MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0

Koichiro Tamura,† Joel Dudley,* Masatoshi Nei,‡ and Sudhir Kumar§**

*Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University; †Department of Biological Sciences, Tokyo Metropolitan University, Tokyo, Japan; ‡Department of Biology and the Institute of Molecular Evolutionary Genetics, The Pennsylvania State University; and §School of Life Sciences, Arizona State University

We announce the release of the fourth version of MEGA software, which expands on the existing facilities for editing DNA sequence data from autosequencers, mining Web-databases, performing automatic and manual sequence alignment, analyzing sequence alignments to estimate evolutionary distances, inferring phylogenetic trees, and testing evolutionary hypotheses. Version 4 includes a unique facility to generate captions, written in figure legend format, in order to provide natural language descriptions of the models and methods used in the analyses. This facility aims to promote a better understanding of the underlying assumptions used in analyses, and of the results generated. Another new feature is the Maximum Composite Likelihood (MCL) method for estimating evolutionary distances between all pairs of sequences simultaneously, with and without incorporating rate variation among sites and substitution pattern heterogeneities among lineages. This MCL method also can be used to estimate transition/transversion bias and nucleotide substitution pattern without knowledge of the phylogenetic tree. This new version is a native 32-bit Windows application with multi-threading and multi-user supports, and it is also available to run in a Linux desktop environment (via the Wine compatibility layer) and on Intel-based Macintosh computers under the Parallels program. The current version of MEGA is available free of charge at http://www.megasoftware.net.

Since the early 1990s, MEGA software functionality has evolved to include the creation and exploration of sequence alignments, the estimation of sequence divergence, the reconstruction and visualization of phylogenetic trees, and the testing of molecular evolutionary hypotheses. The three versions of MEGA have been released, and they integrate Web-based sequence data acquisition and alignment capabilities (fig. 1) with the evolutionary analyses (fig. 2), making it much easier to conduct comparative analyses in a single computing environment (Kumar, Tamura, and Nei 2004). Over time, MEGA has come to enhance the classroom learning experience as its use by researchers, educators, and students in diverse disciplines has expanded (Kumar and Dudley 2007). The fourth version (MEGA4) contains three distinct newly developed functionalities, which are outlined below.

First, we have developed a *Caption Expert* software module that generates descriptions for every result obtained by MEGA4. This description informs the user of all of the options used in the analysis, including the data subset used (e.g., codon positions included), the chosen option for the handling of sites with gaps or missing data, the evolutionary model of substitution (e.g., DNA substitution pattern, uniformity of evolutionary rates among sites, and homogeneity assumption among lineages), and the methods applied for estimating pairwise distances and for inferring and testing phylogeny. The caption also includes specific citations for any method, algorithm, and software used in the given analysis. Two examples of descriptions generated by the *Caption Expert* are shown in figure 3.

The availability of these descriptions is intended to promote a better understanding of the underlying assumptions used in analyses, and of the results produced. This is needed because MEGA's intuitive graphical interface makes it easy for both new and expert users to conduct a variety of computational and statistical analyses. However, some users may not immediately realize the underlying assumptions and data-handling options involved in each analysis. Even expert molecular and population geneticists may not be able to discern all of the assumptions implied. In general, we expect a written description of methods and results to be useful for students and researchers when preparing tables and figures for presentation and publication.

Second, we have now added a Maximum Composite Likelihood (MCL) method for estimating evolutionary distances (d_{ij}) between DNA sequences, which MEGA users frequently employ for inferring phylogenetic trees, divergence times, and average sequence divergences between and within groups of sequences. In this approach, the Composite Log Likelihood (*CL*) obtained as the sum of log likelihood for all sequence pairs in an alignment is maximized by fitting the common parameters for nucleotide substitution pattern (θ) to every sequence pair (i,j): $CL = \sum_{i,j} \ln l(\theta, d_{ij})$ (Tamura, Nei, and Kumar 2004). This approach was previously referred to as the ''Simultaneous Estimation'' (SE) method, because all d_{ij}'s are simultaneously estimated (Tamura, Nei, and Kumar 2004). The MCL approach differs from current approaches for evolutionary distance estimation, wherein each distance is estimated independently of others, either by analytical formulas or by likelihood methods (independent estimation [IE] approach).

The MCL method has many advantages over the IE approach. To begin with, the IE method for estimating evolutionary distance for each pair of sequences will often cause rather large errors unless very long sequences are used. The use of the MCL method reduces these errors considerably, as a single set of parameters estimated from all-sequence pairs is applied to each distance estimation. When distances are estimated with lower errors, distance-based methods for inferring phylogenies are expected to be more accurate. This is indeed the case for the

Key words: selection, genomics, phylogenetics, software, cross-platform.

