Genetics of Aquatic Organisms 7(1), GA550
https://doi.org/10.4194/GA550

PROOF
RESEARCH PAPER

Synthesis of Homo-oligomer Models, Functional Characterization, and Annotation of the Virulent HP33 Protein from the *Vibrio harveyi* Strain of Scale Drop and Muscle Necrosis Disease

Sk Injamamul Islam[1,*], Saloa Sanjida[2], Md. Akib Ferdous[1], Nasim Habib[1]

[1]Jashore University of Science and Technology, Faculty of Biological Science and Technology, Department of Fisheries and Marine Bioscience, Jashore-7408, Bangladesh.
[2]Jashore University of Science and Technology, Faculty of Applied Science and Technology, Department of Environmental Science and Technology, Jashore-7408, Bangladesh.

How to cite

Islam, S.I., Sanjida, S., Ferdous, M.A., Habib, N. (2023). Synthesis of Homo-oligomer Models, Functional Characterization, and Annotation of the Virulent HP33 Protein from the *Vibrio harveyi* Strain of Scale Drop and Muscle Necrosis Disease. *Genetics of Aquatic Organisms, 7(1), GA550*. https://doi.org/10.4194/GA550

Article History
Received 11 September 2022
Accepted 10 January 2023
First Online 24 January 2023

Corresponding Author
Tel.: +6012.2484901
E-mail: 6378506331@student.chula.ac.th

Keywords
Protein
Asian sea bass
V. harveyi
Characterization

Abstract

This study used a comprehensive bioinformatic application to discover the functions of the HP33 protein, which is responsible for the scale drop and muscle necrosis disease (SDMND) in fish. The main objective of the study was to the characterization of the HP33 protein and predict the homo-oligomer models to understand the physical effect of the protein for further research. At first, multiple sequence alignment and sub-cellular localization of the HP33 were predicted by the *in-silico* approach. The result suggests that this putative protein clustered with another hypothetical protein, *Vibrio harveyi,* and is an unstable, nonpolar, and outer membrane protein. Functional analysis of the protein by Pfam, InterProScan, and SMART tools predicts that the HP has a single functional domain that may signify a cluster of biosynthetic genes. The prediction of the active site, as well as the protein-protein interaction, were also predicted in this study. Furthermore, a protein-ligand docking investigation revealed two potential therapeutic compounds (Ferroheme C, Valine) that can be effective against HP33 pathogenesis. In conclusion, the homo-oligomers model's predictions and the ab-initio docking findings will offer important information for an additional immunological investigation, which may be beneficial in a future study on SDMND prophylaxis.

Introduction

Marine aquaculture in southeast Asia relies heavily on Asian sea bass, also known as barramundi (*Lates calcarifer*) (Khang, Phuong, Dat, Knibb, & Nguyen, 2018). The annual amount of Asian sea bass harvested from farms increased from 11,000 to 76,000 tonnes between 1990 and 2015 (Fishsite, 2022). Despite this, Asian sea bass farms have suffered significant financial losses due to disease epidemics. Different bacteria, viruses, and parasites can cause infection in Asian sea bass aquaculture; however, some bacterial diseases can cause 100% mortality at the farm level (Crane & Hyatt, 2011). One of the severe bacterial diseases is scale drop and muscle necrosis (SDMN) which was first discovered in 2017 in Vietnamese sea bass farms (Dong et al., 2017). Infected fish exhibited scale loss, fin rot, and severe muscle necrosis, resulting in a relatively high death rate (40-50%) in infected cages. SDMN-Y6 (often called Y6) was a *Vibrio harveyi* strain frequently found on sick fish samples. Histology confirmed bacterial toxin(s) involvement with dilated renal tubules and sloughing epithelial cells (Dong et al., 2017); similar *Vibrio* species induce hepatopancreatic shrimp necrosis (Sirikharin et al., 2015). Multiple virulent genes and a toxin-carrying prophage were found in *V. harveyi* Y6 linked to SDMN

(Kayansamruaj et al., 2019). Recently, SDMND was observed in farmed Asian sea bass in Thailand, where *V. harveyi* is dominant, and the HP33 protein was identified as the virulence factor (Kwankijudomkul et al., 2021). According to the study, the HP33 protein is a promising vaccine target since it facilitates viral entry into host cells (Kwankijudomkul et al., 2021). HP33 or its encoding gene might also be used as a biomarker in Asian sea bass for early detection of SDMND-related *V. harveyi*. BLAST tests of HP33 against the GenBank database revealed that this gene was present in just seven of the 48 known *V. harveyi* genomes. This suggests that the variety *V. harveyi* with HP33 was recently identified (Kwankijudomkul et al., 2021). However, the purpose of HP33 has yet to be apparent. Regardless, many of the proteins generated by these bacteria are classified as HPs since their structures and biological activities are unknown. These proteins may be helpful, and their annotation can provide new details regarding their structures, routes, and activities. Therefore, bioinformatics techniques may predict and investigate various forms of HP structure, biological activity, and protein interactions. Additionally, HP structural and functional annotation may uncover novel biomarkers and pharmacological targets (Lubec, Afjehi-Sadat, Yang, & John, 2005). Several bioinformatics databases and methods have been used to successfully annotate the activities of putative proteins in various pathogenic bacteria (Turab Naqvi et al., 2017).

