BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 4. 299–306
Advance Access publication April 16, 2008
doi:10.1093/bib/bbn017

MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences

Sudhir Kumar, Masatoshi Nei, Joel Dudley and Koichiro Tamura

Submitted: 18th January 2008; Received (in revised form): 13th March 2008

Abstract

The Molecular Evolutionary Genetics Analysis (MEGA) software is a desktop application designed for comparative analysis of homologous gene sequences either from multigene families or from different species with a special emphasis on inferring evolutionary relationships and patterns of DNA and protein evolution. In addition to the tools for statistical analysis of data, MEGA provides many convenient facilities for the assembly of sequence data sets from files or web-based repositories, and it includes tools for visual presentation of the results obtained in the form of interactive phylogenetic trees and evolutionary distance matrices. Here we discuss the motivation, design principles and priorities that have shaped the development of MEGA. We also discuss how MEGA might evolve in the future to assist researchers in their growing need to analyze large data set using new computational methods.

Keywords: *phylogenetics; genome; evolution; software*

Biologist-friendly software tools are crucial in this age of burgeoning sequence databases. These tools not only make it possible to use computational and statistical methods but also allow scientists to select methods and algorithms best suited to understand the function, evolution and adaptation of genes and species. The Molecular Evolutionary Genetics Analysis (MEGA) software aims to serve both of these purposes in inferring evolutionary relationships of homologous sequences, exploring basic statistical properties of genes and estimating neutral and selective evolutionary divergence among sequences (Figure 1).

The MEGA software project grew out of our own need for employing statistical methods in the phylogenetic analysis of DNA and protein sequences in the early 1990s. At this time, most computer programs available did not allow us to explore the primary data visually and lacked a user-friendly interface. There were two primary general-purpose computer programs for inferring phylogenetic trees. One was PAUP for constructing parsimony trees, and the other was PHYLIP for inferring phylogenetic trees using various character and statistical methods such as maximum likelihood, parsimony and distance methods [1, 2]. These programs were (and continue to be) very useful, but the former lacked statistical methods at that time and the latter did not provide a point-and-click user interface [3–6].

In order to make statistical methods available for phylogenetic analysis in a user-friendly manner,

Corresponding author. Sudhir Kumar, Biodesign Institute, A240, Arizona State University, Tempe, AZ 85287-5301, USA. E-mail: s.kumar@asu.edu

Sudhir Kumar is conducting large-scale analysis of genome sequences and spatial patterns of gene expression, and developing statistical methods and bioinformatics tools. He is the Director of the Center for Evolutionary Functional Genomics at Arizona State University, AZ, USA.

Masatoshi Nei is one of the founders of molecular evolutionary genetics and pursues statistical analysis of molecular and genome evolution. He is the Director of the Institute of Molecular Evolutionary Genetics at Pennsylvania State University, PA, USA.

Joel Dudley is interested in translational bioinformatics, genomic medicine and the application of molecular evolution in translational genomic research. He is a bioinformatics specialist at the Stanford Center for Biomedical Informatics Research at Stanford University, USA.

Koichiro Tamura's research interests are in the area of molecular evolution with emphasis on the pattern of gene and genome evolution, and on development of statistical methods and computer programs. He is an associate professor at the Tokyo Metropolitan University, Japan.

Sequence Mining and Alignment
Web data mining and BLASTing
Importing/editing of Trace files from Sequencers
Integrated codon/protein Alignment Explorer
Manual and native-CLUSTAL alignments
Data Exploretion and Statistics
Evolutionary Explorer for active sequence data
One-click translation of coding sequences
Context-sensitive highlighting of variable and other types of sites
Export data to other formats (e.g., Nexus, Phylip)
Estimation of nucleotide and amino acid compositions
Estimation of relative synonymous codon usage (RSCU)
* Export statistics in Excel, CSV, and text formats
Data Subsets Supported
Selection of groups, genes, and domains
Selection of codon positions and individual sites
Automatic, context-sensitive codon translation
Handling of gaps (pairwise and complete-deletion)
Exploration of Evolutionary Distances
Display of pariwise distances and errors simultaneously
Sorting of rows by name, groups, and drag-and-drop
Computing within and between group distances
* Summarize sequences involved in invalid distances
* Export distances in Excel, CSV, and text formats
Exploration of Phylogenies
Build consensus and condensed trees
Linearized tree and calibrate clock to estimate divergence times
Display most parsimonious ancestral states
Rectangular, radiation, slanted, and circular tree displays
Flip and swap subtrees, and change position of the root
Customize size, fonts, and colors of text and lines
Collapse clusters and associate images with nodes
Scale-bar display for substitutions and time
Import/export trees in newick, TIFF and metafile formats
Descripton of Results and Assumptions
Figure legend style descriptions for all results
Context-dependent format for display
Generates citations/references for all methods employed
Text File Handling
Edits very large files, permits rectangular block edits/cuts
Format conversion from popular formats (e.g., Nexus)
Unlimited Undo, and utilities for reverse complement etc.
Supporting Operating Systems
Windows Multi-user multi-threaded 32-bit application
Virtual PC and emulators (Linux, Mac, SunOS)
Native Linux via WINE application compatibility layer
HTML context-sensitive help on all platforms

