Aquatic Food Studies 1(2), AFS37
https://doi.org/10.4194/AFS37

RESEARCH PAPER

In Silico Functional Annotation of VP 128 Hypothetical Protein from *Vibrio parahaemolyticus*

Moslema Jahan Mou[1,#], Sk Injamamul Islam[2,3,*,#], Sarower Mahfuj[2,3]

[1]University of Rajshahi, Faculty of Life and Earth Science, Department of Genetic Engineering & Biotechnology, Rajshahi-6205, Bangladesh.
[2]Jashore University of Science and Technology, Faculty of Biological Science, Department of Fisheries and Marine Bioscience, Jashore-7408, Bangladesh.
[3]Chulalongkorn University, Faculty of Veterinary Science and Technology, Department of Veterinary Microbiology, Bangkok-10330, Thailand.
[#]These authors share the first authorship.

How to cite

Mou, M.J., Islam, S.I., Mahfuj, S. (2021). In Silico Functional Annotation of VP 128 Hypothetical Protein from *Vibrio parahaemolyticus*. *Aquatic Food Studies*, *1(2), AFS37*. https://doi.org/10.4194/AFS37

Article History
Received 10 August 2021
Accepted 20 October 2021
First Online 22 October 2021

Corresponding Author
Tel.: +0646904801
E-mail:
injamamulislam017@gmail.com

Keywords
Vibrio parahaemolyticus
Hypothetical protein
Characterization
Homology
CASTp

Abstract

Unknown or hypothetical proteins exist, but they have yet to be identified or correlated to gene sequences. Domains of unknown function are proteins that have been identified experimentally but do not have a known functional or structural domain. Using a variety of computational approaches and tools, this research investigated and characterized the likely functional characteristics of a hypothetical protein from *Vibrio parahaemolyticus* (Accession no. QOS18375.1). The physicochemical characteristics, subcellular localization, three-dimensional structure, protein-protein interactions, and functional elucidation of the protein are all available from this in silico perspective. Protein-protein interactions are investigated using the STRING software and resulted that VP128 putative protein interacts strongly with the GlpX type protein Fructose-1,6-bisphosphatase. The in-silico investigation documented the protein's hydrophilic nature with predominantly alpha (α) helices in its secondary structure. Furthermore, the protein, according to the research, features a VP128 domain and is thought to bind ribosomal subunits and the top active sites of the model described also. Therefore, the research findings will facilitate the development of new antibacterial drugs against acute gastroenteritis and other serious diseases by providing a better knowledge of the role of VP128 domain.

Introduction

Hypothetic proteins (HP), which are proteins predicted only from nucleic acid sequences and protein sequences with uncertain functions, make up a major component of viral proteomes (Lubec et al., 2005). Scientists have created a number of methods for predicting protein function using a variety of computational tools. This was accomplished through the use of sequence similarity, phylogenetic analysis, protein-protein interactions, protein—ligand interactions, active site residue similarity, conserved domains, motifs, phosphorylation sites, and gene expression patterns. However, the classical method of inferring function is based on sequence similarity using programs such as BLAST, FASTA and PSI-BLAST (Bharat Siva Varma et al., 2015; Pearson, 2013). HPs are projected proteins based on nucleic acid sequences that require experimental protein chemistry data. Furthermore, these proteins have a low level of similarity to known, annotated proteins (Lubec et al., 2005). Few HPs are conserved throughout phylogenetic lineages and are found in species. HP make up a significant portion of sequenced microbial genomes, but

they have yet to be functionally identified and described at the protein chemistry level (Galperin & Koonin, 2004; Reichart et al., 2020). There are two types of HPs. Uncharacterized protein families (UPFs) are one type, whereas domains of unknown function are another (DUFs). Experimental structures of unknown proteins have been discovered but not characterized or linked to a known gene. DUFs are proteins with no known functional or structural domains that have been identified experimentally. They could have coiled-coil structures or transmembrane areas that prevent function assignment. Analyzing the function of proteins with unknown functions has a number of advantages, including the ability to determine new conformational orientations of 3-dimensional structures, which allows for the evaluation of new domains and motifs, as well as the discovery of new protein mechanisms and cascades. Such new domains could be used as pharmaceutical targets in the future. Furthermore, phylogenetic profiling of proteins in various genomes can be used to predict function (Basu et al., 2011), and high spanning approaches, such as mass spectrometry-based protein complex identification and microarray gene expression profiles (Brown et al., 2000) and systematic synthetic lethal analysis (Goehring et al., 2003), are useful. The idea behind clustering gene-expression patterns is that genes with similar functions are more likely to be expressed together (Yuan et al., 2008). To determine function, scientists employed the neighbor-counting method (Hu et al., 2010). Based on the frequency of its neighbors with certain functions, they attributed a function to an unknown protein. Rather than looking for a simple agreement between the roles of the interacting members. Deng et al. employed a Bayesian technique to determine the likelihood of a hypothetical protein displaying the annotated function (Deng et al., 2002). Many protein domains have unclear activities; yet, they are involved in organisms' metabolic pathways and can have negative consequences. In some cases, mutations such as insertions, deletions, and substitutions can change the function of a protein.

V. parahaemolyticus is a common marine bacterium that can also cause disease in humans. This organism is commonly isolated from a wide range of raw fish, seafood and seafood products. The development of acute gastroenteritis is caused by the consumption of raw or undercooked seafood infected with *V. parahaemolyticus* (Ramamurthy & Nair, 2014). The study's main purpose is to identify and characterize a VP128 protein domain from *V. parahaemolyticus* with unknown function utilizing bioinformatics technologies.

Materials and Methods

Hypothetical Protein (HP) Selection

HPs were found using the phrase ‘‘hypothetical protein" in the NCBI protein database, and the hits were chosen at random to explore the close relatives using blast programs. In order to estimate the function of the hypothesized protein, a similarity search was performed using NCBI blast tools to find proteins that may have structural similarities with it (http://www.ncbi.nlm.nih.gov).

