



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

SPGUP – Sequence analysis, Phylogenetic trees, Genome Diagrams Using Python - a Phylogenetic Tool

¹Prof.Dr. P. K. Srimani and ²Mrs. Kumudavalli M.V

¹Director, R&D Division, B.U., DSI, Bangalore, Karnataka, INDIA

²Department of MCA (BU), DSI, Bangalore, Karnataka, INDIA and Research Scholar, SCSVMV University, Kanchipuram, Tamilnadu, INDIA

Abstract: SPGUP is a software tool and a free GNU/Linux distribution developed by the authors that has the unique feature of providing a set of Phyloinformatic tools for retrieving data from NCBI, to perform Sequence analysis, Phylogenetic tree construction and to build the Genome diagrams. The operating system incorporates most commonly used Phylogenetic software, which has been pre-compiled and pre-configured, allowing the user for direct application of Phylogenetic methods in a stable Linux environment.

Keywords: Sequence, Phylogenetics, Phyloinformatic, operating system

I. INTRODUCTION

Phylogenetic methods are having exponential growth nearly in all fields of biology. As the biological data accumulation rate is increasing, the size of Phylogenetic analyses and the scope of inferences for which they are utilized have both increased enormously. To this end a diverse array of methods that have been developed by the researchers. But the availability of these methods, and the sheer number of software packages which are now regularly employed in Phylogenetic research force academicians, biologists and system administrators to spend huge amounts of time installing, configuring and maintaining them, rather than focusing on research. Phyloinformatic research, in particular, depends largely on processing capacity and highly parallel analyses that are distributed across many computers and processors. As Phylogenetic analysis has become more complex and advanced, careful calibration of operating systems across computers also becomes more complex. While this increasing need of computational power gives researchers the ability to be more creative in assaying complex and time-consuming analyses, installation and configuration of various software tools for these machines becomes highly repetitive, the process will be error-prone and time consuming. This situation or problem can be easily handled with development of a tailor made operating systems for each specific purposes of Phylogenetic analysis. Moreover, in recent days the Phylogenetic methods are experiencing rapid development; hence there is a need for a distribution specifically focusing on this area. SPGUP is one such software tool and an operating system by itself developed using Python ([1], [2]) which heads towards filling this need.

II. RELATED WORK

A thorough survey of the literature pertaining to the subject reveals that very sparse literature is available in this direction. Some recent works include ([1], [2], [3], [4], [5], [6], [7] and [8]). Absolutely no work is available with regard to the present work. Hence, the present innovative work is carried out.

III. NEED AND IMPORTANCE OF THE PROBLEM

As discussed earlier the Phylogenetic analysis has a major application in the field of Biology. Since there are very few operating systems/software packages [3] available in this direction which are aimed at generalizing the methods, and thus do not include many tools or packages essential for Phylogenetic analysis alone. SPGUP is a tool which incorporates methods which are essential for performing basic studies pertaining to sequences and Phylogenetic trees in addition it concentrates on Genome diagrams. A comparative analysis of SPGUP with other available Phylogenetic softwares is essential to ensure and highlight the easy and speedy approach used to obtain various results using SPGUP as compared to other tools. Since not much work in this area has been done, the present investigation is carried out to throw light on the qualitative aspects of the tool.

IV. DATA SET DESCRIPTION

Basically, there are two ways of data selection viz., general and through SPGUP.

i. Data set selection (General Method)

A set of 23 sequences belonging to plants kingdom which contain the protein coded gene 'chalcone synthase' was being selected from NCBI's Nucleotide Database using BLAST as in Table1. The set of sequences under

study was used in CLUSTALW ([4], [5]) to get the resulting tree. Later the tree file was used in Tree View to get the Phylogram and Rectangular cladogram.

ii. Data set selection (Using SPGUP)

The BLAST option in SPGUP tool is used to get set of (23) sequences based on the key-id of the primary sequence having ‘chalcone synthase’. Later it was used in ALIGNMNET option to produce ‘.dnd’ file which is later used to construct various forms of Phylogenetic trees.

