In Silico Structural and Functional Annotation of Nine Essential Hypothetical Proteins from *Streptococcus pneumoniae*

Khairiah Razali[1], Azzmer Azzar Abdul Hamid[1,2], Noor Hasniza Md Zin[1], Noraslinda Muhamad Bunnori[1,2], Hanani Ahmad Yusof[3], Kamarul Rahim Kamarudin[4], and Aisyah Mohamed Rehan[1,2,*]

[1]*Department of Biotechnology, Kulliyyah of Science, International Islamic University Malaysia, Jl. Sultan Ahmad Shah, 25200, Kuantan, Pahang, Malaysia*

[2]*Research Unit for Bioinformatics and Computational Biology (RUBIC), Kulliyyah of Science, International Islamic University Malaysia, Jl. Sultan Ahmad Shah, Kuantan, Pahang, 25200, Malaysia*

[3]*Department of Biomedical Sciences, Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, Jl. Sultan Ahmad Shah, Kuantan, Pahang, 25200, Malaysia*

[4]*Department of Technology and Natural Resources, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Campus, Pagoh Education Hub, Km 1, Jl. Panchor, Muar, Johor Darul Takzim, 84600, Malaysia*

* ***Corresponding author:***

tel: +6012-6848659
email: mraisyah@iium.edu.my

Received: December 13, 2018
Accepted: March 11, 2019

DOI: 10.22146/ijc.41817

Abstract: *The ability of* Streptococcus pneumoniae *to induce infections relies on its virulence factor machinery. A previous CRISPR interference (CRISPRi) study had identified 254 essential proteins that may be responsible for the pathogenicity of* S. pneumoniae *serotype 2 strain D39. However, 39 of them were functionally and structurally uncharacterized. Hence, by using in silico approach, this study aimed to annotate the function and structure of these un-annotated proteins. Initially, all 39 proteins went through primary screening for template availability and pathogenicity. From there, 11 of them were selected and underwent further physicochemical, functional, and structural categorization through an integrated bioinformatics approach by means of amino acid sequence- and structure- based analyses. The obtained data revealed that 9 targeted proteins showed a 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 can help in facilitating a better understanding of pathogenic ability of this microorganism and 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*

■ INTRODUCTION

Streptococcus pneumoniae or pneumococcus is a Gram-positive bacterium under the family of Streptococcaceae. This facultative anaerobe is found mainly at 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 get this disease [1]. Susceptible individuals can develop an invasive pneumococcal infection that can be severe, and in the absence of appropriate antibiotics treatment, may lead to hospitalization, life-long disability, and death [1].

S. pneumoniae is transmitted through the respiratory

route, especially through inhalation of air-borne droplets generated by coughing and sneezing from infected individuals. The colonization of *S. pneumoniae* at host respiratory area can cause pneumonia while its excess to bloodstream enables it to colonize other parts of the body and cause diseases such as otitis media. 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 [2].

In order to cause diseases, pneumococci make use of its virulence factors machinery, which mostly involves its polysaccharide capsule, cell wall, and pneumolysin [3]. Over the past years, prevention and treatment of pneumococcal diseases are through vaccinations and antibiotics, respectively. Example of vaccines and antibiotics are pneumococcal conjugate vaccines (PCV) and amoxicillin, respectively. However, it is found that inappropriate antibiotic prescriptions in treating pneumococcal diseases have led to an increase in antibiotic- and multidrug- resistant pneumococci [4-5]. In addition, currently available vaccines are serotype-specific, and therefore, elicit serotype-specific immunity [6]. The developing countries displayed pneumococcal disease that is caused by a wider spectrum of serotypes as compared to developed countries [7]. Hence, the search for better vaccines and antibiotics with the aim of preventing or treating pneumococcal infections are essential. In order to do so, deep understanding on the virulence factors machinery of *S. pneumoniae* is very much needed.

