Journal of
Data Mining in Genomics & Proteomics

Patel et al., J Data Mining Genomics Proteomics 2014, 5:2
DOI: 10.4172/2153-0602.1000154

Research Article Open Access

Homology Modelling of Conserved rbcL Amino Acid Sequences in Leguminosae Family

Sagar S Patel*, Megha B Vaidya and Dipti B Shah

G. H. Patel Post Graduate Department of Computer Science and Technology, Sardar Patel University, Vallabh Vidyanagar, Gujarat-388120, India

Abstract

This study is focus on Homology modelling of few Leguminosae family species which are found in Gujarat state, INDIA. There are three subfamilies of Leguminosae family which are Fabaceae (Papilionaceae), Caesalpiniaceae and Mimosaceae. Multiple sequence alignment carried out of few species' rbcL protein sequences in each subfamily and conserved amino acid considered for homology modelling. Evolutionarily related proteins have similar sequences and naturally occurring homologous proteins have similar protein structure. It has been shown that three-dimensional protein structure is evolutionarily more conserved than would be expected on the basis of sequence conservation alone; we found that there are few amino acids which are common with same base pairs in each sub-family even though they are from different genus. There is no Protein structure available of conserved amino acids in PDB database of our study so we did homology modelling of three *rbcL* protein sequences (one from each sub family) which are found conserved in Multiple sequence alignment and structure validation with Ramachandran Plot was carried out and CASTp server was used to find out active sites in predicted protein structure and finally function of each predicted protein reported after this homology modeling of few conserved rbcL amino acid sequences in Leguminosae family.

Keywords: Homology modelling; Bioinformatics; Leguminosae family; rbcL

Introduction

Leguminosae family

Leguminosae family contains species of Plants, Herbs, Shrubs, and Trees. Legumes are used as crops, forages and green manures; they also synthesize a wide range of natural products such as flavours, drugs, poisons and dyes. Legumes are able to convert atmospheric nitrogen into nitrogenous compounds useful to plants [1] This is achieved by the presence of root nodules containing bacteria of the genus Rhizobium. These bacteria have a symbiotic relationship with Legumes, fixing free nitrogen for the plants; in return legumes supply the bacteria with a source of fixed carbon produced by photosynthesis. This enables many legumes to survive and compete effectively in nitrogen poor conditions [2,3]. Leguminosae family is further classified into three subfamilies; 1. Fabaceae (Papilionaceae), Caesalpiniaceae and 3. Mimosaceae.

rbcL gene

The most common gene used for plant phylogenetic analyses is the plastid-encoded rbcL gene. This single copy gene is approximately 1430 base pairs in length and is free from length mutations except at the far 3' end. It has fairly conservative rate of evolution. The function of the rbcL gene is to code for the large subunit of ribulose 1, 5 bisphosphate carboxylase/oxygenase (RUBISCO or RuBP Case) [4].

Protein structure

Recent genome sequencing projects have provided massive amount of data, however, many of these genomes are still not fully annotated and consist of genes/proteins with unknown function and structure. This is due to several limitations, such as the cost and time required for experimental approaches [5]. An alternative to laboratory based methods is a bioinformatics approach that utilizes algorithms and databases to estimate protein function. As these algorithms and databases are based on experimental results, they can be an effective means to perform functional and structural annotation of hypothetical proteins. Structures are more evolutionary conserved than sequence; therefore, analysis of three-dimensional (3D) structures holds great potential. Our present study describes the three 3D models of rbcL protein sequences which found conserved in multiple sequence alignment and further three protein structure predicted through homology modelling. In addition sequence and structural analysis and functional annotation were also done [6].

Methodology

In current research, we have considered around 266 species which are found in Gujarat state of India [7,8]. Further we searched each species in NCBI database and finally found around 149 species' information like DNA, Protein and other useful information of leguminosae family [9]. We have only considered rbcL protein sequences for analysis. For calculating physio-chemical properties, Prot Param was used; Various parameters computed by ProtParam included the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY) [7] Secondary structure was also predicted (helix, sheets, and coils) by using PSI Pred [8].

