ISSN 2278 - 3091
International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 5 , No.1 Pages : 09 - 13 (2016)
Special Issue of ICCEIT 2016 - Held on February 16, 2016 in The Solitaire Hotel, Bangalore, India
http://warse.org/IJATCSE/static/pdf/Issue/icceit2016sp03.pdf

Real-time Bioinformatician based on Phylogentic tool

Kumudavalli M.V, Dr. Debabrata Samanta
Department of MCA (BU), DSI, Bangalore, Karnataka, India
kumudamanju@gmail.com

ABSTRACT

Bioinformatics is an interdisciplinary area consisting of Molecular Biology and computer science. The advancements in molecular biology and biotechnology cast the computer science field into new avenues and challenges. The major and recent trend in bioinformatics is pertaining to Phylogenetic analysis. Analysis and usage of various tools is an important stage for any developmental growth towards the biological inferences that has to be drawn in order to justify the research or to obtain the curated data. One such tool is analyzed and applied on a set of data to obtain the Phylogenetic/ Evolutionary tree and other related outputs.

Key words: Phylogenetcs, sequences, Maximium Likelyhood, Alignemnt.

INTRODUCTION

Bioinformatics is a mixture of Biology, Mathematics, Computer Science and Biophysics as depicted in Fig1. As the biological data is growing exponentially the storage and analysis of data has become the major task for the computer world.

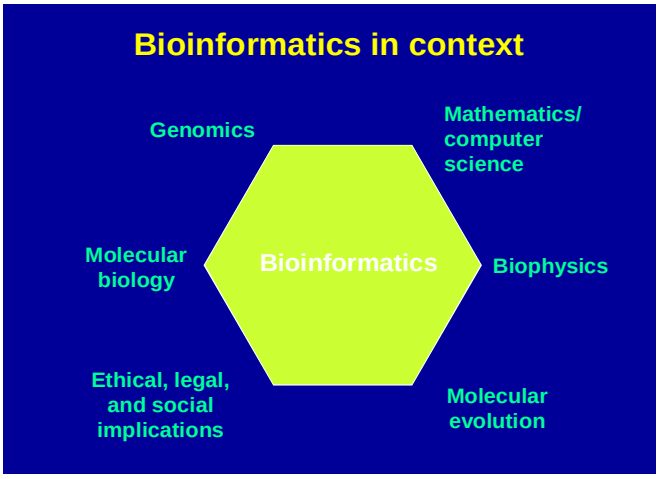

Fig1. Context of Bioinformatics

The essential steps involved in bioinformatics research in terms of computer science are storage and quantitative analysis of sequence, structure, and function of genes and their products. To this end various challenges are faced by the users viz., the methods are in flux and not fully developed, they are scattered and are heterogeneous resources. This challenge is addressed by integrating the different tools and databanks, using of web resources and navigation guides.

The scope of Bioinformatics is divided into two sub fields: tool development and tool application, for which the analysis process is essential. The basic analysis areas are: Sequence analysis, structural analysis and functional analysis. The sequence analysis involves the sequence alignment, sequence databases searching, gene and promoter finding and reconstruction of evolutionary relationships. The evolutionary relationship is generally represented as a tree called as Evolutionary tree or the Phylogenetic tree, Fig2 is an example of evolutionary tree [10].

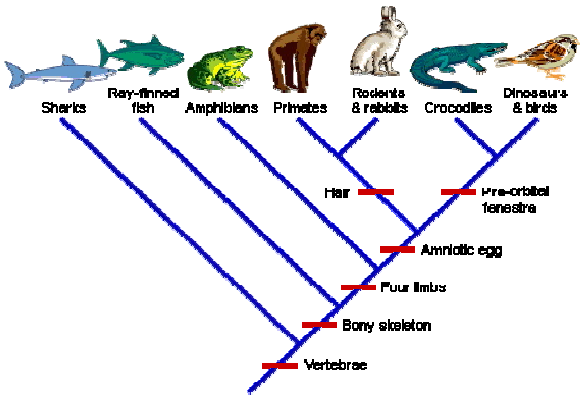

Fig2. Example of basic Evolutionary tree

A phylogeny or evolutionary tree, symbolizes the evolutionary associations among a set of organisms or groups of organisms, called **taxa** that are believed to have a common ancestor. The tips of the phylogenetic tree represent groups of descendent taxa (often species) .The internal nodes of the tree represent the common ancestors of those descendents. The tips are the present and the internal nodes are the past [1], [7]. The edge lengths in some trees

