MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

Article

Functional Annotation of Selected *Streptococcus pneumoniae* Hypothetical Proteins

Khairiah Razali[1,a], Azzmer Azzar Abdul Hamid[1,b], Noor Hasniza Md Zin[1,c], Noraslinda Muhamad Bunnori[1,d], Hanani Ahmad Yusof[2], Kamarul Rahim Kamarudin[3], 'Aisyah Mohamed Rehan[1,e]

[1]*Department of Biotechnology, Kulliyyah of Science, International Islamic University Malaysia (IIUM), Jalan Sultan Ahmad Shah, 25200 Kuantan, Pahang, Malaysia*
E-mail: [a]kyaaahrzl@gmail.com, [b]azzmer@iium.edu.my, [c]hasnizamz@iium.edu.my, [d]noraslinda@iium.edu.my, [e]mraisyah@iium.edu.my

[2]*Department of Biomedical Sciences, Kulliyyah of Allied Health Sciences, International Islamic University Malaysia (IIUM), Jalan Sultan Ahmad Shah, 25200 Kuantan, Pahang, Malaysia*
E-mail: hanani@iium.edu.my

[3]*Department of Technology and Natural Resources, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Pagoh Campus, Pagoh Education Hub, Km 1, Jalan Panchor, 84600 Muar, Johor, Malaysia*
E-mail: kamarulr@uthm.edu.my

***Abstract*— The ability of Streptococcus pneumoniae to induce infection relies on its virulence factor machinery. A previous study has identified essential proteins that might be responsible towards the pathogenicity of S. pneumoniae serotype 2 strain D39. However, 39 of them were yet functionally and structurally uncharacterized. Thus, by using in silico approach, this study aims to annotate the function and the structure of these unannotated proteins. Initially, all 39 targeted proteins went through primary screening for template availability and pathogenicity. From there, 11 of them were selected and were further analyzed on the basis of their physicochemical, functional and structural categorization using an integrated bioinformatics approach by means of amino acid sequence and structure-based analysis. The obtained data suggested that all targeted proteins showed high possibility to be involved in either cell viability or cell pathogenicity mechanism of the bacterium, with SPD_1333 and SPD_1743 being the two most promising proteins to be further studied. Findings from this study could provide a better understanding of the pathogenic ability of this microorganism and thus, enhance drug development and target identification processes in the aim of improving pneumococcal disease control.**

***Keywords*— Hypothetical proteins; *S. pneumoniae strain D39; In silico* analysis of protein; Bioinformatics tools.**

I. INTRODUCTION

Streptococcus pneumoniae or pneumococcus is a Gram-positive bacterium under the family of Streptococcaceae. This facultative anaerobe is found mainly in the upper respiratory tract of human, specifically nose and throat. Despite being one of the normal floras inside a human, this organism is known to be the causative agent of infectious diseases such as pneumococcal pneumonia, meningitis and otitis media. According to the World Health Organization (WHO), in 2015, 16% of the deaths of children under five years old are caused by pneumonia with developing countries being the most prominent to acquire this disease [1].

S. pneumoniae is transmitted through respiratory route, especially through inhalation of air-borne droplets produced by coughing and sneezing from infected individuals. The colonization of *S. pneumoniae* at host respiratory area can cause pneumoniae whilst its excess to bloodstream enables it to colonize other parts of the body and cause other diseases such as otitis media and etc. Once the bacterium has succeeded in invading the bloodstream, it can travel to the blood brain barrier hence attacking the brain and causing pneumococcal meningitis.

In order to cause diseases, pneumococci makes use of its virulence factors machinery, which mostly involves its polysaccharide capsule, cell wall and pneumolysin [2]. Over the past years, pneumococcal diseases have been treated by vaccinations and antibiotics. Examples of vaccines and antibiotics are Pneumococcal Conjugate Vaccine (PCV) 13 and Levofloxacin, respectively. However, previous studies have shown that inappropriate antibiotic prescriptions in treating pneumococcal diseases have led to an increase in antibiotic and multidrug resistant pneumococci [3]. In addition, these drugs are serotype-specific, which means that one type of drug can only be used on a particular strain or serotype [4]. This issue is one of the current concerns among scientists and medical doctors. Thus, it is important to search for better vaccines and antibiotics or other alternatives with the aim of preventing or treating pneumococcal infections. In order to do so, deep understanding on the virulence factors machinery of *S. pneumoniae* is very much needed.