Homology modelling

Homology modelling approach was used to determine the 3D structure of three rbcL conserved protein sequences. BLASTP by Altschul et al. [9] search with default parameters were performed against the Brookhaven Protein Data Bank (PDB) to find suitable templates for homology modelling. For Fabaceae (Papilionaceae) 1RLD; for Caesalpiniaceae 1WDD and for Mimosaceae 1EJ7 were considered as

***Corresponding author:** Sagar Patel, G. H. Patel, Department of Computer Science and Technology, Sardar Patel University, Gujarat-388120, India, Tel: 02692-226802; E-mail: sgr308@gmail.com

Received March 20, 2014; **Accepted** April 26, 2014; **Published** April 29, 2014

Citation: Patel SS, Vaidya MB, Shah DB (2014) Homology Modelling of Conserved rbcL Amino Acid Sequences in Leguminosae Family. J Data Mining Genomics Proteomics 5: 154. doi:10.4172/2153-0602.1000154

the best templates for Homology modelling. Later SPDBV was used for homology model construction.

Protein structure validation

SPDBV generated three structures, which were further validated by using Structural Analysis and Verification Server (SAVS) [10-12] and each three structure were validated in Ramachandran Plot.

Active site prediction

The PDB file constructed was then used for finding the cavities in the protein and for this Computed Atlas of Surface Topography of proteins (CASTp) server was used. CASTp provides an online resource for locating, delineating and measuring concave surface regions on three-dimensional structures of proteins.

The pipeline for the followed methodology is as represented in Chart 1.

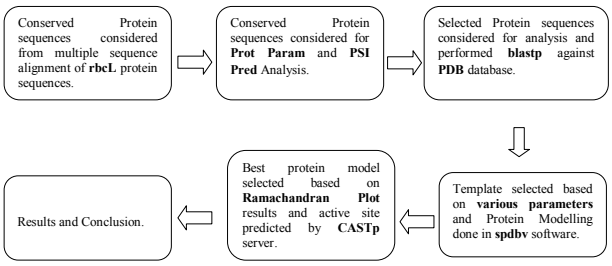

Chart 1: Pipeline for modelling and characterization of the unknown protein.

Results and Discussion

The present study focused on sequence and structural analysis of rbcL protein sequences which are found conserved in Leguminosae Family's subfamilies; for *Fabaceae* subfamily 38%, for *Caesalpiniaceae* 60% and for *Mimosaceae* 54% species were found which had conserved sequences as shown in Table 1. Prot Param was further used to analyze different physiochemical properties from the amino acid sequence which are listed in Table 2.

Results of Prot Param tools shows that protein of Fabaceae subfamily is unstable but stable protein was found in rest of subfamily. While estimated half-life result of Mimosaceae was found very less compare to other two sub-family as shown in Table 2.

Secondary structure analysis was performed using PSI Pred and the three rbcL protein were predicted to contain several helices, coil along with beta sheets as shown in Figures 1a-1c.

Homology modeling and protein structure validation

Homology or comparative modelling is one of the most common structure prediction methods in structural genomics and proteomics. Numerous online servers and tools have become available for homology or comparative modelling of proteins in past years [13]. Despite minimal modifications, one initial step that is common in all modelling tools and servers is to find the best matching template by

Sub family	rbcL protein sequences
Fabaceae	ASKWSPELAAACEVWK
Caesalpiniaceae	SVGFKAGVKDYK
Mimosaceae	RGGLDFTKDDENVNSQPFMR

Table 1: Information of conserved rbcL protein sequences considered for Homology modelling.