Virulence factors play a large role in determining the capability among different strains of *S. pneumoniae* in causing diseases [8]. 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 [9]. Another study further demonstrates that variances in this capsule have raised the number of different pneumococcal strains and serotypes, thus leading to bacterial resistance [10]. Other virulence factors include the cell wall plus several proteins such as hyaluronate lyase, neuraminidase, and pneumolysin [11]. Presently, biotechnology and bioinformatics applications have enabled scientists to completely sequence the bacterial genome and assign structure and function to its proteins and enzymes [12]. Yet, due to the complexity and other constraints, one third of its proteins remain hypothetical with neither structural nor functional elucidations [13]. 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 [14]. With suitable computational and bioinformatics tools as well as an available genome, proteome, and secretome databases, this study is expected to provide insights on the structure and role of hypothetical proteins in virulence factors machinery of *S. pneumoniae,* specifically for strain D39. Acquiring this information will later help researchers to continue with protein expression and purification studies on promising hypothetical protein targets for further analyses. In the longer term, this study will provide a promising platform for drug design and therapeutic studies of pneumonia-related diseases.

■ EXPERIMENTAL SECTION

Sequence Retrieval

The ID name and full sequence of each of the 39 hypothetical proteins were retrieved from UniProtKB (http://www. uniprot.org/) and the National Centre for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/protein/). The corresponding NCBI-protein ID accession number for each of the 39 hypothetical proteins targets identified from [14] is listed in Table S1.

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 [15]. All 39

hypothetical proteins were analyzed by this server for their virulence properties.

Template Availability

Next, the hypothetical proteins were streamed through NCBI BLASTP and PSI- BLAST servers against 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% [16]. 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.

Physicochemical Characteristics

Several physical and chemical parameters (molecular weight, isoelectric point, extinction coefficient, aliphatic index, instability index, and GRAVY) were analyzed using ExPASy ProtParam tool (https://web.expasy.org/protparam/) [17]. These parameters are important in knowing the state of the query protein, especially for means of experimental handling such as for protein isolation and purification.

Conserved Family and Domain

Pfam (https://pfam.xfam.org/) [18] and NCBI CD-Search servers [19] 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 structure of proteins they are similar with.

Subcellular Localization, Trans-Membrane Helices, and Secretome Analyses

PSORT and PSORTb servers (http://www.psort.org/psortb/index.html) [20] were used to predict the subcellular localization of the query protein. Similarly, HMMTOP [21], as well as SignalP [22] and SecretomeP [23] servers, were used to determine the presence of trans-membrane helices and signal peptides, respectively. This information is important in categorizing whether a protein is a membrane protein, secretory protein, or cytoplasmic protein.

Protein-Protein Interaction

STRING website (https://string-db.org/) is an online server that contains protein databases of thousands of organisms and is useful in analyzing protein-protein interactions. STRING currently holds the databases of around 24 million proteins from 5090 organisms [24]. 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.

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 PSIPRED server (bioinf.cs.ucl.ac.uk/psipred/) [25].

Tertiary Structure Prediction

In predicting the tertiary structure, three different servers, namely I-TASSER (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) [26], (PS)2 (ps2.life.nctu.edu.tw/) [27] and ExPASy SWISS-MODEL (https://swissmodel.expasy.org/) [28] were used for each query protein.

All three predicted structures were then validated using Ramachandran plot assessment, Verify3D [29], and QMEAN4 score [30]. From the validation, the best-predicted structure was selected for structural refinement and further analyses.

Structural Refinement

The selected three-dimensional structure was converted from .pdb format to .gro to be subjected to structural refinement by Groningen Machine for Chemical Simulations (GROMACS) software (using force field gromos96 53a6) for improvement [31]. This process includes energy minimization, equilibration, and production stage. Prior to simulation, the box was solvated with water, and the protein system was neutralized. Equilibration and production took 100 ps and 10000 ps of simulation time, respectively.

The refined structure was again validated using the

three aforementioned servers and a graph of root mean square deviation (RMSD) against production time was retrieved. The visualization of the final structure was viewed using PyMOL software (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).