E-mail: s.kumar@asu.edu

Mol. Biol. Evol. 24(8):1596–1599. 2007
doi:10.1093/molbev/msm092
Advance Access publication May 7, 2007

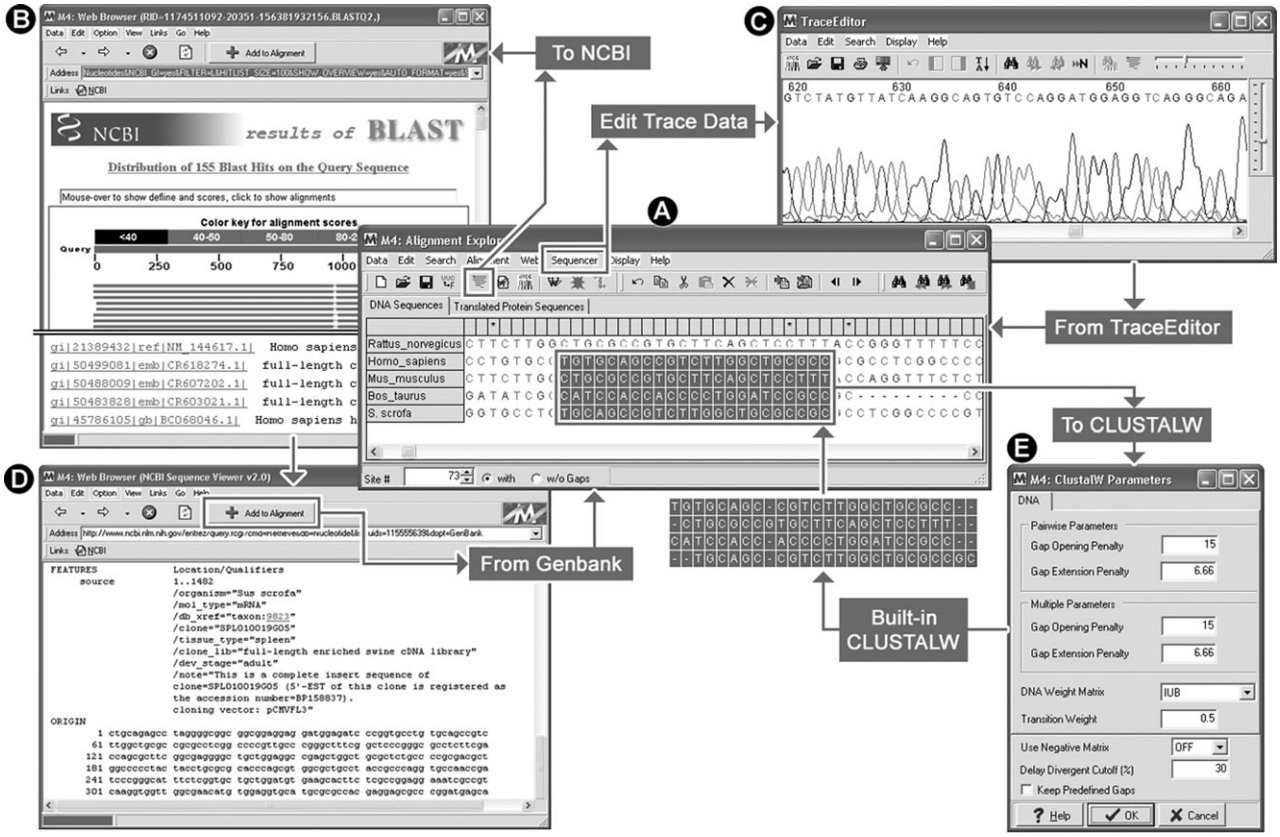

FIG. 1.—Sequence alignment editor and Web-data mining features in MEGA4. In the Alignment Explorer (*A*), the integrated web browser (*B*) permits downloading sequences from online databases directly into the current alignment, without the need for manual cutting-and-pasting and reformatting. The DNA sequences can be translated to the corresponding protein sequences by a single mouse click (*D*), and the protein sequences can be aligned by ClustalW (*E*) (Thompson, Higgins, and Gibson 1994) and adjusted manually by eye. Returning to the nucleotide view automatically aligns the nucleotide sequences according to the protein alignments, and DNA and protein sequence alignments can be exported in a variety of formats for use with other programs. Alignment Editor also contains facilities for editing and importing of trace data files output from DNA sequencers (*C*).