Data types
DNA and protein sequence alignments
Pairwise distance matrices
Unaligned sequences for alignment
Use predefined or create new genetic code tables
Distances
Separate distances by site degeneracy and codon positions
Separation of distances into transitions & transversions
Computation of synonymous and nonsynonymous components
Estimate differences and ratios between distance components
Bootstrap and analytical variances of estimates
Estimate within, between, and net sequence diversity
Maximum Composite Likelihood (MCL) distance
Models
DNA, codon, and protein models
No. of differences and p-distance for all types of differences
Jukes-Cantor and Poisson distances
Tajima-Nei, Tamura 3-parameter, and Tamura-Nei distances
Original and modified Nei-Gojobori codon distances
Original and modified Li-Wu-Lou codon distances
Equal input, Dayhoff and JTT amino acid distances
Incorporate base composition differences between lineages
Incorprate rate variation among sites (Gamma distances)
* Model selection using Likelihood
Pattern of substitution
Homogeneity test for substitution patterns (Disparity Index Test)
Estimation of global 4x4 pattern of nucleotide substitution (MCL)
Estimation of global transition/transversion bias (MCL)
Counts of nucleotide pair frequences between sequences
Tree Inference
NJ and UPGMA with random tie-breaking
ME with Close-Neighbor Interchange (CNI) heuristic search
MP with Max-mini Branch-and-bound search
Mini-mini with search factor and CNI heuristic searches for MP
Fast ordinary least squares branch length estimation
Average over pathway method for parsimony branch lengths
Bootstrap tests of the inferred phylogeny (all methods)
Interior-branch tests for NJ and ME under least-squares
* Maximum Likelihood for large trees (DNA and protein)
Tests of Selection
Codon-based tests of selection (Large and small sample)
Test for sequence pairs, within groups, or average over all
Tajima's test based on segregating sites and sequence diversity
Relative Rate tests & Molecular clocks
Tajima's test for DNA and protein data
Analyze transitions, transversions, or codon positions
Linearized tree for estimating divergtence times (in Tree Explorer)

Figure 1: A list of primary features in MEGA. An asterix marks features that are expected to be included in versions 4.1, 4.2, and 5.

we produced a program with character-based mouse/menu interface for use on DOS systems in 1993 (Figure 2A) [7, 8]. It offered many methods for estimating evolutionary distances from nucleotide and amino acid sequence data, three different methods of phylogenetic inference and statistical tests of inferred phylogenies. In addition, facilities were provided to compute basic statistical properties of DNA and protein sequences, and tools were included for the visual exploration of input sequence data and inferred phylogenies.

Availability of easy-to-use MEGA 1 initiated many biologists to begin exploring the utility of distance-based methods in molecular evolutionary genetic analysis; we received hundreds of request for MEGA 1 in the first year of its release. A case in point is the application of the Neighbor-Joining (NJ) method in phylogenetic inference [9]. By using the NJ method to infer phylogenetic tree in MEGA, researchers often discovered that the NJ method generated a tree quickly for data sets containing many sequences and that differences between NJ

