CABIOS

Vol.10 no.2. 1994
Pages 189-191

MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers

Sudhir Kumar, Koichiro Tamura[1] and Masatoshi Nei[2]

Abstract

A computer program package called MEGA has been developed for estimating evolutionary distances, reconstructing phylogenetic trees and computing basic statistical quantities from molecular data. It is written in C++ and is intended to be used on IBM and IBM-compatible personal computers. In this program, various methods for estimating evolutionary distances from nucleotide and amino acid sequence data, three different methods of phylogenetic inference (UPGMA, neighbor-joining and maximum parsimony) and two statistical tests of topological differences are included. For the maximum parsimony method, new algorithms of branch-and-bound and heuristic searches are implemented. In addition, MEGA computes statistical quantities such as nucleotide and amino acid frequencies, transition/transversion biases, codon frequencies (codon usage tables), and the number of variable sites in specified segments in nucleotide and amino acid sequences. Advanced on-screen sequence data and phylogenetic-tree editors facilitate publication-quality outputs with a wide range of printers. Integrated and interactive designs, on-line context-sensitive helps, and a text-file editor make MEGA easy to use.

Introduction

MEGA (Molecular Evolutionary Genetics Analysis) has been developed to facilitate statistical analyses of molecular evolution by using personal computers. Currently, there are many specialized programs for estimating evolutionary distances between nucleotide or amino acid sequences and reconstructing phylogenetic trees. However, they are usually written for specific methods of analysis and cannot be interconnected easily. MEGA is designed to conduct various statistical analyses in one program and to produce results in publication-quality outputs.

System and requirements

MEGA, v. 1.0, is written and compiled in the Borland C++ and Applications Framework, v. 3.1. This program runs on all IBM personal computers and their compatibles with 640 kbyte RAM in DOS as well as in OS/2 and Microsoft Windows through DOS application capabilities. An event-driven user interface (Figure 1) is implemented with menus and windows by using Borland Turbo Vision, v. 1.0. This interactive interface can be used on most color and monochrome monitors, and it responds to the keyboard as well as to the mouse. Graphics adapters, math coprocessors, and extended or expanded memory are not required to run MEGA, but a hard disk is essential.

MEGA does not limit the amount of data to be analyzed; the size of data is constrained only by the computer memory available from the basic 640 kbyte RAM.

Distance estimation

Most distance estimation methods, which correct for multiple substitutions by taking account of transition/transversion biases, unequal base frequencies (e.g. G + C content variation) and varying substitution rates among sites, are provided in MEGA. These distances are divided into three groups: nucleotide, synonymous–non-synonymous and amino acid.

Nucleotide distances estimate the number of nucleotide substitutions per site between nucleotide sequences. Many methods for estimating the number of nucleotide substitutions per site—such as the proportion of different nucleotides, Jukes and Cantor (1969) distance, Kimura two-parameter (1980) distance, Tajima and Nei (1984) distance, Tamura (1992) distance, and Tamura and Nei (1993) distance—are available in MEGA. For some of these distances, the estimate of the number of transition and transversion substitutions can be also be obtained. The Jin and Nei (1990) and Tamura and Nei (1993) methods of distance estimation are available for the case where the substitution rate varies among sites.

Synonymous–non-synonymous distances are computed by the Nei and Gojobori (1986) method for protein-coding nucleotide sequences by using a genetic code table (nuclear; or mammalian, *Drosophila* or yeast mitochondrial).

Amino acid distances can be estimated for amino acid sequences as well as protein-coding nucleotide sequences. Amino acid distances included in MEGA are the number of amino acid differences, the proportion of different amino acids, the Poisson correction distance (Nei, 1987) and the gamma distance (Nei *et al.*, 1976).