Physicochemical Characterization of the Hypothetical Proteins

The physicochemical properties of the HP in raw sequence format were assessed using the ProtParam tool from the ExPASy service (Gasteiger et al., 2005). The tool computes and provides the molecular weight, theoretical pI, amino acid composition, total number of positive and negative residues, extinction coefficient, instability index, aliphatic index, and grand average of hydropathicity (GRAVY), among other metrics. The extinction coefficient of a protein is a measure of how much light it absorbs at a given wavelength. The instability index is a measurement of a protein's stability in a test tube. An instability index of less than 40 indicates that the situation is stable, whereas a value more than 40 indicates that the situation is unstable. The relative volume filled by aliphatic side chain amino acids is defined as a protein's aliphatic index. The GRAVY value of a peptide or protein is computed by multiplying the total of all amino acid hydropathy values by the number of residues in the sequence (http://web.expasy.org/protparam/protparam-doc.html).

Sequence Analogy

Most basic step in the function prediction of a protein is looking for its structural homologs in different available genomics and proteomics based databases. Popular bioinformatics tool BLASTp was used for this purposes (Mahram & Herbordt, 2010).

Assessment of Secondary Structure

The secondary structural elements of the HP (Accession number QOS18375.1) were predicted through the SOPMA tool (Combet et al., 2000) using the default parameters (window width of 17, number of states of 4, and similarity threshold of 8).

Sub-cellular Localization

It is necessary to obtain information on a protein's subcellular localization in order to predict its cellular function. The outer membrane, inner membrane, periplasm, extracellular space, and cytoplasm all include proteins (Gazi et al., 2016). Virus-PLoc was used to predict viral protein subcellular localization (Shen & Chou, 2007), TMHMM was used to predict transmembrane information about the HPs (Krogh et al., 2001) and HMMTOP (Tusnády & Simon, 1998).

HHpred Model Generation

HHpred (Söding et al., 2006) searches a wide variety of databases, including PDB, SCOP, Pfam, SMART, COGs, and CDD whereas Sequence databases, such as the UniProt non-redundant databases, are examined using conventional sequence methods based. HHpred is also a fast server that uses hidden Markov model pairwise comparison profiles (HMMs) to discover and forecast remote protein homology and structure.

Refinement, and Validation of Three-Dimensional Structures

GalaxyWeb refined the 3D structure of the protein. The validity of the structure is a critical step in homology modeling, which is based on empirically validated 3D protein structures. For basic confirmation, the proposed protein model was submitted to ProSA-web (Wiederstein & Sippl, 2007). The z-score, which represents the overall character of the model, was predicted by the server. If the z-scores of the predicted model are outside the scale of the property for local proteins, the structure is incorrect (Wiederstein & Sippl, 2007). A Ramachandran plot analysis was performed utilizing the Ramachandran Plot Server to establish the overall quality of the protein (https://zlab.umassmed.edu/bu/rama/) (Zhou et al., 2011).

SOSUI Server

SOSUI is a free online tool that determines whether an amino acid sequence supplied is a membrane protein and distinguishes between insoluble and soluble proteins. Four physicochemical parameters are considered in this algorithm: (1) the Kyte and Doolittle hydropathy index, (2) an amphiphilicity index, (3) an amino acid charge index and (4) the length of each sequence. A list of outputs includes a graphic of the hydropathy plot, helical wheel diagrams for every membrane helices, and type of protein (Mitaku et al., 2002).

Model Quality Assessment

Finally, PROCHECK was used to evaluate the quality of the anticipated three-dimensional structure (Laskowski et al., 1993), Verify3D (http://nihserver.mbi.ucla.edu/Verify_3D/) (Eisenberg et al., 1997) and ERRAT Structure Evaluation server (Colovos & Yeates, 1993).

Phylogeny Analysis and Multiple Sequence Alignment

The BioEdit biological sequence alignment editor tool was used to perform multiple sequence alignments between the uncharacterized protein and proteins that shared structural similarities with the uncharacterized protein (Alzohairy, 2011). An older version of Molecular Evolutionary Genetic Study (MEGA) was used to do the phylogenetic analysis (https://megasoftware.net/).

Protein-Protein Interaction Analysis

The functions of proteins depend on interactions between their residues. In this study, we used the STRING database (http://string-db.org/), which uses physical and functional relationships to identify and identify known and anticipated protein interactions. This was based on Genomic Context, high-throughput studies, (Conserved) Co-expression, and Prior Knowledge. This database integrates interaction data from the following sources quantitatively (Franceschini et al., 2013).

Active Site Detection

The Computed Atlas of Surface Topography of Proteins (CASTp) was used to determine the active site of the protein (http://sts.bioengr. uic.edu/castp/). Researchers provided an online resource for locating, identifying and quantifying concave surface areas on three-dimensional protein structures (Dundas et al., 2006).

Results and Discussion

Protein Sequence Retrieval

The protein VP 128 of *V. parahaemolyticus*, which has an unknown function, was obtained from the National Center for Biotechnology Information (NCBI) under the accession number QOS18375.1. The Protein Data Bank does not have the 3D Structure (PDB). As a result, NCBI blastp analysis was performed on the 204 amino acid long protein QOS18375.1 discovered in *V. parahaemolyticus* to provide preliminary data on secondary and tertiary structures. Using several computational approaches and tools, the protein sequence of QOS18375.1 was chosen from the output to identify its potential functional characteristics. The Fasta format of the entry is given in Figure1.

Physicochemical Characterization

Extensive physicochemical characterization is often used to determine the nature of the surface locations responsible for the exceptional catalytic activity. The amino acid sequence of *V. parahaemolyticus* QOS18375.1 was downloaded in FASTA format and utilized as a query sequence for physicochemical parameter determination. QOS18375.1 has an instability index of 40.21 (40), indicating that the protein is unstable (Guruprasad et al., 1990). The protein has a molecular weight of 23328.58 KDa and is acidic (pI 6.21), indicating the presence of ribonuclease A residues based on high extinction coefficient values (10345M^{-1} cm^{-1}) (Gill & von Hippel,

1989). Using these parameters, you can determine where a given protein is located on a 2-D gel (Wilkins et al., 1999). The query protein's higher aliphatic index values (87.89) appear to be a key component in enhanced thermos stability over a wide temperature range. Because the protein is hydrophilic, it has a larger chance of interacting with water (Uddin et al., 2017) as indicated by the lower grand average of hydropathicity (GRAVY) indices value (-0.509) as shown in Table 1.

