International Journal of Biology Research
ISSN: 2455-6548, Impact Factor: RJIF 5.22
www.biologyjournal.in
Volume 2; Issue 3; July 2017; Page No. 77-81

Structural and functional annotation of streptococcus pneumonia hypothetical protein – *In silico* Approach

[*1] S Sugunakala, [2] T Akilandeswari, [3] TR Barath Kumar, [4] S Subasri, [5] N Nivetha, [6] R Akiladevi

[*1] Assistant Professor and Head, Department of Bioinformatics, A.V.C. College (Autonomous), Mannampandal, Mayiladuthurai, Tamil Nadu, India

[2, 3, 4, 5] Department of Biotechnology, A.V.C. College (Autonomous), Mannampandal, Mayiladuthurai, Tamil Nadu, India.

Abstract
The pathogenic and resistant nature of *Streptococcus pneumonia* causing pneumococcal infections can be comprehended by their structurally and functionally annotated sequences. However, many proteins of this organism are not annotated so far and are categorized as hypothetical proteins that means, the *in - vivo* functions for that proteins are yet to be built up. These hypothetical proteins may play a vital role in infectious process and be considered as a drug target. Identification of a good drug target is an important step in drug discovery process. The main aim of this work is to annotate the hypothetical protein in *S. pneumonia* using bioinformatics tools. Initially, the organism specificity and sub cellular localization were analyzed. Then, the domain analysis was carried out which showed the presence of glycosyl dehydrogenase domain in this hypothetical protein. Further, the functional annotation analysis was also carried out which showed that this hypothetical protein plays a role in human - pathogen interaction pathway namely glycan degradation pathway. Since there is no structural evidence for this protein, efforts were made to analyze and to predict the primary, secondary and tertiary structure. Then the predicted tertiary structure was validated and submitted to protein modeling database. Further, the phylogenetic analysis revealed that there was a significant sequential level of conservation among the query hypothetical protein sequence and the template structural sequences. Finally, active site analysis was carried out in which the amino acid residues and their positions were also identified. By knowing the nature of amino acid residues and area of the active sites novel inhibitors may be designed and used to inhibit the host - pathogen interaction occur through glycan degradation pathway and may results in reduced rate of occurrence of pneumococcal infections.

Keywords: streptococcus pneumonia; glycan - degradation pathway; cast p; *In – silico* structure prediction; PMDB

Introduction
In 1881, *Streptococcus pneumoniae* was isolated and identified as the leading cause of many bacterial infections viz bacterial meningitis, bacterial pneumoniae and upper respiratory tract infections [1, 2, 3]. Penicillin was the drug of choice to treat pneumonia. Since 1995, a gradual raise in the development of penicillin resistant strain was documented in India as well as throughout the world [4, 5, 6, 7]. The management of pneumococcal infections caused by these resistant strains was effective with emergence of alternative treatment therapy and it may be achieved by better understanding of the molecular mechanism involved in the host – pathogen interactions. In general, the virulence of *S. pneumoniae* depends upon the activity on the host cell surface carbohydrates. An experimental study also revealed that the genomes of pneumococcus having the potential to process N – linked sugars with an elimination of terminal mannose residues from the host organism and they concluded that the enzymes involved in this reaction play an important role in host pathogen interaction [8]. This may be considered as a potential target to treat pneumococcal infections.

In this study, we identified a novel hypothetical protein WP_000780623 from pathological *Streptococcus pneumonia* - strain SMRU 1923 by pathway analysis using KASS tool from KEGG. This protein involved in the glycan degradation pathway and thus facilitates the host pathogen interaction. Functional and abundant form of glycosylated proteins present in the human being tissues and cells facilitate initial contact for colonization and invasion of pathogenic bacteria. After invasion in to the human host, the pathogenic bacteria process the complex form of N - linked glycans which assists the infection process. At present there are no known experimental evidences supporting structural and functional characteristics which motivates us to carry out an *in - silico* study of this hypothetical protein.

