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

Article

Functional Annotation of Selected *Vibrio cholerae* Hypothetical Proteins

Sarah Nur Syaza Mohd Yunos[1,a], Azzmer Azzar Abdul Hamid[1,b, 2], Noor Hasniza Md Zin[1,c], Noraslinda Muhamad Bunnori[1,d, 2], Hanani Ahmad Yusof[3], Kamarul Rahim Kamarudin[4] and '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]sarahnursyaza94@yahoo.com, [b]azzmer@iium.edu.my, [c]hasnizamz@iium.edu.my, [d]noraslinda@iium.edu.my, [e]mraisyah@iium.edu.my

[2] *Research Unit Bioinformatic and Computational Biology, Kulliyyah of Science, International Islamic University Malaysia (IIUM), Jalan Sultan Ahmad Shah, 25200 Kuantan, Pahang, Malaysia*
E-mail: azzmer@iium.edu.my

[3]*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

[4]*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*— *Vibrio cholerae (V. cholerae)* secreted an enterotoxin that induces acute diarrhea called cholera. Cholera if left untreated, lead to renal failure, shock, hypokalemia and pulmonary edema, which can cause death within hours. This disease occurs due to the bacterial virulence factors machinery. Previous studies have shown that the essential genes of *V. cholerae* O1 biovar El Tor N16961 strain are highly important in the bacterial growth, survival and its virulent properties. However, 45 of the essential genes were categorized as hypothetical genes with no known function and structure. Thus, this *in silico* study aimed to functionally and structurally annotate these essential hypothetical genes. All of the 45 hypothetical genes primarily underwent screening process for its pathogenicity and template availability. After screening, 11 of them were selected for further physicochemical categorization, functional and structural characterization using bioinformatics tools. From the data collected, all of the 11 hypothetical proteins are either involve in translation, cell transportation, cell growth or cell defense mechanism. All of the 11 hypothetical proteins were annotated, with five of them being the most promising proteins for further analysis. The finding of this study could provide an insight on the *V. cholerae* O1 biovar El Tor N16961 mechanism of pathogenesis, which could be useful for target identification for vaccine or drug design in order to reduce the fatality of cholera disease.**

***Keywords*— Hypothetical proteins; *Vibrio cholerae; In silico* analysis of protein; Bioinformatics tools.**

I. INTRODUCTION

Vibrio cholerae (V. cholerae) is a bacterial species that belong to the family of *Vibrionaceae*. This gram-negative, facultative anaerobic, comma-shaped rods with a single polar flagellum bacteria is the agent that secretes enterotoxin that induces severe diarrhea known as cholera [1]. Cholera is an acute diarrheal disease that is caused by ingestion of food or water contaminated with bacteria *V. cholerae*. The major subgroups which caused the outbreak of cholera are *V. cholera* Ol and *V. cholerae* Ol39 [2]. This pandemic disease is mainly caused by the bacterial virulence factor mechanism that colonized the intestinal lumen, which lead to various symptoms such as dehydration and diarrhea. Even though many research and clinical studies have been done, cholera remains to be a serious threat worldwide.

From 2010 till now, cholera has continuously caused significant problem worldwide, with the massive outbreak in Haiti and Yemen, and the sudden endemic disease around the sub-Saharan Africa and southern Asia [3]. This is why the re-emergence of cholera has become a public concern once again. Raise in concern is caused by several factors that includes the recent active cholera occurrence, the emergence of new *V. cholerae* strains that lead to higher severity in clinical symptoms, antimicrobial resistance and antibiotic resistance [4]. Thus, the availability of new and reliable cholera vaccines to elicit protective immunity in targeted population are highly anticipated [4]. To date, there are two

types of oral cholera vaccines available, however, it is in limited quantities and it also has a limited protective efficacy [5]. Presently, researchers are trying to design better drugs and vaccine in order to control this disease. Thus, greater efforts in determining the virulence factors of *V. cholerae* has been taken, in the move to reduce cholera cases and outbreaks.

V. cholerae O1 and O139 major virulence factors are toxin-coregulated pilus (TCP), cholera toxin (CT) and motility. Toxin-coregulated pilus (TCP) is an IV pilus that mediated adherence and the formation of microcolony that is required for intestinal colonization in host such as mice and human. TCP expression is linked to the production of Cholera Toxin (CT). Cholera toxin (CT) is an AB_5 family ADP-ribosyltransferase which caused the profuse rice-watery diarrhea disease. The toxin binds to a specific receptor, monosialosyl ganglioside GM1, on the outer surface of intestinal epithelial cells plasma membrane and secretes an enzymatically active factor that causes the elevation of cyclic adenosine 5-monophosphate (cAMP) production. High cAMP inside the cell will cause excessive secretion of electrolytes and water into the intestinal lumen [1]. Several studies have suggested that flagellar motility also contributes in the mechanism of virulence gene expression [6].