When to connecting genomic and proteomic information, HPs are very important (Ijaq, Chandra, Ray, & Jagannadham, 2022). Large volumes of genomic and transcriptomic data have been collected and stored in online databases during the last several years. Computational analyses of gene structures and nucleic acid sequences have assumed that a large amount of this genomic data encodes a protein. Still, there is no confirmation of its *in vivo* expression (Y. M. Wang et al., 2021). The term "hypothetical protein" (HPs) is often used to describe these molecules. One study found that HPs made up as much as 50% of the proteome in certain species (Minion et al., 2004). Characterizing the structure and function of these unknown functional HPs is a challenge for functional genomics and biology more broadly. Genomic sequence data, structural and functional genomics, and proteomics might benefit from a long-term solution to this issue. In addition to aiding in the creation of potential antibacterial therapies against pathogens and investigating drug resistance, disease, and other biological processes, HP's functional annotation is crucial for understanding a wide variety of disorders (Naveed et al., 2016).

Attributing a function to an HP using a variety of bioinformatics approaches has become simpler as the *in-silico* inquiry has progressed. Therefore, this study aimed to better understand the *V. harveyi* hypothetical protein HP33 by providing its structural and biological role. Furthermore, the study focused on the functional characterization and findings of the best homo-oligomer model for the HP33 to understand the physiological activities of proteins, such as metabolism, signaling, and immune response. Predictions were made based on research on proteins' subcellular localization, secondary structure, and active site. In addition, ab initio homology modeling was used to build a high-quality model of the HP33 for therapeutic targets.

Methods

Finding Similarities and Retrieving Sequences

The sequence of amino acids of putative protein HP33 from the *Vibrio harveyi* Y6 strain has been retrieved from the previous study of Kwankijudomkul et al. 2021 (Kwankijudomkul et al., 2021). Afterward, the sequence was stored in FASTA format and sent for *in-silico* analysis to many prediction services. To provide an initial assessment of the activity of the targeted protein relative to non-redundant proteins, a similarity search was undertaken utilizing the NCBI protein database (Boeckmann et al., 2003) to look for proteins with comparable properties using the BLASTp tool (Johnson et al., 2008).

Phylogenetic Construction, Sequence Alignment, and Physiochemical Properties Analysis

Multiple sequence alignments were done using the biological sequence alignment editor (BioEdit) between the HP33 and proteins with similar structures and compositions (Alzohairy, 2011). This phylogenetic study was conducted using a customized version of MEGA (Molecular Evolutionary Genetics Analysis) (https://megasoftware.net/). ExPASy ProtParam was used to calculate many chemical and physical properties, including molecular weight, theoretical pI, instability index, extinction coefficient, atomic composition, anticipated half-life, aliphatic index, and GRAVY value (Gasteiger et al., 2003).

Subcellular Localization and Virulence Risk Factor Analysis

Subcellular localization was predicted by CELLO (C. Yu & Hwang, 2008). PSORTb subcellular localization estimations were compared to the findings (N. Y. Yu et al., 2010), PSLpred (Bhasin, Garg, & Raghava, 2005), and SOSUIGramN. TMHMM (Möller, Croning, & Apweiler, 2001), HMMTOP (Tusnády & Simon, 2001), and CCTOP (Dobson, Reményi, & Tusnády, 2015) were used for the topology prediction. A transmembrane region is any portion of a protein that is hydrophobic. Moreover, MP3 and VICMpred were used to make predictions about the pathogenicity of HP33. Adding SVM and HMM to the MP3 server may identify pathogenic proteins from metagenomic and genomic datasets with greater accuracy and efficiency (Gupta, Kapil, Dhakan, & Sharma, 2014). Structures, amino acid, and dipeptide

composition of bacterial protein sequences are appraised using SVM-based algorithms in the VICMpred service, resulting in an overall accuracy of 71.75 percent (Saha & Raghava, 2006). The VFDB server leverages the VFanalyzer pipeline to reliably identify potentially dangerous strains to perform a continuous and comprehensive sequence similarity search across the hierarchical prebuilt datasets (Liu, Zheng, Jin, Chen, & Yang, 2019).

Identification of Conserved Domains, Motifs, Folds, Families, and Superfamilies

Function predictions for HP33 were made using various functional databases and methods. These included CDD, Pfam, InterProScan, and SMART. The database of conserved domains was searched (CDD, available at NCBI)(Marchler-Bauer et al., 2005) for preserved domains. The HP33 motifs were investigated using MEME suites (Bailey et al., 2009). Pfam was used to assign the protein's evolutionary links (Finn, 2005) and SuperFamily (Wilson, Madera, Vogel, Chothia, & Gough, 2006) database. InterProScan, a protein sequence analysis and classification program, was used for the functional investigation of the protein (Hunter et al., 2008). The InterProScan software identifies input sequences and compares them to the InterPro protein signature databases (Jones et al., 2014). When comparing input sequences to database entries, SMART looks for sequences with similar domain designs and characteristics (Letunic, Doerks, & Bork, 2012). The PFP-FunD SeqE server (Shen & Chou, 2009) was employed to detect protein folding patterns.

Evaluation of Performance

The accuracy of the in silico approaches used for functional assessment and domain identification was assessed using ROC curve analysis on HP33 (Bradley, 1997). The five tools were examined on three levels to determine their efficiency. In the provided data, there were two columns. In the first column, genuine pessimistic predictions were given binary 0, while accurate optimistic predictions were given binary 1. One to five integer values were assigned to the second column: the more significant the number, the higher the degree of confidence. The input data was sent to the ROC Analysis server (http://www.jrocfit.org) (Alemayehu & Zou, 2012) following format 1. Accuracy, sensitivity, specificity, and ROC curve area under the curve (AUC) were calculated using online ROC software.

