MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment

Sudhir Kumar, Koichiro Tamura and Masatoshi Nei

Date received (in revised form): 1st April 2004

Sudhir Kumar
has been Director of the Center for Evolutionary Functional Genomics in The Biodesign Institute at Arizona State University since 2002. His research interests include development of software, statistical methods and computational tools for comparative sequence analysis. He and Koichiro Tamura are joint first authors of the MEGA3 software.

Koichiro Tamura
is a molecular evolutionist at the Tokyo Metropolitan University. His research interests are in the area of experimental, theoretical and computational molecular evolution.

Masatoshi Nei
is Director of the Institute of Molecular Evolutionary Genetics at the Penn State University since 1990. His research interests are in molecular evolutionary genetics and genomics.

Keywords: *evolution, genomics, software, data mining, sequence alignment, distance, phylogenetics, selection*

Abstract

With its theoretical basis firmly established in molecular evolutionary and population genetics, the comparative DNA and protein sequence analysis plays a central role in reconstructing the evolutionary histories of species and multigene families, estimating rates of molecular evolution, and inferring the nature and extent of selective forces shaping the evolution of genes and genomes. The scope of these investigations has now expanded greatly owing to the development of high-throughput sequencing techniques and novel statistical and computational methods. These methods require easy-to-use computer programs. One such effort has been to produce Molecular Evolutionary Genetics Analysis (MEGA) software, with its focus on facilitating the exploration and analysis of the DNA and protein sequence variation from an evolutionary perspective. Currently in its third major release, MEGA3 contains facilities for automatic and manual sequence alignment, web-based mining of databases, inference of the phylogenetic trees, estimation of evolutionary distances and testing evolutionary hypotheses. This paper provides an overview of the statistical methods, computational tools, and visual exploration modules for data input and the results obtainable in MEGA.

INTRODUCTION

Genome sequencing in large-scale and individual laboratory projects have generated vast amounts of data from diverse organisms. Comparative sequence analyses, performed under the principles of molecular evolutionary genetics, are essential for using these data to build the tree of life, infer the evolutionary patterns of genome and species evolution, and elucidate mechanisms of evolution of various morphological and physiological characters. The need for software to perform these tasks is now well recognised.[1–7] This software must contain fast computational algorithms and useful statistical methods and have an extensive user-friendly interface to enable experimentalists working at the forefront of sequence data generation to discover novel patterns and explore basic sequence attributes.

This need motivated the development of MEGA (Molecular Evolutionary Genetics Analysis software) in the early 1990s. From its inception, our goal for the MEGA software has been to make available a wide variety of statistical and computational methods for comparative sequence analysis in a user-friendly environment.[8–10] The first version of MEGA,[10] released in 1993, was distributed to over 2,000 scientists. The second version, MEGA2,[9] released in 2001, was a complete rewrite of the first version, and took advantage of the manifold increase in computing power of the average desktop computer and the availability of the Microsoft Windows graphical interfaces. The user-friendliness and methodological advances of MEGA2 and the increased scope of the molecular evolutionary analyses performed by the scientific community led to a large increase in the number of users from around the world. A survey of the

Sudhir Kumar,
Life Sciences A-351,
The Biodesign Institute,
Tempe, AZ 85287-4501, USA

Tel: +1 (480) 727 6949
Fax: +1 (480) 965 6899
E-mail: s.kumar@asu.edu

research papers citing the use of MEGA reveals that it has been utilised in diverse disciplines, including AIDS/HIV research, virology, bacteriology and general disease, plant biology, conservation biology, systematics, developmental evolution and population genetics.

MEGA contains automatic and manual sequence alignment facilities

The newly released MEGA3 expands the functionalities of MEGA2 by adding sequence data alignment and assembly features, along with other advancements. The sequence data acquisition is now effectively integrated with the evolutionary analyses, making it much easier to conduct comparative analyses in an integrated computing environment. However, MEGA3 is not intended to be a catalogue of all evolutionary analysis methods. Rather, it is for exploring sequence data from evolutionary perspectives, constructing phylogenetic trees, and testing evolutionary hypotheses, especially for large-scale data sets that have been generated by recent genomics projects.

Built-in web browser streamlines collection of data from GenBank and facilitates BLAST search

The following is a brief overview of the functionality and facilities available in MEGA3. It begins with a description of the newest additions to MEGA – the sequence alignment and data assembly modules – as they constitute the first step in any comparative sequence analysis investigation. This is followed by descriptions of the different types of data that MEGA can analyse, its graphical input and output data explorers, dynamic data subsetting facilities, and the statistical methods and computational tools available for inferring phylogenetic trees and estimating evolutionary distances.

SEQUENCE ACQUISITION AND ALIGNMENTS

Sequence alignment is usually the first step in comparative sequence analysis. It is the process of identifying homologous nucleotide (or amino acid) positions among a set of sequences. Building these alignments involves many steps, including acquiring sequences from databanks, performing computational sequence alignments, and manual fine-tuning of the initial alignment.

Data acquisition in MEGA

Scientists routinely obtain gene sequences from databanks[11,12] using a web browser. Homologous sequences usually are searched in the BLAST procedure by using either a gene name (or other attributes such as the GenBank accession numbers) or a query sequence.[13–15] In both cases, a set of sequences is found and displayed on the computer screen. From this set, researchers may select all or some of the sequences based on specific criteria, for example, taxonomic sampling of chosen species and/or sequence matching score. Usually at this point, investigators begin the mundane, frustrating task of cutting and pasting the sequences from the web-browsers or saving them to files and then processing them for sequence alignment.

To streamline this process, MEGA now includes an integrated web-browsing facility (Figure 1). Researchers can use it in the same way as the commercial web browsers, such as Internet Explorer or Netscape Navigator. Because the MEGA3 web-browsing facility is a wrapper around the full-function HTML browser in the Microsoft Windows operating system, it works even if a commercial browser is not installed on the computer. The MEGA browser therefore can be used as a general-purpose web browser.

In the MEGA web browser, once investigators have generated the list of desired sequences, they click on 'ADD TO ALIGNMENT' whereupon MEGA parses the sequences automatically and sends them to the Alignment Explorer (AE) (Figure 2). This web exploration and data retrieval system will help investigators in their everyday activities without the need to reinvent protocols and allows them to use the novel and modified data searching capabilities provided by GeneBank and other servers without requiring a MEGA3 upgrade.