From the NCBI database, the total number of *V. cholerae* serogroups are 206, and they are classified based on the heat-stable polysaccharides of the somatic (O) antigen. However, from the 206 serogroups, only two were recorded as toxigenic strains, serogroups O1 and O139. Both have been found to be the major contributors of the epidemic cholera outbreaks. There are two biotypes of *V. cholerae* O1, Classical and El Tor, each with two different serotypes, Ogawa and Inaba. Between the Classical and El Tor strains, El Tor remained longer in the environment, as it caused the seventh cholera pandemics that started in 1961 and still ongoing till today. Currently, there are six *V. cholerae* O1 biovar El Tor strains recorded in the National Center for Biotechnology Information (NCBI). However, there is only one strain that has a complete genomic sequence. A study performed in Heiderberg *et al.* [7], determined that a complete genomic sequence of *V. cholerae* O1 biovar El Tor strain N16961 has more than four million base pairs (bp) which made up two circular chromosomes. Work in [8] has normalized the data by removing the biases of the genes location in order to categorize the essential and nonessential genes. This study used high-resolution analysis to determine the essentiality of all genes in the *V. cholerae* genome. Essential genes play important role in bacterial growth, survival and regulations. Using a hidden Markov model (HMM)-based filter, there were 343 *V. cholerae* essential genes and 13% (45 genes) of these genes were categorized as hypothetical proteins [8].

In the past few years, although hundreds of bacterial genomes has been sequenced and stored in the databases, most of its protein functions were still uncharacterized. Due to this factor, there are inclining demand for functional and structural annotation of these unknown proteins which are called “hypothetical proteins” [9]. The hypothetical protein functions are extremely important to molecular biologists. In order to understand the virulence factors machinery of this pathogen, comprehensive knowledge on the proteins functions and structures is important. Currently, there are many bioinformatics tools available to annotate the functional and structural properties of the desired protein. However, studies on these uncharacterized proteins in the databases are still lacking and many remain unknown. As mentioned in Ijaq et al. [10], there were more than 48 million hypothetical proteins sequence recorded in the National Centre for Biotechnology Institute (NCBI) databases during that year [10].

In this study, the objective was to fill the gap between genome sequence information and virulent protein annotation by determining the physicochemical properties of the hypothetical proteins, family and domain prediction, subcellular localization, secretome analysis, protein-protein interaction, three-dimensional protein structure modeling and lastly active sites and ligand binding prediction through computational approaches. This study is significant due to the importance of proper understandings of *V. cholerae* hypothetical protein structures and improving the functional annotation for future research. Computational annotations of the hypothetical proteins of *V. cholerae* are important in providing the insight view of the protein molecular function and structure. Furthermore, the data obtained from this project will offer opportunity for further analysis, such as gene cloning and protein expression to validate the *in silico* findings. The hypothetical proteins were chosen based on the annotated *V. cholerae* O1 biovar El Tor N16961 strain [8] and were analysed using several bioinformatics tools.

II. THE MATERIAL AND METHOD

The methods used in this study include all the bioinformatics program and databases listed in Table I. First, sequence of hypothetical genes were retrieved from the genomic data of *V. cholerae* O1 biovar El Tor N16961 strain and its corresponding protein sequence were analysed, followed by virulence factors analysis, physicochemical characterization, function prediction, protein interaction, structure prediction and lastly active site prediction. The final data of all the proteins were summarized and five suitable hypothetical proteins were annotated.

TABLE I
LIST OF BIOINFORMATICS PROGRAM AND DATABASES FOR FUNCTIONAL AND STRUCTURAL ANNOTATION OF *V. CHOLERAE* HYPOTHETICAL PROTEINS

Methodology	**Bioinformatics Program / Databases**	**Reference**
Sequence retrieval	Universal Protein Knowledgebase (UniProt KB)	[11]
Virulence factors analysis	VICMpred	[12]
	MP3: Prediction of Pathogenic/Virulent Proteins	[13]
Homologs PDB template availability	NCBI Basic Local Alignment Search Tool: Protein (NCBI blastp)	[14]
	NCBI Position-Specific Iterated BLAST (PSI-BLAST)	[14]
Physicochemical characterization	Expert Protein Analysis System: Protein Parameter (ExPASy – ProtParam)	[15]
Domain and family identification	Protein Families Database (Pfam)	[16]
	NCBI Conserved Domain Search Service (CD-Search)	[17]

