Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/17910
DOI URL: http://dx.doi.org/10.21474/IJAR01/17910

RESEARCH ARTICLE

COMPUTATIONAL VALIDATION AND ANALYSIS OF SEMI-QUANTITATIVE DATA USING IN-SILICO APPROACHES

Sai Satya Sri Pulla[1], Vennela Chowdary Gullapalli[1], Pothuri Vishnu[1], Yaswanth Subhash Chowdary Pidikiti[1], Vignaya Maram[1], Nagini Narne[1], Prathika Shalom Kurapati[1], Jaswanth Kalyan Kasa[1] and Reethika Singh Ranwas[2]

1. Department of Biotechnology, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India.
2. Department of Molecular Biology, Cresent Biosciences, ICRAB Campus, Kanpur, Uttar Pradesh, India.

Manuscript Info

Manuscript History
Received: 17 September 2023
Final Accepted: 24 October 2023
Published: November 2023

Key words:-
Computational, EMBOSS, Data, Biology, Computer, Validation

Abstract

With the advent use of computers in every daily life the computational approacheshas collaborative effort between biologists and computer scientists and thus covers a wide variety of traditionalcomputer science domains, including data retrieval, data integration,data cleaning, data modeling, data mining, data warehousing, data managing, ontologies, simulation, parallel computing, agent-based technology, grid computing, and visualization. However, applying each of these domains to biomolecular and biomedicalapplications raises specific and unexpectedly challenging research issues. This review is to provide life scientists and computer scientists with a complete view on biological data management byidentifying specificissues, presenting existing solutions from both academia and industry and providing a framework in which to compare thesesystems.

Introduction:-

Computational approaches and the management of scientific data are critical to support life science discovery.[1] As computational models of proteins, cells, and organisms become increasingly realistic, much biology research will migrate from the wet lab to the computer.[2] Successfully accomplishing the transition to biology in silico, however, requires access to a huge amount of information from across the research community.[3] Much of this information is currently available from publicly accessible data sources, and more is being added daily.[4] Unfortunately, scientists are not currently able to identify easily and exploit this information because of the variety of semantics, interfaces, and data formats used by the underlying data sources.[5]DNA sequences are often modeled as probabilistic phenomena, with the patterns ofinterest being defined as samples drawn from the underlying random process.[6] Forexample, the underlying DNA sequence is modeled as a Markov chain of randomvariables taking on the values (A, C, T, G).[7] The underlying models that define the DNA sequences and their accuracy areultimately a determinant of the accuracy with which the patterns are subsequentlydetected.[8]. Sequence models provide a basis for establishing the significance of patternsobserved, while the pattern models help us look for specific motifs that are of functionalsignificance.[9]

Analyzing Sequences by EMBOSS

Since the beginning of big genome sequencing, initiated by the work on the nematodeCaenhorhabditis elegans, the Staden group has concentrated on developing methods toincrease the efficiency of these large-scale

Corresponding Author:- Reethika Singh Ranwas
Address:- Department of Molecular Biology, Cresent Biosciences, ICRAB Campus, Kanpur, Uttar Pradesh, India.

projects.[10]Features inthe design of EMBOSS program are its flexibility in output formats, its use of a language Ajax Command Definitions (ACD) for specifying the inputs to its programs. The firsttechnical challenge was to parse the ACD to automatically produce suitable dialogue boxes for each EMBOSS program and to prepare SPIN to load the results intomemory. The second problem was to parse these varying results files to display theresults and allow users to interact with them as though they had been produced byinternal SPIN functions.