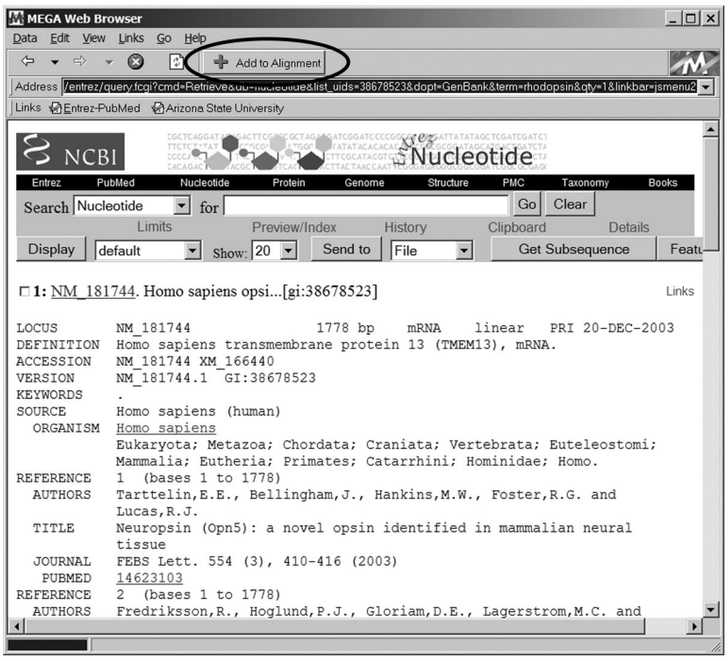


Figure 1: Web browser module in MEGA for accessing databanks and retrieving DNA and protein sequences

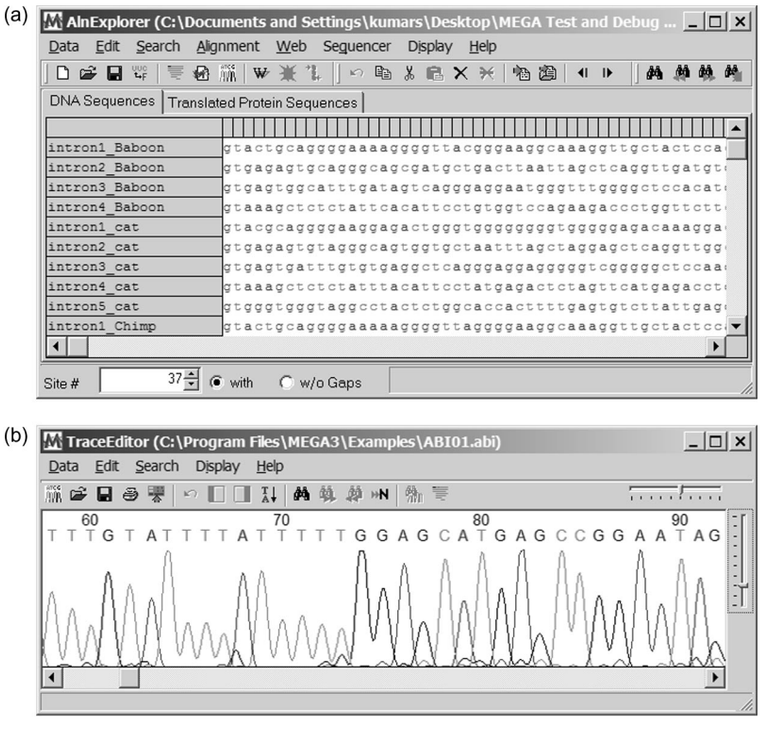


Figure 2: (a) The Alignment Explorer in MEGA for creating, viewing and editing DNA and protein sequence alignments manually and by using ClustalW.[16,17] (b) Sequencer Trace file editor to read ABI and SCF file formats from automated sequencers

Alignment Explorer (AE) Tool

A versatile tool for building DNA and protein sequence alignments requires

- an extensive graphical user interface with facilities to edit sequence data such as the manual insertion of gaps and

reverse complementation of DNA;

- a computational capability for automated multiple sequence alignment;
- services for aligning coding sequences intuitively at DNA as well as protein sequence levels; and
- facilities for the easy importing and exporting of sequence data.

The Alignment Explorer in MEGA contains these features along with a number of other advanced facilities (Figure 2a).

MEGA aligns codons according to their protein sequences

The AE offers two views of the data: DNA and (translated) amino acid sequences. These two views are present in alignment grids (Figure 2a). (For amino acid input sequence data, there is only one such grid.) In the grid, each row represents a single sequence and each column represents a site. Identity across all sequences is indicated by a '*' character in the top row of each column.

Alignment session can be saved for future use

For automated sequence alignment, the AE includes a native implementation of CLUSTALW,[16,17] the most widely used multiple sequence alignment system for DNA and protein data. Sequence alignments can be edited manually and other operations on individual sites, columns and blocks can be performed with just a few mouse clicks. The AE provides unlimited undo capabilities.

CLUSTALW alignment can be done for coding and non-coding regions separately

The AE allows the user to construct alignments intuitively. Users can mark a rectangle (rows and columns of the source sequence) for alignment, invoke the integrated multisequence alignment module (CLUSTALW), specify appropriate alignment parameter values and initiate the sequence alignment. On the completion of the alignment, the AE automatically inserts the aligned sequences back into the source rectangle by expanding or contracting it appropriately. This allows for aligning of different regions of the sequence independently. For example, protein-coding nucleotide regions can be aligned separately from non-coding regions. For the protein-coding regions, users can translate the selected sequences (or chosen rectangle) into protein sequences by a single mouse-click, align the translated protein sequences using CLUSTALW, and then flip back to DNA sequences. The AE automatically adjusts the source nucleotide sequences as per the amino acid sequence alignment. Translated protein sequences can be further aligned manually even before the user comes back to the DNA sequences, thus replacing a multi-step error-prone manual process by a simple and intuitive procedure.

For researchers who would like to complete their task at a later time, MEGA freezes an AE session exactly as it is by saving it to the file in an 'alignment session format'. This facility is also useful for retaining convenient settings for the future construction and expansion of alignments. This system of freezing the AE session is different from writing alignments into text file formats for use in other programs. For this purpose, MEGA provides options to save data to NEXUS,[18,19] MEGA[9,10] and PHYLIP[20] for further analysis.

