BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 12 2001
Pages 1244–1245

# MEGA2: molecular evolutionary genetics analysis software

Sudhir Kumar[1,*], Koichiro Tamura[2], Ingrid B. Jakobsen[3] and Masatoshi Nei[3]

[1]Department of Biology, Arizona State University, Tempe, AZ 85287-1501, USA, [2]Department of Biological Sciences, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji-shi, Tokyo 192-0397, Japan and [3]Department of Biology and the Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, PA 16802, USA

Received on February 27, 2001; revised on March 24 and August 1, 2001; accepted on August 4, 2001

## ABSTRACT

**Summary:** We have developed a new software package, Molecular Evolutionary Genetics Analysis version 2 (MEGA2), for exploring and analyzing aligned DNA or protein sequences from an evolutionary perspective. MEGA2 vastly extends the capabilities of MEGA version 1 by: (1) facilitating analyses of large datasets; (2) enabling creation and analyses of groups of sequences; (3) enabling specification of domains and genes; (4) expanding the repertoire of statistical methods for molecular evolutionary studies; and (5) adding new modules for visual representation of input data and output results on the Microsoft Windows platform.

**Availability:** http://www.megasoftware.net.

**Contact:** s.kumar@asu.edu

Genome sequencing and individual laboratory projects are generating vast amounts of DNA sequence data from diverse organisms. The objective of the Molecular Evolutionary Genetics Analysis version 2 (MEGA2) software development is to provide user-friendly tools for exploring, discovering, and analyzing these data from evolutionary perspectives. The first version of MEGA (Kumar *et al.*, 1994) made many methods of evolutionary analysis easily accessible to the scientific community for research and education, but it was developed keeping in mind the limited computational resources available on the average personal computer in the early 1990s. The development of MEGA2 was undertaken to harness the expanded computing power available on the average desktop today and to fulfill the fast growing need for extensive molecular sequence exploration and analysis software. MEGA2 expands the scope of its predecessor from single gene to genome wide analyses.

MEGA2 can be used to conduct statistical analyses of nucleotide and amino acid sequences and pairwise distances. For these data, MEGA2 allows taxa (and sequences) to be arranged into user-defined groups of individuals, species, or sets of orthologous sequences by using a visual interface. This facility makes it possible to estimate sequence diversity within and between groups for use in population genetic, multigene family phylogenetics, and molecular systematic studies. Furthermore, MEGA2 contains features to analyze sequences at the level of genes and domains. A domain is a section of the sequence and a gene is defined as a collection of one or more domains. Domain and genes can be specified using a visual interface with capability to select genes and domains for individual or separate analyses. These group and gene/domain attributes can be stored in the newly designed data format. Also, MEGA2 now includes functionality for importing data from files written in many different formats, including CLUSTAL, Nexus (e.g. PAUP, MacClade), PHYLIP, and PIR.

An advanced user-friendly feature of MEGA2 is the inclusion of sophisticated input data explorers to browse, edit, summarize, and export the input data subsets. The *Sequence Data Explorer* module displays the aligned sequences and has facilities to change the sequence order and translate/un-translate the protein-coding DNA regions using the chosen genetic code table. (The user can also specify new genetic code tables in the interface and compute statistical attributes of codons.) In this explorer, different types of sites (e.g. conserved, polymorphic, 4-fold degenerate) can be highlighted and various summary statistics (e.g. nucleotide composition, relative synonymous codon usage) computed for all or only highlighted sites. *Distance Data Explorer* displays the input pairwise distances and can be used to calculate overall and group averages based on individual distances.

MEGA2 can compute evolutionary distances based on the observed differences in amino acid and nucleotide sequences (reviewed in Nei and Kumar, 2000; Swofford *et*

[*]To whom correspondence should be addressed.

*al.*, 1996; Page and Holmes, 1998). In addition to the computation of pairwise distances, we have now included the estimation of average pairwise distances within groups, among groups, and for all sequences. MEGA2 uses the bootstrap approach for estimation of variances of all statistical quantities computed and for testing biological hypotheses. For pairwise distances, analytical methods are also included for estimating standard errors.