Domain Identification and Functional Annotation of QOS18375.1

A large number of putative viral proteins with unknown functions have been discovered in members of the *V. parahaemolyticus* family (http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi). There is only one domain found in the VP128 protein which is transcriptional regulator with an ArsR family helix-turn-helix (HTH) domain; Accession no. (COG2345). The domain attaches to DNA at the primary groove, and 1/2-strands form the wing, similar to other winged-HTHs (Saha et al., 2017). By using Expasy Scanprosite server (https://prosite.expasy.org/scanprosite/), no motifs were found for this HP (QOS18375.1) of *V. parahaemolyticus*.

Sequence Analogy Assessment

The sequence similarity of the VP128 domain sequence (QOS18375.1) compared to the Blastp protein structure database against non-redundant (setting default algorithm parameter) (Table 3) and model organisms (setting default algorithm parameter) yielded 5 hits (BlastP https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) (Table 2), indicating the possibility of finding structural features that are similar (a conserved hypothetical protein (VP128) from *Shewanella oneidensis*, *Deinococcus radiodurans*, *Streptomyces* (WP_202492436.1), *Streptomyces* (WP_003976893.1), *Mycobacterium tuberculosis* H37Rv). The percent identities, similarities and E-values are provided in both Table 2 and Table 3. VP128 has a low sequence similarity with few known function proteins, as seen in the Table 2.

Secondary Structure Inquiry

Protein structure and function are strongly connected. Secondary structure components such as helix, coil, sheet, and turn influence protein structure, function, and interaction. On analyzing the protein using various secondary structure prediction tools, the presence of alpha helix is obtained to be dominating in the structure, followed by random coil and extended strand. The presence of beta turn was shown by SOPMA. The analysis indicated the absence of various other secondary structures such as 3_{10} helix, Pi helix, Beta bridge, Bend region, ambigous states. The alpha helix (Hh), extended strand (Ee), beta turn (Tt), and random coil (Cc) of the protein (A0PB24) were predicted by the SOPMA tools to be 127 (62.25%), 20 (9.80%), 14 (6.86%),

Protein Identifier

>QOS18375.1 hypothetical protein VP128_00060 [Vibrio parahaemolyticus]

Amino Acid number: 204

Protein sequence

MKTVDRILQIVKRDGSVTAKQLSSELGMTTMGARQHLQGLEDEGILSIHDVKVKVGRPTRHWSLTQKGHE

QFADRHGELTIQFIEAVEHIFGKDGLDKVTSEREKLTLQNYRQHLDQCESLESKLETLVFLREKEGYMAE

LEQDEHGFILIENHCPICKAATRCPSLCKSELSVFQSLLGDDTTVERTEHIISGQRRCVYRIRA

Figure 1. The Fasta format of hypothetical protein.

Table 1. Physicochemical properties of the hypothetical protein.

Property	Value
Number of amino acids	204
Molecular weight	23328.58 KDa
Theoretical PI	6.21
Total number of negatively charged residues	32
Total number of positively charged residues	28
Ext. Coefficient	10345M^{-1} cm^{-1}
Instability index	40.21
Aliphatic index	87.89
Grand average of hydropathicity (GRAVY)	-0.509

Table 2. Results of a blastp search against NCBI database for VP128 sequences that are against Model organisms.

Protein ID	Protein	Organism	Identity	Similarity	Score	E-value
WP_011073404.1	transcriptional regulator	*Shewanella oneidensis*	100/202(50%)	131/202(64%)	206	2e-66
WP_051618792.1	transcriptional regulator	*Deinococcus radiodurans*	62/203(31%)	108/203(53%)	100	6e-25
WP_202492436.1	MULTISPECIES: transcriptional regulator	*Streptomyces*	54/208(26%)	90/208(43%)	63.5	5e-11
WP_003976893.1	MULTISPECIES: transcriptional regulator	*Streptomyces*	54/208(26%)	90/208(43%)	63.5	5e-11
NP_215976.2	Transcriptional regulator	*Mycobacterium tuberculosis* H37Rv	49/205(24%)	86/205(41%)	58.9	2e-09

Table 3. Output of the blastp search (NCBI database) for similar VP128 sequences against non-redundant (setting default algorithm parameter).

Protein ID	Protein	Organism	Identity	Similarity	Score	E-value
EDM56759	HTH domain family	*Vibrio parahaemolyticus* AQ3810	204/204(100%)	204/204(100%)	425	1e-147
AYF16526.1	HTH domain family	*Vibrio parahaemolyticus*	204/204(100%)	204/204(100%)	419	1e-147
EDM57450.1	HTH domain family	*Vibrio parahaemolyticus* AQ3810	204/204(100%)	204/204(100%)	419	2e-147
WP_005484031.1	MULTISPECIES: transcriptional regulator	*Vibrio*	204/204(100%)	204/204(100%)	417	2e-147
HAS7012953.1	Winged helix-turn-helix transcriptional regulator	*Vibrio parahaemolyticus*	203/204(99%)	204/204(100%)	417	3e-147
WP_140094402.1	transcriptional regulator	*Vibrio parahaemolyticus*	203/204(99%)	204/204(100%)	417	4e-147
EGR0301458.1	transcriptional regulator	*Vibrio parahaemolyticus*	203/204(99%)	204/204(100%)	416	6e-147
WP_069502403.1	transcriptional regulator	*Vibrio parahaemolyticus*	203/204(99%)	204/204(100%)	416	7e-147
MBE3927463.1	transcriptional regulator	*Vibrio parahaemolyticus*	203/204(99%)	204/204(100%)	416	7e-147
WP_005489474.1	transcriptional regulator	*Vibrio parahaemolyticus*	203/204(99%)	204/204(100%)	416	7e-147

and 43 (21.08%), accordingly (Figure 2). The sequence plot from the secondary structure of the HP (Figure 2) represents that most of the protein is extracellular, whereas subcellular localization reports the protein as cytoplasmic. Further studies are required to unleash the nature of the protein.

Subcellular Localization

The subcellular localization of proteins has been used to describe them as therapeutic and vaccination targets (Acharya & Garg, 2016). Proteins found in the cytoplasmic matrix might be used as medication targets, while proteins found in the inner and outer membranes could be used as vaccination targets (Prabhu et al., 2020). The understanding of localization is a crucial step in determining the function of proteins. The HP was anticipated to be present in any of the cell's places (cytoplasm, periplasm, extracellular, inner membrane, and outer membrane) based on the knowledge of trained data sets. Based on the concurrence hits, we estimated the location of the VP128 HP subject for the study. The protein's predicted outcome was discovered in the cytoplasm. Because it is a cytoplasmic protein, the TMHMM server returned no results. The ability to predict membrane proteins is critical for a better understanding of drug targets and the development of effective therapeutic compounds.