Materials and Methods
Target identification and retrieval
In NCBI database (www. ncbi.nlm.nih.gov), 7326 strains of *Streptococcus pneumonia* are available. Among these, strain SMRU 1923 is a recently sequenced pathogenic organism. The whole genome of this strain consists of 3576 genes and 2848 protein sequences in which 407 proteins are classified as hypothetical proteins. An important characteristic for a good drug target is that their uniqueness i.e. it should be an organism specific and not be present in Homo sapiens. In this work, the protein sequence of WP_000780623 was identified as organism specific hypothetical protein by Blast P analysis against Homo sapiens database. During this similarity search parameters namely threshold for expectation value of 0.005, and minimum bit score value of 100 were used.

Sequential analysis of *S. pneumoniae* hypothetical protein
Functional annotation of genome may be made effectively by knowing the sub cellular localization of proteins in organism [9]. CELLO v. 2.5 predictor (http://cello.life.nctu.edu.tw/) was used to predict the localization of hypothetical protein [10]. The *S. pneumoniae* hypothetical protein was searched against

46 bacterial (both gram positive and gram negative) organisms using Database of Essential Genes (DEG available through the URL: (http://tubic.tju.edu.cn/deg/) [11]. Cutoff of 10^{-05} and a minimum bit score of 100 were used as parameters. Similarly, the family and domain analysis were also carried out by using Proteins Families Database (Pfam) [12] & NCBI Conserved Domains Database (NCBI-CDD) [13] respectively. Finally, functional involvement of this hypothetical protein in metabolism was also analyzed by using KASS automatic annotation server [14].

Structure analysis and prediction of active sites of hypothetical protein in *S. pneumoniae*

The primary and secondary structure analysis was performed by using Protparam [15] and Sopma tools [16] from Expasy server (http://www.expasy.org/). The tertiary structure for hypothetical protein was predicted by homology modeling using Protein structure prediction (ps) 2 – v2 web server [17]. (http://ps2.life.nctu.edu.tw/). This predicted protein structure was aligned with the template structure using flexible structure alignment by chaining aligned fragment pairs allowing twists (FATCAT tool) available through the URL: http://fatcat.burnham.org/ [18]. Then, the predicted structure was optimized by using Chiron tool [19] and then validated by using RAMPAGE tool (http: // mordred.bioc.cam.ac.uk / ~rapper / rampage.php). The validated model was deposited in protein modeling database (PMDB) [20] through the URL: (https://bioinformatics.cineca.it/PMDB/). Finally, the active and binding sites were predicted by using CSAT P server [21, 22, 23] (http://sts.bioe.uic.edu/castp/).

Analyzing Phylogenetic relationship between query proteins and template structural sequences

While predicting the tertiary structure for this hypothetical protein using (ps) ^{2} – v2 web server, offers 10 different structural proteins with different similarity score. Multiple sequence alignment was performed by Clustal W and by using Mega software 7.0 software [24], attempts were made to analyze the evolutionary relationship among these proteins along with the *S. pneumoniae* hypothetical protein. The evolutionary history was inferred using the Neighbor-Joining method [25]. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed [26]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches [26]. The evolutionary distances were computed using the number of differences method [27] and are in the units of the number of amino acid differences per sequence. The analysis involved 10 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 404 positions in the final dataset.