In MEGA, gap sites (alignment gap and missing-information

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA

[1]*Current address: Department of Biology, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji-shi, Tokyo 192-03, Japan*

[2]*To whom reprint requests should be sent*

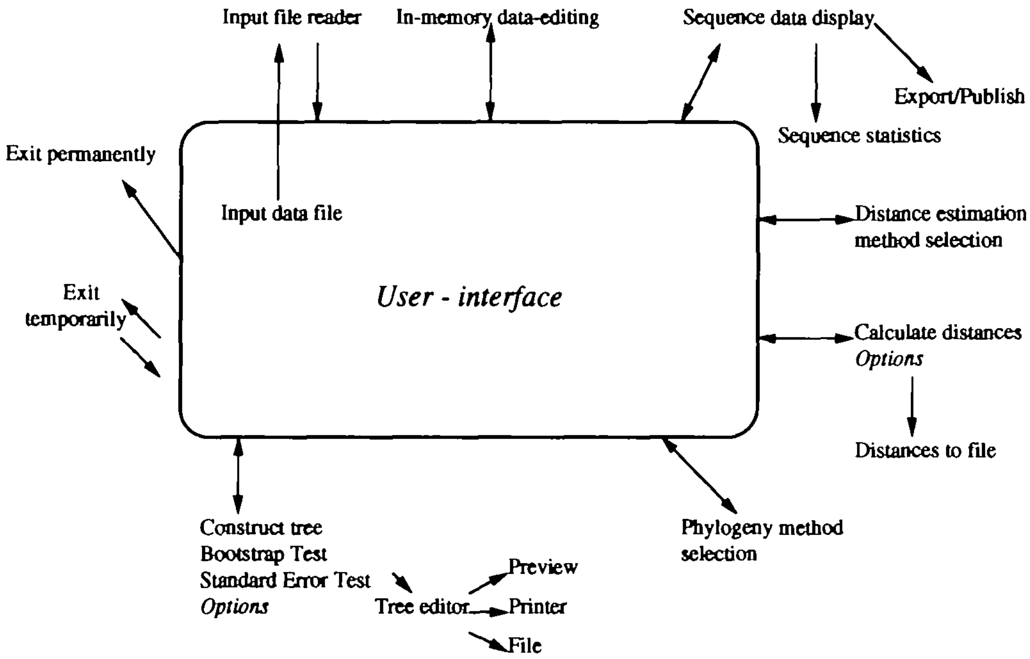

Fig. 1. Event-driven user interface in MEGA

sites) are ignored in distance estimation, but there are two different ways to treat these sites. One way is to delete all of these sites from data analysis, which is called the Complete-Deletion option; the other is to compute a distance for each pair of sequences, ignoring only those gaps that are involved in the comparison. This is called the Pairwise-Deletion option.

Phylogenetic inference

The Unweighted Pair Group Method with Arithmetic Mean (UPGMA; Sneath and Sokal, 1973), the neighbor-joining method (NJ; Saitou and Nei, 1987) and maximum parsimony methods are provided for phylogenetic inference. UPGMA and the neighbor-joining method require a matrix of pairwise distances. By contrast, maximum parsimony methods use information on evolutionary relationships of nucleotides at each site.

UPGMA assumes a constant rate of evolution for all different lineages, whereas the neighbor-joining method requires no such assumption and is known to be quite efficient in recovering true phylogenetic trees (Nei, 1991).

The maximum parsimony method for nucleotide sequences is included in MEGA. It treats all nucleotide changes as unordered and reversible (Eck and Dayhoff, 1966; Fitch, 1971). A new branch-and-bound and a new heuristic search algorithm are implemented. The branch-and-bound algorithm in MEGA reorders the sequences for maximum efficiency. Therefore, the input order of sequences is irrelevant. The heuristic search provided in MEGA is a relaxed version of the branch-and-bound algorithm where the minimum number of substitutions required to explain a part of the topology (partial-tree) is computed at every stage of sequence addition and a local upperbound is computed. A user defined search-factor is then added to the local upperbound, and all partial-trees that are shorter than the new local upperbounds are included in searching for maximum parsimony trees.

MEGA includes two different tests for evaluating the reliability of a tree obtained: the bootstrap test and the standard error test. The bootstrap test (Efron, 1982; Felsenstein, 1985) is provided for both UPGMA and the neighbor-joining methods. In this method, the same number of sites are randomly sampled with replacement from the original sequences, and a phylogenetic tree is constructed from the resampled data. This process is repeated many times, and the reliability of a sequence cluster is evaluated by its relative frequency of the appearance in bootstrap replications. No bootstrap consensus tree is constructed.