Physio-Chemical Property	Fabaceae	Caesalpiniaceae	Mimosaceae
Molecular weight	1776.0 Daltons	1298.5 Daltons	2326.5 Daltons
Theoretical pI	6.18	9.52	4.68
Molecular Formula	$C_{81}H_{122}N_{20}O_{23}S_1$	$C_{60}H_{95}N_{15}O_{17}$	$C_{98}H_{152}N_{30}O_{34}S_1$
Instability index	75.16	-12.83	32.77
Aliphatic index	67.50	56.67	34.00
Estimated half-life	4.4 hours (mammalian reticulocytes, in vitro).	1.9 hours (mammalian reticulocytes, in vitro).	1 hours (mammalian reticulocytes, in vitro).
	>20 hours (yeast, in vivo).	>20 hours (yeast, in vivo).	2 min (yeast, in vivo).
	>10 hours (Escherichia coli, in vivo).	>10 hours (Escherichia coli, in vivo).	2 min (Escherichia coli, in vivo).
Grand average of hydropathicity (GRAVY)	-0.131	-0.425	-1.290
Classification of Protein	Unstable	Stable	Stable

Table 2: Result of Physio-chemical Properties as calculated by Prot Param tool.

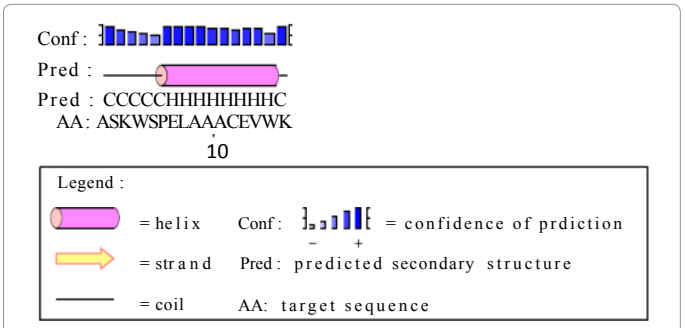

Figure 1a: PSIPred result of *Fabaceace* subfamily.

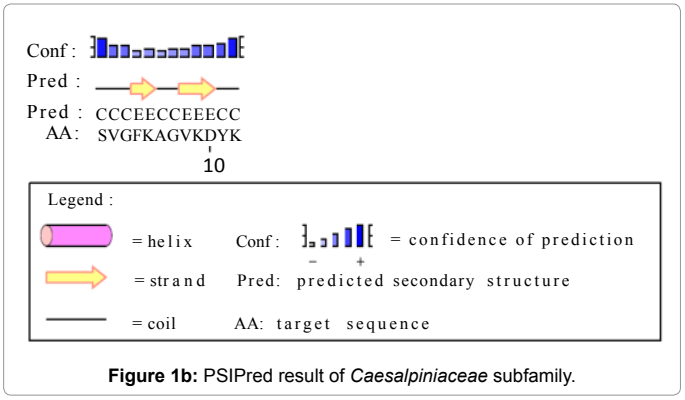

Figure 1b: PSIPred result of *Caesalpiniaceae* subfamily.

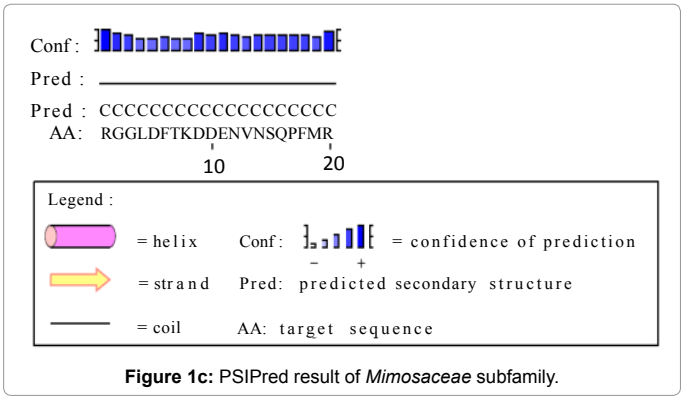

Figure 1c: PSIPred result of *Mimosaceae* subfamily.

performing a sequence homology search with BLASTP [14]. Templates are experimentally determined 3D structures of proteins that share sequence similarity with the query sequence. The template sequence and the protein sequence whose structure is to be determined are aligned using multiple sequence alignment algorithms [15]. A well-defined alignment is very important for the prediction of a reliable 3D structure. BLASTP search was performed for each protein sequence against the PDB to identify templates for homology modelling. Then the query sequence and template ID were given as input for homology modelling using SPDBV. It generated three predicted protein Models which are shown in Figures 2a-2c. From the models retrieved, the selected model along with Ramachandran plot is shown in Figures 3a-3c respectively [16-24]. The final model was selected by checking