ISSN 2278 - 3091
International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 5 , No.1 Pages : 09 - 13 (2016)
Special Issue of ICCEIT 2016 - Held on February 16, 2016 in The Solitaire Hotel, Bangalore, India
http://warse.org/IJATCSE/static/pdf/Issue/icceit2016sp03.pdf

correspond to time estimates – evolutionary time. The tree is generally called as a Phylogram which is a tree in which branch lengths do represent evolutionary time; clades represent true evolutionary history [2].

Various tools are available for construction of Phylogenetic trees, but it requires the usage of many programs/tools as pre requests for the tree construction [15]. To cater to this need the present investigation is carried out to throw some light on the **Phylogeny.fr** tool.

Fig3 is a sample output of the tool used.

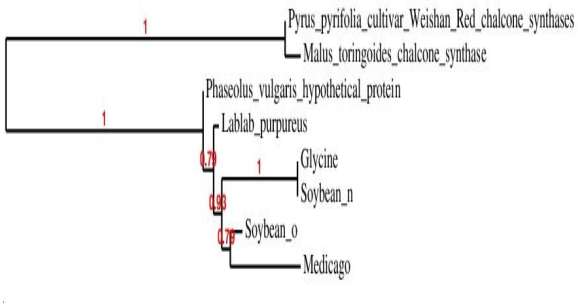

Fig 3. Sample Phylogenetic tree output using the tool.

This paper is organized as follows: Section II deals with related works, Section III discusses the need and importance of the problem, Section IV gives the data set Description, Section V Deals with the methodology, Section VI deals with the experiments and results and Section VII gives the conclusion.

RELATED WORKS

A thorough survey of the literature pertaining to the subject reveals that very spare literature is available in this direction. Some recent works include ([3], [4]). Absolutely no work is available with regard to the present work. Hence, the present investigation is carried out.

NEED AND IMPORTANCE OF THE PROBLEM

As mentioned earlier Bioinformatics finds its scope in sequence alignment and analysis. Sequence analysis is the application of Information Technologies to Molecular Biology. It deals with biological sequences, and processes them to extract considerable information that may yield new insights and strategy in the understanding of biological organisms. Various computer tools/programs are available which give suitable results to the given set of input data [9]. Usage of one such tool **Phylogeny.fr** with input data and the analysis of its various outputs are the main objectives of this research work. **Since not much work pertaining to the phylogeny.fr tool and its applications has been done, the present exploration is carried out to throw light on the qualitative as well as quantitative aspects of the problem.**

DATA SET DESCRIPTION

A set of 23 sequences belonging to plants kingdom was being selected from NCBI's Nucleotide Database using BLAST as in Table1. Later the data set is used in **Phylogeny.fr** tool to obtain the alignment and to get the Phylogenetic tree [12].

Sl.No	LOCUS/ ACCESSION	ORGANISM
1	JN830647.1	Pyrus pyrifolia
2	HQ853494.1	Malus toringoides
3	DQ286037.1	Sorbus aucuparia
4	AF400567.1	Rubus idaeus
5	HQ423171.1	chalcone synthase
6	AB201756.1	Fragaria x ananassa
7	JQ247184.1	Camellia sinensis
8	AM263200.1	Humulus lupulus
9	AJ413277.1	Rhododendron simsii
10	X94706.1	Juglans nigra
11	JN654702.1	Vaccinium corymbosum
12	JQ627646.1	Lonicera japonica
13	AB009350.1	Citrus sinensis
14	JF795272.1	Gossypium hirsutum
15	HQ127337.1	Phlox drummondii
16	EU430077.1	Senna tora
17	AY237728.1	Glycine max
18	L24517.1	Trifolium subterraneum
19	FJ705842.1	Capsicum annuum
20	AY170347.1	Arachis hypogaea
21	DQ208973.1	Cardamine maritime
22	AF112108.1	Barbarea vulgaris
23	AF144530.1	Rorippa amphibian