Result and Discussion

From the Blast P analysis, it is observed that the hypothetical protein was not producing significant similarity with *Homo sapiens* which explains its uniqueness on that organism and may serve as a good therapeutic drug target [28]. Sub cellular identification analysis reveals that this hypothetical protein may be localized in cytoplasmic region and this localization identification offers knowledge about the mechanism of involvement of proteins in disease state and will help to discover some novel drugs [29]. Further, analysis of predicting essentiality using DEG revealed that it is a non essential protein. The domain analysis by Pfam & CDD databases resulted that this hypothetical protein consists of Glycosyl hydrolase family and it is found to be 239 amino acids length which are in the region of 3 - 242. From the KASS server, it is noticed that this hypothetical protein may be involved in the other "glycan degradation pathway". The primary and secondary structural analysis reveals that this hypothetical protein consists of 764 amino acids, mol. weight of 87417.31, Isoelectric point of 5.78, instability index of 37.18, negative GRAVY index of -0.348, and. These values suggest that this protein is (+) vely charged, stable, hydrophilic and soluble protein. Presence of high percent of leucine amino acid indicates that the regions may favor helices. Similarly, the secondary structure analysis reveals that the *S. pneumoniae* hypothetical protein consist of 39.40%, 19.76%, 8.38% and 32.46% of helices, strands, beta turns and coils respectively. The highest percentage contribution of helices favors the protein to fold flexibly and may increase the protein interactions. The process of tertiary structure prediction starts with template identification. Results from template searching process show that the crystal structure of a putative glycoside hydrolase family protein from Bacillus halodurans (PDB ID: 2RDY) produce maximum sequence similarity of 75.21%. Using this PDB structure as a template structure, tertiary structure for *S. pneumoniae* hypothetical protein was modeled. The structure alignment of predicted Vs template structure results show that the two structures are significantly similar with p – value of 0.00^{e+00} and have 727 equivalent positions with an RMSD of 0.77 without twists which proves the reliability of the predicted 3D structure (Fig. 1). Initial validation with RAMPAGE tool results that 90.6% % residues were in favorable regions and 2.6% residues were in the allowed regions. Efforts were made to optimize the residues which are in the allowed regions, the final validation of predicted structure show that 87.4 & 10.6% residues were in favorable and allowed regions respectively. Then the validated model (Fig. 2) was deposited in the Protein Model Database (PMDB) and assigned the PMDB ID as PM 0080983.

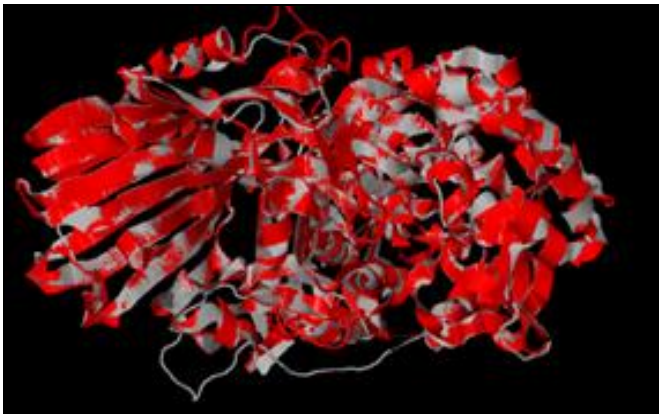

Fig 1: 3D Structure alignment of predicted structure (in grey color) and its template structure PDB ID: 1iysA (in red color).

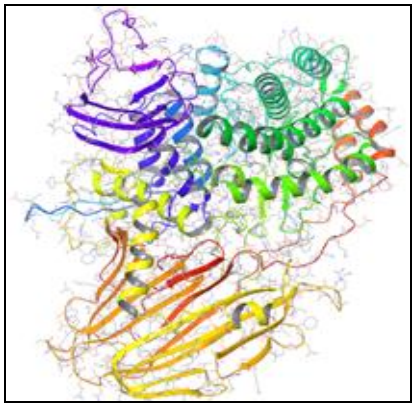

Fig. 2: 3D Structure of target protein predicted by using (ps) [2] – v2 web server

Binding site prediction

Totally 179 pockets were found in the modeled structure. Based on the domain analysis results, it was found that, 179[th] pocket is having the maximum size of area and volume and there were found to be 743.8 & 1176.6 cm^3. The amino acid residues and their positions were also given in Table 1and Fig.3.

Table 1: Results of binding pocket analysis

Pocket Number	Amino acid residues and their positions
179	Trp (430), Tyr (433), Leu (434), Gln (437), Asp (505), Ser (506), Ile (508), Gly (509), Lys (512), Gln (513), Tyr (568), Asn (569), Glu (570), Asp (572), His (574), Lys (575), Ser (682), His (683), His (684), Asn (685), Trp (686), Leu (687), Arg (708), Gly (709), Tyr (711), Gln (734), Lys (735)