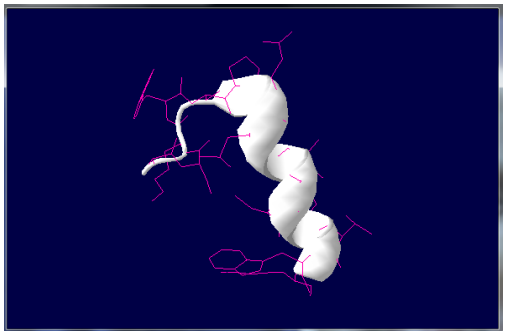

Figure 2a: Protein model of *Fabaceace* subfamily.

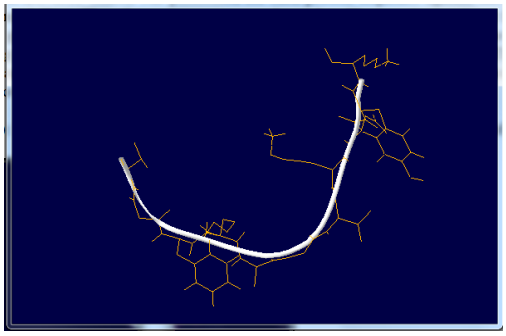

Figure 2b: Protein model of *Caesalpiniaceae* subfamily.

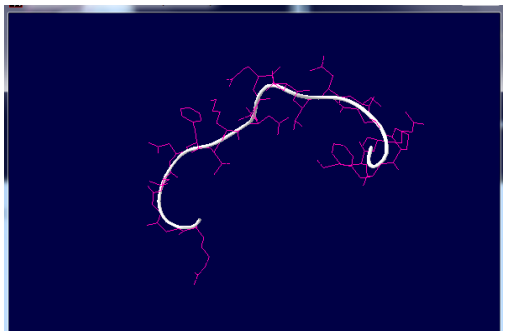

Figure 2c: Protein model of *Mimosaceae* subfamily.

Subfamily	Core region (in %)	Allowed region (in %)	Disallowed region (in %)	Bad Contacts
Fabaceace	100.0	0.0	0.0	0.0
Caesalpiniaceae	100.0	0.0	0.0	0.0
Mimosaceae	80.0	20.0	0.0	0.0

Table 3: Information of Ramachandran Plot.

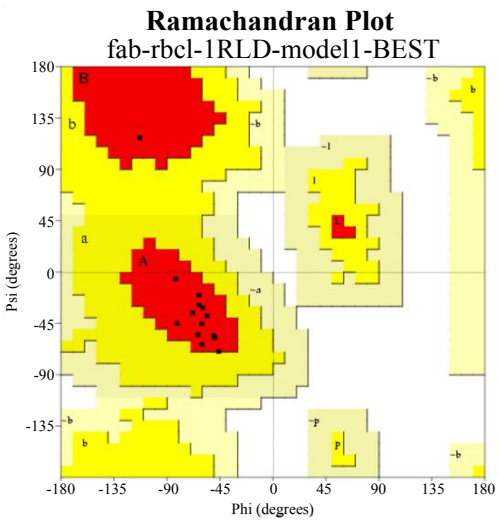

Figure 3a: Ramachandran Plot of *Fabaceace* subfamily.

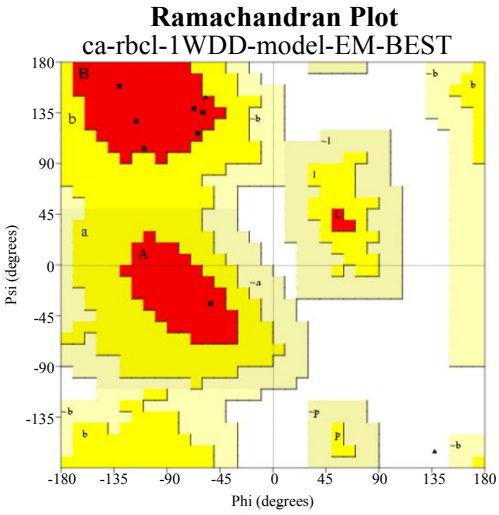

Figure 3b: Ramachandran Plot of *Caesalpiniaceae* subfamily.