Methodology	Bioinformatics program / databases	Reference
Subcellular localization and secretome analyses	Protein Subcellular Localization Prediction Tool (PSORTb)	[18]
	SignalP 4.0 Server (SignalP)	[19]
	SecretomeP 2.0 Server (SecretomeP)	[20]
Protein-protein interaction	Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	[21]
Secondary structure prediction	PSI-BLAST Based Secondary Structure Prediction (PSIPRED)	[22]
Tertiary structure prediction	Iterative Threading Assembly Refinement (I-TASSER)	[23]
	Protein Structure Prediction Server [$(PS)^2$]	[24]
	Expert Protein Analysis System: SWISS-MODEL (ExPaSy SWISS-MODEL)	[25]
Tertiary structure validation	RAMPAGE: Ramachandran Plot Assessment	[26]
	Verify3D: Assessment of Protein Models with Three-Dimensional Profiles	[27]
	ExPaSy SWISS-MODEL: QMEAN4	[28]

A. Sequence Retrieval

Based on the ID number of the 45 genes, the genome of *V. cholerae* was analysed in NCBI website (https://www.ncbi.nlm.nih.gov) and found that all of the genes were present and characterized as hypothetical genes. For the further characterization process that follows, their fasta sequences were retrieved from UniProt (http://www.uniprot.org).

B. Homologs PDB Structure Availability

BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to examine the availability of structural homologs in Protein Data Bank (PDB) [29]. This search was performed together with Position-Specific Iterated BLAST (PSI-BLAST) to scan a set of predetermined position-specific scoring matrices of the desired protein [30]. Homolog structures that have a template with query coverage higher than 50% and identity of 30% to 70% were chosen.

C. Virulence Factors Analysis

Based on the essential hypothetical protein categorized by reference [8], all of the 45 genes were subjected to VICMpred (http://crdd.osd d.net/raghava/vicmpred/) and MP3 (http://metagenomics.iise rb.ac.in/mp3/algorithm.php) servers to identify the virulence factors. Virulent proteins were described as potential targets for developing drugs or vaccine as they involve in the infection and colonization of the pathogenic bacteria. Proteins that are responsible in virulence-associated factors were chosen for annotation.

D. Physicochemical Categorization

The hypothetical proteins physicochemical properties were determined by using ExPASy ProtParam (https://web.expasy.org/protparam/). This server theoretically measures the physicochemical characterization of a protein, such as theoretical isoelectric point (pI), molecular weight, extinction coefficient, total number of positive and negative residues, instability index, aliphatic index and grand average hydropathicity (GRAVY) [1].

E. Hypothetical Protein Domain(s) and Family(s)

The server that was used were to study protein domain and family of the hypothetical protein was Pfam (https://pfam.xfam.org). Pfam is a software designed as a comprehensive and accurate collection of protein domains families [31]. The data obtained was then further analyzed to compare the conserved domain by using the CD-Search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

F. Sub-cellular Localization

Then, the hypothetical proteins were subjected to sub-cellular localization analysis. The knowledge of sub-cellular localization is important in characterizing a protein as drug or vaccine target. Protein that localized in the cytoplasm can act as possible drug targets, whilst the surface membrane proteins can be considered as potential vaccine targets [29].

G. Protein-protein Interaction

The protein interaction with other protein in the cell was studied by using the Search Tool for Retrieval of Interacting Genes (STRING) (https://string-db.org) [21]. STRING consists of a large repository of protein-protein interactions that involves functional interactions, stable complexes and regulatory interactions among proteins. This server enables us to understand the individual function of the hypothetical proteins.

H. Secondary Structure Prediction

Before predicting the tertiary structure of the query protein, the fasta sequence was analyzed using PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/) to determine its secondary structure. The secondary structure of a protein is mainly defined by the pattern of hydrogen bonding between the backbone amino and carboxyl group. This prediction gives information on numbers of alpha helices, beta sheets and loops present in shaping a protein structure.

I. Tertiary Structure Prediction

The protein tertiary structure was predicted using three different servers. These are the I-TASSER (https://zhanglab.ccm b.med.umich.edu/I-TASSER/), ExPASy SWISS-MODEL (https://swissmodel.expasy.org) and $(PS)^2$ (http://ps2.life.nc tu.edu.tw). The servers were used to compare which server gives the best result of the predicted structure.

III. RESULTS AND DISCUSSION

A. Target Selection

From the study stated in Chao et al. [8], there are 45 hypothetical proteins of *V. cholerae* categorized as essential proteins. The protein's amino acid sequence was subjected to NCBI BLAST to retrieve the available homologue(s) template for the proteins structure. Hypothetical proteins that were found to have homologue template with query coverage higher than 50% and identity range from 30% to 70% was chosen for this study (Table II). These hypothetical proteins were then analysed to predict its pathogenicity by

using VICMpred and MP3. From this screening process, 11 proteins were selected for further analysis.