Prediction of 2D and 3D Structure, Refinement, Validation, and Assessment of Model Quality

Protein secondary structure predictions were made using PROTEUS Structure Prediction Server 2.0 (Montgomerie, Sundararaj, Gallin, & Wishart, 2006). Its algorithm employs artificial neural networks and machine learning methods. Protein secondary structures (β sheets, α helixes, and coils) are predicted using a server-side algorithm with a front-end website. The 3D structure of the target protein was predicted using the RaptorX server (http://raptorx.uchicago.edu/) (Xu, McPartlon, & Li, 2021). GalaxyWeb was used to improve the protein's 3D structure. Homology modeling relies on 3D protein structures that have been experimentally verified; hence the structure's reliability is essential. ProSA-web was consulted for basic validation of the proposed protein model. The server predicted the z-score, which indicates the model's overall personality (Mou, Islam, & Mahfuj, 2021). If the z-scores of the expected model fall outside the range for local proteins, the structure is incorrect (Islam & Mou, 2022). To evaluate protein quality, the Ramachandran Plot Server was used (https://zlab.umassmed.edu/bu/rama/). Three online tools, PROCHECK, Verify3D, and ERRAT Structure Evaluation, were then used to analyze the resulting 3D model.

Protein-protein Interaction Analysis

The STRING database, a pre-computed international repository for gathering and studying protein-protein connections, was established because of the importance of context information (von Mering et al., 2003). Therefore, STRING includes a one-of-a-kind scoring system that compares various connections to a standard reference set, yielding just one confidence score per projection. It is much simpler to explore the modularity of biological processes when using the graphical representation of weighted protein connections derived from a network. This representation offers a high-level view of functional linkage (Sivashankari & Shanmughavel, 2006). This research used the STRING database, which can be found online at http://string-db.org/. This database analyzes the structural and functional links between proteins to identify known and predicted protein interactions. High-throughput research, genomic context (Conserved) This judgment was made based on co-expression and prior knowledge. The following interaction data sources are quantitatively included in this database (Szklarczyk et al., 2015).

Protein Disulfide Bonds

It is impossible for a protein to fold into a functional and stable structure without forming disulfide bonds between its cysteine residues. Disulfide linkages throughout a hypothetical protein were predicted using CYSPRED and DIANA to get insight into the experimental structural analysis and protein stability. With CYSPRED, you may see whether the cysteine residues in your query protein form disulfide bridges or links. Starting with the residue chain's non-binding state, CYSPRED is a neural network-based

predictor that has been taught to correctly discriminate the bonding states of cysteine in proteins (Grützner et al., 2009). DIANA was also used since it helps anticipate disulfide linkages in a protein sequence input. Appropriately estimating disulfide bridges is critical for understanding the function of a hypothetical protein and tertiary prediction approaches (Ferrè & Clote, 2005). The tertiary structure of hypothetical proteins can be used as a reference when identifying docking sites; it can help us create medicines to treat diseases linked to potential gene alterations.

Ligand Binding Site Prediction

The Galaxy server was used to predict protein-ligand binding sites in hypothetical proteins. To determine where a ligand will attach to a protein, GalaxySite uses a technique called protein-ligand docking. The structure can be either experimental (with or without ligand) or model (with or without ligand). GalaxySite employs the GalaxyTBM approach to predict the structure without a refinement step if a protein sequence is supplied. The complex structures of comparable proteins discovered by HHsearch predict the binding ligands. A ligand docking approach called LigDockCSA is then used to predict the protein-ligand complex structures (Heo, Shin, Lee, & Seok, 2014).

Homo-oligomer Models Prediction and Detection of the Active Sites

The GalaxyHomomer service (http://galaxy.seoklab.org/homomer) predicts a protein's homo-oligomer structure from its monomer amino-acid sequence. Homo-oligomerization often occurs in nature and is linked to the physiological activities of proteins, such as metabolism, signaling, and immunity. Information on the homo-oligomer structure is critical for gaining molecular-level knowledge of protein functions and regulation (Baek, Park, Heo, Park, & Seok, 2017). If you give GalaxyHomomer an amino-acid sequence, it will generate five models by probing for sequence, structure, and ab initio docking similarities. Model 1 automatically recognizes and remodels less dependable loop or terminal sections (known as ULR), and the whole oligomer structure is simplified using the Galaxy Refine complex. If the monomer composition is supplied as input, five oligomer models are built using a structural similarity-based technique and ab initio docking. During oligomer modeling, those sections are adjusted if the user supplies less reliable elements of the input structure.

Using the Computed Atlas of Surface Topography of Proteins (CASTp), the active site of this protein was found (Dundas et al., 2006). It is a web-based application that helps you find, define, and measure concave surface regions on 3D protein structures. In a comprehensive, systematic, and quantitative fashion, CASTp retrieves the topographical characteristics of a protein. It is feasible to detect and quantify functional sites outside and inside a protein's three-dimensional structure. Because of this, it is now a necessary tool for determining which parts of proteins and amino acids play an essential role in ligand binding (Islam & Mou, 2022; W. Tian, C. Chen, X. Lei, J. Zhao, & J. Liang, 2018).

Results and Discussion

Similarity Identification, Multiple Sequence Alignment, and Phylogeny Analysis

Similarities with other proteins were discovered using BLASTp against a non-redundant database (Table 1). Multiple sequence alignment was utilized to connect the FASTA sequences of the putative protein HP33 to homologous known proteins. Phylogenetic analysis was employed to back up homology evaluations of proteins at the complex and subunit levels. The alignment and BLAST data were used to produce a phylogenetic tree that provides a similar picture of the protein (Figure 1). Branches are also compared based on distance. From the phylogenetic tree, it is evident that the HP33 protein sequence grouped with the hypothetical protein of *V. harveyi* (Accession No. 009697735).