MEGA has a Sequencer Trace File editor

MEGA also includes facilities for handling trace files (Figure 2b). Users can view and edit the trace data (electropherogram) produced by the automated DNA sequencers. This viewer/editor can read and edit data in ABI (Applied Biosystems)and SCF (Source Comparison interchange format or Staden format) trace file formats. The displayed sequences can be added directly into AE or sent to the web browser for conducting BLAST searches, among other things. Therefore, AE is a versatile tool for building and expanding sequence alignments.

INPUT DATA AND FORMATS

MEGA facilitates the molecular evolutionary analysis of DNA and amino acid sequences and of pairwise distance matrices. For these purposes, sequences

User can designate genes, domains and sites in special categories

(or taxa) can be placed into separate groups, such as orthologous groups of genes in a multigene family data set or different sets of related taxa. For sequence data, researchers can define domains (continuous blocks of nucleotides, codons or amino acid sites) and genes (collections of domains). Domains can be coding or non-coding, and the codon start for coding domains can be specified. MEGA also allows for the assignment of individual nucleotide sites, codons or amino acid sites into user-defined categories, each category being represented by a single character (eg a letter or a digit). This facilitates analyses requiring collections of non-contiguous sites, such as the DNA binding or antigen recognition sites. All of these data attributes (eg domains, genes, groups and category labels) can be set visually using simple drag-and-drop operations. They also can be read from and saved to ASCII text files. An example of an input data file, showing how the sequence and site attributes are specified within the data 'on the spot', is given in Figure 3.

Sequences can be assembled into groups for statistical analysis

Construction of data subsets

Pairwise- and Complete-deletion options are available for analysing sequences with alignment gaps

In MEGA, data subsetting operations can be achieved by using simple point-and-click manoeuvres. For the first level of data subset construction, users can create subsets containing virtually any combination of sequences (taxa), including groups, domains and genes. At the time of analysis, users can include or exclude data with missing information and/or with alignment gaps, which, along with the selection of nucleotide codon positions and site categories, constitutes the second level of data subset construction. MEGA can remove all sites (or codons) containing missing data and alignment gaps prior to analysis (called the complete-deletion) or dynamically as the need arises during the analysis (called the pairwise-deletion).[2,10] The third level of data subsetting and transformation is done automatically in the codon-by-codon analyses. If the selected data subset contains non-coding as well as coding regions, MEGA automatically extracts all complete codons. Similarly, if the analysis requires translation into amino acid sequences, MEGA will do it automatically.

These three levels of data handling are designed to eliminate error-prone manual data editing. They provide a powerful, yet simple, framework for generating desired data subsets. When any specific analysis is

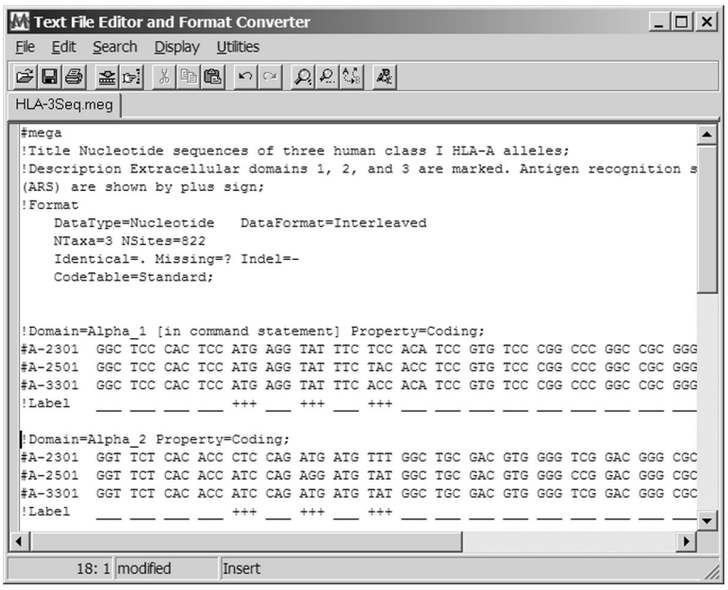


Figure 3: An example of a sequence data file in the MEGA format

carried out, MEGA concatenates all the selected domains and genes.

Format conversion and Text Editor

MEGA supports conversions from several different file formats into the MEGA format. Currently supported data file formats include CLUSTAL,[16] NEXUS (PAUP, McClade),[19] PHYLIP (interleaved and non-interleaved),[20] GCG,[21] FastA, PIR, NBRF, MSF, IG and XML (NCBI).[11] The format conversion utility is available in the text file editor. The text file editor is useful for creating and editing ASCII text files and is automatically invoked by MEGA if the input data file processing modules detect errors in the data file format.

MEGA text editor can handle very large text and data files, and supports rectangular cut-and-paste and unlimited undo

The MEGA Text File Editor (Figure 3) is similar to the Microsoft Windows' NotePad and WordPad accessories. Two features of Text Editor are likely to be especially useful to genomics researchers. First is its ability to open and edit very large files, even larger than many millions of bytes and containing many large contigs. Second, it provides the user with the ability to select text in *rectangles* and do cut-and-paste operations on these selections. In addition, the Text Editor contains utilities to remove or insert spaces in the text and do a reverse complement on any selected text! Therefore, a number of simple utilities in the Text Editor can come in handy for routine text file editing for molecular sequence handling.

Exploring sequence data

Positions with special attributes can be highlighted in Data Explorer

An important feature of MEGA is the presence of an input Sequence Data Explorer (SDE), which allows investigators to browse attributes of sequence data and export those data to other formats. Researchers may also use it to view an alignment to make sure that the data has been interpreted correctly by MEGA.

The SDE displays sequences in a two-dimensional grid (Figure 4). Faint grey boxes outlining each codon mark protein-coding regions of the DNA sequences. These segments can be easily translated and untranslated by a click of a button. Sequences can be arranged in this viewer by drag-and-drop operations; clicking the box preceding the sequence name in the left column changes its inclusion status. All sequences and sites not currently selected are greyed out.

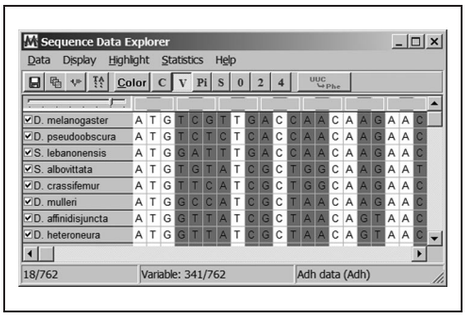


Figure 4: Input Data Explorer for sequences

The SDE can compute many different attributes for columns (sites), including site variability (invariable, variable, singleton, parsimony-informative) and degeneracy (0-, 2- and 4-fold).[8,9] Users can highlight sites with these attributes, one attribute at any given time. These attributes are dynamically computed, so if some sequences in the current data subset are excluded or included, the attributes will be updated immediately.