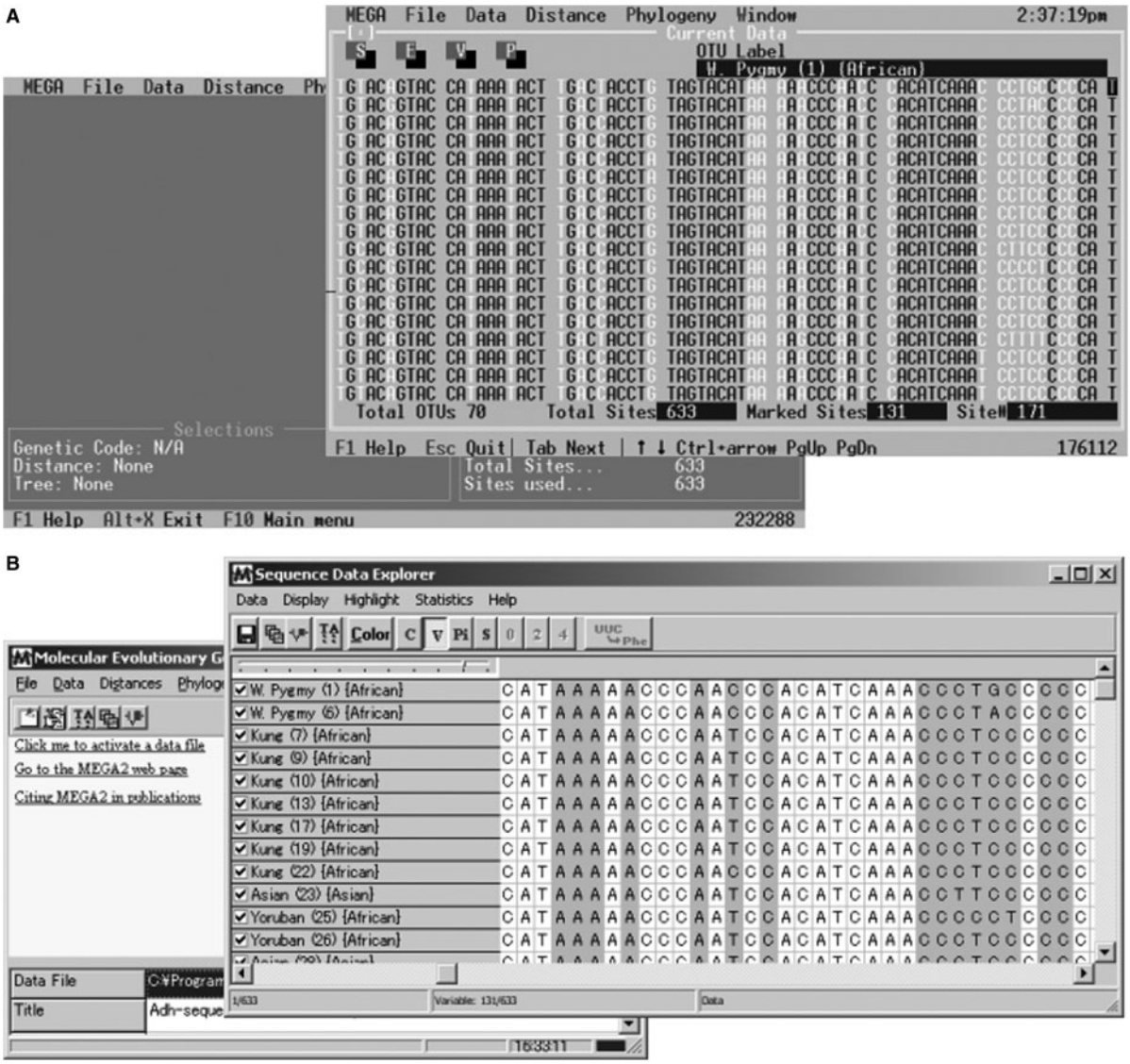

Figure 2: Screenshots of the first version of MEGA containg a character-based point-and-click interface for use on DOS (**A**). MEGA I evolved into a Windows program (MEGA 2) with an extensive GUI (**B**). The main MEGA windows are overlaid with the Input Sequence Data Explorer, in which columns containing at least two different types of nucleotides (variable sites) are highlighted.

trees and those produced by other time-consuming methods were localized to parts of the trees that were usually statistically weakly supported. These first-hand experiences spurred a more widespread appreciation of statistical methods for phylogenetic analysis, which was clearly reflected in the fast-growing citations of the NJ method (over 16 000 citations to date) and software packages, such as MEGA 1, providing access to statistical methods in phylogenetics (Figure 1 in [10]).