Active Site and Ligand Prediction

The refined tertiary model was sent to metaPocket 2.0 (projects.biotec.tu-dresden.de/metapocket/) for active site prediction [32]. This server allows the prediction of the top three possible ligand binding sites of the query protein. In addition, the refined structure was also sent to COACH server (https://zhanglab.ccmb.med.umich.edu/COACH/) for the prediction of possible ligand that may bind to the active sites of the query protein [33].

■ RESULTS AND DISCUSSION

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}) while the lowest is 2980 M^{-1} cm^{-1}, which belongs to SPD_0878. Moreover, in term 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, are given in Table 1.

Protein Domains and Families

The initial step in understanding the 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 a specific domain(s) and family(s), while no record or identification was found on SPD_0965 and SPD_1898 (listed in Table 2). This may be due to their short amino acid length (52 and 59 residues, respectively). A study shows that mini-proteins (those with residues of not more than 100 amino acids) are difficult to be analyzed experimentally and computationally due to their small sizes and short gene lengths [34].

Subcellular Localization and Secretome Analyses

Determination of protein subcellular location is significant, especially for target identification [35]. Furthermore, location prediction can give an idea on the role of a query protein and whether it is categorized as a cytoplasmic, membrane, or secretory protein. Plus, it is also important to locate the presence of trans-membrane helices and signal peptide because the positive prediction of these two can further validate a protein's function in secretory or extracellular interactions [36].

Analyses done to all subjects of this study revealed

Table 1. Physicochemical characteristics by ExPASy ProtParam. EC: Extinction Coefficient; AI: Aliphatic Index; II: Instability Index; GRAVY; grand average of hydropathy

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

Table 2. 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

five proteins to be at a cytoplasmic location, another five at the cell membrane and one indecisive. In term of the presence of trans-membrane helices, three proteins were predicted to have one transmembrane helix (SPD_0402, SPD_1346 and SPD_1898), one protein with two transmembrane helices (SPD_1288) and another one protein with three trans-membrane helices (SPD_1392) while the remaining six have 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 said to be responsible in secretory pathway mechanism (listed in Table 3).

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 [24]. Hence, knowing the relationship between a hypothetical protein and other proteins can give insights into its possible function or role. In accordance with this, the analysis of protein-protein interaction by STRING gave information on the types of relation (neighborhood, co-occurrence, text-mining, and experimental) between the query protein and others.

Table 3. Subcellular, trans-membrane helices, signal peptide, and secretome analyses

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

Table 4 shows 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 was revealed that most of the proteins involve directly with the virulence machinery of *S. pnuemoniae*.

Secondary and Tertiary Structure Prediction and Refinement

Another pivotal aspect to consider when annotating protein functional properties is its two- and three-dimensional structure. This prediction revealed possible shape or folding (helices, strands, and loops) of a query protein from its amino acid sequence. The knowledge on protein structure enables further identification on important protein characteristics such as active sites and binding ligands. Structural refinement, on the other hand, is crucial in improving the predicted structure to minimize the energy, thus obtaining more native protein folding [37].

In this study, all structures were successfully predicted and refined except for two proteins (SPD_1346 and SPD_0878) due to large atomistic structure. Based on the graph of root mean square deviation (RMSD) against

Table 4. Protein-protein interactions by STRING

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

time during the production stage, six out of nine refined structures had reached the plateau stage. However, the remaining three (SPD_0402, SPD_1288, and SPD_1392) still had an increasing RMSD, suggesting longer production time is needed (Table 5).

Structural Validation

Structural validation was done to verify the quality of predicted models. After the refinement process, final structures were again subjected to structural validation through Ramachandran plot assessment (to visualize the distribution of torsion angles in a protein structure), QMEAN4 (to describe the likelihood that a predicted model is of comparable quality to experimental structure) and Verify3D (to verify the propensity of protein's sequence with its predicted three- dimensional structure). In QMEAN4, the closer the score to 0 indicates better model quality. In Verify 3D, a score of above 80% indicates a high tendency of the sequence to take shape like its predicted structure.