Table I: Contains the sequence details

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 annum
20	AY170347.1	Arachis hypogaea
21	DQ208973.1	Cardamine maritime
22	AF112108.1	Barbarea vulgaris
23	AF144530.1	Rorippa amphibian

V. METHODOLOGY

The traditional methods which use the existing software tools require the following steps for Phylogenetic tree construction:

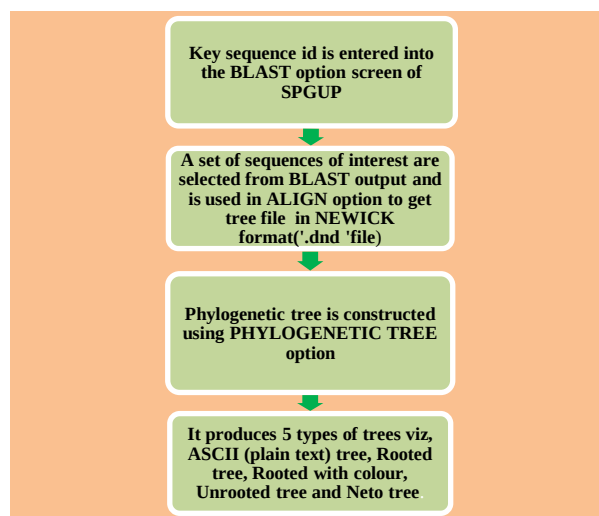
Step1: The initial sequence having ‘chalcone synthase’ is being selected form NCBI’s data base, it is then used in BLAST program of NCBI to get the set of similar sequences [4].

Step2: The set of sequences under study is pasted onto the CLUSTALW tool window for alignment. The tool also gives the Tree output in NEWICK format, and as a Phylogram.

Step3: The tree file in NEWICK format is later used in TREE VIEW software tool, which is to be pre-configured before use to get various types of pictorial representations of the tree.

This approach does not predict anything about the different tree formats viz., ASCII tree, Unrooted and Neto format tree, but in case where SPGUP tool is used, exhaustive results pertaining to the study are obtained with minimum effort and maximum efficiency. These are illustrated in the forthcoming sections.

However, in the present case only the following steps are needed for the results pertaining to Phylogenetic trees construction.



Pictorial representation of the workflow of SPGUP

VI. EXPERIMENTS AND RESULTS

In this section, the results of the present investigation using both the traditional method (Figures 1, 2 and 3) and the SPGUP tool (Figures 4, 5, 6, 7 and 8) are presented. The experiments are conducted as per the steps mentioned in section V. Figure1 shows the Phylogenetic tree as a cladogram which is obtained by using the CLUSTALW tool. A cladogram is a simple tree depicting only relationships between terminal nodes. A cladogram can also show inferred character changes [6].

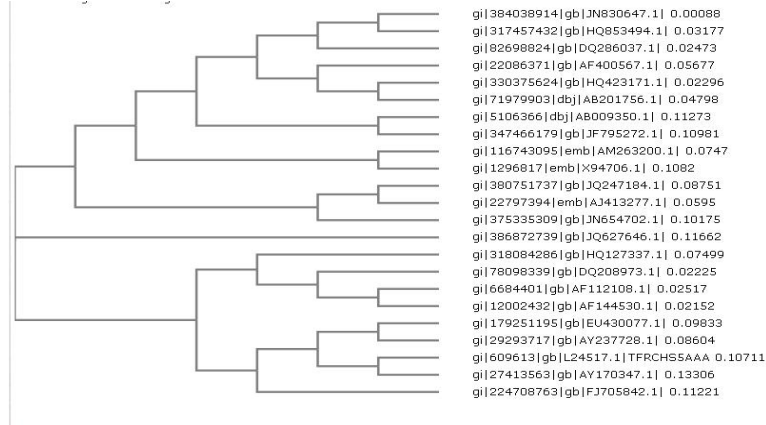


Fig. 1 Cladogram using CLUSTALW

The aligned sequences of the data set are further used in TREE VIEW tool which is exclusively used for tree drawing. It takes the input file in NEWICK format. Figures 2 and 3 represent the Phylogenetic tree in a Phylogram and Rectangular cladogram format.