TABLE 1 - THE SEQUENCE DETAILS

METHODOLOGY

The first step in Phylogenetic tree construction is to retrieve the sequence of study form the database and to find the sequences which are similar to the target sequence. Therefore the BLAST program is used. BLAST (Basic Local Alignment Search Tool) is a heuristic method to find the highest scoring locally optimal alignments between a query sequence and a database. **blastn**: program is used to compares a nucleotide query sequence against a nucleotide

ISSN 2278 - 3091
International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 5 , No.1 Pages : 09 - 13 (2016)
Special Issue of ICCEIT 2016 - Held on February 16, 2016 in The Solitaire Hotel, Bangalore, India
http://warse.org/IJATCSE/static/pdf/Issue/icceit2016sp03.pdf

sequence database to get the other sequences in the table. Later the data set is used in **Phylogeny.fr** which is a Phylogenetic tool.

Phylogeny.fr has been designed to provide a high performance platform that transparently chains programs relevant to Phylogenetic analysis in a comprehensive and flexible pipeline. Although there are various tools available for the tree construction, it all depends on few other programs before the final tree construction [11]. So the philosophy of Phylogeny.fr is to support biologists with no experience in phylogeny in analyzing their data in a robust way. The Phylogeny.fr platform offers a phylogeny pipeline which can be executed through three main modes: "One click mode", "Advanced mode" and "A la carte mode". In the present work the data set consisting of 23 sequences of plant kingdom is used in "one click mode" for the sequence alignment and the reconstruction of Phylogenetic tree [13].

The "one click mode", targets users who do not wish to deal with program and constraint selection. By default, the pipeline is already set up to run and connect programs recognized for their accurateness and speed. It uses MUSCLE program for multiple alignment and PhyML for phylogeny. Finally it uses TreeDyn which offers many tree customization options compared to other tree rendering tools and especially for tree annotations, for tree drawing to reconstruct a robust Phylogenetic tree from the set of sequences [8].

The "one click mode" follows the steps as in Fig4 below:

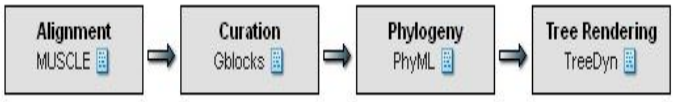

Fig 4. Work flow of one click mode

EXPERIMENTS AND RESULTS

The experimental procedure follows the steps mentioned below:

Step1: The DNA sequence of chalcone synthase was taken from Nucleotide database of NCBI.

Step 2: Remaining group of 22 DNA sequences were selected by using BLAST program as in Table 1.

Step3: The sequences are saved in FASTA format in a text file.

Step4: The data file is loaded onto the **Phylogeny.fr** "One click mode" window.

Step5: once the data is set it starts the alignment process using MUSCLE program.

Step6: It curates the aligned sequences, before the tree is constructed.

Step7: Phylogenetic tree is constructed using Maximum Likelihood tree building algorithm [5], [14].

Step8: At last the tree rendering takes place using TreeDyn. i.e it reconstructs the Phylogenetic tree using Retree program from PHYLIP package[6].