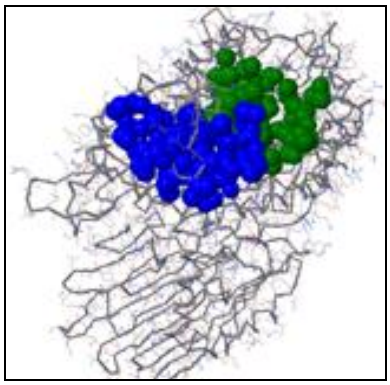

Fig 3: Prediction of Binding site

Phylogenetic analysis

The phylogenetic tree displayed highest degree of similarity between the studied hypothetical protein and its related proteins for homology modeling obtained by protein blast from NCBI. All of them clustering with the other bacterial species with bootstrapping confidence level more than 45%. The corresponding branch lengths also reveal the closed similarity between studied hypothetical protein from *Streptococcus pneumonia* with glycoside hydrolase of Bacillus halodurans. The phylogenetic tree clearly discriminates the other homologous proteins forming separate clades from the other bacterial species but structurally lower similarity (Fig. 4).

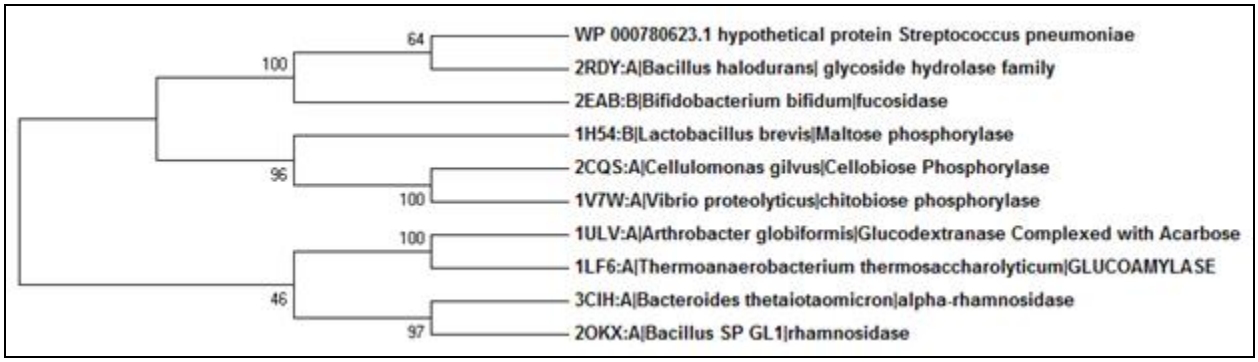

Fig 4: Evolutionary relationships of Homologous proteins for strtucture prediction.

Conclusion

In this study, by using Bioinformatics approach, 3D structure for a hypothetical protein WP_000780623 of *Streptococcus pneumoniae* was modeled and possible functional annotations were also predicted. The diseases caused by *S. pneumoniae* were generally treated with beta lactam antibiotics like penicillin. Due to the rapid development of resistant property of *S. pneumoniae* against antibiotics, there will be an emergence of alternative therapy to treat pneumococcal infections effectively. From this functional identification study, we noticed that this hypothetical protein possesses glycosyl transferase activity and it involves in glycan degradation pathway. Degradation of glycans into simple sugars are the major food source for the host adopted pathogenic pneumococcal organism with which it will have the stability and strongly invade into the human host to produce diseases like meningitis. The invasion as well as the stability will be destroyed if we inhibit the protein involved in glycan degradation process. To develop effective inhibitors, the three dimensional structure of the protein is essential. Since there is a lack of experimental 3D structure for this protein, in this work, the 3D structure was modeled by using 2RDY – chain A (crystal structure of a putative glycoside hydrolase family protein from Bacillus halodurans) as template structure. From the phylogenetic analysis it was inferred that the tertiary structure for the *S. pneumoniae* hypothetical protein was modeled with both the structurally and sequentially closely related template model. Further, the superimposition of modeled structure along with template structure also gave the minimum RMSD value which revealed that the features of modeled structure can be considered effectively. Optimization was also carried out and finally, the modeled structure was deposited in the PMDB database. The identified structural insights of binding site of modeled protein may help the medicinal chemists to design an appropriate drug compound to treat infections caused by *S. pneumonia* effectively.