Protein Binding Sites and Gene Ontology (GO) Prediction

Nine different protein binding sites were identified at positions 4-15; 21-38; 58-62; 73-78; 80-82; 91;92; 132-135;150-155; 324 (Figure 3).

The functional aspects of biological process ontology (Table 4), molecular functional ontology (Table 5) and cellular component ontology were predicted and categorized using gene ontology (Table 6). Molecular functional ontology (Table 5) calculated as DNA binding (37%), Biological process ontology (Table 4) detected as regulation of transcription, DNA-templated (37%); transcription initiation from RNA polymerase II promoter (37%). Cellular component ontology (Table 6) predicted as cytoplasm (37%); cell part (37%); cell (37%); intracellular (37%); intracellular part (37%); intracellular organelle (17%); membrane-bounded organelle (15%) and mitochondrion (15%).

Homology Identification

HHpred was employed to create a distant homology model. According to the HHPred output (Figure 4), novel sequence similarities with the VP128 protein were discovered, such as the virus protein's 204aa protein domain, RPS5, MRGBP,1-204aa human membrane protein domain, fusion associated small

```
SOPMA:
Alpha helix      (Hh) :   127 is  62.25%
3₁₀ helix        (Gg) :     0 is   0.00%
Pi helix         (Ii) :     0 is   0.00%
Beta bridge      (Bb) :     0 is   0.00%
Extended strand  (Ee) :    20 is   9.80%
Beta turn        (Tt) :    14 is   6.86%
Bend region      (Ss) :     0 is   0.00%
Random coil      (Cc) :    43 is  21.08%
Ambiguous states (?) :     0 is   0.00%
Other states         :     0 is   0.00%
```

Figure 2. SOPMA score for secondary structure elements

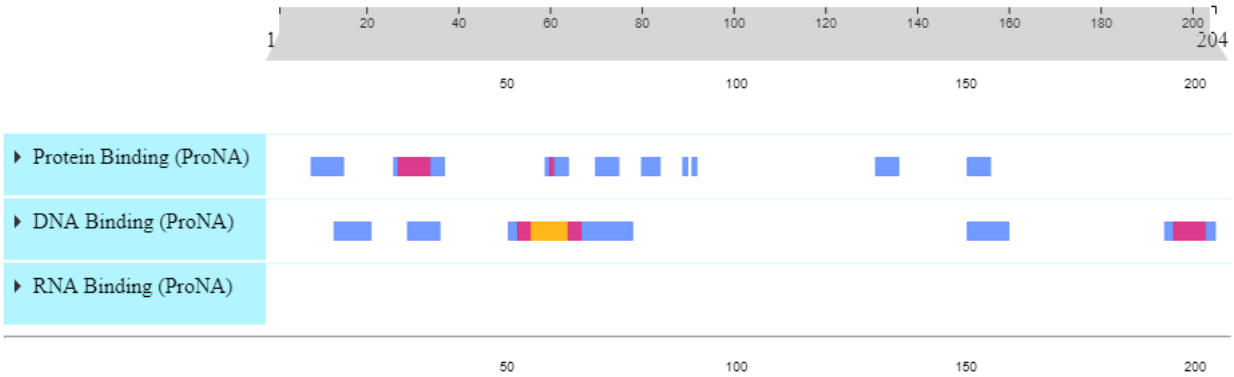

Figure 3. Protein-Protein and polynucleotide binding sites

Table 4. Biological process ontology.

GO ID	GO Term	Reliability (%)
GO: 0006355	regulation of transcription, DNA-templated	37%
GO: 0006367	transcription initiation from RNA polymerase II promoter	37%

Table 5. Molecular functional ontology.

GO ID	GO Term	Reliability (%)
GO: 0003677	DNA binding	37%

Table 6. Cellular component ontology.

GO ID	GO Term	Reliability (%)
GO: 0005737	Cytoplasm	37%
GO: 0005739	Cell part	37%
GO: 0005740	Cell	37%
GO: 0005732	Intracellular	37%
GO: 0005733	Intracellular part	37%
GO: 0005730	Intracellular organelle	15%
GO: 0005735	Membrane-bounded organelle	15%
GO: 0005741	Mitochondrion	15%

transmembrane (FAST) protein (THOC7), and plant subtilisin-like protease (26s). New sequence similarities were discovered with the VP128 protein, such as with regulatory proteins, YSLB protein, HTH type transcriptional proteins, DNA binding protein, gene activator AphA, and PadR-family transcriptional protein, based on the HHPred output (Figure 4).

At the Modeller server, the homology derived model was produced from HHPred using 3G3Z as the template. The target sequence of QOS18375.1 was entered into the HHpred Template Selection tool in FASTA format, and the most active template was chosen as the first of 250 hits with a probability rate of 100%, and the tertiary modeled protein structure was saved in PDB format predicted by Modeller (Figure 5).

Model Validation

The Galaxy Refine server was used to refine the protein's projected tertiary structure, resulting in five refined models with more amino acid residues in the preferred location**.** The scores shown above indicate the enhanced model's caliber as compared to the other variants. In Discovery Studio, refine model 1 was chosen and visualized (Figure 6).

The before and after improved VP128 protein models were validated using the Ramachandran Plot Server and the ProSA-Web web server. According to the Ramachandran plot server, 98.864% of the structure before refinement was in the favorable zone. The rampage server generated a better result after

```
Query         QOS18375.1 hypothetical protein VP128_00060 [Vibrio parahaemolyticus]
Match_columns 204
No_of_seqs    219 out of 2463
Neff          11.5078
Searched_HMMs 61623
```