TABLE II
THE HOMOLOGUE TEMPLATE AND PATHOGENICITY ANALYSIS OF 11 SELECTED HYPOTHETICAL PROTEINS OF *V. CHOLERAE*

Gene ID	Homology Template		Virulence Analysis	
	Query coverage (%)	Identity similarity (%)	VICMpred	MP3
VC_0004	99	58	Virulence factor	Non-pathogenic
VC_0358	98	33	Cellular process	Pathogenic
VC_0849	93	30	Cellular process	Pathogenic
VC_1884	54	46	Metabolism molecule	Pathogenic
VC_2500	98	43	Metabolism molecule	Pathogenic
VC_2499	99	49	Metabolism molecule	Non-pathogenic
VC_0357	100	50	Cellular process	Non-pathogenic
VC_0519	100	34	Cellular process	Non-pathogenic
VC_1127	99	53	Metabolism molecule	Non-pathogenic
VC_1912	95	51	Cellular process	Non-pathogenic
VC_0850	85	52	Cellular process	Non-pathogenic

B. Physicochemical Characteristics

Physicochemical characteristics of the hypothetical proteins were summarised in Table III. Seven out of the eleven hypothetical proteins possessed a pI value of lower than 7.0 which indicates that the proteins have acidic side chains with extra negative charge. Whilst, higher pI value such as 9.08 and 9.05 for VC_1127 and VC_0850, respectively, showed that the protein has a basic side chain with extra positive charge. The instability index of all the eleven protein showed that only five proteins (VC_0358, VC_1884, VC_2500, VC_2499 and VC_0519) were stable whilst the rest were classified as not stable. These unstable proteins are highly sensitive and easily precipitate if it is not handled properly. The unstable proteins may require additional steps such as denaturation prior to the isolation and purification process.

TABLE III
PHYSICOCHEMICAL CHARACTERIZATION OF SELECTED HYPOTHETICAL PROTEIN BY EXPASY PROTPARAM SERVER

Gene ID	MW (kDa)	pI	(+)	(-)	II	AI	GRAVY
VC_0004	60.62	6.20	43	46	41.87	91.07	-0.035
VC_0358	10.36	5.06	6	12	29.58	120.99	0.185
VC_0849	15.88	6.11	14	15	42.73	77.85	0.047
VC_1884	43.75	6.23	28	30	36.43	114.31	0.452
VC_2500	40.89	8.53	35	33	39.55	117..54	0.347
VC_2499	39.19	6.77	32	32	28.47	119.16	0.528
VC_0357	13.27	4.61	5	16	46.85	109.92	0.090
VC_0519	16.07	5.3	22	25	38.80	101.63	-0.214
VC_1127	22.77	9.08	24	19	46.37	106.15	0.010
VC_1912	44.55	7.25	49	49	45.98	98.12	-0.339
VC_0850	11.61	9.05	17	15	68.10	106.14	-0.350

C. Protein Domains and Families

Protein domain is the conserved region of a protein known for a certain role of the protein. Family is a group of protein that share similar evolutionary origin with related functions. From the 11 selected hypothetical protein, 10 of them were classified into a specific domain(s) or family(s), however, there were no record of protein family for protein VC_1127 (Table IV).

TABLE IV
IDENTIFICATION OF HYPOTHETICAL PROTEINS DOMAINS AND FAMILIES

Gene ID	Pfam	CD-search
VC_0004	YidC periplasmic domain 60Kd inner membrane protein	60 kDa inner membrane protein
VC_0358	DsrH like protein family	DsrH family protein
VC_0849	Polyketide cyclase / dehydrase and lipid transport	START/RHO_alpha_C/ PITP/Bet_v1/CoxG/Cal C (SRPBCC) ligand-binding domain
VC_1884	MacB-like periplasmic core domain (MacB_PCD) FtsX-like permease family (FtsX)	MacB_PCD super family
VC_2500	Predicted permease YjgP/YjgQ family	LPS export ABC transporter permease LptF (LptF_YjgP)
VC_2499	Predicted permease YjgP/YjgQ family	Lipopolysaccharide ABC transporter permease LptG (YjgP_YjgQ)
VC_0357	DsrE/DsrF-like family	DsrE family Sulfur relay protein TusC/DsrF
VC_0519	Yqey-like protein	Yqey-like protein
VC_1127	Protein of unknown function (DUF489)	Lysogenization protein HflD (DUF489)
VC_1912	Tetratricopeptide repeat (TPR_7)	Protein Classification lipopolysaccharide assembly protein LapB (YciM)
VC_0850	RnfH family Ubiquitin	TGS domain