Neighbor-Joining method (Saitou and Nei 1987), as the use of the MCL distances leads to a much higher accuracy (Tamura, Nei, and Kumar 2004). Even when the topologies estimated are the same, the use of the MCL distances often gives higher bootstrap values for the estimated phylogenetic tree compared to the use of IE distances, which is evident from the example given in figure 4 A (MCL: bold, IE: italics).

In addition, the IE distances are not always estimable when pairwise distances are calculated between very distantly related sequences, because the arguments of logarithms in the analytical formulas may become negative by chance. The probability of occurrence of such inapplicable cases increases as the number of sequences in the data increases, the evolutionary distances become larger, and the substitution pattern becomes more complex (Tamura, Nei, and Kumar 2004). The use of the MCL method eliminates this problem effectively and allows for the use of sophisticated models in inferring phylogenies from an increasingly larger number of diverse sequences.

MEGA4 implements the MCL approach for estimating distances between sequence pairs, average distances between and within groups, and average pairs overall with their variances estimated by a bootstrap approach. Our implementation of the MCL method allows for the consideration of substitution rate variation from site to site, using an approximation of the gamma distribution of evolutionary rates, and the incorporation of heterogeneity of base composition in different species/sequences. The user also has the flexibility to estimate the numbers of transition and transversion type substitutions per site separately. Naturally, the MCL distances can be used for inferring phylogenies by the distance-based methods, along with the bootstrap tests of phylogenies.

MEGA4 implements the MCL approach under the Tamura-Nei (1993) substitution model, in which the rates of two types of transitional substitutions (between purines [a_1] and between pyrimidines [a_2]) and the rate of transversional substitutions (b) are considered separately by taking into account the unequal frequencies of four nucleotides (base composition bias). The MCL estimates of the transition/transversion rate ratio have been found to be close to the true values in previous simulation experiments (Tamura, Nei, and Kumar 2004). We have employed this feature to provide users with a facility to compute the relative rates of substitutions between nucleotides based on the MCL estimates of a_1, a_2, b, and on the observed frequencies of the four nucleotides under the Tamura-Nei (1993) model (fig. 3C). For ease of comparison, we have expressed these substitution rates as relative frequencies of substitutions