Basic statistical quantities

MEGA can compute base frequencies, codon usage bias, and nucleotide pair frequencies between sequences

The SDE also contains facilities for computing the nucleotide and amino acid frequencies and relative codon usage biases (RSCU)[22] in sequences. For pairs of sequences, the SDE provides the frequencies of different pairs of nucleotides (10 or 16 pairs) and the transition/transversion ratios. These computations can be carried out for all the data or only for the sites that are highlighted.

ESTIMATING EVOLUTIONARY DISTANCES

Estimating the number of nucleotide or amino acid substitutions needed to

compute evolutionary distances is one of the most important subjects in molecular evolutionary genetics and comparative genomics. Evolutionary distances are required for reconstructing phylogenetic trees, assessing sequence diversity within and between groups of sequences, and estimating times of species divergence, among other things.[2,3]

MEGA can handle base composition and transition/transversion biases as well as the substitution pattern heterogeneity among lineages

MEGA contains many statistical methods for estimating the evolutionary distance (actual number of substitutions per site) between sequences based on the observed number of differences. The methods included correct for multiple substitutions by taking into account the transition/transversion bias, unequal base frequencies, varying substitution rates among sites, and heterogeneous substitution patterns among lineages.[2,3,23] Researchers can choose any of these options from a simple dialogue box (Figure 5).

MEGA divides distances into three groups – nucleotide, synonymous–non-synonymous and amino acid – based on the properties of the sequence data and the type of substitutions being considered. Nucleotide distances estimate the number of nucleotide substitutions per site between DNA sequences. Analytical formulas for estimating these distances under many substitution models are included in MEGA (Table 1). Under some models, numbers of transition and transversion substitutions per site also can be estimated separately.[23–26] For these cases, MEGA also provides facilities for computing the transition/transversion bias. For all models that contain parameters to account for base frequency bias,[25–27] MEGA includes the Tamura–Kumar[23] modification of the original distance estimation formulas to account for the differences (if any) in base composition biases among lineages (heterogeneity of substitution patterns). Analytical formulas for incorporating rate variation among sites (according to a gamma distribution) under the above-mentioned models are also included (see Table 1). The user needs to provide the shape parameter of the gamma distribution for this purpose, which can be computed using a variety of methods in other software, for example Yang,[31] Gu and Zhang[32] and Yang and Kumar.[33]

Synonymous and non-synonymous distances are computed by comparing codons between sequences using a specified genetic code table. A non-

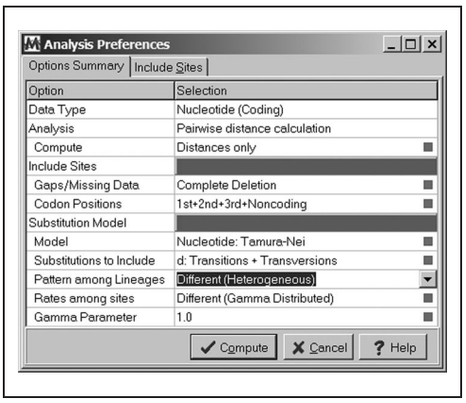


Figure 5: Distance method selection environment in MEGA

Table 1: Nucleotide substitution models for estimating evolutionary distances

Substitution model	Transition/ transversion bias	Base frequency bas	Rate variation among sites	Heterogeneous patterns among lineages
Jukes–Cantor[28]			Yes	
Kimura[24]	Yes		Yes	
Tamura[26]	Yes	Yes (G+C)	Yes	Yes
Tajima–Nei[27]		Yes	Yes	Yes
Tamura–Nei[25]	Yes	Yes	Yes	Yes
Log-Det[23,29,30]	Yes	Yes		Yes

Synonymous and non-synonymous distances can be computed

synonymous change is a substitution in a codon that causes a different amino acid to be encoded; a position in a codon in which non-synonymous substitutions occur is a non-synonymous site. The number of non-synonymous changes per non-synonymous site is the non-synonymous distance (d_N). A synonymous change is a substitution that does not change the encoded amino acid, and a synonymous distance (d_S) is defined in the same way as the non-synonymous distance. (Users have the flexibility of choosing a genetic code from the pre-loaded collection of all known genetic codes; they can also add a new genetic code table.)

For computing synonymous and non-synonymous distances, both the Nei–Gojobori[34] and Li–Wu–Luo[35] methods and their modifications,[9,36–40] which account for transition/transversion bias, are available in MEGA. With these methods, users can compute a number of quantities, including the numbers of substitutions per site at only synonymous sites, only non-synonymous sites, only four-fold-degenerate sites, and only zero-fold-degenerate sites. In addition, one also can estimate the difference between synonymous and non-synonymous distances.

Standard errors can be computed by bootstrap and analytical methods

MEGA can be used to estimate distances for amino acid sequences as well as the nucleotide sequences of protein-coding regions. It automatically translates the coding domains of sequences into amino acid sequences. Analytical formulas are included for Poisson correction distance,[2] equal input model,[3,41] and distances based on widely used Dayhoff[42] and Jones–Thornton–Taylor (JTT)[43] substitution matrices. Distances under Poisson correction and equal input models are computed using analytical formulas, whereas Dayhoff and JTT distances are computed using iterative procedures under a maximum likelihood formulation (as is done in PHYLIP[20]), in which the substitution rate matrix (between amino acids) published by Dayhoff[42] and JTT[43] are used. (Other substitution matrices will be included in the future.) For the equal input model, the heterogeneity of amino acid substitution patterns between sequences can be considered in estimating distances.[23] For all amino acid distances, methods are included to account for rate variation among amino acid positions in distance estimation.

PAM and JTT matrix based protein distances are now available in MEGA

As mentioned earlier, in MEGA researchers can arrange sequences into groups. Not all sequences need to belong to a group, but each sequence can belong only to one group. Once these groups are specified, the computational options for the estimating the average evolutionary divergence within and between groups become available. If the groups correspond to populations, users also have the option of using procedures that are specifically meant for population data, through the compute Sequence Diversity options in the Distances menu in MEGA.