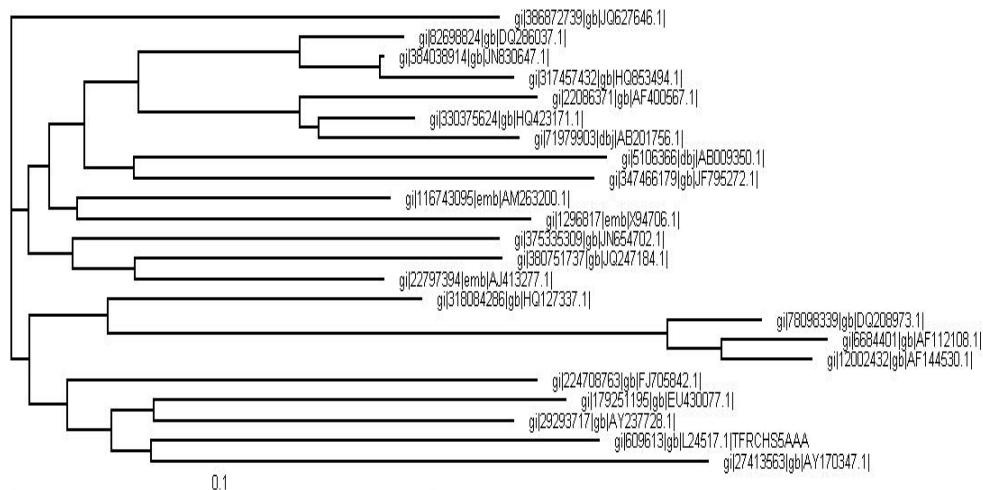


Fig.2 Phylogram Using TREE VIEW

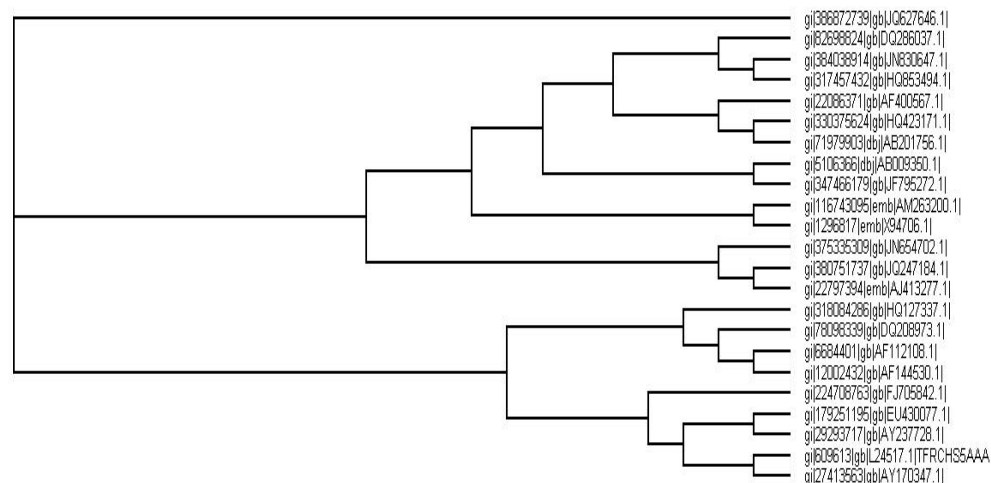


Fig. 3 Rectangular Cladogram using TREE VIEW

Following are the results obtained from SPGUP tool. The SPGUP tool allows the user to construct various forms of Phylogenetic tree with minimum effort. Figure4 represents a basic Phylogram. A Phylogram or additive tree which has additional information, in the sense that edge lengths are drawn proportional to some attributes. The evolutionary tree for the group of 23 sequences under study is as below