In 1995, the release of Microsoft Windows 95 led to the first stable graphical user interface (GUI) for Intel-CPU-based personal computers, which were called IBM compatibles. We decided to evolve MEGA's character-based user-interface from a 640×480 pixel size screen to a GUI. Such GUIs enhanced the user experience by providing more screen space for displaying larger amounts of data and results, allowed for a more sophisticated and intuitive display of information, and enabled more intuitive and efficient data manipulations. Furthermore, the expanded memory space (from 640 kilobytes to up to 3 Gigabytes) in Windows made it possible to analyze both a larger number of sequences (>10 000) and sequences of increased length (millions of base pairs and amino acids).

MEGA 2, released in 2001, further improved the capabilities of MEGA 1 by facilitating analyses of larger data sets, enabling analyses of grouped sequences, specification of multiple domains and genes, and expansion of the repertoire of statistical methods for molecular evolutionary studies (Figure 2B) [11]. It was made available over the Internet (http://www.megasoftware.net) and was downloaded by over 35 000 users (unique e-mail addresses) between 2001–04. Over time, the use of the Windows version has been replacing the DOS version.

In 2004, the release of MEGA 3 addressed the long-standing need for making the sequence data retrieval and alignment less frustrating and less error-prone [12]. Now researchers could edit DNA sequence data from auto-sequencers, retrieve data from web databases, and perform automatic and manual sequence alignments in MEGA. This integrated sequence data acquisition and evolutionary analysis tool was downloaded by 50 000 unique e-mail addresses from 2004 to 2007. The next release in 2007 (MEGA 4) offered facilities to generate detailed captions for many different types of analyses and results (discussed in detail later), a maximum composite likelihood method for evolutionary distance estimation, multi-threading and multi-user support, and Linux support via the Wine application compatibility layer [13]. It has already seen over 15 000 downloads in the first 6 months of its release.

These increasing numbers of downloads per year have translated into expanded use of MEGA, as indicated by an accelerating number of publications citing MEGA (~10 000 from 1993 to 2007; Figure 3). The ISI Web of Science indicates that over 250 journals have published papers citing MEGA across a diverse spectrum of biological sciences. This growing impact of MEGA testifies to the widespread utility of statistical methods in deciphering evolutionary patterns and inferring the phylogenetic relationships of DNA and protein sequences in these times of momentous sequence data abundance. Indeed, a select group of phylogenetic analysis software tools has experienced extensive growth over the last few years (see Figure 1 in [10]).

In the following, we describe our biologist-centric philosophy behind the user-interface design, priorities for programming new methods and tools, and future plan of the MEGA project. Here we focus on conceptual aspects though some technical details are discussed.

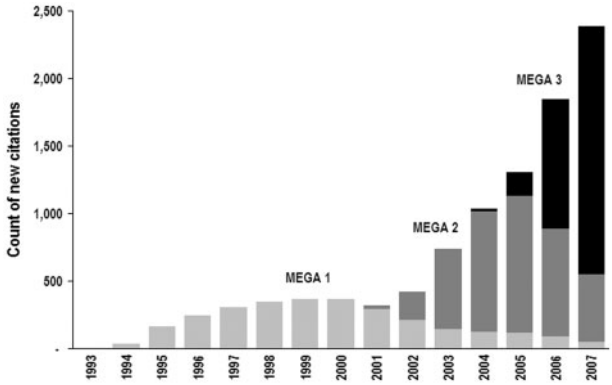

Figure 3: Number of MEGA citations each year (ISI Web of Science, February 2008 edition).

BIOLOGIST-CENTRIC DESIGNS IN MEGA

Simple user-interface and context-dependence

The availability of expansive screen display area in modern computing environments entices developers into presenting access to all of the software's functionality to the user up front in the form of complex hierarchical menu systems. This often leads to an over-populated interface and, consequently, steep learning curve for new users. In MEGA, we circumvented this pitfall by programming the user interface to render itself dynamically: it only displays buttons and menu options to the user that are context appropriate for the currently active data set and analysis conditions. Users specify models of sequence evolution and the data subset to employ only when needed by the program for calculations. Many new users have shared that they are able to learn MEGA functionality without much assistance, which we attribute in part to this context-dependent interface model. The imprint of context-dependence principle is seen throughout MEGA. For example, the distribution of the computational functionalities and display properties into input data explorers and output result explorers is also a product of the context-dependence design imperative, as it enables the user to conduct simple downstream analyses easily using the results presented.

