast genomes of nine different species.

Ex 4.0.1: Start MEGA 4 by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Activate the data in the Chloroplast_Martin.meg file by using the *File|Open* command.

We will begin with the bootstrap test for the neighbor-joining tree.

Ex 4.1.1: Select the *Tests|Bootstrap Test of Phylogeny|Neighbor Joining Tree* command from the main application menu.

Ex 4.1.2: An analysis preferences dialog box appears. Use the *Models* pulldown to ensure that the *Amino Acid|p-distance* model is selected. Note that only the *Amino Acid* submenu is available.

Ex 4.1.3: Click "OK" to accept the default values for the rest of the options. A progress indicator provides the progress of the test as well as the details of your analysis preferences.

Ex 4.1.4: Once the computation is complete, the Tree Explorer appears and display two tree tabs. The first tab is the original Neighbor-Joining tree and the second is the Bootstrap consensus tree.

Ex 4.1.5: To produce a condensed tree, use the *Compute|Condensed Tree menu* command from the Tree Explorer menu. This tree shows all the branches that are supported at the default cutoff value of $BCL \geq 50$. To change this value, select the *View|Options* menu command and click the *cutoff values* tab. Select the *Compute|Condensed Tree* menu command and the NJ tree will reappear.

Ex 4.1.7: Print this tree. (see Ex 3.3.1 - Ex 3.3.2)

Ex 4.1.8: Select the *File|Exit Tree Explorer* (**Ctrl-Q**) command to exit the Tree Explorer. A warning box will inform you that your tree data has not been saved. Click "OK" to close Tree Explorer without saving the tree session.

For neighbor-joining trees, you may conduct the standard error test for every interior branch by using the *Phylogeny|Neighbor-Joining* command. In MEGA, this test is available for the *p*-distance, Poisson Correction and Gamma distance for amino acid sequences. Since we did the above analysis for the p-distance, we will use the same distance estimation method to compare the results from the bootstrap and standard error tests.

Ex 4.2.1: Go to the *Phylogeny* menu and select the *Neighbor-Joining* command to produce an analysis preferences dialog box. In the *Models* pulldown, be sure that p-distance is the model chosen. Click on the *Test of Phylogeny* tab to reveal the test options. Under the *Test of Inferred Phylogeny* option group, check the *Interior*

Branch Test option.

Ex 4.2.2: Click "OK" to begin the computation. A progress indicator will appear briefly. The neighbor-joining tree with confidence probabilities (*CP*) from the standard error test of branch lengths is displayed on the screen.

Ex 4.2.3: Compare the *CP* values on this tree with the *BCL* values of the tree that you printed in the previous procedure.

Now exit MEGA using the Alt + X command.

2.35 Working With Genes and Domains

Ex 5.0.1: Start MEGA 4 by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Ex 5.0.2: Activate the data present in the Contigs.meg file by using the File|Open command.

We will now examine how to define and edit gene and domain definitions

Ex 5.1.1: Select the Data|Setup/Select Genes & Domains menu command.

Ex 5.1.2: Delete the Data domain by right clicking on it and selecting Delete Gene/Domain from the popup menu.

Ex 5.1.3: Right-click on the Genes/Domains item in the Names column and select Add New Domain. Right-click on the new domain and select Edit Name from the popup menu and set the name to "Exon1".

Ex 5.1.4: Select the ellipsis button next to the question mark in the **From** column to set the first site of the domain. When the site selection window appears select site number 1 and push the "OK" button.

Ex 5.1.5: Select the ellipsis button in the **To** column to set the last site of the domain. When the site selection window appears select site number 3918 and push the "OK" button.

Ex 5.1.6: Check the box in the **Coding** column to indicate that this domain is protein coding.

Ex 5.1.7: Add two more domains to the Genes/Domains item. One of these domains will be named "Intron1" and will begin at site 3919 and end at site 5191. The other will be named "Exon2" and will begin at site 5192 and end at site 8421. Be sure to check the checkbox in the **Coding** column for "Exon2" to indicate a protein-coding domain.

Ex 5.1.8: Right-click on the Genes/Domains item and select Insert New Gene from the popup menu. Change the name of this gene to "Predicted Gene" and click-and-drag all of the domains to this new gene such that they are displayed as children of the "Predicted Gene" node in the display tree.

Ex 5.1.9: Press the "Close" button at the bottom of the window to exit the Gene/Domain manager.

Now we use these domain definitions to restrict analyses when computing pairwise distances.

Ex 5.2.1: Select the *Distances|Compute Pairwise* menu item from the main menu and make sure a Nucleotide model is selected.

Ex 5.2.2: On the *Include Sites* tab make sure that the "Noncoding sites" option does not have a checkmark next to it. Go back to the main menu and press the "Compute" button to begin the analysis.

Ex 5.2.3: When the computation is complete the Distance Explorer will display the pairwise distance computed using only the sequence data from exonic domains of the "Predicted Gene".

2.36 Test of Positive Selection

In this example, we present various analyses of protein-coding nucleotide sequences for five alleles from the human HLA-A locus (Nei and Hughes 1991).

Ex 6.0.1: Start MEGA 4 by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Ex 6.1.1: Activate the data present in the **HLA_3Seq.meg** file by using the *File|Open* command.

Ex 6.1.2: Now that the data file is active, note that various details about the data file are displayed at the bottom of the main application window and more menu items have become available on the main menu.

Let us compute the synonymous and nonsynonymous distances appropriate for studying positive Darwinian selection in this set of antigen recognition codons.

Ex 6.2.1: Select the *Tests|Codon-based Tests of Selection|Z-Test* menu command. An analysis preferences dialog appears. Use the *Models* pulldown in the *Options Summary tab* to select *Syn-Nonsysnonymous|Nei-Gojobori Method|p-distance* model. In the *Test Selection* tab, select *Positive Selection* from the pulldown and select the *Overall Average* analysis type. Click the *Include Sites* tab and make sure that the *Pairwise Deletion* option is selected.

Ex 6.2.2: Click on "OK" to accept the default values for the remaining options. A progress indicator appears briefly; the computation results are displayed in a results window in grid format.

Ex 6.2.3: The *Prob* column contains the probability computed (must be <0.05 for hypothesis rejection at 5% level) and the *Stat* column contains the statistic used to compute the probability. The difference in synonymous and nonsynonymous substitutions should be significant at the 5% level.

Ex 6.2.4: Exit MEGA and deactivate the active data file using the **Alt** + **X** command.

2.37 Managing Taxa With Groups

Ex 7.0.1: Start MEGA 4 by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Ex 7.0.2: Activate the data present in the Crab_rRNA.meg file by using the File|Open command.

We will now examine how to define and edit groups of taxa

Ex 7.1.2: Select the Data|Setup/Select Taxa & Groups menu command.

Ex 7.1.3: Press the "New Group" button found below the Taxa/Groups pane to add a new group to the data. Name this new group "Pagurus".

Ex 7.1.4: While holding the Control button on the keyboard, click on all of the Pagurus species in the Ungrouped Taxa pane to highlight them. When they are all highlighted press the left-facing arrow button found on the vertical toolbar between the two windowpanes.

Ex 7.1.5: Select the "All" group in the Taxa/Groups pane and press the "New Group" button to add a second group. Name this group "Non-Pagurus". Add the remaining unassigned taxa to this group and press the "Close" button at the bottom of the window to exit this view.

Ex 7.1.6: Now that groups have been defined the *Compute Within Group Mean*, *Compute Between Group Means*, and *Compute Net Between Group Means* menu commands from the *Distance* menu item may be used to analyze the data.

2.38 Computing Statistical Quantities for Nucleotide Sequences

In this exercise, we illustrate the use of the *Data Explorer* for computing various statistical quantities of nucleotide sequences. In addition, we explain shortcuts for obtaining frequently used commands, methods of accessing on-line help, and the distinction between enabled and disabled commands.

Ex 8.0.1: Start *MEGA 4* by double-clicking on the *MEGA* desktop icon or by using the Windows start-menu to click on the *MEGA* icon located in the programs folder.

We now will examine the contents of the file Drosophila_Adh.meg by using the built-in *Text Editor*.

Ex 8.1.1: Click on the *File* menu item to expand the menu options. To activate the text editor either click on it or press the **F3** key on your keyboard. In the text editor, use the *File*|*Open* command to open the **Drosophila_Adh.meg** file.

Ex 8.1.2: Examine the **Drosophila_Adh.meg** file to reveal the #mega format specifier, title, OTU names [what is an OTU?], and the interleaved sequence data.

Ex 8.1.3: We advise that you exit the text editor before proceeding with data

analysis. Select the *File* menu item from the text editor's menu and click the *Exit* option from the expanded menu. If the editor asks you if you would like to save the changes that you have made to the file, select *No*.

To study statistical quantities of the data in the file Drosophila_Adh.meg, we must first activate it.

Ex 8.2.1: You can activate a data file using the link in the main application window or select the *File* menu item from the main menu and click the *Open Data* option from the expanded menu. You may also press the **F5** key on your keyboard. All of these methods will display a standard Windows open file dialog box.

Ex 8.2.2: Open the **Drosophila_Adh.meg** data file under the **Examples** folder.

Ex 8.2.3: A progress dialog box will appear briefly. When the data file is active, details about it are displayed at the bottom of the main application window. More menu items now are available on the main menu.

Examine the main menu. Now that the data file is active, the menu items *Data, Distances,* and *Tests* have become available.

We now will use Data Explorer to compute some basic statistics for these data.

Ex 8.3.1: Select the *Data|Data Explorer* command or press the **F4** key.

Ex 8.3.2: DNA sequences are displayed on the screen in a grid format. Use the arrow keys (??) or the mouse to move from site to site; note a change in the bottom-left corner of the *Site#* display. Use the up and down (??) arrow keys or the mouse to move between OTUs. The *Total Sites* view on the bottom-left panel displays the sequence length and the Highlighted *Sites* displays "None" because no special site attributes are yet highlighted.

Ex 8.3.3: To highlight variable sites, select the *Highlight|Variable Sites* option, click the button labeled "V" from the shortcut bar below the menu, or press the **V** key. All sites that are variable are highlighted, and the number in the Highlighted *Sites* display changes. When you press **V** again, the sites return to the normal color and *Highlighted Sites* displays "None."

Ex 8.3.4: Now to highlight the parsimony-informative, press the **P** key, click on the button labeled "Pi" from the shortcut bar below the menu, or select the *Highlight/Parsim-info* menu command. To highlight 0, 2, and 4-fold degenerate sites, press the **0**, **2**, or **4** keys, respectively, click on the corresponding button from the shortcut bar below the menu, or select the corresponding command from the highlight menu.

Ex 8.3.5: To compute the nucleotide base frequencies, select the *Statistics|Nucleotide Composition* menu command. This will calculate the composition and display the results of the calculation in a text file using the built-in text editor.

Ex 8.3.6: To compute codon usage, select the *Statistics|Codon Usage* menu command. This will calculate the codon usage and display the results of the calculation in a text file using the built-in text editor.

Ex 8.3.7: To compute nucleotide pair frequencies, select the *Statistics*|*Nucleotide Pair Frequencies*|*Directional* or the *Statistics*|*Nucleotide Pair Frequencies*|*Unidirectional* menu command. This will calculate the pair frequencies and display the results of the calculation in a text file using the built-in text editor.

Ex 8.3.8: To translate these protein-coding sequences into amino acid sequences and back, press the **T** key or select the *Data*|*Translate/Untranslate* menu command from the Data Explorer menu.

Ex 8.3.9: Once the sequences are translated, calculate the amino acid composition by selecting the *Statistics*|*Amino Acid Composition* menu command from the Data Explorer Menu.

Ex 8.3.10: To shut down MEGA, select the *File*|*Exit* menu command from the main MEGA application window and close the data file.

2.39 Constructing Trees from Distance Data

This example introduces procedures for selecting options from menus, opening files in the read-only mode, activating a distance data file, and building trees from the distance data.

Ex 9.0.1: Start MEGA by double-clicking on the MEGA desktop icon or by using the Windows start-menu to click on the MEGA icon located in the programs folder.

Ex 9.0.2: A *Splash screen* appears, which displays the current version of MEGA.

Ex 9.0.3: This *Splash screen* automatically disappears and the MEGA application becomes available.

In this example, we use the data in the Hum_Dist.meg file. Although we will not edit the file, we will use MEGA's built-in text editor to examine its contents before we proceed further.

Ex 9.2.1: Click on *File* menu to expand the menu options. Click on the menu item labeled *Text Editor* or press on the **F3** key to activate the built-in text editor.

Ex 9.2.2: Use the *Text Editor* to view the contents of the **Hum_Dist.meg** file. To open a file with the *Text Editor* , click on the folder icon below the main menu or on the *File* menu item, then choose *Open* from the expanded menu. You may also use the key combination **Ctrl+O** to open a file. All of these options will lead you to a standard Windows open file dialog box. Use this dialog box to locate the **Examples** folder and open the **Hum_Dist.meg** file. After you open the file with the dialog box, you will see the file contents displayed in the *Text Editor* window.

Ex 9.2.3: Examine the contents of the data file and then exit the *Text Editor* before proceeding with data analysis. Select the *File* menu item from the *Text Editor 's* menu and click *Exit* from the expanded menu. If the editor asks you if you would like to save your changes, select *No*.

A data file must be activated before an analysis can be performed. (Note that opening a file for browsing or editing is different from activating it for analysis.) Now we will activate the Hum_Dist.meg data file.

Ex 9.3.1: You can activate a data file by using the link in the main application window or by selecting the *File* menu item from the main menu and clicking the *Open Data* option from the expanded menu. You also can press the **F5** key on your keyboard. All of these methods will display a standard Windows open file dialog box.

Ex 9.3.2: Use the open file dialog box to locate and open the **Hum_Dist.meg** file located in the **Examples** folder. After you have selected the file for opening, a progress indicator will appear briefly.

Ex 9.3.3: When the data file is active, details about it are displayed at the bottom of the main application window. More menu items now are available on the main menu.

Now we will make a phylogenetic tree from the distance data.

Ex 9.4.1: From the expanded menu in the *Phylogeny* menu, select the *Neighbor-joining* command.

Ex 9.4.2: A confirmation window will appear, indicating that MEGA is ready to conduct the requested analysis. Click on the button labeled "OK;" a progress meter will appear briefly.

Ex 9.4.3: The *Tree Explorer* will instantly display a neighbor-joining tree on the screen. To exit the *Tree Explorer*, select the *File* menu item from the *Tree Explorer* menu and click the *Exit Tree Explorer* option from the expanded menu. The *Tree Explorer* will ask you if you would like to save the tree data. If you save the tree, you can use *Tree Explorer* to view and manipulate it in the future.

With this, let us end this session of MEGA.

Ex 9.5.1: Go to the *Data* menu and click on the *Close Data* command. The program will inquire if you would like the data to be inactivated. Select "Yes."

Ex 9.5.2: To exit MEGA, press **Alt** + **X,** or select the *Exit* command from the expanded *File* menu.

3 Part II: Assembling Data for Analysis

Text File Editor and Format Converter

MEGA includes a *Text File Editor*, which is useful for creating and editing ASCII text files. It is invoked automatically by *MEGA* if the input data file processing modules detect errors in the data file format. In this case, you should make appropriate changes and save the data file.

The text editor is straightforward if you are familiar with programs like Notepad. Click on the section you wish to change, type in the new text, or select text to cut, copy or paste. Only the display font can be used in a document. You can have as many different text editor windows open at one time and you may close them independently. However, if you have a file open in the *Text Editor*, you should save it and close the *Text Editor* window before trying to use that data file for analysis in *MEGA*. Otherwise, *MEGA* may not have the most up-to-date version of the data.

The *Text File Editor and Format converter* is a sophisticated tool with numerous special capabilities that include:

- **Large files** –The ability to operate on files of virtually unlimited size and line lengths.
- **General purpose** –Used to view/edit any ASCII text file.
- **Undo/ReDo** –The availability of an unlimited depth of undo/redo options
- **Search/Replace** –Searches for and does block replacements for arbitrary strings.
- **Clipboard** – Supports familiar clipboard *cut*, *copy*, and *paste* operations.
- **Normal and Column blocks** – Supports regular contiguous line blocks and columnar blocks. This is quite useful while manually aligning sequences in the *Text Editor*.
- **Drag/Drop** – Moves text with the familiar cut and paste operations or you can select the text and then move it with the mouse.
- **Screenshots** –Creates screen snapshots for teaching and documentation purposes directly from the edit window.
- **Printing** –Prints the contents of the edit file.

The *Text Editor* contains a menu bar, a toolbar, and a status bar.

The Menu bar

Menu	Description

File menu	The File Menu contains the functions that are most commonly used to open, save, rename, print, and close files. (Although there is no separate "rename" function available, you can rename a file by choosing the Save As… menu item and giving the file a different name before you save it.)
Edit menu	The Edit Menu contains functions that are commonly used to manipulate blocks of text. Many of the edit menu items interact with the Windows Clipboard, which is a *hidden window* that allows various selections to be copied and pasted across documents and applications.
Search menu	The Search Menu has several functions that allow you to perform searches and replacements of text strings. You can also jump directly to a specific line number in the file.
Display menu	The Display Menu contains functions that affect the visual display of files in the edit windows.
Utilities menu	The Utilities Menu contains several functions that make this editor especially useful for working with files containing molecular sequence data (note that the *MEGA 4* editor does not try to understand the contained data, it simply operates on the text, assuming that the user knows what (s)he is doing.

Toolbar

The Toolbar contains shortcuts to some frequently used menu commands.

Status Bar

The Status bar is positioned at the bottom of the editor window. It shows the position of the cursor (line number and position in the line), whether the file has been edited, and the status of some keyboard keys (CAPS, NUM, and SCROLL lock).

Hotkeys and Shortcut keys

Many menu items have a *hotkey* and/or a *shortcut* key. These are special key combinations that are helpful for people who are more comfortable using a keyboard than the mouse. *Hotkeys* are identified by an underscore character in the name of the menu item, e.g., "File", "New". These allow you to hold down the Alt-key, which is usually found next to the space bar on the keyboard, then hit the underlined letter to produce the same action as if you clicked that name with the mouse. We show this using the notation <Alt>+key – e.g., the hotkey for the file menu item is shown as <Alt>+F. Be sure that you depress both keys together, holding the <Alt> key down a little bit longer than the letter key. (Some people try

hitting both keys simultaneously, as if they're hitting two keys on a piano keyboard. Quite often, this approach does not produce the desired results.)

For instance, you could create a new file by clicking the mouse on the "File" menu item, then clicking on the "New" item beneath it. Using hotkeys, you could type <Alt>+F followed by <Alt>+N. Or, more simply, while you're holding down the <Alt> key, hit the 'F' key followed by the 'N' key, then release the <Alt> key.

You might notice that several menu items, e.g., the New Item on the File menu, show something to the right that looks like 'Ctrl+N'. This is called a *Shortcut* key sequence. Whereas executing a command with hotkeys often requires several keystrokes, shortcut keys can do the same thing with just one keystroke. Shortcut keys work the same as hotkeys, using the <Ctrl> key instead of the <Alt> key. To create a new file, for example, you can hold down the <Ctrl> key and hit the 'N' key, which is shown as <Ctrl>+N here. (In the menus, this appears simply as 'Ctrl+N'.)

Not all menu items have associated shortcut keys because there are only 26 shortcut keys, one for each letter of the alphabet. Hotkeys, in contrast, are localized to each menu and submenu. For hotkeys to work, the menu item must be visible whereas *shortcut keys work at any time.* For instance, if you are typing data into a text file and want to create a note in a new window, you may simply hit the shortcut key sequence, <Ctrl>+N to generate a new window. After you type the note, you can hit <Ctrl>+S to save it, give it a file name, hit the enter key [this part doesn't make sense]; then you can hit the <Alt>+F+C hotkey sequence to close the file (there is no shortcut key for closing a file).

3.1 Trace Data File Viewer/Editor

Using this function, you can view and edit trace data produced by an automated DNA sequencer in ABI and Staden file formats. The sequences displayed can be added directly into the Alignment Explorer or sent to the Web Browser for conducting BLAST searches.

A brief description of various functions available in the Trace Data file Viewer/Editor are as follows:

Data menu

Open File in New Window: Launches a new instance to view/edit another file.

Open File: Allows you to select another file to view/edit in the current window.

Save File: Save the current data to a file in Staden format.

Print: Prints the current trace data, excluding all masked regions.

Add to Alignment Explorer: DNA sequence data, excluding all masked regions, is sent to the *Alignment Explorer* and appears as a new sequence at the end of the current alignment.

Exit: Closes the current window.

Edit menu

Undo: Use this command to Undo one or more previous actions.

Copy: This menu provides options to (1) copy DNA sequences from FASTA or plain text formats to the clipboard and (2) copy the exact portion you are viewing of the currently displayed trace image to the clipboard in the Windows Enhanced Meta File format. For FASTA format copying, both the sequence name and the DNA data will be copied, excluding the masked regions. To copy only the selected portion of the sequence, use the plain text copy command (If nothing is selected, then the plain text command will copy the entire sequence, except for the masked regions).

Mask Upstream: Mask or unmask region to the left (upstream) of the cursor.

Mask Downstream: Mask or unmask region to the right (downstream) of the cursor.

Reverse Complement: Reverse complements the entire sequence.

Search menu

Find: Finds a specified query sequence.

Find Next: Finds the next occurrence of the query sequence. To specify the query sequence, first use the **Find** menu command.

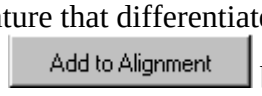
Find Previous: Finds the previous occurrence of a query sequence. To specify the query sequence, first use the **Find** command.

Next N: Go to the next indeterminate (N) nucleotide.

Search in File: This command searches another file, which you specify, for the selected sequence in the current window. It can be used when you are assembling sequence subclones to build a contig.

Do BLAST Search: Launch web browser to BLAST the currently selected sequence. If nothing is selected, the entire sequence, excluding the masked regions, will be used.

3.2 Web Browser

MEGA contains a fully functional *Web Browser* to assist users in sequence data retrieval and web exploration. The most important feature that differentiates this web browser from other browsers (e.g., Netscape or Internet Explorer) is the 

button. Pressing this button causes the *MEGA* web explorer to extract sequence data from the currently displayed web page and send it to the Alignment Explorer 's alignment grid, where it will be inserted as new sequences. At present, the *MEGA* web browser can interpret data displayed in FASTA format or in the default format at the NCBI website. (You can ask the NCBI website to display the data in the FASTA format by using the **Display** option on the web page shown.) (We plan to enhance this functionality further in version 3.1.)

Furthermore, the *MEGA* web browser provides a genomics database, exploration oriented interface for web searching. (In fact this is almost the same functionality as in the most recent versions of the Internet Explorer.)

This causes the web browser window to navigate back to the web location found before the current site in the explorer location history.

This causes the web browser window to navigate forward to the web location found after the current site in the explorer location history.

This causes the web browser to terminate loading a web location.

This causes the web browser to reload the current web location.

This causes the web browser to extract sequence data from the current web page and send it to *Alignment Builder's* alignment grid as new sequence rows. If the web explorer is unable to find properly formatted sequence data in the current web page a warning box will appear.

Address Field

The web location, or address field, is located in the second toolbar. This field contains the URL of the current web location as well as a pull down list of previously visited URLs. If a new URL is entered into the box and the **Enter** key is pressed, the web explorer will attempt to navigate to the entered URL.

Links

This toolbar provides shortcuts to a selection of websites.

There are number of menus in the web browser, including **Data**, **Edit**, **View**, **Links**, **Go**, and **Help**. These menus provide access to routine functionalities, which are self-explanatory in use.

3.3 Some Text Editor Utilities

3.31 Open Saved Alignment Session

Alignment | Open Saved Alignment Session…

Use this command to display a previously saved Alignment Explorer session (saved in a filename with .**MAS** extension).

3.32 Copy Screenshot to Clipboard

Utilities | *Copy Screenshot to Clipboard*

This item presents three other options for selecting the format of an image that is being copied to the clipboard. Once it is copied, it can be pasted in any other graphic or word processing program.

Bitmap Format: This is the common Windows Bitmap (BMP) Format.

Windows Metafile Format: This selects the Windows Metafile Format (WMF)

Enhanced Metafile Format: This selects the Windows Enhanced Metafile Format.

3.33 Format Selected Sequence

Utilities | *Convert to Mega Format*

This submenu presents four other menu items that offer some common ways of reformatting text.

Merge Multiple Lines: This is used to merge several separate lines into one long (very wide) line

Remove Spaces/Digits: This is used to remove spaces and digits from a genetic sequence.

Insert Spaces Every 3: This is used to break the selected text into three-character chunks (e.g., codons). Note that it does not remove any already existing spaces.

Insert Spaces Every 10: This is used to break the selected text into ten-character chunks.

3.34 Reverse Complement

Utilities | *Reverse Complement*

This item reverses the order of characters in the selected block and then replaces each character by its complement. Only A, T, U, C, and G are complemented; the rest of the characters are left as they are. Please use it carefully as *MEGA 4* does not validate whether the characters in the selected block are nucleotides.

3.35 Convert to Mega Format (in Text Editor)

Utilities | *Convert to Mega Format*

This item converts the sequence data in the current edit window, or in a selected file, into a *MEGA* format file. It brings up a dialog box, which allows you to choose the file and/or the format for this purpose. *MEGA 4* converts the data file and displays the converted data in the editor.

Files written in a number of popular data formats can be converted into MEGA

format. *MEGA 4* supports the conversion of CLUSTAL, NEXUS (PAUP, MacClade), PHYLIP, GCG, FASTA, PIR, NBRF, MSF, IG, and XML formats. Details about how MEGA reads and converts these file formats are given in the section Importing Data from Other Formats.

3.4 Building Sequence Alignments

3.41 Alignment Explorer

The *Alignment Explorer* provides options to (1) view and manually edit alignments and (2) generate alignments using a built-in CLUSTALW implementation (for the complete sequence or data in any rectangular region). The *Alignment Explorer* also provides tools for exploring web-based databases (e.g., NCBI Query and BLAST searches) and retrieving desired sequence data directly into the current alignment.

The *Alignment Explorer* has the following menus in its main menu: *Data, Edit, Search, Alignment, Web, Sequencer, Display, and Help.* In addition, there are Toolbars that provide quick access to many *Alignment Explorer* functions. The main *Alignment Explorer* window contains up to two alignment grids.

For amino acid input sequence data, the *Alignment Explorer* provides only one view. However, it offers two views of DNA sequence data: the DNA Sequences grid and the Translated Protein Sequences grid. These two views are present in alignment grids in the two tabs with each grid displaying the sequence data for the current alignment. Each row represents a single sequence and each column represents a site. A "*" character is used to indicate site columns, exhibiting consensus across all sequences. An entire sequence may be selected by clicking on the gray sequence label cell found to the left of the sequence data. An entire site may be selected by clicking on the gray cell found above the site column. The alignment grid has the ability to assign a unique color to each unique nucleotide or amino acid and it can display a background color for each cell in the grid. This behavior can be controlled from the *Display* menu item found in the main menu. Please note that when the ClustalW alignment algorithm is initiated, it only will align the sites currently selected in the alignment grids. Multiple sites may be selected by clicking and then dragging the mouse within the grid. Note that all of the manual or automatic alignment procedures carried out in the Protein Sequences grid will be imposed on the corresponding DNA sequences as soon as you flip to the DNA sequence grid. Even more importantly, the *Alignment Explorer* provides unlimited UNDO capabilities.

3.42 Aligning coding sequences via protein sequences

MEGA 3 provides a convenient method for aligning coding sequences based on the alignment of protein sequences. In order to accomplish this you use the *Alignment Explorer* to load a data file containing protein-coding sequences. If you click on

the *Translated Protein Sequences* tab you will see that the protein-coding sequences are automatically translated into their respective protein sequence. With this tab active select the Alignment|Align by ClustalW menu item or click on the "W" tool bar icon to begin the alignment of the translated protein sequences. Once the alignment of the translated protein sequences completes, click on the *DNA Sequences* tab and you'll find that *Alignment Explorer* automatically aligned the protein-coding sequences according to the aligned translated protein sequences. Any manual adjustments made to the translated protein sequence alignment will also be reflected in the protein-coding sequence tab.

3.43 CLUSTALW

About CLUSTALW

ClustalW is a widely used system for aligning any number of homologous nucleotide or protein sequences. For multi-sequence alignments, ClustalW uses progressive alignment methods. In these, the most similar sequences, that is, those with the best alignment score, are aligned first. Then progressively more distant groups of sequences are aligned until a global alignment is obtained. This heuristic approach is necessary because finding the global optimal solution is prohibitive in both memory and time requirements. ClustalW performs very well in practice. The algorithm starts by computing a rough distance matrix between each pair of sequences based on pairwise sequence alignment scores. These scores are computed using the pairwise alignment parameters for DNA and protein sequences. Next, the algorithm uses the neighbor-joining method with midpoint rooting to create a guide tree, which is used to generate a global alignment. The guide tree serves as a rough template for clades that tend to share insertion and deletion features. This generally provides a close-to-optimal result, especially when the data set contains sequences with varied degrees of divergence, so the guide tree is less sensitive to noise.

See:

Higgins D., Thompson J., Gibson T. Thompson J. D., Higgins D. G., Gibson T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.
Nucleic Acids Res. 22:4673-4680. (1994)

CLUSTALW Options (DNA)

This dialog box displays a single tab containing a set of organized parameters that are used by ClustalW to align the DNA sequences. If you are aligning protein-coding sequences, please note that CLUSTALW **will not** respect the codon positions and may insert alignment gaps within codons. For aligning cDNA or sequence data containing codons, we recommend that you align the translated protein sequences (see Aligning coding sequences via protein sequences).

In this dialog box, you will see the following options:

Parameters for Pairwise Sequence Alignment

Gap Opening Penalty: The penalty for opening a gap in the alignment. Increasing this value makes the gaps less frequent.

Gap Extension Penalty: The penalty for extending a gap by one residue. Increasing this value will make the gaps shorter. Terminal gaps are not penalized.

Parameters for Multiple Sequence Alignment

Gap Opening Penalty: The penalty for opening a gap in the alignment. Increasing this value makes the gaps less frequent.

Gap Extension Penalty: The penalty for extending a gap by one residue. Increasing this value will make the gaps shorter. Terminal gaps are not penalized.

Common Parameters

DNA Weight Matrix: The scores assigned to matches and mismatches (including IUB ambiguity codes).

Transition Weight: Gives transitions a weight between 0 and 1. A weight of zero means that the transitions are scored as mismatches, while a weight of 1 gives the transitions the match score. For distantly-related DNA sequences, the weight should be near zero; for closely-related sequences, it can be useful to assign a higher score.

Use Negative Matrix: Enabled negative weight matrix values will be used if they are found; otherwise the matrix will be automatically adjusted to all positive values.

Delay Divergent Cutoff (%): Delays the alignment of the most distantly-related sequences until after the most closely-related sequences have been aligned. The setting shows the percent identity level required to delay the addition of a sequence. Sequences that are less identical than this level will be aligned later.

Keep Predefined Gaps: When checked, alignment positions in which ANY of the sequences have a gap will be ignored.

NOTE: All Definitions are derived from the CLUSTALW manual.

CLUSTALW Options (Protein)

This dialog box displays a single tab containing a set of organized parameters that are used by ClustalW to align DNA sequences. If you are aligning protein-coding sequences, please note that CLUSTALW **will not** respect the codon positions and may insert alignment gaps within codons. For aligning cDNA or sequence data containing codons, we recommend that you align the translated protein sequences (see Aligning coding sequences via protein sequences).

In this dialog box, you will see the following options:

Parameters for Pairwise Sequence Alignment

Gap Opening Penalty: The penalty for opening a gap in the alignment. Increasing this value makes the gaps less frequent.

Gap Extension Penalty: The penalty for extending a gap by one residue. Increasing this value will make the gaps shorter. Terminal gaps are not penalized.

Parameters for Multiple Sequence Alignment

Gap Opening Penalty: The penalty for opening a gap in the alignment. Increasing this value makes the gaps less frequent.

Gap Extension Penalty: The penalty for extending a gap by one residue. Increasing this value will make the gaps shorter. Terminal gaps are not penalized.

Common Parameters

DNA Weight Matrix: The scores assigned to matches and mismatches (including IUB ambiguity codes).

Residue-specific Penalties: Amino acid specific gap penalties that reduce or increase the gap opening penalties at each position or sequence in the alignment. For example, positions that are rich in glycine are more likely to have an adjacent gap than positions that are rich in valine. See the documentation for details.

Hydrophilic Penalties: Used to increase the chances of a gap within a run (5 or more residues) of hydrophilic amino acids; these are likely to be loop or random coil regions in which gaps are more common.

Gap Separation Distance: Tries to decrease the chances of gaps being too close to each other. Gaps that are less than this distance apart are penalized more than other gaps. This does not prevent close gaps; it makes them less frequent, promoting a block-like appearance of the alignment.

Use Negative Matrix: When enabled negative weight matrix values will be used if they are found; otherwise the matrix will be automatically adjusted to all positive values.

Delay Divergent Cutoff (%): Delays the alignment of the most distantly-related sequences until after the alignment of the most closely-related sequences. The setting shows the percent identity level required to delay the addition of a sequence; sequences that are less identical than this level will be aligned later.

Keep Predefined Gaps: When checked, any alignment positions in which ANY of the sequences have a gap will be ignored.

NOTE: All definitions are derived from CLUSTALW manual.

3.44 BLAST

About BLAST

BLAST is a widely used tool for finding matches to a query sequence within a large sequence database, such as Genbank. BLAST is designed to look for local alignments, *i.e.* maximal regions of high similarity between the query sequence and the database sequences, allowing for insertions and deletions of sites. Although the optimal solution to this problem is computationally intractable, BLAST uses carefully designed and tested heuristics that enable it to perform searches very rapidly (often in seconds). For each comparison, BLAST reports a goodness score and an estimate of the expected number of matches with an equal or higher score than would be found by chance, given the characteristics of the sequences. When this expected value is very small, the sequence from the database is considered a "hit" and a likely homologue to the query sequence. Versions of BLAST are available for protein and DNA sequences and are made accessible in *MEGA* via the Web Browser.

See:

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

Do BLAST Search

Alignment | Do BLAST Search

Use this to launch the BLAST search in the *MEGA* Web Browser. The web-browser is displayed with the BLAST facility at the NCBI website.

3.45 Menu Items in the Alignment Explorer

Toolbars in Alignment Explorer

Basic Functions

This prepares *Alignment Builder* for a new alignment. Any sequence data currently loaded into *Alignment Builder* is discarded.

This activates the *Open File* dialog window. It is used to send sequence data from a properly formatted file into *Alignment Builder*.

This activates the *Save Alignment Session* dialog window. It may be used to save the current state of the *Alignment Builder* into a file so that it may be restored in the future.

This causes nucleotide sequences currently loaded into *Alignment Builder* to be translated into their respective amino acid sequences.

Web/Data Explorer Functions

This displays the NCBI BLAST web site in the Web Explorer tab window. If a sequence in the sequence grid is selected prior to clicking this button, the Web Explorer will auto-fill the BLAST query window with the selected sequence data.

This displays the default database (GenBank) in the Web Explorer tab window.

This activates the *Open Trace File* dialog window, which may be used to open and view a sequencer file. The sequence data from the sequencer file then can be sent into *Alignment Explorer*.

Alignment Functions

This displays the ClustalW parameters dialog window, which is used to configure ClustalW and initiate the alignment of the selected sequence data. If you do not select sequence data prior to clicking this button, a message box will appear asking if you would like to select all of the currently loaded sequences.

This marks or unmarks the currently selected single site in the alignment grid. Each sequence in the alignment may have only one site marked at a time. Modifications can be made to the alignment by marking two or more sites and then aligning them using the *Align Marked Sites* function.

This button aligns marked sites. Two or more sites must be marked in order for this function to have an effect.

Search Functions

This activates the *Find Motif* search box. When this box appears, it asks you to enter a motif sequence (a small subsequence of a larger sequence) as the search term. After the search term is entered, the *Alignment Builder* finds each occurrence of the search term and indicates it with yellow highlighting. For example, if you were to enter the motif "AGA" as the search term, then each occurrence of "AGA" across all sequences in the sequence grid would be highlighted in yellow.

This searches towards the beginning of the current sequence for the first occurrence of the motif search term. If no motif search has been performed prior to clicking this button, the *Find Motif* search box will appear.

This searches towards the end of the current sequence for the first occurrence of the motif search term. If no motif search has been performed prior to clicking this button, the *Find Motif* search box will appear.

This locates the marked site in the current sequence. If no site has been marked, a warning box will appear.

Editing Functions

This undoes the last *Alignment Builder* action.

This copies the current selection to the clipboard. It may be used to copy a single base, a block of bases, or entire sequences to the clipboard.

This removes the current selection from the *Alignment Builder* and sends it to the clipboard. This function can affect a single base, a block of bases, or entire sequences.

This pastes the contents of the clipboard into the *Alignment Builder*. If the clipboard contains a block of bases, it will be pasted into the builder starting at the point of the current selection. If the clipboard contains complete sequences they will be added to the current alignment. For example, if the contents of a FASTA file were copied to the clipboard from a web browser, it would be pasted into *Alignment Builder* as a new sequence in the alignment.

This deletes a block of selected bases from the alignment grid.

This deletes gap-only sites (sites containing a gap across all sequences in the alignment grid) from a selected block of bases.

Sequence Data Insertion Functions

This creates a new, empty sequence row in the alignment grid. A label and sequence data must be provided for this new row.

This activates an *Open File* dialog box that allows for the selection of a sequence data file. Once a suitable sequence data file is selected, its contents will be imported into *Alignment Builder* as new sequence rows in the alignment grid.

Site Number display on the status bar

Site # The **Site #** field indicates the site represented by the current selection. If the **w/o Gaps** radio button is selected, then the *Alignment Builder* will disregard the shifting affect of gaps when determining gap sites. If a block of sites are selected, then this field will contain the site # for the first site in the block. If an entire sequence is selected this field will contain the site # for the last site in the sequence.

Alignment Menu (in Alignment Explorer)

This menu provides access to commands for editing the sequence data in the alignment grid. The commands are:

Align by ClustalW: This option is used to align the DNA or protein sequence included in the current selection on the alignment grid. You will be prompted for the alignment parameters (DNA or Protein) to be used in ClustalW; to accept the parameters, press "OK". This initiates the ClustalW alignment system. *Alignment Builder* then aligns the current selection in the alignment grid using the accepted parameters.