Physicochemical Features

The average isotopic weights of the amino acids found in a protein are added to the isotopic mass of one water molecule, which is then multiplied by ProtParam to get the molecular weight of the protein (Wilkins et al., 1999). Ser 31, Leu 27, Glu 23, Gly 20, Thr 20, Asp 19, Val 18, Ile 18, Lys 18, Gln 17, Ala 17, Asn 15, Pro 12, Phe 11, Met 8, Arg 8, Tyr 8, Cys 3 and Trp 3 are the most prevalent amino acids in the protein. The calculated molecular weight was 32987.00 Da, and the predicted pI was 4.63, indicating that the protein was negatively

Table 1. Non-redundant sequencing generated a protein with comparable characteristics

Protein ID	HP33			
	Organism	Protein Name	Identity (%)	e value
WP_009697735.1	*Vibrio harveyi*	hypothetical protein	99.33	0.00
WP_045488642.1	*Vibrio harveyi*	hypothetical protein	99.00	0.00
WP_045455267.1	*Vibrio campbellii*	hypothetical protein	96.99	0.00
WP_104035359.1	*Vibrio jasicida*	hypothetical protein	65.02	0.00
WP_045495220.1	*Vibrio hyugaensis*	hypothetical protein	64.29	0.00

charged. The pKa values of amino acids are used to figure out the pI of a protein. It plays a crucial role in determining how a protein responds to changes in pH (Pace, Grimsley, & Scholtz, 2009). There were 26 positively charged (Arg + Lys) and 42 negatively charged (Asp + Glu) residues. When your protein is in a test tube, an instability index tells you how stable it is. A protein with an instability index of less than 40 is projected to be stable, whereas a score of more than 40 indicates that the protein might be unstable (Gamage, Gunaratne, Periyannan, & Russell, 2019). The calculated instability index of 37.34 classed the protein as inconsistent. The aliphatic index is the number of amino acids in a protein with an aliphatic side chain, such as leucine, alanine, valine, and isoleucine (Nehete, Bhambar, Narkhede, & Gawali, 2013). The proteins' 81.84 aliphatic index suggests that they retain their structural integrity throughout a broad temperature range. The gravitational constant was found to be -0.370. In this case, the protein is nonpolar since the GRAVY test was negative. One obtains a GRAVY value by multiplying the hydropathy scores of each amino acid residue by the total number of residues in the sequence. The material is more hydrophobic if it has a more excellent positive score (J. Zhang et al., 2016). *In vitro*, mammalian reticulocytes had a half-life of 30 hours, yeast had a half-life of > 20 hours, and Escherichia coli had a half-life of > 10 hours. And the protein's molecular formula was discovered to be $C_{1447}H_{2286}N_{382}O_{475}S_{11}$.

Hypothetical Protein Functional Annotation

This putative protein sequence was found to include a domain belonging to a family of uncharacterized functions using the conserved domain search tool (accession No. DUF6068). Two more domain search tools, InterProScan and Pfam, were used to confirm the results. The Pfam server predicted the Sodium: a family of unknown functions at positions 1-65 and 193-276 with an e-value of 6.8e and 15e, respectively. This domain represents a gene in a cluster involved in biosynthesis (BGC). According to MIBiG, this BGC is part of the non-ribosomal peptide (NRP) and polyketide biosynthesis families. This family appears to be dominated by cystobacterineae and includes a protein from the *Myxococcus virescens* benzamide biosynthetic gene cluster (Wenzel et al., 2015). According to the PFP-FunDSeqE service, the protein in concern has a predicted fold type that is immunoglobulin-like. The immunoglobulin (Ig)-like the domain is a protein domain comparable to the Ig domains of immunoglobulins in terms of its amino acid sequence and three-dimensional structure. Ig domains are the only kind of domains that include the immunoglobulin fold, which consists of 70–110 amino acids. In conventional Ig-like domains, seven to ten strands are spread out over two sheets in a structure, and connectivity is considered typical. The MEME suite explored the *V. harveyi* Y6 strain HP33 protein motif. The study focused on five different topics for HP33. Additionally, the MEME program could predict width, positions, and E-value. In the case of HP33, there was either one instance (of a contributing motif site) per sequence or none. The E-value came out to be 7.59e-30 after being calculated. It was discovered that the pattern's width ranged between 6 and 50 pixels (inclusive). The motif analysis for HP33 is shown in Figure 2.

Figure 1. Multiple sequence phylogenetic tree

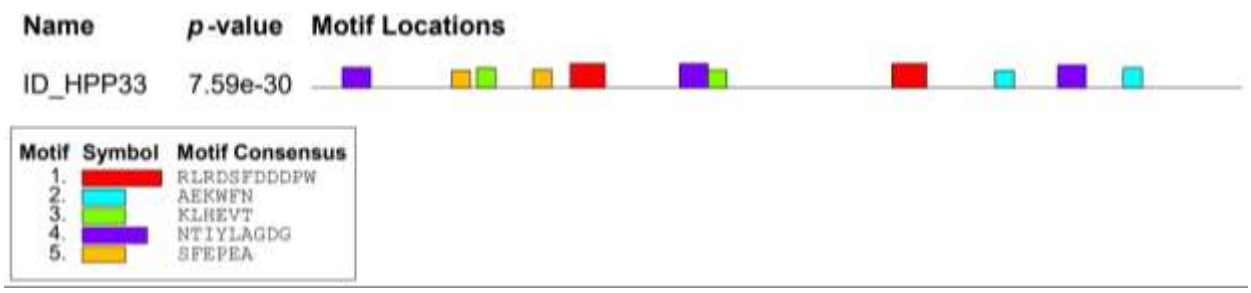

Figure 2: Motif analysis of the HP33

Evaluation of Performance

ROC analysis demonstrated high reliability and trust for the five *in-silico* tools and servers (Kumar, Maan, Singh, & Kaur, 2017). The prediction confidence was deemed high when two or more tools anticipated the same result for HP33. The accuracy, sensitivity, and specificity of the tools employed in functional annotation were reported at 95.96 percent, 96.804 percent, and 97.762 percent, respectively (Table 2 and Figure 3), indicating that the results were acceptable (Ahmed, 2022).