Exploration of basic attributes of input data

All versions of MEGA contain visual modules for browsing, editing and computing basic statistical quantities for the input data. For example, the user

can calculate base frequencies and relative synonymous codon usage for all positions across all selected sequences or for only positions they highlight (Figure 2). These basic statistical quantities are necessary to assess the DNA and protein sequence variability, location of positions that harbor evolutionary change and inequality of the usage of 4 nucleotides, 20 amino acid residues and 64 codons. MEGA 4.1 supports exporting of statistical results (and even sequence alignments) to Microsoft Excel and to CSV formats for further analyses and graphical representations.

Also, input data explorers contain functions to select/eliminate specific genes, domains and species for analysis. Therefore, MEGA separates the building of the primary data subsets from the evolutionary analysis of data. In addition, users are offered many context-dependent data subset options, including the selection of codon positions to include, automatic translation of codons and the handling of sites containing alignment gaps by removing them for sequence pairs (pairwise-deletion option) or completely (complete-deletion option) across all sequences.

Result explorers

MEGA contains many visual explorers for results produced, which offer built-in capabilities to carry out additional calculations and prepare results for publication. The Tree Explorer is a prime example. It contains facilities for showing multiple representations of the tree, building consensus trees, compressing sub-trees to focus on higher level relationships among sequences, printing trees as Windows metafiles and TIFF files and exporting them in the Newick-compatible format for use in other evolutionary analysis programs. In MEGA 2, we added functions to compute 'linearized' trees [14] and apply molecular clock calibrations to estimate divergence times for all branching points in the tree displayed. The Tree Explorer also estimates ancestral states of nucleotide or amino acids on each node in the tree using the maximum parsimony method. In MEGA 3, capabilities for associating and displaying images for individual taxa and clusters were included to make it easier to produce an informative phylogenetic representation fit for publication.

The Distance Matrix Explorer, first made available in MEGA 2, is another visual result explorer that displays pairwise distances along with the estimates of standard errors, with flexibility to rearrange species (sequences or taxa) by dragging-and-dropping. It has the built-in capacity to calculate overall, within and between group averages when the user has arranged taxa into groups. When displaying results from tests of selection, it can highlight all pairs of sequences (or groups of sequences) in which the null hypothesis can be rejected at a 5% level. In MEGA 4.1, this explorer supports the export of distance matrices and tables to Microsoft Excel and to CSV formats for additional statistical analysis.

In addition, Distance Matrix Explorer in MEGA 4.1 has the capability to identify all pairs of sequences for which the evolutionary distances are not calculable due to the lack of any common sites in the sequence alignment. This frequently happens when applying phylogenetic analyses to data sets containing a large number of distantly related DNA and protein sequences. Researchers need to quickly find and eliminate some of these offending sequences to be able to build phylogenetic trees. Therefore, our focus is on enhancing the capability of result explorers to aid researchers in analysis of large data sets.

Intuitive data assembly and sequence alignments

Most bench scientists use a web browser to obtain gene sequences from databanks in a complex process that results in the mundane and frustrating task of cutting and pasting sequences from the web browsers, or saving them to files before processing them for sequence alignment. In developing MEGA 3, we emulated and enhanced the researchers' data assembly and alignment workflow under our *assist-rather-than-reinvent* design principle. We began by integrating a fully functional web browsing facility that had the ability to download sequence data directly into MEGA with a single click, which replaced a time-consuming, multi-step and error-prone manual process with a simple and intuitive procedure (see [12] for a description). Because experimental biologists frequently work with trace data, we also included facilities for viewing and editing the trace files (electropherogram) produced by the automated DNA sequencers.

MEGA 3 also featured a full function Alignment Explorer containing an extensive graphical user interface for handling and aligning sequence data sets gathered in MEGA through the web browsing facility and by importing data from FASTA and trace files [12]. It included a native implementation of

selected CLUSTALW source code [15] for automated sequence alignment with facilities for the manual refinement of the alignments. Because biologists often need to align subsets of sequences and positions in a larger alignment, we programmed Alignment Editor to align any user-selected rectangular region in all, or a subset of sequences, and insert it in the context of the larger alignment. For the protein-coding regions, users can click to translate the selected sequences (or selected rectangular subset) into protein sequences, align the translated protein sequences using CLUSTALW and switch their context back to DNA sequences—all with a few mouse clicks. The rapidly increasing user base of MEGA 3 (Figure 3) and the use of MEGA's Alignment Explorer in the most recent edition of a popular book [16] are a testament to the long-standing need for user-friendly and intuitive sequence alignment environment.