No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	3G3Z_A Transcriptional regulat	99.5	1.5E-12	2.5E-17	85.4	13.3	96	3-101	32-127 (145)
2	5FRY_A POSITIVE PHENOL-DEGRADA	99.4	8.5E-11	1.4E-15	82.1	13.9	126	79-204	46-191 (211)
3	3AJ1_E Cellulose synthase oper	99.3	1.4E-10	2.2E-15	77.3	13.8	108	70-179	15-127 (167)
4	3NJC_B YSLB protein; NESG, PSI	99.3	7.1E-11	1.2E-15	78.4	10.6	115	79-204	31-152 (158)
5	3BDD_A Regulatory protein MarR	99.3	2.9E-10	4.7E-15	74.3	13.2	102	2-105	31-132 (142)
6	3BDD_B Regulatory protein MarR	99.3	2.9E-10	4.7E-15	74.3	13.2	102	2-105	31-132 (142)
7	2OSO_A Hypothetical protein MJ	99.3	1.4E-10	2.3E-15	77.6	12.0	116	79-204	32-157 (163)
8	6L79_A Phenol regulator MopR;	99.3	2.4E-10	3.9E-15	80.9	13.4	128	77-204	48-195 (229)
9	5HLI_A MarR family transcripti	99.3	9.1E-11	1.5E-15	77.4	10.4	105	2-109	41-145 (149)
10	1Z91_A Organic hydroperoxide r	99.3	1.3E-10	2.1E-15	76.4	10.7	99	2-105	40-138 (147)
11	5AIP_A TRANSCRIPTIONAL REGULAT	99.3	6.3E-10	1E-14	72.9	13.6	102	1-105	35-136 (146)
12	3ZPL_F PUTATIVE MARR-FAMILY TR	99.3	3.4E-10	5.5E-15	76.9	12.6	102	2-105	56-159 (177)
13	2FXA_C Protease production reg	99.3	8E-10	1.3E-14	77.0	14.2	128	2-134	48-175 (207)
14	2FBH_A transcriptional regulat	99.2	5.5E-10	8.9E-15	73.3	12.4	102	1-105	36-138 (146)
15	3BJA_A Transcriptional regulat	99.2	7.1E-10	1.1E-14	72.1	12.5	101	2-105	33-133 (139)
16	3ZMD_A PUTATIVE TRANSCRIPTIONA	99.2	8.5E-10	1.4E-14	75.0	13.0	102	2-105	63-166 (178)
17	3ZMD_D PUTATIVE TRANSCRIPTIONA	99.2	8.5E-10	1.4E-14	75.0	13.0	102	2-105	63-166 (178)
18	2A61_B transcriptional regulat	99.2	1.4E-09	2.2E-14	71.3	13.3	101	2-105	33-133 (145)
19	1YYV_B putative transcriptiona	99.2	7E-10	1.1E-14	71.6	11.6	93	2-98	35-128 (131)
20	6PCO_A MarR-family transcripti	99.2	1.2E-09	1.9E-14	75.5	12.8	101	2-105	81-181 (195)
21	5HS7_A HTH-type transcriptiona	99.2	2.1E-10	3.4E-15	71.6	8.2	93	2-98	17-110 (110)
22	3U1D_A uncharacterized protein	99.2	9.8E-10	1.6E-14	72.7	11.3	99	2-103	29-136 (151)
23	2HZT_B Putative HTH-type trans	99.2	1.1E-09	1.8E-14	68.0	10.8	88	1-92	13-101 (107)
24	2HZT_C Putative HTH-type trans	99.2	1.1E-09	1.8E-14	68.0	10.8	88	1-92	13-101 (107)
25	5DD8_B Transcriptional regulat	99.2	1.8E-09	3E-14	73.6	12.8	101	2-105	69-172 (181)
26	1BJA_B TRANSCRIPTION REGULATOR	99.2	6.2E-10	1E-14	67.6	9.3	73	1-83	15-88 (95)
27	7BZE_A HTH-type transcriptiona	99.2	1E-09	1.7E-14	69.9	10.7	87	2-90	26-115 (123)
28	5E1W_A Transcriptional regulat	99.2	2E-09	3.3E-14	73.2	12.7	101	2-105	57-158 (180)

Figure 4. HHpred output displaying probable similarities with QOS18375.1

Figure 5. Predicted three-dimensional structure of the hypothetical protein

Figure 6. Validation of model

refinement, with 98.96% of residues in highly preferred regions Table 7. The validation quality and potential faults in a basic tertiary structure model are assessed using the ProSA-web server. Validation of the final VP128 protein model reveals a Z-score of -5.39 which indicate the model is significant **(**Table 7).

Model Quality Assessment

VERIFY3D was used to evaluate the model's quality. The tertiary structure of the projected two-chain protein model (3G3Z) passed the Verify 3D structure assessment experiment with ease, and we received other results such as PROCHECK and ERRAT. The VP128 domains are found in many eubacteria, archaebacteria, and pathogens, and they play an important role in peptidoglycan production, therefore researchers might be interested in developing new drugs based on them.

From 110 to 170 amino acids, good quality was reported in VERIFY3D. (Figure7). The overall quality factor in ERRAT was 95.604. Two lines are shown on the error axis to illustrate the confidence with which regions that exceed that error value can be rejected. When expressed as a percentage of the protein for which the estimated error value falls below the 95% rejection criterion, strong high-resolution structures produce values of around 95% or higher, whereas lesser resolutions (2.5 to 3A) produce values of around 91% (Figure 8).

Sequence Alignment and Phylogeny Analysis

The FASTA sequences of the uncharacterized protein (QOS18375.1) and homologous annotated proteins were aligned using multiple sequence alignment (Figure 9). In order to evaluate the evolutionary relationship between the query protein and its homologs, we compiled a phylogenetic tree based on amino acid sequences (Figure 10). Phylogenetic analysis was also used to establish the homology assessment between the proteins, down to the complex and subunit level. Because they were discovered in the same class as the query protein, the phylogram allowed us infer that the closely related homologs with the query protein had comparable functions. Figure 10 shows a phylogenetic tree based on alignment and BLAST results that give a similar perspective about the protein**.**

Protein-Protein Interaction

The STRING protein-protein interaction network indicated that our putative protein interacts strongly with the GlpX type protein Fructose-1,6-bisphosphatase; this connection suggests that the protein may serve as a DNA/RNA hydrase.

Active Site Detection

The modeled protein's CASTp v.3.0 algorithm predicted 17 distinct active sites (Figure 12). CASTp is a database server that can detect areas on proteins, define their boundary, compute the areas' area, and find the areas' dimensions. Pockets on protein surfaces and vacuums hidden within proteins are involved. Surfaces of solvent accessible molecules (surface of Richards) and molecular surfaces (surface of Connolly) are used to define a pocket and volume spectrum or vacuum. CASTp might be used to investigate protein operational zones and surface characteristics. CASTp provides a graphical, user-interface flexible, dynamic display as well as on-the-fly measurement of user-submitted constructions (Tian et al., 2018). The top active sites of the modeled protein were identified between the area of 238.338 and the volume of 197.309 (Figure 12). Figure 12 shows the protein's anticipated active site together with its amino acid residues.