Generally, based on Ramachandran plot assessment, 80% of residues of all structures fell in the favored region except for the two unrefined structures, SPD_0878 (79.0%) and SPD_1346 (63.9%). Similarly, QMEAN4 score obtained by all proteins showed value ranges between -0.09 to -5.61, with two outliers, again from SPD_0878 (-8.18) and SPD_1346 (-13.54). Lastly, when subjected to Verify 3D, only four out of eleven proteins (SPD_0965, SPD_0402, SPD_1333, and SPD_1743) had a score of above 80%. The summary of all assessments is shown in Table 5.

Active Sites and Ligand Prediction

It is essential to know the active sites and possible binding ligand of a query protein. Based on the predicted ligand, further interpretation on the functional property of the protein can be made more precisely. The type of ligand that binds to a particular protein determines its function in a cellular mechanism or pathway. Plus, it is also important for drug designing purpose [38].

Throughout the eleven subjects of study, the active sites and possible ligand of all proteins were able to be identified except for SPD_1346 (Table 6). This exception is due to the low validity of the unrefined model of SPD_1346 tertiary structure.

In terms of ligand binding prediction (Table 6), four proteins (SPD_1333, SPD_1392, SPD_0339, and SPD_0878) showed possibility in the involvement of the cell pathogenicity mechanism and five proteins (SPD_0965, SPD_0131, SPD_0402, SPD_1288 and SPD_1743) showed possible involvement in cell viability mechanism of *S. pneumoniae* strain D39. The remaining two proteins (SPD_1898 and SPD_1346) were inconclusive due to poor results of the binding ligand prediction. This limitation may need further *in silico* studies such as molecular docking to confirm their protein-ligand interactions.

Table 5. Summary of structural validation of final modeled structures

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

Table 6. Predicted function for each protein based on their binding ligand

Gene ID	Binding ligand			Predicted function
	Cell viability	Cell pathogenicity	Unknown	
SPD_0965	Glucose			Glucose metabolism
SPD_0131	ATP molecule			Energy production
SPD_0402	2,4-Dichlorophenol			Phenol metabolism
SPD_1333		Dilysine-containing molecule		Inhibits eukaryotic motifs
SPD_1288	Glycine			Survival and growth
SPD_1898			Null	-
SPD_1392		Cordycepin triphosphate (COTP)		Inhibits RNA chain elongation
SPD_1743	ADP molecule			ATPase activity
SPD_0339		Activator protein -1 (AP-1)		Stress adaptation
SPD_0878		Phosphatidylcholine (PC)		Stress adaptation
SPD_1346			Null	-

Potential Drug Design Candidates

The screening process of 39 essential hypothetical proteins revealed that 11 of them are suitable to be target proteins. In general, sequence- and structure- based analyses showed that the targeted proteins are diverse in terms of their physicochemical characteristics, structures, and functions. Overall, two proteins (SPD_1333 and SPD_1743) showed convenience in their assessments hence making them the best potential drug design candidates out of all 11 proteins.

For SPD_1333, the protein family prediction revealed that this protein contains a sequence of lactonase family member along with residues 4-335. The sequence encodes for 6-phosphogluconolactonase, an enzyme that hydrolyzes 6-phosphogluconolactone to 6-phosphogluconate in carbohydrate metabolism via pentose phosphate [39]. This pathway is important in synthesizing nucleotides and nucleic acids vital to cells. Hence, this suggests the role of SPD_1333 in maintaining cell mechanisms and viability.

Next, the secondary structure prediction of SPD_1333 by PSIPRED server showed that the protein contains 28 possible strands interconnected by loops with very good confidence. Besides, the tertiary model structured by I-TASSER server found that SPD_1333 is structurally closed to 3HFQ_A protein (99.1% alignment, 48.2% similarity). Based on Fig. 1(a), the structure reveals the formation made by these 28 strands, thus making up a seven-bladed beta propeller structure (name as described by Pfam server).

