

MEGA Software Celebrates Silver Anniversary

In the past 25 years, stretching from the dawn of personal computing to the social media age, the “always free” molecular evolutionary genetics analysis (MEGA) tool has been downloaded 1.6 million times worldwide.

In 1993, the very first version of Molecular Evolutionary Genetics Analysis (**MEGA**) software was developed to meet the growing needs for nucleic acid and protein sequence analysis tools among evolutionary and molecular biologists. MEGA was developed by Sudhir Kumar and Koichiro Tamura in the laboratory of Masatoshi Nei, who guided and supported MEGA development throughout its early years.

On the occasion of MEGA’s silver anniversary, Kumar and Tamura reflected on the origins and refinement of the software over the years, as well as its continuing impact on the field of evolutionary biology.

“I began my doctoral studies in the Nei lab in January 1991,” said Kumar. “The Nei lab routinely pioneered new methods and implemented them in computer programs so others could use them. MEGA’s development escalated that tradition to fulfill the need for a user-friendly tool that enabled molecular phylogenetic analysis using evolutionary distances.”

MEGA was designed from the start to be “always free” to the public, and “developed to facilitate statistical analyses of molecular evolution by using personal computers.”

The MEGA software was first announced to the world in a three-page publication in the journal *Computer Applications in the Biosciences*, or *CABIOS* (later changed to *Bioinformatics*).

“Within a few years, we received over 2,500 requests for MEGA by postal mail,” said Kumar. “The first version was distributed on 5.25- or 3.5-in floppy diskettes, which have long since gone extinct. We also, by necessity, ended up running a small print and distribute shop in the lab to mail out the handy 140-page instruction manual we developed, which included a lot of useful information on newer statistical methods in computational evolutionary biology, including distance estimation methods, phylogenetic inference algorithms, and basic sequence statistics.”

MEGA 1.0 was designed to handle single gene comparisons across species. It was written in C++ programming language and intended to be used on market leading IBM and IBM-compatible personal computers, which were predominantly built with 486 processors and had a maximum of 640 kb RAM memory.

“MEGA is designed to conduct various statistical analyses in one program and to produce results in publication-quality outputs,” wrote Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei.

An interactive user interface, they wrote, “can be used on most color and monochrome monitors, and it responds to

the keyboard as well as to the mouse”—all fairly new PC features at the time. On-screen sequence data and phylogenetic-tree viewers facilitated publication-quality outputs with a wide range of dot-matrix printers, which were also considered state-of-the-art.

Both “data mining” and bioinformatics represented a fairly new way of studying biology, using only the silicon power of computer microprocessors, later dubbed science in silico.

Decades of advancement in computational biology methods have included estimating evolutionary distances, reconstructing phylogenetic trees and computing basic statistical quantities from molecular data, all of which has been facilitated by an explosion of accessible sequence data deposited in GenBank. The NIH genetic sequence database began in 1982 and surpassed 150,000 genomic sequences and 150 million bases by 1993.

In 1993, genomics was in its infancy. The first genome had yet to be sequenced and the Human Genome Project has just getting off the ground. It would not be until two more years when J. Craig Venter and colleagues published the first completely sequenced genome of a self-replicating, free-living organism—the bacteria *Haemophilus influenzae*.

“The sequence data began to accumulate in large amounts, and everyone needed new analytical methods and tools to harness these data. MEGA was the right program at the right time to enable scientists to take advantage of the PC revolution in their data analyses,” said Kumar, who completed his Ph.D. in the Nei lab in 1996, and has since remained the steady, driving force behind MEGA.

The creation of MEGA also coincided with the launch of the World Wide Web, which was dominated at the time by Mosaic, the first browser to truly popularize the Web and the Internet.

MEGA Sequel

Software that is not updated can quickly face obsolescence, so the launch of MEGA2 8 years later required a major overhaul to capitalize on the convergent technological forces driving the advancement of evolutionary analysis at breakneck speed.

Computational power was doubling every two years as predicted by Moore’s law, and this power was essential to handling the data produced by a growing number of gene and genome projects (including model organisms *C. elegans*, *S. cerevisiae*, and *D. melanogaster*, and the first draft of the Human Genome Project). By 2001, GenBank reached the milestones of 15 million genomic sequences and 15 billion bases deposited.

“The first version of MEGA made many versions 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," wrote the authors in their 2001 *Bioinformatics* article.

"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 analysis."

"The second version of MEGA 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," said Koichiro Tamura, who has co-led MEGA software development for over two decades.

Starting with version 2, MEGA has been continuously and freely available to download from www.megasoftware.net. From these humble beginnings of the personal computer age, MEGA has become one of the most-cited and most-downloaded evolutionary software tools. The number of MEGA downloads now exceeds 1.6 million across nearly 200 countries.

MEGA for All

Like clockwork, a new iteration of MEGA has emerged every few years, each with some unique features. At the same time, MEGA has not been merely intended to be a catalogue of methods, but rather, provide a seamless user experience. MEGA 3 (2004) integrated sequence data alignment, MEGA 4 (2007) added an expert system to generate figure legends, MEGA 5 (2011) added maximum likelihood methods for molecular phylogenetics, and MEGA 6 (2013) added methods and tools for estimating divergence times and analyzing gene duplicates.

In addition to these releases, the MEGA team also produced MEGA-CC (2012) to address a growing need in the research community to apply MEGA for batch processing a large number of datasets. This was followed by MEGA-MD (2014) for human mutation diagnosis. MEGA was the first software that offered methods for both phylogenomics and phylomedicine.