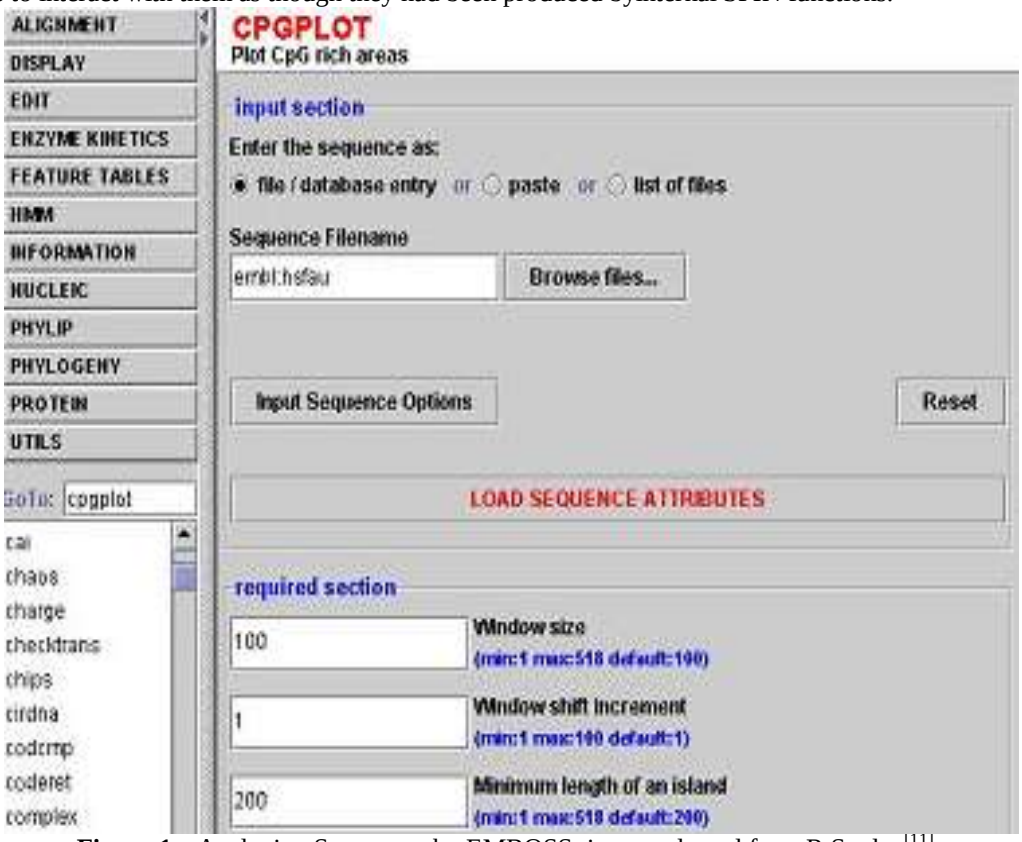

Figure 1:- Analyzing Sequences by EMBOSS, image adapted from R Staden[11].

EMBOSS

EMBOSS is a free Open Source software analysis package specially developed forthe needs of the molecular biology user community.[12] The software uses data in a variety of formats and even allows transparent retrieval of sequence data from the web. Asextensive libraries are provided with the package, it is a platform that allows otherscientists to develop and release software in true open-source spirit. EMBOSS alsointegrates a range of currently available packages and tools for sequence analysis intoa seamless whole.[13] At the time of writing EMBOSS contained over 100 programs for sequence alignment, rapid database searching with sequence patterns, proteinmotif identification, nucleotide sequence pattern analysis,codon usage analysis for smallgenomes, rapid identification of sequence patterns in large scale sequence sets andpresentation tools for publication.