Lastly, COACH server predicted a dilysine-containing peptide molecule (xKxKxx) binding at pocket 2 of the query protein (Fig. 1(b)). This finding strengthens the suggestion on the possibility of SPD_1333 to be a membrane protein because dilysine motif plays a role in conferring the localization of this kind of protein [40]. Many types of motifs, including the dilysine motif, are known to mimic eukaryotic motifs, thus enabling pathogenic bacteria to disturb host's cellular functions [41]. From this, it is suggested that the ability of SPD_1333 to bind with dilysine motif-containing molecule may give benefit to this protein in accomplishing its virulence mission inside the host.

SPD_1743, on the other hand, contains TsaE domain, which is a domain that falls under P-loop NTPase superfamily and significant in synthesizing threonylcarbomyl adenosine biosynthesis protein. This protein is indirectly responsible in the N6-threonylcarbamoyladenosine (t(6)A) pathway. A study found that t6A involves in decoding accuracy of mRNA codons during protein synthesis [42]. Evidently, a defect in t(6)A pathway can lead to increased frame shift events, wrong start codon selections and occurrence of pleiotropic phenotypes [43]. Hence this shows that TsaE domain is important in maintaining cell viability.

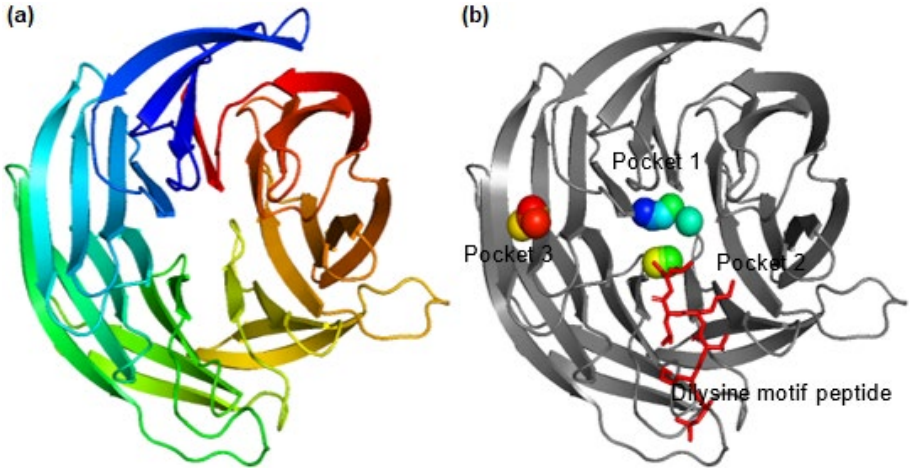

Fig 1. (a) Refined three-dimensional structure of SPD_1333 with twenty-eight strands; (b) Ligand binding site prediction by metaPocket server. Pocket 1, 2, and 3 represents the top three predicted sites. Ligand molecule containing dilysine peptide motif is represented in sticks

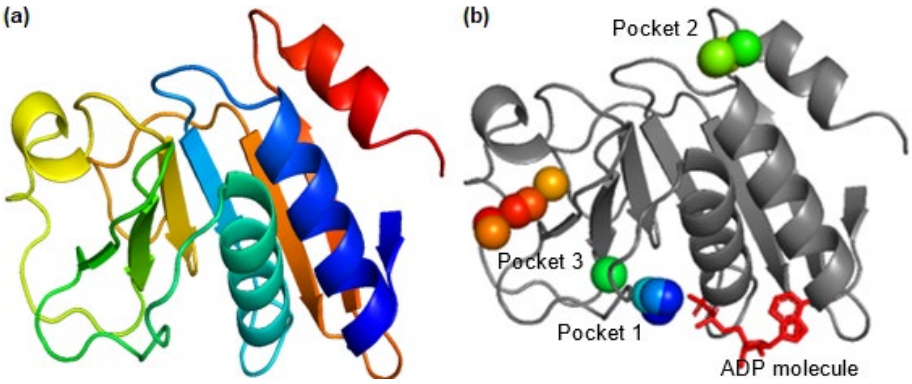

Fig 2. (a) Refined three-dimensional structure of SPD_1743 with four helices and seven strands; (b) Ligand binding site prediction by metaPocket server. Pocket 1, 2, and 3 (spheres) represents the top three predicted sites. Adenosine diphosphate (ADP) molecule is presented in sticks

In term of the secondary structure prediction, PSIPRED predicted SPD_1473 to have four helices and seven strands altogether. Similarly, tertiary structure prediction and refinement by I-TASSER and GROMACS using a functionally unknown 1HTW_A as a template (98.6% aligned, 27.6% identity) showed the same number of helices and strands (Fig. 2(a)).