Nature of Subcellular Localization

We need to know the subcellular localization of hypothetical proteins to understand their function because various cellular sites reflect distinct responsibilities. This data can also be used to develop a medication targeting the target protein (J. Wang, Sung, Krishnan, & Li, 2005). PSORTb and PSLpred validated CELLO's prediction of subcellular localization analysis. The outer membrane was predicted to be the HP's subcellular site (Table 3). The Outer Membrane protein, unlike THMM, is not expected to include transmembrane helices. These data all point to the protein being an Outer Membrane protein (Islam, Sanjida, Mou, Mahfuj, & Nasir, 2022). The native sub-cellular compartment of a protein is one aspect of its function. Thus, predicting localization is an essential step toward predicting function.

Virulence Factor Prediction

When transmitting a disease, pathogens must create virulence factors that allow them to evade the host's immune response. To effectively develop vaccines and use reverse vaccinology, understanding the molecular pathways behind pathogenicity is essential (Chaudhuri & Ramachandran, 2014). Virulence factors are important for microbial pathogenesis. A mutation of a virulence factor from a virulent pathogen will attenuate the pathogen strain ("Front-matter," 2015). Using the MP3 server, HP33 has been discovered as a virulence factor and pathogenic protein. According to the VICMpred server, this protein is involved in cellular activities (Ahmed, 2022).

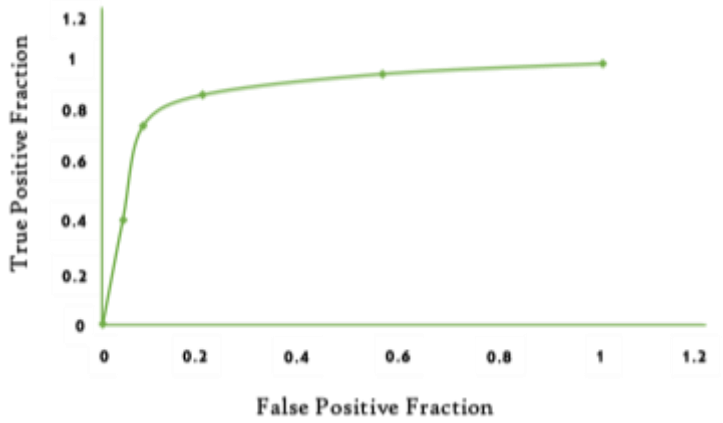

Figure 3. Utilizing bioinformatics tools to analyze HP33 functional annotations statistically

Table 2. ROC curve assessment analysis

SL. NO.	Tools/Servers	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC area
1.	InterProScan	100	100	100	1
2.	Pfam	100	100	100	.98
3.	CDD	96.7	97.1	98.1	.79
4.	SMART	89.3	90.0	93.7	0.4
5.	MEME suites	93.8	96.92	97.01	.85
	Average	95.96	96.804	97.762	.804

Table 3. Sub-cellular localization of hypothetical protein predicting from different servers

No.	Analysis	Result
1.	CELLO 2.5	Outer Membrane
2.	PSORTb	Outer Membrane
3.	PSLpred	Outer Membrane
4.	TMHMM 2.0	No transmembrane helices present
5.	HMMTOP	One transmembrane helices present (6-25)
6.	CCTOP	No Transmembrane protein

Secondary Structure Analysis

Secondary structure in proteins is formed through intermolecular and intramolecular hydrogen bonding between the amide groups in the primary structure. Protein's two most critical secondary structures are α helices and β sheets. The right-handed helix configuration of the alpha helix. Carbonyl (CO) and amino (NH) hydrogen bonds stabilize the fourth amino acid, the C-terminal amino acid. Beta sheets are two-dimensional structures formed when beta strands are hydrogen-bonded together (Han, Zhang, Ishida, & Froimowicz, 2017). The PROTEUS Structure Prediction Server 2.0 study found that the percentages of α-helices, β-sheet content, coil content, the number of sequence alignments used for ab-initio projections, and the overall confidence value were 22%, 32%, 45%, 1%, and 75.3%, respectively.

3D Structure Prediction, Model Quality Refinement, and Assessment

Hypothetical proteins' biochemical or biophysical roles can be deduced from their structures (Bernstein et al., 1977). The 3D structure of uncharacterized and hypothetical proteins can aid in function assignment. Since protein folding patterns are often conserved throughout evolution, structure-based comparisons may discover homologs when sequence-based comparisons are unproductive (Sivashankari & Shanmughavel, 2006). As a result, structure-based molecular function attribution is a potential technique for assigning biological proteins and discovering new motifs on a large scale. Protein model 1 was selected after using the RaptorX server to predict the protein's three-dimensional structure of interest. RaptorX predicts 3D structures for protein sequences without Protein Data Bank homologs (PDB). According to RaptorX, secondary and tertiary structures, solvent accessibility, disordered areas, and solvent accessibility are all predicted using a sequence input (Källberg, Margaryan, Wang, Ma, & Xu, 2014). In Discovery Studio, the tertiary and refined model 1 were chosen and visualized (Figure 4A). Ramachandran plots, which illustrate the distribution of and angles concerning the model boundaries, were used by PROCHECK to evaluate the stability of the galaxy server-enhanced model (Figure 4B). In the most desired locations, a viable model covers 84.9% of the residues. Verify3D and ERRAT were used to validate a 3D structural model of the target sequence, which was then compared to the established model (Islam, Sanjida, Mahfuj, Islam, & Mou, 2022). The model has a unique environmental profile, as evidenced by the fact that 82.23 % of residues have an average score of ≥0.2 on the Verify3D graph and an overall quality factor of 71.6157 in ERRAT. The 3D structure was later changed via the YASARA energy minimization server. The estimated energy was –61,130.4 kJ/mol before energy minimization, but it was decreased to –254,513.5 kJ/mol after energy minimization (by three rounds of the steepest descent technique), making the modeled structure more stable. In addition, the ProSA web server analysis yielded a Z score of -4.09, indicating that the model is valid (Figure 4C).