Mark/Unmark Site: This marks or unmarks a single site in the alignment grid. Each sequence in the alignment may only have one site marked at a time. Modifications can be

made to the alignment by marking two or more sites and then aligning them using the *Align Marked Sites* function.

Align Marked Sites: This aligns marked sites. Two or more sites in the alignment must be marked for this function to have an effect.

Unmark All Sites: This item unmarks all currently marked sites across all sequences in the alignment grid.

Delete Gap-Only Sites: This item deletes gap-only sites (site columns containing gaps across all sequences) from the alignment grid.

Auto-Fill Gaps: If this item is checked, then the *Alignment Builder* will ensure that all sequences in the alignment grid are the same length by padding shorter sequences with gaps at the end.

Display Menu (in Alignment Explorer)

This menu provides access to commands that control the display of toolbars in the alignment grid. The commands in this menu are:

Toolbars: This contains a submenu of the toolbars found in *Alignment Explorer*. If an item is checked, then its toolbar will be visible within the *Alignment Explorer* window.

Use Colors: If checked, *Alignment Explorer* displays each unique base using a unique color indicating the base type.

Background Color: If checked, then *Alignment Explorer* colors the background of each base with a unique color that represents the base type.

Font: The *Font* dialog window can be used to select the font used by *Alignment Explorer* for displaying the sequence data in the alignment grid.

Edit Menu (in Alignment Explorer)

This menu provides access to commands for editing the sequence data in the alignment grid. The commands in this menu are:

Undo: This undoes the last *Alignment Explorer* action.

Copy: This copies the current selection to the clipboard. It may be used to copy a single base, a block of bases, or entire sequences.

Cut: This removes the current selection from the *Alignment Explorer* and sends it to the clipboard. This function can affect a single base, a block of bases, or entire sequences.

Paste: This pastes the contents of the clipboard into the *Alignment Explorer*. If the clipboard contains a block of bases, they will be pasted into the builder, starting at the point of the current selection. If the clipboard contains complete sequences, they will be added to the current alignment. For example, if the contents of a FASTA file are copied from a web browser to the clipboard, they will be pasted

into the *Alignment Explorer* as a new sequence in the alignment.

Delete: This deletes a block of selected bases from the alignment grid.

Delete Gaps: This deletes gaps from a selected block of bases.

Insert Blank Sequence: This creates a new, empty sequence row in the alignment grid. A label and sequence data must be provided for this new row.

Insert Sequence From File: This activates an *Open File* dialog box that allows for the selection of a sequence data file. Once a suitable sequence data file is selected, its contents will be imported into *Alignment Explorer* as new sequence rows in the alignment grid.

Select Site(s): This selects the entire site column for each site within the current selection in the alignment grid.

Select Sequences: This selects the entire sequence for each site within the current selection in the alignment grid.

Select all: This selects all of the sites in the alignment grid.

Allow Base Editing: If this item is checked, it changes the base values for all cells in the alignment grid. If it is not checked, then all bases in the alignment grid are treated as read-only.

Data Menu (in Alignment Explorer)

This menu provides commands for creating a new alignment, opening/closing sequence data files, saving alignment sessions to a file, exporting sequence data to a file, changing alignment sequence properties, reverse complimenting sequences in the alignment, and exiting *Alignment Explorer*. The commands in this menu are:

Create New Alignment: This tells *Alignment Explorer* to prepare for a new alignment. Any sequence data currently loaded into *Alignment Builder* is discarded.

Open: This submenu provides two options: opening an existing sequence alignment session (previously saved from *Alignment Explorer*), and reading a text file containing sequences in one of many formats (including, MEGA, PAUP, FASTA, NBRF, etc.). Based on the option you choose, you will be prompted for the file name that you wish to read.

Close: This closes the currently active data in the *Alignment Explorer*.

Save Session: This allows you to save the current sequence alignment to an alignment session. You will be requested to give a file name to write the data to.

Export Alignment: This allows you to export the current sequence alignment to a file. You can choose to export the file to MEGA or FASTA formats.

DNA Sequences: Use this item to specify that the input data is DNA. If DNA is

selected, then all sites are treated as nucleotides. The *Translated Protein Sequences* tab contains the protein sequences. If the data is non-coding, then ignore the second tab, as it has no affect on the on the DNA sequence tab. However, any changes you make in the *Protein Sequence* tab are applied to the *DNA Sequences* tab window. Note that you can UNDO these changes by using the undo button.

Protein Sequences: Use this item to specify that the input data is amino acid sequences. If selected, then all sites are treated as amino acid residues.

Translate/Untranslate: This item only will be available if protein-coding DNA sequences are available in the alignment grid. It will translate protein-coding DNA sequences into their respective amino acid sequences using the selected genetic code table.

Select Genetic Code Table: This displays the *Select Genetic Code* dialog window, which can select the genetic code table that is used when translating protein-coding DNA sequence data.

Reverse Complement: This becomes available when an entire sequence of row(s) is selected. It will update the selected rows to contain the reverse compliment of the originally selected sequence(s).

Exit Alignment Explorer: This closes the *Alignment Explorer* window and returns to the main *MEGA* application window. When selected, a message box appears asking if you would like to save the current alignment session to a file. Then a second message box appears asking if you would like to save the current alignment to a *MEGA* file. If the current alignment is saved to a *MEGA* file, a third message box will appear asking if you would like to open the saved *MEGA* file in the main *MEGA* application.

Search Menu (in Alignment Explorer)

This menu allows searching for sequence motifs and marked sites. The commands in this menu are:

Find Motif: This activates the *Find Motif* search box. When this box appears, it asks you to enter a motif sequence (a small subsequence of a larger sequence) as the search term. After you enter the search term, the *Alignment Explorer* finds each occurrence of it and indicates it with yellow highlighting. For example, if you enter the motif "AGA" as the search term, then each occurrence of "AGA" across all sequences in the sequence grid would be highlighted in yellow.

Find Next: This searches for the first occurrence of the motif search term towards the end of the current sequence. If no motif search has been performed prior to clicking this button, the *Find Motif* search box will appear.

Find Previous: This searches towards the beginning of the current sequence for the first occurrence of the motif search term. If no motif search has been performed prior to clicking this button, the *Find Motif* search box will appear.

Find Marked Site: This locates the marked site in the current sequence. If no site

has been marked for this sequence, a warning box will appear.

Highlight Motif: If this item is checked, then all occurrences of the text search term (motif) are highlighted in the alignment grid.

Sequencer Menu (in Alignment Explorer)

Edit Sequencer File: This item displays the *Open File* dialog box used to open a sequencer data file. Once opened, the sequencer data file is displayed in the Trace Data File Viewer/Editor. This editor allows you to view and edit trace data produced by the automated DNA sequencer. It reads and edits data in ABI and Staden file formats and the sequences displayed can be added directly into the Alignment Explorer or send to the Web Browser for conducting BLAST searches.

Web Menu (in Alignment Explorer)

This menu provides access to commands for querying GenBank and doing a BLAST search, as well as access to the MEGA web Browser. The commands in this menu are:

Query Gene Banks: This item starts the Web Browser and accesses the NCBI home page (http://www.ncbi.nlm.nih.gov).

Do BLAST Search: This item starts the Web Browser and accesses the NCBI BLAST query page. If you select a sequence in the alignment grid prior to selecting this item, the web browser will automatically copy the selected sequence data into the search field.

Show Browser: This item will show the Web Browser.

4 Part III: Input Data Types and File Format

4.1 MEGA Input Data Formats

4.11 MEGA Format

For *MEGA* to read and interpret your data correctly, it should be formatted according to a set of rules. All input data files are basic ASCII-text files, which may contain DNA sequence, protein sequence, evolutionary distance, or phylogenetic tree data. Most word processing packages (e.g., Microsoft Word, WordPerfect, Notepad, WordPad) allow you to edit and save ASCII text files, which are usually marked with a **.TXT** extension. After creating the file, you should change this extension to **.MEG**, so that you can distinguish between your data files and the other text files. Because the organizational details vary for different types of data, we discuss the data formats for molecular sequences, distances, and phylogenetic trees separately. However, there are a number of features that are common to all *MEGA* data files.

4.12 General Conventions

Common Features

The first line must contain the keyword `#MEGA` to indicate that the data file is in *MEGA* format. The data file may contain a succinct description of the data (called Title) included in the file on the second line. The `Title statement` is written according to a set of rules and is copied from *MEGA* to every output file. In the long run, an informative title will allow you to easily recognize your past work.

The data file may also contain a more descriptive multi-line account of the data in the `Description` statement, which is written after the `Title` statement. The `Description` statement also is written according to a set of rules. Unlike the Title statement, the Description statement is not copied from *MEGA* to every output file.

In addition, the data file may also contain a `Format` statement, which includes information on the type of data present in the file and some of its attributes. The `Format` statement should be generally written after the `Title` or the `Description` statement. Writing a format statement requires knowledge of the keywords used to identify different types of data and data attributes.

All taxa names must be written according to a set of rules.

Comments can be written anywhere in the data file and can span multiple lines. They must always be enclosed in square brackets ([and]) brackets and can be nested.

Writing Comments

Comments can be placed anywhere in the data file as long as they are contained within a pair of square brackets [like this]. Nested comments are allowed [[like] this].

Key Words

MEGA supports a number of keywords, in addition to MEGA and TITLE, for writing instructions in the format and command statements. These key words can be written in any combination of lower- and upper-case letters. For writing instructions, follow the style given in the examples along with the keyword description for different types of data.

Rules for Taxa Names

Distance matrices as well as sequence data may come from species, populations, or individuals. These evolutionary entities are designated as OTUs (Operational Taxonomic Units) or taxa. Each taxon must have an identification tag, i.e., a taxon Iabel. In the input files prepared for use in *MEGA*, these labels should be written according to the following conventions:

'#' Sign

Every Iabel must be written on a new line, and a '#' sign must precede the label. There are no restrictions on the length of the Iabels in the datafile, but *MEGA 4* will truncate all labels longer than 40 characters. These labels are not required to be unique, although identical labels may result in ambiguities and should be avoided.

Characters to use in labels

Taxa labels must start with alphanumeric characters (0-9, a-z, and A-Z) or a special character: dash (-), plus (+) or period (.). After the first character, taxa labels may contain the following additional special characters: underscore (_), asterisk (*), colon (:), round open and close brackets ( ), vertical line (|),back slash (\), and forward slash (/).

For multiple word labels, an underscore can be used to represent a blank space. All underscores are converted into blank spaces, and subsequent displays of the labels show this change. For example, E._coli becomes E. coli.

Rules for Title Statement

A Title statement must be written on the line following the `#mega`. It always begins with `!Title` and ends with a semicolon.

```
#mega
!Title  This is an example title;
```

A title statement may not occupy more than one line of text. It must not contain a semicolon inside the statement, although it must contain one at the end of the statement.

Rules for Description Statement

A `Description` statement is written after the Title statement. It always begins with `!Description` and ends with a semicolon.

```
#mega

!Title  This is an example title;
!Description  This is detailed information the data file;
```

A description statement may occupy multiple lines of text. It must not contain a semicolon inside the statement, although it must contain one at the end of the statement.

Rules for the Format Statement

A format statement contains one or more command statements. A command statement contains a command and a valid setting keyword (command=keyword format). For example, the command statement `DataType=Nucleotide` tells *MEGA* that nucleotide sequence data is contained in the file. Based on the DataType setting, different types of keywords are valid.

Keywords for Sequence Data

Keywords for Distance Data

Keywords for Tree Data

4.13 Sequence Input Data

General Considerations (Sequence Data)

The sequence data must consist of two or more sequences of equal length. All sequences must be aligned and you may use the in-built alignment system for this purpose. Nucleotide and amino acid sequences should be written in IUPAC single-letter codes. Sequences can be written in any combination of upper- and lower-case letters. Special symbols for alignment gaps, missing data, and identical sites also can be included in the sequences.

Special Symbols

Blank spaces and tabs are frequently used to format data files, so they are simply ignored by *MEGA*. ASCII characters such as the period (**.**), dash (**-**), and question mark (**?**), are generally used as special symbols to represent identity to the first sequence, alignment gaps, and missing data, respectively.

IUPAC single letter codes

Nucleotide or amino acid sequences should be written in IUPAC single-letter codes. The single-letter codes supported in *MEGA* are as follows.

Symbols	Name	Remarks
DNA/RNA		
A	Adenine	Purine
G	Guanine	Purine
C	Cytosine	Pyrimidine

T	Thymine	Pyrimidine
U	Uracil	Pyrimidine
R	Purine	A or G
Y	Pyrimindine	C or T/U
M		A or C
K		G or T
S	Strong	C or G
W	Weak	A or T
H	Not G	A or C or T
B	Not A	C or G or T
V	Not U/T	A or C or G
D	Not C	A or G or T
N	Ambiguous	A or C or G or T

Protein

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic Acid	Asp
E	Glutamic Acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met

N	Asparagine	Asn
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
V	Valine	Val
W	Tryptophan	Trp
Y	Tyrosine	Tyr
*	Termination	*

Keywords for Format Statement (Sequence data)

Command	**Setting**	**Remark**	**Example**
DataType	DNA, RNA, nucleotide, protein	Specifies the type of data in the file	`DataType`=DNA
NSeqs	A count	Number of sequences	`NSeqs`=85
NTaxa	A count	Synonymous with `NSeqs`	`NTaxa`=85
NSites	A count	Number of nucleotides or amino acids	`Nsites`=4592
Property	Exon, Intron, Coding, Noncoding, and End.	Specifies whether a domain is protein coding. `Exon` and `Coding` are synonymous, as are `Intron` and `Noncoding`. `End` specifies that the domain with the given name ends at	`Property`=cyt_b

		this point.	
Indel	single character	Use dash (-) to identify insertion/deletions in sequence alignments	`Indel` = -
Identical	single character	Use period (.) to show identify with the first sequence.	`Identical` = .
Match Char	single character	Synonymous with the `identical` keyword.	`MatchChar` = .
Missing	single character	Use a question mark (?) to indicate missing data.	`Missing` = ?
CodeTable	A name	This instruction gives the name of the code table for the protein coding domains of the data	`CodeTable` = Standard

Defining Genes and Domains

Writing Command Statements for Defining Genes and Domains

The *MEGA* format easily can designate genes and domains within the molecular sequence data. In this format, attributes of different sites (and groups of sites, termed domains) are specified within the data "on the spot" rather than in an attributes block before or after the actual data, as is the case in some other data formats. An example of a three-sequence dataset written in *MEGA* format is shown below. The sequences consist of three genes named FirstGene, SecondGene, and ThirdGene for two groups of organisms *Setup/Select Genes/Domain* (Mammals and Birds). (Note that the genes and domains can also be defined interactively through a dialog box.)

```
!Gene=FirstGene  Domain=Exon1 Property=Coding;

#Human_{Mammal}  ATGGTTTCTAGTCAGGTCACCATGATAGGTCTCAAT

#Mouse_{Mammal}  ATGGTTTCTAGTCAGGTCACCATGATAGGTCCCAAT

#Chicken_{Aves}  ATGGTTTCTAGTCAGCTCACCATGATAGGTCTCAAT


!Gene=SecondGene  Domain=Intron Property=Noncoding;

#Human      ATTCCCAGGGAATTCCCGGGGGGTTTAAGGCCCCTTTAAAGAAAGAT
```

```
#Mouse      GTAGCGCGCGTCGTCAGAGCTCCCAAGGGTAGCAGTCACAGAAAGAT
#Chicken    GTAAAAAAAAAAGTCAGAGCTCCCCCCAATATATATCACAGAAAGAT

!Gene=ThirdGene  Domain=Exon2  Property=Coding;
#Human       ATCTGCTCTCGAGTACTGATACAAATGACTTCTGCGTACAACTGA
#Mouse       ATCTGATCTCGTGTGCTGGTACGAATGATTTCTGCGTTCAACTGA
#Chicken     ATCTGCTCTCGAGTACTGCTACCAATGACTTCTGCGTACAACTGA
```

Keywords for Command Statements (Genes/Domains)

Command	Setting	Remark	Example
`Domain`	A name	This instruction defines a domain with the given name	`Domain`=first_exon
`Gene`	A name	This instruction defines a gene with the given name	`Gene`=cytb
`Property`	Exon, Intron, Coding, Noncoding, and End.	This instruction specifies the protein-coding attribute for a domain. Keywords `Exon` and `Coding` are synonymous; similarly `Intron` and `Noncoding` are synonymous. `End` specifies the domain in which the given name has ended.	`Property`=cytb
`CodonStart`	A number	This instruction specifies the site where the next 1st-codon position will be found in a protein-coding domain.	`CodonStart`=2

Defining Groups

Writing Command Statements for Defining Groups of Taxa

The *MEGA* format allows you to assign different taxa to groups in a sequence as well as to distance data files. In this case, the name of the group is written in a set of curly brackets following the taxa name. The group name can be attached to the taxa name using an underscore or just can be appended. It is important to note that there should be no spaces

between the taxa name and group name. (Note that the groups of taxa can also be defined interactively through a dialog box.) In the following, we show an example in which human and mouse are designated as the members of the group Mammal and chicken belongs to group Aves.

```
!Gene=FirstGene  Domain=Exon1  Property=Coding;
#Human_{Mammal}  ATGGTTTCTAGTCAGGTCACCATGATAGGTCTCAAT
#Mouse_{Mammal}  ATGGTTTCTAGTCAGGTCACCATGATAGGTCCCAAT
#Chicken_{Aves}  ATGGTTTCTAGTCAGCTCACCATGATAGGTCTCAAT

!Gene=SecondGene  Domain=Intron Property=Noncoding;
#Human     ATTCCCAGGGAATTCCCGGGGGGTTTAAGGCCCCTTTAAAGAAAGAT
#Mouse     GTAGCGCGCGTCGTCAGAGCTCCCAAGGGTAGCAGTCACAGAAAGAT
#Chicken   GTAAAAAAAAAAGTCAGAGCTCCCCCCAATATATATCACAGAAAGAT

!Gene=ThirdGene  Domain=Exon2  Property=Coding;
#Human      ATCTGCTCTCGAGTACTGATACAAATGACTTCTGCGTACAACTGA
#Mouse      ATCTGATCTCGTGTGCTGGTACGAATGATTTCTGCGTTCAACTGA
#Chicken    ATCTGCTCTCGAGTACTGCTACCAATGACTTCTGCGTACAACTGA
```

Setup/Select Taxa & Groups

Data | Setup/Select Taxa & Groups

This invokes the *Setup/Select Taxa & Groups* dialog box for including or excluding taxa, defining groups of taxa, and editing names of taxa and groups.

Labelling Individual Sites

4.14 Site Label

The individual sites in nucleotide or amino acid data can be labeled to construct non-contiguous sets of sites. The Setup Genes and Domains dialog can be used to assign or edit site labels, in addition to specifying them in the input data files. This is shown in the following example of three-sequences in which the sites in the ThirdGene are labeled with a '+' mark. An underscore marks an absence of any labels.

!Gene=FirstGene Domain=Exon1 Property=Coding;

#Human_{Mammal} ATGGTTTCTAGTCAGGTCACCATGATAGGTCTCAAT

#Mouse_{Mammal} ATGGTTTCTAGTCAGGTCACCATGATAGGTCCCAAT

#Chicken_{Aves} ATGGTTTCTAGTCAGCTCACCATGATAGGTCTCAAT

!Gene=SecondGene Domain=AnIntron Property=Noncoding;

#Human
ATTCCCAGGGAATTCCCGGGGGGTTTAAGGCCCCTTTAAAGAAAGAT

#Mouse
GTAGCGCGCGTCGTCAGAGCTCCCAAGGGTAGCAGTCACAGAAAGAT

#Chicken
GTAAAAAAAAAAGTCAGAGCTCCCCCCAATATATATCACAGAAAGAT

!Gene=ThirdGene Domain=Exon2 Property=Coding;

#Human
ATCTGCTCTCGAGTACTGATACAAATGACTTCTGCGTACAACTGA

#Mouse
ATCTGATCTCGTGTGCTGGTACGAATGATTTCTGCGTTCAACTGA

#Chicken
ATCTGCTCTCGAGTACTGCTACCAATGACTTCTGCGTACAACTGA

!Label +++__-+++-a-+++-L-+++-k-+++123+++-_-+++---+++;

Each site can be associated with only one label. A label can be a letter or a number.

For analyses that require codons, *MEGA* includes only those codons in which all three positions are given the same label. This site labeling system facilitates the analysis of specific sites, as often is required for comparing sequences of regulatory elements, intron-splice sites, and antigen recognition sites in the genes of applications such as the Major Histocompatibility Complex.

4.15 Labeled Sites

Sites in a sequence alignment can be categorized and labeled with user-defined symbols. Each category is represented by a letter or a number. Each site can be assigned to only one category, although any combination of categories can be selected for analysis.

Labeled sites work independently of and in addition to genes and domains, thus allowing complex subsets of sites to be defined easily.

4.16 Distance Input Data

General Considerations (Distance Data Formats)

For a set of *m* sequences (or taxa), there are *m*(*m*-1)/2 pairwise distances. These distances can be arranged either in the lower-left or in the upper-right triangular matrix. After writing the `#mega,!Title,!Description`, and `!Format` commands (some of which are optional), you then need to write all the taxa names (see below). Taxa names are followed by the distance matrix. An example of a matrix is:

```
#one
#two
#three
#four
#five

1.0 2.0  3.0  4.0
 3.0     2.5  4.6
    1.3  3.6
         4.2
```

In the above example, pairwise distances are written in the upper triangular matrix (upper-right format). Two alternate distance matrix formats are:

Lower-left matrix

```
d1
2

d1   d2
3    3

d1   d2   d3
4    4    4

d1   d2   d3   d4
5    5    5    5
```

Upper-right matrix

```
d1   d1   d1   d15
2    3    4

     d2   d2   d25
     3    4

          d3   d35
          4

               d45
```

Keywords for Format Statement (Distance data)

Command	Setting	Remark	Example
DataType	Distance	Specifies that the distance data is in the file	DataType=distance
NSeqs	A count	Number of sequences	NSeqs=85
NTaxa	A count	Same as NSeqs	NTaxa=85
DataFormat	Lowerleft, upperright	Specifies whether the data is in lower left triangular matrix or the upper right triangular matrix	DataFormat=lowerleft

Examples below show the lower-left and the upper-right formats for a five-sequence dataset. Note that in each case the distances are organized in a different order.

Lower-left matrix

d12			
d13	d23		
d14	d24	d34	
d15	d25	d35	d45

Upper-right matrix

d12	d13	d14	d15
	d23	d24	d25
		d34	d35
			d45

Defining Groups

4.17 Tree Input Data

Tree Data

* This section of the online help will be available in future updates of MEGA.

Display Newick Trees from File

Phylogeny | Display Newick Trees from File…

Use this to retrieve and display one or more trees written in Newick format. Multiple trees can be displayed, and their consensus built, in the *Tree Explorer*. *MEGA* supports the display of Newick format trees containing branch lengths as well as bootstrap or other counts (note that the Newick formats do not contain the total number of bootstrap replications conducted).

4.2 Importing Data from other Formats

4.21 Importing Data From Other Formats

MEGA 4 supports conversions from several different file formats into *MEGA 4* format. Each format is indicated by the file extension used. Supported formats include:

Extension	**File type**
. aln	CLUSTAL
. nexus	PAUP, MacClade
. phylip	PHYLIP Interleaved
. phylip2	PHYLIP Noninterleaved
. gcg	GCG format
. fasta	FASTA format
. pir	PIR format
. nbrf	NBRF format
. msf	MSF format
. ig	IG format
. xml	Internet (NCBI) XML format

The following sections briefly describe each of these formats and how *MEGA* handles their conversion.

COMMON FILE CONVERSION ATTRIBUTES

The default input formats are determined by a file's extension (e.g., a file with the extension of ".ig" is initially assumed to be in "IG" input format). However, you have the

option to specify any format for any file; the file extension is simply used as an initial guide. Note that the specification of an incorrect file format most often results in an erroneous conversion or other unexpected error.

Input file types can include any of the following characters in their sequence data:

- The letters: a-z,A-Z for DNA and protein sequences
- Peroid `(.)`
- Hyphen `(-)`
- The space character
- Question mark (?).

Depending on their context, all other characters encountered in input files are either ignored or are interpreted as specific non-sequence data, such as comments, headers, etc.

The first line of all converted files is always: `#Mega`

The second line of all converted file is always: `!Title: <filename>`

where `<filename>` is the name of the input file.

The third line of all converted files is blank.

Many formats can specify the length of the sequences contained within them. The *MEG*A conversion utility ignores these data and does not check to see if the sequences are as long as they are purported to be.

4.22 Convert To MEGA Format (Main File Menu)

File | Convert to MEGA Format

This item allows you to choose the file and/or the format that you would like to use to convert a given sequence data file into a *MEGA* format. It converts the data file and displays the converted data in the editor.

Files written in a number of popular data formats can be converted into *MEGA* format. MEGA 4 supports conversion of CLUSTAL, NEXUS (PAUP, MacClade), PHYLIP, GCG, FASTA, PIR, NBRF, MSF, IG, and XML formats. Details about how MEGA reads and converts these file formats are given in the section Importing Data from Other Formats.

4.23 Format Specific Notes

Converting CLUSTAL Format

Converting CLUSTAL Format

The sequence alignment outputs from CLUSTAL software often are given the default extension .ALN. CLUSTAL is an interleaved format. In a page-wide arrangement the sequence name is in the first column and a part of the sequence's data is right justified. An example of the CLUSTAL format follows:

CLUSTAL X (1.8) multiple sequence alignment

```
Q9Y2J0_Has            ------------
MTDTVFSNSSNRWMYPSDRPLQSNDKEQLQAGWSVHPG
Q06846_RP3A_BOVIN           ------------
MTDTVFSSSSSRWMCPSDRPLQSNDKEQLQTGWSVHPS
JX0338_rabphilin-3A-mouse   ------------MTDTVVN----
RWMYPGDGPLQSNDKEQLQAGWSVHPG

Q9Y2J0_Has            GQPDRQRKQEELTDEEKEIINRVIARAEKMEEMEQER--
IGRLVDRLENM
Q06846_RP3A_BOVIN
     GQPDRQRKQEELTDEEKEIINRVIARAEKMEEMEQER--IGRLVDRLENM
JX0338_rabphilin-3A-mouse
     AQTDRQRKQEELTDEEKEIINRVIARAEKMEAMEQER--IGRLVDRLETM
```

The CLUSTAL file above would be converted by *MEGA 4* into the following format:

```
#mega

Title: Bigrab2.aln



#Q9Y2J0_Hsa

------------MTDTVFSNSSNRWMYPSDRPLQSNDKEQLQAGWSVHPG

GQPDRQRKQEELTDEEKEIINRVIARAEKMEEMEQER--IGRLVDRLENM

RKNVAGDGVNRCILCGEQLGMLGSACVVCEDCKKNVCTKCGVET-NNRLH


#Q06846_RP3A_BOVIN

------------MTDTVFSSSSSRWMCPSDRPLQSNDKEQLQTGWSVHPS

GQPDRQRKQEELTDEEKEIINRVIARAEKMEEMEQER--IGRLVDRLENM
```

```
RKNVAGDGVNRCILCGEQLGMLGSACVVCEDCKKNVCTKCGVETSNNRPH


#JX0338_rabphilin-3A-mouse
------------MTDTVVN----RWMYPGDGPLQSNDKEQLQAGWSVHPG
AQTDRQRKQEELTDEEKEIINRVIARAEKMEAMEQER--IGRLVDRLETM
RKNVAGDGVNRCILCGEQLGMLGSACVVCEDCKKNVCTKCGVETSNNRPH
```

Converting FASTA format

Converting FASTA format

The FASTA file format is very simple and is quite similar to the *MEGA* file format. This is an example of a sample input file:

```
>G019uabh 400 bp
ATACATCATAACACTACTTCCTACCCATAAGCTCCTTTTAACTTGTTAAAGTCTTGCTTG
AATTAAAGACTTGTTTAAACACAAAAATTTAGAGTTTTACTCAACAAAAGTGATTGATTG
ATTGATTGATTGATTGATGGTTTACAGTAGGACTTCATTCTAGTCATTATAGCTGCTGGC
AGTATAACTGGCCAGCCTTTAATACATTGCTGCTTAGAGTCAAAGCATGTACTTAGAGTT
GGTATGATTTATCTTTTTGGTCTTCTATAGCCTCCTTCCCCATCCCCATCAGTCTTAATC
AGTCTTGTTACGTTATGACTAATCTTTGGGGATTGTGCAGAATGTTATTTTAGATAAGCA
AAACGAGCAAAATGGGGAGTTACTTATATTTCTTTAAAGC
>G028uaah 268 bp
CATAAGCTCCTTTTAACTTGTTAAAGTCTTGCTTGAATTAAAGACTTGTTTAAACACAAA
ATTTAGACTTTTACTCAACAAAAGTGATTGATTGATTGATTGATTGATTGATGGTTTACA
GTAGGACTTCATTCTAGTCATTATAGCTGCTGGCAGTATAACTGGCCAGCCTTTAATACA
TTGCTGCTTAGAGTCAAAGCATGTACTTAGAGTTGGTATGATTTATCTTTTTGGTCTTCT
ATAGCCTCCTTCCCCATCCCATCAGTCT
```

The *MEGA* file converter looks for a line that begin with a greater-than sign (‘>’), replaces it with a pound sign (‘#’), takes the word following the pound sign as the sequence name, deletes the rest of the line, and takes the following lines (up to the next line beginning with

a '>') as the sequence data. The MEGA file above would convert as follows:

```
#mega
Title: infile.fasta

#G019uabh
ATACATCATAACACTACTTCCTACCCATAAGCTCCTTTTAACTTGTTAAAGTCTTGCTTG
AATTAAAGACTTGTTTAAACACAAAAATTTAGAGTTTTACTCAACAAAAGTGATTGATTG
ATTGATTGATTGATTGATGGTTTACAGTAGGACTTCATTCTAGTCATTATAGCTGCTGGC
AGTATAACTGGCCAGCCTTTAATACATTGCTGCTTAGAGTCAAAGCATGTACTTAGAGTT
GGTATGATTTATCTTTTTGGTCTTCTATAGCCTCCTTCCCCATCCCCATCAGTCTTAATC
AGTCTTGTTACGTTATGACTAATCTTTGGGGATTGTGCAGAATGTTATTTTAGATAAGCA
AAACGAGCAAAATGGGGAGTTACTTATATTTCTTTAAAGC
#G028uaah
CATAAGCTCCTTTTAACTTGTTAAAGTCTTGCTTGAATTAAAGACTTGTTTAAACACAAA
ATTTAGACTTTTACTCAACAAAAGTGATTGATTGATTGATTGATTGATGGTTTACA
GTAGGACTTCATTCTAGTCATTATAGCTGCTGGCAGTATAACTGGCCAGCCTTTAATACA
TTGCTGCTTAGAGTCAAAGCATGTACTTAGAGTTGGTATGATTTATCTTTTTGGTCTTCT
ATAGCCTCCTTCCCCATCCCATCAGTCT
```

Convert GCG Format

Converting GCG Format

These files consist of one or more groups of non-blank lines separated by one or more blank lines; the non-blank lines look similar to this:

```
Chloroflex

    Chloroflex  Length: 428  Mon Sep 25 17:34:20 MDT 2000
Check: 0  ..

   1 MSKEHVQTIA TDDVSKNGHT PPTNASTPPY PFVAIVGQAE LKLALLLCVV

  51 NPTIGGVMVM GHRGTAKSTA VRALAAMLPP IKAVAGCPYS CAPDRTAGLC
```

```
 101 DQCRALEQQS GKTKKPAVIN IPVPVVDLPL GATEDRVCGT LDIERALTQG
 151 VQAFAPGLLA RANRGFLYID EVNLLEDHLV DVLLDVAASG VNVVEREGVS
 201 VRHPARFVLV GSGNPEEGDL RPQLLDRFGL HARITTITDV SERVEIVKRR
 251 REYDADPFAF VEKWAKETQK LQRKIKQAQR RLPEVILPDP VLYKIAELCV
 301 KLEVDGHRGE LTLARA.ATA LAALEGRNEV TVQDVRRIAV LALRHRLRKD
 351 PLETQD.... ...DAVRIER AVEEVLVP.. .......... ..........
 401 .......... .......... ........
```

The "Check" tag near the end of a line signifies the first line in a new sequence expression. The name of the sequence is obtained from the preceding line; the following lines, up to the next blank line, are accepted as the sequence. For each line in the sequence, the leading digits are stripped off, and the rest of the line is used. The following shows a conversion of the above sequence.

```
#mega
Title: infile.gcg

#Chloroflex
MSKEHVQTIA TDDVSKNGHT PPTNASTPPY PFVAIVGQAE LKLALLLCVV
NPTIGGVMVM GHRGTAKSTA VRALAAMLPP IKAVAGCPYS CAPDRTAGLC
DQCRALEQQS GKTKKPAVIN IPVPVVDLPL GATEDRVCGT LDIERALTQG
VQAFAPGLLA RANRGFLYID EVNLLEDHLV DVLLDVAASG VNVVEREGVS
VRHPARFVLV GSGNPEEGDL RPQLLDRFGL HARITTITDV SERVEIVKRR
REYDADPFAF VEKWAKETQK LQRKIKQAQR RLPEVILPDP VLYKIAELCV
KLEVDGHRGE LTLARA.ATA LAALEGRNEV TVQDVRRIAV LALRHRLRKD
PLETQD.... ...DAVRIER AVEEVLVP.. .......... ..........
.......... .......... ........
```

Converting IG Format Files

IG Format

These files consist of one or more groups of non-blank lines separated by one or more

blank lines. The following is an example of the non-blank lines:

```
;G028uaah  240 bases
G028uaah
CATAAGCTCCTTTTAACTTGTTAAAGTCTTGCTTGAATTAAAGACTTGTT
TAAACACAAAATTTAGACTTTTACTCAACAAAAGTGATTGATTGATTGAT
```

The first line in each group begins with a semicolon. This line is ignored by *MEGA 4*. The following line (e.g., `G028uaah` above) is treated as the name of the sequence. Subsequent lines, until the next semicolon, are taken as the sequence. *MEGA 4* recognizes the letters a-z and A-Z for DNA and protein sequences and only a few special characters, such as period [.], hyphen [-], space, and question mark [?]. Depending on their context, all other characters in the input files are either ignored or are interpreted as specific non-sequence data, such as comments, headers, etc.