Transparency of assumptions and explanation of results

Intuitive GUI makes it easy for both novice and expert users to conduct a variety of computational and statistical analyses in MEGA. However, users must select which data subset to use (e.g., codon positions or domains), choose whether to delete all positions containing alignment gaps and missing data, specify an evolutionary model of DNA or amino acid substitution and choose whether to assume uniformity of evolutionary rates among sites or not. In MEGA 2, these options were available for selection in a series of tabs in a Dialog box, with selections in only one tab displayed at any time. Soon after its release, we realized that many users did not change the default options while analyzing data. Default options are not suitable in all circumstances. So, we revamped the main Analysis Preferences dialog box in MEGA 3 in order to make the user more aware of all the choices available at the same time. This development removed the burden on users of flipping through tabs to examine all options, enhancing the users' knowledge regarding the underlying assumptions and data-handling options chosen in each analysis.

In MEGA 4, we expanded the transparency of choices and assumptions by adding a new Caption Expert system. Caption Expert generates detailed descriptions for every result produced by MEGA, which informs the user of all the options selected and includes specific citations for any method, algorithm and software employed in the given analysis [13]. Furthermore, it gives the number of sequences used, the number of aligned positions included and the units of the resulting statistical quantities. All captions are context-dependent; that is, they change with the type of result displayed and the way the user is viewing the results. For example, if a user is viewing a phylogeny without branch lengths (cladograms), then Caption Expert writes about the branching pattern rather than the number of substitutions or the units in which the branch lengths are expressed. Therefore, we follow the *what-you-see-is-what-you-get* principle in generating descriptions. All descriptions appear in natural language text, which is aimed at promoting a better understanding of the underlying assumptions and finer details of the results presented. Such written descriptions of methods and results are useful for archival purposes, and they should aid students and researchers in preparing tables and figures for presentations.

Facilities for saving user sessions

In MEGA, we added functions in Alignment (MEGA 3) and Tree Explorers (MEGA 2) to save the current state of the visual module to be loaded later. For example, researchers often require multiple days to complete aligning sequences, frequently needing to add sequences to an alignment as new data as they become available. By saving the current Alignment Explorer session to a file, the users can return to a precise, visual snapshot of any previous alignment, along with any specified display or alignment parameters. This will assist researchers in assembling larger data sets incrementally and will not require saving of the data in flat text files that lose visual information associated with the sequence alignment and the phylogeny display. In the future, we plan to add a facility to save the input data session in order to obviate the need to save intermediate data subsets (selection of taxa, groups and genes) and other settings to text files; text files are cumbersome for very large and complex data sets.

DEVELOPMENTS IN MEGA 5 AND BEYOND

We are currently planning to make a number of fundamental enhancements prompted by the growing needs of contemporary researchers to

analyze large data sets, to use MEGA on multiple platforms and to utilize additional statistical and computational methods. In the following, we discuss a few of these aspects briefly.

Comparative genomics using maximum likelihood methods

Historically, MEGA has included likelihood methods for estimating evolutionary distances between sequence pairs as well as distance-based and Maximum Parsimony methods for inferring phylogenetic trees. Many novel methodological and algorithmic developments, combined with a manifold increase in the computing power on average desktops, have made it possible to carry out computationally intensive statistical methods of phylogenetic analysis based on the Maximum Likelihood (ML) principle [3, 5, 6]. Therefore, we plan to dramatically expand the repertoire of ML methods in MEGA. Starting with MEGA 5, we will gradually add facilities for selecting the best-fit model of DNA and protein substitution, estimating the extent of rate variation among sites, testing molecular clocks among species and paralogous genes, reconstructing nucleotides and amino acids in the ancestral sequences, and inferring phylogenetic trees. These additions will provide an integrated solution for analysis of molecular sequences using a variety of statistical methods.

Bioinformatics for biologists in MEGA

An increasingly greater number of biologists are now interested in automating their analyses in MEGA. This automation need has arisen from the profound success of sequencing efforts, which have fundamentally altered the nature and scope of investigations by evolutionary and molecular biologists. Biologists are routinely analyzing large sequence data sets, which, until recently, used to be the exclusive domain of bioinformatics investigators.