Analysis of Protein-protein Interactions and Protein Disulfide Bonds

Proteins commonly interact in a mutually dependent manner to perform a comparable function. To trigger transcription, for example, transcription factors interact with one another. As a result, the activities of proteins may be inferred from their interactions with other proteins (Sivashankari & Shanmughavel, 2006). Interactions between residues determine protein functioning. To estimate the protein's probable functional interactions, we employed the STRING 10.0 algorithm (Figure 5). Supplementary

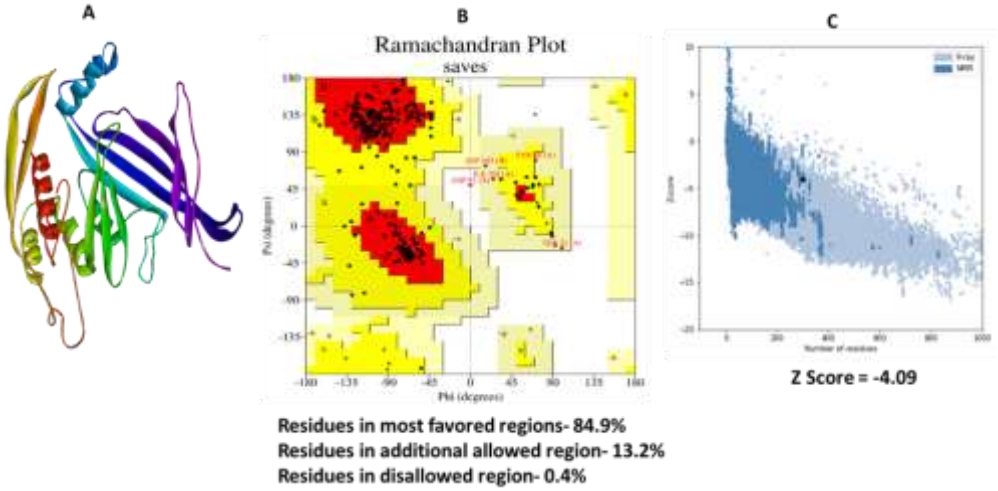

Figure 4. (A) Predicted tertiary structure of the hypothetical protein, (B) Ramachandran plot analysis of the refine model, and (C) Z-score results of the refine model from ProSA server.

Table 1 lists the discovered functional partners along with their scores. Furthermore, the most common HP33 functional partner proteins have yet to be annotated in the database, suggesting that HP33 is a unique functional protein from the *V. harveyi* Y6 strain. Similarly, disulfide linkages play a crucial role in protein folding from a structural and functional standpoint. Therefore, research into protein disulfide bonds is essential for elucidating their advanced structure and function. As a result, protein disulfide bond analysis is crucial for revealing proteins' higher structure and biological processes. Protein characterization during biopharmaceutical manufacture relies heavily on disulfide bonding because improper disulfide bond formation or exchange might encourage antibody aggregation (L. Zhang, Chou, & Moo-Young, 2011). The quality of the tertiary model predicted by CYSPRED indicates that it contained two bonding states and one non-bonding state (Islam, Mou, Sanjida, & Mahfuj, 2022a). Additionally, three binding cystine sequences were predicted by DIANA in different regions of the protein (Table 4). These servers hypothesized that the HP33 protein's high-order structure was stabilized by a disulfide bond between three cysteine residues locked away within the protein.

Ligand Binding Interactions

Target models were matched with the PDB file of the best-predicted domain-A model to predict ligand binding sites on the Galaxy server. A galaxy server suggested three models with different ligands. The findings are likewise divided into three portions by the Galaxy server. Templates for protein-ligand complexes predicted ligand-binding residues and model binding poses (Table 5; Figure 6 (A, B)). LIGPLOT was used to examine interactions at the anticipated ligand-binding site. Table 5 summarizes the results of the protein-ligand interaction investigation. Figure 7 depicts the most likely protein-ligand binding poses and a template model for another protein-ligand complex. The definition of residue-ligand interaction impacts ligand-binding residues (Islam, Mou, Sanjida, & Mahfuj, 2022b). A binding site residue is one in which the distance between an amino acid residue and a ligand atom is smaller than the total of the two atoms' van der Waals radii + 0.5Å (Chen et al., 2016). Furthermore, the ligand HEC is a tiny molecule known as Ferroheme C (DrugBank Accession Number DB03317). The chemical formula $C_{34}H_{36}FeN_4O_4S_2$ has a molecular weight of roughly 684.65 KDa and a monoisotopic value of 684.152734. Valine, on the other hand, is a branched-chain essential amino acid with stimulant action (DrugBank Accession Number DB00161). It aids in the development of muscle and the restoration of damaged tissues. It represents a critical step in penicillin biosynthesis (Arakawa et al., 2010). $C_5H_{11}NO_2$, the chemical formula, has a molecular weight of 117.5 KDa and a monoisotopic value of 117.078978601.