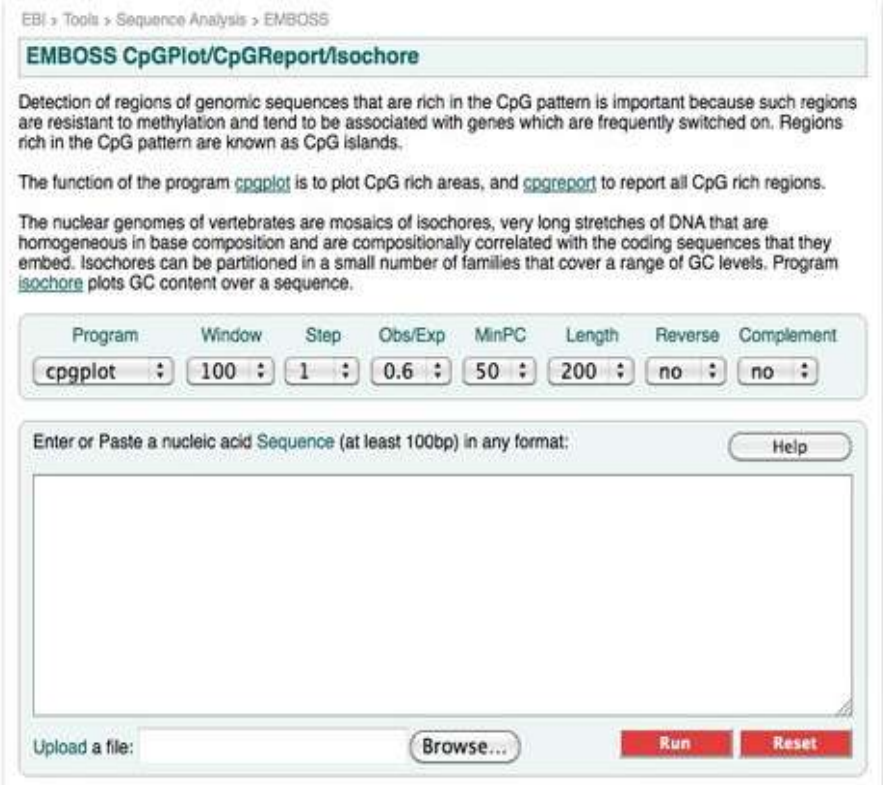

Figure 2:- Retrieval of sequence data from the web usingEMBOSS.

Databases Are Autonomous

Biological data sources represent a loose collection of autonomous websites, each with its own governing body and infrastructure.[14] These sites vary in almost every possible instance such as computer platform, access, and data management system. Much of the available biological data exists in legacy systems in which there are no structured information management systems. These data sources are inconsistent at the semantic level, and often, there is no adequate attendant meta-data specification. Scientific literature, images, and other free-text documents are commonly stored in unstructuredor semi-structured formats (plain text files, HTML or XML files, binary files). Genomic, microarray gene expression, and proteomic data are routinely storedin conventional spreadsheet programs or in structured relational databases (Oracle, Sybase, DB2, Informix).[15] Major data depository centers have implementedvarious data formats for operations; the National Center for Biotechnology Information (NCBI) has adopted the highly nested data system ASN.1 (Abstract SyntaxNotation) for the general storage of gene, protein, and genomic information; the United States Department of Agriculture (USDA) Plant Genome Data and Information Center has adopted the object-oriented, data management systems and interface.[16]New databases spring up at a rapid rate and older databases disappear.In response to the advance of biological research and technology, the overall features of biological data sources are subjected to continuous changes includingdata content and data schema.

Functional data visualization

The construction, visualization and interpretation of phylogenetic trees are instrumental for biological analysis in multiple fields, including evolutionary biology, genetics and comparative genomics.[17] Recently published software tools allow for the use of interactive elements and comprehensive analytics in association with these trees by means of the latest web technologies. The commonly used phyloXML standard has been extended with various elements of the complex type as permitted by the current phyloXML XSD schema. The newly created elements <taxonomies> and <domains> are used to specify taxonomy and domain colours, descriptions and links. In order to reference between the <phylogeny> element of the current standard and the new elements, the <code> and <name> sub-elements are used for the <taxonomy> and the <domain> element, respectively.[18] Colours are represented as HEX values (e.g. 0zGGBBRR). Furthermore, graphs like a pie, a binary and a multi-bar chart as well as a heat map or a boxplot can be displayed next to the leaf nodes by using the newly

defined <graph> elements.PhyD3 is implemented as a flexible and lightweight tool allowing for the display of interactive and complex phylogenetic trees in a web-based environment without security-based limitations and without the need for external plugins.[19] The implementation is fast and responsive to user interaction with the tree elements' display parameters being easily changed through user-friendly access controls.