Estimation of standard errors

For pairwise distances, MEGA contains analytical formulas for computing standard errors whenever available. However, analytical formulas for standard errors are approximate because they are based on a number of simplifying assumptions, including the assumption of large sample size. Furthermore, in many cases it is not possible to derive analytical expressions for computing variances and covariances. For instance, when average distances within and among groups of sequences are computed, formulas for computing the covariances required for obtaining the standard errors are either unavailable or are too cumbersome. In such cases, the bootstrap method[2,3,44] provides a convenient approach to computing the standard errors. This method does not require assumptions about the underlying distributions of the estimated distances and is made available in MEGA for the estimation of standard errors for all pairwise as well as average distances.

Disparity index and test for substitution pattern homogeneity

MEGA can examine the extent of substitution pattern heterogeneity

Methods for computing evolutionary distances that take into account the nucleotide (or amino acid) frequency bias for correcting multiple substitutions assume that the pattern of nucleotide substitution has remained the same throughout the evolutionary history of examined sequences. As mentioned above, MEGA includes improved versions of many of these methods, because if the homogeneity assumption is not satisfied, the distance estimate and all dependent inferences will be biased. Therefore, it is important to examine whether sequences in a data set have evolved with the same pattern of substitution. A knowledge of sequences and species evolving with heterogeneous patterns is also useful in understanding shifts in mutational patterns and selective pressures.[45,46]

MEGA can conduct codon-based tests for detecting positive selection

MEGA contains the disparity index test[46] for this purpose. This test is conducted using pairs of sequences (obviating the need to know the phylogenetic history) and it does not require knowledge of the actual pattern of substitution. In addition to the *P*-value of the disparity index test for rejecting the null hypothesis of homogeneous pattern, MEGA provides options for computing the compositional distance (based on the observed difference between nucleotide frequencies in two sequences) and the value of the disparity index normalised by the number of sites used in each pairwise comparison.

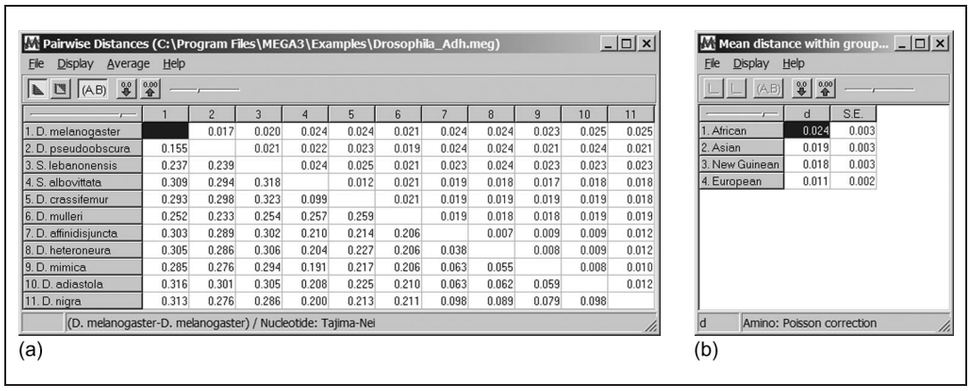
Exploring distance matrices

Results from evolutionary distance estimations are displayed in the Distance Matrix Explorer (Figure 6). This provides facilities for rearranging taxa by a simple drag-and-drop procedure and allows users to calculate overall and group averages based on individual distances when group information is available. It is capable of presenting standard errors and has options for displaying pairwise distances in the lower-left or upper-right matrices. Facilities for exporting distance matrices are also included.

TESTS OF SELECTION

For testing hypotheses of neutral and adaptive evolution at the molecular level, MEGA conducts codon-based tests that compare synonymous (d_S) and non-synonymous distances (d_N). In this case, the null hypothesis of $d_S = d_N$ (strict neutral evolution) can be tested by constructing a normal-deviate test using either the Nei–Gojobori[34] or Li–Wu–Lou[35] methods and their modifications.[9,36–40] In addition to testing $d_S = d_N$ for individual sequence pairs, researchers can conduct an overall test in which the average d_S is compared with the average d_N over the same set of sequence pairs. When groups of sequences are defined, tests of selection within groups can be conducted. In all of these cases, the bootstrap method is used to generate the standard error of the test statistic for conducting the normal deviate test.

For pairwise sequence analysis only, MEGA also provides the Fisher exact test[47]



(a) (b)

Figure 6: Distance matrix viewer showing distances and their standard errors for sequence pairs (a) and for within group averages (b)

for comparing the relative abundance of synonymous and non-synonymous substitutions. This is because when the number of differences between sequences is small, the normal-deviate test becomes too liberal in rejecting the null hypothesis.[47] MEGA also can compute the test statistic for Tajima's[48] test of neutrality. Therefore, selection can be detected and tested in different ways in MEGA.

INFERRING PHYLOGENETIC TREES

Phylogenetic trees infer the evolutionary relationships of species and patterns of gene duplications in multigene families. They are also important for elucidating the patterns and processes of molecular evolution through studies of adaptive and neutral evolutionary changes. MEGA contains both distance-based and maximum parsimony (MP) methods for phylogenetic reconstruction. It includes the Unweighted Pair Group Method with Arithmetic Mean (UPGMA),[49] the Neighbour-Joining (NJ)[50] method, and the Minimum Evolution (ME)[51,52] method for inferring phylogenetic trees using distance matrices. UPGMA is an agglomerative algorithm in which the tree is inferred, assuming constancy of the rate of evolution for all lineages. It should be used only if this assumption is satisfied. MEGA contains a non-parametric test of the molecular clock to compare the rate of evolution in two sequences, given an outgroup sequence.[53,54] The power of this test is similar to the Muse–Weir maximum likelihood ratio test.[53,55]

Neighbour-joining (NJ), UPGMA, minimum evolution (ME) and maximum parsimony (MP) methods are available

The NJ method does not make any assumption about rate constancy and constructs the tree hierarchically, such that the sum of branch lengths (S) under the ordinary least squares method is minimised in each step of taxa clustering. This is a computationally inexpensive method under the minimum evolution principle, and its accuracy is known to be comparable to other more time-consuming ME algorithms.[2,56–58] However, MEGA also contains the Close-Neighbour-Interchange (CNI) heuristic search algorithms for finding the optimal tree under the minimum evolution criterion. In this case, a temporary tree, such as the NJ tree, is generated and then all of the topologies that are different from this temporary tree by a topological distance[59] of 2 and 4 are examined. If a more optimal tree is found, then this process is repeated and all of the topologies previously examined are avoided. This stops when there are no more topologies to examine. The topologies with the smallest S-value are then chosen as the final trees. For fast computation of S-values during the ME tree search, we employ a dynamic procedure,[60] which appears to speed up the search by a factor of $m/10$, where m is the number of sequences. For all topologies, MEGA provides estimates of the branch lengths computed during the UPGMA, NJ and ME search procedures.