Homo-oligomer Models Prediction and Ab Initio Docking Results Analysis

By running the Naccess program on the Galaxy server, we may get the total area of the interface ($Å^2$) between any two chains (Baek et al., 2017). The protein sequence identity between a query and a template is shown in Figure 9. It goes from 0 (unique) to 100 (individual) (identical). Table 6 displays the oligomer templates, subunit count, and interface area. Predicted models based on ab initio docking are shown as a docking score, whereas those based on templates offer either sequence identity or structural similarity. The structural similarities between the query protein and

Table 4: CYSPRED and DIANA predict cysteine residues important in disulfide bonding

CYSPRED				DIANA	
Cysteine	Prediction	Reliability	Distance	Score	Bonded cysteine
CYS 23	Bonding State	4	189	0.01073	IALYGCGGGGS-ADNAQCKTTWS
CYS 212	Bonding State	8	228	0.99116	IALYGCGGGGS-TYNLVCDGMEL
CYS 251	NON-Bonding State	9	39	0.01104	ADNAQCKTTWS-TYNLVCDGMEL

Table 5: Predicted ligand-binding residues

Ligand Name	Molecular Weight	DrugBank Name	Binding Residues
HEC	618.50	Ferroheme C	272F 273I 274E 275Q 276V 277E 287R 288E 289T 290K 292T
VAL	117.15	Valine	3K 4R 7L 25S 26G 27S 90Q 92D 93P

Table 6. Ab initio Docking Results of Five homo-oligomer models of HP33

Model No.	Number of subunits	Interface area $Å^2$	Docking score
1	2 mer	1200.9	2024.170
2	2 mer	2721.9	1951.639
3	2 mer	1179.9	1855.651
4	2 mer	1469.9	1852.032
5	4 mer	6666.6	1269.693