The results are obtained in various forms which can be used as required by the user. It produces an output file which contains the tree in Newick format as below:

**((((Glycine:0.10768,Trifolium:0.25157)0.7200000000:0.04504,Arachis:0.89266)
0.9760000000:0.16749,Senna:0.2274)0.9920000000:0.09086,(Lonicera:0.31156,
(Capsicum:0.63003,((((Camellia:0.21288,Rhododendr:0.18541)0.2950000000:0.03202,
Phlox:0.41986)0.7730000000:0.04038,Vaccinium:0.22157)0.8710000000:0.0673,
((((((Fragaria:0.10272,Rosa:0.03644)0.9800000000:0.05851,Rubus:0.06597)
0.9900000000:0.10667,((Malus:0.03755,Pyrus:0.01876)0.8260000000:0.01584,
Sorbus:0.05052)0.9960000000:0.1196)0.9750000000:0.08244,(Citrus:0.25986,
Gossypium:0.30507)0.5320000000:0.06773)0.8940000000:0.05744,(Juglans:0.12569,
(Cardamine:0.01974,(Rorippa:0.03373,Barbarea:0.01955)0.7110000000:0.01664)
1.0000000000:0.67389)0.9680000000:0.21205)0.2650000000:0.01546,Humulus:0.24306)
0.9250000000:0.0626)0.0000000000:0.0)0.6480000000:0.04223)0.9920000000:0.09439);**

The tree image is given in PDF, PNG and SVG format, and it is user's choice to select and store the appropriate format of the image. The output Tree of the current data set under study is given as in Fig5.

ISSN 2278 - 3091
International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 5 , No.1 Pages : 09 - 13 (2016)
Special Issue of ICCEIT 2016 - Held on February 16, 2016 in The Solitaire Hotel, Bangalore, India
http://warse.org/IJATCSE/static/pdf/Issue/icceit2016sp03.pdf

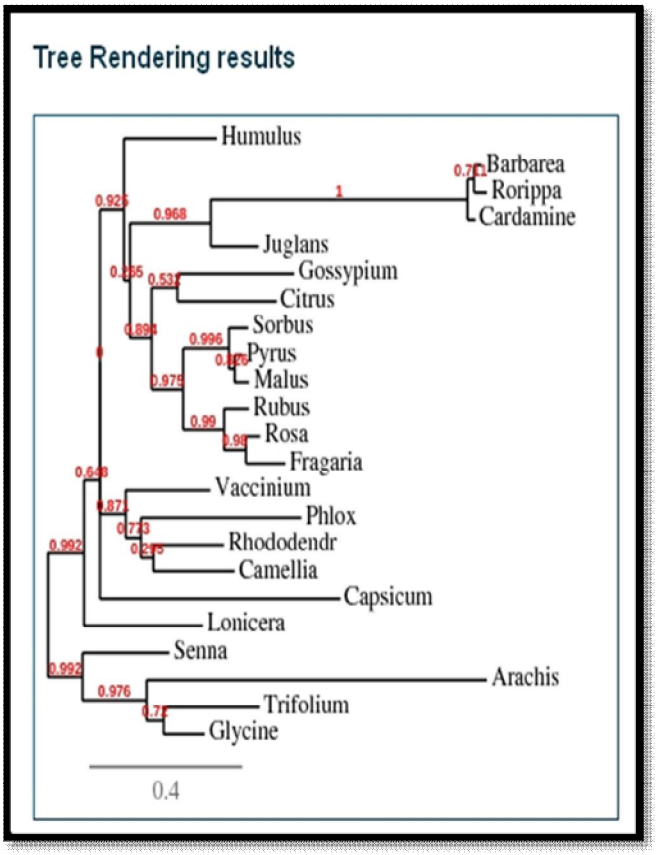

Fig 5. Phylogenetic tree (the branch length is proportional to the number of substitutions per site).

CONCLUSION

As the evolutionary/Phylogenetic tree gives the basic and important relations among the species, it is the most common problem to be addressed in field of Biology. The requirement of efficient tool for the Phylogenetic analysis is eased with the **Phylogeny.fr** tool. It is an efficient and an easy to use tool which makes the job of a user easy by providing the essential sub programs required for the tree construction as a pipeline. It is a web based service which caters to the need of Phylogenetics in a simple manner. It produces an accurate and a perfect Phylogenetic tree for the data set/Sequences provided.