Virulence factors are the ones that are responsible in the capability of *S. pneumoniae* to cause diseases. Understanding pneumococci virulence factors machinery demands full knowledge of its proteins and components involved. The most important factor in the virulence of this organism is its polysaccharide capsule [5]. A study has demonstrated that variances in this capsule have raised the number of different pneumococcal strains and serotypes, thus leading to bacterial resistance [6]. Other virulence factors include the cell wall and several other proteins such as hyaluronate lyase, neuraminidase and pneumolysin. Presently, biotechnology and bioinformatics applications have enabled scientists to completely sequence bacterial genome and assign structure and function to its proteins and enzymes [7]. Yet, due to complexity and time constraint, one third of *S. pneumoniae* proteins still remain hypothetical with neither structural nor functional elucidations [8]. This problem has limited the potential of designing drugs capable of fighting pneumococci-related diseases.

Hence, this study aims to fill the gap between genome sequence information and virulent protein annotation by interpreting physicochemical characteristics, structures and functions of selected hypothetical proteins from the previously identified essential proteins of *S. pneumoniae* strain D39 [9]. By using suitable computational and bioinformatics tools as well as available genome, proteome and secretome databases, this study is expected to provide an insight on the structure and role of hypothetical proteins in the virulence factors machinery of *S. pneumoniae*, specifically strain D39. Acquiring this information will later help researchers to continue with protein expression and purification studies on promising hypothetical protein targets for further analysis. In the longer term, this study will provide a promising platform for drug design and therapeutic studies of pneumonia-related diseases.

II. THE MATERIAL AND METHOD

A. Sequence Retrieval

The ID name and the full sequence of each of the 39 hypothetical proteins was retrieved from UniProtKB (http://www.uniprot.org/) and the National Centre for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/). Column 1 and 2 of Table I show protein ID used by the previous study [9] and NCBI, respectively.

B. Virulence Prediction

MP3 server (http://metagenomics. iiserb.ac.in/mp3/) uses Support Vector Machines (SVM) or Hidden Markov Model (HMM) to calculate the algorithm and predict the pathogenesis of query protein. All 39 hypothetical proteins were analyzed by this server for their virulence properties.

C. Template Availability

Next, the hypothetical proteins were streamed through the NCBI blast and PSI-BLAST servers against the Protein Data Bank (PDB) proteins database for the search of homology. Hypothetical proteins having the template aligned at above 50% and similarity of 30 to 70% were of concern. It has been widely accepted that two proteins are considered homologous if their sequence similarity is beyond 30% [10]. At the end of the selection process, 11 out of 39 hypothetical proteins of *S. pneumoniae* strain D39 were selected to be the subjects of study.

TABLE I
LIST OF 39 HYPOTHETICAL PROTEINS WITH THEIR RESPECTIVE ID NAMES

No.	Ref [11]	NCBI Website	No.	Ref [11]	NCBI Website
1	SPD_0008	ABJ54275.1	21	SPD_0880	ABJ54796.1
2	SPD_0403	ABJ55288.1	22	SPD_1136	ABJ55057.1
3	SPD_0408	ABJ54600.1	23	SPD_1197	ABJ54310.1
4	SPD_0965	ABJ54319.1	24	SPD_1198	ABJ54015.1
5	SPD_0990	ABJ53736.1	25	SPD_1288	ABJ54428.1
6	SPD_1392	ABJ54564.1	26	SPD_1333	ABJ53729.1
7	SPD_1405	ABJ54327.1	27	SPD_1346	ABJ53954.1
8	SPD_1435	ABJ53912.1	28	SPD_1391	ABJ55369.1
9	SPD_1522	ABJ53766.1	29	SPD_1416	ABJ54839.1
10	SPD_2029	ABJ53876.1	30	SPD_1417	ABJ54239.1
11	SPD_0131	ABJ53868.1	31	SPD_1549	ABJ55165.1
12	SPD_0339	ABJ54095.1	32	SPD_1560	ABJ55503.1
13	SPD_0350	ABJ55277.1	33	SPD_1672	ABJ55016.1
14	SPD_0394	ABJ54531.1	34	SPD_1706	ABJ54739.1
15	SPD_0402	ABJ54394.1	35	SPD_1743	ABJ54209.1
16	SPD_0476	ABJ55230.1	36	SPD_1803	ABJ53653.1
17	SPD_0478	ABJ54971.1	37	SPD_1898	ABJ55068.1
18	SPD_0675	ABJ53899.1	38	SPD_2043	ABJ54886.1
19	sufD	ABJ54405.1	39	SPD_2044	ABJ53789.1
20	SPD_0878	ABJ53627.1			

D. Physicochemical Characteristics

Several physical and chemical parameters (molecular weight, isoelectric point, extinction coefficient, aliphatic index, instability index and GRAVY) were analysed using ExPASy ProtParam tool (https://web.expasy.org/protparam/). These parameters are important in knowing the state of query protein, especially for means of experimental handling such as for protein isolation and purification.