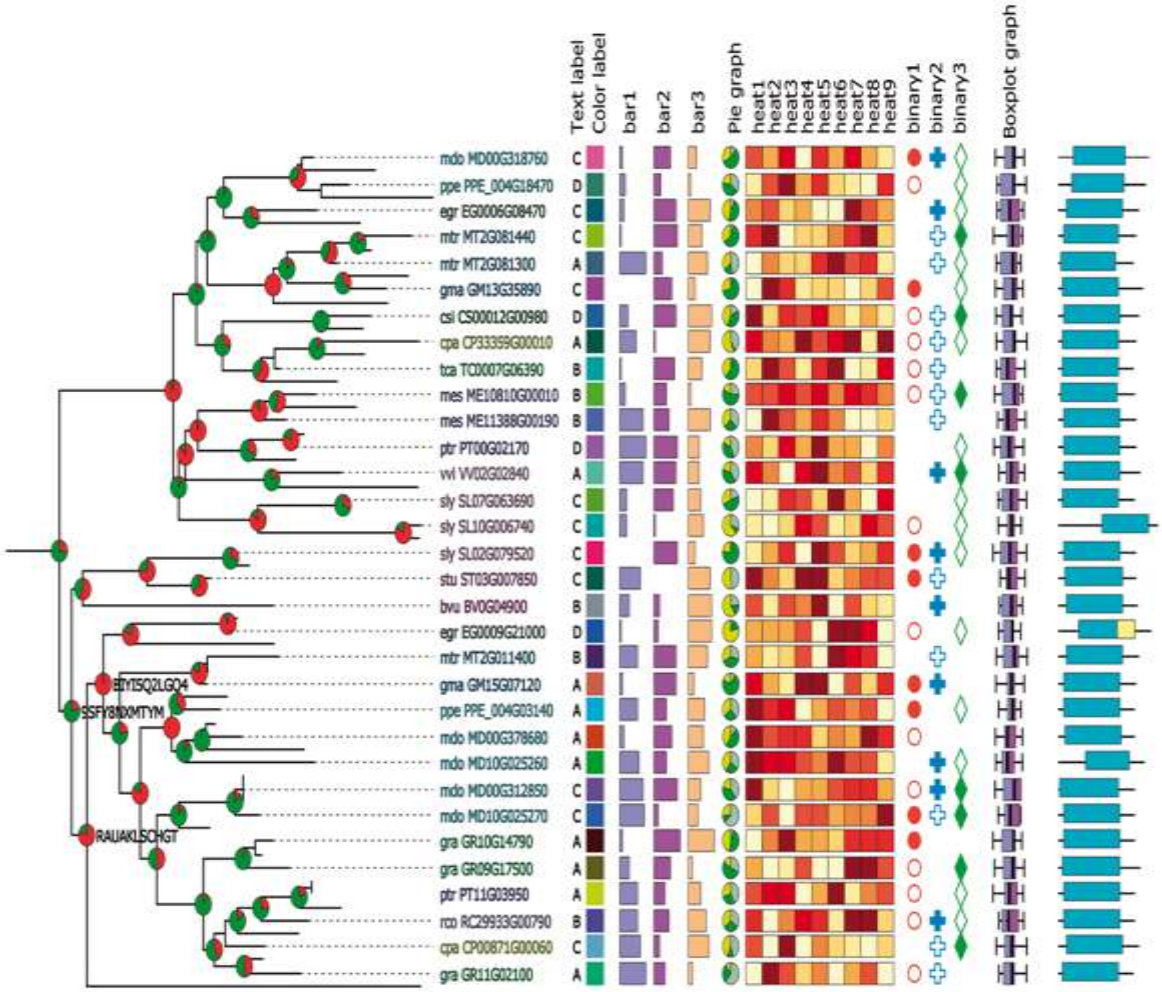

Figure 3:- Visualization of artificial data using inner node pie charts, taxonomy colorization, leaf node labelsand domain architecture.