ACKNOWLEDGEMENTS

One of the authors Mrs. Kumudavalli M.V acknowledges Dayananda Sagar Institutions, Bangalore, Karnataka and SCSVMV University, Kanchipuram, Tamilnadu, India for providing the facilities for carrying out the research work.

REFERENCES

[1] Teresa Przytycka., "Stability of Characters and Construction of Phylogenetic trees," Journal of Computational Biology, 14(5): 539-549, 2007

[2] Thorpe and W.J. Dickinson., "The Use of Regulatory Patterns In Constructing Phylogenies," Systematic Zoology, 37(2):97-105, 1988

[3] Dereeper A.*, Guignon V.*, Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. *Phylogeny.fr: robust phylogenetic analysis for the non-specialist.* Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W465-9. Epub 2008 Apr 19. (**PubMed**) ***: joint first authors**

[4] Srimani P. K, Kumudavalli M.V., "A Computational Analysis of the Phylogenetic Trees of Some Eukaryotes Sequences," International Journal of Current Research, Vol. 4,Issue, 05, pp. 206-210, May, 2012.

[5] Naruya Saitou., "A Theoretical Study of The Underestimation of Branch Lengths by The Maximum Parsimony Principal," Systematic Zoology, 38(1):1-6, 1989.

[6] Felsenstein J., "Inferring phylogenies from protein sequences by parsimony, distance, and likelyhood methods," Methods in Enzymology, 266:418-427,1996.- PHYLIP Package.

[7] Joshi S.P, Gupta V.S, Aggarwal R.K, Ranjekar P.K, D.S.Brar., "Genetic diversity and phylogenetic relationship as revealed by inter simple sequence repeat (ISSR) polymorphism in the genus Oryza.", Theor Appl Genet, Vol 100: Pg 1311–1320, Springer –Verlag, 2000.

[8] Davis J. P., Akella1 S., Waddell2 P. H., "Accelerating Phylogenetics Computing on the Desktop: Experiments with Executing UPGMA in Programmable Logic", Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, USA, September 1-5, 2004.

[9] Anne Kupczok, Arndt Von Haeseler, and Steffen Klaere, "An Exact Algorithm for the Geodesic Distance between Phylogenetic Trees ",Journal Of Computational Biology, Volume 15, Number 6, Pp. 577–591, 2008.

[10] Priyank Raj Katariya, Sathish S Vadhiyar, "Phylogenetic Predictions on Grids", Fifth IEEE International Conference on e-Science, 2009.

[11] Hu Yushan1,2, Luo Lei1,2, Liu Weijia1,2 and Chen Xiaoguang1, "Sequence analysis of the groEL gene and its potential application in identification of pathogenic bacteria ", African Journal of Microbiology Research Vol. 4(16), pp. 1733-1741, 18 August, 2010.

[12] Taran Granta,b,* and Arnold G. Klugec,*, "Data exploration in phylogenetic inference: Scientific, heuristic, or neither", Cladistics 19, 379–418, 2003.

[13] Manoj Giri1*, Dipti Jindal2 , Savita Kumari2, Sarla Kumari3, Devender Singh4, Jawahar Lal5 and Neena Jaggi6, "SEALI: A sequence alignment tool",Journal of Bioinformatics and Sequence Analysis, Vol. 2(3), pp. 30-35, July 2010.

[14] David L. Swofford,1,6 Peter J. Waddell,2 John P. Huelsenbeck,3 ,Peter G. Foster,1,7 Paul O. Lewis,4 and James S. Rogers5, "Bias in Phylogenetic Estimation and Its Relevance to the Choice between Parsimony and Likelihood Methods", Syst. Biol. 50(4):525–539, 2001.

[15] Nicolas Salamin,1,2 Trevor R. Hodkinson,1 and Vincent Savolainen2, "Towards Building the Tree of Life: A Simulation Study for All Angiosperm Genera", Systematic Biology. 54(2):183–196, 2005.