E. Conserved Family and Domain

Pfam (https://pfam.xfam.org/) and NCBI CD-Search servers were used to predict possible domain or family of a query protein. Domain and family are able to give insight into the possible role or interaction that may be associated with the query protein by looking at the function and the structure of proteins they are similar with.

F. Subcellular Localization, Trans-Membrane Helices and Secretome Analysis

PSORT and PSORTb servers (http://www.psort.org/psortb/index.html) were used to predict the subcellular localization of the query protein. Similarly, HMMTOP as well as SignalP and SecretomeP servers were used to determine the presence of trans-membrane helices and signal peptides, respectively. All information is important in categorizing whether a protein is a membrane protein, secretory protein or cytoplasmic protein.

G. Protein-Protein Interaction

The STRING website (https://string-db.org/) is an online server that contains protein databases of thousands of organisms and is useful in analysing protein-protein interactions. STRING currently holds the databases of around five million proteins from thousand organisms [11]. By using

STRING, the interactions between the query protein and other surrounding proteins were accessed. This enables the identification of functional and regulatory interactions among proteins.

H. Secondary Structure Prediction

The initial structural annotation of the query protein was determined by predicting its secondary structure. The prediction allows the information on how many possible helices, strands and loops are present in shaping the query protein. This step was done using the PSIPRED server (bioinf.cs.ucl.ac.uk/psipred/).

I. Tertiary Structure Prediction

In predicting the tertiary structure, three different servers, namely I-TASSER (https://zhanglab.ccmb.med.umich.edu/I-TASSER/), (PS)2 (ps2.life.nctu.edu.tw/) and ExPASy SWISS-MODEL (https://swissmodel.expasy.org/) were used for each query protein.

All predicted structures from these three servers were then validated using Ramachandran plot assessment, Verify3D and QMEAN4 score. From the validation, the best predicted structure was selected for structural refinement and further analysis.

III. RESULTS AND DISCUSSION

The screening process of 39 essential hypothetical proteins revealed that only 11 of them are suitable to be targeted proteins. In general, sequence and structure-based analysis showed that the targeted proteins are diverse in terms of their physicochemical characteristics, structures and functions.

A. Physicochemical Characteristics

The physicochemical characteristics analysis revealed that the isoelectric point (pI) value for all selected hypothetical proteins from this study fell between the ranges of 4.59 to 9.40. Next, the highest extinction coefficient (EC) belongs to SPD_1346 (38740 M^{-1} cm^{-1}) whilst the lowest is 2980 M-1 cm-1, which belongs to SPD_0878. Furthermore, in terms of the instability index (II), 6 out of 11 proteins (SPD_0965, SPD_0402, SPD_1333, SPD_1392, SPD_1743 and SPD_0339) were predicted to be stable inside a test tube. Unstable proteins may require additional steps such as denaturation prior to isolation and purification. Other details on the parameters of each protein, such as the molecular weight, aliphatic index and GRAVY value, given in Table II.

TABLE III
PHYSICOCHEMICAL CHARACTERISTICS BY EXPASY PROTPARAM

Gene ID	MW (Da)	pI	EC (M^{-1} cm^{-1})	AI	II	GRAVY
SPD_0965	5961.67	8.19	5500	65.77	15.95	-0.956
SPD_0131	9288.36	4.59	8940	82.34	66.60	-0.812
SPD_0402	12868.69	4.90	5960	107.02	13.89	0.134
SPD_1333	37756.89	5.08	33030	81.01	32.35	-0.434
SPD_1288	8258.25	9.40	8480	160.68	42.66	1.442
SPD_1898	7229.30	8.82	8480	80.85	50.00	-0.949
SPD_1392	30129.38	7.92	26930	121.66	31.99	0.362
SPD_1743	16401.74	4.73	15930	108.84	34.76	-0.190
SPD_0339	12575.29	4.74	4470	90.37	36.24	-0.456
SPD_0878	18970.57	4.88	2980	89.69	53.18	-0.852
SPD_1346	60797.40	5.07	38740	81.78	50.96	-0.528