The example converts to *MEGA* file format as follows:

```
#mega
!Title: filename
#G019uabh
ATACATCATAACACTACTTCCTACCCATAAGCTCCTTTTAACTTGTTAAA
GTCTTGCTTGAATTAAAGACTTGTTTAAACACAAAAATTTAGAGTTTTAC
```

Converting MSF Format

Converting MSF format

The MSF format is an interleaved format that is designed to simplify the comparison of sequences with similar lengths.

```
    G006uaah  MSF: 240  Type: N  Wed Sep 20 12:57:06 MDT
    2000  Check: 0 ..
```

```
Name: G019uabh      Len:  400  Check:   0  Weight:  1.00
```

```
     Name: G028uaah         Len:   268  Check:    0  Weight:
```

```
1.00

 Name: G022uabh          Len:   257  Check:    0  Weight:
1.00

 Name: G023uabh          Len:   347  Check:    0  Weight:
1.00

 Name: G006uaah          Len:   240  Check:    0  Weight:
1.00

//


       G019uabh ATACATCATA ACACTACTTC CTACCCATAA
GCTCCTTTTA ACTTGTTAAA

       G028uaah CATAAGCTCC TTTTAACTTG TTAAAGTCTT
GCTTGAATTA AAGACTTGTT

       G022uabh TATTTTAGAG ACCCAAGTTT TTGACCTTTT
CCATGTTTAC ATCAATCCTG

       G023uabh AATAAATACC AAAAAAATAG TATATCTACA
TAGAATTTCA CATAAAATAA

       G006uaah ACATAAAATA AACTGTTTTC TATGTGAAAA
TTAACCTANN ATATGCTTTG

       G019uabh GTCTTGCTTG AATTAAAGAC TTGTTTAAAC
ACAAAAATTT AGAGTTTTAC

       G028uaah TAAACACAAA ATTTAGACTT TTACTCAACA
AAAGTGATTG ATTGATTGAT

       G022uabh TAGGTGATTG GGCAGCCATT TAAGTATTAT
TATAGACATT TTCACTATCC

       G023uabh ACTGTTTTCT ATGTGAAAAT TAACCTAAAA
ATATGCTTTG CTTATGTTTA

       G006uaah CTTATGTTTA AGATGTCATG CTTTTTATCA
GTTGAGGAGT TCAGCTTAAT

       G019uabh TCAACAAAAG TGATTGATTG ATTGATTGAT
TGATTGATGG TTTACAGTAG

       G028uaah TGATTGATTG ATGGTTTACA GTAGGACTTC
ATTCTAGTCA TTATAGCTGC

       G022uabh CATTAAAACC CTTTATGCCC ATACATCATA
ACACTACTTC CTACCCATAA

       G023uabh AGATGTCATG CTTTTTATCA GTTGAGGAGT
TCAGCTTAAT AATCCTCTAC

       G006uaah AATCCTCTAA GATCTTAAAC AAATAGGAAA
AAAACTAAAA GTAGAAAATG
```

```
       G019uabh GACTTCATTC TAGTCATTAT AGCTGCTGGC
AGTATAACTG GCCAGCCTTT

       G028uaah TGGCAGTATA ACTGGCCAGC CTTTAATACA
TTGCTGCTTA GAGTCAAAGC

       G022uabh GCTCCTTTTA ACTTGTTAAA GTCTTGCTTG
AATTAAAGAC TTGTTTAAAC

       G023uabh GATCTTAAAC AAATAGGAAA AAAACTAAAA
GTAGAAAATG GAAATAAAAT

       G006uaah GAAATAAAAT GTCAAAGCAT TTCTACCACT
CAGAATTGAT CTTATAACAT

       G019uabh AATACATTGC TGCTTAGAGT CAAAGCATGT
ACTTAGAGTT GGTATGATTT

       G028uaah ATGTACTTAG AGTTGGTATG ATTTATCTTT
TTGGTCTTCT ATAGCCTCCT

       G022uabh ACAAAATTTA GACTTTTACT CAACAAAAGT
GATTGATTGA TTGATTGATT

       G023uabh GTCAAAGCAT TTCTACCACT CAGAATTGAT
CTTATAACAT GAAATGCTTT

       G006uaah GAAATGCTTT TTAAAAGAAA ATATTAAAGT
TAAACTCCCC

       G019uabh ATCTTTTTGG TCTTCTATAG CCTCCTTCCC
CATCCCCATC AGTCTTAATC

       G028uaah TCCCCATCCC ATCAGTCT

       G022uabh GATTGAT

       G023uabh TTAAAAGAAA ATATTAAAGT TAAACTCCCC
TATTTTGCTC GTTTTTGCTT

       G019uabh AGTCTTGTTA CGTTATGACT AATCTTTGGG
GATTGTGCAG AATGTTATTT

       G023uabh ATCTAAAATA CATTCTGCAC AATCCCCAAA
GATTGATCAT ACGTTAC

       G019uabh TAGATAAGCA AAACGAGCAA AATGGGGAGT
TACTTATATT TCTTTAAAGC
```

The *MEGA* format converter "unravels" the interleaved data by extracting each line beginning with the first name, then those beginning with the second name, and so on, ultimately producing a corresponding file that looks like this:

```
#mega

Title: thisfile.msf
```

```
#G019uabh
ATACATCATA ACACTACTTC CTACCCATAA GCTCCTTTTA ACTTGTTAAA
GTCTTGCTTG AATTAAAGAC TTGTTTAAAC ACAAAAATTT AGAGTTTTAC
TCAACAAAAG TGATTGATTG ATTGATTGAT TGATTGATGG TTTACAGTAG
GACTTCATTC TAGTCATTAT AGCTGCTGGC AGTATAACTG GCCAGCCTTT
AATACATTGC TGCTTAGAGT CAAAGCATGT ACTTAGAGTT GGTATGATTT
ATCTTTTTGG TCTTCTATAG CCTCCTTCCC CATCCCCATC AGTCTTAATC
AGTCTTGTTA CGTTATGACT AATCTTTGGG GATTGTGCAG AATGTTATTT
TAGATAAGCA AAACGAGCAA AATGGGGAGT TACTTATATT TCTTTAAAGC

#G028uaah
CATAAGCTCC TTTTAACTTG TTAAAGTCTT GCTTGAATTA AAGACTTGTT
TAAACACAAA ATTTAGACTT TTACTCAACA AAAGTGATTG ATTGATTGAT
TGATTGATTG ATGGTTTACA GTAGGACTTC ATTCTAGTCA TTATAGCTGC
TGGCAGTATA ACTGGCCAGC CTTTAATACA TTGCTGCTTA GAGTCAAAGC
ATGTACTTAG AGTTGGTATG ATTTATCTTT TTGGTCTTCT ATAGCCTCCT
TCCCCATCCC ATCAGTCT

#G022uabh
TATTTTAGAG ACCCAAGTTT TTGACCTTTT CCATGTTTAC ATCAATCCTG
TAGGTGATTG GGCAGCCATT TAAGTATTAT TATAGACATT TTCACTATCC
CATTAAAACC CTTTATGCCC ATACATCATA ACACTACTTC CTACCCATAA
GCTCCTTTTA ACTTGTTAAA GTCTTGCTTG AATTAAAGAC TTGTTTAAAC
ACAAAATTTA GACTTTTACT CAACAAAAGT GATTGATTGA TTGATTGATT
GATTGAT

#G023uabh
AATAAATACC AAAAAAATAG TATATCTACA TAGAATTTCA CATAAAATAA
ACTGTTTTCT ATGTGAAAAT TAACCTAAAA ATATGCTTTG CTTATGTTTA
AGATGTCATG CTTTTTATCA GTTGAGGAGT TCAGCTTAAT AATCCTCTAC
```

```
GATCTTAAAC AAATAGGAAA AAAACTAAAA GTAGAAAATG GAAATAAAAT
GTCAAAGCAT TTCTACCACT CAGAATTGAT CTTATAACAT GAAATGCTTT
TTAAAAGAAA ATATTAAAGT TAAACTCCCC TATTTTGCTC GTTTTTGCTT
ATCTAAAATA CATTCTGCAC AATCCCCAAA GATTGATCAT ACGTTAC

#G006uaah
ACATAAAATA AACTGTTTTC TATGTGAAAA TTAACCTANN ATATGCTTTG
CTTATGTTTA AGATGTCATG CTTTTTATCA GTTGAGGAGT TCAGCTTAAT
AATCCTCTAA GATCTTAAAC AAATAGGAAA AAAACTAAAA GTAGAAAATG
GAAATAAAAT GTCAAAGCAT TTCTACCACT CAGAATTGAT CTTATAACAT
GAAATGCTTT TTAAAAGAAA ATATTAAAGT TAAACTCCCC
```

Converting NBRF Format

Converting NBRF Format

NBRF files consist of one or more groups of non-blank lines separated by one or more blank lines; the non-blank lines look similar to this:

```
>P1;Chloroflex
Chloroflex  428 bases
 MSKEHVQTIA TDDVSKNGHT PPTNASTPPY PFVAIVGQAE LKLALLLCVV
 NPTIGGVMVM GHRGTAKSTA VRALAAMLPP IKAVAGCPYS CAPDRTAGLC
 DQCRALEQQS GKTKKPAVIN IPVPVVDLPL GATEDRVCGT LDIERALTQG
 VQAFAPGLLA RANRGFLYID EVNLLEDHLV DVLLDVAASG VNVVEREGVS
 VRHPARFVLV GSGNPEEGDL RPQLLDRFGL HARITTITDV SERVEIVKRR
 REYDADPFAF VEKWAKETQK LQRKIKQAQR RLPEVILPDP VLYKIAELCV
 KLEVDGHRGE LTLARA-ATA LAALEGRNEV TVQDVRRIAV LALRHRLRKD
 PLETQD---- ---DAVRIER AVEEVLVP-- ---------- ----------
 ---------- ---------- --------*
```

Each group begins with a line starting with a greater-than symbol ('>'). This line is ignored. The first word in the following line (e.g., `Chloroflex` above) is treated as the name of the sequence; the rest of that line is ignored Subsequent lines are taken as the sequence. This example would be converted to the *MEGA* file format as follows:

```
#mega
!Title: filename

#Chloroflex
MSKEHVQTIA TDDVSKNGHT PPTNASTPPY PFVAIVGQAE LKLALLLCVV
NPTIGGVMVM GHRGTAKSTA VRALAAMLPP IKAVAGCPYS CAPDRTAGLC
DQCRALEQQS GKTKKPAVIN IPVPVVDLPL GATEDRVCGT LDIERALTQG
VQAFAPGLLA RANRGFLYID EVNLLEDHLV DVLLDVAASG VNVVEREGVS
VRHPARFVLV GSGNPEEGDL RPQLLDRFGL HARITTITDV SERVEIVKRR
REYDADPFAF VEKWAKETQK LQRKIKQAQR RLPEVILPDP VLYKIAELCV
KLEVDGHRGE LTLARA-ATA LAALEGRNEV TVQDVRRIAV LALRHRLRKD
PLETQD---- ---DAVRIER AVEEVLVP-- ---------- ----------
---------- ---------- --------
```

Converting Nexus Format

Format: nexus

The NEXUS file format has a header with lines identifying the name of each of the sequences in the file, followed by lines that begin with the sequence name and some data. An example of part of an input file is:

```
#NEXUS
BEGIN DATA;
 DIMENSIONS NTAX=17 NCHAR=428;
 FORMAT DATATYPE=PROTEIN INTERLEAVE MISSING=-;
[Name: Chloroflex       Len:   428  Check:    0]
[Name: Rcapsulatu       Len:   428  Check:    0]
```

```
MATRIX

          Chloroflex MSKEHVQTIATDDVSKNGHT
PPTNASTPPYPFVAIVGQAE

          Rcapsulatu ---------MTTAVARLQPS
ASGAKTRPVFPFSAIVGQED


          Chloroflex DQCRALEQQSGKTKKPAVIN
IPVPVVDLPLGATEDRVCGT

          Rcapsulatu DWATVLS-----TN---VIR
KPTPVVDLPLGVSEDRVVGA
```

The *MEGA 4* conversion function looks for all the lines starting with the "`[Name:`" flag and takes the following word as a sequence name. The conversion function then scans through the data looking for all lines starting with each of the identified names and places them on the output. This appears as follows:

```
#mega

Title: infile.nexus

#Chloroflex

MSKEHVQTIATDDVSKNGHT PPTNASTPPYPFVAIVGQAE

DQCRALEQQSGKTKKPAVIN IPVPVVDLPLGATEDRVCGT


#Rcapsulatu

---------MTTAVARLQPS ASGAKTRPVFPFSAIVGQED

DWATVLS-----TN---VIR KPTPVVDLPLGVSEDRVVGA
```

Converting PHYLIP (interleaved) Format

Converting the PHYLIP interleaved file format

The PHYLIP format is interleaved, similar to the MSF format. It consists of a line of numeric data, which is ignored by *MEGA 4*, followed by a group of one or more lines of text. The text begins with a sequence name in the first column and is followed by the initial part of each sequence; the group is terminated by a blank line. The number of lines in subsequent groups of data is similar to the first group. Each line is a continuation of the identified sequence and begins in the same position as in the first group. The following might be observed at the beginning of a PHYLIP data file:

```
2 2000 I

G019uabh      ATACATCATA ACACTACTTC CTACCCATAA GCTCCTTTTA
ACTTGTTAAA

G028uaah      CATAAGCTCC TTTTAACTTG TTAAAGTCTT GCTTGAATTA
AAGACTTGTT


              GTCTTGCTTG AATTAAAGAC TTGTTTAAAC ACAAAAATTT
AGAGTTTTAC

              TAAACACAAA ATTTAGACTT TTACTCAACA AAAGTGATTG
ATTGATTGAT


              TCAACAAAAG TGATTGATTG ATTGATTGAT TGATTGATGG
TTTACAGTAG

              TGATTGATTG ATGGTTTACA GTAGGACTTC ATTCTAGTCA
TTATAGCTGC
```

MEGA 4 would convert this data as follows:

```
#mega

Title: cap-data.phylip


#G019uabh

ATACATCATA ACACTACTTC CTACCCATAA GCTCCTTTTA ACTTGTTAAA

GTCTTGCTTG AATTAAAGAC TTGTTTAAAC ACAAAAATTT AGAGTTTTAC

TCAACAAAAG TGATTGATTG ATTGATTGAT TGATTGATGG TTTACAGTAG

#G028uaah

CATAAGCTCC TTTTAACTTG TTAAAGTCTT GCTTGAATTA AAGACTTGTT

TAAACACAAA ATTTAGACTT TTACTCAACA AAAGTGATTG ATTGATTGAT

TGATTGATTG ATGGTTTACA GTAGGACTTC ATTCTAGTCA TTATAGCTGC
```

Converting PHYLIP (Noninterleaved) Format

Converting PHYLIP non-interleaved format

While otherwise similar to the PHYLIP interleaved format, this format is not interleaved. For example:

```
0 0 I

G019uabh      ATACATCATA ACACTACTTC CTACCCATAA GCTCCTTTTA
ACTTGTTAAA

              GTCTTGCTTG AATTAAAGAC TTGTTTAAAC ACAAAAATTT
AGAGTTTTAC

              TCAACAAAAG TGATTGATTG ATTGATTGAT TGATTGATGG
TTTACAGTAG

              GACTTCATTC TAGTCATTAT AGCTGCTGGC AGTATAACTG
GCCAGCCTTT

              AATACATTGC TGCTTAGAGT CAAAGCATGT ACTTAGAGTT


G028uaah      CATAAGCTCC TTTTAACTTG TTAAAGTCTT GCTTGAATTA
AAGACTTGTT

              TAAACACAAA ATTTAGACTT TTACTCAACA AAAGTGATTG
ATTGATTGAT

              TGATTGATTG ATGGTTTACA GTAGGACTTC ATTCTAGTCA
TTATAGCTGC

              TGGCAGTATA ACTGGCCAGC CTTTAATACA TTGCTGCTTA
GAGTCAAAGC

              ATGTACTTAG AGTTGGTATG ATTTATCTTT TTGGTCTTCT
```

This file would be converted to *MEGA* format as follows:

```
#mega

Title: infile.phylip2


#G019uabh

ATACATCATA ACACTACTTC CTACCCATAA GCTCCTTTTA ACTTGTTAAA

GTCTTGCTTG AATTAAAGAC TTGTTTAAAC ACAAAAATTT AGAGTTTTAC

TCAACAAAAG TGATTGATTG ATTGATTGAT TGATTGATGG TTTACAGTAG
```

```
GACTTCATTC TAGTCATTAT AGCTGCTGGC AGTATAACTG GCCAGCCTTT
AATACATTGC TGCTTAGAGT CAAAGCATGT ACTTAGAGTT

#G028uaah
CATAAGCTCC TTTTAACTTG TTAAAGTCTT GCTTGAATTA AAGACTTGTT
TAAACACAAA ATTTAGACTT TTACTCAACA AAAGTGATTG ATTGATTGAT
TGATTGATTG ATGGTTTACA GTAGGACTTC ATTCTAGTCA TTATAGCTGC
TGGCAGTATA ACTGGCCAGC CTTTAATACA TTGCTGCTTA GAGTCAAAGC
ATGTACTTAG AGTTGGTATG ATTTATCTTT TTGGTCTTCT
```

Converting PIR Format

Converting PIR Format

These files consist of groups of non-blank lines that look similar to this:

```
ENTRY           G006uaah
TITLE           G019uabh 400 bp 240 bases
SEQUENCE

                5        10        15        20        25
30

      1 A C A T A A A A T A A A C T G T T T T C T A T G T G
A A A A

     31 T T A A C C T A N N A T A T G C T T T G C T T A T G
T T T A

     61 A G A T G T C A T G C T T T T T A T C A G T T G A G
G A G T

     91 T C A G C T T A A T A A T C C T C T A A G A T C T T
A A A C

    121 A A A T A G G A A A A A A A C T A A A A G T A G A A
A A T G

    151 G A A A T A A A A T G T C A A A G C A T T T C T A C
C A C T

    181 C A G A A T T G A T C T T A T A A C A T G A A A T G
C T T T
```

```
    211 T T A A A A G A A A A T A T T A A A G T T A A A C T
C C C C
```

The *MEGA* format converter looks for the "ENTRY" tag and treats the following string as the sequence name, e.g., `G006uaah` above. The remaining lines have their digits and spaces removed; any non-sequence characters also are deleted. *MEGA* would convert the above sequence as follows:

```
#mega

Title: filename.pir


#G006uaah

ACATAAAATAAACTGTTTTCTATGTGAAAA

TTAACCTANNATATGCTTTGCTTATGTTTA

AGATGTCATGCTTTTTATCAGTTGAGGAGT

TCAGCTTAATAATCCTCTAAGATCTTAAAC

AAATAGGAAAAAAACTAAAAGTAGAAAATG

GAAATAAAATGTCAAAGCATTTCTACCACT

CAGAATTGATCTTATAACATGAAATGCTTT

TTAAAAGAAAATATTAAAGTTAAACTCCCC
```

Converting XML format

Converting XML Format

These files consist of a group of XML tags and attribute values. A DOCTYPE header may or may not be present. This is a relatively new format and is subject to revision. The *MEGA* input converter for XML file formats does not implement a full parser; it only looks for a few specific tags that might be present. For example, an XML file might contain the following data:

```
<Bioseq-set>
  <Bioseq>
    <name>G019uabh</name>
    <length>240</length>
```

```
    <mol>DNA</mol>

    <cksum>302C447C</cksum>

    <seq-
data>ATACATCATAACACTACTTCCTACCCATAAGCTCCTTTTAACTTGTTAAAGTCTT
GCTTGAATT

AAAGACTTGTTTAAACACAAAAATTTAGAGTTTTACTCAACAAAAGTGATTGATTGATTG
ATTGATTGATTGATGGTT

TACAGTAGGACTTCATTCTAGTCATTATAGCTGCTGGCAGTATAACTGGCCAGCCTTTAA
TACATTGCTGCTTAGAGT

CAAAGCATGTACTTAGAGTT</seq-data>

  </Bioseq>

</Bioseq-set>
```

The *MEGA* format converter looks for the following two tags:

```
<name>G019uabh</name>

<seq-data>ATACATCATAACACTAC. . .</seq-data>
```

If it finds these tags, it uses the text between the `<name>. . .</name>` tags as the sequence name, and the text between the `<seq-data>. . .</seq-data>` tags as the sequence data corresponding to that name. The conversion of the above XML block into *MEGA* format would look like this:

```
#Mega

Title: filename.xml


#G019uabh

ATACATCATAACACTACTTCCTACCCATAAGCTCCTTTTAACTTGTTAAAGTCTTGCTTG
AATT

AAAGACTTGTTTAAACACAAAAATTTAGAGTTTTACTCAACAAAAGTGATTGATTGATTG
ATTGATTGATTGATGGTT

TACAGTAGGACTTCATTCTAGTCATTATAGCTGCTGGCAGTATAACTGGCCAGCCTTTAA
TACATTGCTGCTTAGAGT
```

4.3 Genetic Code Tables

4.31 Built-in Genetic Codes

MEGA 4 contains four commonly used genetic code tables: (1) Standard, (2) Vertebrate mitochondrial, (3) Drosophila mitochondrial, and (4) Yeast mitochondrial. They can be used as templates to create additional genetic code tables using the *Genetic Code Selector*. Genetic codes for these four built-in tables in one letter code are given below.

	Code Table					**Code Table**			
Codon	1	2	3	4	Codon	1	2	3	4
UUU	F	F	F	F	AUU	I	I	I	I
UUC	F	F	F	F	AUC	I	I	I	I
UUA	L	L	L	L	AUA	I	M	M	I
UUG	L	L	L	L	AUG	M	M	M	M
UCU	S	S	S	S	ACU	T	T	T	T
UCC	S	S	S	S	ACC	T	T	T	T
UCA	S	S	S	S	ACA	T	T	T	T
UCG	S	S	S	S	ACG	T	T	T	T
UAU	Y	Y	Y	Y	AAU	N	N	N	N
UAC	Y	Y	Y	Y	AAC	N	N	N	N
UAA	*	*	*	*	AAA	K	K	K	K
UAG	*	*	*	*	AAG	K	K	K	K
UGU	C	C	C	C	AGU	S	S	S	S
UGC	C	C	C	C	AGC	S	S	S	S
UGA	*	W	W	W	AGA	R	*	S	R
UGG	W	W	W	W	AGG	R	*	S	R
CUU	L	L	L	T	GUU	V	V	V	V
CUC	L	L	L	T	GUC	V	V	V	V
CUA	L	L	L	T	GUA	V	V	V	V
CUG	L	L	L	T	GUG	V	V	V	V
CCU	P	P	P	P	GCU	A	A	A	A
CCC	P	P	P	P	GCC	A	A	A	A
CCA	P	P	P	P	GCA	A	A	A	A

CCG	P	P	P	P	GCG	A	A	A	A
CAU	H	H	H	H	GAU	D	D	D	D
CAC	H	H	H	H	GAC	D	D	D	D
CAA	Q	Q	Q	Q	GAA	E	E	E	E
CAG	Q	Q	Q	Q	GAG	E	E	E	E
CGU	R	R	R	R	GGU	G	G	G	G
CGC	R	R	R	R	GGC	G	G	G	G
CGA	R	R	R	R	GGA	G	G	G	G
CGG	R	R	R	R	GGG	G	G	G	G

4.32 Adding/Modifying Genetic Code Tables

You may add new genetic code tables and/or edit existing code tables in the *Genetic Code Selector*. All changes made will be remembered by *MEGA* for all future analyses.

4.33 Computing Statistical Attributes (Genetic Code)

There is a significant amount of redundancy in the genetic code because most amino acids are encoded by multiple codons. Therefore, it is interesting to know the degeneracy of each codon position in all codons. In *MEGA 4* this information can be computed for any code table in the *Genetic Code Selector*. In addition to the degeneracy of the codon positions, *MEGA 4* writes the number of synonymous sites and the number of nonsynonymous sites for each codon using the Nei and Gojobori (1986) method. An example of the results obtained for the standard genetic code is given below:

Code Table: Standard

Method: Nei-Gojobori (1986) methodology

S = *No. of synonymous sites*

N = No. of nonsynonymous sites

Codon	**No. of Sites for codon**		**Redundancy**		
	S	**N**	**Pos 1st**	**Pos 2n**	**Pos 3rd**

				d	
UUU (F)	0.333	2.667	0	0	2
UUC (F)	0.333	2.667	0	0	2
UUA (L)	0.667	2.333	2	0	2
UUG (L)	0.667	2.333	2	0	2
UCU (S)	1	2	0	0	4
UCC (S)	1	2	0	0	4
UCA (S)	1	2	0	0	4
UCG (S)	1	2	0	0	4
UAU (Y)	1	2	0	0	2
UAC (Y)	1	2	0	0	2
UAA (*)	0	3	0	0	0
UAG (*)	0	3	0	0	0
UGU (C)	0.5	2.5	0	0	2
UGC (C)	0.5	2.5	0	0	2
UGA (*)	0	3	0	0	0
UGG (W)	0	3	0	0	0
CUU (L)	1	2	0	0	4
CUC (L)	1	2	0	0	4
CUA (L)	1.333	1.667	2	0	4

Select Genetic Code Table

Data | Select Genetic Code Table

Use the *Select Genetic Code Table* dialog from the *Data* menu to select the genetic code used by the protein-coding nucleotide sequence data. This also allows you to add genetic codes to the list, edit existing codes, and compute a few simple statistical properties of the chosen genetic code. This option becomes visible when you open a data set containing nucleotide sequences.

4.34 Code Table Editor

The *Code Table Editor* allows you to create new genetic codes and to edit existing genetic codes. It contains the code of the highlighted genetic code table from the previous window. To name the new genetic code or to change an existing code, click in the 'Name' box and type the new name.

The genetic code in this editor is set up intuitively. To save space, only the amino acid encoded by a codon is shown. The first position of the codon is shown on the left, the second position on the top, and the third position on the right. To find the codon for any given entry on the screen, position your mouse over the desired amino acid and wait for a moment; a yellow hint will be displayed.

To change the amino acid encoded by any codon, click and scroll down to choose the desired amino acid. Alternatively, once the codon has been selected, type in the first letter of the name of the amino acid and the program will jump to that part of the list. To indicate a stop codon, select '***' or type *.

Once you have made all the required changes to the name and codons, click *OK*. Otherwise, click *Cancel*. We recommend that you check the altered genetic code using the *View* option to make sure that the changes have been properly interpreted by *MEGA*.

4.4 Viewing and Exploring Input Data

4.41 Sequence Data Explorer

4.42 Sequence Data Explorer

The *Sequence Data Explorer* shows the aligned sequence data. You can scroll along the alignment using the scrollbar at the bottom right hand side of the explorer window. The *Sequence Data Explorer* provides a number of useful functionalities for exploring the statistical attributes of the data and also for selecting data subsets.

This explorer consists of a number of regions as follows:

Menu Bar

Data menu

Display menu

Highlight menu

Statistics menu

Help: This item brings up the help file for the *Sequence Data Explorer*.

Tool Bar

The tool bar provides quick access to the following menu items:

- *General Utilities*
  -  : This brings up the *Exporting Sequence Data* dialog box, which contains options to control how *MEGA* writes the output data.
  - **Color**: This brings up a color palette selection box with which you can choose the color to be displayed in the highlighted sites.
  -  : This brings up the dialog box for setting up and selecting domains and genes.
  -  : This brings up the dialog box for setting up, editing, and selecting taxa and groups of taxa.
-  : This toggle replaces the nucleotide (amino acid) at a site with the identical symbol (e.g. a dot) if the site contains the same nucleotide (amino acid).
- *Highlighting Sites*
  - **C**: If this button is pressed, then all constant sites will be highlighted. A count of the highlighted sites will be displayed on the status bar.
  - **V**: If this button is pressed, then all variable sites will be highlighted. A count of the highlighted sites will be displayed on the status bar.
  - **Pi**: If this button is pressed, then all parsimony-informative sites will be highlighted. A count of the highlighted sites will be displayed on the status bar.
  - **S**: If this button is pressed, then all singleton sites will be highlighted. A count of the highlighted sites will be displayed on the status bar.
  - **0**: If this button is pressed, then sites will be highlighted only if they are zero-fold degenerate sites in all sequences displayed. A count of highlighted sites will be displayed on the status bar. (This button is

available only if the dataset contains protein coding DNA sequences).

- **2**: If this button is pressed, then sites will be highlighted only if they are two-fold degenerate sites in all sequences displayed. A count of highlighted sites will be displayed on the status bar. (This button is available only if the dataset contains protein coding DNA sequences).
- **4**: If this button is pressed, then sites will be highlighted only if they are four-fold degenerate sites in all sequences displayed. A count of highlighted sites will be displayed on the status bar. (This button is available only if the dataset contains protein coding DNA sequences).
- 

: This button provides the facility to translate codons in the sequence data into amino acid sequences and back. All protein-coding regions will be automatically identified and translated for display. When the translated sequence is already displayed, then issuing this command displays the original nucleotide sequences (including all coding and non-coding regions). Depending on the data displayed (translated or nucleotide), relevant menu options in the *Sequence Data Explorer* become enabled. Note that the translated/un-translated status in this data explorer does not have any impact on the options for analysis available in *MEGA* (e.g., *Distances* or *Phylogeny* menus), as *MEGA* provides all possible options for your dataset at all times.

The 2-Dimensional Data Grid

Fixed Row: This is the first row in the data grid. It is used to display the nucleotides (or amino acids) in the first sequence when you have chosen to show their identity using a special character. For protein coding regions, it also clearly marks the first, second, and the third codon positions.

Fixed Column: This is the first and the leftmost column in the data grid. It is always visible, even when you are scrolling through sites. The column contains the sequence names and an associated check box. You can check or uncheck this box to include or exclude a sequence from analysis. Also in this column, you can drag-and-drop sequences to sort them.

Rest of the Grid: Cells to the right of and below the first row contain the nucleotides or amino acids of the input data. Note that all cells are drawn in light color if they contain data corresponding to unselected sequences or genes or domains.

Status Bar

This section displays the location of the focused site and the total sequence length. It also shows the site label, if any, and a count of the highlighted sites.

Data Menu

Data Menu

This allows you to explore the active data set, and establish various data attributes, and data subset options.

Data Menu (in Sequence Data Explorer)

This menu provides commands for working with selected data in the *Sequence Data Explorer*

The commands in this menu are:

Write Data to File	Brings up the *Exporting Sequence Data* dialog box.
Translate/Untranslate	Translates protein-coding nucleotide sequences into protein sequences, and back to nucleotide sequences.
Select Genetic Code Table	Brings up the *Select Genetic Code* dialog box, in which you can select, edit or add a genetic code table.
Setup/Select Genes and Domains	Brings up the *Sequence Data Organizer*, in which you can define and edit genes and domains.
Setup/Select Taxa and Groups	Brings up the *Select/Edit Taxa and Groups* dialog, in which you can edit taxa and define groups of taxa.
Quit Data Viewer	Takes the user back to the main interface.

Translate/Untranslate (in Sequence Data Explorer)

Data | Translate/Untranslate

This command is available only if the data contain protein-coding nucleotide sequences. It automatically extracts all protein-coding domains for translation and displays the corresponding protein sequence. If the translated sequence is already displayed, then issuing this command displays the original nucleotide sequences, including all coding and non-coding regions. Depending on the data displayed (translated or nucleotide), relevant menu options in the Sequence Data Explorer are enabled. However, translated and un-translated status does not have any impact on the analytical options available in *MEGA* (e.g., *Distances* or *Phylogeny* menus), as *MEGA* provides all possible options for your dataset at all times.

Select Genetic Code Table (in Sequence Data Explorer)

Data | Select Genetic Code Table

Select Genetic Code Table, can be invoked from within the *Data* menu in *Sequence Data Explorer*, and is also available in the main interface directly in the *Data*

Menu.

Setup/Select Taxa & Groups (in Sequence Data Explorer)

Data | *Setup/Select Taxa & Groups*

Setup/Select Taxa & Groups, can be invoked from within the *Data* menu in *Sequence Data Explorer*, and is also available in the main interface directly in the *Data* Menu.

Setup/Select Genes & Domains (in Sequence Data Explorer)

Data | *Setup/Select Genes & Domains*

Setup/Select Genes & Domains, can be invoked from within the *Data* menu in *Sequence Data Explorer*, and is also available in the main interface directly in the *Data* Menu.

Export Data (in Sequence Data Explorer)

Data | *Export Data*

The *Exporting Sequence Data* dialog box first displays an edit box for entering a title for the sequence data being exported. The default name is the original name of the data set, if there was one. Below the title is a space for entering a brief description of the data set being exported.

Next is the option for determining the format of the data set being exported; *MEGA* currently allows the user to export the data in *MEGA*, PAUP 3.0 and PAUP 4.0 (Nexus, Interleaved in both cases), and PHYLIP 3.0 (Interleaved). tA the end of each line, is "Writing site numbers." The three options available are to not write any number, to write one for each site, or to write the site number of the last site.

Other options in this dialog box include the number of sites per line, which codon position(s) is to be used and whether noncoding regions should be included, and whether the output is to be interleaved. For missing or ambiguous data and alignment gaps, there are four options: include all such data, exclude all such data, exclude or include sites with missing or ambiguous data only, and exclude sites with alignment gaps only.

Quit Data Viewer

Data | *Quit Data Viewer*

This command closes the *Sequence Data Explorer*, and takes the user back to main interface.

Display Menu

Display Menu (in Sequence Data Explorer)

This menu provides commands for adjusting the display of DNA and protein sequences in the grid.

The commands in this menu are:

- *Show only selected sequences*: To work only in a subset of the sequences in the data set, use the check boxes to select the sequences of interest.
- *Use Identical Symbol*: If this site contains the same nucleotide (amino acid) as appears in the first sequence in the list, this command replaces the nucleotide (amino acid) symbol with a dot (.). If you uncheck this option, the *Sequence Data Explorer* displays the single letter code for the nucleotide (amino acid).
- *Color Cells*: This option displays the sequences such that consecutive sites with the same nucleotide (amino acid) have the same background color.
- *Sort Sequences*: The sequences in the data set can be sorted based on several options: sequence names, group names, group and sequence names, or as per the order in the *Select/Edit Taxa Groups* dialog box.
- *Restore input order*: This option resets any changes in the order of the displayed sequences (due to sorting, etc.) back to that in the input data file.
- *Show Sequence Name*: The name of the sequences can be displayed or hidden by checking or unchecking this option. If the sequences have been grouped, then unchecking this option causes only the group name to be retained. If no groups have been made, then no name is displayed.
- *Show Group Name*. This option can be used to display or hide group names if the taxa have been categorized into groups.
- *Change Font*. Brings up the *Font* dialog box, allowing the user to choose the type, style, size, etc. of the font to display the sequences.

Restore Input Order

Display | Restore Input Order

Choosing this restores the order in *Sequence Data Explorer* to that in the input text file.

Show Only Selected Sequences

Display | Show only Selected Sequences

The check boxes in the left column of the display grid can be used to select or deselect sequences for analysis. Subsequent use of the "Show Only Selected Sequences" option in the *Display* menu of *Sequence Data Explorer* hides all the deselected sequences and displays only the selected ones.

Color Cells

Display | *Color cells*

This command colors individual cells in the two-dimensional display grid according to the nucleotide or amino acid it contains. A list of default colors, based on the biochemical properties of the residues, is given below. In a future version, these colors will be customizable by the user.

For DNA sequences:

Symbol	Color
A	Yellow
G	Fuchsia
C	Olive
T	Green
U	Green

For amino acid sequences:

Symbol	Color	Symbol	Color
A	Yellow	M	Yellow
C	Olive	N	Green
D	Aqua	P	Blue
E	Aqua	Q	Green
F	Yellow	R	Red
G	Fuchsia	S	Green
H	Teal	T	Green
I	Yellow	V	Yellow

K	Red	W	Green
L	Yellow	Y	Lime

Use Identical Symbol

Display | *Use Identical Symbol*

Data that contain multiple aligned sequences may be easier to view if, when the nucleotide (amino acid) is the same as that in the corresponding site in the first sequence, the nucleotide (amino acid) is replaced by a dot. Choosing this option again brings back the nucleotide (amino acid) single-letter codes.

Show Sequence Names

Display | *Show Sequence Names*

This option displays the full sequence names in *Sequence Data Explorer*

Show Group Names

Display | *Show Group Names*

This option displays the full group names in *Sequence Data Explorer* if the sequences have been grouped in *Select/Edit Taxa Groups*

Change Font...

Display | *Change Font…*

This command brings up the *Change Font* dialog box, which allows you to change the display font, including font type, style and size. Options to strikeout or underline selected parts of the sequences are also available. There is also an option for using different scripts, although the only option currently available is "Western". Finally the "Sample" window displays the effects of your choices

Sort Sequences

Display | *Sort Sequences*

The sequences in the data set can be sorted based on several options: sequence name, group name, group and sequence names, or as per the order in the Select/Edit Taxa Groups dialog box.

Sort Sequences by Group Name

Display | Sort Sequences | By Group Name

Sequences that have been grouped in *Select/Edit Taxa Groups* can be sorted by the alphabetical order of group names or numerical order of group ID numbers. If the group names contain both a name and a number, the numerical order will be nested within the alphabetical order.

Sort Sequences by Group and Sequence Names

Display | Sort Sequences | By Group and Sequence Names

Sequences that have been grouped in *Select/Edit Taxa Groups* can be sorted by the alphabetical order of group names or the numerical order of group ID numbers. If the group names contain both a name and a number, the numerical order is nested within the alphabetical order. The sequences can be further arranged by sorting the sequence names within the group names.

Sort Sequences As per Taxa/Group Organizer

Display | Sort Sequences | As per Taxa/Group Organizer

The sequence/group order seen in *Select/Edit Taxa Groups* is initially the same as the order in the input text file. However, this order can be changed by dragging-and-dropping. Choose this option if you wish to see the data in the same order in the *Sequence Data Explorer* as in *Select/Edit Taxa Groups*.

Sort Sequences By Sequence Name

Display | Sort Sequences | By Sequence Name

The sequences are sorted by the alphabetical order of sequence names or the numerical order of sequence ID numbers. If the sequence names contain both a name and a number, then the sorting is done with the numerical order nested within the alphabetical order.

Highlight Menu

Highlight Menu (in Sequence Data Explorer)

This menu can be used to highlight certain types of sites. The options are constant sites, *variable sites, parsimony-informative sites, singleton sites, 0-fold, 2-fold* and *4-fold degenerate* sites.

Highlight Conserved Sites

Highlight | Conserved Sites

Use this command to highlight constant sites

Highlight Variable Sites

Highlight | Variable Sites

Use this command to highlight variable sites sites.

Highlight Singleton Sites

Highlight | Singleton Sites

Use this command to highlight singleton sites.

Highlight Parsimony Informative Sites

Highlight | Parsim-Info Sites

Use this command to highlight parsimony-informative sites.

Highlight 0-fold Degenerate Sites

Highlight | 0-fold Degenerate Sites

Use this command to highlight 0-fold degenerate sites.

Highlight 2-fold Degenerate Sites

Highlight | 2-fold Degenerate Sites

Use this command to highlight 2-fold degenerate sites. The command is visible only if the data consists of nucleotide sequences.

Highlight 4-fold Degenerate Sites

Highlight | 4-fold Degenerate Sites

Use this command to highlight 4-fold degenerate sites. The command is visible only if the data consists of nucleotide sequences.

Statistics Menu

Statistics Menu (in Sequence Data Explorer)

Various summary statistics of the sequences can be computed and displayed using this menu. The commands are:

Nucleotide Composition.

Nucleotide Pair Frequencies.

Codon Usage.

Amino Acid Composition.

Use All Selected Sites.

Use only Highlighted Sites. Sites can be selected according to various criteria (see Highlight Sites), and analysis can be performed only on the chosen subset of sites.

Nucleotide Composition

Statistics | *Nucleotide Composition*

This command is visible only if the data consist of nucleotide sequences. *MEGA* computes the base frequencies for each sequence as well as an overall average. These will be displayed by domain in a *Text Editor* domain (if the domains have been defined in Setup/Select Genes & Domains).

Nucleotide Pair Frequencies

Statistics | *Nucleotide Pair Frequencies*

This command is visible only if the data consists of nucleotide sequences. There are two options available: one in which the nucleotide acid pairs are counted bidirectionally site-by-site for the two sequences (giving rise to 16 different nucleotide pairs), the other, in which the pairs are counted unidirectionally (10 nucleotide pairs). *MEGA* will compute the frequencies of these quantities for each sequence as well as an overall average. They will be displayed in a Text Editor domain by domain (if domains have been defined in Setup/Select Genes & Domains).

Codon Usage

Statistics | *Codon Usage*

This command is visible only if the data contains protein-coding nucleotide sequences. *MEGA 4* computes the percent codon usage and the RCSU values for each codon for all sequences included in the dataset. Results will be displayed in a Text Editor domain (if domains have been defined in Setup/Select Genes & Domains).

Amino Acid Composition

Statistics | *Amino acid Composition*

This command is visible only if the data consists of amino acid sequences or if the translated protein coding nucleotide sequences are displayed. *MEGA* will compute the amino acid frequencies for each sequence as well as an overall average, which will be displayed in a Text Editor domain (if domains have been defined in Setup/Select Genes & Domains).

Use All Selected Sites

Statistics | *Use All Selected Sites*

Analysis is conducted on all sites in the sequences, irrespective of whether any sites have been labeled or highlighted.

Use only Highlighted Sites

Statistics | *Use only Highlighted Sites*

Sites can be selected according to various criteria (see Highlight Sites), and analyses will be performed only on the chosen subset of sites. All statistical attributes will be based on these sites.

4.43 Distance Data Explorer

Distance Data Explorer

The *Distance Data Explorer* shows the pair-wise distance data. This explorer is flexible and it provides useful functionalities for computing within group, among group, and overall averages, as well as facilities for selecting data subsets.

This explorer consists of a number of regions as follows:

Menu Bar

File menu

Display menu

Average menu.