The standard error test (Rzhetsky and Nei, 1992, 1993) is for evaluating the reliability of a neighbor-joining tree. In this method, once an NJ tree is obtained, the branch lengths of the tree are re-estimated by using the ordinary least-squares method, and the standard errors of these estimates are computed. To test the statistical significance of a branch with length b and standard error $s(b)$, the t-test is used to obtain the confidence probability (CP).

Basic statistical quantities

MEGA has a feature to compute the nucleotide frequencies and codon frequencies (codon usage tables) for each or all nucleotide sequences used. Nucleotide frequencies can also be computed for the first, second and third nucleotide positions of codons. Numbers of 2- or 4-fold redundant codons in each sequence or all sequences can also be computed. In addition, nucleotide pair frequencies, transition/transversion ratios, the number of variable sites in specified segments, and the number of total variable and informative sites can be computed for sequence data.

Input and output

Both sequence data and distance matrix data can be entered in MEGA as ASCII text files. Nucleotide and amino acid sequences must be written in IUPAC single-letter codes. Only aligned sequences are acceptable and they should be presented in either interleaved (blockwise) or non-interleaved (continuous) format. Alignment gaps and missing data sites are also allowed in input sequences. Distance matrices can be presented either in the lower-left or in the upper-right triangular matrix. From the input data set, any subset may be selected for analysis. Some OTUs as well as some specific sites (domains) may be selected from the original data without modifying the input data file.

Nucleotide sequences can be translated into amino acid sequences, and both can be displayed on the screen. They can be written in files that can be directly used in other programs, such as PAUP (Swofford, 1993) and PHYLIP (Felsenstein, 1993). The variable, informative, and 2- or 4-fold redundant sites can be highlighted in sequences displayed on the screen.

The phylogenetic-tree editor facilitates the relocation of root, the adjustment of tree size, and the flipping and swapping of branches on the screen. The edited tree can be stored in text files, printed as graphic images on a wide range of printers (9-pin dot matrix to PostScript), and previewed on the screen with the EGA, VGA and Hercules graphics adapters.

Availability

An instruction manual (140 pp.) accompanies the MEGA program. It provides instructions for getting started with MEGA and descriptions of various input data and file formats. It also gives a brief description of distance estimation methods, phylogenetic inference algorithms, basic sequence statistics, elements of the interactive user interface, a menu-command reference, and a walk-through tutorial for easy learning and usage.

MEGA v. 1.0 is distributed with a nominal fee to defray the cost of producing the user manual and diskette(s), and the mailing expenses. However, for anyone who is unable to pay the fee for some reason, MEGA will be provided free of charge on receipt of a letter explaining the circumstances. MEGA will not be sent by electronic mail because the accompanying printed manual cannot be included in this case. For technical enquiries or to obtain an order form, please contact the authors.

References

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

Efron,B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. SIAM, Philadelphia.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein,J. (1993) *PHYLIP: Phylogeny Inference Package*, version 3.5. University of Washington, Seattle, WA.

Fitch,W.M. (1971) Towards defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

Jin,L. and Nei,M. (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, **7**, 82–102.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Kimura,M. (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Nei,M. (1991) Relative efficiencies of different tree making methods for molecular data. In Miyamoto,M.M. and Cracraft,J.L. (eds), *Recent Advances in Phylogenetic Studies of DNA Sequences*. Oxford University Press, Oxford, pp. 90–128.

Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.

Nei,M., Chakraborty,R. and Fuerst,P.A. (1976) Infinite allele model with varying mutation rate. *Proc. Nat. Acad. Sci. USA*, **73**, 4164–4168.

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

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

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy*. Freeman, San Francisco.

Swofford,D.L. (1993) *PAUP. Phylogenetic Analysis Using Parsimony (PAUP)*, version 3.1. University of Illinois, Champaign.

Tajima,F. and Nei,M. (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.*, **1**, 269–285.

Tamura,K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition–transversion and G+C-content biases. *Mol. Biol. Evol.*, **9**, 678–687.

Tamura,K. and Nei,M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitchondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**, 512–526.

Received October 17, 1993; accepted December 16, 1993