B. Protein Domains and Families

The initial step in understanding functional property of a protein is to determine its domain and family. From this study, out of eleven selected hypothetical proteins, nine of them were classified into specific domain(s) and family(s) with no record or identification was found on SPD_0965 and SPD_1898 (listed in Table III). This might be due to the length of their amino acid that is short, 52 and 59 residues, respectively. A study showed that mini-proteins (those with residues of not more than 100 amino acids) are difficult to be analysed experimentally and computationally due to their small sizes and short gene lengths [12].

TABLE IIIII
CONSERVED FAMILY(S) AND DOMAIN(S) BY PFAM AND NCBI CD-SEARCH

Gene ID	Pfam and NCBI CD-search	Description
SPD_0965	-	-
SPD_0131	DUF1447 family	Protein of unknown function
SPD_0402	Asp23 superfamily, YloU family	Alkaline shock protein, cell envelope-related function
SPD_1333	Lactonase family	Lactonase, 7-bladed beta-propeller, carbohydrate transport and metabolism
SPD_1288	DUF4059 family	Protein of unknown function
SPD_1898	-	-
SPD_1392	DisA_N family	Diadenylate cyclase (c-di-AMP synthetase), DisA bacterial checkpoint controller nucleotide-binding
SPD_1743	P-loop NTPase superfamily, TsaE domain	Threonylcarbamoyl adenosine biosynthesis protein TsaE
SPD_0339	DivIVA family	Cell division protein
SPD_0878	HTH_24 domain, DUF536 family	Winged helix-turn-helix DNA binding, Protein of unknown function
SPD_1346	YceG-like family	Cell division protein YceG

C. Sub-Cellular Localization and Secretome Analysis

Determination of protein subcellular location is significant, especially for target identification [13]. Furthermore, location prediction can give an idea about the role of a query protein and their types such as cytoplasmic, membrane or secretory protein. It is also important to locate the presence of trans-membrane helices and signal peptide as a positive prediction of these two can further validate a protein's function in secretory or extracellular interactions [14].

Analysis done to 11 selected proteins of this study revealed five proteins to be at cytoplasmic location whilst, another five at cell membrane and one indecisive. In terms of the presence of trans-membrane helices, three proteins were predicted to have one trans-membrane helix (SPD_0402, SPD_1346 and SPD_1898), one protein with two trans- membrane helices (SPD_1288) and another one protein with three trans-membrane helices (SPD_1392) whilst the remaining six possessed no trans-membrane helix. None of the proteins were predicted to own a signal peptide and five out of eleven proteins (SPD_0402, SPD_1333, SPD_1346, SPD_1288 and SPD_1392) were thought to be responsible in secretory pathway mechanism (listed in Table IV). Due to the absence of signal peptide, it is not presented in Table IV.

TABLE IV
SUBCELLULAR, TRANS-MEMBRANE HELICES AND SECRETOME ANALYSIS

Gene ID	Subcellular Localization		Trans-membrane Helices	Secretome Analysis
	PSORT	PSORTb	HMMTOP	SecretomeP (score)
SPD_0965	Bacterial cytoplasm	Extracellular	-	No (0.400)
SPD_0131	Bacterial cytoplasm	Cytoplasmic	-	No (0.100)
SPD_0402	Bacterial membrane	Cytoplasmic membrane	One (18-37)	Possibly (0.654)
SPD_1333	Bacterial cytoplasm	Cytoplasmic	-	Possibly (0.775)
SPD_1288	Bacterial membrane	Cytoplasmic membrane	Two (12-30, 51-72)	Possibly (0.950)
SPD_1898	Bacterial membrane	Unknown	One (4-20)	No (0.078)
SPD_1392	Bacterial membrane	Cytoplasmic membrane	Three (6-25, 34-54, 59-78)	Possibly (0.847)
SPD_1743	Bacterial cytoplasm	Cytoplasmic	-	No (0.057)
SPD_0339	Bacterial cytoplasm	Cytoplasmic	-	No (0.050)
SPD_0878	Bacterial cytoplasm	Cytoplasmic	-	No (0.089)
SPD_1346	Bacterial membrane	Unknown	One (188-206)	Possibly (0.927)

D. Protein-Protein Interaction

The involvement of a protein in virulence factor machinery is pretty much influenced by its interactions with other proteins. Some proteins work in synergy in order to perform vital cellular functions [13]. Hence, knowing the relationship between a hypothetical protein and other proteins can give insights into its possible function or role. In accordance to this, the analysis of protein-protein interaction by STRING provide information on the types of relation (neighborhood, co-occurrence, text-mining and experimental) between query protein and others.