Help: This item brings up the help file.

Tool Bar

The tool bar provides quick access to a number of menu items.

- *General Utilities*
  -  : This icon brings up the *Options dialog box* to export the distance matrix as a text file with options to control how *MEGA* writes the output data.

-  : This button brings up the dialog box for setting up, editing, and selecting taxa and groups of taxa.

- *Distance Display Precision*

  -  : With each click of this button, the precision of the distance display is decreased by one decimal place.

  -  : With each click of this button, the precision of the distance display is decreased by one decimal place.

  - Column Sizer: This is a slider that can be used to increase or decrease the width of the columns that show the pairwise distances.

The 2-Dimensional Data Grid

This grid displays the pair-wise distances between all the sequences in the data in the form of a lower or upper triangular matrix. The names of the sequences and groups are the row-headers; the column headers are numbered from 1 to *m*, *m* being the number of sequences. There is a column sizer button for the row-headers, so you can increase or decrease the column size to accommodate the full name of the sequences and groups.

- *Fixed Row*: This is the first row in the data grid that displays the column number.

- *Fixed Column*: This is the first and the leftmost column in the data grid and contains taxa names. Even if you scroll past the initial screen this column will always be visible. To include a taxon in the data set for analysis, check the associated box. In this column, you also can drag-and-drop taxa names to sort them in the desired manner.

- *Rest of the Grid*: The cells to the right of the first column and below the first row contain the nucleotides or amino acids of the input data. Note that all cells containing data corresponding to unselected sequences or genes/domains are drawn in a light color.

Status bar

The status bar shows the sequence pair corresponding to the position of the cursor when the cursor is on any distance value in the display.

File Menu (in Distance Data Explorer)

The *File* menu consists of three commands:

- *Select & Edit Taxa/Groups*: This brings up a dialog box to categorize the taxa into groups.

- *Export/Print Distances*: This brings up a dialog box for writing pairwise distances as a text file, with a choice of several formats.

- *Quit Viewer*: This closes the *Distance Data Explorer*.

Display Menu (in Distance Data Explorer)

The *Display* menu consists of four main commands:

- *Show Only Selected Taxa*: This is a toggle, showing a matrix of all or only selected taxa.
- *Sort Taxa*: This provides a submenu for sorting the order of taxa in one of three ways: by input order, by taxon name or by group name.
- *Show Group Names*: This is a toggle for displaying or hiding the group name next to the name of each taxon, when available.
- *Change Font*: This brings up the dialog box, which allows you to choose the type and size of the font used to display the distance values.

Average Menu (in Distance Data Explorer)

This menu is used for the computation of average values using the selected taxa. The following averaging options are available:

Overall: This computes and displays the overall average.

Within groups: This is enabled only if at least one group is defined. For each group, an arithmetic average is computed for all valid pairwise comparisons and results are displayed in the *Distance Matrix Explorer*. All incalculable within-group averages are shown with a red "n/c".

Between groups: This is enabled only if at least two groups of taxa are defined. For each between group average, an arithmetic average is computed for all valid inter-group pairwise comparisons and results are displayed in the *Distance Matrix Explorer*. All incalculable within group averages are shown with a red "n/c".

Net Between Groups: This computes ***net*** average distances between groups of taxa and is enabled only if at least two groups of taxa with at least two taxa each are defined. The net average distance between two groups is given by

$$d\mathrm{A} = d\mathrm{XY} - (d\mathrm{X} - d\mathrm{Y})/2$$

where, dXY is the average distance between groups X and Y, and dX and dY are the mean within-group distances. All incalculable within group averages are shown with a red "n/c".

Options dialog box

At the top of the options dialog box is an option for the output format (Publication and *MEGA*) with the type of information that is output (distances) mentioned beneath. Below this is the option for outputting the distance data as a lower left triangular matrix or an upper right triangular matrix. On the right are options for specifying the number of decimal places for the pairwise distances in the output,

and the maximum number of distances per line in the matrix.

In addition there are three buttons, one to print or save the output, one to quit the *Options dialog box* without exporting the data (*Cancel*), and the third to bring up the help file (this file). The *Print/Save* button brings up the *Distances Display Box*, where the distances are displayed as specified, with various options to edit, print and save the output.

4.44 Text Editor

Using Text Editor

File Menu

New (in Text Editor)

File | *New*

Use this command to create a new file in the *Text Editor*.

Open (in Text Editor)

File | *Open*

Use this command to open an existing file in the *Text Editor*.

Reopen (in Text Editor)

File | *Reopen*

Choose this command to reopen a recently closed text file from the most-recently-used-files list. When you close a text file in the *Text Editor*, it is added to the Reopen list.

Select All (in Text Editor)

Edit | *Select All*

This is used to select (highlight) everything in the displayed file.

Go to Line (in Text Editor)

Edit | *Go to Line #*

This opens a small dialog box that allows you to enter a number indicating the line

to which you want to move.

Show Line Numbers (in Text Editor)

Display | *Show Line Numbers*

This item can be checked (on) or un-checked (off) to show whether line numbers are displayed next to the lines.

Word Wrap (in Text Editor)

Display | *Word Wrap*

This item can be checked (on) or un-checked (off) to show whether lines in the edit window are automatically wrapped around based on the current window's width.

Save (in Text Editor)

File | *Save*

This allows you to save the file currently being edited.

Save As (in Text Editor)

File | *Save As*

This command brings up the *Save As* dialog box, which allows you to choose the directory, the filename and extension, and the type of file you wish to save. To make a file suitable for loading as data in *MEGA*, you should save the file in *MEGA* format (it is a plain ASCII text file). If there is already another file with the same name, it will be overwritten

Print (in Text Editor)

File | *Print*

This command will print the currently displayed file to the selected printer.

Close File (in Text Editor)

File | *Close File*

This closes the current file.

Exit Editor (in Text Editor)

File | *Exit Editor*

This closes the currently open file. If the file was modified, but the modifications have not been saved, *MEGA* will ask whether to discard the changes. Note that this command exits the *Text Editor* only, not *MEGA*.

Delete (in Text Editor)

Edit | *Delete*

This deletes the selected (highlighted) text. It is NOT copied to the clipboard.

Edit Menu

Cut (in Text Editor)

Edit | *Cut*

This command places a copy of the selected text on the *Windows* clipboard, removing the original string. To paste the contents on the clipboard, use the *Paste* command.

Copy (in Text Editor)

Edit | *Copy*

This places a copy of the selected text on the *Windows* clipboard, leaving the original string untouched. To paste the contents on the clipboard, use the *Paste* command.

Paste (in Text Editor)

Edit | *Paste*

This inserts the most recently copied text present on the *Windows* clipboard.

Undo (in Text Editor)

Edit | *Undo*

Choose this command to undo your most recent action. Repeated use of this command will undo each action, starting with the most recent and going to the oldest. it has unlimited depth.

Font (in Text Editor)

Display | *Set Font*

Choose this command to activate a dialog box with which you can change the display font used by the *Text Editor*. Since an ASCII text file does not have a font attribute, it simply contains the text in the file. Therefore the change in the font only affects the display. The new font is remembered by *MEGA* as your preferred display font for the *Text Editor*.

Search Menu

Find (in Text Editor)

Search | *Find*

Choose this command to display the *Find Text* dialog box.

Find Again (in Text Editor)

Search | *Find Again*

Choose this to repeat the last *Find* command.

Replace (in Text Editor)

Search | *Replace*

This brings up a *Search and Replace* dialog box, which allows you to replace a text string in the file currently being edited.

4.5 Visual Tools for Data Management

4.51 Setup/Select Genes & Domains

Data | Setup/Select Genes & Domains

The *Setup/Select Genes & Domains* dialog box allows you to view, specify, and edit genes and domains and to label sites.

4.52 Groups of taxa

A group of taxa is a set of one or more taxa. Members of a group can be specified in the input data file, and created and edited in the *Setup Taxa and Groups* dialog.

Groups of taxa often are constructed based on their evolutionary relatedness. For example, sequences may be grouped based on the geographic origin of the source individual, or sequences from a multigene family may be arranged into groups consisting of orthologous sequences.

4.53 Data Subset Selection

Sequence Data Subset Selection

Any subset of sequence data can be selected for analysis using the options in the *Data* menu. You may:

1. *Select Taxa* (sequences) *or Groups* of taxa through the *Setup/Select Taxa & Groups* dialog box,
2. *Choose Domains and Genes* through the *Setup/Select Genes & Domains* dialog box,

   Items 1 and 2 lead to the construction of a primary data subset, which is maintained until it is modified in the two dialog boxes mentioned in the above items or in the *Sequence Data Explorer*.
3. Select any combination of Codon Positions to use through the *Analysis Preferences/Options dialog box* from the *Data* | *Select Preferences* menu item in the main interface.
4. Choose to include only the Labeled Sites through the *Data* | *Select Preferences* menu item.
5. Decide to enforce Complete-Deletion or Pairwise-Deletion of the missing data and alignment gaps.

Items 3, 4, and 5 provide the second level of data subset options. You are given relevant choices immediately prior to the start of the analysis. Therefore, these choices are secondary in nature and are specific to the currently requested analysis. The *Analysis Preferences* dialog box remembers them for your convenience and provides them as a default the next time you conduct an analysis that utilizes those options.

Distance Data Subset Selection

You may select *Select Taxa* (sequences) or *Groups* of taxa through the *Setup/Select Taxa & Groups* dialog box to construct a distance matrix. You also can select sequences in the *Distance Data Explorer* by clicking on the check marks next to the taxa names.

5 Part IV: Evolutionary Analysis

5.1 Computing Basic Statistical Quantities for Sequence Data

5.11 Basic Sequence Statistics

In the study of molecular evolution, it often is necessary to know some basic statistical quantities, such as nucleotide frequencies, codon frequencies, and transition/transversion ratios. The statistical quantities that can be computed by *MEGA* are discussed in this section.

5.12 Nucleotide and Amino Acid Compositions

The relative frequencies of the four nucleotides (nucleotide composition) or of the 20 amino acid residues (amino acid composition) can be computed for one specific sequence or for all sequences. For the coding regions of DNA, additional columns are presented for the nucleotide compositions at the first, second, and third codon positions. All results are presented domain-by-domain, if the dataset contains multiple domains. Results for the amino acid composition are presented in a similar tabular form.

5.2 Computing Evolutionary Distances

5.21 Distance Models

Models for estimating distances

The evolutionary distance between a pair of sequences usually is measured by the number of nucleotide (or amino acid) substitutions occurring between them. Evolutionary distances are fundamental for the study of molecular evolution and are useful for phylogenetic reconstructions and the estimation of divergence times. Most of the widely used methods for distance estimation for nucleotide and amino acid sequences are included in *MEGA*. In the following three sections, we present a brief discussion of these methods: nucleotide substitutions, synonymous-nonsynonymous substitutions, and amino acid substitutions. Further details of these methods and general guidelines for the use of these methods are given in Nei and Kumar (2000). Note that in addition to the distance estimates, *MEGA 4* also computes the standard errors of the estimates using the analytical formulas and the bootstrap method.

Distance methods included in *MEGA* in divided in three categories (Nucleotide, Syn-nonsynonymous, and Amino acid):

Nucleotide

Sequences are compared nucleotide-by-nucleotide. These distances can be computed for protein coding and non-coding nucleotide sequences.

No. of differences

p-distance

***Jukes-Cantor** Model*

with Rate Uniformity Among Sites

with Rate Variation Among Sites

***Tajima-Nei** Model*

with Rate Uniformity and Pattern Homogeneity

with Rate Variation Among Sites

with Pattern Heterogeneity Between Lineages

with Rate Variation and Pattern Heterogeneity Heterogeneity

***Kimura 2-Parameter** Model*

with Same Rate Among Sites

with Rate Variation Among Sites)

***Tamura 3-Parameter** Model*

with Rate Uniformily and Pattern Homogeneity

with Rate Variation Among Sites

with Pattern Heterogeneity Between Lineages

with Rate Variation and Pattern Heterogeneity

***Tamura-Nei** Model*

With Rate Uniformity and Pattern Homogeneity

with Rate Variation Among Sites

with Pattern Heterogeneity Between Lineages

with Rate Variation and Pattern Heterogeneity

***Log-Det** Method*

with Pattern Heterogeneity Between Lineages

***Maximum Composite Likelihood** Model*

with Rate Uniformity and Pattern Homogeneity

with Rate Variation Among Sites

with Pattern Heterogeneity Between Lineages

with Rate Variation and Pattern Heterogeneity

Syn-Nonsynonymous

Sequences are compared codon-by-codon. These distances can only be computed for protein-coding sequences or domains.

Nei-Gojobori Method

Modified Nei-Gojobori Method

Li-Wu-Luo Method

Pamilo-Bianchi-Li Method

Kumar Method

Amino Acid

Amino acid sequences are compared residue-by-residue. These distances can be computed for protein sequences and protein-coding nucleotide sequences. In the latter case, protein-coding nucleotide sequences are automatically translated using the selected genetic code table.

No. of differences

p-distance

Poisson Model

with Rate Uniformily Among Sites

with Rate Variation Among Sites

Equal Input Model

with Rate Uniformity and Pattern Homogeneity

with Rate Variation Among Sites

with Pattern Heterogeneity Between Lineages

with Rate Variation and Pattern Heterogeneity

Dayhoff and JTT Models

with Rate Uniformity Among Sites

with Rate Variation Among Sites

Nucleotide Substitution Models

No. of differences (Nucleotide)

This distance is the number of sites at which the two compared sequences differ. If you are using the pairwise deletion option for handling gaps and missing data, it is important to realize that this count does not normalize the number of differences based on the number of valid sites compared, if the sequences contain alignment gaps. Therefore, we recommend that if you use this distance you use the *complete-deletion option.*

For this distance, *MEGA* provides facilities for computing the following quantities:

***d*: Transitions + Transversions** : Number of different nucleotide sites.

***s*: Transitions only** : Number of nucleotide sites with transitional differences.

***v*: Transversions only** : Number of nucleotide sites with transversional differences.

R* = *s*/*v : Transition/transversions ratio.

***L*: No of valid common sites**: Number of compared sites.

Formulas for computing these quantities and their variances are as follows.

Var(d) = $n_d(L-n_d)/L$

Var(s) = $s(L-s)/L$

Var(v) = $v(L-v)/L$

$R = s/v$

Var(R) = $\left[c_1^2P + c_2^2Q - (c_1P + c_2Q)^2\right]/L$

where $c_1 = 1/s$ and $c_2 = -s/v^2$

P and Q are the proportion of sites showing transitional and transversional differences, respectively.

See also Nei and Kumar (2000), page 33.

p-distance (Nucleotide)

This distance is the proportion (p) of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared. It does not make any correction for multiple substitutions at the same site, substitution rate biases (for example, differences in the transitional and transversional rates), or differences in evolutionary rates among sites.

MEGA provides facilities for computing following *p*-distances and related quantities:

***d*: Transitions + Transversions** : Proportion of nucleotide sites that are different.

***s*: Transitions only** : Proportion of nucleotide sites with transitional differences.

***v*: Transversions only** : Proportion of nucleotide sites with transversional differences.

R* = *s*/*v : Transition/transversions ratio.

***L*: No of valid common sites**: Number of sites compared.

Formulas for computing these quantities are as follows:

Quantity	Formula	Variance
p,	n_d / L,	$p(1-p)/L$
s,	P,	$s(1-s)/L$
v,	Q,	$v(1-v)/L$
R,	P/Q,	$\left[c_1^2 P + c_2^2 Q - (c_1 P + c_2 Q)^2\right]/L$

where $c_1 = 1/s$ and $c_2 = -s/v^2$

P and *Q* are the proportion of sites showing transitional and transversional differences, respectively.

See also Nei and Kumar (2000), page 33.

Jukes-Cantor distance

In the Jukes and Cantor (1969) model, the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G. As is shown below, the multiple hit correction equation for this model produces a maximum likelihood estimate of the number of nucleotide substitutions between two sequences. It assumes an equality of substitution rates among sites (see the related gamma distance), equal nucleotide frequencies, and it does not correct for higher rate of transitional substitutions as compared to transversional substitutions.

The Jukes-Cantor model

	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

MEGA provides facilities for computing the following quantities:

***d*: Transitions + Transversions** : Number of nucleotide substitutions per site.

***L*: No of valid common sites**: Number of sites compared.

Formulas for computing these quantities are as follows:

Distance

$$d = -\tfrac{3}{4}\log_e\left(1 - \tfrac{4}{3}p\right)$$

where *p* is the proportion of sites with different nucleotides.

Variance

$$\mathrm{Var}(d) = p(1-p)/\left[\left(1-\tfrac{4}{3}p\right)^2 L\right]$$

See also Nei and Kumar (2000), page 36.

Tajima-Nei distance

In real data, nucleotide frequencies often deviate substantially from 0.25. In this case the Tajima-Nei distance (Tajima and Nei 1984) gives a better estimate of the number of nucleotide substitutions than the Jukes-Cantor distance. Note that this assumes an equality of substitution rates among sites and between transitional and transversional substitutions.

The Felsenstein-Tajima-Nei model

	A	T	C	G
A	-	αg_T	αg_C	αg_G
T	αg_A	-	αg_C	αg_G
C	αg_A	αg_T	-	αg_G
G	αg_A	αg_T	αg_C	-

MEGA provides facilities for computing the following quantities for this method:

***d*: Transitions + Transversions** : Number of nucleotide substitutions per site.

***L*: No of valid common sites**: Number of sites compared.

Formulas for computing these quantities are as follows:

Distance

$$d = -b\log_e\left(1 - p/b\right)$$

where *p* is the proportion of sites with different nucleotides and

$$b = \frac{1}{2}\left[1 - \sum_{i=1}^{4} g_i^2 + p^2/c\right],$$
$$c = \sum_{i=1}^{3}\sum_{j=i+1}^{4} \frac{x_{ij}^2}{2g_i g_j}.$$

where *x*ij is the relative frequency of the nucleotide pair *i* and *j*, *gi*'s are the nucleotide frequencies.

Variance

$$\mathrm{Var}(d) = b^2 p(1-p)/\left[(b-p)^2 L\right]$$

See also Nei and Kumar (2000), page 38.

Kimura 2-parameter distance

Kimura's two parameter model (1980) corrects for multiple hits, taking into account transitional and transversional substitution rates, while assuming that the four nucleotide frequencies are the same and that rates of substitution do not vary among sites (see related Gamma distance).

The Kimura 2-parameter model

	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-

MEGA 4 provides facilities for computing the following quantities:

Quantity	**Description**
d: Transitions + Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = *s*/*v*	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

Formulas for computing these quantities are as follows:

Distances

$$d = -\tfrac{1}{2}\log_e(w_1) - \tfrac{1}{4}\log_e(w_2)$$

$$s = -\tfrac{1}{2}\log_e(w_1) + \tfrac{1}{4}\log_e(w_2)$$

$$v = -\tfrac{1}{2}\log_e(w_2)$$

$$R = s/v$$

where *P* and *Q* are the frequencies of sites with transitional and transversional differences

respectively, and

$$w_1 = 1 - 2P - Q$$
$$w_2 = 1 - 2Q$$

Variances

$$\mathrm{Var}(d) = \left[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2\right] / L$$
$$\mathrm{Var}(s) = \left[c_1^2 P + c_4^2 Q - (c_1 P + c_4 Q)^2\right] / L$$
$$\mathrm{Var}(v) = \left[c_2^2 Q(1-Q)\right] / L$$
$$\mathrm{Var}(R) = \left[c_5^2 P + c_6^2 Q - (c_5 P + c_6 Q)^2\right] / L$$

where

$$c_1 = 1/w_1,$$
$$c_2 = 1/w_2,$$
$$c_3 = \tfrac{1}{2}(c_1 + c_2),$$
$$c_4 = \tfrac{1}{2}(c_1 - c_2),$$
$$c_5 = c_1/v,$$
$$c_6 = (c_4 - c_2 R)/v$$

-

See also Nei and Kumar (2000), page 37.

Tamura 3-parameter distance

Tamura's 3-parameter model corrects for multiple hits, taking into account differences in transitional and transversional rates and G+C-content bias (1992). It assumes an equality of substitution rates among sites.

The Tamura 3-parameter model

	A	T	C	G
A	-	β(1-θ)	βθ	αθ
T	β(1-θ)	-	αθ	βθ
C	β(1-θ)	α(1-θ)	-	βθ
G	α(1-θ)	β(1-θ)	βθ	-

MEGA 4 provides facilities for computing the following quantities:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.

s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
$R = s/v$	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

The formulas for computing these quantities are as follows:

Distances

$$d = -w_\theta \log_e(w_1) - \tfrac{1}{2}(1 - w_\theta)\log_e(w_2)$$

$$s = -w_\theta \log_e(w_1) + \tfrac{1}{2} w_\theta \log_e(w_2)$$

$$v = -\tfrac{1}{2}\log_e(w_2)$$

$$R = s/v$$

where *P* and *Q* are the proportion of sites with transitional and transversional differences respectively, and

$$\theta = g_C + g_G,$$
$$w_\theta = 2\theta(1 - \theta),$$
$$w_1 = 1 - P/w_\theta - Q,$$
$$w_2 = 1 - 2Q,$$

Variances

$$\mathrm{Var}(d) = \left[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2\right]/L$$

$$\mathrm{Var}(s) = \left[c_1^2 P + c_4^2 Q - (c_1 P + c_4 Q)^2\right]/L$$

$$\mathrm{Var}(v) = \left[c_2^2 Q(1-Q)\right]/L$$

$$\mathrm{Var}(R) = \left[c_5^2 P + c_6^2 Q - (c_5 P + c_6 Q)^2\right]/L$$

where

$$c_1 = 1/w_1,$$
$$c_2 = 1/w_2,$$
$$c_3 = w_\theta c_1 + (1 - w_\theta)c_1,$$
$$c_4 = w_\theta (c_1 - c_2),$$
$$c_5 = c_1/v,$$
$$c_6 = (c_4 - c_2 R)/v.$$

See also Nei and Kumar (2000), page 39.

Tamura-Nei distance

The Tamura-Nei model (1993) corrects for multiple hits, taking into account the differences in substitution rate between nucleotides and the inequality of nucleotide frequencies. It distinguishes between transitional substitution rates between purines and transversional substitution rates between pyrimidines. It also assumes equality of substitution rates among sites (see related gamma model).

The Tamura-Nei model

	A	T	C	G
A	-	βg_T	βg_C	$\alpha_1 g_G$
T	βg_A	-	$\alpha_2 g_C$	βg_G
C	βg_A	$\alpha_2 g_T$	-	βg_G
G	$\alpha_1 g_A$	βg_T	βg_C	-

MEGA 4 provides facilities for computing the following quantities for this method:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = *s*/*v*	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

Formulas for computing these quantities are as follows:

Distances

$$d = -k_1 \log_e(w_1) - k_2 \log_e(w_2) - k_3 \log_e(w_3)$$

$$s = -k_1 \log_e(w_1) - k_2 \log_e(w_2) - (k_3 - 2 g_R g_Y) \log_e(w_3)$$

$$v = -2 g_R g_Y \log_e(w_3)$$

$$R = s/v$$

where *P*1 and *P*2 are the proportions of transitional differences between nucleotides A and G, and between T and C, respectively, *Q* is the proportion of transversional differences, *g*A, *g*C, *g*G, *g*T, are the respective frequencies of A, C, G and T, *gR* = *gA* + *gG*, *gY*, = *gT* + *gC*, and

$$k_1 = 2g_A g_G / g_R,$$
$$k_2 = 2g_T g_C / g_Y,$$
$$k_3 = 2(g_R g_Y - g_A g_G g_Y / g_R - g_T g_C g_R / g_Y),$$
$$k_4 = 2(g_A g_G + g_T g_C + g_R g_Y),$$
$$w_1 = 1 - P_1 / k_1 - Q / 2g_R,$$
$$w_2 = 1 - P_2 / k_2 - Q / 2g_Y,$$
$$w_3 = 1 - Q / 2g_R g_Y,$$

Variances

$$\mathrm{Var}(d) = \left[c_1^2 P_1 + c_2^2 P_2 + c_4^2 Q - (c_1 P_1 + c_2 P_2 + c_4 Q)^2\right] / L$$
$$\mathrm{Var}(s) = \left[c_1^2 P_1 + c_2^2 P_2 + c_5^2 Q - (c_1 P_1 + c_2 P_2 + c_5 Q)^2\right] / L$$
$$\mathrm{Var}(v) = \left[c_3^2 Q(1-Q)\right] / L$$
$$\mathrm{Var}(R) = \left[c_6^2 P_1 + c_7^2 P_2 + c_8^2 Q - (c_6 P_1 + c_7 P_2 + c_8 Q)^2\right] / L$$

where

$$c_1 = 1/w_1,$$
$$c_2 = 1/w_2,$$
$$c_3 = 1/w_3,$$
$$c_4 = k_1 c_1 / 2g_R + k_2 c_2 / 2g_Y + k_3 c_3 / (2g_R g_Y),$$
$$c_5 = k_1 c_1 / 2g_R + k_2 c_2 / 2g_Y + k_3 c_3 / 2g_R g_Y - c_3,$$
$$c_6 = c_1 / v,$$
$$c_7 = c_2 / v,$$
$$c_8 = (c_5 - c_3 R) / v.$$

See also Nei and Kumar (2000), page 40.

Maximum Composite Likelihood Method

A composite likelihood is defined as a sum of related log-likelihoods. Since all pairwise distances in a distance matrix have correlations due to the phylogenetic relationships among the sequences, the sum of their log-likelihoods is a composite likelihood. Tamura et al. (2004) showed that pairwise distances and the related substitution parameters are accurately estimated by maximizing the composite likelihood. They also found that, unlike the cases of ordinary independent estimation of each pairwise disntace, a complicated model had virtually no disadvantage in the composite likelihood method for phylogenetic analyses. Therefore, only the Tamura-Nei (1993) model is available for this method in MEGA4 (see related Tamura-Nei distance). It assumes equality of substitution pattern among lineages and of substitution rates among sites (see related gamma model and heterogeneous patterns).

Gamma Distances

Computing the Gamma Parameter (a)

In the computation of gamma distances, it is necessary to know the gamma parameter (*a*).

This parameter may be estimated from the dataset under consideration or you may use the value obtained from previous studies. For estimating *a*, a substantial number of sequences is necessary; if the number of sequences used is small, the estimate has a downward bias (Zhang and Gu 1998). The current release of *MEGA 4* does not contain any programs for estimating *a*; however we plan to make them available in the future. Therefore you need to use another program for estimating the *a* value. Some of the frequently used programs that include this facility are PAUP* (Swofford 1998) for DNA sequences, PAML and PAMP programs for DNA and protein sequences (Yang 1999), and GAMMA programs from Gu and Zhang (1997).

Equal Input Model (Gamma)

In real data, amino acid frequencies usually vary among the different kinds of amino acids and substitution rates are not uniform among sites. In this case, the correction based on the equal input model gives a better estimate of the number of amino acid substitutions than the Poisson correction distance. The rate variation among sites is modeled using the Gamma distribution; for computing this distance you will need to provide a gamma parameter (*a*).

MEGA provides facilities for computing the following quantities:

Quantity	**Description**
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

Formulas used are:

Distance

$$d = ba\left[(1 - p/b)^{-1/a} - 1\right]$$

where *p* is the proportion of different amino acid sites, *a* is the gamma parameter, *gi* is the frequency of amino acid *i*, and

$$b = 1 - \sum_i g_i^2,$$

Variance

$$\mathrm{Var}(d) = p(1-p)/\left[(1 - p/b)^{-2(1+1/a)} L\right]$$

Jukes-Cantor Gamma distance

In the Jukes and Cantor (1969) model, the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G. The multiple hit correction equation for this model, which is given below, produces a maximum likelihood estimate of the number of nucleotide substitutions between two sequences, while relaxing the assumption that all

sites are evolving at the same rate. However, it assumes equal nucleotide frequencies and does not correct for higher rate of transitional substitutions as compared to transversional substitutions. If the rate variation among sites is modeled using the Gamma distribution, you will need to provide a gamma parameter (*a*) for computing this distance.

The Jukes-Cantor model

	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

MEGA provides facilities for computing the following *p*-distances and related quantities:

***d*: Transitions + Transversions** : Number of nucleotide substitutions per site.

***L*: No of valid common sites**: Number of sites compared.

The formulas for computing these quantities are as follows:

Distance

$$d = \tfrac{3}{4} a\left[\left(1 - \tfrac{4}{3} p\right)^{-1/a} - 1\right]$$

where *p* is the proportion of sites with different nucleotides and *a* is the gamma parameter.

Variance

$$\mathrm{Var}(d) = p(1-p) \Big/ \left[\left(1 - \tfrac{4}{3} p\right)^{-2(1+1/a)} L\right]$$

See also Nei and Kumar (2000), page 36 and estimating gamma parameter.

Kimura gamma distance

Kimura's two-parameter gamma model corrects for multiple hits, taking into account transitional and transversional substitution rates and differences in substitution rates among sites. Evolutionary rates among sites are modeled using the Gamma distribution, and you will need to provide a gamma parameter for computing this distance.

The Kimura 2-parameter model

	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-

MEGA 4 provides facilities for computing the following quantities:

Quantity	**Description**
d: Transitions + Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = *s*/*v*	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

The formulas for computing these quantities are as follows:

Distances

$$d = \frac{a}{2}\left[(w_1)^{-1/a} + \tfrac{1}{2}(w_2)^{-1/a} - \tfrac{3}{2}\right]$$

$$s = \frac{a}{2}\left[(w_1)^{-1/a} - \tfrac{1}{2}(w_2)^{-1/a} - \tfrac{1}{2}\right]$$

$$v = \frac{a}{2}\left[(w_2)^{-1/a} - 1\right]$$

$$R = s/v$$

where *P* and *Q* are the respective total frequencies of transition type pairs and transversion type pairs, *a* is the gamma parameter, and

$$w_1 = 1 - 2P - Q$$
$$w_2 = 1 - 2Q$$

Variances

$$\mathrm{Var}(d) = \left[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2\right] / L$$

$$\mathrm{Var}(s) = \left[c_1^2 P + c_4^2 Q - (c_1 P + c_4 Q)^2\right] / L$$

$$\mathrm{Var}(v) = \left[c_2^2 Q(1-Q)\right] / L$$

$$\mathrm{Var}(R) = \left[c_5^2 P + c_6^2 Q - (c_5 P + c_6 Q)^2\right] / L$$

where

$$c_1 = w_1^{-(1+1/a)},$$
$$c_2 = w_2^{-(1+1/a)},$$
$$c_3 = \tfrac{1}{2}(c_1 + c_2),$$
$$c_4 = \tfrac{1}{2}(c_1 - c_2),$$
$$c_5 = c_1 / v,$$
$$c_6 = (c_4 - c_2 R)/v$$

See also Nei and Kumar (2000), page 44 and estimating gamma parameter.

Tajima Nei distance (Gamma rates)

In real data, nucleotide frequencies often deviate substantially from 0.25. In this case the Tajima-Nei distance (Tajima and Nei 1984) gives a better estimate of the number of nucleotide substitutions than the Jukes-Cantor distance. Note that this assumes an equality of substitution rates among sites and between transitional and transversional substitutions. The rate variation among sites is modeled using the gamma distribution, and you will need to provide a gamma parameter (*a*) for computing this distance.

The Felsenstein-Tajima-Nei model

	A	T	C	G
A	-	αg_T	αg_C	αg_G
T	αg_A	-	αg_C	αg_G
C	αg_A	αg_T	-	αg_G
G	αg_A	αg_T	αg_C	-

MEGA provides facilities for computing the following quantities for this method:

***d*: Transitions + Transversions** : Number of nucleotide substitutions per site.

***L*: No of valid common sites**: Number of sites compared.

The formulas for computing these quantities are as follows:

Distance

$$d = ba\left[(1 - p/b)^{-1/a} - 1\right]$$

where *p* is the proportion of sites with different nucleotides, *a* is the gamma parameter, and

$$b = \frac{1}{2}\left[1 - \sum_{i=1}^{4} g_i^2 + p^2 / c\right],$$
$$c = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{x_{ij}^2}{2 g_i g_j}.$$

where *x*ij is the relative frequency of the nucleotide pair *i* and *j*, *gi*'s are the nucleotide frequencies.

Variance

$$\mathrm{Var}(d) = p(1-p)/\left[(1 - p/b)^{-2(1+1/\alpha)} L\right]$$

Tamura-Nei gamma distance

The Tamura-Nei (1993) distance with the gamma model corrects for multiple hits, taking into account the different rates of substitution between nucleotides and the inequality of nucleotide frequencies. In this distance, evolutionary rates among sites are modeled using the gamma distribution. You will need to provide a gamma parameter for computing this distance.

The Tamura-Nei model

	A	T	C	G
A	-	βg_T	βg_C	$\alpha_1 g_G$
T	βg_A	-	$\alpha_2 g_C$	βg_G
C	βg_A	$\alpha_2 g_T$	-	βg_G
G	$\alpha_1 g_A$	βg_T	βg_C	-

MEGA 4 provides facilities for computing the following quantities for this method:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = *s*/*v*	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

The formulas for computing these quantities are as follows:

Distances

$$d = a\left[k_1(w_1)^{-1/a} + k_2(w_2)^{-1/a} + k_3(w_3)^{-1/a} - k_4\right]$$

$$s = a\left[k_1(w_1)^{-1/a} + k_2(w_2)^{-1/a} + (k_3 - 2g_R g_Y)(w_3)^{-1/a} - (k_4 - 2g_R g_Y)\right]$$

$$v = 2ag_R g_Y\left[(w_3)^{-1/a} - 1\right]$$

$$R = s/v$$

where *P*1 and *P*2 are the proportions of transitional differences between nucleotides A and G, and between T and C, respectively, *Q* is the proportion of transversional differences, *g*A, *g*C, *g*G, *g*T, are the respective frequencies of A, C, G and T, *gR* = *gA* + *gG*, *gY*, = *gT* + *gC*, *a* is the gamma parameter and

$$k_1 = 2g_A g_G / g_R,$$
$$k_2 = 2g_T g_C / g_Y,$$
$$k_3 = 2(g_R g_Y - g_A g_G g_Y / g_R - g_T g_C g_R / g_Y),$$
$$k_4 = 2(g_A g_G + g_T g_C + g_R g_Y),$$
$$w_1 = 1 - P_1 / k_1 - Q/2g_R,$$
$$w_2 = 1 - P_2 / k_2 - Q/2g_Y,$$
$$w_3 = 1 - Q/2g_R g_Y,$$

Variances

$$\mathrm{Var}(d) = \left[c_1^2 P_1 + c_2^2 P_2 + c_4^2 Q - (c_1 P_1 + c_2 P_2 + c_4 Q)^2\right]/L$$

$$\mathrm{Var}(s) = \left[c_1^2 P_1 + c_2^2 P_2 + c_5^2 Q - (c_1 P_1 + c_2 P_2 + c_5 Q)^2\right]/L$$

$$\mathrm{Var}(v) = \left[c_3^2 Q(1-Q)\right]/L$$

$$\mathrm{Var}(R) = \left[c_6^2 P_1 + c_7^2 P_2 + c_8^2 Q - (c_6 P_1 + c_7 P_2 + c_8 Q)^2\right]/L$$

where

$$c_1 = (w_1)^{-(1+1/a)},$$
$$c_2 = (w_2)^{-(1+1/a)},$$
$$c_3 = (w_3)^{-(1+1/a)},$$
$$c_4 = k_1 c_1/2g_R + k_2 c_2/2g_Y + k_3 c_3/2g_R g_Y,$$
$$c_5 = k_1 c_1/2g_R + k_2 c_2/2g_Y + k_3 c_3/2g_R g_Y - c_3,$$
$$c_6 = c_1/v,$$
$$c_7 = c_2/v,$$
$$c_8 = (c_5 - c_3 R)/v.$$

See also Nei and Kumar (2000), page 45 and estimating gamma parameter.

Tamura 3-parameter (Gamma)

Tamura's 3-parameter model corrects for multiple hits, taking into account the differences in transitional and transversional rates and the G+C-content bias (1992). Evolutionary

rates among sites are modeled using the gamma distribution, and you will need to provide a gamma parameter for computing this distance.

The Tamura 3-parameter model

	A	T	C	G
A	-	$\beta(1-\theta)$	$\beta\theta$	$\alpha\theta$
T	$\beta(1-\theta)$	-	$\alpha\theta$	$\beta\theta$
C	$\beta(1-\theta)$	$\alpha(1-\theta)$	-	$\beta\theta$
G	$\alpha(1-\theta)$	$\beta(1-\theta)$	$\beta\theta$	-

MEGA 4 provides facilities for computing the following quantities:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = *s*/*v*	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

The formulas for computing these quantities are as follows:

Distances

$$d = \frac{a}{2}\left[w_\theta w_1^{-1/a} + (1 - w_\theta)w_2^{-1/a} - w_\theta - 1\right]$$

$$s = \frac{a}{2}\left[w_\theta w_1^{-1/a} - w_\theta w_2^{-1/a} - w_\theta\right]$$

$$v = \frac{a}{2}\left[w_2^{-1/a} - 1\right]$$

$$R = s/v$$

where *P* and *Q* are the proportion of sites with transitional and transversional differences, respectively, a is the gamma parameter, and

$$\theta = g_C + g_G,$$
$$w_\theta = 2\theta(1 - \theta),$$
$$w_1 = 1 - P/w_\theta - Q,$$
$$w_2 = 1 - 2Q,$$

Variances

$$\mathrm{Var}(d) = \left[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2\right] / L$$

$$\mathrm{Var}(s) = \left[c_1^2 P + c_4^2 Q - (c_1 P + c_4 Q)^2\right] / L$$

$$\mathrm{Var}(v) = \left[c_2^2 Q(1-Q)\right] / L$$

$$\mathrm{Var}(R) = \left[c_5^2 P + c_6^2 Q - (c_5 P + c_6 Q)^2\right] / L$$

where

$$c_1 = w_1^{-(1+1/\alpha)},$$
$$c_2 = w_2^{-(1+1/\alpha)},$$
$$c_3 = w_\theta c_1 + (1 - w_\theta) c_2,$$
$$c_4 = w_\theta (c_1 - c_2),$$
$$c_5 = c_1 / v,$$
$$c_6 = (c_4 - c_2 R) / v.$$

Maximum Composite Likelihood (Gamma Rates)

The Tamura-Nei (1993) distance with the gamma model estimated by the composite likelihood method (Tamura et al. 2004) corrects for multiple hits, taking into account the different rates of substitution between nucleotides and the inequality of nucleotide frequencies. In this distance, evolutionary rates among sites are modeled using the gamma distribution. You will need to provide a gamma parameter for computing this distance. See related Tamura-Nei gamma distance.