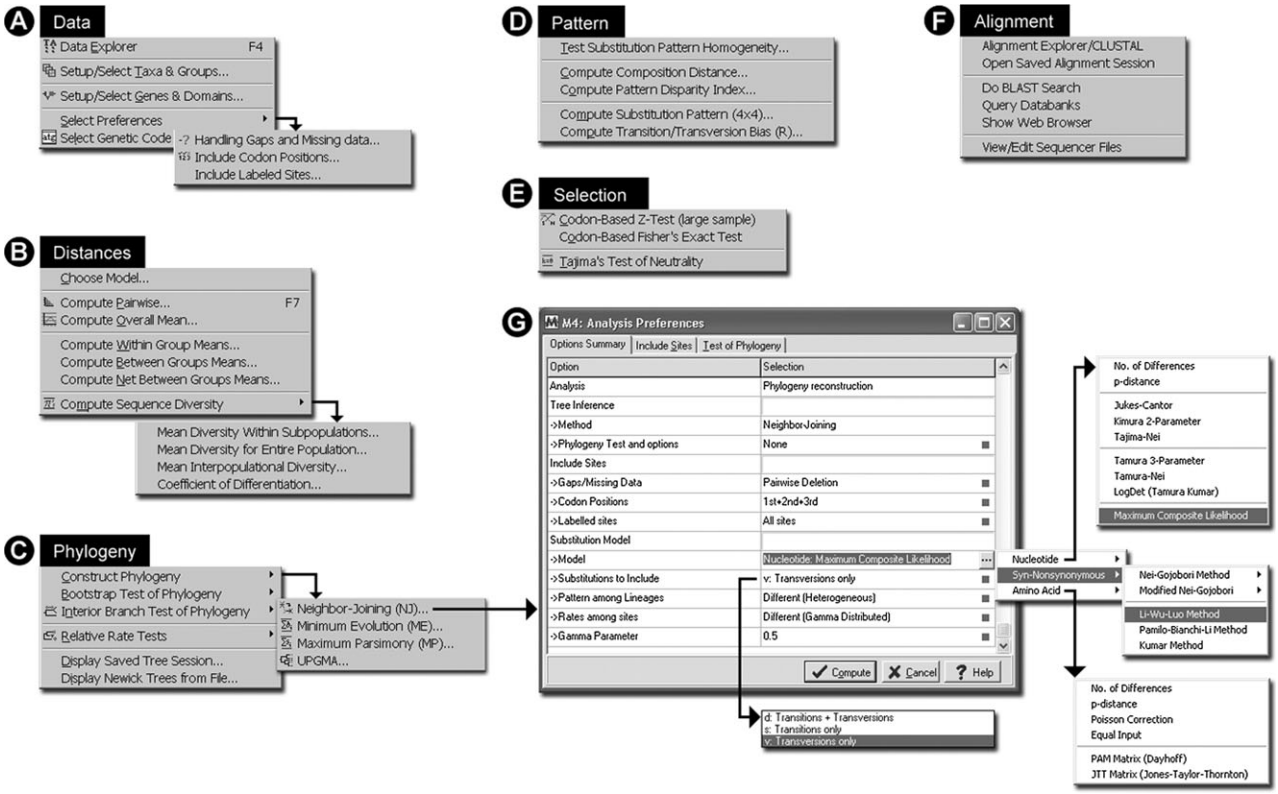

FIG. 2.—A collection of menus that provide access to many different data analysis options in MEGA4, including exploration of input data set (*A*), estimation of evolutionary distances (*B*), inferring and testing phylogenetic trees (*C*), tests of homogeneity of substitution patterns and its estimation (*D*), tests of selection (*E*), alignment of DNA and protein sequences (*F*), and the dialog box that provides users with options to select model of substitution and data sub-setting options (*G*).

between nucleotides such that the sum of all frequencies is 100 (see also Gojobori, Li, and Graur 1982).

Third, we have now programmed MEGA4 to run on some versions of Linux through the Wine software compatibility layer (www.winehq.org). The first advancement alleviates the problem of performance degradation (and the need to purchase Windows emulation software) when using MEGA on Linux. Wine is neither a hardware nor a software emulator, but an open source tool that allows for the native execution of Windows applications on Linux. Our tests of MEGA4 running on Linux show the display, stability, and performance to be highly satisfactory and comparable to the native Windows system (fig. 4B). Furthermore, investigators now report MEGA4 running on

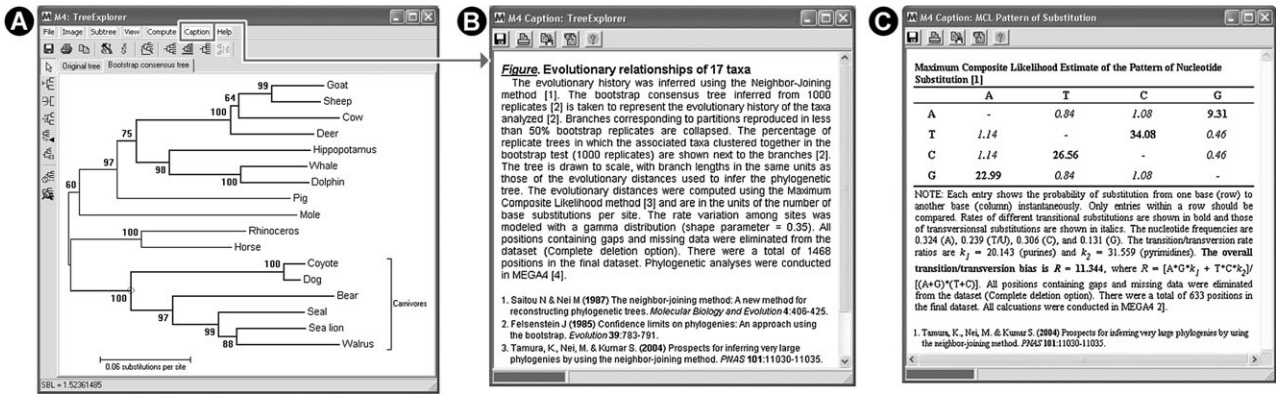

FIG. 3.—The Tree-Explorer displaying a Neighbor-Joining tree of mitochondrial 16S rRNA sequences (*A*), and the description generated by the *Caption Expert* (*B*). Estimates of the relative probabilities of nucleotide substitutions for 70 control-region sequences of human mitochondrial DNA sequences are shown in (*C*). The gamma shape parameter (a= 0.35) was estimated using the Yang and Kumar (1996) method, and the rest of the analysis details are given in (*B*). It is worth noting that the Tree Explorer shown in (*A*) includes a high-resolution tree drawing facility that includes displaying trees in a variety of formats, with options to display/hide branch lengths as well as clade confidence labels, and re-rooting and rearranging trees, among other functionalities. MEGA4 can export the drawings to graphics programs, and can export trees in Newick format for use by other programs. Furthermore, MEGA can import and draw trees from Newick format files that have been estimated by other programs (see fig. 2*C*).