Finally, ligand prediction by COACH predicted an adenosine diphosphate (ADP) molecule to bind at pocket 1 of the protein (Fig. 2(b)). The binding property is structurally similar to a protein of unknown function isolated from *Haemophilus influenza* (1HTW_A). Adenosine diphosphate, or alternatively known as adenosine pyrophosphate, is a hydrolyzed form of adenosine triphosphate (ATP), an organic molecule that involves in vital cellular processes such as cell respiration [44]. As evidently proven by a study done to 1HTW_A protein [45], the binding probability of SPD_1743 with the hydrolyzed ATP molecule may suggest its function in ATPase activity. Targeting SPD_1743 may alter its function in maintaining cell respiration hence provoking the viability of *S. pneumoniae* strain D39.

■ CONCLUSION

The analyses done on all eleven proteins revealed that seven of the proteins are classified under protein domain or family that involves in either pathogenicity or viability of *S. pneumoniae*. Furthermore, based on the

binding ligand assessment, five out of eleven proteins were strongly predicted to be involved in the pathogenesis and four in the survival mechanism of *S. pneumoniae* strain D39. Finally, the sequence- and structure- based assessments also showed that SPD_1333 (predicted to involve in cell pathogenicity mechanism) and SPD_1743 (predicted to involve in cell viability mechanism) are the best candidates to be further studied.

By using *in silico* sequence- and structure- based approaches, this study had successfully filled the information gap of previously un-annotated essential proteins in *S. pneumoniae* strain D39 by predicting probable physicochemical, functional and structural properties of selected hypothetical proteins.

■ ACKNOWLEDGMENTS

We would like to thank all staff at the Kulliyyah of Science, International Islamic University Malaysia, for their assistance. This study is funded by the IIUM RIGS research grant (RIGS16-312-0476) and FRGS research grant from the Malaysian Ministry of Education (FRGS/1/2016/SKK11/UIAM/02/1).

■ REFERENCES

[1] World Health Organization, *Pneumococcal Disease*, https://www.who.int/biologicals/vaccines/pneumococcal/en/, accessed on February 19, 2019.

[2] Weiser, J.N., Ferreira, D.M., and Paton, J.C., 2018, *Streptococcus pneumoniae*: Transmission, colonization and invasion, *Nat. Rev. Microbiol.*, 16 (6), 355–367.

[3] Henriques-Normark, B., and Tuomanen, E.I., 2013, The pneumococcus: Epidemiology, microbiology, and pathogenesis, *Cold Spring Harb. Perspect. Med.*, 3 (7), a010215.

[4] Song, J.H., 2013, Advances in pneumococcal antibiotic resistance, *Expert Rev. Respir. Med.*, 7 (5), 491–498.

[5] Cherazard, R., Epstein, M., Doan, T.L., Salim, T., Bharti, S., and Smith, M.A., 2017, Antimicrobial resistant *Streptococcus pneumoniae*, *Am. J. Ther.*, 24 (3), e361–e369.

[6] Lipsitch, M., and Siber, G.R., 2016, How can vaccines contribute to solving the antimicrobial resistance problem?, *MBio*, 7 (3), 00428-16.

[7] Rodgers, G.L., and Klugman, K.P., 2016, Surveillance of the impact of pneumococcal conjugate vaccines in developing countries, *Hum. Vaccin. Immunother.*, 12 (2), 417–420.

[8] Mitchell, A.M., and Mitchell, T.J., 2010, *Streptococcus pneumoniae*: Virulence factors and variation, *Clin. Microbiol. Infect.*, 16 (5), 411–418.