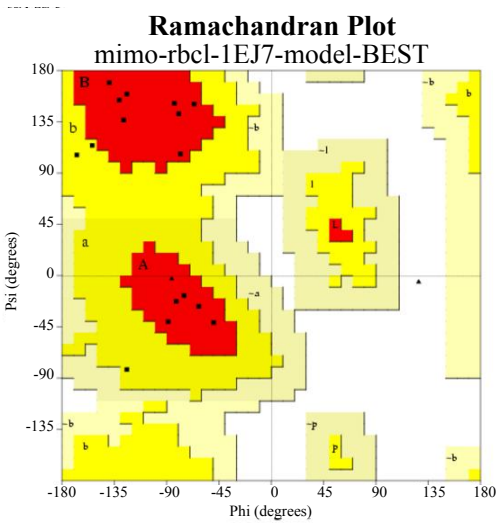

Figure 3c: Ramachandran Plot of *Mimosaceae* subfamily.

various parameters and these are shown in Table 3. These parameters included percentage of amino acids in core, allowed and disallowed regions along with no of bad contacts.

Active site prediction

Active site signifies the functional region of the protein. During the active site prediction with the help of CASTp, it was observed that few pockets were predicted in *Caesalpiniaceae* and *Mimosaceae* subfamily protein structure but no pocket found in *Fabaceace* subfamily protein structure. Some of the predicted pockets are as shown in Figures 4a-4c.

Conclusion

We have used homology modelling approach to propose the 3D structure and possible functions for the conserved rbcL protein sequences which are found in Leguminosae Family. The function of protein can be understood better by its structure and structure of rbcL protein is already known so the function of these fragments of conserved sequences are confirmed by taking following templates; for Fabaceae (*Papilionaceae*) subfamily, 1RLD; for *Caesalpiniaceae* subfamily, 1WDD and for *Mimosaceae* subfamily, 1EJ7 for Homology modelling. Later SPDBV was used for homology model construction. With the help of above findings we found that each conserved protein involved in important function like in *Caesalpiniaceae* subfamily predicted protein structure has site which is heterodimer interface [polypeptide binding] and disulfide bond found within that particular structure. In *Mimosaceae* subfamily, predicted protein structure has few sites which are heterodimer interface [polypeptide binding], active catalytic residue site and metal binding site [ion binding] and no active site found in *Fabaceae* subfamily predicted protein structure. So, these particular predicted protein structures has many important feature as described above and found common in selected species of study in each subfamily of Leguminosae family and these protein sequences can be used for classification of Leguminosae Family species as protein sequences are found conserved in each subfamily. So, if your protein sequence has one of the conserved protein sequences as described in this study then it might be fall within that particular subfamily of Leguminosae Family.

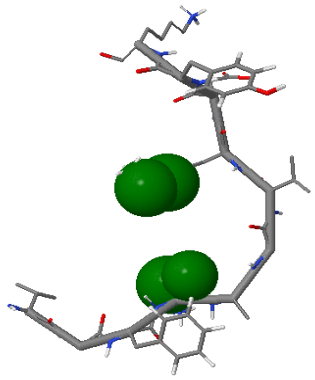

Figure 4a: Result of CASTp server of *Caesalpiniaceae* subfamily.

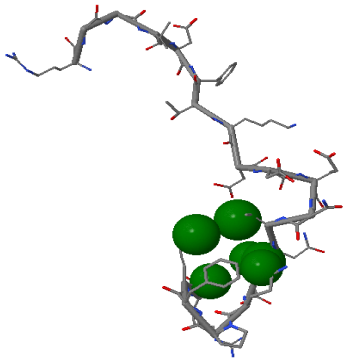

Figure 4b: Result of CASTp server of *Mimosaceae* subfamily.