PhyD3 provides import and export tools to facilitate greater interoperability. Using the import tool users can supply trees in Newick and phyloXML formats,[20] with optional numerical data, which can be easily converted to the extended phyloXML format with graph annotations. Export capabilities provide the conversion of the current tree visualization into vector graphics (SVG format) and bitmaps (PNG format), as well as the extended phyloXML data itself.

Conclusion:-

Computers and biologists have to work together to address the level of challenges presented by the inherent complexity and vast scales of time and spacecovered by the life sciences. The opportunities for biological science research in the 21st century require a robust, comprehensive information integration infrastructure underlying all aspects of research. As we discussed in our review, substantial progress has been made for data integration at the technical andarchitectural level. However, data integration at the semantic level remains a major challenge. Before we will be able to seize any of these opportunities, the biologyand bioinformatics communities have to overcome the current limitations in metadata specification, maintenance of data provenance and data quality, consistentsemantics and ontology, and web presentations.

References:-

1. Michener WK, Jones MB. Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution. 2012 Feb 1;27(2):85-93.
2. Motta S, Pappalardo F. Mathematical modeling of biological systems. Briefings in Bioinformatics. 2012 Oct 14;14(4):411-22.
3. Zengler K, Palsson BO. A road map for the development of community systems (CoSy) biology. Nature Reviews Microbiology. 2012 May;10(5):366.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic acids research. 2008 Oct 21;37(suppl_1):D26-31.
5. Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC. DiscoveryLink: A system for integrated access to life sciences data sources. IBM systems Journal. 2001;40(2):489-511.
6. Singh GB. Statistical modeling of DNA sequences and patterns. InIntroduction to Bioinformatics 2003 (pp. 357-373). Humana Press, Totowa, NJ.
7. Leroux BG. Maximum-likelihood estimation for hidden Markov models. Stochastic processes and their applications. 1992 Feb 1;40(1):127-43.
8. Bell O, Tiwari VK, Thomä NH, Schübeler D. Determinants and dynamics of genome accessibility. Nature Reviews Genetics. 2011 Aug;12(8):554.
9. Singh GB, Singh H. Databases, models, and algorithms for functional genomics. Molecular biotechnology. 2005 Feb 1;29(2):165-83.
10. Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Staden R, Halloran N, Green P, Thierry-Mieg J, Qiu L, Dear S. The C. elegans genome sequencing project: a beginning. Nature. 1992 Mar;356(6364):37.
11. Staden R, Judge DP, Bonfield JK. Analyzing sequences using the Staden package and EMBOSS. InIntroduction to bioinformatics 2003 (pp. 393-410). Humana Press, Totowa, NJ.
12. Gilbert D. Bioinformatics software resources. Briefings in bioinformatics. 2004 Sep 1;5(3):300-4.
13. Tiwari A, Sekhar AK. Workflow based framework for life science informatics. Computational biology and chemistry. 2007 Oct 1;31(5-6):305-19.
14. Robbins RJ. Bioinformatics: Essential Infrastructure for Global Biology1. Journal of Computational Biology. 1996;3(3):465-78.
15. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic acids research. 2006 Nov 11;35(suppl_1):D760-5.
16. Barry PT. Abstract syntax notation-one (ASN. 1). InIEE Tutorial Colloquium on Formal Methods and Notations Applicable to Telecommunications 1992 Mar 19 (pp. 2-1). IET.
17. Kreft Ł, Botzki A, Coppens F, Vandepoele K, Van Bel M. PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. Bioinformatics. 2017 May 19;33(18):2946-7.
18. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S. CDD: specific functional annotation with the Conserved Domain Database. Nucleic acids research. 2008 Nov 4;37(suppl_1):D205-10.
19. Montesino R, Fenz S, Baluja W. SIEM-based framework for security controls automation. Information Management & Computer Security. 2012 Oct 5;20(4):248-63.
20. Zhang H, Gao S, Lercher MJ, Hu S, Chen WH. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic acids research. 2012 Jun 12;40(W1):W569-72.