Conclusion

The VP128 domain has a crucial role as an antiviral target and binds to ribosomal subunits, according to the research. VP128 was also discovered to be a soluble protein with one exposed domain. The predicted protein's output is long, indicating that the studied hypothetical protein is soluble. The existence and distribution of VP128 domains across a wide range of

Table 7. Ramachandran Plot Server and ProSA-Web online server results

Parameters		Initial Model	Refine Model	Remarks
Ramachandran	Highly Preferred	98.864%	98.96%	Significant
	Preferred	0.00%	0.00%	Significant
	Questionable	1.136%	1.04%	Significant
ProSA Web	Z-Score	-5.02	-5.39	Significant

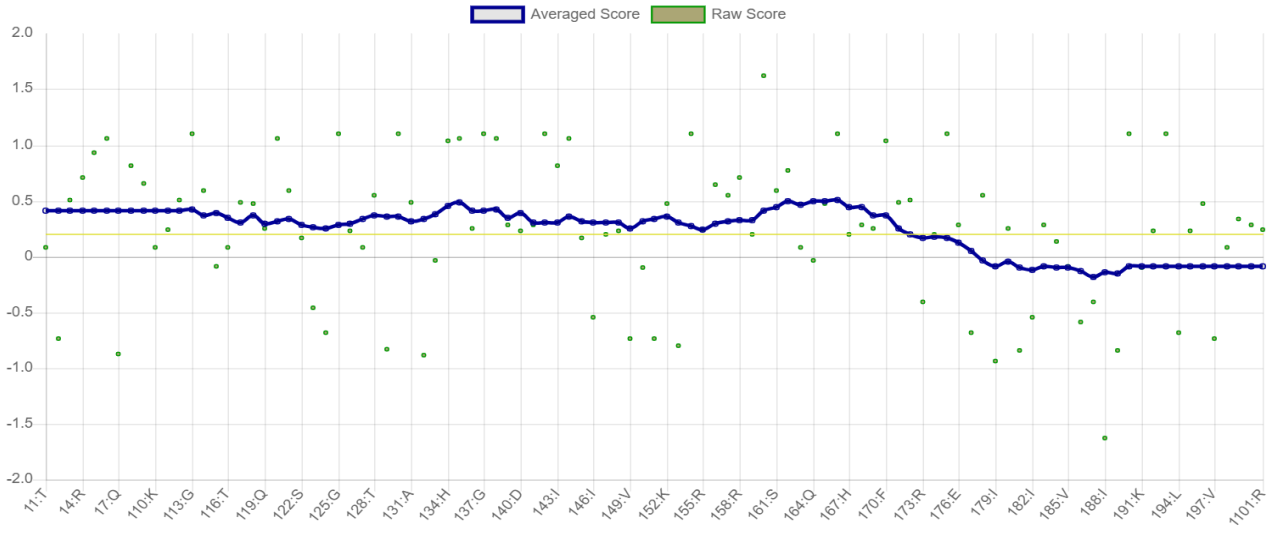

Figure 7. Model quality obtained by VERIFY3D.

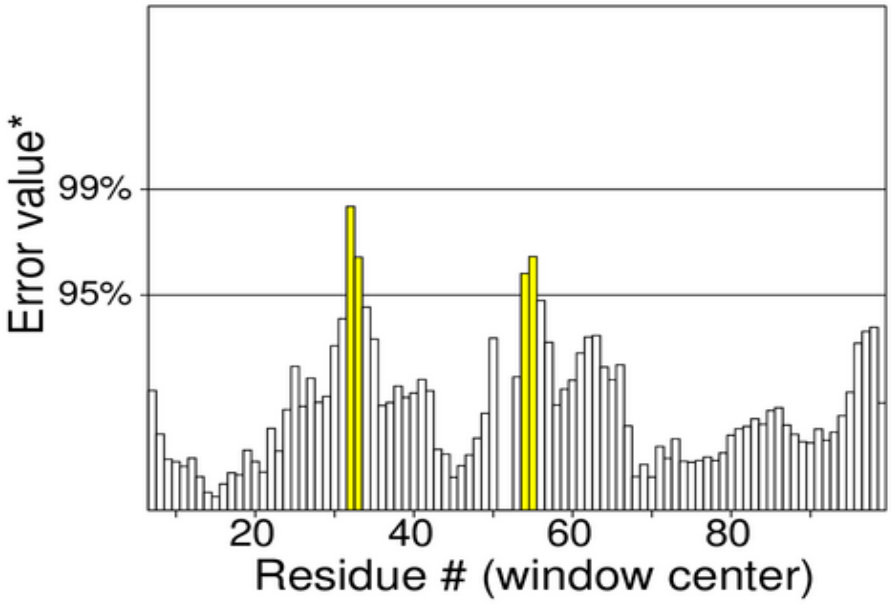

Figure 8. Model quality obtained by ERRAT.

Table 8. Quality of the protein model (3G3Z) assessed by ANOLEA (Melo & Feytmans, 1998)

Features	Chain A	Chain B
Amino acids with high energy	3-4; 13-15; 22; 26; 65; 69-70; 77; 79; 82-83; 113-114; 123-124; 127; 133; 141-142	3-4; 11; 13-14; 18; 22; 65; 70; 77; 79; 82-83; 85; 93; 96; 113-114; 118; 124; 127; 133; 141-142
Total amino acids with high energy	22	24
Total number of atoms	1111	1111
Total number of non-local atomic interactions	6628	6582
Total non-local energy of the protein (E/kT units)	-447	-423
Non-local normalized energy Z-score	-0.74	-0.49