Heterogeneous Patterns

Tajima Nei Distance (Heterogeneous patterns)

In real data, nucleotide frequencies often deviate substantially from 0.25. In this case the Tajima-Nei distance (Tajima and Nei 1984) gives a better estimate of the number of nucleotide substitutions than the Jukes-Cantor distance. Note that this assumes an equality of substitution rates among sites and between transitional and transversional substitutions. When the nucleotide frequencies are different between the sequences, the modified formula (Tamura and Kumar 2002) relaxes the assumption of substitution pattern homogeneity.

The Felsenstein-Tajima-Nei model

	A	T	C	G
A	-	αg_T	αg_C	αg_G
T	αg_A	-	αg_C	αg_G
C	αg_A	αg_T	-	αg_G
G	αg_A	αg_T	αg_C	-

MEGA provides facilities for computing the following quantities for this method:

***d*: Transitions + Transversions** : Number of nucleotide substitutions per site.

***L*: No of valid common sites**: Number of sites compared.

Formulas for computing these quantities are as follows:

Distance

$$d = -b\log_e(1 - p/c)$$

where *p* is the proportion of sites with different nucleotides and

$$b = 1 - \sum_{i=1}^{4} g_i^2,$$
$$c = 1 - \sum_{i=1}^{4}\sum_{j\neq i}^{4} g_{Ii}g_{Ij}.$$

where *x*ij is the relative frequency of the nucleotide pair *i* and *j*, *gi*'s are the nucleotide frequencies.

Variance can be estimated by the bootstrap method.

Tamura 3 parameter (Heterogeneous patterns)

Tamura's 3-parameter model corrects for multiple hits, taking into account the differences in transitional and transversional rates and the G+C-content bias (1992). It assumes an equality of substitution rates among sites. When the G+C-contents are different between the sequences, the modified formula (Tamura and Kumar 2002) relaxes the assumption of substitution pattern homogeneity.

The Tamura 3-parameter model

	A	T	C	G
A	-	β(1-θ)	βθ	αθ
T	β(1-θ)	-	αθ	βθ
C	β(1-θ)	α(1-θ)	-	βθ
G	α(1-θ)	β(1-θ)	βθ	-

MEGA 4 provides facilities for computing the following quantities:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.

s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = s/v	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

Formulas for computing these quantities are as follows:

Distances

$$d = -w_\theta \log_e(w_1) - \tfrac{1}{2}(1 - w_\theta)\log_e(w_2)$$

$$s = -w_\theta \log_e(w_1) + \tfrac{1}{2} w_\theta \log_e(w_2)$$

$$v = -\tfrac{1}{2}\log_e(w_2)$$

$$R = s/v$$

where *P* and *Q* are the proportion of sites with transitional and transversional differences, respectively, and

$$\begin{aligned}
\theta_1 &= g_{1C} + g_{1G},\\
\theta_2 &= g_{2C} + g_{2G},\\
\theta &= (\theta_1 + \theta_2)/2,\\
w_\theta &= 2\theta(1-\theta),\\
w_1 &= 1 - P/\left[\theta_1(1-\theta_2) + \theta_2(1-\theta_1)\right] - Q,\\
w_2 &= 1 - 2Q.
\end{aligned}$$

The variances can be estimated by the bootstrap method. .

Tamura-Nei distance (Heterogeneous Patterns)

The Tamura-Nei model (1993) corrects for multiple hits, taking into account the substitution rate differences between nucleotides and the inequality of nucleotide frequencies. It distinguishes between transitional substitution rates between purines and transversional substitution rates between pyrimidines. It assumes an equality of substitution rates among sites (see related gamma model). When nucleotide frequencies are different between the sequences, the modified formula (Tamura and Kumar 2002) relaxes the assumption of substitution pattern homogeneity.

The Tamura-Nei model

	A	T	C	G
A	-	βg_T	βg_C	$\alpha_1 g_G$
T	βg_A	-	$\alpha_2 g_C$	βg_G
C	βg_A	$\alpha_2 g_T$	-	βg_G
G	$\alpha_1 g_A$	βg_T	βg_C	-

MEGA 4 provides facilities for computing the following quantities for this method:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = *s*/*v*	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

Formulas for computing these quantities are as follows:

Distances

$$d = -k_1 \log_e(w_1) - k_2 \log_e(w_2) - k_3 \log_e(w_3)$$

$$s = -k_1 \log_e(w_1) - k_2 \log_e(w_2) - (k_3 - 2 g_R g_Y) \log_e(w_3)$$

$$v = -2 g_R g_Y \log_e(w_3)$$

$$R = s/v$$

where *P*1 and *P*2 are the proportions of transitional differences between nucleotides A and G, and between T and C, respectively, *Q* is the proportion of transversional differences, *gXA*, *gXC*, *gXG*, *gXT*, are the respective frequencies of A, C, G and T of sequence *X*, *gXR* = *gXA* + *gXG* and *gXY* = *gXT* + *gXC*, *gA*, *gC*, *gG*, *gT*, *gR*, and *gY* are the average frequencies of the pair of sequences, and

$$k_1 = 2 g_A g_G / g_R,$$
$$k_2 = 2 g_T g_C / g_Y,$$
$$k_3 = 2 (g_R g_Y - g_A g_G g_Y / g_R - g_T g_C g_R / g_Y),$$
$$w_1 = 1 - P_1 / (g_{1A} g_{2G} + g_{1G} g_{2A}) - Q/2 g_R,$$
$$w_2 = 1 - P_2 / (g_{1T} g_{2C} + g_{1C} g_{2T}) - Q/2 g_Y,$$
$$w_3 = 1 - Q / (g_{1R} g_{2Y} + g_{1Y} g_{2R}).$$

The variances can be estimated by the bootstrap method.

Maximum Composite Likelihood (Heterogeneous Patterns)

The Tamura-Nei distance (1993) estimated by the composite likelihood method (Tamura et al. 2004) corrects for multiple hits, taking into account the substitution rate differences between nucleotides and the inequality of nucleotide frequencies. When the nucleotide frequencies between the sequences are different, the expected proportions of observed differences (P1, P2, Q) in the computation of the composite likelihood can be obtained by the modified formulas according to Tamura and Kumar (2002) to relax the assumption of the substitution pattern homogeneity. See related Tamura-Nei distance (Heterogeneous Patterns).

Gamma Rates

Equal Input Model (Gamma rates and Heterogeneous Patterns)

In real data, amino acid frequencies usually vary among different kind of amino acids. Therefore, the correction based on the equal input model gives a better estimate of the number of amino acid substitutions than the Poisson correction distance. If you are computing the rate variation among sites using the Gamma distribution, you will need to provide a gamma parameter (*a*). When the amino acid frequencies are different between the sequences, the modified formula (Tamura and Kumar 2002) relaxes the estimation bias.

MEGA provides facilities for computing the following quantities:

Quantity	**Description**
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

Formulas used are:

Distance

$$d = ba\left[(1 - p/c)^{-1/a} - 1\right]$$

where *p* is the proportion of different amino acid sites, *a* is the gamma parameter, *gXi* is the frequency of amino acid *i* for sequence *X*, *gi* is the average frequency for the pair of the sequences, and

$$b = 1 - \sum_i g_i^2,$$

$$c = 1 - \sum_i \sum_j g_{Ii} g_{Ij}.$$

The variance of *d* can be estimated by the bootstrap method.

Tajima Nei Distance (Gamma Rates and Heterogeneous patterns)

In real data, nucleotide frequencies often deviate substantially from 0.25. In this case the Tajima-Nei distance (Tajima and Nei 1984) gives a better estimate of the number of nucleotide substitutions than the Jukes-Cantor distance. Note that this assumes an equality of substitution rates among sites and between transitional and transversional substitutions. The rate variation among sites is modeled using the gamma distribution, and you will need to provide a gamma parameter (*a*) for computing this distance. When the nucleotide frequencies are different between the sequences, the modified formula (Tamura and Kumar 2002) relaxes the assumption of substitution pattern homogeneity.

The Felsenstein-Tajima-Nei model

	A	T	C	G
A	-	αg_T	αg_C	αg_G
T	αg_A	-	αg_C	αg_G
C	αg_A	αg_T	-	αg_G
G	αg_A	αg_T	αg_C	-

MEGA provides facilities for computing the following quantities for this method:

***d*: Transitions + Transversions** : Number of nucleotide substitutions per site.

***L*: No of valid common sites**: Number of sites compared.

The formulas for computing these quantities are as follows:

Distance

$$d = ba\left[\left(1 - p/c\right)^{-1/a} - 1\right]$$

where *p* is the proportion of sites with different nucleotides, *a* is the gamma parameter, and

$$b = 1 - \sum_{i=1}^{4} g_i^2,$$

$$c = 1 - \sum_{i=1}^{4}\sum_{j\neq i}^{4} g_{Ii} g_{Ij}$$

where *x*ij is the relative frequency of the nucleotide pair *i* and *j*, *gi*'s are the nucleotide frequencies.

Variance can be estimated by the bootstrap method.

Tamura-Nei distance (Gamma rates and Heterogeneous patterns)

The Tamura-Nei (1993) distance with the gamma model corrects for multiple hits, taking

into account the rate substitution differences between nucleotides and the inequality of nucleotide frequencies. In this distance, evolutionary rates among sites are modeled using the gamma distribution. You will need to provide a gamma parameter for computing this distance. When the nucleotide frequencies between the sequences are different, the modified formula (Tamura and Kumar 2002) relaxes the assumption of the substitution pattern homogeneity.

The Tamura-Nei model

	A	T	C	G
A	-	βg_T	βg_C	$\alpha_1 g_G$
T	βg_A	-	$\alpha_2 g_C$	βg_G
C	βg_A	$\alpha_2 g_T$	-	βg_G
G	$\alpha_1 g_A$	βg_T	βg_C	-

MEGA 4 provides facilities for computing the following quantities for this method:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = s/v	Transition/transversions ratio.
L: No of valid common sites	Number of sites compared.

The formulas for computing these quantities are as follows:

Distances

$$d = a\left[k_1(w_1)^{-1/a} + k_2(w_2)^{-1/a} + k_3(w_3)^{-1/a} - k_4\right]$$

$$s = a\left[k_1(w_1)^{-1/a} + k_2(w_2)^{-1/a} + (k_3 - 2g_R g_Y)(w_3)^{-1/a} - (k_4 - 2g_R g_Y)\right]$$

$$v = 2ag_R g_Y\left[(w_3)^{-1/a} - 1\right]$$

$$R = s/v$$

where *P*1 and *P*2 are the proportions of transitional differences between nucleotides A and G, and between T and C, respectively, *Q* is the proportion of transversional differences, *gXA*, *gXC*, *gXG*, *gXT*, are the respective frequencies of A, C, G and T of sequence *X*, *gXR* = *gXA* + *gXG* and *gXY* = *gXT* + *gXC*, *gA*, *gC*, *gG*, *gT*, *gR*, and *gY* are the average frequencies of the pair of sequences, *a* is the gamma parameter and

$$k_1 = 2g_A g_G / g_R,$$
$$k_2 = 2g_T g_C / g_Y,$$
$$k_3 = 2(g_R g_Y - g_A g_G g_Y / g_R - g_T g_C g_R / g_Y),$$
$$k_4 = 2(g_A g_G + g_T g_C + g_R g_Y),$$
$$w_1 = 1 - P_1 / (g_{1A} g_{2G} + g_{1G} g_{2A}) - Q/2g_R,$$
$$w_2 = 1 - P_2 / (g_{1T} g_{2C} + g_{1C} g_{2T}) - Q/2g_Y,$$
$$w_3 = 1 - Q / (g_{1R} g_{2Y} + g_{1Y} g_{2R}).$$

The variances can be estimated by the bootstrap method.

Tamura 3 parameter (Gamma rates and Heterogeneous patterns)

Tamura's 3-parameter model corrects for multiple hits, taking into account the differences in transitional and transversional rates and the G+C-content bias (1992). Evolutionary rates among sites are modeled using the gamma distribution, and you will need to provide a gamma parameter for computing this distance. When the G+C-contents between the sequences are different, the modified formula (Tamura and Kumar 2002) relaxes the assumption of substitution pattern homogeneity.

The Tamura 3-parameter model

	A	T	C	G
A	-	β(1-θ)	βθ	αθ
T	β(1-θ)	-	αθ	βθ
C	β(1-θ)	α(1-θ)	-	βθ
G	α(1-θ)	β(1-θ)	βθ	-

MEGA 4 provides facilities for computing the following quantities:

Quantity	**Description**
d: Transitions & Transversions	Number of nucleotide substitutions per site.
s: Transitions only	Number of transitional substitutions per site.
v: Transversions only	Number of transversional substitutions per site.
R = *s*/*v*	Transition/transversion ratio.
L: No of valid common sites	Number of sites compared.

Formulas for computing these quantities are as follows:

Distances

$$d = \frac{a}{2}\left[w_\theta w_1^{-1/a} + (1 - w_\theta) w_2^{-1/a} - w_\theta - 1\right]$$

$$s = \frac{a}{2}\left[w_\theta w_1^{-1/a} - w_\theta w_2^{-1/a} - w_\theta\right]$$

$$v = \frac{a}{2}\left[w_2^{-1/a} - 1\right]$$

$$R = s/v$$

where *P* and *Q* are the proportion of sites with transitional and transversional differences, respectively, a is the gamma parameter, and

$$\begin{aligned}
\theta_1 &= g_{1C} + g_{1G}, \\
\theta_2 &= g_{2C} + g_{2G}, \\
\theta &= (\theta_1 + \theta_2)/2, \\
w_\theta &= 2\theta(1 - \theta), \\
w_1 &= 1 - P/\left[\theta_1(1 - \theta_2) + \theta_2(1 - \theta_1)\right] - Q, \\
w_2 &= 1 - 2Q.
\end{aligned}$$

The variances can be estimated by the bootstrap method.

Maximum Composite Likelihood (Gamma Rates and Heterogeneous Patterns)

The Tamura-Nei (1993) distance estimated by the composite likelihood method (Tamura et al. 2004) with the gamma model corrects for multiple hits, taking into account the rate substitution differences between nucleotides and the inequality of nucleotide frequencies. In this distance, evolutionary rates among sites are modeled using the gamma distribution. You will need to provide a gamma parameter for computing this distance. When the nucleotide frequencies between the sequences are different, the expected proportions of observed differences (P1, P2, Q) in the computation of the composite likelihood can be obtained by the modified formulas according to Tamura and Kumar (2002) to relax the assumption of the substitution pattern homogeneity.

Amino Acid Substitution Models

No. of differences (Amino acids)

This distance is the number of sites at which two sequences being compared are different. If the sequences contain alignment gaps or missing data and you are using the pairwise deletion option, you must realize that this count does not normalize the number of differences based on the number of valid sites compared. Therefore, if you use this distance, we recommend that you use the complete-deletion option.

MEGA computes the following quantities:

Quantity	Description
d: distance	Number of sites different.
L: No of valid common sites	Number of sites compared.

The formulas used are:

Quantity	Formula	Variance
n_d	None	$n_d(L-n_d)/L$

See also Nei and Kumar (2000), page 18.

p-distance (Amino acids)

This distance is the proportion (*p*) of amino acid sites at which the two sequences to be compared are different. It is obtained by dividing the number of amino acid differences by the total number of sites compared. It does not make any correction for multiple substitutions at the same site or differences in evolutionary rates among sites.

MEGA provides facilities to compute the following quantities:

Quantity	Description
d: distance	Proportion of amino acid sites different.
L: No of valid common sites	Number of sites compared.

The formulas used are:

Quantity	Formula	Variance
p	n_d/L	$p(1-p)/L$

where n_d is the number of amino acids that are different between two aligned sequences.

See also Nei and Kumar (2000), page 18.

Equal Input Model (Amino acids)

In real data, frequencies usually vary among different kind of amino acids. In this case, the correction based on the equal input model gives a better estimate of the number of amino acid substitutions than the Poisson correction distance. Note that this assumes an equality of substitution rates among sites and the homogeneity of substitution patterns between

lineages.

MEGA provides facilities to compute the following quantities:

Quantity	**Description**
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

The formulas used are:

Distance

$$d = -b\log(1 - p/b)$$

where *p* is the proportion of different amino acid sites, *gi* is the frequency of amino acid *i*, and

$$b = 1 - \sum_i g_i^2,$$

Variance

$$\mathrm{Var}(d) = b^2 p(1-p)/\left[(b-p)^2 L\right]$$

Poisson Correction (PC) distance

The Poisson correction distance assumes equality of substitution rates among sites and equal amino acid frequencies while correcting for multiple substitutions at the same site.

MEGA provides facilities to compute the following quantities:

Quantity	**Description**
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

Formulas used are:

Quantity	**Formula**	**Variance**
d	$-\ln(1-p)$	$p/\left[(1-p)L\right]$

See also Nei and Kumar (2000), page 20.

Dayhoff and JTT Models

The PAM and JTT distances correct for multiple substitutions based on the model of amino acid substitution described as substitution-rate matrices. The PAM distance uses the PAM 001 matrix (p. 348 in Dayhoff 1979) and the JTT distance uses the JTT matrix (Jones et al. 1992). Using a substitution-rate matrix (*Q*), the matrix (*F*), which consists of the observed proportions of amino acid pairs between a pair of sequences with their divergence time *t*, is given by the following equation

$$F(t) = Ae^{2tQ},$$

where *A* denotes the diagonal matrix of the equilibrium amino acid frequencies for *Q*. From this equation, the evolutionary distance $d = 2tQ$ can be iteratively computed by a maximum-likelihood method. The eigen values for the PAM and JTT matrices required in this computation were obtained from the program source code of PHYLIP version 3.6 (Felsenstein et al. 1993-2001).

MEGA provides facilities for computing the following quantities:

Quantity	**Description**
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

The variance of *d* can be estimated by the bootstrap method.

Gamma Distances

Dayhoff and JTT distances (Gamma rates)

The PAM and JTT distances correct for multiple substitutions based on a model of amino acid substitution described as substitution-rate matrices. The PAM distance uses PAM 001 matrix (p. 348 in Dayhoff 1979) and the JTT distance uses JTT matrix (Jones et al. 1992). The matrix (*F*) uses a substitution-rate matrix (*Q*) and the gamma distribution with parameter *a* for the rate variation among sites. It consists of the observed proportions of amino acid pairs with their divergence time *t*, given by the following equation

$$F(t) = A\left(\frac{a}{a - 2tQ}\right)^{a},$$

where *A* denotes the diagonal matrix of the equilibrium amino acid frequencies for *Q*. From this equation, the evolutionary distance $d = 2tQ$ can be computed iteratively by a maximum-likelihood method. The eigen values for the PAM and JTT matrices required in this computation were obtained from the program source code of PHYLIP version 3.6 (Felsenstein et al. 1993-2001).

MEGA provides facilities for computing the following quantities:

Quantity	Description
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

The variance of *d* can be estimated by the bootstrap method.

Gamma distance (Amino acids)

The Gamma distance improves upon the Poisson correction distance by taking care of the inequality of the substitution rates among sites. For this purpose, you will need to provide the gamma shape parameter (*a*).

For estimating the Dayhoff distance, use *a* = 2.25 (see Nei and Kumar [2000], page 21 for details).

For computing Grishin's distance, use *a* = 0.65. 23 (see Nei and Kumar [2000], page 23 for details)

MEGA provides facilities to compute the following quantities:

Quantity	Description
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

Formulas used are:

Quantity	Formula	Variance
d	$a\left[(1-p)^{-1/a}-1\right]$	$p\left[(1-p)^{-(1+2/a)}\right]/L$

See also Nei and Kumar (2000), page 23 and estimating gamma parameter.

Heterogeneous Patterns

Equal Input Model (Heterogeneous Patterns)

In real data, amino acid frequencies usually vary among different kinds of amino acids. In

this case, a correction based on the equal input model gives a better estimate of the number of amino acid substitutions than does the Poisson correction distance. Note that this assumes an equality of substitution rates among sites. When the amino acid frequencies are different between the sequences, the modified formula (Tamura and Kumar 2002) relaxes the estimation bias.

MEGA provides facilities for computing the following quantities:

Quantity	**Description**
d: distance	Number of amino acid substitutions per site.
L: No of valid common sites	Number of sites compared.

Formulas used are:

Distance

$$d = -b\log(1 - p/c)$$

where *p* is the proportion of different amino acid sites, *gXi* is the frequency of amino acid *i* for sequence *X*, *gi* is the average frequency for the pair of the sequences, and

$$b = 1 - \sum_i g_i^2,$$
$$c = 1 - \sum_i \sum_j g_{Xi} g_{Yj}$$

The variance of *d* can be estimated by the bootstrap method.

Synonymouse and Nonsynonymous Substitution Models

Nei-Gojobori Method

This method computes the numbers of synonymous and nonsynonymous substitutions and the numbers of potentially synonymous and potentially nonsynonymous sites (Nei and Gojobori 1986). Based on these estimates, *MEGA* can be asked to produce the following quantities:

Number of differences (*Sd* or *Nd*)

These are simple counts of the number of synonymous (*Sd*) and nonsynonymous (*Nd*) differences. To compare these two numbers, you must use the p-distance because the number of potential synonymous sites is much smaller than the number of nonsynonymous sites.

p-distance (*pS* or *pN*)

The count of the number of synonymous differences (*Sd*) is normalized using the

possible number of synonymous sites (*S*). A similar computation can be made for nonsynonymous differences.

Jukes-Cantor correction (*dS* or *dN*)

The *p*-distances computed above can be corrected to account for multiple substitutions at the same site.

Difference between synonymous and nonsynonymous distances

MEGA 4 can compute differences between the synonymous and nonsynonymous distances. These statistics are useful in conducting tests for selection.

Number of Sites (*S* or *N*)

The numbers of potential synonymous and nonsynonymous sites can be computed using this option. For each pair of sequences, the average number of synonymous or nonsynonymous sites is reported.

The formulas for computing these quantities are:

Quantity	**Formula**	**Variance**
p_S	S_d/S	$V(p_S) = p_S(1-p_S)/S$
p_N	N_d/N	$V(p_N) = p_N(1-p_N)/N$
d_S	$-\frac{3}{4}\ln\left(1-\frac{4}{3}p_S\right)$	$V(d_S) = p_S(1-p_S)\Big/\left[\left(1-\frac{4}{3}p_S\right)^2 S\right]$
d_N	$-\frac{3}{4}\ln\left(1-\frac{4}{3}p_N\right)$	$V(d_N) = p_N(1-p_N)\Big/\left[\left(1-\frac{4}{3}p_N\right)^2 N\right]$
Dp	$p_N - p_S$	$V(p_N)+V(p_S)$
Dd	$d_N - d_S$	$V(d_N)+V(d_S)$

See also Nei and Kumar (2000), page 52

Modified Nei-Gojobori Method

The modified Nei-Gojobori distance differs from the original Nei-Gojobori formulation in one way: transitional and transversional substitutions are no longer assumed to occur with the same frequency. Thus the user is requested to provide the Transition/Transversion (*R*) ratio. When $R = 0.5$, this method becomes identical to the Nei-Gojobori method. When $R > 0.5$, the number of synonymous sites is less than estimated using Nei-Gojobori method and consequently, the number of nonsynonymous sites will be larger than estimated with

the original Nei-Gojobori (Nei and Gojobori 1986) approach.

Number of differences (*Sd* or *Nd*)

These are counts of the numbers of synonymous (*Sd*) and nonsynonymous (*Nd*) differences. To compare these two numbers you must use the p-distance because the number of potential synonymous sites is much smaller than the number of nonsynonymous sites.

p-distance (*pS* or *pN*)

The count of the number of synonymous differences (*Sd*) is normalized using the number of potential synonymous sites (*S*). A similar computation can be made for nonsynonymous differences.

Jukes-Cantor correction (*dS* or *dN*)

The *p*-distances computed above can be corrected to account for multiple substitutions at the same site.

Difference between synonymous and nonsynonymous distances

MEGA 4 can compute differences between synonymous and nonsynonymous distances. These statistics are useful when conducting tests for selection.

Number of Sites (*S* or *N*)

Numbers of potentially synonymous and nonsynonymous sites can be computed using this option. For each pair of sequences, the average number of synonymous or nonsynonymous sites is reported.

The formulas for computing these quantities are:

Quantity	*Formula*	**Variance**
p_S	S_d / S_R	$V(p_s) = p_S(1-p_S)/S_R$
p_N	N_d / N_R	$V(p_N) = p_N(1-p_N)/N_R$
d_S	$-3/4 \ln(1-4/3\, p_S)$	$V(d_S) = p_S(1-p_S)\Big/\left[(1-4/3\, p_S)^2 S_R\right]$
d_N	$-3/4 \ln(1-4/3\, p_N)$	$V(d_N) = p_N(1-p_N)\Big/\left[(1-4/3\, p_N)^2 N_R\right]$
	$p_N - p_S$	$V(p_N)+V(p_S)$
D	$d_N - d_S$	$V(d_N)+V(d_S)$

See also Nei and Kumar (2000), page 52.

Li-Wu-Luo Method

In this method (Li et al 1985), each site in a codon is allocated to 0-fold, 2-fold or 4-fold degenerate categories. For computing distances, all 0-fold and two-thirds of the 2-fold sites are considered nonsynonymous, whereas one-third of the 2-fold and all of the 4-fold sites are considered synonymous. The observed transitional and transversional differences between codons then are partitioned into those occurring at 0-fold, 2-fold and 4-fold degenerate sites. Based on this information, the following quantities can be estimated.

Synonymous distance

This is the number of synonymous substitutions per synonymous site.

Nonsynonymous distance

This is the number of nonsynonymous substitutions per nonsynonymous site.

Substitutions at the 4-fold degenerate sites

This is the number of substitutions per 4-fold degenerate site; it is useful for measuring the rate of neutral evolution.

Substitutions at the 0-fold degenerate sites

This is the number of substitutions per 0-fold degenerate site; it is useful for measuring the rate of amino acid sequence evolution.

Number of 4-fold degenerate sites

This is the estimate of the number of 4-fold degenerate sites, computed by averaging the number of 4-fold degenerate sites in the two sequences, compared.

Number of 0-fold degenerate sites

This is the estimate of the number of 0-fold degenerate sites, computed by averaging the number of 0-fold degenerate sites in the two sequences, compared.

Difference between synonymous and nonsynonymous distances

This computes the differences between the synonymous and nonsynonymous distances. These statistics are useful for conducting tests of selection.

The formulas for computing these quantities are:

Quant	**Formula**	**Variance**

ity

d_S	$\frac{3\left[L_2A_2 + L_4(A_4 + B_4)\right]}{(L_2 + 3L_4)}$	$\frac{9\left[L_2^2V(A_2) + L_4^2V(A_4 + B_4)\right]}{(L_2 + 3L_4)^2}$
d_N	$\frac{3\left[L_2B_2 + L_0(A_0 + B_0)\right]}{(2L_2 + 3L_0)}$	$\frac{9\left[L_2^2V(B_2) + L_0^2V(A_0 + B_0)\right]}{(2L_2 + 3L_0)^2}$
d_4	$A_4 + B_4$	$\left[a_4^2P_4 + k_4^2Q_4 - \left(a_4P_4 + k_4^2Q_4\right)^2\right] / L$
d_0	$A_0 + B_0$	$\left[a_0^2P_0 + k_0^2Q_0 - \left(a_0P_0 + k_0^2Q_0\right)^2\right] / L$
D	$d_N - d_S$	$V(d_N) + V(d_S)$

Here,

L_0, L_2 and L_4 are the number of 0-fold, 2-fold and 4-fold degenerate sites, respectively.

$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i)$, and

$B_i = \frac{1}{2}\ln(b_i)$, where

$a_i = 1/(1 - 2P_i - Q_i)$, $b_i = 1/(1 - 2Q_i)$, $c_i = (a_i - b_i)/2$, $k_i = (a_i + b_i)/2$

Pi and *Qi* are the proportions of i-fold degenerate sites that show transitional and transversional differences, respectively.

$V(A_i) = \left[a_i^2P_i + c_i^2Q_i - (a_iP_i + c_iQ_i)^2\right] / L_i$,

$V(B_i) = b_i^2Q_i(1 - Q_i)/L_i$

See also Nei and Kumar (2000), page 62.

Pamilo-Bianchi-Li Method

This method (Pamilo and Bianchi 1993; Li 1993) is a modification of Li, Wu and Luo's method. The only difference concerns the allocation of 2-fold sites to synonymous and nonsynonymous categories. Rather than assuming an equal transition and transversion rate, the rate is inferred from the observed number of transitions and transversions at the 4-fold

degenerate sites. Based on this information, the following quantities can be estimated:

Synonymous distance

This is the number of synonymous substitutions per synonymous site.

Nonsynonymous distance

This is the number of nonsynonymous substitutions per nonsynonymous site.

Substitutions at the 4-fold degenerate sites (*d4*)

This is the number of substitutions per 4-fold degenerate site; it is useful for measuring the rate of neutral evolution.

Substitutions at the 0-fold degenerate sites (*d0*)

This is the number of substitutions per 0-fold degenerate site; it is useful for measuring the rate of amino acid sequence evolution.

Number of 4-fold degenerate sites(*L4*)

The estimate of the number of 4-fold degenerate sites, computed by averaging the number of 4-fold degenerate sites in the two sequences, compared.

Number of 0-fold degenerate sites (*L0*)

The estimate of the number of 0-fold degenerate sites, computed by averaging the number of 0-fold degenerate sites in the two sequences, compared.

Difference between synonymous and nonsynonymous distances (*D*)

This computes the differences between the synonymous and nonsynonymous distances. These statistics are useful for conducting tests of selection.

The formulas for computing these quantities are:

Quantity	**Formula**	**Variance**
d_S	$B_4 + \frac{(L_2 A_2 + L_4 A_4)}{(L_2 + L_4)}$	$V(B_4) + \frac{\left[L_2^2 V(A_2) + L_4^2 V(A_4)\right]}{(L_2 + L_4)^2} - \frac{b_4 Q_4 \left[2a_4 P_4 - c_4(1 - Q_4)\right]}{(L_2 + L_4)}$
d_N	$A_0 + \frac{(L_0 B_0 + L_2 B_2)}{(L_0 + L_2)}$	$V(A_0) + \frac{\left[L_0^2 V(B_0) + L_2^2 V(B_2)\right]}{(L_0 + L_2)^2} - \frac{b_0 Q_0 \left[2a_0 P_0 - c_0(1 - Q_0)\right]}{(L_0 + L_2)}$

d4	$A_4 + B_4$	$\left[a_4^2 P_4 + k_4^2 Q_4 - \left(a_4 P_4 + k_4^2 Q_4\right)^2\right] / L$
d0	$A_0 + B_0$	$\left[a_0^2 P_0 + k_0^2 Q_0 - \left(a_0 P_0 + k_0^2 Q_0\right)^2\right] / L$
D	$d_S - d_N$	$V(d_S) + V(d_N) - 2\mathrm{cov}(d_S, d_N)$
Ai	$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(i$	$V(A_i) = \left[a_i^2 P_i + c_i^2 Q_i - \left(a_i P_i + c_i Q_i\right)^2\right] / L_i$
Bi	$B_i = \frac{1}{2}\ln(b_i)$	$V(B_i) = b_i^2 Q_i (1 - Q_i) / L_i$

Here,

L_0, L_2 and L_4 are the number of 0-fold, 2-fold and 4-fold degenerate sites, respectively.

$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i)$, and

$B_i = \frac{1}{2}\ln(b_i)$, where

$a_i = 1/(1 - 2P_i - Q_i)$, $b_i = 1/(1 - 2Q_i)$, $c_i = (a_i - b_i)/2$, $k_i = (a_i + b_i)/2$

Pi and *Qi* are the proportions of i-fold degenerate sites that show transitional and transversional differences, respectively.

$V(A_i) = \left[a_i^2 P_i + c_i^2 Q_i - \left(a_i P_i + c_i Q_i\right)^2\right] / L_i$,

$V(B_i) = b_i^2 Q_i (1 - Q_i) / L_i$

See also Nei and Kumar (2000), page 64.

Kumar Method

This method is a modification of the Pamilo-Bianchi-Li and Comeron (1995) methods and is able to handle some problematic degeneracy class assignments (see a detailed description below). It computes the following quantities:

Synonymous distance

This is the number of synonymous substitutions per synonymous site.

Nonsynonymous distance

This is the number of nonsynonymous substitutions per nonsynonymous site.

Substitutions at the 4-fold degenerate sites

This is the number of substitutions per 4-fold degenerate site. It is useful for measuring the rate of neutral evolution.

Substitutions at the 0-fold degenerate sites

This is the number of substitutions per 0-fold degenerate site. It is useful for measuring the rate of amino acid sequence evolution.

Number of 4-fold degenerate sites

This is the estimate of the number of 4-fold degenerate sites, computed by averaging the number of 4-fold degenerate sites in the two sequences, compared.

Number of 0-fold degenerate sites

This is the estimate of the number of 0-fold degenerate sites, computed by averaging the number of 0-fold degenerate sites in the two sequences, compared.

Difference between synonymous and nonsynonymous distances

This computes the differences between the synonymous and nonsynonymous distances. These statistics are useful for conducting tests of selection.

Kumar's modification of the PBL method:

The treatment of arginine and isoleucine codons in the Li-Wu-Luo and the Pamilo-Bianchi-Li methods is arbitrary, which sometimes creates a problem because the arginine codons occur quite frequently. Comeron (1995) addressed this problem by dividing the 2-fold degenerate sites into two groups: 2S-fold and 2V-fold. The 2S-fold refers to sites in which the transitional change is synonymous and the two transversional changes are nonsynonymous, whereas the 2V-fold represents sites in which the transitional change is nonsynonymous and the transversional changes are synonymous. Although these definitions help in correcting some of the inaccurate classifications of synonymous and nonsynonymous sites (e.g., methionine codons), they do not solve the problem completely. For example, consider mutations in the first nucleotide position of the arginine codon: CGG produces TGG (Trp), AGG (Arg), or GGG (Gly). The transitional change (C to T) results in a nonsynonymous change. Of the two transversional substitutions, one (C to A) results in a synonymous change, while the other (C to G) results in a nonsynonymous change. Therefore, this nucleotide site is neither a 2S-fold nor a 2V-fold site. Thus, the first position of three arginine codons (CGU, CGC, and CGA) and the third position of two isoleucine codons (ATT and ATC) cannot be assigned to any of the Comeron (1995) categories. For this reason, Comeron (personal communication) used a more complicated classification of codons when he wrote his computer program. For example, the first position of arginine codon CGG was assigned to a 2V-fold site with a probability of one-

third and to a 0-fold site with a probability of two-thirds. Similar assignments are used by W.-H. Li (personal communication) in his computer program.

Since the nucleotide site assignments discussed above are quite arbitrary and may not apply to all known genetic code tables, Kumar developed another method that uses the PBL method for any genetic code table. In this version, nucleotide sites are first classified into 0-fold, 2-fold, and 4-fold degenerate sites. The 2-fold degenerate sites are further subdivided into simple 2-fold and complex 2-fold degenerate sites. Simple 2-fold sites are those at which the transitional change results in a synonymous substitution and the two transversional changes result in nonsynonymous substitutions. All other 2-fold sites, including those for the three isoleucine codons, belong to the complex 2-fold site category. If we use this definition, all nucleotide sites can be classified into the five groups shown in the following table.

Table.

Degeneracy ->	0-fold	Simple 2-fold	Complex 2-fold		4-fold
No. of sites ->	*L*0	*L*2S	*L*2C		*L*4
			Syn	*Nonsyn*	
Transition (*s*)	*s*0	*s*2	*s*2S	*s*2N	*s*4
Transversion (*v*)	*v*0	*V*2	*v*2S	*v*2N	*v*4

Here, *L*0, *L*2S, *L*2C, and *L*4 are the numbers of 0-fold, simple 2-fold, complex 2-fold, and 4-fold degenerate sites, respectively.

Once this table is filled using the observed counts for a given pair of sequences, we compute the proportions of transitional (*P*i) and transversional (Qi) differences for the *i*-fold degenerate site in the following way:

$$P_0 = \frac{s_0 + s_{2N}}{L_0 + L_{2C}}, \qquad Q_0 = \frac{v_0}{L_0},$$

$$P_2 = \frac{s_0 + s_{2S}}{L_{2S} + L_{2C}}, \qquad Q_2 = \frac{v_2 + v_{2N}}{L_{2S} + L_{2C}},$$

$$P_4 = \frac{s_4}{L_4}, \qquad Q_4 = \frac{v_4 + v_{2S}}{L_4 + L_{2C}}$$

From these quantities, we compute the *A*i and *B*i as in the PBL method. Then using *L*2 = *L*2C + *L*2S, we apply the formulas for the PBL method.

See also Nei and Kumar (2000), page 64.

5.22 Specifying Distance Estimation Options

Analysis Preferences (Distance Computation)

In this dialog box you can select and view the desired options in the **Options Summary**. Options are organized in logical sections. A lime square in the right-most cell in a row indicates that you have a choice regarding the attribute in that row The three primary sets of options available in this dialog box are:

Analysis

Compute

Use this to specify whether to compute *Distances only* or *Distances and Standard Errors*. If you select the latter, then you are given a choice as to how to compute it in the *Standard Error Computation* box.

Standard Error Computation By

This row is visible only if you have chosen *Distances and Std. Err* in the *Compute* row. You may choose to use analytical formulas or the bootstrap method to calculate standard errors depending on the type of distance computed. Whenever the standard errors are estimated by the bootstrap method, you will be prompted for the number of bootstrap replicates and a random number seed.

When you compute average distance or diversity, only the bootstrap method is available for computing standard errors.

Include Sites

These are options for handling gaps or missing data, including or excluding codon positions, and restricting the analysis to labeled sites, if applicable.

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing information before the calculation begins (Complete-deletion option). Alternatively, you may choose to retain all such sites initially, excluding them as necessary in the pairwise distance estimation (Pairwise-deletion option).