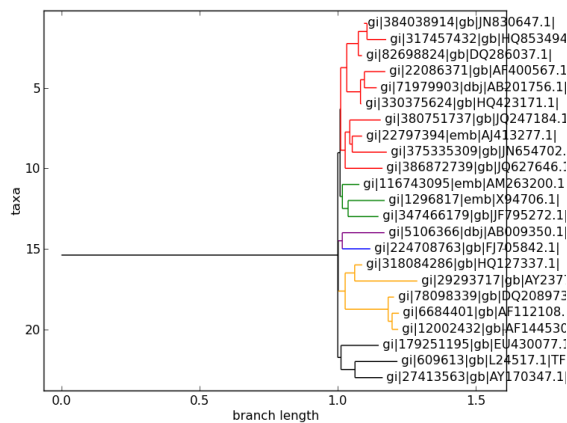


Fig. 4 ASCII tree (Phylogram) using SPGUP

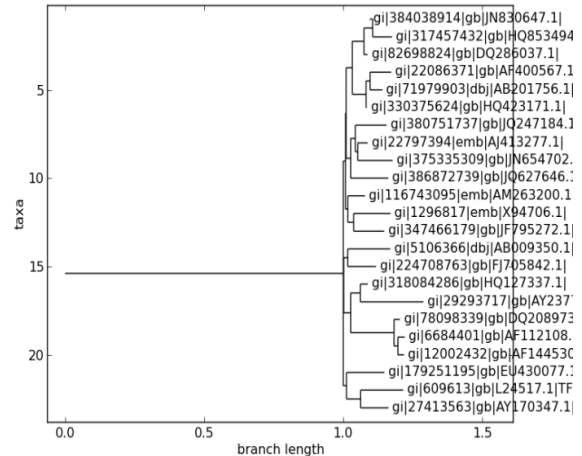


Fig.5 Rooted tree (Phylogram) with color using SPGUP

A Phylogram can be rooted/unrooted , Figure 5 presents the Phylogram which is a rooted tree drawn with colored edges representing the node to which they belong. Figures 6 and 7 depict the rooted and unrooted trees respectively in the radial format [7] using SPGUP tool.

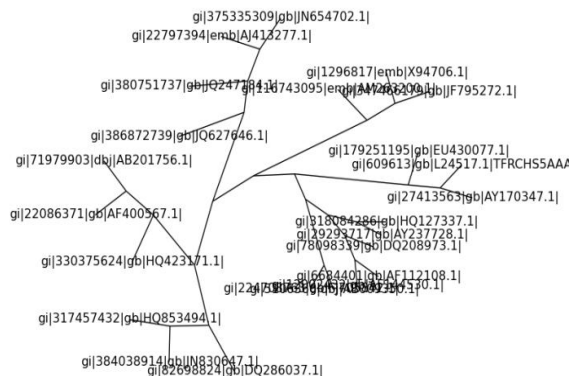


Fig. 6 Rooted tree using SPGUP

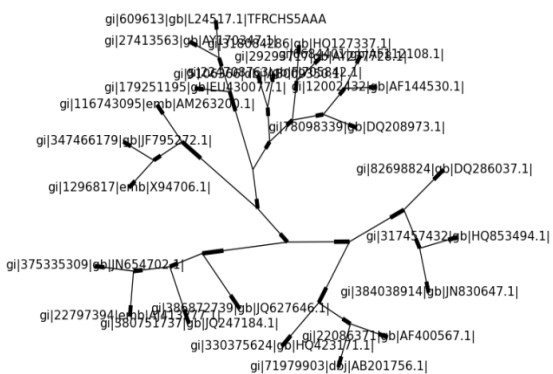


Fig. 7 Unrooted tree using SPGUP

Figure 8 is a special format of Phylogenetic tree generally drawn for a very large data set.