As compared to the first version, MEGA2 contains a more extensive collection of methods to estimate the number of synonymous substitutions per synonymous site ($d_S$) and the number of nonsynonymous substitutions per nonsynonymous site ($d_N$). We have added the modified Nei–Gojobori method as well as the the Li–Wu–Lou method and its modification (PBL: Pamilo-Bianchi and Li method; reviewed in Nei and Kumar, 2000). The PBL method is further modified in MEGA2 for use with any genetic code table because substitutions at the 2-fold degenerate positions in some codons (e.g. arganine) are not easily classified as synonymous or nonsynonymous (Comeron, 1995; Nei and Kumar, 2000). W.-H. Li and J. Comeron (personal communication) have developed solutions for such problems, specifically for the 'Standard' genetic code table in their computer programs. Since these classifications do not apply to all genetic code tables, we have developed a consistent strategy in which we subdivide the 2-fold degenerate sites into simple and complex categories. The simple 2-fold degenerate sites are those at which the only transitional change results in a synonymous substitution and the two transversional changes result in nonsynonymous substitutions. All other 2-fold degenerate sites belong to the complex 2-fold site category. We now score the number of transitional and transversional substitutions for these five site categories and compute the proportions of transitional ($P_i$) and transversional ($Q_i$) differences for the 0, 2, and 4-fold degeneracy classes as follows:

$$P_0 = \frac{s_0 + s_{2N}}{L_0 + L_{2C}},\ Q_0 = \frac{v_0}{L_0},\ P_2 = \frac{s_0 + s_{2S}}{L_{2S} + L_{2C}},$$
$$Q_2 = \frac{v_2 + v_{2N}}{L_{2S} + L_{2C}},\ P_4 = \frac{s_4}{L_4},\ Q_4 = \frac{v_4 + v_{2S}}{L_4 + L_{2C}},$$

where $L_0$, $L_2$, $L_{2S}$, $L_{2C}$, and $L_4$ are the numbers of 0-fold, 2-fold, simple 2-fold, complex 2-fold, and 4-fold degenerate sites; and $s$ and $v$ refer to the transitional and transversional differences observed in the respective categories.

With these quantities, the number of transitional substitutions per site ($A_i$) and the tranversional substitutions per site ($B_i$) are computed using equation (4.9) in Nei and Kumar (2000). Given $A_i$'s and $B_i$'s, and the equation $L_2 = L_{2C} + L_{2S}$, we can apply equation (4.11) for the PBL method in Nei and Kumar (2000) for estimating the evolutionary distance.

We have expanded the choice of tree-building methods, keeping in mind that the future data sets will likely consist of a large number of sequences from multiple genes. UPGMA, Neighbor-Joining, and Minimum Evolution, and Maximum Parsimony methods are included (Nei and Kumar, 2000; Swofford *et al.*, 1996). The bootstrap test and the interior branch length tests are included to examine the reliability of the inferred phylogenies (reviewed in Nei and Kumar, 2000). MEGA2 also has the capability to conduct relative rate tests for examining the molecular clock hypothesis, for studying positive Darwinian selection by examining if $d_N$ is greater than $d_S$, and conducting test of neutrality (see website http://www.megasoftware.net for details).

In addition to an ASCII-text file editor, MEGA2 includes a *Distance Matrix Explorer* to display pairwise distances along with their standard error estimates and a *Tree Explorer* with facilities for multiple representations of the tree and the construction of consensus trees. In the *Tree Explorer*, users can compute linearized trees (Takezaki *et al.*, 1995) and provide molecular clock calibrations to estimate divergence times for all branching points in the tree displayed. The tree display can be modified by compressing subtrees such that the higher level relationships in trees can be shown. Printing and exporting of trees in the Newick-compatible format and as windows-metafiles are available to ensure compatibility with other evolutionary analysis and graphics editing programs.

## ACKNOWLEDGEMENTS

We thank colleagues, students, and volunteers for testing the pre-release versions and to Sudhindra Gadagkar for comments on an earlier draft. This project is supported by research grants from NIH, NSF and Burroughs–Wellcome Fund to S.K. and NIH, NSF and NASA to M.N.

## REFERENCES

Comeron,J.M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.*, **41**, 1152–1159.

Kumar,S., Tamura,K. and Nei,M. (1994) MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput. Appl. Biosci.*, **10**, 189–191.

Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Page,R.D.M. and Holmes,E.C. (1998) *Molecular Evolution: a Phylogenetic Approach*. Blackwell, Oxford.

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) Phylogenetic inference. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics*, 2nd edn, Sinauer, Sunderland, MA, pp. 407–514.

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