Codon Positions

Click on the ellipses or the lime square, for the option of selecting any combination of 1^{st}, 2^{nd}, 3^{rd}, and non-coding positions for analysis. This option is available only if the nucleotide sequences contain protein-coding regions and you have selected a nucleotide-by-nucleotide analysis.

Labeled Sites

This option is available only if some or all of the sites have associated labels. By clicking on the ellipses, you will be provided with the option of including sites with

selected labels. If you choose to include only labeled sites, then these sites will be the first extracted from the data. Then all other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon to be incorporated in the analysis.

Substitution Model

In this set of options, you choose the various attributes of the substitution models.

Model

Here you select a stochastic model for estimating evolutionary distance by clicking on the ellipses to the right of the currently selected model (click on the lime square to select this row first). This will reveal a menu containing many different distance methods and models.

Substitutions to Include

Depending on the distance model or method selected, the evolutionary distance can be teased into two or more components. By clicking on the drop-down button (first click on the lime square to select this row), you will be provided with a list of components relevant to the chosen model.

Transition/Transversion Ratio

This option will be visible if the chosen model requires you to provide a value for the Transition/Transversion ratio (R).

Pattern among Lineages

This option becomes available if the selected model has formulas that allow the relaxation of the assumption of homogeneity of substitution patterns among lineages.

Rates among Sites

This option becomes available if the selected distance model has formulas that allow rate variation among sites. If you choose gamma-distributed rates, then the Gamma parameter option becomes visible.

Distance Model Options

With this option, you can choose the general attributes of the substitution models for DNA and protein sequence evolution.

Model

You can select a stochastic model for estimating evolutionary distances by clicking on the ellipses to the right of the currently selected model (click on the lime square to select this row first). This will reveal a menu containing many different distance methods/models.

Transition/Transversion Ratio

This option will be visible if the chosen model requires you to provide a value for the Transition/Transversion ratio (R).

Pattern among Lineages

This option becomes available if the distance model you have selected has formulas that allow the relaxation of the assumption of homogeneity of substitution patterns among lineages.

Rates among Sites

This option becomes available if the distance model you have selected has formulas that allow rate variation among sites. If you choose gamma distributed rates, then the Gamma parameter option becomes visible.

Bootstrap method to compute standard error of distance estimates

When you choose the bootstrap method for estimating the standard error, you must specify the number of replicates and the seed for the psuedorandom number generator. In each bootstrap replicate, the desired quantity is estimated and the standard deviation of the original values are computed (see Nei and Kumar [2000], page 25 for details).

It is possible that in some bootstrap replicates the quantity you desire is not calculable for statistical or technical reasons. In these cases, *MEGA* will discard the results of the bootstrap replicates and its final estimate will be the results of all valid replicates. This means that the number of bootstrap replicates used can be smaller than the number specified by the user. However, if the number of valid bootstrap replicates is < 25, then *MEGA* will report that the standard error cannot be computed (an "n/c" swill appear in the result window).

5.23 Compute Pariwise

5.24 Compute Means

5.25 Compute Sequence Diversity

5.3 Constructing Phylogenetic Trees

5.31 Phylogenetic Inference

Reconstruction of the evolutionary history of genes and species is currently one of the most important subjects in molecular evolution. If reliable phylogenies are produced, they will shed light on the sequence of evolutionary events that generated the present day diversity of genes and species and help us to understand the mechanisms of evolution as well as the history of organisms.

Phylogenetic relationships of genes or organisms usually are presented in a treelike form with a root, which is called a *rooted tree*. It also is possible to draw a tree without a root, which is called an *unrooted tree*. The branching pattern of a tree is called a topology.

There are numerous methods for constructing phylogenetic trees from molecular data (Nei and Kumar 2000). They can be classified into *Distance methods*, *Parsimony methods*, and *Likelihood methods*. These methods are explained in Swofford et al. 1996, Li (1997), Page and Holmes (1998), and Nei and Kumar (2000).

5.32 NJ/UPGMA Methods

Analysis Preferences (NJ/UPGMA)

In this dialog box, you can view and select desired options in the **Options Summary**. Options are organized in logical sections. A lime square in the right cell of a row indicates that you have a choice for that attribute. The three primary sets of options available in this dialog box are:

Phylogeny Test and Options

To assess the reliability of a phylogenetic tree, *MEGA* provides two different types of tests: the *Bootstrap test* and the *Interior branch test*. Both of these tests use the bootstrap re-sampling strategy, so you need to enter the *number of replicates* and a starting *random seed*. For a given data set applicable tests and the phylogeny inference method are enabled.

Include Sites

These are options for handling gaps and missing data, including or excluding codon positions, and restricting the analysis to labeled sites, if applicable.

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing-information before the calculation begins using the Complete-deletion option. Alternatively, you may choose to retain all such sites initially, excluding them as necessary using the Pairwise-deletion option.

Codon Positions

By clicking on the ellipses or the lime square, you may select any combination of 1^{st}, 2^{nd}, 3^{rd}, and non-coding positions for analysis. This option is available only if the nucleotide sequences contain protein-coding regions and you have selected a nucleotide-by-nucleotide analysis.

Labeled Sites

This option is available only if there are labels associated with some or all of the sites in the data. By clicking on the ellipses, you will have the option of including sites with selected labels. If you chose to include only labeled sites, then these sites will be first extracted from the data and all other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon in the analysis.

Substitution Model

In this set of options, you can choose various attributes of the substitution models for DNA and protein sequences.

Model

By clicking on the ellipses to the right of the currently selected model, you may select a stochastic model (method) for estimating evolutionary distance (click on the lime square to select this row first). This will reveal a menu containing many different distance methods and models.

Substitutions to Include

Depending on the distance model or method selected, the evolutionary distance can be teased into two or more components. By clicking on the drop-down button (first click on the lime square to select this row), you will be provided with a list of components relevant to the chosen model.

Transition/Transversion Ratio

This option will be visible if the chosen model requires you to provide a value for the Transition/Transversion ratio (R).

Pattern among Lineages

This option becomes available if the selected model has formulas that allow the relaxation of the assumption of homogeneity of substitution patterns among lineages.

Rates among Sites

This option becomes available if the selected distance model has formulas that allow rate variation among sites. If you choose gamma-distributed rates, then the Gamma parameter option becomes visible.

5.33 Minimum Evolution Method

Minimum Evolution

In the ME method, distance measures that correct for multiple hits at the same sites are used, and a topology showing the smallest value of the sum of all branches (*S*) is chosen as an estimate of the correct tree. However, the construction of a minimum evolution tree is time-consuming because, in principle, the *S* values for all topologies must be evaluated. The number of possible topologies (unrooted trees) rapidly increases with the number of taxa so it becomes very difficult to examine all topologies. In this case, one may use the neighbor-joining method. While the NJ tree is usually the same as the ME tree, when the number of taxa is small the difference between the NJ and ME trees can be substantial (reviewed in Nei and Kumar 2000). In this case if a long DNA or amino acid sequence is used, the ME tree is preferable. When the number of nucleotides or amino acids used is relatively small, the NJ method generates the correct toplogy more often than does the ME method (Nei et al. 1998, Takahashi and Nei 2000). In *MEGA*, we have provided the close-neighbor-interchange search to examine the neighborhood of the NJ tree to find the potential ME tree.

Analysis Preferences (Minimum Evolution)

In this dialog box you can select and view desired options in the **Options Summary**. Options are organized in logical sections. A lime square in the right cell of a row indicates that you have a choice for that particular attribute. The primary sets of options available in this dialog box are:

Tree Inference

Phylogeny Test and Options

To assess the reliability of a phylogenetic tree, *MEGA* provides two different types of tests: the *Bootstrap test* and the *Interior branch test*. Both of these tests use the bootstrap re-sampling strategy, so you need to enter the *number of replicates* and a starting *random seed*. For a given data set, applicable tests and the phylogeny inference method are enabled.

Search Options

This sets the extensiveness of the heuristic search for the Minimum Evolution (ME) tree. *MEGA* employs the Close-Neighbor-Interchange (CNI) algorithm for finding the ME tree. It is a branch swapping method, which begins with an initial NJ tree.

Include Sites

These are options for handling gaps and missing data, including or excluding codon positions, and restricting the analysis to labeled sites, if applicable.

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing information before the calculation begins using Complete-deletion option. Alternatively, you may choose to retain all such sites initially, excluding them as necessary using the (Pairwise-deletion option).

Codon Positions

By clicking on the ellipses or the lime square, you may select any combination of 1^{st}, 2^{nd}, 3^{rd}, and non-coding positions for analysis. This option is available only if the nucleotide sequences contain protein-coding regions and you have selected a nucleotide-by-nucleotide analysis.

Labeled Sites

This option is available only if there are labels associated with some or all of the sites in the data. By clicking on the ellipses, you will have the option of including sites with selected labels. If you chose to include only labeled sites, then these sites will be first extracted from the data and all other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon in the analysis.

Substitution Model

In this set of options, you can choose various attributes of the substitution models for DNA and protein sequences.

Model

By clicking on the ellipses to the right of the currently selected model, you may select a stochastic model for estimating evolutionary distance (click on the lime square to select this row first). This will reveal a menu containing many different distance methods and models.

Substitutions to Include

Depending on the distance model or method selected, the evolutionary distance can be teased into two or more components. By clicking on the drop-down button (first click on the lime square to select this row), you will be provided with a list of components relevant to the chosen model.

Transition/Transversion Ratio

This option will be visible if the chosen model requires you to provide a value for the Transition/Transversion ratio (R).

Pattern among Lineages

This option becomes available if the selected model has formulas that allow the relaxation of the assumption of homogeneity of substitution patterns among lineages.

Rates among Sites

This option becomes available if the selected distance model has formulas that allow rate variation among sites. If you choose gamma-distributed rates, then the Gamma parameter option becomes visible.

5.34 Maximum Parsimony (MP) Method

5.35 Branch-and-Bound algorithm

The branch-and-bound algorithm is used to find all the MP trees . It guarantees to find all the MP trees without conducting an exhaustive search. *MEGA* also employs the Max-mini branch-and-bound search, which is described in detail in Kumar et al. (1993) and Nei and Kumar (2000, page 123).

Alignment Gaps and Sites with Missing Information

In *MEGA*, gap sites are ignored in the MP analysis, but there are two different ways to treat these sites. One is to delete all of these sites from data analysis. This option, called the *Complete-Deletion* option, is generally desirable because different regions of DNA or amino acid sequences often evolve under different evolutionary forces. However, if the number of nucleotides (or amino acids) involved in a gap is small and gaps are distributed more or less randomly, you may include all such sites and treat them as missing data. Therefore, gaps and missing data are never used in computing tree lengths in *MEGA 4*.

Consensus Tree

The MP method produces many equally parsimonious trees. Choosing this command produces a composite tree that is a consensus among all such trees, for example, either as a strict consensus, in which all conflicting branching patterns among the trees are resolved by making those nodes multifurcating or as a Majority-Rule consensus, in which conflicting branching patterns are resolved by selecting the pattern seen in more than 50% of the trees.

(Details are given in Nei and Kumar [2000], page 130).

Analysis Preferences (Maximum Parsimony)

This dialog box contains four overlapping pages, with each page marked by *Tab*s running across the top. You can go to any page by simply clicking on the Tab. Each tab page organizes a set of logically related options. Information from all the pages is used in the requested analysis, so it is important that you examine the options selected in each tab before pressing *OK* to proceed with analysis.

Phylogeny Test and Options

To assess the reliability of the MP trees, *MEGA* provides the bootstrap test. You need to enter the *number of replicates* and a starting *random seed* for this test.

Search Options

Use this to select between the branch-and-bound and the heuristic (close-neighbor interchange) searches. For the branch-and-bound search, an optimized Max-mini branch-and-bound algorithm is used. While this algorithm is guaranteed to find all the MP trees, a branch-and-bound search often is too time consuming for more than 15 sequences, although this number varies from data set to data set. Alternatively, you may use the heuristic search (Close-Neighbor-Interchange)., a branch swapping method that begins with a given initial tree. You may automatically obtain a set of initial trees by using the Min-mini algorithm with a given search factor. Alternatively, you can use the random addition option to produce the initial trees.

Include Sites

This provides options for handling gaps and missing data in the analysis, specifying inclusion and exclusion of codon positions, and restricting the analysis to only some types of labeled sites (if applicable).

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing-information before the parsimony analysis begins using the Complete-deletion option. Alternatively, you may choose to retain all such sites. In this case, all missing-information and alignment gap sites are treated as missing data in the calculation of tree length.

Codon Positions

By clicking on the ellipses (or the lime square), you may select any combination of 1^{st}, 2^{nd}, 3^{rd}, and non-coding positions for analysis. This option is available *only if* the

nucleotide sequences contain protein-coding regions. If they do, you can choose between the analysis of nucleotide sequences or translated protein sequences. If you choose the latter, *MEGA* will translate all protein-coding regions into amino acid sequences and conduct the protein sequence parsimony analysis.

Labeled Sites

This option is available only if there are labels associated with some or all of the sites in the data. By clicking on the ellipses, you will have the option of including sites with selected labels. If you choose to include only labeled sites, then these sites will be the first extracted from the data and all other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon to be incorporated in the analysis.

Heuristic Search

5.36 Min-mini algorithm

This is a heuristic search algorithm for finding the MP tree, and is somewhat similar to the branch-and bound search method. However, in this algorithm, many trees that are unlikely to have a small local tree length are eliminated from the computation of their *L* values. Thus while the algorithm speeds up the search for the MP tree, as compared to the branch-and-bound search, the final tree or trees may not be the true MP tree(s). The user can specify a search factor to control the extensiveness of the search and *MEGA* adds the user specified search factor to the current local upper bound. Of course, the larger the search factor, the slower the search, since many more trees will be examined.

(See also Nei & Kumar (2000), pages 122, 125)

Close-Neighbor-Interchange (CNI)

In any method, examining all possible topologies is very time consuming. This algorithm reduces the time spent searching by first producing a temporary tree, (e.g., an NJ tree when an ME tree is being sought), and then examining all of the topologies that are different from this temporary tree by a topological distance of d_T = 2 and 4. If this is repeated many times, and all the topologies previously examined are avoided, one can usually obtain the tree being sought.

For the MP method, the CNI search can start with a tree generated by the random addition of sequences. This process can be repeated multiple times to find the MP tree.

See Nei & Kumar (2000) for details.

5.37 Maximum Composite Likelihood Method

Maximum Composite Likelihood Method

Maximum Composite Likelihood (MCL) method is used for estimating evolutionary distances between all pair of sequences simultaneously, with and without incorporating rate variation among sites and substitution pattern heterogeneities among lineages. It can also be used to estimate transition/transversion bias and nucleotide substitution pattern without requiring a priori knowledge of the phylogenetic tree.

5.38 Statistical Tests of a Tree Obtained

General Comments on Statistical Tests

There are two different types of methods for testing the reliability of an obtained tree. One is to test the topological difference between the tree and its closely related tree by using a certain quantity, for example, the sum of all branch lengths in the minimum evolution method. This type of test examines the reliability of every interior branch of the tree, and is generally a conservative test as compared to other tests included in *MEGA*.

The other type of test examines the reliability of each interior branch whether or not it is significantly different from 0. If a particular interior branch is not significantly different from 0, we cannot exclude the possibility of a trifurcation of the associated branches or that the other types of bifurcating trees can be generated by changing the splitting order of the three branches involved. Therefore, in *MEGA* we implement the bootstrap procedure for estimating the standard error of the interior branch and test the deviation of the branch length from 0 (Dopazo 1994).

The third type of test is the bootstrap test, in which the reliability of a given branch pattern is ascertained by examining the frequency of its occurrence in a large number of trees, each based on the resampled dataset.

Details of these procedures are given in Nei and Kumar (2000, chapter 9).

Condensed Trees

When several interior branches of a phylogenetic tree have low statistical support (*PC* or *PB*) values, it often is useful to produce a multifurcating tree by assuming that all interior branches have a branch length equal to 0. We call this multifurcating tree a *condensed tree*. In *MEGA*, condensed trees can be produced for any level of *PC* or *PB* value. For example, if there are several branches with *PC* or *PB* values of less than 50%, a condensed tree with the 50% *PC* or *PB* level will have a multifurcating tree with all its branch lengths reduced to 0.

Since branches of low significance are eliminated to form a condensed tree, this tree emphasizes the reliable portions of branching patterns. However, this tree has one drawback. Since some branches are reduced to 0, it is difficult to draw a tree with proper branch lengths for the remaining portion. Therefore we give our attention only to the topology so the branch lengths of a condensed tree in *MEGA* are not proportional to the number of nucleotide or amino acid substitutions.

Note that, although they may look similar, condensed trees are different from the consensus trees mentioned earlier. A consensus tree is produced from many equally parsimonious trees, whereas a condensed tree is merely a simplified version of a tree. A condensed tree can be produced for any type of tree (NJ, ME, UPGMA, MP, or maximum-likelihood tree).

See also Nei and Kumar (2000) page 175.

Interior Branch Tests

Interior Branch Test of Phylogeny

Phylogeny | Interior Branch Test of Phylogeny

A *t*-test, which is computed using the bootstrap procedure, is constructed based on the interior branch length and its standard error and is available only for the NJ and Minimum Evolution trees. *MEGA* shows the confidence probability in the Tree Explorer; if this value is greater than 95% for a given branch, then the inferred length for that branch is considered significantly positive.

See Nei and Kumar (2000) (chapter 9) for further details.

Neighbor Joining (Construct Phylogeny)

Phylogeny | Construct Phylogeny | Neighbor-Joining…

This command is used to construct a neighbor-joining (NJ) tree (Saitou & Nei 1987). The NJ method is a simplified version of the minimum evolution (ME) method, which uses distance measures to correct for multiple hits at the same sites, and chooses a topology showing the smallest value of the sum of all branches as an estimate of the correct tree. However, the construction of an ME tree is time-consuming because, in principle, the *S* values for all topologies have to be evaluated and the number of possible topologies (unrooted trees) rapidly increases with the number of taxa.

With the NJ method, the *S* value is not computed for all or many topologies. The examination of different topologies is imbedded in the algorithm, so that only one final tree is produced. This method does not require the assumption of a constant rate of evolution so it produces an unrooted tree. However, for ease of inspection, *MEGA* displays NJ trees in a manner similar to rooted trees. The algorithm of the NJ method is somewhat complicated and is explained in detail in Nei and Kumar (2000).

For constructing the NJ tree, *MEGA* may request that you specify the distance estimation method, subset of sites to include, and whether to conduct a test of the inferred tree through an *Analysis Preferences* dialog box.

Bootstrap Tests

Bootstrap Test of Phylogeny

Phylogeny | Bootstrap Test of Phylogeny

One of the most commonly used tests of the reliability of an inferred tree is Felsenstein's

(1985) bootstrap test, which is evaluated using Efron's (1982) bootstrap resampling technique. If there are *m* sequences, each with *n* nucleotides (or codons or amino acids), a phylogenetic tree can be reconstructed using some tree building method. From each sequence, *n* nucleotides are randomly chosen with replacements, giving rise to *m* rows of *n* columns each. These now constitute a new set of sequences. A tree is then reconstructed with these new sequences using the same tree building method as before. Next the topology of this tree is compared to that of the original tree. Each interior branch of the original tree that is different from the bootstrap tree the sequences it partitions is given a score of 0; all other interior branches are given the value 1. This procedure of resampling the sites and the subsequent tree reconstruction is repeated several hundred times, and the percentage of times each interior branch is given a value of 1 is noted. This is known as the bootstrap value. As a general rule, if the bootstrap value for a given interior branch is 95% or higher, then the topology at that branch is considered "correct". See Nei and Kumar (2000) (chapter 9) for further details.

This test is available for four different methods: Neighbor Joining, Minimum Evolution, Maximum Parsimony, and UPGMA.

5.39 Handling Missing Data and Alignment Gaps

Alignment Gaps and Sites with Missing Information

Gaps often are inserted during the alignment of homologous regions of sequences and represent deletions or insertions (indels). They introduce some complications in distance estimation. Furthermore, sites with missing information sometimes result from experimental difficulties; they present the same alignment problems as gaps. In the following discussion, both of these situations are treated in the same way.

In *MEGA*, there are two ways to treat gaps. One is to delete all of these sites from the data analysis. This option, called the *Complete-Deletion*, is generally desirable because different regions of DNA or amino acid sequences evolve under different evolutionary forces. The second method is relevant if the number of nucleotides involved in a gap is small and if the gaps are distributed more or less randomly. In that case it may be possible to compute a distance for each pair of sequences, ignoring only those gaps that are involved in the comparison; this option is called *Pairwise-Deletion*. The following table illustrates the effect of these options on distance estimation with the following three sequences:

```
          1         10         20
seq1      A-AC-GGAT-AGGA-ATAAA
seq2      AT-CC?GATAA?GAAAAC-A
seq3      ATTCC-GA?TACGATA-AGA          Total sites = 20.
```

Here, the alignment gaps are indicated with a hyphen (-) and the missing information sites are denoted by a question mark (?).

Complete-Deletion and *Pairwise-Deletion* options

		Differences/Comparisons		
Option	**Sequence Data**	**(1, 2)**	**(1, 3)**	**(2,3)**
Compl	1. A C GA A GA A A A	1/1	0/1	1/10

ete deletion	2. A C GA A GA A C A 3. A C GA A GA A A A	0	0	
Pairwise Deletion	1. A-AC-GGAT-AGGA-ATAAA 2. AT-CC?GATAA?GAAAAC-A 3. ATTCC-GA?TACGATA-AGA	2/12	3/13	3/14

In the above table, the number of compared sites varies with pairwise comparisons in the *Pairwise-Deletion* option, but remains the same for pairwise comparisons in the *Complete-Deletion* option. In this data set, more information can be obtained by using the *Pairwise-Deletion* option. In practice, however, different regions of nucleotide or amino acid sequences often evolve differently, in which case, the *Complete-Deletion* option is preferable.

Include Sites Option

With this command you can set the options for handling gaps and missing data in the analysis, such as including or excluding codon positions, and restricting the analysis to only some types of labeled sites, if applicable.

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing information before the parsimony analysis begins (Complete-deletion option). Alternatively, you may choose to retain all such sites. In this case, all missing-information and alignment gap sites are treated as missing data in the calculation of tree length.

Codon Positions

By clicking on the ellipses (revealed by clicking on the lime-colored square), you will be provided with the option of selecting any combination of 1^{st}, 2^{nd}, 3^{rd}, and non-coding positions for analysis. This option is available only if the nucleotide sequences contain protein-coding regions. If it does, you can choose between the analysis of nucleotide sequences or translated protein sequences. If the latter is chosen, *MEGA* will translate all protein-coding regions into amino acid sequences and conduct the protein sequence parsimony analysis.

Labeled Sites

This option is available only if you have labels associated with some or all of the sites in the data. By clicking on the ellipses, you will be provided with the option of including sites with selected labels. If you choose to include only labeled sites, then these sites will be the first extracted from the data. Then all other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon to be incorporated in the analysis.

5.4 Tests of Selection

5.41 Synonymous/Nonsynonymous Tests

Large Sample Tests of Selection

One way to test whether positive selection is operating on a gene is to compare the relative abundance of synonymous and nonsynonymous substitutions that have occurred in the gene sequences. For a pair of sequences, this is done by first estimating the number of synonymous substitutions per synonymous site (dS) and the number of nonsynonymous substitutions per nonsynonymous site (dN), and their variances: Var(dS) and Var(dN), respectively. With this information, we can test the null hypothesis that H0: dN = dS using a Z-test:

$$Z = (dN - dS) / SQRT(Var(dS) + Var(dN))$$

The level of significance at which the null hypothesis is rejected depends on the alternative hypothesis (H1) .

H0: dN = dS

H1:	(a)	dN ≠ dS	(test of neutrality).
	(b)	dN > dS	(positive selection).
	(c)	dN < dS	(purifying selection).

For alternative hypotheses (b) and (c), we use a one-tailed test and for (a) we use a two-tailed test. These three tests can be conducted directly for pairs of sequences, overall sequences, or within groups of sequences. For testing for selection in a pairwise manner, you can compute the variance of (dN - dS) by using either the analytical formulas or the bootstrap resampling method.

For data sets containing more than two sequences, you can compute the average number of synonymous substitutions and the average number of nonsynonymous substitutions to conduct a Z-test in manner similar to the one mentioned above. The variance of the difference between these two quantities is estimated by the bootstrap method (Nei and Kumar [2000], page 56).

Analysis Preferences (Z-test of Selection)

In this dialog box, you can view and select options in the **Options Summary**. Options are organized in logical sections. A lime square in the right cell of a row indicates that you have a choice for that particular attribute. The three primary sets of options available in this dialog box are:

Analysis

Hypothesis to Test

One way to test whether positive selection is operating on a gene is to compare the relative abundance of synonymous and nonsynonymous substitutions within the gene sequences. For a pair of sequences, this is done by first estimating the number of synonymous substitutions per synonymous site (dS) and the number of nonsynonymous substitutions per nonsynonymous site (dN), and their variances: Var(dS) and Var(dN), respectively. With this information, we can test the null

hypothesis that H0: dN = dS using a Z-test:

$Z = (d\text{N} - d\text{S}) / \text{SQRT}(\text{Var}(d\text{S}) + \text{Var}(d\text{N}))$

The level of significance at which the null hypothesis is rejected depends on the alternative hypothesis (H1):

H0: dN = dS

H1:	(a)	dN ≠ dS	(test of neutrality).
	(d)	dN > dS	(positive selection).
	(e)	dN < dS	(purifying selection).

For alternative hypotheses (b) and (c), we use a one-tailed test and for (a) we use a two-tailed test. These three tests can be conducted directly for pairs of sequences, overall sequences, or within groups of sequences. For testing for selection in a pairwise manner, you can compute the variance of (dN - dS) by using either the analytical formulas or the bootstrap resampling method.

For data sets containing more than two sequences, you can compute the average number of synonymous substitutions and the average number of nonsynonymous substitutions to conduct a Z-test in a manner similar to the one mentioned above. The variance of the difference between these two quantities can be estimated by the bootstrap method (Nei and Kumar [2000], page 56).

Analysis Scope

Use this option to specify whether to conduct an analysis for sequence pairs, an overall average, or within sequence groups (if sequence groups are specified).

Std. Err. Computation by

Depending on the scope of the analysis (pairwise versus other), you may compute standard errors using analytical formulas or the bootstrap method. Whenever standard errors are estimated by the bootstrap method, you will be prompted for the number of bootstrap replicates and a random number seed.

When the selected test involves the computation of average distance, only the bootstrap method is available for computing standard errors.

Include Sites

These are options for handling gaps and missing data and restricting the analysis to labeled sites, if applicable.

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing information before the calculation begins (Complete-deletion option). Alternatively, you may choose to retain all such sites initially, excluding them as necessary in the pairwise distance estimation (Pairwise-deletion option).

Labeled Sites

This option is available only if there are labels associated with some or all of

the sites in the data. By clicking on the ellipses, you will have the option of including sites with selected labels. If you chose to include only labeled sites, they will be first extracted from the data and all of the other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon in the analysis.

Substitution Model

In this set of options, you can choose various attributes of the substitution models for DNA and protein sequences.

Model

By clicking on the ellipses to the right of the currently selected model, you may select a stochastic model for estimating evolutionary distance (click on the lime square to select this row first). This will reveal a menu containing many different distance methods and models.

Transition/Transversion Ratio

This option will be visible if the chosen model requires you to provide a value for the Transition/Transversion ratio (R).

Analysis Preferences (Fisher's Exact Test)

When the numbers of codons or the total numbers of synonymous and/or nonsynonymous substitutions are small, the large sample Z-test is too liberal in rejecting the null hypothesis. In these cases, tests of selection can be conducted to examine the null hypothesis of the neutral evolution. Only the Nei-Gojobori and Modified Nei-Gojobori methods can be used for this test because it requires the direct computation of the numbers of synonymous and nonsynonymous differences, and the number of synonymous and nonsynonymous sites. It should be used only when sequences show a small number of differences. To conduct Fisher's Exact Test, you need to specify two specific options:

Analysis

Hypothesis to Test

This tests for positive selection (dN > dS) and can only be computed for sequence pairs.

Include Sites

These options handle gaps and missing data and restrict the analysis to labeled sites, if applicable.

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing information before the calculation begins by using the Complete-deletion option. Alternatively, you may choose to retain all such sites initially, excluding them as necessary using the Pairwise-deletion option.

Labeled Sites

This option is available only if there are labels associated with some or all of the sites in the data. By clicking on the ellipses, you will have the option of

including sites with selected labels. If you chose to include only labeled sites, then these sites first will be extracted from the data then all other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon in the analysis.

Substitution Model

In this set of options, you choose various attributes of the substitution models for DNA and protein sequences.

Model

By clicking on the ellipses to the right of the currently selected model, you may select a stochastic model for estimating evolutionary distance (click on the lime square to select this row first). This will reveal a menu containing two different options: the original or modified Nei & Gojobori methods.

Transition/Transversion Ratio

This option will be visible if the chosen model requires you to provide a value for the Transition/Transversion ratio (R).

Analysis Preferences (Pattern Homogeneity Analysis)

In this dialog box, you can select and view options in the **Options Summary**. Options are organized in logical sections and a lime square on the right cell in a row indicates that you have a choice for that particular attribute. The two primary sets of options available in this dialog box are to compute the composition distance, disparity index, or to test the homogeneity of substitution pattern (Kumar and Gadagkar 2001).

Calculate

Use this to specify whether to compute Composition Distance , Disparity Index, or to test the homogeneity of evolutionary patterns. If the test is selected, *MEGA* will conduct the Monte-Carlo analysis, for which you need to provide the number of replicates and a starting random seed.

Include Sites

These are options for handling gaps and missing data, including or excluding codon positions, and restricting the analysis to labeled sites (if applicable).

Gaps and Missing Data

You may choose to remove all sites containing alignment gaps and missing information before the calculation begins by using the Complete-deletion option. Alternatively, you may choose to retain all such sites initially, excluding them as necessary by using the Pairwise-deletion option.

Codon Positions

By clicking on the ellipses or the lime square, you may select any combination of 1^{st}, 2^{nd}, 3^{rd}, and non-coding positions for analysis. This option is available only if the nucleotide sequences contain protein-coding regions and you have selected a nucleotide-by-nucleotide analysis. If they do, you also can choose between the analysis of nucleotide sequences or translated protein sequences. If the latter is chosen, *MEGA* will translate all protein-coding regions into amino acid sequences and conduct the protein sequence analysis.

Labeled Sites

This option is available only if there are labels associated with some or all of the sites in the data. By clicking on the ellipses, you will have the option of including sites with selected labels. If you chose to include only labeled sites, then these sites first will be extracted from the data and all other options mentioned above will be enforced. Note that labels associated with all three positions in the codon must be included for a full codon in the analysis.

5.42 Other Tests

Tajima's Test of Neutrality

Selection | Tajima's Test of Neutrality

This conducts Tajima's test of neutrality (Tajima 1989), which compares the number of segregating sites per site with the nucleotide diversity. (A site is considered segregating if, in a comparison of *m* sequences, there are two or more nucleotides at that site; nucleotide diversity is defined as the average number of nucleotide differences per site between two sequences). If all the alleles are selectively neutral, then the product *4Nv* (where *N* is the effective population size and *v* is the mutation rate per site) can be estimated in two ways, and the difference in the estimate obtained provides an indication of non-neutral evolution. Please see Nei and Kumar (2000) (page 260-261) for further description.

5.5 Molecular Clock Test

Tajima's Test (Relative Rate)

Phylogeny | Relative Rate Tests | Tajima's Test

Use this to conduct Tajima's relative rate test (Tajima 1993), which works in the following way. Consider three sequences, 1, 2 and 3, and let 3 be the outgroup. Let *nijk* be the observed number of sites in which sequences 1, 2 and 3 have nucleotides *i, j* and *k*. Under the molecular clock hypothesis, E(*nijk*) = E(*njik*) irrespective of the substitution model and whether or not the substitution rate varies with the site. If this hypothesis is rejected, then the molecular clock hypothesis can be rejected for this set of sequences.

In response to this command, you can select the three sequences for conducting Tajima's test. For nucleotide sequences, this test offers the flexibility of using only transitions, only transversions, or both. If the data is protein coding, then you can choose to analyze translated sequences or any combination of codon positions by clicking on the 'Data for Analysis' button.

See Nei and Kumar (2000) (page 193-196) for further description and an example.

5.6 Substitution Pattern

5.61 Pattern Menu

Pattern Menu

This menu provides access to the test for examining the substitution pattern homogeneity between sequences (Kumar and Gadagkar 2001) and computing the two statistics related to this test (pairwise sequence composition distance and the disparity index) (Kumar and Gadagkar 2001).

Compute Substitution Pattern

Pattern | Compute Substitution Pattern

After selection, the Analysis Preference window will pop out with all the options. For Include Sites, you can decide how to deal with Gaps/Missing Data; For Substitution Model, you can choose Pattern among Lineages and Rates among sites from the dropdowns. Click Compute button to start the calculation. MEGA caption expert with full figure legends will present the result. You can save and print the results from this window.

5.62 Compute Pattern Disparity Index

Pattern | Compute Pattern Disparity Index

Under the menu Pattern, select Compute Pattern Disparity Index, an Analysis Preferences window will show up with options for how to deal with Gaps/Missing Data and Condon Positions, click button Compute after selection to start the calculation. The progress bar will appear to show the progress of calculation. A Disparity Index window shows the results.

5.63 Compute Composition Distance

Pattern | Compute Composite Distance

The Analysis Preferences window will open, for the option Gaps/Missing Data you can click the dropdown to choose between "Complete Deletion" and "Pairwise Deletion". After the clicking Compute button, the Composition Distance window will pop out with the results.

5.64 Compute Transition/Transversion Bias

5.65 Pattern | Compute Transition/Transversion Bias ®

Select ComputeTransition/Trasversion Bias from Pattern menu, the Analysis Preference window will appear. For option Include Sites, you can choose how to deal with Gaps/Missing Data and Condon Positions. For option Substitution Model, you can choose how to define "pattern among Lineages" and "Rates among sites". Click Compute button after selection. MEGA caption expert will present the result with full table/figure legends.

6 Part V: Visualizing and Exploring Data and Results

6.1 Distance Matrix Explorer

6.11 Distance Matrix Explorer

The *Distance Matrix Explorer* is used to display results from the pairwise distance calculations. It is an intelligent viewer with the flexibility of altering display modes and functionalities and for computing within groups, among groups, and overall averages.

This explorer consists of a number of regions as follows:

Menu Bar

File Menu

Display Menu

Average Menu

Help: This button brings up the help file.

Tool Bar

The tool bar provides quick access to a number of menu items.

- *General Utilities*
  - Lower-left Triangle button: Click this icon to display pairwise distances in the lower-left matrix. If standard errors (or other statistics) are shown, they will be displayed in the upper-right.
  - Upper-right Triangle button: Click this icon to display pairwise distances in the upper-right matrix. If standard errors (or other statistics) also areshown, they will be displayed in the lower-left.
  - (A,B): This button is an on-off switch to write or hide the name of the highlighted taxa pair. The taxa pair is displayed in the status bar below.
- *Distance Display Precision*
  -  : This decreases the precision of the distance display by one decimal place with each click of the button.
  -  : This increases the precision of the distance display by one decimal place with each click of the button.
  - Column Sizer: This is a slider that increases or decreases the width of the columns showing the pairwise distances.

The 2-Dimensional Data Grid

This grid displays the pairwise distances between taxa (or within groups etc.) in the form of a lower or upper triangular matrix. The taxa names are the row-headers; the column headers are numbered from 1 to *m*, with *m* being the number of taxa. There is a column sizer for the row-headers, so that you can increase or decrease the column size to accommodate the full name of the sequences or groups.

- *Fixed Row*: This is the first row in the data grid and displays the column number.
- *Fixed Column*: This is the first and leftmost column in the data grid. This column is always visible even if you scroll past the initial screen. It contains taxa names and an associated check box. To include or exclude taxa from analysis, you can check or uncheck this box. In this column, you can drag-and-drop taxa names to sort them.
- *Rest of the Grid*: Cells to the right of the first column and below the first row contain the nucleotides or amino acids of the input data. Note that all cells are drawn in light color if they contain data corresponding to unselected sequences or genes and domains.

Status bar

The left sub-panel shows the name of the statistic for the currently selected value. In the next panel, the status bar shows the taxa-pair name for the selected value.

6.12 Average Menu (in Distance Matrix Explorer)

With this menu, you can compute the following average values:

Overall: Computes and displays the overall average.

Within groups: Thisitem is enabled only if at least one group is defined. For each group, an arithmetic average is computed for all valid pairwise comparisons and the results are displayed in the *Distance Matrix Explorer*. All incalculable within-group averages are shown with an "n/c" in red.

Between Groups: This item is enabled only if at least two groups of taxa are defined. For each between-group average, an arithmetic average is computed for all valid inter-group pairwise comparisons and results are displayed in the *Distance Matrix Explorer*. All incalculable within-group averages are shown with an "n/c" in red.

Net Between Groups: This item is enabled only if at least two groups of taxa are defined. It computes ***net*** average distances between groups of taxa. This value is given by

$$dA = dXY - (dX - dY)/2$$

where dXY is the average distance between groups X and Y, and dX and dY are the mean within-group distances. You must have at least two groups of taxa with a minimum of two taxa each for this option to work. All incalculable within-group averages are shown with a red "n/c".

6.13 Display Menu (in Distance Matrix Explorer)

The display menu consists of four main commands:

- *Show Pair Name*: This is a toggle to write or hide the name of the taxa pair highlighted,

which is displayed in the status bar below.

- *Sort Taxa*: This provides a submenu for sorting the order of taxa in one of three ways: by input order, by taxon name or by group name.
- *Show Names*: This is a toggle for displaying or hiding the taxa name.
- *Show Group Names*: This is a toggle for displaying or hiding the group name next to the name of each taxon, when available.
- *Change Font*: This brings up the dialog box that allows you to choose the type and size of the font for displaying the distance values.

6.14 File Menu (in Distance Matrix Explorer)

The file menu consists of three commands:

- *Show Input Data Title*: This displays the title of the input data.
- *Show Analysis Description:* This displays various options used to calculate the quantities displayed in the *Matrix Explorer*.
- *Export/Print Distances*: This brings up a dialog box for writing pairwise distances as a text file, with a choice of several formats.
- *Quit Viewer*: This exits the *Distance Data Explorer*.