For MEGA7, "we performed a significant upgrade of MEGA, which was necessary to speed up the data-crunching time and memory usage with 64-bit processors, and much larger memory space to handle gigabytes of data, so now people can analyze an ever-larger amount of sequences," said Tamura. MEGA7 is the most sophisticated and powerful version yet, designed to extend its capability to analyses of more complex and large DNA data sets on Microsoft Windows.

For Kumar, Tamura, and Nei, free access to MEGA for the scientific community is a key to propelling worldwide evolutionary discoveries. "MEGA has been freely available now for 25 years, and for any use, spanning research, teaching, and industry. We enable people throughout the world, including developing nations, to use fundamental technologies that are needed to address these burgeoning sequence databases.

Everyone in the world should be able to use evolutionary and genomics tools to analyze the wealth of information that is being produced relating the genomes of humans to pathogens, to disease traits, to uncover our similarities and differences. It will take all of our global efforts to do so. The most important thing is to develop user-friendly, sophisticated software for use by all."

And, in the last quarter century (1993–2018), the various iterations of MEGA have been cited in >120,000 research publications, spanning a diverse range of biological research disciplines. MEGA has helped investigators worldwide to make discoveries in a wide range of areas such as infectious diseases, evolutionary biology, and functional genomics.

Never resting on their laurels, Kumar, Tamura, and their development team remain hard at work, actively seeking input from users around the world for developing the next version of MEGA. "The overwhelming feedback was that MEGA needs to be available for use on Linux and macOS," said Glen Stecher, who joined MEGA software team 7 years ago.

This led to MEGA X, newest version of MEGA. MEGA X is the second major overhaul that will run natively on Linux and Mac computer systems in addition to MS Windows. The use of 'X' in this version name is a dual signifier: MEGA X is actually the 10th iteration of MEGA (including MEGA-CC and MEGA-MD releases) and MEGA is now a cross-platform application. MEGA X for Linux and Windows is announced in this issue, with a macOS version to follow within the next few months.

These core software enhancements will contribute to the longevity of MEGA for future use in molecular evolution, bioinformatics, functional genomics, computational biology, and basic biomedicine applications.

Joseph Caspermeyer^{1,*}

¹MBE Press Office

*Corresponding author: E-mail: MBEpress@gmail.com.
doi:10.1093/molbev/msy098

Appendix: Chronology of a quarter century of MEGA developments.

1993: MEGA 1 was programmed in the original Borland C++ using the Turbo Vision compiler. It was meant for use on Personal Computers running MS-DOS (Microsoft Disk Operating System [MS-DOS]). It had a character-based graphical user-interface, which had menus, data and tree explorers, and support for using the mouse. It was a 16-bit application. Its use has been cited in over 3,000 publications.

Bioinformatics 10:189–191.

2001: MEGA 2 was a complete rewrite of MEGA for use on Microsoft Windows Graphical User Interface. It was a 32-bit application, which increased its sequence analysis capabilities dramatically. It has been cited in over 8,500 publications.

Bioinformatics 17:1244–1245.

2004: MEGA 3 added sequence data alignment and assembly features, and effectively integrated sequence data acquisition

from databases via internet with the evolutionary analyses. Its use has been cited in >13,000 publications.

Brief Bioinformatics 9:299–306.

2007: MEGA 4 incorporated new computational methods and an expert system to generate figure legends for every result presented by MEGA 4. This feature was added to promote a better understanding of the underlying assumptions used in analyses, and of the results produced. It was the first iteration to be published in MBE and the most successful software update yet, cited >30,000 times. It was ranked number 45 among the top-100 most cited articles of all time by Nature (October 29, 2014).

Mol Biol Evol. 24:1596–1599.

2011: MEGA 5 added maximum likelihood methods for molecular phylogenetics. In this version, the user interface was redesigned to be activity driven to make it easier for the use of both beginners and experienced scientists. This has been the most-cited version of MEGA, with >35,000 citations to date.

Mol Biol Evol. 28:2731–2739.

2012: MEGA-CC made available the computing core of the MEGA software as a stand-alone program to address the growing need in the research community to apply MEGA for batch processing a large number of datasets and to integrate it into their analysis workflows. It has been cited in >130 publications.

Bioinformatics 28:2685–2686.

2013: MEGA 6 added methods and tools for building molecular evolutionary trees scaled to time (timetrees). A TimeTree Wizard with an intuitive step-by-step graphical interface was introduced to enable users to provide many inputs needed to build timetrees. It has been cited >22,000 times.

Mol Biol Evol. 30:2725–2729.

2014: MEGA-MD introduced a tool to forecast the deleteriousness of amino acid substitutions using evolutionary methods. This facility is available in all subsequent releases of MEGA.

Bioinformatics 30:1305–1307.

2016: MEGA 7 upgraded MEGA's architecture to 64-bit computing and added functionalities to predict gene duplication events in gene family trees. Its use has been cited in over 4,800 publications.

Mol Biol Evol. 33:1870–1874.

2018: MEGA X is a major rewrite of MEGA source code and graphical user interface in order to make it work seamlessly across multiple computing platforms. Here the "X" marks both the 10th iteration of MEGA (including CC and MD releases) and that it is now a cross-platform application. MEGA for Linux and Windows is released.

Mol Biol Evol. 35:1547–1549.