The first hypothetical protein, VC_0004 belongs to YidC periplasmic domain that has a function as membrane protein insertase independent of the Sec protein-conducting channel. YidC can also assist in the lateral integration and folding of membrane proteins that insert into the membrane via the Sec pathway [32]. Sec pathway possesses many roles and one of them is to promote transportation of virulence proteins [33]. For protein VC_0358, its domain is DsrH which involved in oxidation of intracellular sulphur (Table IV), however the clear role of this domain are remain elusive [34]. VC_0849 belongs in SRPBCC (START/RHO_alpha_C/PITP/Bet_v1/ CoxG/CalC). SRPBCC domain has a deep hydrophobic ligand-binding pocket. A previous study [35] showed that the *V. cholerae* mechanism of adhesion is controlled by both specific and nonspecific interaction. Nonspecific hydrophobic interactions such as SRPBCC can assist in regulating the adherence of *V. cholerae* in human [36]. CD-SEARCH predicted that protein VC_1127 belongs to lysogenization protein HfID family. This family plays an important role in toxigenic effect of CTXϕ lysogenic bacteriophage that carries genes encoding the pilus

colonization factor TCP [37]. The domain and family of protein VC_1912 predicted by Pfam exhibited that the protein belongs to tetratricopeptide repeats (TPR_7) whilst CD-SEARCH suggested that the protein is a lipopolysaccharide assembly protein LapB (YciM).

D. Subcellular Localization and Secretome Analyses

Based from the analysis, five proteins were predicted to localize in the cytoplasmic whilst, five in the inner membrane with one protein located at the plasma membrane. *V. cholerae* invades the epithelial lining cells of the host by excreting a certain type molecules. These secreted proteins can promote cell adhesion, recognition and invasion. From the SignalP server, none of the protein has a signal peptide whilst, analysis using SecretomeP server showed that only one protein which is VC_0004 could involve in the secretory pathway (Table V).

TABLE V
HYPOTHETICAL PROTEINS SUBCELLULAR LOCALIZATION AND SIGNAL PEPTIDES ANALYSIS

Gene ID	Subcellular Localization Prediction		Signal Peptides	
	PSORT	**PSORTb**	**SignalP**	**SecretomeP**
VC_0004	Bacterial inner membrane	Cytoplasmic membrane	No	Possibly (0.598)
VC_0358	Bacterial cytoplasm	Unknown	No	No (0.044)
VC_0849	Bacterial cytoplasm	Unknown	No	No (0.249)
VC_1884	Plasma membrane	Cytoplasmic membrane	No	No (0.153)
VC_2500	Bacterial inner membrane	Cytoplasmic membrane	No	No (0.099)
VC_2499	Bacterial inner membrane	Cytoplasmic membrane	No	No (0.220)
VC_0357	Bacterial cytoplasm	Unknown	No	No (0.021)
VC_0519	Bacterial cytoplasm	Cytoplasmic	No	No (0.078)
VC_1127	Bacterial inner membrane	Unknown	No	No (0.140)
VC_1912	Bacterial inner membrane	Cytoplasmic	No	No (0.157)
VC_0850	Bacterial cytoplasm	Cytoplasmic	No	No (0.086)

E. Protein-Protein Interaction

The data for the protein-protein network that have close interaction with the hypothetical proteins from STRING server is summarised in Table VI. The involvement of the protein in virulence factor machinery is influenced by its interactions with other proteins. Some proteins work in synergy in order to perform vital cellular functions. Hence, knowing the relationship between a hypothetical protein and other proteins could provide an insight into its possible function or role. The analysis of protein-protein interaction by STRING gives information on the types of relation (neighborhood, co-occurrence, text-mining and experimental) between the query protein and others.

Out of the 11 hypothetical proteins studied, five of them have been linked in either direct or indirect relationships to the pathogenic pathway of the bacteria (Table VI). Protein VC_0004, interact closely with FtsY, and SecY which are proteins that play important roles in protein secretion system. FtsY is chaperone that delivers protein to SecA, which is a membrane transporter. Then, this receptor will act as a motor to push the protein across the membrane via specific protein channel such as SecY and SecE [33]. Since many pathogenicity factors are secreted, the respective protein channels could be a potential drug target. VC_0358 relates closely with VC_0354 which is FKBP-type peptidylprolyl isomerase that many studies claimed that it plays as a secondary role in virulence such as improper folding or secretion of virulence factors [37]. VC_0849 interacts closely to proteins such as ubiG and ubiE, enzyme that catalyzes the chemical reaction that produce ubiquinone-9. Ubiquinone is a compound that facilitates the electron-transfer mechanism in living cells such as VcDsbA. *In vitro* assay showed that VcDsbA participate in the redox pathway that senses the presence of the bile salts in the small intestine that activates virulence gene expression in *V. cholerae*.