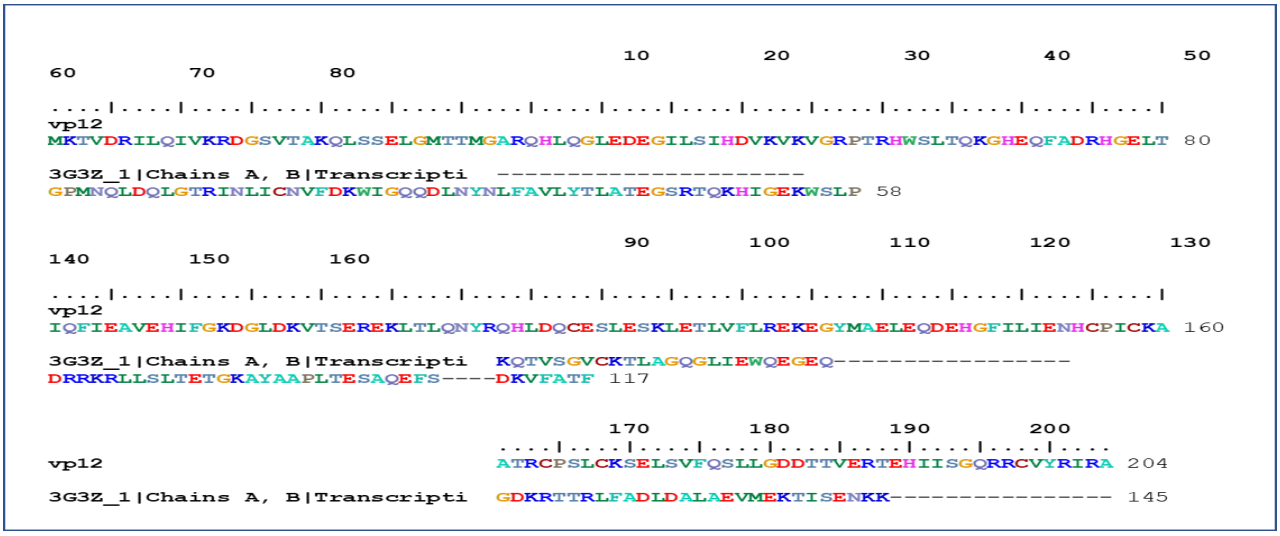

Figure 9. Sequence alignment with homolog protein sequence.

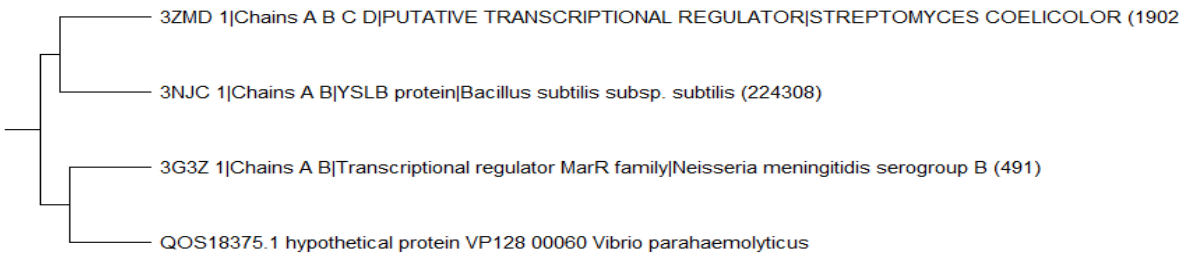

Figure 10. Multiple sequence phylogenetic tree.

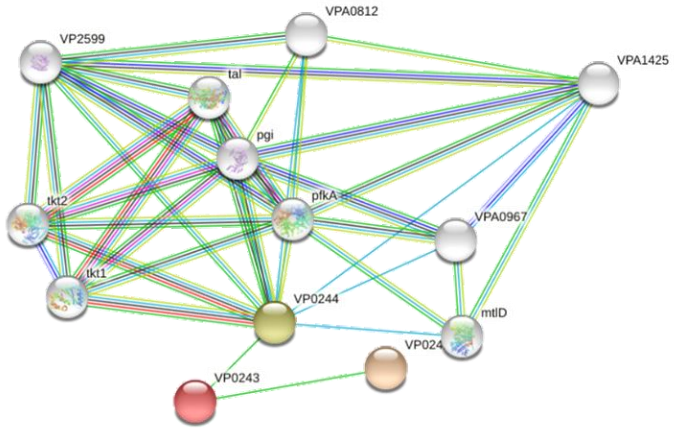

Figure 11. String network analysis of the hypothetical protein, indicates as VP0244.

Figure 12. Active location of the VP 128 hypothetical protein (A) The blue sphere represents the protein's active site. (B) The active site of amino acid residues (Blue color)

bacteria and diseases suggests that new antibacterial drugs could be developed. More research is being done, such as protein-ligand docking studies, to identify the representative amino acids involved for ligand binding. The 3G3Z domain is found in a wide range of pathogens and diseases, and it plays an important function in DNA transcription, therefore it may be of interest to researchers looking to produce new drugs against acute gastroenteritis.

Ethical Statement

This article does not contain any studies involving animals performed by any of the authors.

Funding Information

There is no any funding institution for this study.

Author Contribution

Conceptualization, M.M. and SK; methodology, SK. and M.M.; software, SK.; validation, M.M., and SK.; formal analysis, SK. and S.M.; investigation, SK.; resources, SK.; data curation, M.M. and SK.; writing—original draft preparation, M.M.; writing—review and editing, SK. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The author(s) declare that they have no known competing financial or non-financial, professional, or personal conflicts that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author sincerely grateful to the ASEAN and Non-ASEAN scholarship authority at Chulalongkorn University, Thailand as giving financial support for pursuing masters studies and the third author also highly grateful to the Chulalongkorn University, Thailand and National Science and Technology Development Agency (NSTDA) for providing financial assistance for PhD study.

References

Acharya, A., & Garg, L. C. (2016). Drug Target Identification and Prioritization for Treatment of Ovine Foot Rot: An In Silico Approach. *International journal of genomics*, *2016*, 7361361-7361361. https://doi.org/10.1155/2016/7361361

Alzohairy, A. (2011). BioEdit: An important software for molecular biology. *GERF Bulletin of Biosciences*, *2*, 60-61.