However, a majority of researchers using MEGA do not wish to trade their GUI environments for the cumbersome command line interfaces and learn unintuitive commands [10]. While the use of scripting in large-scale analysis provides a general solution to many bioinformatics problems, we find that researchers commonly turn to scripting to repeat the same analysis for different genes and genomic regions. Therefore, we plan to add a point-and-click iteration system that will enable scientists to launch the same analysis for distinct and overlapping data subsets, including different genes, codon positions, sliding windows and groups of sequences, without the use of scripting languages or the need to learn text command. Using this system, researchers will be able to build gene-by-gene phylogenies (and their consensus), estimate average sequence divergence for individual genes in multigene data sets, conduct gene-by-gene tests of selection, and estimate substitution parameters (e.g. transition/transversion rate ratios and G + C-contents) for different genomic segments in MEGA 5 and later releases.

MEGA as a cross-platform web application

Many biologists have expressed interest in using MEGA on non-Windows platforms (e.g. Macintosh and Linux). The porting of MEGA source code, especially the GUI, to multiple platforms has not been feasible; we estimate that this will require multiple years of development and debugging. However, the recent rise of the web browser as the centerpiece of the desktop as well as the cultural shift among computer users in regards to web-based application utilization, offers a unique opportunity to develop cross-platform solutions. The user interface experience offered by Web 2.0 applications has evolved beyond the static HTML-based displays and clunky, graphical Java applications embedded into HTML-based web pages. The Rich User Interface (RUI) technologies made available in the Web 2.0 era have enabled the creation of web-based analogues of common desktop applications. This has the potential to help us bring the full MEGA GUI experience to the web-desktop (WebTop). The general availability of advanced, open-source JavaScript frameworks would allow us to implement many dynamic elements (i.e. contextual menus and tabular data displays) characteristic of the desktop application user experience. Therefore, our plan is to make available WebTop MEGA version in the future. MEGA for the WebTop will also use a web service model so the user can experience a perpetual upgrade cycle wherein improvements and bug fixes can be implemented and delivered whenever they are available.

In summary, MEGA is an integrated workbench for biologists for mining data from the web, aligning sequences, conducting phylogenetic analyses, testing evolutionary hypothesis and generating publication quality displays and descriptions.

Key Points

- Software developers need to go beyond user-friendliness to a biologist-centric approach for building tools for researchers.
- The burden is on bioinformatics software developers to inform users about the precise nature of the results generated.
- There is a need to use context-dependence in the design of user interface and adopt an assist-not-reinvent philosophy for integrating biologists' workflows.
- The advent of Web Top (Web 2.0 associated technologies) offers fresh avenues for developing new and porting existing software for use on multiple operating systems.

Acknowledgements

Support for this work comes, in part, from the National Institutes of Health (S.K., M.N.) and Japan Society of Promotion of Science (K.T., S.K.). We are grateful to Kristi Garboushian for her editorial comments and Wayne Parkhurst for assistance with graphics.

References

1. Swofford D. *PAUP: Phylogenetic Analysis Using Maximum Parsimony*. Illinois: Illinois Natural History Survey Champaign, 1990.
2. Felsenstein J. *PHYLIP: Phylogenetic Inference Package*. Seattle, WA: University of Washington, 1993.
3. Felsenstein J. *Inferring phylogenies*. Sunderland, MA: Sinauer Associates, 2004.
4. Nei M. *Molecular Evolutionary Genetics*. New York: Columbia University Press, 1987.
5. Nei M, Kumar S. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press, 2000.
6. Yang Z. *Computational Molecular Evolution*. Oxford: Oxford University Press, 2006.
7. Kumar S, Tamura K, Nei M. *Manual for MEGA: Molecular Evolutionary Genetics Analysis Software*. University Park, PA: Pennsylvania State University, 1993.
8. Kumar S, Tamura K, Nei M. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput Appl Biosci* 1994;**10**:189–91.
9. Saitou N, Nei M. The neighbor-joining method – a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**:406–25.
10. Kumar S, Dudley J. Bioinformatics software for biologists in the genomics era. *Bioinformatics* 2007;**23**:1713–7.
11. Kumar S, Tamura K, Jakobsen IB, *et al*. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 2001; **17**:1244–5.
12. Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 2004;**5**:150–63.
13. Tamura K, Dudley J, Nei M, *et al*. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007;**24**:1596–9.
14. Takezaki N, Rzhetsky A, Nei M. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol* 1995;**12**: 823–33.
15. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* 1996; **266**:383–402.
16. Hall BG. *Phylogenetic trees made easy: a how-to manual*. Sunderland, MA: Sinauer Associates, 2007.