References

1. Fass RJ. Aetiology and treatment of community-acquired pneumonia in adults: an historical perspective. Journal of Antimicrobial Chemotherapy. 1993; 32:17-27.
2. Meier PS, Utz S, Aebi S, Mühlemann K. Low-level resistance to rifampin in Streptococcus pneumoni. Antimicrobial Agents and Chemotherapy. 2003; 47:863-868.
3. von Mollendorf C, Cohen C, de Gouveia L, Quan V, Meiring S, Feldman C, *et al.* Factors associated with ceftriaxone nonsusceptibility of Streptococcus pneumoniae: analysis of South African national surveillance data. 2003 to 2010. Antimicrobial Agents and Chemotherapy. 2014; 58:3293-3305.
4. Thomas K, Group IBISI, Network I C E. Prospective multicentre hospital surveillance of Streptococcus pneumoniae disease in India. The lancet. 1999; 353:1216-1221.
5. Song JH, Jung SI, Ko KS, Kim NY, Son JS, Chang HH, *et al.* High prevalence of antimicrobial resistance among clinical Streptococcus pneumoniae isolates in Asia (an ANSORP study). Antimicrobial Agents and chemotherapy. 2004; 48:2101-2107.
6. Capoor M R, Nair D, Aggarwal P & Gupta B. Rapid diagnosis of community-acquired pneumonia using the Bac T/alert 3D system. Brazialian Journal of Infectious Dieases. 2006; 10:352-356.
7. Wattal C, Oberoi JK, Pruthi PK, Gupta S. Nasopharyngeal carriage of Streptococcus pneumonia. Indian Journal of Pediatrics. 2007; 74:905-907.
8. Robb M, Hobbs JK, Woodiga SA, Shapiro-Ward S, Suits M D, McGregor N, *et al.* Molecular Characterization of N-glycan Degradation and Transport in Streptococcus pneumoniae and Its Contribution to Virulence. PLoS pathogens, 2017, 13.
9. Nancy YY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, *et al.* PSORTb 3.0: improved protein sub cellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010; 26:1608-1615.
10. Yu CS, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. Proteins Structure Function Bioinformatics. 2006; 64:643-651.
11. Luo H, Lin Y, Gao F, Zhang CT, Zhang R, DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic acids Research, 2014; 42.
12. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE. The Pfam protein families database, Nucleic Acids Research, 2009; 985.
13. Marchler-Bauer A, Lu S, Anderson J B, Chitsaz F, Derbyshire M K, DeWeese-Scott C Y. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research, 2011; 39.
14. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research, 2007, 35.
15. Gasteiger E, Hoogland C, Gattiker A, Duvaud S E, Wilkins MR, Appel RD *et al.* Protein identification and analysis tools on the ExPASy server. Humana Press, 2005, 571-607.
16. Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Computer Applications in the Bioscience. 1995; 11:681-684.
17. Chen CC, Hwang JK, Yang JM. (PS) ^{2}-v2: template-based protein structure prediction server. Bioinformatics, 2009, 10.
18. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics. 2003; 19:246-255.
19. Ramachandran S, Kota P, Ding F, Dokholyan NV. Proteins Structure Function Bioinformatics. 2011; 79:260-271.
20. Castrignano T, De Meo PDO, Cozzetto D, Talamo IG, Tramontano A. The PMDB protein model database. Nucliec Acids Research, 2006, 34.
21. J Liang H, Edelsbrunner C. Woodward. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Science. 1998; 7:1884-1897.
22. J Liang H, Edelsbrunner P, Fu P, Sudhakar V, Subramaniam S. Analytical shape computing of

macromolecules II: identification and computation of inaccessible cavities inside proteins. Proteins. 1998; 33:18-29.
23. J Liang H, Edelsbrunner P, Fu P, Sudhakar V, Subramaniam S. Analytical shape computing of macromolecules I: molecular area and volume through alpha shape. Proteins. 1998; 33:1-17.
24. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology Evolution, 2016.
25. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology Evolution. 1987; 4:406-425.
26. Felsensteinb J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 1985, 783-791.
27. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. Evolving genes and proteins. 1965; 97:97-166.
28. Butt AM, Batool M, Tong Y. Homology modeling comparative genomics and functional annotation of Mycoplasma genitalium hypothetical protein MG_ 237, Bioinformation. 2011; 7:299-303.
29. Imai K, Asakawa N, suji T, Akazawa F, Ino A, Sonoyama M, *et al.* SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. Bioinformation. 2008; 2:417-421.