Basu, M. K., Selengut, J. D., & Haft, D. H. (2011). ProPhylo: partial phylogenetic profiling to guide protein family construction and assignment of biological process. *BMC bioinformatics*, *12*, 434-434. https://doi.org/10.1186/1471-2105-12-434

Bharat Siva Varma, P., Adimulam, Y. B., & Kodukula, S. (2015). In silico functional annotation of a hypothetical protein from Staphylococcus aureus. *Journal of Infection and Public Health*, *8*(6), 526-532. https://doi.org/https://doi.org/10.1016/j.jiph.2015.03.007

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, *97*(1), 262-267. https://doi.org/10.1073/pnas.97.1.262

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci*, *2*(9), 1511-1519. https://doi.org/10.1002/pro.5560020916

Combet, C., Blanchet, C., Geourjon, C., & Deléage, G. (2000). NPS@: Network Protein Sequence Analysis. *Trends in Biochemical Sciences*, *25*(3), 147-150. https://doi.org/10.1016/S0968-0004(99)01540-6

Deng, M., Zhang, K., Mehta, S., Chen, T., & Sun, F. (2002). Prediction of protein function using protein-protein interaction data. *Proc IEEE Comput Soc Bioinform Conf*, *1*, 197-206.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res*, *34*(Web Server issue), W116-118. https://doi.org/10.1093/nar/gkl282

Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol*, *277*, 396-404. https://doi.org/10.1016/s0076-6879(97)77022-8

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., & Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, *41*(Database issue), D808-815. https://doi.org/10.1093/nar/gks1094

Galperin, M. Y., & Koonin, E. V. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res*, *32*(18), 5452-5463. https://doi.org/10.1093/nar/gkh885

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, 571-607.

Gazi, M. A., Kibria, M. G., Mahfuz, M., Islam, M. R., Ghosh, P., Afsar, M. N., Khan, M. A., & Ahmed, T. (2016). Functional, structural and epitopic prediction of hypothetical proteins of Mycobacterium tuberculosis H37Rv: An in silico approach for prioritizing the targets. *Gene*, *591*(2), 442-455. https://doi.org/10.1016/j.gene.2016.06.057

Gill, S. C., & von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Analytical Biochemistry*, *182*(2), 319-326. https://doi.org/https://doi.org/10.1016/0003-2697(89)90602-7

Goehring, A. S., Mitchell, D. A., Tong, A. H., Keniry, M. E., Boone, C., & Sprague, G. F., Jr. (2003). Synthetic lethal analysis implicates Ste20p, a p21-activated potein kinase, in polarisome activation. *Mol Biol Cell*, *14*(4), 1501-1516. https://doi.org/10.1091/mbc.e02-06-0348

Guruprasad, K., Reddy, B. V., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng*, *4*(2), 155-161. https://doi.org/10.1093/protein/4.2.155

Hu, P., Jiang, H., & Emili, A. (2010). Predicting protein functions by relaxation labelling protein interaction network. *BMC bioinformatics*, *11 Suppl 1*(Suppl 1), S64-S64. https://doi.org/10.1186/1471-2105-11-S1-S64

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, *305*(3), 567-580. https://doi.org/10.1006/jmbi.2000.4315

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283-291. https://doi.org/https://doi.org/10.1107/S0021889892009944

Lubec, G., Afjehi-Sadat, L., Yang, J.-W., & John, J. P. P. (2005). Searching for hypothetical proteins: Theory and practice based upon original data and literature. *Progress in Neurobiology*, *77*(1), 90-127. https://doi.org/https://doi.org/10.1016/j.pneurobio.2005.10.001

Mahram, A., & Herbordt, M. (2010). *Fast and accurate NCBI BLASTP: acceleration with multiphase FPGA-based prefiltering*. https://doi.org/10.1145/1810085.1810099

Melo, F., & Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy11Edited by J. Thornton. *Journal of Molecular Biology*, *277*(5), 1141-1152.https://doi.org/https://doi.org/10.1006/jmbi.1998.1665

Mitaku, S., Hirokawa, T., & Tsuji, T. (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics*, *18*(4), 608-616. https://doi.org/10.1093/bioinformatics/18.4.608

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, *Chapter 3*, Unit3.1-Unit3.1. https://doi.org/10.1002/0471250953.bi0301s42

Prabhu, D., Rajamanikandan, S., Anusha, S. B., Chowdary, M. S., Veerapandiyan, M., & Jeyakanthan, J. (2020). In silico Functional Annotation and Characterization of Hypothetical Proteins from Serratia marcescens FGI94. *Biology Bulletin*, *47*(4), 319-331. https://doi.org/10.1134/S1062359020300019

Ramamurthy, T., & Nair, G. B. (2014). Bacteria: Vibrio parahaemolyticus. In Y. Motarjemi (Ed.), *Encyclopedia of Food Safety* (pp. 555-563). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-378612-8.00118-9

Reichart, N. J., Jay, Z. J., Krukenberg, V., Parker, A. E., Spietz, R. L., & Hatzenpichler, R. (2020). Activity-based cell sorting reveals responses of uncultured archaea and bacteria to substrate amendment. *The ISME Journal*, *14*(11), 2851-2861. https://doi.org/10.1038/s41396-020-00749-1

Saha, R. P., Samanta, S., Patra, S., Sarkar, D., Saha, A., & Singh, M. K. (2017). Metal homeostasis in bacteria: the role of ArsR–SmtB family of transcriptional repressors in combating varying metal concentrations in the environment. *BioMetals*, *30*(4), 459-503. https://doi.org/10.1007/s10534-017-0020-3

Shen, H. B., & Chou, K. C. (2007). Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers*, *85*(3), 233-240. https://doi.org/10.1002/bip.20640

Söding, J., Remmert, M., Biegert, A., & Lupas, A. N. (2006). HHsenser: exhaustive transitive profile search using HMM-HMM comparison. *Nucleic Acids Res*, *34*(Web Server issue), W374-378. https://doi.org/10.1093/nar/gkl195

Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Research*, *46*(W1), W363-W367. https://doi.org/10.1093/nar/gky473

Tusnády, G. E., & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol*, *283*(2), 489-506. https://doi.org/10.1006/jmbi.1998.2107

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids research*, *35*(suppl_2), W407-W410. https://doi.org/10.1093/nar/gkm290

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., & Hochstrasser, D. F. (1999). Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol*, *112*, 531-552. https://doi.org/10.1385/1-59259-584-7:531

Yuan, Y., Li, C.-T., & Wilson, R. (2008). Partial mixture model for tight clustering of gene expression time-course. *BMC bioinformatics*, *9*, 287-287. https://doi.org/10.1186/1471-2105-9-287

Zhou, A., O'Hern, C., & Regan, L. (2011). Revisiting the Ramachandran plot from a new angle. *Protein science: a publication of the Protein Society*, *20*, 1166-1171. https://doi.org/10.1002/pro.644