One of the protein interacted with VC_1127 is VC_1836 a translocation protein TolB. TolB is present in almost all Gram-negative bacteria. It is a periplasmic component of the Tol-Pal system that connects the cytoplasmic membrane with the outer membrane. The essentiality of *tolB* gene was shown in a study [38], which demonstrated that the depletion of TolB, inhibits the viability of a gram-negative bacteria, *P. aeruginosa, in vitro* and markedly reduces its persistence as well as its pathogenicity in an animal infection model. It also showed reduction in resistance to human serum and several antibiotics [39]. Lastly, VC_1912 have close relationships with protein VC_0118. VC_0118 (uroporphyrin-III C-methyltransferase) is a multifunctional protein. It is one of the possible drug target protein of *Vibrio parahaemolyticus* in the Drug Target Protein Database (DTP) [40].

Thus, the five hypothetical proteins mentioned, which were VC_0004, VC_0358, VC_0849, VC_1127 and VC_1912 can be a potential target protein as some of their neighboring proteins involved either directly or indirectly with the bacteria's virulence factor machinery.

TABLE VI
PROTEIN-PROTEIN INTERACTION OF HYPOTHETICAL PROTEIN WITH FUNCTIONALLY IMPORTANT PROTEIN USING STRING SERVER

Gene ID	STRING
VC_0004	**secY** Description: Preprotein translocase subunit SecY. (0.988) **ftsY** Description: Cell division protein FtsY (0.948)
VC_0358	**VC0354** Description: FKBP-type peptidylprolyl isomerase. (0.526) **VC1356** Description: Sulfur relay, TusE/DsrC/DsvC family protein. (0.721) **tusD** Description: Sulfur transfer complex subunit TusD. (0.975)
VC_0849	**ubiA** Description: 4-hydroxybenzoate octaprenyltransferase. (0.835) **nadK** Description: Inorganic polyphosphate/ATP-NAD kinase. (0.745) **fabG** Description: 3-ketoacyl-ACP reductase. (0.788)

Gene ID	STRING
VC_1884	**VC2252** Description: Outer membrane protein assembly factor YaeT. (0.893) **VC1107** Description: Outer membrane lipocarrier protein LolA. (0.970)
VC_2500	**VC2528** Description: ABC transporter ATP-binding protein. (0.984) **VC1959** Description: Septum formation. (0.747)
VC_2499	**VC2528** Description: ABC transporter ATP-binding protein. (0.997) **VC2252** Description: Outer membrane protein assembly factor YaeT. (0.770)
VC_0357	**VC0356** Description: Sulfur transfer complex subunit TusD. (0.999) **VC0359** Description: Ribosomal protein S12. (0.678)
VC_0519	**VC2459** Description: DNA repair protein RecO. (0.513) **VC_0517** Description: RNA polymerase sigma factor RpoD. (0.678)
VC_1127	**VC1126** Description: Adenylosuccinate lyase. (0.858) **VC1836** Description: Translocation protein TolB. (0.467)
VC_1912	**VC0118** Description: uroporphyrin-III C-methyltransferase. (0.659) **VC1914** Description: Integration host factor subunit beta. (0.724)
VC_0850	**VC0847** Description: Phage family integrase. (0.745) **VC1016** Description: Electron transport complex protein RnfB. (0.875)

F. Structure Prediction

The protein structures were predicted using three different servers, I-TASSER, ExPASy SWISS-MODEL and $(PS)^2$. The structural models from all these three servers were compared and the best protein structure model is selected for further structure refinement. This will improve the quality of structure models (low QMEAN4 score or high Verify3D percentage) and allow better prediction of their active site and possible ligand binding. Table VII showed the quality of best protein models for each protein.

TABLE VII
SUITABLE PROTEIN THREE-DIMENSIONAL STRUCTURE TEMPLATE RETRIEVED FROM DATABASE

Gene ID	PDB ID	Species	QMEAN4 Score	Verify3D
VC_0004	3wvf.1.A	*Escherichia coli*	-3.14	78.72%
VC_0358	2d1p.1.C	*Escherichia coli*	-1.00	92.31%
VC_0849	1t17.1.A	*Caulobacter crescentus*	-5.08	83.33%
VC_1884	5naa.1.A	*Escherichia coli*	-1.43	51.48%
VC_2500	5175.1.C	*Klebsiella pneumoniae*	-8.34	44.54%
VC_2499	5175.1.D	*Klebsiella pneumoniae*	-5.56	48.30%
VC_0357	2d1p.B	*Escherichia coli*	-1.94	90.68%
VC_0519	1ng6.1.A	*Bacillus subtilis*	0.16	84.35%
VC_1127	1sdi.1.A	*Escherichia coli*	-0.96	98.54%
VC_1912	4zlh.1.A	*Escherichia coli*	-0.83	75.37%
VC_0850	2hj1.1.B	*Haemophilus influenzae*	-1.04	21.79%

IV. CONCLUSIONS

The study showed that from eleven hypothetical proteins selected, five of the proteins involved in the bacterial pathogenicity, whilst other are essential for bacterial survival. All of the proteins are located either in the cytoplasmic or the plasma membrane of the cell. Five proteins, namely VC_0004, VC_0358, VC_0849, VC_1127 and VC_1912 were suggested to be suitable target proteins for experimental analyses.