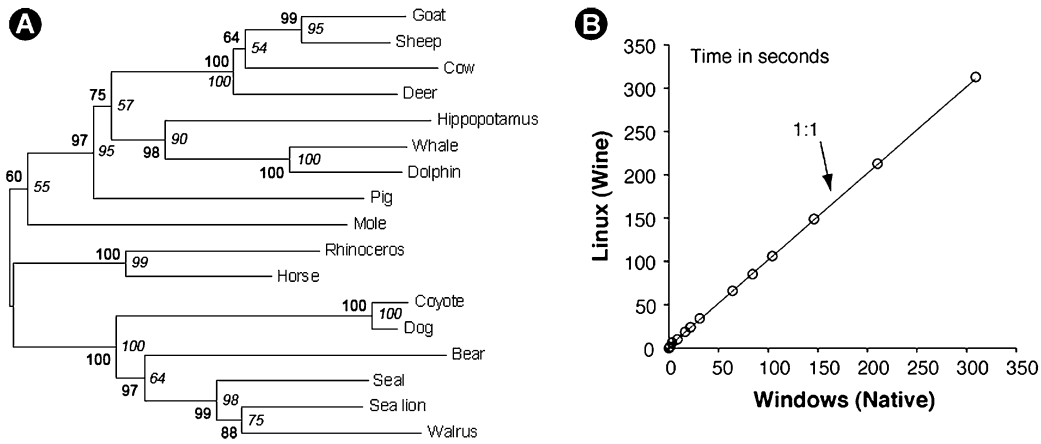

FIG. 4.—(*A*) Bootstrap support for the branching order of 16 Laurasiatheria species reconstructed with MCL approach (bold) and without MCL approach (italics) under the Tamura-Nei (1993) model (see figure 3B for rest of the analysis details). The 16S rRNA sequences used were downloaded from GenBank and were aligned in MEGA4 using CLUSTALW (accession numbers: AJ428578, NC004029, X72004, AF303109, NC008093, DQ480502, X97336, X79547, DQ534707, AJ554051, AJ554061, NC000889, NC007704, AB074968, NC005044, and NC001941). (*B*) Comparison of MEGA4 performance benchmarks on Windows and Linux (with Wine application compatibility layer). Identical hardware configuration was used, and example data sets included in the MEGA4 installation were employed. The results show that computations executed under Wine are penalized by about 2 s, which is attributable to the need for Wine's initialization.

Intel-based Macintosh computers under the Parallels program as well as it does on Windows-native personal computers (see Hall 2007). The Parallels program is a native solution for Macintosh computers that permits them to simultaneously run Windows and Macintosh software.

We have also built support for a multi-user environment, which will allow each user of the same computer to keep his/her customized settings, including file locations, window sizes, choice of genetic code table, and previously used analysis options. This feature will facilitate educational and laboratory usage, where a single computer is often shared by multiple users.

In conclusion, MEGA4 now contains a wide array of functionalities for the molecular evolutionary analysis of data (http://www.megasoftware.net/features.html). It is useful to note that while we are continuously adding new methods and functions to MEGA, we do not intend to make it a catalog of all evolutionary analysis methods available. Rather, it is anticipated to become a workbench for the exploration of sequence data from evolutionary perspectives.

Acknowledgments

We thank the colleagues, students, and volunteers who spent countless hours testing the early release versions of MEGA; almost all facets of MEGA's design and implementation benefited from their comments. We thank Ms. Linwei Wu for assistance with MEGA Web site and for handling bugs, and Ms. Kristi Garboushian for editorial support. We thank the two reviewers for suggesting many useful text additions, which have been included in the figure 1 legend and in the text. We also thank Drs. Masafumi Nozawa and Barry Hall for comments on an earlier version of this manuscript. The MEGA software project is supported by research grants from National Institutes of Health (S.K. and M.N.) and from Japan Society for Promotion of Sciences (K.T.).

Literature Cited

Gojobori T, Li WH, Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol. 18:360–369.

Hall BG. Phylogenetic trees made easy: A how-to manual. Sunderland (MA): Sinauer Associates.

Kumar S, Dudley J. 2007. Bioinformatics for biologists in the genomics era. Bioinformatics. 10.1093/bioinformatics/btm239.

Kumar S, Tamura K, Nei M. 2004. MEGA3: an integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5:150–163.

Saitou N, Nei M. 1987. The Neighbor-Joining method—a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol Biol Evol. 10: 512–526.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the Neighbor-Joining method. Proc Natl Acad Sci USA. 101:11030–11035.

Thompson JD, Higgins DG, Gibson TJ. 1994. ClustalW—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.

Yang Z, Kumar S. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol Biol Evol. 13:650–659.

William Martin, Associate Editor

Accepted May 2, 2007