6.2 Sequence Data Explorer

6.21 Data Menu

6.22 Display Menu

6.23 Highlight Menu

6.24 Statistics Menu

6.3 Tree Explorer

6.31 Tree Explorer

Phylogeny | Any tree-building option

The *Tree Explorer* displays the evolutionary tree based on the options used to compute or display the phylogeny. The main menu of the *Tree Explorer* has the following items:

File MenuHC_File_Menu_in_Tree_Explorer

Image MenuHC_Image_Menu_in_Tree_Explorer

Subtree MenuHC_Subtree_in_Tree_Explorer

View MenuHC_View_Menu_in_Tree_Explorer

Compute MenuHC_Compute_in_Tree_Explorer

6.32 Information Box

The information box in the *Tree Explorer* lists the various statistical attributes of the displayed tree with the branch or node highlighted. It usually contains multiple tabs.

General. This reminds the user of the number of taxa (and groups, if any) and of the strategy used to deal with gaps and missing data.

Tree. This contains information about the type of tree –rooted/unrooted, and the sum of branch lengths, SBL, or the treelength. In addition, information about the total number of trees and the tree number of the current tree is displayed.

Branch. In the *Tree Explorer* window you may click on a branch or on a node of the tree. If you click on a branch, this tab displays its location in terms of the two nodes it connects. (Leaf taxa are numbered in the order in which they appear in the input data file.) This window also displays the length of the selected branch. If you click on a node, the internal identification number of that node is displayed.

6.33 File Menu (in Tree Explorer)

This menu has the following options:

Save*:* This brings up the *Save As* dialog box and saves all the information currently held by the *Tree Explorer* to a file in a binary format. This feature allows you to retrieve the current *Tree Explorer* session for tree manipulation and printing.

Export Current Tree*:* This writes the topology of the current tree in the *MEGA* tree format to a specified file. Note that only the branching pattern is stored.

Export All Trees*:* This writes the topologies of all trees in the *MEGA* tree format to a specified file. Note that only the branching pattern is stored.

Show Information: This brings up the *Information* dialog box.

Print: This brings up the *Print* dialog box and prints the current tree in the displayed size; if the displayed tree is larger than the page size, it will be printed on multiple pages.

Print in a sheet*:* This brings up the *Print* dialog box and prints the current tree,

after restricting the size of the printed tree to one sheet. The current tree also can be printed using the button on the toolbar.

Printer Setup*:* This allows the user to setup the printer.

Exit Tree Explorer*:* This exits the *Tree Exp*lorer.

6.34 Image Menu (in Tree Explorer)

The image menu contains three options:

Copy to Clipboard: This copies the tree image to the clipboard, which can also be done by simultaneously pressing *Ctrl* and 'C' keys. You then can paste the copied image into any other Windows application (e.g., PowerPoint and Word).

Save as Enhanced metafile: This option saves the image as an enhanced windows metafile (**.EMF**). It brings up the *Save As* dialog box to specify the filename.

Save as TIFF: This option saves the tree image as Tagged Image File Format (TIFF) with 400dpi resolution and without LZW compression. TIFF is a popular raster graphics format widely supported by image-manipulation softwares such as Adobe Photoshop. Note that one cannot edit each tree part in this format as in the cases of EMF and PDF, while the graphics quality and cross-platform compatibility are beter. Also, users should notice that the file size is much larger than EMF and PDF and it usually becomes tens or hundreds of mega bytes. It brings up the Save As dialog box to specify the filename.

Loan Taxon Images: This option automatically associates images to each taxon. To use it, you will be prompted for the directory where the bitmap images (in BMP format) reside. For each taxon, the image file must have a BMP extension and the filename must be identical to the taxon name displayed in the *Tree Explorer*. All of the valid images that are found will be retrieved and displayed.

6.35 Subtree Menu (in Tree Explorer)

This menu contains the tree manipulation options *Swap*, *Flip* and *Compress/Expand*. In addition, by clicking on the corresponding items in the menu (for which there are tool buttons on the left), you can specify the root of the tree, and display a subtree (a portion of the tree defined by a given internal branch) in a separate window.

Choosing 'Divergence Time' transforms the cursor to an arrow below which is the icon associated with the divergence time option. To obtain the evolutionary rate of a specific lineage, you should point the cursor to that branch and click. On the other hand, if you are interested in the average evolutionary rate of a cluster of two

or more taxa, then you should click at the node at the common ancestor of the cluster. Either way, *MEGA* brings up the *Divergence Time* dialog box, which displays the evolutionary rate information for a given divergence time.

Many of these functionalities are also available through tools in the toolbar on the left side of the displayed tree.

6.36 Subtree Drawing Options (in Tree Explorer)

This dialog box provides choices options for changing various visual attributes for the selected subtree. If the *Overwrite Downstream* option is checked, any subtree drawing options that have been applied to downstream nodes within the current subtree will be overwritten.

Property Tab:

Name/Caption: This section allows you to provide an alphanumeric caption for the selected node.

Node/Subtree Marker: This section provides elements for changing the shape and color of the selected subtree node marker. If the *Apply to Taxon Markers* option is checked, the selected shape and color options will be applied to all taxon markers contained within the subtree.

Branch Line: This section provides various drawing options that will be applied to the branch lines of the selected subtree.

Display Tab:

Display Caption: If checked, the node caption, if set within the **Property Tab**, will be displayed.

Align Vertically: If checked,?????

Display Bracket: If checked, this item will display a bracket that encompasses the selected subtree using the configured bracket drawing options.

Display Taxon Names: If checked, the taxon names attributed to the leaf nodes will be displayed.

Display Node Markers: If checked, any node markers that were configured within the **Property Tab** will be displayed.

Display Taxon Markers: If checked, any taxon markers that were configured within the **Property Tab** will be displayed.

Compress Subtree: If checked, the selected subtree will be compressed and rendered as a graphical vector according to the configured drawing options.

Image Tab:

Display Image: If checked, the Tree Explorer will display an image, if loaded, at the configured position relative to the subtree node caption text.

6.37 Cutoff Values Tab

In this tab, you can specify a cut-off level for the condensed or consensus trees. Appropriate options become available depending on the trees displayed.

6.38 Divergence Time Dialog Box

This dialog box allows the user to specify the evolutionary rate for constructing linearized trees. This can be done by providing the evolutionary rate directly or by providing the divergence time for the given node.

6.39 View Menu (in Tree Explorer)

This menu brings up several viewing options:

Topology only*:* This displays the tree in the form of relationships among the taxa, ignoring the branch lengths.

Root on Midpoint*:* This roots the tree on the midpoint of the longest path between two taxa.

Arrange Taxa*:* This allows you to arrange the taxa in the tree based on the order of taxa in the input data file or to produce a tree that looks "balanced."

Tree/Branch Style*:* This allows you to select the display of the tree in one of three styles: *Traditional*, *Radiation*, or *Circle*. For Traditional, there are three additional options: *Rectangular*, *Straight* or *Curved*.

Show/Hide*:* This allows you to display or hide the following information: taxon label, taxon marker, statistics (e.g., bootstrap values), branch lengths, or scale bar.

Fonts*:* This allows you to choose features such as font type and size for information, including the taxon label, statistics, and scale bar.

Options*:* This brings up the *Option dialog box*, which provides control over various aspects of the tree drawing, including individual branches, the taxon names, and the scale bar.

6.310 Options dialog box (in Tree Explorer)

Through this dialog box, you can specify various drawing attributes for the tree. All options are organized in five tabs.

TreeHC_Tree_tab_in_Format_dialog_box

BranchHC_Branch_tab_in_Format_dialog_box

LabelsHC_Taxon_Name_tab_in_Format_dialog_box

ScaleHC_Scale_Bar_tab_in_Format_dialog_box

CutoffCutoff_Values_Tab_in_format_dialog_box

6.311 Tree tab (in Options dialog box)

This allows you to manipulate aspects of the tree, depending on the style you used to draw the tree. For instance, if you used the traditional rectangular style, then you can manipulate the taxon separation distance, branch length, or tree width, in the number of pixels. This tab also contains a schematic of a tree illustrating these features.

6.312 Branch tab (in Options dialog box)

This tab has options for the following aspects of the tree:

Line Width. This allows the user to choose the width of the lines.

Display Statistics/Frequency. This presents the options to *Hide* or *Show* the statistics and frequency, to choose the font, or to alter the placement of the numbers by manipulating the horizontal and vertical positions.

Display Branch Length. This presents the option to *Show* the branch length or *Hide* it if it is shorter than a specified length, to alter the placement of the written branch lengths, and to choose the number of decimal places for writing the branch lengths.

6.313 Labels tab (in Options dialog box)

This tab has options for the following:

Display Taxon Names. Presents the option to show (checked) or hide (unchecked) the label and to choose the font.

Display Markers. Allows you to draw small symbols along with or instead of taxa names in the tree. Two combo boxes and a list allow you to select the marker graphics and its color.

6.314 Scale Bar tab (in Options dialog box)

This tab has options:

Line Width. This drop-down menu allows you to choose the width of the line and the font size used in the scale bar. ***Show Distance Scale***. This allows you to show or hide the scale bar distance, to enter the unit used and to choose its length and the interval between tick marks.

Show Time Scale. This presents the option of showing or hiding the divergence time in the scale bar, and to enter the units used. You also can determine the interval between two major ticks and two minor ticks. To activate this option the divergence time for a node or the evolutionary rate must be given.

6.315 Compute Menu (in Tree Explorer)

This performs various tree computations, including Condensed tree, Linearized tree, and Consensus tree, and allows you to estimate the divergence time for each node using the molecular clock.

6.4 Caption Expert

6.41 Creating Data Captions with Caption Expert

MEGA includes a *Caption Expert* system that provides the ability to generate detailed, publication-quality captions from analysis results. The Caption Expert system is available for every type of analysis result that can be generated using MEGA (distance matrix, phylogeny, tests, etc). When invoked, the Caption Expert system will analyze the properties of the analysis results and provide a caption title followed by a detailed explanation of the analysis results. The caption text will reveal the properties of the data that underwent the analysis as well as the assumptions and parameters relevant to the computational methods that were employed. In some cases a data table will be included in the output. The resulting caption text is displayed in its own window allowing it to be printed directly, or copied and pasted into external applications such as Microsoft® Word or PowerPoint.

7 Appendix

7.1 Frequently Asked Questions

7.11 Computing statistics on only highlighted sites in Data Explorer

Go to the *Statistics* menu in the *Sequence Data Explorer*, and click on *Use highlighted sites only*. Now all statistical quantities computed using the *Statistics* menu will be based only on the highlighted sites.

7.12 Finding the number of sites in pairwise comparisons

If you want to find the number of sites between pairs of sequences or the average number of sites, then go to the *Distance* menu and select the desired distance type. Then in *Substitutions to Include,* select anoption regarding the number of sites.

7.13 Get more information about the codon based Z-test for selection

The codon based Z-test for selection can be done in two places. First, you can use the *Tests | Codon Based tests of selection | Z-test (large sample)* option to find the probability that the null hypothesis will be rejected, in addition to the actual value of the Z-statistic. Alternatively, if you want to know the difference between s and n (synonymous and nonsynonymous substitutions and their variance, you can go to the *Distances | Pairwise* menu option and in the distance computation dialog, select an appropriate method (e.g., Nei-Gojobori method) and then choose s-n (or n-s depending on your need) from the *Substitutions to include* menu. Also, you can choose to compute standard error.

7.14 Menus in MEGA are so short; where are all the options?

Our aim in developing the objectively driven user-interface of *MEGA 4* has been a clutter-free work environment that asks the user for information on a need-to-know basis Although this modular analytical tool looks simple, behind each menu item is a wide range of useful options and tools that come with enhancements that are designed to reduce the amount of time needed for mundane non-technical tasks. Consider, for example, the *Sequence Data Explorer*. This unique module is hidden away when you don't want it but is always working behind the scenes. It allows you to view the data in various ways, export data subsets, and compute many important basic statistical quantities. Another interesting module is the *Genetic Code selector,* which allows you to choose the depth at which you wish to work with a code table. With it you can select a desired code table, add new data to and edit the existing code table, view the selected code table in a conventional format, compute the degeneracy for each site in every codon, and compute the number of potentially synonymous and nonsynonymous sites for each codon. In addition, you can always find help by checking the help index.

7.15 Writing only 4-fold degenerate sites to an output file

All sequence data subset facilities are accessible through the *Export Data* command in the *Sequence Data Explorer*. To write 4-fold degenerate sites to a file, highlight the 4-fold degenerate sites on the screen and then select *Export Data*. In that command, choose to write only the highlighted sites. For example, if you select to write only the third codon positions, all 4-fold degenerate sites found in the third codon positions will be written to the file.

7.2 Main Menu Items and Dialogs Reference

7.21 Main MEGA Menus

Main MEGA Window

The main window in *MEGA* contains a menu bar, a toolbar, and a data description window (DDW). The menu bar may contain two or more menu items depending on whether a data file is active and on the type of data being analyzed.

Menu Bar

Menus: Description

File menu	Use the *File* menu commands to open, save, close data for analysis and for editing text files
Data menu	Use the *View* menu commands to display the active data, edit different data attributes, and compute basic statistical properties.
Distances menu	Use the *Distance* menu commands to calculate evolutionary distances and diversity.
Pattern menu	Use this menu to conduct tests and compute statistics regarding the substitution pattern homogeneity among lineages.
Selection *menu*	Use this menu to conduct tests of selection.
Phylogeny menu	Use the *Phylogeny* menu commands to calculate evolutionary trees, test their reliability, and view saved trees.
Alignment menu	Use this menu to construct sequence alignments and explore the world-wide-web.
Help menu	Use the *Help* menu to access the online help system, which is displayed in a special help window.

Toolbar

This contains shortcuts to some frequently used menu commands, such as those in the

Data menu.

Data Description window

This displays a summary of the currently active data set.

File Data

Open Data

File | Open Data

Choose this command to load a data file for analysis. A dialog box will appear to allow you to give the data file name. *MEGA* will first read the data file to check if it contains the Format command (see MEGA format), which specifies certain attributes of the input data (e.g., type of data). If *MEGA* does not find sufficient information in the format command, it will request the necessary information through an *Input Data Format* dialog.

If you attempt to open a dataset from a file and *MEGA* detects inconsistencies or errors in the format, it will open the file in the *text editor*, allowing you to make changes in the text file so that it conforms to the *MEGA* format.

Once a data file is opened successfully, the *Open Data* command will be disabled, some of the file's basic attributes will be displayed the bottom of the main window. To enable the *Open D*ata command, close the currently active data using the *File* | *Close Data* command.

Open dialog box

Use the open dialog box to load new data into *MEGA* for analysis.

Property	Description
Look In	Lists the current directory. Use the drop-down list to select a different drive or directory.
Files	Displays all files in the current directory matching the wildcards given in *File Name* or the file type in *Files Of Type*. You can display a list of files (default) or you can show details for each file.
File Name	Enter the data file name you want to load or type in the wildcards to use as filters.
Files of Type	Choose the type of data file you want to open. At present *MEGA* allows you to load in MEGA format files only, which should usually have the **.MEG** extension.
Up One Level	Click this button to move you directory level up from the current directory.
Create New Folder	Click this button to create a new subdirectory in the current directory.
List	Click this button to view a list of files and directories in the current directory.

Details Click this button to view a list of files and directories along with time stamp, size, and attribute information.

Export Data

File | Export Data

This command activates the appropriate input data explorer, presents a dialog box for specifying options and a file for writing the currently active data subset in a chosen format.

Reopen Data

File | Reopen Data

This reopens a recently closed data file from the submenu, which shows the names of the five most recently used data files.

Close Data

File | Close Data

This deactivates the currently open data file. Before issuing this command, save any modifications that you wish to retain by exporting the data through the data explorer (*Data | Data Explorer*).

This command is enabled only if a dataset is loaded in *MEGA*.

Exit

File | Exit

This command closes the currently active data file and all other windows. If you want to save changes to the data set displayed on the screen, before issuing this command you must choose *File | Export Data* and *Print* or *Save*. Note that *MEGA* does not automatically save changes made to active data to the original data file.

Printer Setup

File | Printer Setup

Choose this command to change the properties of your printer.

File Menu

File Menu

This allows you to perform various important tasks, including activating a data file, closing a data file, editing text files, and exiting *MEGA*.

Data Menu

Data Explorer

Data | Data Explorer

Data Exploreris used to view the currently active data set, calculate its basic statistical attributes, export it in formats compatible with other programs, and define subsets for analysis. Depending on the currently active data type, one of the following explorers will be available:

Data Type	Explorer
DNA, RNA, Protein sequences	*Sequence Data Explorer*
Evolutionary divergence	*Distance Data Explorer*

Include Codon Positions

Data | Select Preferences | Include Codon Positions

Use this menu item to specify the codon positions you would like to include in the nucleotide sequence analysis. You can include any combination of 1st, 2nd, 3rd positions and non-coding sites. The specified options are used only if you conduct a nucleotide-by-nucleotide site analysis. If relevant, you will be given this choice in the dialog box that appears in response to a requested analysis (e.g., distance computation or phylogenetic reconstruction). Thus, you have the flexibility to select or change appropriate options at the time of the analysis.

Include Labeled Sites

Data | Select Preferences | Include Labeled Sites

Use this to specify whether to include only the labeled sites in the analysis and, if so, which ones. This option is available only if you have some sites labeled. If relevant, you also will be given this choice in the dialog box that appears in response to a requested analysis (e.g., a distance computation or phylogenetic reconstruction). Thus, you have the flexibility to select or change appropriate options at the time of the analysis.

Handling Gaps and Missing Data

Data | Select Preferences | Handling Gaps and Missing Data

Use this to specify whether to use the *Pairwise-Deletion* or the *Complete- Deletion* option for handling alignment gaps and missing data. You also can specify these options in the dialog box that appears in response to a requested analysis (e.g., distance computation or phylogenetic reconstruction). Therefore, you have the flexibility to select or change appropriate options at the time of the analysis.

Select Preferences

Data | Select Preferences

This submenu specifies (1) how the alignment gaps and missing data will be handled, (2) which codon positions will be used, and (3) whether to restrict the analysis to the sites with selected labels. One or more of these options may be disabled depending on the attributes of the data set. For instance, the selection of codon positions is not valid when amino acid sequence data is being analyzed.

These options also are available in the *Options* dialog box that appears in response to a requested analysis (e.g., distance computation or phylogenetic reconstruction). Thus, you have the flexibility to select and change appropriate options at the time of the analysis.

Distances Menu

Distances Menu

Distances Menu

Use this menu to compute: pairwise and average distances between sequences; within, between, and net average distances among groups; and sequence diversity statistics for data from multiple populations.

Choose Model

Distances | Choose Model…

Choose this to select a specific model of change for computing distances. The model also can be chosen or changed in the dialog box that appears when you request an analysis, such as distance computation or phylogenetic reconstruction.

Compute Pairwise

Distances | Compute Pairwise…

Choose this to compute the distances and standard errors between pairs of taxa. A *Select Distance Options* dialog, in which you can choose the desired distance estimation method and other relevant options, will appear.

Compute Overall Mean

Distances | Compute Overall Mean…

This calculates the mean pairwise distance and standard error for the set of sequences under study. The overall mean is the arithmetic mean of all individual pairwise distances between taxa. A *Select Distance Options* dialog, in which you can choose the desired distance estimation method and other relevant options, will appear. Before using the bootstrap method to compute standard error, please read how *MEGA* implements the bootstrap method for this purpose.

Compute Within Groups Mean

Distances | Compute Within Groups Means…

This computes the mean pairwise distances within groups of taxa. The within group

means are arithmetic means of all individual pairwise distances between taxa within a group. A *Select Distance Options* dialog, in which you can choose the desired distance estimation method and other relevant options, appears. You must have at least one group of taxa, with a minimum of two taxa defined, to utilize this option.

How to define groups of taxa.

Compute Sequence Diversity

Distances | Compute Sequence Diversity

The *Sequence Diversity* submenu provides four commands for computing the population and subpopulation diversities that are useful in molecular population genetics studies. First, you define a group, using a population of sequences. Unlike the generic averages of within group, between group, and net between group distances calculated using other commands in the *Distances menu*, formulas used in the following commands are those used specifically in population genetics analyses.

The commands are:

Mean Diversity within Subpopulations

In a subpopulation, the mean diversity is defined as

$\pi_i = \frac{q}{q-1}\sum_{i=1}^{q}\sum_{j=1}^{q} x_i x_j d_{ij}$, where x_i is the frequency of *i*-th sequence in the sample from subpopulation *i*, and *q* is the number of different sequences in this subpopulation.

Mean Diversity for Entire Population

For the entire population, the mean diversity is defined as

$\pi_T = \frac{q}{q-1}\sum_{i=1}^{q}\sum_{j=1}^{q} \bar{x}_i \bar{x}_j d_{ij}$, where $\bar{x}_i$ is the estimate of average frequency of the *i*-th allele in the entire population, and *q* is the number of different sequences in the entire sample.

Mean Interpopulational Diversity

The estimate of interpopulational diversity is given by

$\delta_{ST} = \pi_T - \pi_S$.

Coefficient of Differentiation

The estimate of the proportion of interpopulational diversity is given by

$N_{ST} = \delta_{ST} / \pi_T$.

Compute Net Between Groups Means

Distances | Compute Net Between Groups Means…

This command computes the ***net*** average distances between groups of taxa. The net average distance between two groups is given by

$$dA = dXY - ((dX + dY)/2)$$

where, dXY is the average distance between groups X and Y, and dX and dY are the mean within-group distances. A *Select Distance Options* dialog, in which you can choose the desired distance estimation method and other relevant options, will appear.

You must have at least two groups of taxa with a minimum of two taxa each for this option to work.

How to define groups of taxa.

Compute Between Groups Means

Distances | Compute Between Groups Means…

This computes the average distances between groups of taxa. The average distance is the arithmetic mean of all pairwise distances between two groups in the inter-group comparisons. A *Select Distance Options* dialog, in which you can choose the desired distance estimation method and other relevant options, will appear. You must have at least two groups of taxa for this option to work.

How to define groups of taxa.

Phylogeny Menu

Phylogeny Menu

Phylogeny Menu

Use the *Phylogeny* menu to construct phylogenetic trees, infer their reliability using the bootstrap and interior branch tests, conduct molecular clock tests, and view previously constructed trees.

Display Saved Tree Session

Phylogeny | Display Saved Tree Session…

Use this command to display a previously saved *Tree Explorer* session (saved in a filename with .**MTS** extension).

Relative Rate Tests

Relative Rate Tests

Phylogeny | Relative Rate Tests

This submenu provides access to a test of the constancy of evolutionary rates between two sequences or clusters of sequences, using an outgroup sequence.

Construct Phylogeny

Construct Phylogeny

Phylogeny | Construct Phylogeny

This submenu contains commands for constructing Neighbor Joining, Minimum Evolution, Maximum Parsimony, and UPGMA trees.

Neighbor-Joining (NJ) Method

This method (Saitou and Nei 1987) is a simplified version of the minimum evolution (ME) method (Rzhetsky and Nei 1992). The ME method uses distance measures that correct for multiple hits at the same sites; it chooses a topology showing the smallest value of the sum of all branches (*S*) as an estimate of the correct tree. However, construction of an ME tree is time-consuming because, in principle, the *S* values for all topologies must be evaluated. Because the number of possible topologies (unrooted trees) rapidly increases with the number of taxa, it becomes very difficult to examine all topologies.

In the case of the NJ method, the *S* value is not computed for all or many topologies, but the examination of different topologies is embedded in the algorithm, so that only one final tree is produced. The algorithm of the NJ method is somewhat complicated and is explained in detail in Nei and Kumar (2000, page 103).

The NJ method produces an unrooted tree because it does not require the assumption of a constant rate of evolution. Finding the root requires an outgroup taxon. In the absence of outgroup taxa, the root is sometimes given at the midpoint of the longest distance connecting two taxa in the tree, which is referred to as mid-point rooting.

Minimum Evolution (Construct Phylogeny)

Phylogeny | Construct Phylogeny | Minimum Evolution…

This command is used to construct a phylogenetic tree under the minimum evolution criterion. In this method the sum, *S*, of all branch length estimates, *i.e.*,

$S = ?bi,$

is computed for all plausible topologies, and the topology that has the smallest *S* value is chosen as the best tree: the ME tree. This criterion does not require the assumption of evolutionary rate constancy as needed in the UPGMA analysis. Therefore the inferred phylogenetic tree is an unrooted tree, even though, for ease of inspection, it is often displayed in a manner similar to rooted trees.

MEGA employs the Close-Neighbor-Interchange (CNI) algorithm to find the ME tree. This is a branch swapping method, which begins with a given initial tree. You can ask *MEGA* to automatically construct a Neighbor-Joining (NJ) tree and use that as the starting tree. Alternatively, you can provide your own topology. Note that the final tree produced after this search is not guaranteed to be the ME tree. These options are available in the *ME Tree Tab* of the *Analysis Preferences* dialog box, which is displayed before the phylogenetic analysis begins. This dialog box also allows you to specify the distance estimation method, subset of sites to include, and whether to conduct a test of the inferred

tree.

Maximum Parsimony

Maximum Parsimony (MP) Method

Maximum parsimony (MP) methods originally were developed for morphological characters, and there are many different versions (see Nei and Kumar [2000] for a review). In *MEGA*, we consider both of these methods for nucleotide and amino acid sequence data (Eck and Dayhoff 1966; Fitch 1971).

For constructing an MP tree, only sites at which there are at least two different kinds of nucleotides or amino acids, each represented at least twice, are used (parsimony-informative sites). Other variable sites are not used for constructing an MP tree, although they are informative for distance and maximum-likelihood methods.

MEGA estimates MP tree branch lengths by using the average pathway method for unrooted trees (see Nei and Kumar [2000], page 132).

To search for MP Trees, *MEGA* provides three different types of searches: the max-mini branch-and-bound search, min-mini heuristic search, and close-neighbor-interchange heuristic search. Only the branch-and-bound search is guaranteed to find all the MP trees, but it takes prohibitive amount of time if the number of sequences is large (>15). For details, please see chapter 7 in Nei and Kumar (2000)

Maximum Parsimony (Construct Phylogeny)

Phylogeny | Construct Phylogeny | Maximum Parsimony…

This command is used to construct phylogenetic trees under the maximum parsimony criterion. For a given topology, the sum of the minimum possible substitutions over all sites is known as the Tree Length. The topology with the minimum tree length is known as the Maximum Parsimony tree.

The phylogenetic tree(s) inferred using this criterion are unrooted trees, even though, for ease of inspection, they are often displayed in a manner similar to rooted trees.

MEGA includes the Max-mini branch-and-bound search, which is guaranteed to find all the MP trees. However, it is often too time consuming for more than 15 sequences. In those cases, you should use the Close-Neighbor-Interchange (CNI) algorithm to find the MP tree. CNI is a branch swapping method that begins with a given initial tree. You can ask *MEGA* to automatically obtain a set of initial trees by using the Min-mini algorithm with a given search factor. Alternatively, you can produce the initial trees by providing your own topology or by using the random addition option. These options are available in the *MP Tree Tab* of the *Options* dialog box and are displayed before the phylogenetic analysis begins. Note that these CNI branch-swapping procedures may not produce the best MP trees or all the MP trees.

By default, all nucleotide (or amino acid) changes are weighted equally in *MEGA* (standard parsimony). However, for nucleotide sequences, you have the option of

conducting a transversion parsimony analysis in which only transversional changes are considered for calculating the tree length. In addition, through the *Analysis Preferences/Options* dialog box, you are given options on which subset of sites to include, and whether to conduct a test of the inferred tree .

UPGMA

UPGMA

This method assumes that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. An interesting aspect of this method is that it produces a tree that mimics a species tree, with the branch lengths for two OTUs being the same after their separation. Because of the assumption of a constant rate of evolution, this method produces a rooted tree, though it is possible to remove the root for certain purposes. The algorithm for UPGMA is discussed in detail in Nei and Kumar (2000, page 87).

UPGMA (Construct Phylogeny)

Phylogeny | Construct Phylogeny | UPGMA…

This command is used to construct a UPGMA tree. This tree-making method assumes that the rate of evolution has remained constant throughout the evolutionary history of the included taxa. Therefore, it produces a rooted tree.

If your input data is a distance matrix, then using this command makes *MEGA* proceed directly to constructing and displaying the UPGMA tree. In all other instances, you will be asked in an *Analysis Preferences* dialog box to specify the distance estimation method, subset of sites to include, and whether to conduct a test of the inferred tree.

Pattern Menu

Selection Menu

Selection Menu

Selection Menu

This menu provides access to codon-based tests of selection as well as to Tajima's test of neutrality.

Codon Based Z-Test (large sample)

Selection | Codon Based Z-test (large sample)

One way to test whether positive selection is operating on a gene is to compare the relative abundance of synonymous and nonsynonymous substitutions that have occurred in the gene sequences. For a pair of sequences, this is done by first estimating the number of synonymous substitutions per synonymous site (dS) and the number of nonsynonymous substitutions per nonsynonymous site (dN), and their variances: Var(dS) and Var(dN),

respectively. With this information, we can test the null hypothesis that H0: *d*N = *d*S using a Z-test:

$$Z = (dN - dS) / SQRT(Var(dS) + Var(dN))$$

The level of significance at which the null hypothesis is rejected depends on the alternative hypothesis (H1)

H0: *d*N = *d*S

H1:	(a)	*d*N ? *d*S	(test of neutrality).
	(b)	*d*N > *d*S	(positive selection).
	(c)	*d*N < *d*S	(purifying selection).

For alternative hypotheses (b) and (c), we use a one-tailed test and for (a) we use a two-tailed test. These three tests can be conducted directly for pairs of sequences, overall sequences, or within groups of sequences. For testing for selection in a pairwise manner, you can compute the variance of (*d*N - *d*S) by using either the analytical formulas or the bootstrap resampling method.

For data sets containing more than two sequences, you can compute the average number of synonymous substitutions and the average number of nonsynonymous substitutions to conduct a Z-test in a manner similar to the one mentioned above. The variance of the difference between these two quantities is estimated by the bootstrap method (See Nei and Kumar (2000) page 55).

Codon Based Fisher's Exact Test

Selection | Codon Based Fisher's Exact Test

This provides a test of selection based on the comparison of the numbers of synonymous and nonsynonymous substitutions between sequences. Use this command to conduct a small sample test of positive selection (Zhang et al. 1997): a one-tailed Fisher's Exact test. If the resulting *P* -value is less than 0.05, then the null hypothesis of neutral evolution (strictly neutral and purifying selection) is rejected. If the observed number of synonymous differences per synonymous site (*p*S) exceeds the number of nonsynonymous differences per nonsynonymous site (*p*N) then *MEGA* sets *P* = 1 to indicate purifying selection, rather than positive selection.

See Nei and Kumar (2000) (page 56) for further description and an example.

Alignment Menu

Alignment Menu

Alignment Menu

This menu provides access to options for viewing and building DNA and protein sequence alignments and for exploring the web based databases (e.g., NCBI Query and BLAST searches) in the *MEGA* environment.

Alignment Explorer/CLUSTAL

Alignment | Alignment Explorer/CLUSTAL

This option displays the Alignment Explorer, which can be used to view and build DNA and protein sequence alignments and to explore the web based databases (e.g., NCBI Query and BLAST searches) in the *MEGA* environment.

Query Databanks

Alignment | Query Databanks

Use this to open the MEGA web-browser to search the NCBI and other web sites for sequence data.

Show Web Browser

Alignment | Show Web Browser

Use this option to launch the *MEGA* Web Browser.

View/Edit Sequencer Files

Alignment | View/Edit Sequencer Files

Use this option to view/edit the sequence data in ABI (*.abi and **.**ab1) and Staden (.scf) files. The Alignment Explorer provides this option directly.

Help Menu

Help Menu

Help Menu

This menu provides access to the help index as well as the *About* dialog box, which provides version information for *MEGA*.

Index

Help | Index

This command provides access to the help file index and keyword searching facilities.

About

Help | About…

This command will display the *About* dialog box showing the copyright, authors, and version information for *MEGA*.

7.22 MEGA Dialogs

Input Data Format Dialog

The Input Data Format dialog is displayed if *MEGA* does not find enough information about the type of data included in the input file.

Data Type

This displays the list of data types that *MEGA* is able to analyze. Highlight the current data type by clicking on it. Depending on the type of data selected, you may need to provide information about the following additional items.

For Sequence Data

- Missing Data

  Character used to show missing data in the data file; it should be set to a question mark (?).

- Alignment Gap

  Character used to represent gaps inserted in the multiple sequence alignment; it is set to a dash (-) by default.

- Identical Symbol

  Character used to represent identity with the first sequence in the data files; it is set to a dot (.) by default.

For Pairwise Distance Data

- Missing Data

  Character used to show missing data in the data file; it should be set to a question mark (?).

- Matrix Format

  Choose the lower-left or upper-right distance matrix for the pairwise distance data type.

Note: To avoid having to answer these questions every time you read your data file, save the data by exporting it in *MEGA* format.

Setup/Select Taxa & Groups Dialog

This dialog box has two sub-windows (*Taxa/Groups* and *Ungrouped Taxa*), a panel bar between them containing a few buttons, and a command panel, with the lower part containing the *Add*, *Delete*, *Close*, and *Help* buttons.

Taxa/Groups sub-window on the left: It shows all the currently defined taxa and group names hierarchically. If a taxon has been assigned to a group, it will appear connected to that group. Groups may be displayed in a collapsed format (indicated by a + mark before their name). You can click '+' to expand the group to a listing of the taxa contained in it, and click '–' to collapse the group to only view the group name. Groups that do not contain any members do not have this box. Next is a checkbox indicating whether a given group or taxon will be included in an analysis. Following that is an icon indicating a taxon (single box) or a group (layer of boxes). Grayed out check boxes are used to indicate that some of the taxa in a group are selected and others are unselected. You can rearrange the order of taxa and groups using drag-and-drop. However, note that this order is not automatically used in the *Data Explorer*. To enforce this order, use the *Sort* command in the *Data Explorer*.

Ungrouped Taxa Sub-window on the right: This shows the names of all the taxa that do not belong to any of the groups to facilitate your ability to move taxa into groups. If this sub-window does not appear on your screen, then hold and drag the lower right corner of the dialog box to expand its width to unhide it.

Middle Command Panel: This resides between the above-mentioned two sub-windows and contains a splitter on its right edge. You can grab the splitter and move it to change the proportion of the space taken by the two sub-windows. In this panel left and right arrow buttons are used to add or remove taxa from the groups. Clicking the hand-with-a-pencil icon with a highlighted taxon or group name will allow you to edit that name.

Lower Command Panel: In the lower part of the *Select/Edit Taxa/Groups* window are buttons that are used to add and/or delete groups. The '+' and '–' buttons are also present on the middle command panel.

Buttons	**Description**
Add	Creates a new group.
Delete	Deletes the currently selected group. Any taxa that were assigned to the group will become freestanding.
Ungroup	Makes all the taxa in the selected group freestanding, but does not remove the group from the list.
Close	Closes the dialog box.
Help	Brings up help regarding the dialog box.

How to perform functions:

Function	**Description**
Creating a new group	Click on the *Add* button. Click on the highlighted name of the group and type in a

	new name.
Deleting a group	Select the group and click the *Delete* button. Any taxa that were assigned to this group will become freestanding.
Adding taxa to a group	Drag-and-drop the taxon on the desired group or select one or more taxa in the *Ungrouped Taxa* window and click on the left arrow button on the middle command panel.
Removing a taxon from a group	Click on the taxon and drag-and-drop it into a group (or outside all groups). Or, select the taxon and click on the right arrow button on the middle command panel.
Include/Exclude taxa or groups	Click the checkbox next to the group or taxa name.

Setup/Select Genes & Domains Dialog

Use the Gene & Domain Editor to inspect, define, and select domains, and genes, and labels for individual sites.

The Genes & Domains dialog consists of two tabs: *Define/Edit/Select* and *Site Labels*.

Define/Edit/Select tab

This tab contains a hierarchical listing of gene and domain names with the corresponding information organized into four columns for amino acid sequences and six columns for nucleotide sequences.

Gene and domain name listing

Each line in this display contains a small 'expand/contract' box, a checkbox, a gene/domain icon, and the name of the gene or domain. The 'expand/contract' box allows you to display or hide the information below a given gene. The checkbox shows if the gene or domain is currently selected for analysis. All defined genes and domains appear below the *Genes\Domain* node in the hierarchy. All domain names are shown with a yellow background. The *Independent* node shows the number of Independent sites, which are not assigned to any domains or genes.

If your input data file does not contain any domains, then *MEGA* automatically creates a domain called Data. If you wish to create new domains, you should delete the Data domain to make all sites independent. Remember that only independent sites can be assigned to domains, and sites cannot be assigned to multiple domains. Genes are simply collections of domains, and thus gene boundaries are decided based on the domains contained in them. The *MEGA* gene and domain organizer is flexible and is designed to enable you to specify genes and domains as they appear in a genome. For instance, a sequence may contain one or more genes, each of which may contain one or more domains. In between genes, there may be intergenic domains. In addition, within or

between genes or domains, there may be sites that are not members of any domain.

At the bottom of this tab, you will find a toolbar with many drop-down menu buttons, which can be used to *Add/Insert* new genes or domains. The add and insert operations differ in the following way. If you add a gene or domain, then the new gene or domain will be added at the end of the list to which the currently focused gene or domain belongs. If you insert a gene (or domain), it will be inserted by shifting all the following genes or domains down. *Add* and *Insert* commands are context sensitive.

You can rearrange the relative position of genes and domains by drag-and-drop operations.

Inspecting/modifying attributes of genes and domains

When you start, all genes and domains are shown. Click on the '+' in the expand/contract box to expand the listing for each gene to its domains. Click on the '-' to collapse to the gene. To select and deselect genes or domains from analysis, click in the corresponding checkbox. When a gene is selected but some domains within the gene are not, the checkbox for the gene will be grayed. If you deselect a gene, all domains within that gene are automatically deselected.

On the right side of the gene and domain hierarchy, you will find at least four columns of information for each domain and gene. All information shown for genes is computed based on the domains contained.