The finding of this study can be useful for future works and experimental analysis. With all the computational data of the hypothetical proteins such as its physicochemical properties, predicted function and structural model, the role of each protein was identified. For long term purposes, it could help in modulating new target identification and drug discovery to control cholera, thus, reduce this lethal epidemic disease worldwide.

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] 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), pp. 53–9, 2015.

[2] F. R. Chowdhury, Z. Nur, N. Hassan, L. Seidlein, and S. Dunachie, "Pandemics, pathogenicity and changing molecular epidemiology of cholera in the era of global warming," *Annals of Clinical Microbiology and Antimicrobials*, vol. 16(10), pp. 1-6, 2017.

[3] A. K. Siddique, and R. Cash, "Cholera outbreaks in the classical biotype era," *Current Topics in Microbiology and Immunology*, vol. 379, p. 1–16, 2014.

[4] World Health Organization, "Cholera vaccines: WHO position paper - August 2017," *Weekly Epidemiological Record*, vol. 2017(92), pp. 477–500, 2017.

[5] L. M. Bilung, Y. S. Fuh, V. Linang, A. Benjamin, M. Vincent, K. Apun, S. Lihan, and C. S. Lin, "Genomic diversity of cholera outbreak strains in East Malaysia," *Malaysian Journal of Medicine and Health Sciences*, vol. 10(2), pp. 19–26, 2014.

[6] A. J. Silva and J. A. Benitez, "Vibrio cholerae Biofilms and Cholera Pathogenesis," *PLoS Neglected Tropical Diseases*, vol. 10(2), pp. 1-25, 2016.

[7] J. F. Heidelberg, J. A. Elsen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, L. Umayam, S. R. Gill, K. E. Nelson, T. D. Read, H. Tettelin, D. Richardson, M. D. Ermolaeva, J. Vamathevan, S. Bass, Q. Halving, I. Dragol, P. Sellers, L. McDonald, T. Utterback, R. D. Fleishmann, W. C. Nierman, O. White, S. L. Saizberg, H. O. Smith, R. R. Colwell, J. J. Mekalanos, C. J. Venter, and C. M. Fraser, "DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae," *Nature*,