[9] Hyams, C., Camberlein, E., Cohen, J.M., Bax, K., and Brown, J.S., 2010, The *Streptococcus pneumoniae* capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms, *Infect. Immun.*, 78 (2), 704–715.

[10] Mostowy, R., Croucher, N.J., Hanage, W.P., Harris, S.R., Bentley, S., and Fraser, C., 2014, Heterogeneity in the frequency and characteristics of homologous recombination in pneumococcal evolution, *PLoS Genet.*, 10 (5), 1004300.

[11] Jedrzejas, M.J., 2001, Pneumococcal virulence factors: Structure and function, *Microbiol. Mol. Biol. Rev.*, 65 (2), 187–207.

[12] Dahlström, K.M., 2015, From Protein Structure to Function with Bioinformatics, *Dissertation*, Faculty of Science and Engineering, Åbo Akademi University, Turku, Finland.

[13] Wuchty, S., Rajagopala, S.V., Blazie, S.M., Parrish, J.R., Khuri, S., Finley, R.L., and Uetz, P., 2017, The protein interactome of *Streptococcus pneumoniae* and bacterial meta-interactomes improve function predictions, *mSystems*, 2 (3), 00019-17.

[14] Liu, X., Kjos, M., Sorg, R.A., Veening, J., van Kessel, S.P., Zhang, J., Knoops, K., Slager, J., Domenech, A., and Gallay, C., 2017, High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae, *Mol. Syst. Biol.*, 13 (5), 931.

[15] Gupta, A., Kapil, R., Dhakan, D.B., and Sharma, V.K., 2014, MP3: A software tool for the prediction of pathogenic proteins in genomic and metagenomic data, *PLoS One*, 9 (4), e93907.

[16] Pearson, W.R., 2013, An introduction to sequence similarity ("homology") searching, *Curr. Protoc.*

Bioinf., 42 (1), 3.1.1–3.1.8.

[17] Bairoch, A., Gattiker, A., Wilkins, M.R., Gasteiger, E., Duvaud, S., Appel, R.D., and Hoogland, C., 2009, "Protein Identification and Analysis Tools on the ExPASy Server", in *The Proteomics Protocols Handbook*, Eds. Walker, J.M., Humana Press, 571–607.

[18] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E., and Finn, R.D., 2019, The Pfam protein families database in 2019, *Nucleic Acids Res.*, 47 (D1), D427–D432.

[19] Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., and Bryant, S.H., 2017, CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures, *Nucleic Acids Res.*, 45 (D1), D200–D203.

[20] Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., and Brinkman, F.S.L., 2010, PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, *Bioinformatics*, 26 (13), 1608–1615.

[21] Tusnády, G.E., and Simon, I., 2001, The HMMTOP transmembrane topology prediction server, *Bioinformatics*, 17 (9), 849–850.

[22] Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H., 2019, SignalP 5.0 improves signal peptide predictions using deep neural networks, *Nat. Biotechnol.*, 37, 420–423.

[23] Bendtsen, J.D., Jensen, L.J., Blom, N., von Heijne, G., and Brunak, S., 2004, Feature-based prediction of non-classical and leaderless protein secretion, *Protein Eng. Des. Sel.*, 17 (4), 349–356.

[24] Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J., and von Mering, C., 2019, STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic Acids Res.*, 47 (D1), D607–D613.

[25] Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K., and Jones, D.T., 2013, Scalable web services for the PSIPRED Protein Analysis Workbench, *Nucleic Acids Res.*, 41 (W1), W349–W357.

[26] Zhang, Y., 2008, I-TASSER server for protein 3D structure prediction, *BMC Bioinf.*, 9 (1), 40.

[27] Chen, C.C., Hwang, J.K., and Yang, J.M., 2006, $(PS)^2$: Protein structure prediction server, *Nucleic Acids Res.*, 34 (Web Server), W152–W157.