MEGA has a variety of search algorithms for finding ME and MP trees

For the MP criterion, MEGA treats all nucleotide and amino acid changes as unordered and reversible.[61,62] Branch-and-bound (max-mini algorithm[8]) as well as heuristic search methods (min-mini[8] and CNI[2]) for finding the optimal tree under the MP criterion are included. Optimisations such as the dynamic estimation of the cost function and the single column discrepancy procedures are implemented to reduce the computational time requirements.[63] For the MP trees inferred, MEGA provides estimates of branch lengths under the MP criterion using the average pathway method for unrooted trees.[2,64,65]

In the current version, MEGA automatically concatenates all selected genes and domains. This concatenation approach is known to be quite effective in inferring the correct tree.[66]

Robustness of inferred phylogeny

It is important to conduct statistical tests to know the reliability of the inferred multigene as well as organismal phylogenies. MEGA provides two types of tests: the bootstrap[67] and interior

branch length[51,68] tests. The bootstrap test is the most commonly used test for evaluating the reliability of inferred trees; it is made available in MEGA for the ME, MP, NJ and UPGMA methods. Both the majority-rule consensus[67] as well as condensed[10] trees, which are computed using the frequency of different phylogenetic partitions in the trees inferred from individual bootstrap replicates, are made available.

Bootstrap and interior branch tests for phylogenies are available

In the interior branch test, the robustness of the estimated tree is examined by evaluating each interior branch individually. MEGA computes the ordinary least squares estimate of a given branch length[51] and uses the bootstrap approach to compute its standard error.[68] Then a normal deviate test is conducted to determine if the branch length is significantly greater than 0. The interior branch test is known to be much less conservative than the bootstrap test, which is conservative in most cases.[69,70] However, Nei and Kumar[2] suggest that the interior branch tests are likely to be most suitable for closely related sequences.

Exploring phylogenetic trees

Tree Explorer can now import trees from the Newick format text files

The Tree Explorer (TE) made available in MEGA is an advanced tool in which the same phylogenetic tree can be visualised in a number of ways such as, a topology without branch lengths (eg cladogram), with estimated branch lengths (eg phylogram), and in a linearised[71] fashion. Also, the TE can construct consensus and condensed trees when multiple topologies are available. In version 3, MEGA can now import and display Newick format trees.

The TE provides functionality to annotate trees or reduce large trees by defining groups such that the higher-level relationships can be highlighted (Figure 7). When sequences belong to a group, MEGA automatically prepares the display to mark subtrees in which sequences from the group are included. Users can associate special symbols with different groups and with individual taxa. Even bitmap images, associated with groups or individual taxa, can be included for display.

The TE provides extensive flexibility to change the look of the displayed topology. Users can re-root the tree, add a scale bar, add branch lengths, and change font and line colours in many different combinations. Users also can freeze a TE session exactly as it is by saving it to a file in a 'tree session format'. This facility is useful for retaining convenient settings for future tree

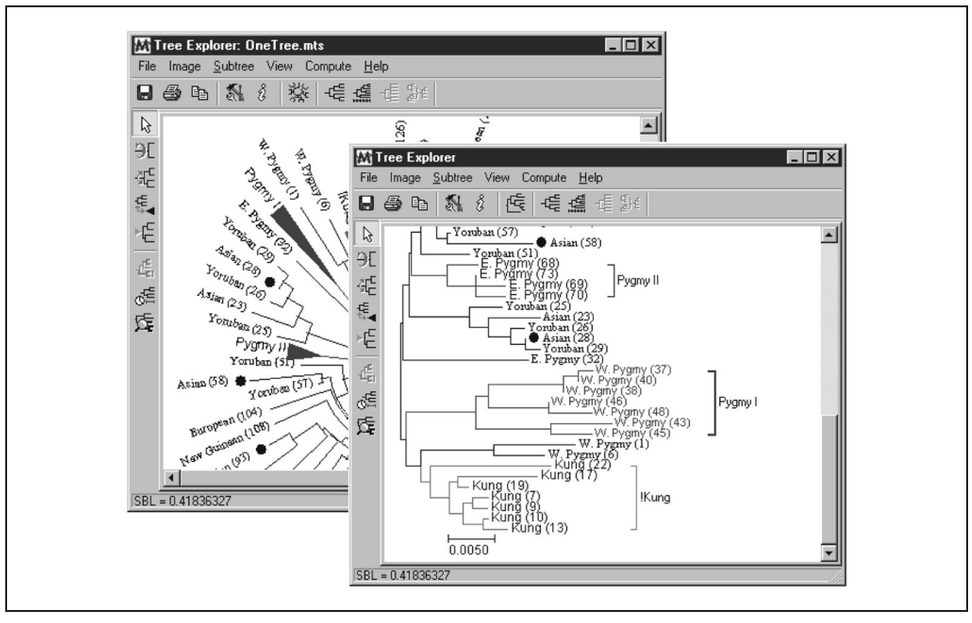


Figure 7: The Tree Explorer in MEGA

Trees can be exported in graphics and text formats

displays. This is different from the procedure of writing trees into text file formats for use in other programs or for graphics editing. For that purpose, MEGA provides options for printing and exporting trees in the Newick-compatible format and as Windows enhanced metafiles. Users simply 'Copy' the displayed tree image and 'Paste' it into Microsoft Word or PowerPoint in the Windows environment.

SOFTWARE PLATFORM AND AVAILABILITY

MEGA is available free at www.megasoftware.net

The current release of MEGA (MEGA3) was developed for use on the Microsoft Windows operating systems. MEGA3 is a native 32-bit multithreaded program with no built-in constraints on the number of sequences or the sequence length. MEGA3 also can be used on various Mac OSs by running it under the Virtual PC emulators of Windows 95 or later Windows releases. It can be obtained from the website[72] free of charge for educational and research purposes. Context-sensitive help is included with the MEGA installation and is available on-line through the MEGA website. To help users become quickly familiar with the program, 'A Walk through MEGA' section is included in the help files.