vol. 406(6795), pp. 477–483, 2000.
[8] M. C. Chao, J. R. Pritchard, Y. J. Zhang, E. J. Rubin, J. Livny, B. M. Davis, and M. K. Waldor, "High-resolution definition of the Vibrio cholerae essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data," *Nucleic Acids Research*, vol. 41(19), pp. 9033–9048, 2013.
[9] M. A. Gazi, M. G. Kibria, M. Mahfuz, M. R. Islam, P. Ghosh, M. N. A. Afsar, M. A. Khan, and T. Ahmed, "Functional, structural and epitopic prediction of hypothetical proteins of Mycobacterium tuberculosis H37Rv: An in silico approach for prioritizing the targets," *Gene*, vol. 591(2), pp. 442–455, 2016.
[10] J. Ijaq, M. Chandrasekharan, R. Poddar, N. Bethi, and V. S. Sundararajan, "Annotation and curation of uncharacterized proteins-challenges," *Frontiers in Genetics*, vol. 6(119), pp. 1-7, 2015.
[11] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh, "UniProt: the Universal Protein knowledgebase.," *Nucleic Acids Research*, vol. 32(Database issue), pp. D115-D119, 2004.
[12] S. Saha, and G. P. S. Raghava, "VICMpred: An SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition," *Genomics, Proteomics Bioinformatics*, vol. 4(1), pp. 42–47, 2006.
[13] A. Gupta, R. Kapil, D. B. Dhakan, and V. K. Sharma, "MP3: A software tool for the prediction of pathogenic proteins in genomic and metagenomic data," *PLOS One*, vol. 9(4), pp. 1-11, 2014.
[14] G. M. Boratyn, C. Camacho, P. S. Cooper, G. Coulouris, A. Fong, N. Ma, T. L. Madden, W. T. Matten, S. D. McGinnis, Y. Merezhuk, Y. Raytselis, E. W. Sayers, T. Tao, J. Ye, and I. Zaretskaya, "BLAST: a more efficient report with usability improvements.," *Nucleic Acids Res.*, vol. 41(W1), pp. W29-W33, 2013.
[15] A. B. Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Severine Duvaud, Marc R. Wlkins, and Ron D. Appel, *Protein Identification and Analysis Tools on the ExPASy Server*, chap. The Proteomics Protocols Handbook. New Jersey, United States of America: Humana Press, vol. 112, 2005, pp. 571–615.
[16] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: Towards a more sustainable future," *Nucleic Acids Research*, vol. 44(D1), pp. D279–D285, 2016.
[17] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, L. Y. Geer, and S. H. Bryant, "CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures," *Nucleic Acids Research*, vol. 45(D1), pp. D200–D203, 2017.
[18] N. Y. Yu, J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. Cenk Sahinalp, M. Ester, L. J. Foster, and F. S. L. Brinkman, "PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics*, vol. 26(13), pp. 1608–1615, 2010.
[19] T. N. Petersen, S. Brunak, G. Von Heijne, and H. Nielsen, "SignalP 4.0: Discriminating signal peptides from transmembrane regions," *Nature Methods*, vol. 8(10), pp. 785–786, 2011.
[20] J. D. Bendtsen, L. Kiemer, A. Fausbøll and S. Brunak, "Non-classical protein secretion in bacteria," *BMC Microbiology*, vol. 5(58), 2005.
[21] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. Von Mering, "STRING v10: Protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43(D1), pp. D447–D452, 2015.
[22] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16(4), pp. 404–405, 2000.
[23] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER suite: Protein structure and function prediction," *Nature Methods*, vol. 12(1), pp. 7–8, 2014.
[24] C. C. Chen, J. K. Hwang, and J. M. Yang, "(PS)2: Protein structure prediction server," *Nucleic Acids Research*, vol. 34(Suppl._2), pp. W152–W157, 2006.
[25] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, "SWISS-MODEL: An automated protein homology-modeling server," *Nucleic Acids Research*, vol. 31(13), pp. 3381–3385, 2003.
[26] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, "Structure validation by Cα geometry: φ,ψ and Cβ deviation," *Proteins: Structure, Function, and Genetics*, vol. 50(3), pp. 437–450, 2003.
[27] D. Eisenberg, R. Lüthy, and J. U. Bowie, "VERIFY3D: Assessment of protein models with three-dimensional profiles," *Methods Enzymology*, vol. 277, pp. 396–406, 1997.
[28] P. Benkert, S. C. E. Tosatto, and D. Schomburg, "QMEAN: A comprehensive scoring function for model quality assessment," *Proteins: Structure, Function, and Bioinformatics*, vol. 71(1), pp. 261–277, 2008.
[29] M. Shahbaaz, M. I. Hassan, and F. Ahmad, "Functional annotation of conserved hypothetical proteins from Haemophilus influenzae Rd KW20," *PLOS One*, vol. 8(12), pp. 1-16, 2013.
[30] D. Gore, and A. Raut, "Computational function and structural annotations for hypothetical proteins of Bacillus anthracis," *Biofrontiers*, vol. 1(1), pp. 27-36, 2018.
[31] A. P Bidkar, "In-silico structural and functional analysis of hypothetical proteins of Leptospira Interrogans," *Biochemical Pharmacology*, vol. 03(3), 2014.
[32] W. R. Pearson, "An introduction to sequence similarity ('homology') searching," *Current Protocols in Bioinformatics*, sup. 42, pp. 3.1.1-3.1.8, 2013.
[33] R. E. Dalbey, and A. Kuhn, "YidC family members are involved in the membrane insertion, lateral integration, folding, and assembly of membrane proteins," *Journal of Cell Biology*, vol. 166(6), pp. 769–774, Sep. 2004.
[34] E. R. Green, and J. Mecsas, "Bacterial secretion systems: An overview.," *Microbiology Spectrum*, vol. 4(1), pp. 215–239, 2016.
[35] O. Niderman-Meyer, T. Zeidman, E. Shimoni, and Y. Kashi, "Mechanisms involved in governing adherence of Vibrio cholerae to granular starch," *Applied and Environmental Microbiology*, vol. 76(4), pp. 1034–1043, 2010.
[36] A. Kuznetsov, "Modularity and distribution of Sulfur metabolism genes in bacterial populations: Search and design," *Journal of Computer Science & Systems Biology*, vol. 3(5), pp. 91-106, 2010.
[37] F. Fan, and B. Kan, "Survival and proliferation of the lysogenic bacteriophage CTXΦ in Vibrio cholerae," *Virologica Sinica*, vol. 30(1), pp. 19–25, 2015.
[38] A. Lo Sciuto, R. Fernández-Piñar, L. Bertuccini, F. Iosi, F. Superti, and F. Imperi, "The periplasmic protein TolB as a potential drug target in Pseudomonas aeruginosa," *PLOS One*, vol. 9(8), 2014.
[39] S. Behrens-Kneip, "The role of SurA factor in outer membrane protein transport and virulence," *International Journal of Medical Microbiology*, vol. 300(7), pp. 421–428, 2010.
[40] F. Kiefer, K. Arnold, M. Künzli, L. Bordoli, and T. Schwede, "The SWISS-MODEL repository and associated resources," *Nucleic Acids Research*, vol. 37(Database Issue), pp. D387-392, 2009.