Table V in this study has shown the top three proteins of the highest interaction with the query protein. The score given for each interaction was in the range of 0 to 1, with 1 being the strongest interaction. From the analysis, it showed that most of the proteins being studied were directly involved with the virulence machinery of *S. pnuemoniae*.

TABLE V
PROTEIN-PROTEIN INTERACTIONS BY STRING

Gene ID	Interacting Protein	Protein Function
SPD_0965	Obg protein, CpoA protein	Modulates vital processes, Saccharides biosynthesis
SPD_0131	Ribonuclease J, MecA protein, DivIB protein	Hydrolyses β-lactam antibiotics, Involves in bacterial pathogenesis, Cell wall synthesis
SPD_0402	SPD_0403, SPD_1388	Catalyses glycerol metabolic processes, Key regulator for virulence of Gram-positive bacteria
SPD_1333	Zwf protein, Gnd protein, SPD_1330	Carbohydrate degradation process, Carbohydrate degradation process, ATP-binding cassette transporter
SPD_1288	TrxB protein, SPD_1290, SPD_1293	Catalyses the reduction of thioredoxin, ABC transporter, Involves in aminoglycoside antibiotics resistance mechanism
SPD_1898	SPD_1899, SPD_1897, SPD_1895	Purine nucleotide biosynthesis, Purine nucleotide biosynthesis, Protein biosynthesis
SPD_1392	GlmM protein, SPD_2032, SPD_1393	Catalyses peptidoglycan biosynthesis, Involves in c-di-AMP homeostasis, Catalyses disulphide bonds formation
SPD_1743	TsaD protein, NnrD protein, Recombinase A	Involves in tRNA processing machinery, Involves in bacterial stress adaptation, Responses to β-lactam antibiotics
SPD_0339	EzrA protein, RecU protein, Pbp2 protein	Essential for growth, cell division and cell size homeostasis, Involves in DNA damage repair mechanism, Involves in methicillin resistance mechanism
SPD_0878	MtnN protein, SPD_0875, GlmU protein	Involves in virulence machinery of Gram-negative bacteria, Controls cell homeostasis, Cell membrane synthesis
SPD_1346	GreA protein, MurC protein	Regulates RNA polymerase activity, Involves in peptidoglycan biosynthesis

E. Secondary and Tertiary Structure Prediction

Another important aspect to consider during annotating protein functional properties is its two and three dimensional structure. This prediction revealed the possible shape or folding (helices, strands and loops) of a query protein from its amino acid sequence. The knowledge of protein structure enables further identification of important protein characteristics such as active sites and binding ligands. In this study, all structures were successfully predicted, except for two proteins (SPD_1346 and SPD_0878) due to their large atomistic structure (Table VI).

TABLE VI
SUMMARY OF STRUCTURAL VALIDATION OF FINAL MODELLED STRUCTURES

Gene ID	Ramachandran Plot Assessment			QMEAN4 Score	Verify3D (%)
	Favored Region (%)	Allowed Region (%)	Outlier Region (%)		
SPD_0965	83.7	10.2	6.1	-3.97	100.00
SPD_0131	86.5	8.1	5.4	-3.09	57.14
SPD_0402	89.0	6.8	4.2	-2.15	100.00
SPD_1333	87.4	9.0	3.6	-3.20	100.00
SPD_1288	93.0	7.0	0.0	-3.49	8.11
SPD_1898	87.5	8.9	3.6	-3.57	71.19
SPD_1392	82.1	10.8	7.1	-5.61	75.28
SPD_1743	89.6	9.0	1.4	-0.95	100.00
SPD_0339	96.6	3.4	0.0	-0.09	2.44
SPD_0878	79.5	11.2	9.3	-8.18	22.09
SPD_1346	63.9	23.7	12.4	-13.54	16.88

IV. CONCLUSIONS

The analysis which has been done on all eleven proteins showed that, respectively, four and three of the proteins are classified under domain or family that could be possibly involved in pathogenicity and viability of *S. pneumoniae*. Furthermore, five out of eleven proteins were strongly predicted to be involved in the pathogenesis whilst, four in the survival mechanism of *S. pneumoniae* strain D39. Structure refinement, active sites and ligand binding prediction are currently pursued, and this will narrow down the potential candidates to be further exploited. By using *in silico* sequence and structure-based approaches, the main contribution of this study draws on the gap of previously unannotated essential proteins by predicting probable physicochemical, functional and structural properties of selected hypothetical proteins.

