MEGA: A Bio Computational Software for Sequence and Phylogenetic Analysis

Vipan Kumar[1], Apurba Dey[2], Amarpal Singh[3]

***Abstract*—In the last five years, Biocomputing has moved into central position in molecular biology research. Enormous improvements in genetic mapping and sequencing technology have led to the accumulation of vast amount of biological information in the database. With the advent of this extensive repertoire of raw sequence information, the next major challenge for a modern researcher is to interpret this biological information. At the remarkable rate of progress that is being made in sequencing and phylogenetic organisms, wet research techniques alone cannot keep up with the influx of genomic information. Molecular Evolutionary Genetic Analysis (MEGA) is user friendly bio-computational software for sequence analysis and phylogenetic analysis.**

MEGA developed with the aim to bridge the gap between wet lab result and significance that can characterize by nucleotide and amino acid to produce scoring and evolutionary relationship. In this paper an attempt had been made to analysis the growth and evolution of MEGA software from MEGA1 to MEGA 4 and its advantages over other bioinformatics tool box.

***Index Terms*—Biocomputing, Sequence Alignment, Scoring, MEGA**

I. Introduction

MEGA is software tool with the functional task of providing annotation of biologically relevant information from a nucleotide or proteomic sequence. Its' goal is to provide a simple but powerful interface for the analysis and mining of genomic information while seamlessly handling the nonscientific complexities of interfacing with hardware, computational clusters, software packages, raw data, and file formats.

The design of the MEGA software package (Molecular Evolutionary Genetic Analysis) is engineered to be an 'extensible framework' that facilitates expansion with any bioinformatics software tool approaching point and click simplicity. MEGA 4 is specifically designed to reduce the time needed for mundane tasks in data analysis and to provide statistical methods of molecular evolutionary genetic analysis. MEGA 4 also includes distance matrix and phylogeny explorers well as advanced graphical modules for the visual representation of input data and output result.

Manuscript received 10 Januarary, 2009 This work was supported in part by the Department of Chemical Engineering & Bio Technology.

1. Vipan Kumar ,Senior Lecturer, Department of Chemical Engineering & Bio-Technology ,Gurdaspur-143521,Punjab.India (corresponding author to phone:91-98882-20918; fax:91-1874-221463; e-mail: vipan752002@ indiatimes.com)
2. Apurba Dey Dr., Professor, Department of Biotechnology, NIT, Durgapur, West Bengal, India (e-mail:apurbadey2003@yahoo.co.in).
3. Amarpal Singh Dr. Department of Electronics Engineering & Communication, Gurdaspur-143521,Punjab.India (e-mail:a_singh@yahoo.com).

II. Literature Review & Evolution Of MEGA

MEGA Version 1

The first version of MEGA (1994) made many methods of evolutionary analysis easily accessible to the scientific community for research and education.MEGA1 was developed keeping in mind the limited computational resources available on the average personal computer (RAM) in the early 1990. MEGA1 launched without sophisticated features because low memory of the system. It exclude Sequence alignment construction (alignment editor, Motif searching and BLAST sequences from alignment), multiple sequence alignment, substitution pattern homogeneity test (Monte Carlo test) and Synonymous/Non-synonymous (Modified Nei-Gojobori method).In addition of that UPGMA, part of phylogeny is not considered in version 1.Moreover the MEGA 1 used the DOS operating platform.

MEGA Version 2

As compared to the first version, MEGA2 contains a more extensive collection of methods to estimate the number of synonymous/Non-synonymous substitutions per synonymous site and non-synonymous substitutions respectively. MEGA 2 had added the modified Nei-Gojobori method as well as Li-Wu-Lou methods and its modification. MEGA 2 expanded with 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, Maximum evolution and maximum parsimony methods are also included in this version. Molecular evolutionary genetic analysis version 2 broadly has capabilities (1) analyses of large dataset (2) creation and analysis's of groups of sequences (3) expanding the repertoire of statistical methods for molecular evolutionary studies (4) new modules for visual representation of input data and output results on Microsoft window platform.

A survey of the research papers citing the use of MEGA reveals that it has been utilised in diverse disciplines, including AIDS/HIV research, virology, bacteriology and general disease, plant biology, conservation biology, systematics, developmental evolution and population genetics.

MEGA Version 3

The newly released MEGA3 expands the functionalities of MEGA2 by adding sequence data alignment and assembly

features, along with other advancements. The sequence data acquisition is now effectively integrated with the evolutionary analyses, making it much easier to conduct comparative analyses in an integrated computing environment. MEGA3 is not trial to be a catalogue of evolutionary analysis methods. Rather, it is for exploring sequence data from evolutionary perspectives, constructing phylogenetic trees, and testing evolutionary hypotheses, especially for large-scale data sets that have been generated by recent genomics projects. The advantages of MEGA 3 are the sequence alignment and data assembly modules – as they constitute the first step in any comparative sequence analysis investigation.

This is followed by descriptions of the different types of data that MEGA can analyze its graphical input and output data explorers, dynamic data sub setting facilities, and the statistical methods and computational tools available for inferring phylogenetic trees and estimating evolutionary distances.

MEGA Version 4

The fourth version (MEGA4) contains three distinct newly developed functionalities, Firstly, 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 pat- tern, 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. The significance of this description is 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.

Second, MEGA 4 added a Maximum Composite Likelihood (MCL) method for estimating evolutionary distances (dij) 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 (h) to every sequence pair (i,j).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.

Third, we have now programmed MEGA4 to run on some versions of Linux through the Wine software compatibility layer. 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. MEGA4 running on Linux show the display, stability, and performance to be highly satisfactory and comparable to the native Windows system .Further more, investigators now report MEGA4 running on Intel-based Macintosh computers under the Parallels program as well as it does on Windows-native personal computers .The Parallels program is a native solution for Macintosh computers that permits them to simultaneously run Windows and Macintosh software.