MEGA runs smoothly on MacOS under Virtual, PC and other Windows emulator environments

Acknowledgments

We thank colleagues, students and volunteers for spending countless hours testing various pre-releases of MEGA. Their comments influenced and improved all facets of user interface and methodology implementation. Also, the efforts by two software development associates, Joel Dudley and David Schwartz, are much appreciated. MEGA software project has been supported by research grants from NIH, NSF and Burroughs-Wellcome Fund to S.K. and NIH and NSF to M.N.

References

1. Collins, F. S., Morgan, M. and Patrinos, A. (2003), 'The Human Genome Project: Lessons from large-scale biology', *Science*, Vol. 300, pp. 286–290.
2. Nei, M. and Kumar, S. (2000), 'Molecular Evolution and Phylogenetics', Oxford University Press, New York.
3. Felsenstein, J. (2003), 'Inferring Phylogeny', Sinauer Associates, Sunderland, MA.
4. Koonin, E. V., Aravind, L. and Kondrashov, A. S. (2000), 'The impact of comparative genomics on our understanding of evolution', *Cell*, Vol. 101, pp. 573–576.
5. Ureta-Vidal, A., Ettwiller, L. and Birney, E. (2003), 'Comparative genomics: Genome-wide analysis in metazoan eukaryotes', *Nat. Rev. Genet.*, Vol. 4, pp. 251–262.
6. Chain, P., Kurtz, S., Ohlebusch, E. and Slezak, T. (2003), 'An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges', *Brief. Bioinform.*, Vol. 4, pp. 105–123.
7. Goodman, N. (2002), 'Biological data becomes computer literate: New advances in bioinformatics', *Curr. Opin. Biotechnol.*, Vol. 13, pp. 68–71.
8. Kumar, S., Tamura, K. and Nei, M. (1994), 'MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers', *Comput. Appl. Biosci.*, Vol. 10, pp. 189–191.
9. Kumar, S., Tamura, K., Jakobsen, I. B. and Nei, M. (2001), 'MEGA2: Molecular Evolutionary Genetics Analysis software', *Bioinformatics*, Vol. 17, pp. 1244–1245.
10. Kumar, S., Tamura, K. and Nei, M. (1993), 'Manual for MEGA: Molecular Evolutionary Genetics Analysis Software', Pennsylvania State University, University Park, PA.
11. NCBI, URL: http://www.ncbi.nih.nlm.gov
12. EMBL, URL: http://www.ebi.ac.uk/embl/index.html
13. Altschul, S. F., Madden, T.L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25, pp. 3389–3402.
14. Schaffer, A. A., Aravind, L., Madden, T. L. *et al.* (2001), 'Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements', *Nucleic Acids Res.*, Vol. 29, pp. 2994–3005.
15. Altschul, S. F. and Koonin, E. V. (1998), 'Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases', *Trends Biochem. Sci.*, Vol. 23, pp. 444–447.
16. Jeanmougin, F., Thompson, J. D., Gouy, M. *et al.* (1998), 'Multiple sequence alignment with Clustal x', *Trends Biochem. Sci.*, Vol. 23, pp. 403–405.
17. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994), 'Clustal-W – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice', *Nucleic Acids Res.*, Vol. 22, pp. 4673–4680.

18. Maddison, D. R., Swofford, D. L. and Maddison, W. P. (1997), 'Nexus: An extensible file format for systematic information', *Syst. Biol.*, Vol. 46, pp. 590–621.
19. Swofford, D. L. (1998), 'PAUP*: Phylogenetic Analysis Using Parsimony (and other methods)', Sinauer Associates, Sunderland, MA.
20. Felsenstein, J. (1993), 'PHYLIP: Phylogeny Inference Package', University of Washington, Seattle, WA.
21. GCG, URL: http://www.accelrys.com/about/gcg.html
22. Sharp, P. M., Tuohy, T. M. F. and Mosurski, K. R. (1986), 'Codon usage in yeast – cluster-analysis clearly differentiates highly and lowly expressed genes', *Nucleic Acids Res.*, Vol. 14, pp. 5125–5143.
23. Tamura, K. and Kumar, S. (2002), 'Evolutionary distance estimation under heterogeneous substitution pattern among lineages', *Mol. Biol. Evol.*, Vol. 19, pp. 1727–1736.
24. Kimura, M. (1980), 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', *J. Mol. Evol.*, Vol. 16, pp. 111–120.
25. Tamura, K. and Nei, M. (1993), 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees', *Mol. Biol. Evol.*, Vol. 10, pp. 512–526.
26. Tamura, K. (1992), 'The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial-DNA', *Mol. Biol. Evol.*, Vol. 9, pp. 814–825.
27. Tajima, F. and Nei, M. (1983), 'Estimation of evolutionary distance between nucleotide sequences', *Mol. Biol. Evol.*, Vol. 1, pp. 269–285.
28. Jukes, T. H. and Cantor, C. R. (1969), 'Evolution of protein molecules' in Munro, R. E., ed, 'Mammalian Protein Metabolism', Academic Press, New York, pp. 21–132.
29. Lake, J. A. (1994), 'Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances', *Proc. Natl Acad. Sci. USA*, Vol. 91, pp. 1455–1459.
30. Lockhart, P. J., Steel, M. A., Hendy, M. D. and Penny, D. (1994), 'Recovering evolutionary trees under a more realistic model of sequence evolution', *Mol. Biol. Evol.*, Vol. 11, pp. 605–612.
31. Yang, Z. (1997), 'PAML: A program package for phylogenetic analysis by maximum likelihood', *Comput. Appl. Biosci.*, Vol. 13, pp. 555–556.
32. Gu, X. and Zhang, J. (1997), 'A simple method for estimating the parameter of substitution rate variation among sites', *Mol. Biol. Evol.*, Vol. 14, pp. 1106–1113.
33. Yang, Z. and Kumar, S. (1996), 'Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites', *Mol. Biol. Evol.*, Vol. 13, pp. 650–659.
34. Nei, M. and Gojobori, T. (1986), 'Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions', *Mol. Biol. Evol.*, Vol. 3, pp. 418–426.
35. Li, W. H., Wu, C. I. and 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', *Mol. Biol. Evol.*, Vol. 2, pp. 150–174.
36. Li, W. H. (1993), 'Unbiased estimation of the rates of synonymous and nonsynonymous substitution', *J. Mol. Evol.*, Vol. 36, pp. 96–99.
37. Zhang, J., Rosenberg, H. F. and Nei, M. (1998), 'Positive Darwinian selection after gene duplication in primate ribonuclease genes', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 3708–3713.
38. Comeron, J. M. (1995), 'A method for estimating the numbers of synonymous and nonsynonymous substitutions per site', *J. Mol. Evol.*, Vol. 41, pp. 1152–1159.
39. Ina, Y. (1995), 'New methods for estimating the numbers of synonymous and nonsynonymous substitutions', *J. Mol. Evol.*, Vol. 40, pp. 190–226.
40. Pamilo, P. and Bianchi, N. O. (1993), 'Evolution of the *Zfx* and *Zfy* genes: Rates and interdependence between the genes', *Mol. Biol. Evol.*, Vol. 10, pp. 271–281.
41. Tajima, F. and Nei, M. (1984), 'Estimation of evolutionary distance between nucleotide-sequences', *Mol. Biol. Evol.*, Vol. 1, pp. 269–285.
42. Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978), 'Composition of proteins', in Dayhoff, M. O., Ed., 'Atlas of protein Sequence and Structure', National Biomedical Research Foundation, Silver Spring, MD, pp. 345–352.
43. Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992), 'The rapid generation of mutation data matrices from protein sequences', *Comput. Appl. Biosci.*, Vol. 8, pp. 275–282.
44. Efron, B. (1982), 'The jackknife, the bootstrap, and other resampling plans', Society for Industrial and Applied Mathematics, Philadelphia, PA.
45. Kumar, S. and Subramanian, S. (2002), 'Mutation rates in mammalian genomes', *Proc. Natl Acad. Sci. USA*, Vol. 99, pp. 803–808.