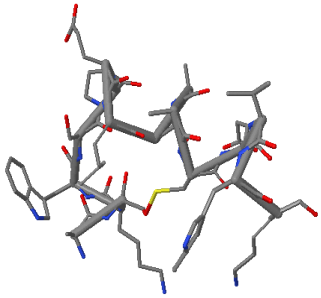

Figure 4c: Result of CASTp server of *Fabaceace* subfamily.

Acknowledgement

We are heartily thankful to Prof. (Dr.) P.V. Virparia, Director, GDCST, Sardar Patel University, Vallabh Vidyanagar, for providing us facilities for the research work. We are also thankful to DST-PURSE program and Center for Interdisciplinary Studies in Science and Technology (CISST), Sardar Patel University, Vallabh Vidyanagar, Gujarat (India) for providing financial assistance in the form of fellowship.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215: 403-4l0.
2. Binkowski TA, Naghibzadeh S, Liang J (2003) CASTp: Computed Atlas of Surface Topography of proteins. Nucleic Acids Res 31: 3352-3355.
3. Butt AM, Khan IB, Haq F, Tong Y (2011) De novo structural modeling and computational sequence analysis of a bacteriocin protein isolated from Rhizobium leguminosarumbv. viciae strain LC-3. African Journal of Biotechnology 10: 38.
4. Tandon G, Sharma R, Mishra AK, Chandrasekharan H (2013) Structural and Functional Annotation of Uncharacterized Protein in *Triticumaestivum*. International Journal of Bioinformatics and Biological Science 1: 209-219.
5. Shah (1978) Flora of Gujarat State. Publ. by Sardar Patel University, Vallabh Vidyanagar, Anand, India.
6. Oza GM, Kishore SR (2006) Biodiversity of Gujarat Forest Trees.PublBy INSONA, Vadodara, India.
7. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, et al. (2005) Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (edn): The Proteomics Protocols Handbook, Humana Press 571-607.
8. Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 23: 272.
9. Hayashi K, Kawano S (2000) Molecular systematics of Lilium and allied genera (Liliaceae): phylogenetic relationships among Lilium and related genera based on the rbcL and matK gene sequence data. Plant Species Biology 15: 73-93.
10. Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with threedimensional profiles. Nature 356: 83-85.
11. Wojciechowski MF, Lavin M, Sanderson MJ (2004) A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. Am J Bot 91: 1846-1862.

12. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404-405.

13. Laskoswki RA, MacArthur MW, Moss DS, Thorton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystl 26: 283-291.

14. http://www.amazon.com/Leguminous-Trees-Anand-District-Bioinformatics/dp/3843369828

15. http://saspjournals.com/wp-content/uploads/2014/03/SJAVS-1250-57.pdf

16. Patel Sagar, Panchal Hetal Kumar (2014) Bioinformatics Information of Leguminosae Family in Gujarat State. International Journal of Agriculture, Environment and Biotechnology 7: 11-15.

17. Sagar Patel, Hetalkumar Panchal (2014) Evolutionary studies of few species belonging to Leguminosae family based on RBCL gene. Discovery 9: 38-50.

18. Sagar Patel, Panchal H (2013) Leguminobase: A Tool To Get Information of Some Leguminosae Family Members From Ncbi Database in Journal of Advanced Bioinformatics Applications and Research 4: 54-59.

19. Sagar Patel, Panchal H, Anjaria K (2012) DNA Sequence analysis by ORF FINDER & GENOMATIX Tool: Bioinformatics Analysis of some tree species of Leguminosae Family. 922- 926.

20. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6470264&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6470264

21. Sagar Patel, Panchal H, Smart J, Anjaria K (2013) Distribution of Leguminosae family members in Gujarat State of India: Bioinformatics Approach in International Journal of Computer Science and Management Research 2: 2184-2189.

22. Sagar Patel, PanchalH, Smart J, Anjaria K (2013) Species Information Retrieval Tool: A Bioinformatics tool for Leguminosae family in International Journal of Bioinformatics and Biological Science 2: 187-194.

23. http://www.en.wikipedia.org

24. http://www.theplantlist.org/browse/A/Leguminosae