MEGA 4 also 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

III. METHOD

We are utilizing a series of techniques for the prediction of specific characteristics of proteins and nucleotide. Each analysis package is treated as a modularized component that can be dynamically incorporated in MEGA to facilitate extensibility and allow updating output data for research of sequence alignment (pair-wise and multiple) and new methods are developed for evolutionary analysis.

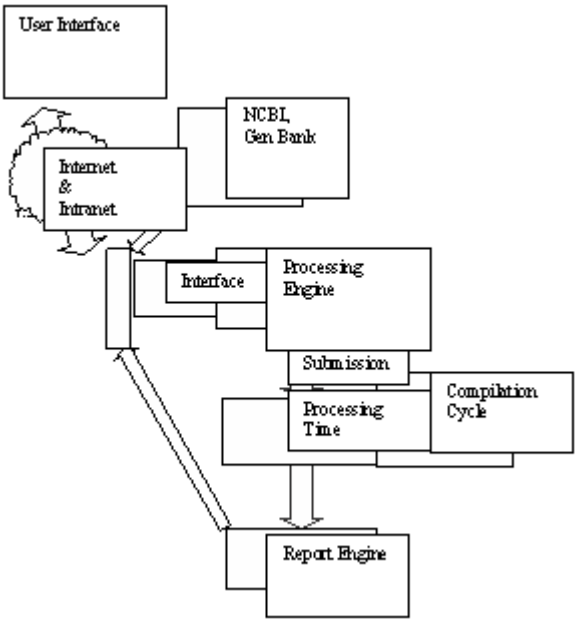

Figure1: Working Model of MEGA Software

IV. Approach

Included in the initial release of MEGA, software have developed or incorporated existing modules for the analysis of the following qualities:

- Creating Multiple Sequence Alignment
- Sequence Conversion from Nucleotide to Protein
- Computing Basic Statistical Quantities for Sequence Data
- Constructing Phylogenetic Trees
- Distance Matrix Explore
- Physiological / Chemical characteristics
- Computing Evolutionary Distance
- Synonymous and Non-synonymous Test

Details as to the nature of modules, external resources, databases and software tools incorporated in the MEGA are listed as a web resource.

V. Implementation

MEGA is composed of three individual separate components, namely - A graphical user interface, a processing engine, and a report engine. The graphical user interface provides the control point that the user interacts with on a personal computer. This component is developed to be run on any desktop computer (Macintosh –Windows using Virtual PC, Sun Workstation-Soft Windows 95 and Linux-Window using VMWare) using a multiplatform for uniformity.

The processing engine is authenticate, manages the forking and execution of individual analyses or generation of shell scripts for the intelligent submission to a processing queue. Queues currently supported are National Centre for Biotechnology Information (NCBI), Clustal, Blast and Fasta.

The report engine combines resulting output to attempt to interpret make intelligent summaries along with presenting the raw analyses output in a failure format. Moreover, a user can specify and filter for specific characteristics of amino acids and nucleic sequence, which enable a user to utilize MEGA as user friendly search tool.

VI. Conclusion

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 MEGA latest version is continuously adding new methods and functions up to MEGA 4. MEGA, 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. MEGA have produced the foundation of a powerful, easy to use, extensible software package that is able to make a working hypothesis of characteristics of a sequence and evolutionary analysis. MEGA produces annotation results that are interpreted by computer and presented in a useful manner, more so than a conglomeration of separate analyses.

References

[1] Kumar, S., Tamura, K. and Nei, M. (1994), 'MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers', Computational Applied. Bioscience. Vol. 10, pp. 189 – 191.

[2]. Kumar, S., Tamura, K., Jakobsen, I. B. and Nei, M. (2001), 'MEGA2: Molecular Evolutionary Genetics Analysis software', Bioinformatics, Vol. 17, pp. 1244 – 1245. 10. Kumar, S., Tamura, K. and Nei, M. (1993), 'Manual for MEGA: Molecular Evolutionary Genetics Analysis Software', Pennsylvania State University, University Park, PA.

[3].Tamura, K. and Kumar, S. (2002),'Evolutionary distance estimation under heterogeneous substitution pattern among lineages', Mol. Biol. Evol., Vol. 19, pp. 1727 –1736.

[4]. Kimura, M. (1980), 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', J. Mol. Evol., Vol. 16, pp. 111 – 120.

[5]. Tamura, K. and Nei, M. (1993), 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees', Mol. Biol. Evol.,Vol. 10, pp. 512 – 526.

[6]. Tamura, K. (1992), 'The rate and pattern of nucleotide substitution in Drosophila mitochondrial-DNA', Mol. Biol. Evol., Vol. 9, pp. 814 – 825.

[7] Tajima, F. and Nei, M. (1983), 'Estimation of evolutionary distance between nucleotide sequences', Mol. Biol. Evol., Vol. 1, pp. 269 –285.

[8] .Kumar, S. (1996), 'A stepwise algorithm for finding minimum evolution trees', Mol. Biol. Evol., Vol. 13, pp. 584 – 593.

[9]. Kumar, S. and Gadagkar, S. R. (2000), 'Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies', J. Mol.Evol., Vol. 51, pp. 544 – 553.

[10]. Kumar S, Tamura K, Nei M. 2004. 'MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment' Brief Bioinformatics. 5:150–163.

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

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

[13] 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.

[14.] Thompson JD, Higgins DG, Gibson TJ. 1994. 'Clustal W 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.

[15]. 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.