ACKNOWLEDGEMENT

We would like to thank all staff at Kulliyyah of Science, International Islamic University Malaysia for their assistance. This study is funded by RAGS 14-036-0099 research grant from the Malaysian Ministry of Education and IIUM RIGS research grant (RIGS16-312-0476).

REFERENCES

[1] World Health Organization. (2016) *Pneumococcal Disease*. [Online]. Available: http://www.who.int/biologicals/vaccines/pneumococcal/en/

[2] N. Henriques, Birgitta, and I. T. Elaine. "The pneumococcus: epidemiology, microbiology and pathogenesis." *Cold Spring Harbor Perspectives in Medicine*, vol. 3(7), p. a010215, 2013.

[3] C. M. Gant, A. W. Rosingh, J. L. López-Hontangas, M. van der Heijden, F. González-Morán, J. J. E. Bijlsma, E. Canton, and RedMiva (Network of Microbiological Vigilance of Comunidad Valenciana), "Serotype distribution and antimicrobial resistance of invasive pneumococcal disease strains in the Comunidad Valenciana, Spain, during the winter of 2009-2010: Low PCV7 coverage and high levofloxacin resistance," *Antimicrobial Agents and Chemotherapy*, vol. 56(9), pp. 4988-4989, 2012.

[4] M. Lipsitch, and R. S. George, "How can vaccines contribute to solving the antimicrobial resistance problem?" *mBio*, vol. 7(3), p. e00428-16, 2016.

[5] C. Hyams, C. Emilie, M. C. Jonathan, B. Katie, and J. S. Brown, "The Streptococcus pneumoniae capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms," *Infection and Immunity*, vol. 78(2), pp. 704-715, 2010.

[6] R. Mostowy, J. C. Nicholas, P. H. William, R. H. Simon, B. Stephen, and F. Christophe, "Heterogeneity in the frequency and characteristics of homologous recombination in pneumococcal evolution," *PLOS Genetics*, vol. 10(5), p. e1004300, 2014.

[7] K. M. Dahlström. (2015) *From Protein Structure to Function with Bioinformatics*. [Online]. Available: https://www.doria.fi/handle/10024/117295

[8] S. Wuchty, S. V. Rajagopala, S. M. Blazie, J. R. Parrish, S. Khuri, R. L. Finley, and P. Uetz. "The protein interactome of Streptococcus pneumoniae and bacterial meta-interactomes improve function predictions," *MSystems*, vol. 2(3), p. e00019-17, 2017.

[9] X. Liu, G. Clement, K. Morten, D. Arnau, S. Jelle, P. V. K. Sebastiaan, K. Kèvin, A. S. Robin, J. R. Zhang, and J. W. Veening, "High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae," *Molecular Systems Biology*, vol. 13(5), p. 931, 2017.

[10] W. R. Pearson, "An introduction to sequence similarity ("homology") searching." *Current Protocols in Bioinformatics*, vol. 42(1), pp. 3-1, 2013.

[11] M. S. Islam, S. M. Shahik, M. Sohel, N. I. A. Patwary, and M.A. Hasan, "In silico structural and functional annotation of hypothetical proteins of Vibrio cholerae O139," *Genomics & Informatics*, vol. 13(2), p. 53, 2015.

[12] F. Wang, X. Jingfa, P. Linlin, Y. Ming, Z. Guoqiang, J. Shouguang, and Y. Jun, "A systematic survey of mini-proteins in bacteria and archaea," *PLOS One*, vol. 3(12), p. e4027, 2018.

[13] S. Wan, D. Yucong, and Z. Quan, "HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source," *Proteomics*, vol. 17(17-18), p. 1700262, 2017.

[14] M. J. Davis, A. H. Kelly, C. Francis, J. L. Fink, F. Zhang, K. Takeya, and K. Chikatoshi, "Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units," *PLOS Genetics*, vol. 2(4), p. e46, 2006.