The first two columns show the site number in the sequence where the domain begins (*From* column) and where it ends (*To* column). The total number of sites shown next to the *To* column indicates the total number of sites automatically computed, based on the range of information given in the previous two columns. A question mark (?) shows that the domain exists but that the range of sites is not yet specified.

To specify or change sites that belong to a given domain, click on the domain name. The corresponding rows in the *From* and the *To* columns contain a button with three dots (ellipses). To change the start site, click on the ellipses in the *From* column. This will bring up a small *Site Picker* dialog box with which you can highlight the desired site and click *OK*. In this viewer, you will see that sites have different background colors. A white background marks independent sites, a red background indicates that the site is used by another domain, and a yellow background shows that the current site belongs to the domain being edited. To cancel any changes, click on *Cancel* in the *Site Picker* dialog box.

For nucleotide sequences, two additional columns are found in the *Define/Edit/Select tab*: the *Coding?* column and the *Codon Start* column. A check-mark in the *Coding?* column shows that a given domain is protein coding. If it is checked, then the next column allows you to specify whether the first site in the domain is in the first, second, or the third codon position.

Site Labels Tab

This tab displays sequences and allows you to label individual sites. To do this, change the default underscore (_) in the topmost line to the label of choice and give it a light green background. The site number will be displayed below in a window, next to which is shown the name of the domain, along with gene, name. Labeled sites can be selected or deselected for analysis.

To change or give a label to a site, click on the site and type in the character you wish to mark it with. You can use the left and right arrow buttons on the keyboard to move to and then label adjacent sites. To change a label, simply overtype it. To remove a label, use the spacebar to type a space.

Example

Imagine an alignment consisting of a genomic sequence, including a gene and its upstream and downstream regions. You can define each intron and exon as a domain, and then define the overall gene, assigning the exons and introns to that gene. The upstream and downstream regions also can be defined as domains, or possibly multiple domains, depending on the analysis you wish to perform. These domains do not have to be assigned to any gene. Furthermore, some sites may be left unassigned, as independent sites. These can be scattered throughout the sequence and can be included or excluded from analysis as a group. If you have a complicated patterns of sites you wish to analyze as groups, and the domaingene approach is unsuitable, you should assign a category to these sites, which can be specified in addition to the groups and domains.

Select Genetic Code Table Dialog

This dialog selects the desired genetic code, and edits and displays the properties of the genetic codes. At present only one genetic code can be selected in *MEGA* at any given time; it is used for all coding regions in all sequences in the data set.

To select a genetic code, click in the square box to its left.

You can also highlight any genetic code by clicking on the text.

You can then use the following buttons found along the top of the dialog box:

Button	Description
Add	Creates a new genetic code table. A *code table editor* will be shown with the genetic code of the currently highlighted code table loaded.
Delete	Removes the highlighted genetic code from the list. Note that the standard genetic code cannot be deleted.
Edit	Modifies the highlighted genetic code or its name. The *code table editor* will be invoked for editing the genetic code.
View	Displays the highlighted genetic code in a printable format.
Statistic	Displays the number of synonymous and non-synonymous sites for the codons of the highlighted genetic code following the Nei-Gojobori (1986)

s method. The degeneracy values for the first, second, and third codon positions are displayed following Li et al. (1985).

7.3 Error Messages

7.31 Blank Names Are Not Permitted

As this error message suggests, you cannot leave the name of a sequence, taxa, domain, or gene blank.

7.32 Data File Parsing Error

An error occurred while parsing the input data file. Pay close attention to the message provided, then look for the error that occurred just prior to the event indicated in the file.

7.33 Dayhoff/JTT Distance Could Not Be Computed

The Dayhoff/JTT matrix-based correction could not be applied for one or more pairs of sequences. If you wish to know which pair(s), use the *Distances|Pairwise* option. They will be shown in the *Distance Matrix Dialog* with a red n/c (not computable).

7.34 Domains Cannot Overlap

Any given site can belong to only one domain, at most. If you would like to assign a site or range of sites belonging to one domain to a second domain, you must first change or delete the definition of the first domain.

7.35 Equal Input Correction Failed

This error message means that, the Equal Input Model-based correction could not be applied for the amino acid distances estimation. If you wish to know which pair(s) of sequences has this problem, use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c (not computable).

7.36 Fisher's Exact Test Has Failed

Fisher's exact test uses estimates of the number of synonymous sites (S), the number of nonsynonymous sites (N), the number of synonymous differences (Sd), and the number of nonsynonymous differences (Nd). It fails for a number of reasons. If thenumbers are very large, some mathematical functions may not be able to handle them, although we have tried to avoid this by using logarithms of factorials. To diagnose the problem, compute S, N, Sd, and Nd using the *Distances|Pairwise* option four times. If you still cannot find the problem, please contact us

7.37 Gamma Distance Failed Because p > 0.99

For amino acid distance estimation, if the proportion of amino acids between two sequences that are different has exceeded 99%, the gamma distance cannot be calculated. To know which pair(s) of sequences has this problem, use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c.

7.38 Gene Names Must Be Unique

MEGA requires that all gene names in a genome be unique, although, for convenience, many domains can have the same name. For example, you may want to give the name Exon-1 to the first exon in all genes.

7.39 Inapplicable Computation Requested

You have requested a computation that is not allowed or is unavailable for the currently active dataset. If you think that this is in error, then please report this potential software bug to us.

7.310 Incorrect Command Used

The selected command or option is not valid here. Please look at the brief description provided in the error message window to determine the nature of the problem.

7.311 Invalid special symbol in molecular sequences

Unique ASCII characters, except letters and '*', can be used as special symbols for alignment gaps, missing data, and identical sites. Frequently used symbols for identical sites, alignment gaps, and missing data are '.', '-', and '?', respectively. This error message means that you have attempted to use the same symbols for two or more of these types of sites, or a chosen symbol is not appropriate. For example, do not use N (the ambiguous site symbol for DNA/RNA sequences), or X (the ambiguous site symbol for protein sequences) because they are already available as the IUPAC symbols for molecular sequences.

7.312 Jukes-Cantor Distance Failed

The Jukes-Cantor correction is used to calculate nucleotide distances and synonymous and nonsynonymous substitution distances. If the proportion of sites that are different (nucleotides, synonymous, or nonsynonymous) is greater than or equal to 75%, the Jukes-Cantor correction cannot be applied. If you see this error message, then this has happened for one or more pairs in your data. If you wish to know which pair(s), use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c.

7.313 Kimura Distance Failed

The Kimura (1980) distance correction is used in a number of operations, including calculating nucleotide distances and synonymous and nonsynonymous substitution distances. These formulas cannot be applied if the argument in the logarithm approaches zero or becomes negative. If you see this error message, then this has happened for one or more pairs in your data. If you wish to know which pair(s), use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c.

7.314 LogDet Distance Could Not Be Computed

The formula used for calculating distances contains many log terms. If some of their arguments approach zero too closely or become negative the LogDet correction cannot be applied. If you wish to know which pair(s) of sequences has this problem, use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c (not computable).

7.315 Missing data or invalid distances in the matrix

The selected set of taxa contains one or more pairs for which the evolutionary distance is either invalid or not available. Please inspect the distance data in the *Data Explorer* to identify those pairs and remove one or more taxa, as needed.

7.316 No Common Sites

For the sequences and data subset options selected, *MEGA* found zero common sites. If you selected the **complete deletion** option then you might achieve better results using the **pairwise deletion** option, as **complete deletion** removes all sites containing a gap in any part of the alignment. If you selected the **pairwise deletion** option then MEGA was unable to calculate the distance between one or several of the sequence pairs in the alignment. To identify such pairs compute a pairwise distance matrix using the p-distance method and look for the word "n/c" in place of the pairwise distance value.

7.317 Not Enough Groups Selected

The currently active dataset or subset does not contain enough groups to conduct the desired analysis. Please define or select more groups using the *Setup Taxa and Groups Dialog*.

7.318 Not Enough Taxa Selected

The currently active dataset or subset does not contain enough sequences or taxa to conduct the desired analysis. Please add or select more sequences.

7.319 Not Yet Implemented

The task you requested was not activated. This function either was not be available in your release of *MEGA* or needs to be activated by us. Please contact the authors and report this software bug at your earliest convenience.

7.320 p distance is found to be > 1

This peculiar situation can occur in the computation of the proportion of synonymous (or nonsynonymous) substitutions per site, especially when the number of included codons is small. If you wish to know which pair(s) of sequences has this problem, please use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c.

The Kimura (1980) distance correction is used in a number of operations, including calculating nucleotide distances and synonymous and nonsynonymous substitution distances. These formulas cannot be applied if the argument in the logarithm approaches zero or becomes negative. If you see this error message, then this has happened for one or more pairs in your data. If you wish to know which pair(s), use the *Distances|Pairwise*

option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c.

7.321 Poisson Correction Failed because p > 0.99

For an amino acid estimation of distances, the proportion of amino acids that differ between two sequences has exceeded 99% and the Poisson correction distance formula cannot be applied. If you wish to know which pair(s) of sequences has this problem, use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c (not computable).

7.322 Tajima-Nei Distance Could Not Be Computed

For one or more pairs of sequences, the Tajima-Nei correction could not be applied, which usually occurs if the argument in the log term of the formula becomes too close to zero. If you wish to know which pair(s) of sequences has this problem, use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c (not computable).

7.323 Tamura (1992) Distance Could Not Be Computed

For one or more pairs of sequences, the Tajima-Nei correction could not be applied. This usually occurs if the argument in the log term of the formula becomes too close to zero or if it is negative, or if the G+C-content is 0% or 100%. If you wish to know which pair(s) of sequences has this problem, use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c (not computable).

7.324 Tamura-Nei Distance Could Not Be Computed

The Tamura-Nei distance formula contains many log terms. If some of their arguments approach zero too closely or become negative, the Tamura-Nei model correction cannot be applied. If you wish to know which pair(s) of sequences has this problem, use the *Distances|Pairwise* option. All such pairs will be shown in the *Distance Matrix Dialog* with a red n/c (not computable).

7.325 Unexpected Error

While carrying out the requested task, an unexpected error has occurred in *MEGA*. Please contact the authors and report this software bug as soon as possible. We will try to solve the problem at the earliest possible time.

7.326 User Stopped Computation

You have aborted the current process by pressing the *Stop process* button on the progress indicator.

7.4 Glossary

7.41 ABI File Format

The ABI File Format is a binary file that is produced by ABI sequencer software. This data file, referred to as a "trace file" is viewable in MEGA's Trace File Editor, which is part of the Alignment Explorer.

7.42 Alignment Gaps

Phylogenetic analysis on two or more DNA or amino acid sequences requires that the sequences be aligned so that the substitutions can be accurately enumerated. During alignment, gaps must be introduced in sequences that have undergone deletions or insertions. These gaps are known as alignment gaps or indels.

7.43 Alignment session

When working in MEGA's Alignment Explorer you can choose to save the current state of all data and settings in the alignment explorer to a file so you can archive your work, or save it to resume editing in the future. An alignment session is a binary file format that is saved with the .MAS file extension.

7.44 Bifurcating Tree

A bifurcating tree is one in which each ancestral lineage gives rise to exactly two descendent lineages. A tree with only bifurcating nodes is called a bifurcating tree.

7.45 Branch

A branch is a line connecting either two internal nodes to each other or an external node to an internal node in a phylogenetic tree. The length of a branch denotes the genetic distance (e.g., number of substitutions per unit time) between the two taxa it connects.

7.46 ClustalW

nClustalW is a general purpose multiple sequence alignment program for DNA or proteins. You can learn more about ClustalW by visiting its website (http://www.ebi.ac.uk/clustalw/).

7.47 Codon

A codon is triplet of nucleotides that codes for a specific amino acid.

7.48 Codon Usage

There are 64 (43) possible codons that code for 20 amino acids (and stop signals) so one amino acid may be encoded by several codons (e.g., serine is encoded by six codons in nuclear genes). It is therefore interesting to know the codon usage for each amino acid. In *MEGA*, the numbers of the 64 codons used in a gene can be computed either for one specific sequence or for all examined sequences. In addition to the codon frequencies, *MEGA* also writes the Sharp et al. (1986) relative synonymous codon usage (RSCU) statistic (see Nei and Kumar 2000, page 11).

7.49 Complete-Deletion Option

In the complete-deletion option, sites containing missing data or alignment gaps are removed before the analysis begins. This is in contrast to the pairwise-deletion option in which sites are removed during the analysis as the need arises (e.g., pairwise distance computation).

7.410 Composition Distance

Composition distance is a measure of the difference in nucleotide (or amino acid) composition for a given pair of sequences. It is one half the sum of squared difference in counts of bases (or residues). MEGA 4 computes and presents the Composition Distance per site, which is given by the total composition distance between two sequences divided by the number of positions compared, excluding gaps and missing data.

7.411 Compress/Uncompress

This command changes the cursor to the 'Compress/Uncompress' icon. If you click on an interior branch, *MEGA* will prompt you to give a name to the group that will be formed. It then will compress all the lineages defined by this branch into a solid elongated triangle whose thickness is proportional to the number of taxa condensed. Clicking on the branch again will uncompress it.

The cursor may be reverted to the arrow by clicking on the arrow icon on the left hand side of the *Tree Explorer*.

7.412 Condensed Tree

When interior branches in a phylogenetic tree do not have statistically significant lengths, choosing this command condenses the tree into a topology in which each branch with less than the desired statistical significance is collapsed.

7.413 Constant Site

A site containing the same nucleotide or amino acid in all sequences is referred to as a constant site. *MEGA* identifies a site as a constant site only if at least two sequences contain unambiguous nucleotides or amino acids.

7.414 Degeneracy

0-fold degenerate sites are those at which all changes are nonsynonymous.

2-fold degenerate sites are those at which one out of three changes is synonymous. (All sites at which two out of three changes are synonymous also are included in this category.)

4-fold degenerate sites are those at which all changes are synonymous.

7.415 Disparity Index

Disparity Index measures the observed difference in substitution patterns for a pair of sequences. It works by comparing the nucleotide (or amino acid) frequencies in

given pair of sequences and using the number of observed differences between sequences. MEGA 4 computes and presents the Disparity Index per site, which is given by the total disparity index between two sequences divided by the number of positions compared, excluding gaps and missing data. It is more powerful than a chi-square test of the equality of base frequencies between sequences.

7.416 Domains

A domain is a continuous block of sites in a sequence alignment. A domain can be free-standing or assigned to genes and protein-coding (e.g., exons) or non-coding (e.g., introns). Domains can be defined in the input data, and can be defined and edited in the *Setup Genes Domains* dialog.

7.417 Exon

A protein-coding gene typically consists of multiple coding regions, known as exons, interspersed with non-coding DNA (introns)

7.418 Extant Taxa

The taxa whose sequences, other genetic information or morphological characters, etc. are being used for a phylogenetic analysis are known as extant taxa, irrespective of whether the individuals or species to whom the sequences and other information belong are extant or extinct.

7.419 Flip

This command changes the cursor to the 'Flip' icon. Then, if you click on an interior branch, *MEGA* reverses the order of the lineages defined by this branch.

The cursor will revert to the arrow if you click on the arrow icon on the left hand side of the *Tree Explorer*.

7.420 Format command

A format command in a data file begins with `!Format` and contains at least the data type included in the file.

7.421 Gamma parameter

According to the gamma distribution, the substitution rate often varies from site to site within a sequence. The shape of this distribution is determined by a value known as the gamma parameter, which is also known as the shape parameter.

7.422 Gene

A gene is a collection of domains. The domains included in a gene need not be consecutive or of the same type. Genes and domains can be defined in the input data, and can be defined and edited in the *Setup Genes and Domains* dialog. Genes can be selected or unselected from an analysis. When a gene is unselected, all its

domains are automatically unselected. However, a gene can be selected, with some of its domains unselected.

7.423 Genetic Codes

A genetic code table specifies the amino acid residues encoded by the various codons. Vertebrate mitochondria, Drosophila mitochondria, and yeast mitochondria all have their own genetic code tables, which are slightly different from the most common table, the Standard Genetic Code Table.

7.424 Indels

Phylogenetic analysis on two or more DNA or amino acid sequences requires that the sequences be aligned so that the substitutions can be accurately enumerated. During the alignment, gaps must be introduced in sequences that have undergone deletions or insertions. These gaps are known as alignment gaps, or indels.

7.425 Independent Sites

In a sequence alignment, all sites that have not been assigned to any gene or domain are classified as independent.

7.426 Inferred Tree

A tree reconstructed from the observed sequence or other appropriate data using any tree-making method (such as UPGMA, NJ, ME, or MP) is known as an inferred or reconstructed tree.

7.427 Intron

Introns are the non-coding segments of DNA in a gene that are interspersed among the exons.

7.428 Maximum Composite Likelihood

In general, a composite likelihood is defined as a sum of log-likelihoods for related estimates. In MEGA4, the maximum composite likelihood is used for describing the sum of log-likelihoods for all pairwise distances in a distance matrix (Tamura et al. 2004) estimated by using the Tamura-Nei (1993) model (see related Tamura-Nei distance). Further information is in the Maximum Composite Likelihood Method.

7.429 Max-mini branch-and-bound search

This is an algorithm for searching for the MP tree using the branch-and bound search method. See Nei & Kumar (2000) for details.

7.430 Maximum Parsimony Principle

For any given topology, the sum of the minimum possible substitutions over all

sites is known as the tree length for that topology. The topology with the minimum tree length is known as the Maximum Parsimony tree.

7.431 Mid-point rooting

In the mid-point rooting method, the root of an unrooted tree is placed at the mid-point of the longest distance between two taxa in a tree.

7.432 Monophyletic

A cluster of taxa that shared a common ancestor comparatively recently in the evolutionary history of a phylogenetic tree is monophyletic. The term reflects the close relationship of the taxa with each other.

7.433 mRNA

Protein-coding genes are first *transcribed* into messenger RNAs (mRNA), which are, in turn, *translated* into amino acid sequences to make proteins.

7.434 NCBI

jAn acronym that stands for "National Center for Biotechnology Information". NCBI is a federally funded resource for molecular biology information. NCBI creates databases, conducts research in computational biology, develops software and tools for analyzing genome data, and disseminates biomedical information. You can find out more about NCBI by visiting the NCBI website (http://www.ncbi.nlm.nih.gov).

7.435 Newick Format

NEWICK is a simple format used to write out trees in a text file. While this is a hard-to-read format for humans, it is very useful for exchanging trees between different types of software. An example of the contents of a NEWICK format tree file is given below (note that semi-colon is needed to end the tree). Further information on this format can be found at Joe Felsenstein's website.

```
((raccoon, bear),((sea_lion,seal),((monkey,cat), weasel)),dog);

The above tree with branch lengths will look as follows:

((raccoon:19.19959,bear:6.80041):0.84600,((sea_lion:11.99700,
seal:12.00300):7.52973,((monkey:100.85930,cat:47.14069):20.59201,
weasel:18.87953):2.09460):3.87382,dog:25.46154);

If you wish to specify bootstrap values then they could appear
before the branch lengths (e.g., in .DND files produced by CLUSTAL)
or after the branch lengths (e.g., in .PHB files produced by
CLUSTAL).  In these cases, the format might look like:

((raccoon:19.19959,bear:6.80041)50:0.84600,((sea_lion:11.99700,
seal:12.00300)100:7.52973,((monkey:100.85930,cat:47.14069)80:20.59
201, weasel:18.87953)75:2.09460)50:3.87382,dog:25.46154);
```

or

((raccoon:19.19959,bear:6.80041):0.84600[50],((sea_lion:11.99700, seal:12.00300):7.52973[100],((monkey:100.85930,cat:47.14069):20.59201[80], weasel:18.87953):2.09460[75]):3.87382[50],dog:25.46154);

7.436 Node

A node in a phylogenetic tree represents a taxon, the *external* or *terminal nodes* represent the extant taxa and the *internal nodes* represent the ancestral taxa.

7.437 Nonsynonymous change

A nucleotide change is nonsynonymous if it changes the amino acid encoded by the original codon. A nucleotide site in which one or more changes are nonsynonymous is referred to as a nonsynonymous site. If only one of three possible nucleotide changes at that site is nonsynonymous, then the site is 1/3 nonsynonymous. If two of three nucleotide changes are nonsynonymous, then the site is 2/3 nonsynonymous. And, if all three possible nucleotide changes are nonsynonymous, then the site is completely nonsynonymous.

7.438 Nucleotide Pair Frequencies

When two nucleotide sequences are compared, the frequencies of 10 or 16 different types of nucleotide pairs can be computed. In *MEGA*, these frequencies are presented in a text file.

7.439 OLS branch length estimates

The ordinary least squares estimate of a branch length (*b*) is given by

$$b = \sum_{i<j} w_{ij} d_{ij},$$

where *d*ij is the pairwise distance between sequences *i* and *j*. The coefficients *w*ij's depend on whether the branch under consideration is internal or external.

Coefficients wij's for an internal branch

$$w_{ij} = \begin{cases} -1/(2m_A m_B), & \text{if } i \text{ belongs to cluster A and } j \text{ to cluster B} \\ (m_B m_C + m_A m_D)/[(m_A + m_B)(m_C + m_D)(2m_A m_C)], & \text{if } i \text{ belongs to cluster A and } j \text{ to cluster C} \\ (m_A m_C + m_B m_D)/[(m_A + m_B)(m_C + m_D)(2m_A m_D)], & \text{if } i \text{ belongs to cluster A and } j \text{ to cluster D} \\ (m_A m_C + m_B m_D)/[(m_A + m_B)(m_C + m_D)(2m_B m_C)], & \text{if } i \text{ belongs to cluster B and } j \text{ to cluster C} \\ (m_B m_C + m_A m_D)/[(m_A + m_B)(m_C + m_D)(2m_B m_D)], & \text{if } i \text{ belongs to cluster B and } j \text{ to cluster D} \\ -1/(2m_C m_D), & \text{if } i \text{ belongs to cluster C and } j \text{ to cluster D} \\ 0, & \text{if both } i \text{ and j belong to the same cluster} \end{cases}$$

where, *m*A, *m*B, *m*C, and *m*C are the numbers of sequences in clusters A, B, C, and D, respectively.

Coefficients wij's for an external branch

$$w_{ij} = \begin{cases} 1/(2m_A), & \text{if } i \text{ is the only member of cluster C and } j \text{ belongs to cluster A} \\ 1/(2m_B), & \text{if } i \text{ is the only member of cluster C and } j \text{ belongs to cluster B} \\ -1/(2m_A m_B), & \text{if } i \text{ belongs to cluster A and } j \text{ to cluster B} \\ 0, & \text{if both } i \text{ and j belong to the same cluster} \end{cases}$$

where, *m*A and *m*B are the numbers of sequences in clusters A and B.

7.440 Orthologous Genes

Two genes are said to be orthologous if they are the result of a speciation event.

7.441 Outgroup

An outgroup is a sequence (or set of sequences) that is known to be a sister taxa to all other sequences in the dataset.

7.442 Pairwise-deletion option

In the pairwise-deletion option, sites containing missing data or alignment gaps are removed from the analysis as the need arises (e.g., pairwise distance computation). This is in contrast to the complete-deletion option in which all such sites are removed prior to the analysis.

7.443 Parsimony-informative site

A site is parsimony-informative if it contains at least two types of nucleotides (or amino acids), and at least two of them occur with a minimum frequency of two.

7.444 Polypeptide

A polypeptide is a chain of many amino acids.

7.445 Positive selection

At the DNA sequence level, positive selection refers to selection in favor of nonsynonymous substitutions. In this case, the evolutionary distance based on nonsynonymous substitutions is expected to be greater than synonymous substitutions.

7.446 Protein parsimony

A Maximum Parsimony analysis on protein sequences is known as protein parsimony.

7.447 Purifying selection

Purifying selection refers to selection against nonsynonymous substitutions at the DNA level. In this case, the evolutionary distance based on synonymous substitutions is expected to be greater than the distance based on nonsynonymous substitutions.

7.448 Purines

The nucleotides adenine (A) and guanine (G) are known as purines.

7.449 Pyrimidines

The nucleotides cytosine (C) and thymine (T) are known as pyrimidines.

7.450 Random addition trees

This refers to the generation of random initial trees for a heuristic search to find MP trees. In this case, a tree is generated by randomly selecting a sequence and adding it to the growing tree on a randomly-selected branch.

7.451 Rooted Tree

A rooted tree is one in which the root of the phylogenetic tree is determined by using the mid-point rooting or outgroups sequences.

7.452 RSCU

Many amino acids are coded by more than one codon; thus multiple codons for a given amino acid are synonymous. However, many genes display a non-random usage of synonymous codons for specific amino acids. A measure of the extent of this non-randomness is given by the Relative Synonymous Codon Usage (RSCU) (Sharp et al. 1986).

The RSCU for a particular codon (i) is given by

$$RSCU_i = X_i / \sum\sum X_i /n$$

where X_i is the number of times the i_{th} codon has been used for a given amino acid, and *n* is the number of synonymous codons for that amino acid.

7.453 Singleton Sites

A singleton site contains at least two types of nucleotides (or amino acids) with, at most, one occurring multiple times. *MEGA* identifies a site as a singleton site if at least three sequences contain unambiguous nucleotides or amino acids.

7.454 Staden

The Staden file format is used to store data from DNA sequencing instruments. Each file contains the data for a single reading and includes the called sequence as well as additional data obtained from the reading. This file format was first described in

Dear, S and Staden, R. "A Standard file format for data from DNA sequencing instruments", DNA Sequence 3, 107-110, (1992)

MEGA is able to display the contents of a Staden-formatted trace file using MEGA's Trace File Editor, which is part of the Alignment Explorer.

7.455 Statements in input files

All statements in *MEGA* files start with an exclamation mark (!) and end with a semicolon (;). They are useful in specifying various attributes of the data and the data file. There are three common statements for all types of data: `Title`, `Format`, and `Description`. There also are other statements that can be used in *MEGA* files, depending on the type of data being analyzed.

7.456 Swap

This command changes the cursor to the 'Flip' icon. Then, you click on an interior branch, *MEGA* swaps the two subtrees defined by this branch. If each of the subtrees is an individual taxon, then Swap is the same as Flip.

The cursor will revert to the arrow if you click on the arrow icon on the left-hand side of the *Tree Explorer*.

7.457 Synonymous change

A nucleotide change is synonymous if it does not cause the codon to code for a different amino acid. A nucleotide site in which one or more changes is synonymous is referred to as a synonymous site. If only one of three possible nucleotide changes at that site is synonymous, then the site is 1/3 synonymous. If two of three nucleotide changes are synonymous, then the site is 2/3 synonymous and 1/3 nonsynonymous. And, if all three possible nucleotide changes are synonymous, then the site is completely synonymous.

7.458 Taxa

A taxon is the individual unit whose evolutionary relationship is being investigated. Depending on the study, "taxa" may refer to species, populations, individuals, or sequences within an individual.

7.459 Topological distance

The topological distance quantifies the extent of topological differences between two given trees. For unrooted, bifurcating trees, this distance is twice the number of interior branches at which the taxa are partitioned differently.

7.460 Topology

The branching pattern of a tree is its topology.

7.461 Transition

A transition occurs when a purine is substituted by a purine, or a pyrimidine by a pyrimidine.

7.462 Transition Matrix

A transition matrix specifies the probability of every possible substitution among the nucleotides or amino acids.

7.463 Transition/Transversion Ratio (R)

This is the ratio of the number of transitions to the number of transversions for a pair of sequences. *R* becomes 0.5 when there is no bias towards either transitional or transversional substitution because, when the two kinds of substitution are equally probable, there are twice as many possible transversions as transitions. *MEGA* allows you to conduct an analysis of your data with a specified value of R.

Note that *R* should not be confused with the ratio of the transition and transversion rates ($k = \alpha/\beta$).

7.464 Translation

Translation is the process whereby each codon in the mRNA is translated into a particular amino acid, according to the genetic code specific to the species and its DNA, and added to the growing polypeptide chain.

7.465 Transversion

A change from a purine to a pyrimidine, or vice versa, is a transversion.

7.466 Tree length

Tree length is the criterion used by the Maximum Parsimony method to search for the best tree. It is defined as the sum of the minimum numbers of substitutions over all sites for the given topology.

To compute the tree length for the unweighted parsimony method, we use the procedure described in Fitch (1971), which is based on the two rules described below. For a given site these rules are applied to each node and the sum of substitutions over all nodes and over all sites is taken. Note that the estimation of the minimum number of substitutions is not affected by the position of the root.

Rule 1. When the two descendent nodes of an ancestral node have some states (nucleotides or amino acids) in common, the ancestral node is assigned to the set of common states. In this case, the most parsimonious explanation does not require any substitutions.

Rule 2. When the two descendant nodes have no states in common, then all states in the descendent nodes are combined to form the set of possible states at the ancestral node. In this case, one substitution is required.

7.467 Unrooted tree

An unrooted tree is one in which no assumption is made regarding the ancestor of all the taxa in the tree.

7.468 Variable site

A variable site contains at least two types of nucleotides or amino acids. Some variable sites can be singleton or parsimony-informative. A site that is not variable is referred to as a constant site.

7.5 Reference

Comeron JM (**1995**) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution* **41**:1152-1159.

Dayhoff MO (**1978**) Survey of new data and computer methods of analysis. In Dayhoff MO, ed., *Atlas of Protein Sequence and Structure*, vol. **5**, supp. 3, pp. 29, National Biomedical Research Foundation, Silver Springs, Maryland.

Schwarz R & Dayhoff M (**1979**) Matrices for detecting distant relationships. In Dayhoff M, editor, *Atlas of protein sequences*, pages 353 - 58. National Biomedical Research Foundation.

DeBry RW (**1992**) The consistency of several phylogeny-inference methods under varying evolutionary rates. *Molecular Biology and Evolution* **9**:537-551.

Dopazo J (**1994**) Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. *Journal of Molecular Evolution* **38**:300-304.

Eck RV & Dayhoff MO (**1966**) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Springs, Maryland.

Efron B (**1982**) *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, SIAM, Philadelphia.

Estabrook GF, Johnson CS & McMorris FR (**1975**) An idealized concept of the true cladistic character. *Mathematical Biosciences* **23**:263-272.

Felsenstein J (**1978**) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**:401-410.

Felsenstein J (**1985**) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-791.

Felsenstein J (**1986**) Distance Methods: Reply to Farris. *Cladistics* **2**:130-143.

Felsenstein J (**1988**) Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22**:521-565.

Felsenstein J (1993) Phylogeny Inference Package (PHYLIP). Version 3.5. University of Washington, Seattle.

Felsenstein J & Kishino H (**1993**) Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology* **42**:193-200.

Fitch WM (**1971**) Towards defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology* **20**:406-416.

Fitch WM & Margoliash E (**1967**) Construction of phylogenetic trees. *Science* **155**:279 284.

Goldman N (**1993**) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**:182-198.

Gu X & Zhang J (**1997**) A simple method for estimating the parameter of substitution rate variation among sites. *Molecular Biology and Evolution* **15**:1106-1113.

Hedges SB, Kumar S, Tamura K, & Stoneking M (**1992**). Human origins and analysis of mitochondrial DNA sequences. *Science* **255**:737-739.

Hendy MD & Penny (D) (**1982**) Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* **59**:277-290.

Hendy M D & Penny D (**1989**) A framework for the quantitative study of evolutionary trees. *Systematic Zoology* **38**:297-309.

Hillis DM & Bull JJ (**1993**) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**:182-192.

Hillis DM, Moritz C & Mable BK (**1996**) *Molecular Systematics*. 2 edition. Sunderland, MA: Sinauer Associates, Inc.

Jones DT, Taylor WR & Thornton JM (**1992**) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**: 275-82.

Jukes TH & Cantor CR (**1969**) Evolution of protein molecules. In Munro HN, editor, *Mammalian Protein Metabolism*, pp. 21-132, Academic Press, New York.

Kimura M (**1980**) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**:111-120.

Kishino H & Hasegawa M (**1989**) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **29**:170- 179.

Kumar, S. and S. R. Gadagkar (**2001**) Disparity Index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* **158**:1321-1327.

Kumar S, Tamura K & Nei M (**1993**) *MEGA: Molecular Evolutionary Genetics Analysis*. Pennsylvania State University, University Park, PA.

Kumar S, Tamura K & Nei M (**2004**) MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. *Briefings in Bioinformatics* **5**:150-163.

Lake JA (**1987**) A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution* **4**:167-191.

Li W-H (**1993**) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**:96-99.

Li W-H (**1997**) *Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Li W-H, Wu C-I & Luo C-C (**1985**) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**:150-174.

Maddison WP & Maddison DR (**1992**) MacClade: Analysis of phylogeny and character evolution. Version 3. Sinauer Associates, Sunderland, Massachusetts.

Nei M (**1986**) Stochastic errors in DNA evolution and molecular phylogeny. In Gershowitz H, Rucknagel DL, & Tashian RE, editors, *Evolutionary Perspectives and the New Genetics*. pp. 133-147. Alan R. Liss, New York.

Nei M (**1991**) Relative efficiencies of different tree making methods for molecular data. In Miyamoto MM and Cracraft JL, editors, *Recent Advances in Phylogenetic Studies of DNA Sequences*, pp. 90-128. Oxford University Press, Oxford.

Nei M & Gojobori T (**1986**) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**:418-426.

Nei M & Jin L (**1989**) Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution* **6**:290-300.

Nei M & Kumar S (**2000**) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nei M, Chakraborty R & Fuerst PA (**1976**) Infinite allele model with varying mutation rate. *Proceedings of National Academy of Sciences* (*USA*) **73**:4164-4168.

Nei M, Stephens JC & Saitou N (**1985**) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Molecular Biology and Evolution* **2**:66-85.

Nei M, Kumar S & Takahashi (**1998**) The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proceedings of National Academy of Sciences* (*USA*) **95**:12390-12397

Page RDM & Holmes EC (**1998**) *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford, U.K.

Pamilo P& Bianchi NO (**1993**) Evolution of the Zfx and Zfy, genes: Rates and interdependence

between the genes. *Molecular Biology and Evolution* **10**:271-281.

Penny D & Hendy MD (**1985**) The use of tree comparison metrics. *Systematic Zoology* **34**:75-82.

Pamilo P & Nei M (**1988**) Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**:568-583.

Press WH, Flannery BP, Teukolsky *SA* & Vetterling WT (**1989**) *Numerical Recipes in Pascal: The Art of Scientific Computing*. Cambridge University Press, New York.

Purdom PW, Bradford PG, Tamura K & Kumar S (**2000**) Single column discrepancy and dynamic max-mini optimizations for quickly finding the most parsimonious evolutionary trees. *Bioinformatics* **16**:140-151.

Rzhetsky A & Nei M (**1992**) A simple method for estimating and testing minimum evolution trees. *Molecular Biology and Evolution* **9**:945-967.

Rzhetsky A & Nei M (**1993**) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution* **10**:1073-1095.

Saitou N & Nei M (**1987**) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406-425.

Sankoff D & Cedergren RJ (**1983**) Simultaneous comparison of three or more sequences related by a tree. In Sankoff D & Kruskal JB, editors., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pp. 253-263. Addison-Wesley, Reading, Massachusetts.

Sharp PM, Tuohy TMF & Mosurski KR (**1986**) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**:5125-5143.

Sneath PHA & Sokal RR (**1973**) *Numerical Taxonomy. Freeman*, San Francisco.

Sourdis J & Krimbas C (**1987**) Accuracy of phylogenetic trees estimated from DNA sequence data. *Molecular Biology and Evolution* **4**:159-166.

Sourdis J & Nei M (**1988**) Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Molecular Biology and Evolution* **5**:298-311.

Studier, J. A. and K. L. Keppler. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution* **5**:729-731.

Swofford DL (**1993**) Phylogenetic Analysis Using Parsimony (PAUP), Version 3.1.1. University of Illinois, Champaign.

Swofford DL (**1998**) *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)* Sunderland, MA: Sinauer Associates.

Swofford DL, Olsen GJ, Waddell PJ & Hillis DM (**1996**). Phylogenetic Inference. In Hiillis DM, Moritz D, and Mable BK, editors, *Molecular Systematics*, pp. 407-514.Sinauer Associates, Sunderland, Massachusetts.

Tajima F (**1983**) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**:437-460.

Tajima F (**1989**) Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics* **123**:585-595.

Tajima F (**1993**) Simple methods for testing molecular clock hypothesis. *Genetics* **135**:599-607.

Tajima F & Nei M (**1982**) Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *Journal of Molecular Evolution* **18**:823-833.

Tajima F & Nei M (**1984**) Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* **1**:269-285.

Takahashi K & Nei M (**2000**) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution* **17**:1251-1258.

Takezaki N, Rzhetsky A & Nei M (**2004**) Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution* **12**:823-833.

Tamura K (**1992**) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Molecular Biology and Evolution* **9**:678-687.

Tamura K (**1994**) Model selection in the estimation of the number of nucleotide substitutions. *Molecular Biology and Evolution* **11**:154-157.

Tamura K and S Kumar (**2002**) Evolutionary distance estimation under heterogeneous substitution pattern among lineages *Molecular Biology and Evolution* **19**:1727-1736.

Tamura K & Nei M (**1993**) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**:512-526.

Tamura K, Nei M & Kumar S (**2004**) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences* (*USA*) **101**:11030-11035.

Tanaka T & Nei M (**1989**) Positive Darwinian selection observed at the variable-region genes of immunoglobulins. *Molecular Biology and Evolution* **6**:447-459.

Tateno Y, Nei M & Tajima F (**1982**) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *Journal of Molecular Evolution* **18**:387-404.

Tateno Y, Takezaki N & Nei M (**1994**) Relative efficiencies of the maximum likelihood, neighbor-joining, and maximum parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution* **11**:261-277.

Yang Z (**1999**) PAML: Phylogenetic analysis by maximum likelihood, Version 2.*0*. University College London, London.

Zhang J & Gu X (**1998**) Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* **149**:1615-1625.

Zhang J, Kumar S & Nei M (**1997**) Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Molecular Biology and Evolution* **14**:1335-1338.

Zhang J, Rosenberg HF & Nei M (**1998**). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences* (*USA*) **95**:3708-3713.

Zuckerkandl E & Pauling L (**1965**) Evolutionary divergence and convergence in proteins, pp. 97-166 in *Evolving Genes and Proteins*, edited by V. Bryson and H.J. Vogel. Academic Press, New York.

Index

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z