46. Kumar, S. and Gadagkar, S. R. (2001), 'Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences', *Genetics*, Vol. 158, pp. 1321–1327.

47. Zhang, J., Kumar, S. and Nei, M. (1997), 'Small-sample tests of episodic adaptive evolution: A case study of primate lysozymes', *Mol. Biol. Evol.*, Vol. 14, pp. 1335–1338.

48. Tajima, F. (1989), 'Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics*, Vol. 123, pp. 585–595.

49. Sneath, P. H. A. and Sokal, R. R. (1973), 'Numerical Taxonomy; The Principles and Practice of Numerical Classification', W. H. Freeman, San Francisco, CA.

50. Saitou, N. and Nei, M. (1987), 'The Neighbor-Joining Method – a new method for reconstructing phylogenetic trees', *Mol. Biol. Evol.*, Vol. 4, pp. 406–425.

51. Rzhetsky, A. and Nei, M. (1992), 'A simple method for estimating and testing minimum-evolution trees', *Mol. Biol. Evol.*, Vol. 9, pp. 945–967.

52. Rzhetsky, A. and Nei, M. (1993), 'Theoretical foundation of the minimum-evolution method of phylogenetic inference', *Mol. Biol. Evol.*, Vol. 10, pp. 1073–1095.

53. Tajima, F. (1993), 'Simple methods for testing the molecular evolutionary clock hypothesis', *Genetics*, Vol. 135, pp. 599–607.

54. Gu, X. and Li, W. H. (1992), 'Higher rates of amino acid substitution in rodents than in humans', *Mol. Phylogenet. Evol.*, Vol. 1, pp. 211–214.

55. Muse, S. V. and Weir, B. S. (1992), 'Testing for equality of evolutionary rates', *Genetics*, Vol. 132, pp. 269–276.

56. Kumar, S. (1996), 'A stepwise algorithm for finding minimum evolution trees', *Mol. Biol. Evol.*, Vol. 13, pp. 584–593.

57. Kumar, S. and Gadagkar, S. R. (2000), 'Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies', *J. Mol. Evol.*, Vol. 51, pp. 544–553.

58. Takahashi, K. and 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', *Mol. Biol. Evol.*, Vol. 17, pp. 1251–1258.

59. Robinson, D. F. and Foulds, L. R. (1981), 'Comparison of phylogenetic trees', *Math. Biosci.*, Vol. 53, pp. 131–147.

60. Bryant, D. and Waddell, P. (1998), 'Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees', *Mol. Biol. Evol.*, Vol. 15, pp. 1346–1359.

61. Eck, R. V. and Dayhoff, M. O. (1966), 'Atlas of Protein Sequence and Structure', National Biomedical Research Foundation, Silver Springs, MD.

62. Fitch, W. M. (1971), 'Toward defining course of evolution – minimum change for a specific tree topology', *Syst. Zool.*, Vol. 20, pp. 406-416.

63. Purdom, P. W., Bradford, P. G., Tamura, K. and Kumar, S. (2000), 'Single column discrepancy and dynamic max-mini optimizations for quickly finding the most parsimonious evolutionary trees', *Bioinformatics*, Vol. 16, pp. 140–151.

64. Maddison, W. P. and Maddison, D. R. (1992), 'MacClade: Analysis of phylogeny and character evolution', Sinauer Associates, Sunderland, MA.

65. Yang, Z., Kumar, S. and Nei, M. (1995), 'A new method of inference of ancestral nucleotide and amino acid sequences', *Genetics*, Vol. 141, pp. 1641–1650.

66. Rokas, A., Williams, B. L., King, N. and Carroll, S. B. (2003), 'Genome-scale approaches to resolving incongruence in molecular phylogenies', *Nature*, Vol. 425, pp. 798–804.

67. Felsenstein, J. (1985), 'Confidence limits on phylogenies: An approach using the bootstrap', *Evolution*, Vol. 39, pp. 783–791.

68. Dopazo, J. (1994), 'Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach', *J. Mol. Evol.*, Vol. 38, pp. 300–304.

69. Sitnikova, T. (1996), 'Bootstrap method of interior-branch test for phylogenetic trees', *Mol. Biol. Evol.*, Vol. 13, pp. 605–611.

70. Sitnikova, T., Rzhetsky, A. and Nei, M. (1995), 'Interior-branched and bootstrap tests of phylogenetic trees', *Mol. Biol. Evol.*, Vol. 12, pp. 319–333.

71. Takezaki, N., Rzhetsky, A. and Nei, M. (1995), 'Phylogenetic test of the molecular clock and linearized trees', *Mol. Biol. Evol.*, Vol. 12, pp. 823–833.

72. URL: http://www.megasoftware.net