[28] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T., 2018, SWISS-MODEL: Homology modelling of protein structures and complexes, *Nucleic Acids Res.*, 46 (W1), W296–W303.

[29] Eisenberg, D., Lüthy, R., and Bowie, J.U., 1997, VERIFY3D: Assessment of protein models with three-dimensional profiles, *Methods Enzymol.*, 277, 396–404.

[30] Benkert, P., Biasini, M., and Schwede, T., 2011, Toward the estimation of the absolute quality of individual protein structure models, *Bioinformatics*, 27 (3), 343–350.

[31] Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B., and Lindahl, E., 2015, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX*, 1-2, 19–25.

[32] Huang, B., 2009, MetaPocket: A meta approach to improve protein ligand binding site prediction, *OMICS*, 13 (4), 325–330.

[33] Yang, J., Roy, A., and Zhang, Y., 2013, Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment,

Bioinformatics, 29 (20), 2588–2595.

[34] Wang, F., Xiao, J., Pan, L., Yang, M., Zhang, G., Jin, S., and Yu, J., 2008, A systematic survey of mini-proteins in bacteria and archaea, *PLoS One*, 3 (12), e4027.

[35] Wan, S., Duan, Y., and Zou, Q., 2017, HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source, *Proteomics*, 17 (17-18), 1700262.

[36] Davis, M.J., Hanson, K.A., Clark, F., Fink, J.L., Zhang, F., Kasukawa, T., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Teasdale, R.D., 2006, Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units, *PLoS Genet.*, 2 (4), e46.

[37] Rydzewski, J., Jakubowski, R., and Nowak, W., 2015, Communication: Entropic measure to prevent energy over-minimization in molecular dynamics simulations, *J. Chem. Phys.*, 143 (17), 171103.

[38] Chen, J., Almo, S.C., and Wu, Y., 2017, General principles of binding between cell surface receptors and multi-specific ligands: A computational study, *PLOS Comput. Biol.*, 13 (10), e1005805.

[39] Miclet, E., Stoven, V., Michels, P.A.M., Opperdoes, F.R., Lallemand, J.Y., and Duffieux, F., 2001, NMR spectroscopic analysis of the first two steps of the pentose-phosphate pathway elucidates the role of 6-phosphogluconolactonase, *J. Biol. Chem.*, 276 (37), 34840–34846.

[40] Andersson, H., Kappeler, F., and Hauri, H.P., 1999, Protein targeting to endoplasmic reticulum by dilysine signals involves direct retention in addition to retrieval, *J. Biol. Chem.*, 274 (21), 15080–15084.

[41] Ruhanen, H., Hurley, D., Ghosh, A., O'Brien, K.T., Johnston, C.R., and Shields, D.C., 2014, Potential of known and short prokaryotic protein motifs as a basis for novel peptide-based antibacterial therapeutics: a computational survey, *Front. Microbiol.*, 5, 4.

[42] Luthra, A., Swinehart, W., Bayooz, S., Phan, P., Stec, B., Iwata-Reuyl, D., and Swairjo, M.A., 2018, Structure and mechanism of a bacterial t6A biosynthesis system, *Nucleic Acids Res.*, 46 (3), 1395–1411.

[43] Miyauchi, K., Kimura, S., and Suzuki, T., 2013, A cyclic form of N6-threonylcarbamoyladenosine as a widely distributed tRNA hypermodification, *Nat. Chem. Biol.*, 9 (2), 105–111.

[44] Bugreev, D.V., and Mazin, A.V., 2004, Ca^{2+} activates human homologous recombination protein Rad51 by modulating its ATPase activity, *Proc. Natl. Acad. Sci. U.S.A.*, 101 (27), 9988–9993.

[45] Teplyakov, A., Obmolova, G., Tordova, M., Thanki, N., Bonander, N., Eisenstein, E., Howard, A.J., and Gilliland, G.L., 2002, Crystal structure of the YjeE protein from *Haemophilus influenzae*: A putative ATPase involved in cell wall synthesis, *Proteins Struct. Funct. Genet.*, 48 (2), 220–226.