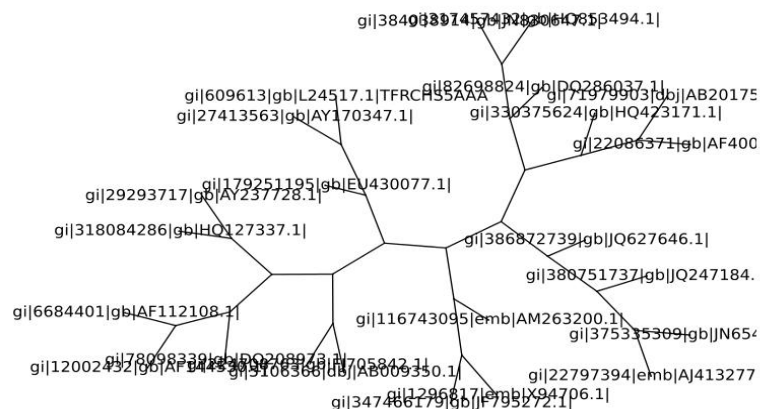


Fig. 8 Neto format tree (for very large data set) using SPGUP

VII. CONCLUSION

The present investigation with the dataset of 23 sequences was carried out to throw some light on the existing traditional methods/software tools which require the knowledge of many tools with their setup and usage as compared to the tool SPGUP developed by the authors.

In the traditional methods, the user has to follow a lot of steps in order to construct the Phylogenetic trees i.e. use of BLAST for finding the similar sequences, followed by CLUSTALW tool for aligning the sequences, in order to get the Newick format tree. Later the tree file in '.dnd' form is used in another tool called TREE VIEW for getting different forms of Phylogenetic trees. These methods take a huge computational time with dependencies on various tools, which is of concern to the user. Compared to traditional methods SPGUP is a Phyloinformatic tool developed by the authors and is a pipeline of tools required for the tree construction, which has a GUI based approach.

The special features of SPGUP tool are: (i) Different forms of trees can be constructed for the same data set. (ii) Less number of steps is involved in the computational process. (iii) It can be used for both Nucleotide (DNA, RNA) and Protein sequences. (iv) In order to increase the scope and applications an integrated approach is employed in the design of SPGUP, comprising of Python, Networkx, Graphviz, etc. (v) The SPGUP tool is first of its kind and has lot of scope for further research.

These special features have made it very user friendly for both the biologists and computer/IT professionals. The versatility of this tool is that, it has better knowledge acquisition, assimilation and dissemination.

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] <http://www.python.org>
- [2] <http://www.open-bio.org>
- [3] Stephen A. Smith^{1,*} and Casey W. Dunn², "Phyutility: a phyloinformatics tool for trees, alignments and molecular data", *BIOINFORMATICS*, APPLICATIONS NOTE Vol. 24 no. 5 2008, pages 715–716 doi:10.1093/bioinformatics/btm619
- [4] S. B. Needleman and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology*, 48:443-453, 1970.
- [5] Elodine Nedelec,¹ Thomas Moncion,^{2,3} Elisabeth Gassiat,¹ Bruno Bossard,^{2,4} Guillemette Duchateau-Nguyen,⁵ Alain Denise,^{2,5} and Michel Termier⁵. "A Pair-wise Alignment Algorithm Which Favors Clusters of Blocks". *Journal of Computational Biology*, 12, Number 1, 33-47, 2005.
- [6] Thompson J.D., Higgins, D.G., and Gibson, T.J. "Clustalw; improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucl. Acids Res.* 22, 4673- 4680, 1994.
- [7] 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.
- [8] Valdimir Makarenkov¹ and Bruno Leclerc². "Comparison of Additive Tree using Circular Orders". *Journal of Computational Biology*, 7, 731-744, 2005.