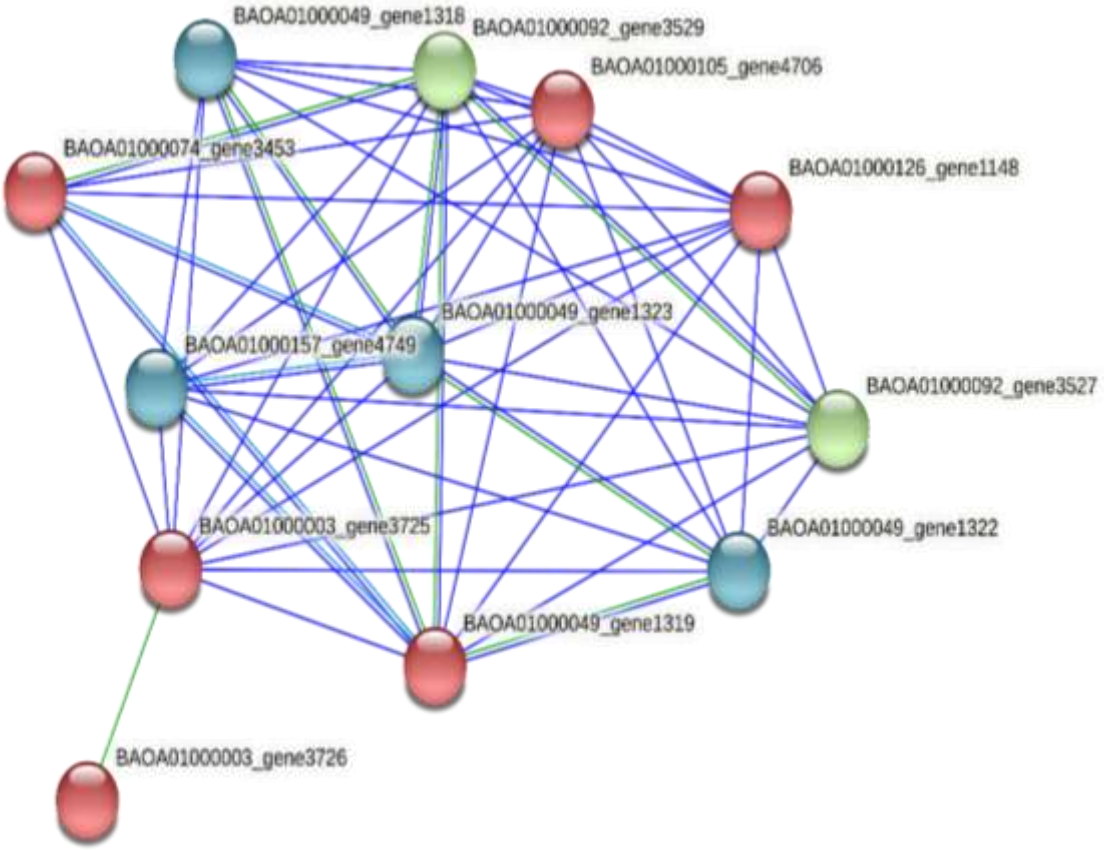

Figure 5: String (Protein-protein interactions) analysis of hypothetical protein

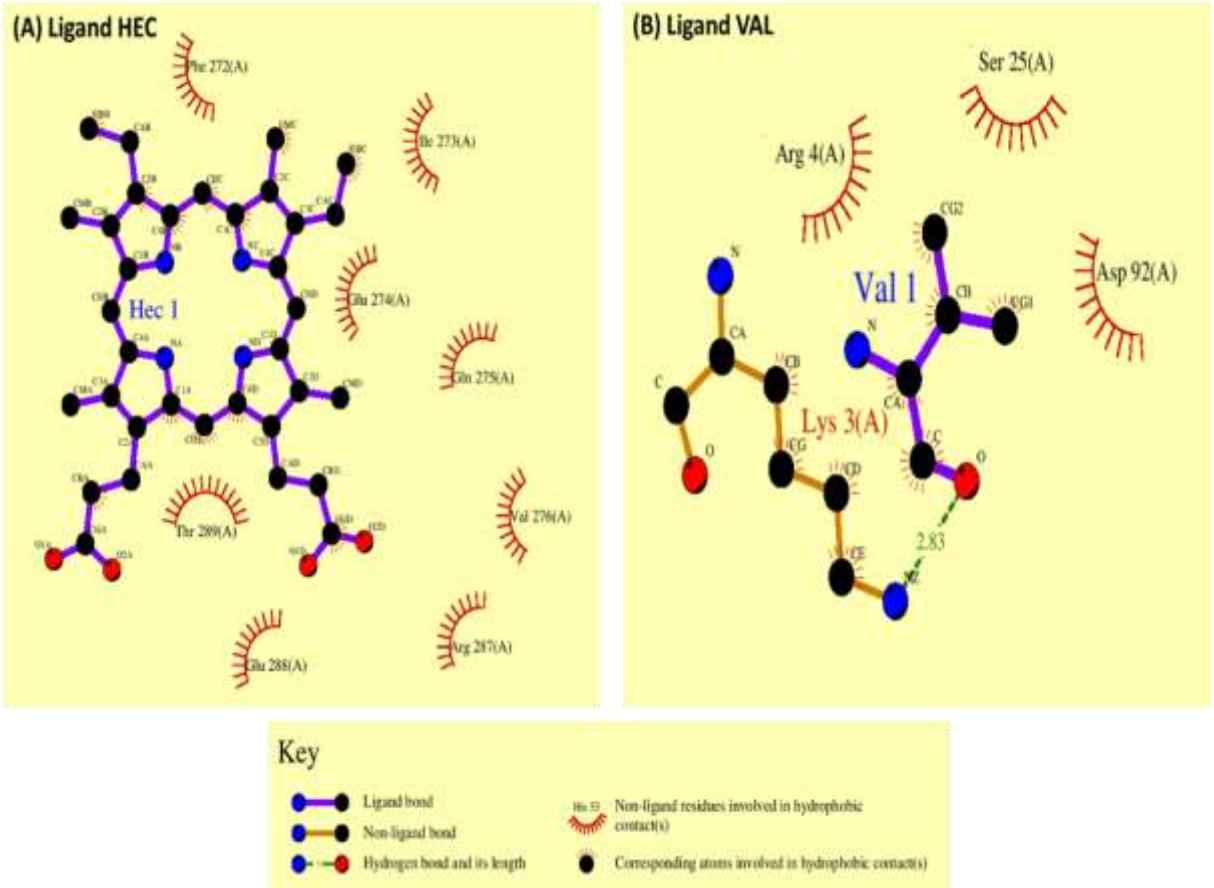

Figure 6. Interaction analysis of HP33 protein with two ligand molecules (A) Ligand HEC and (B) Ligand VAL

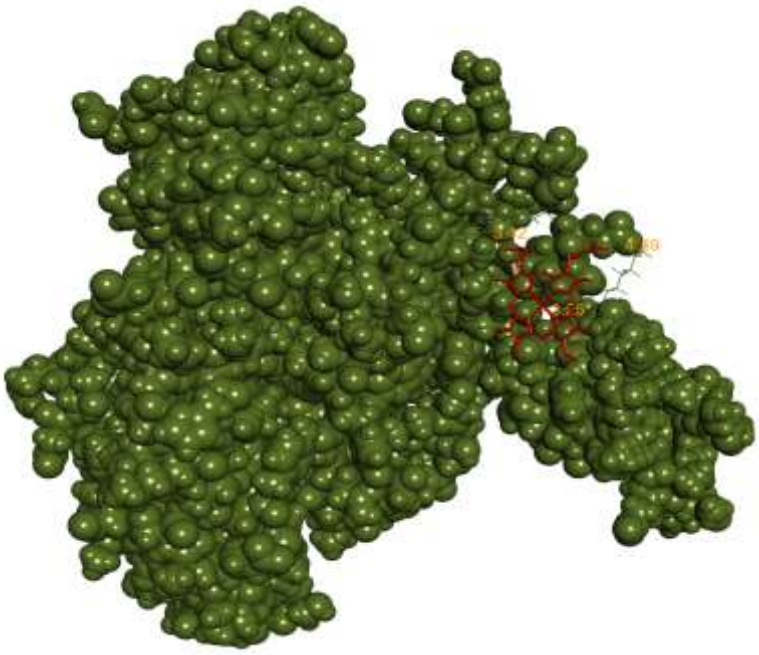

Figure 7. (A) Predicted binding pose.

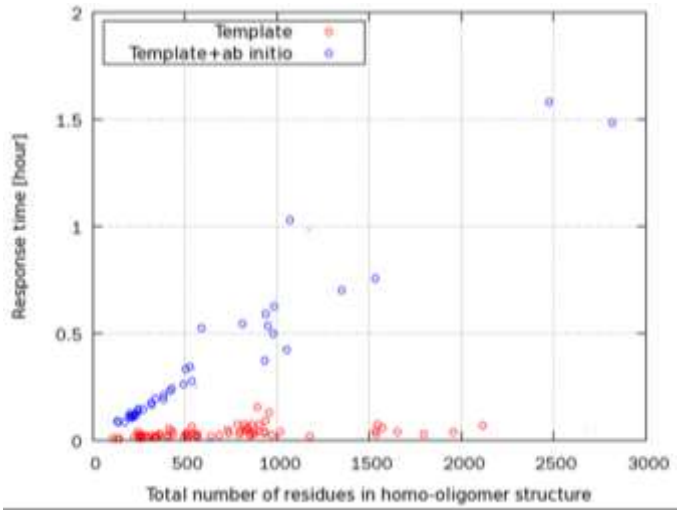

Figure 8. Response time on the total number of residues in homo-oligomers models

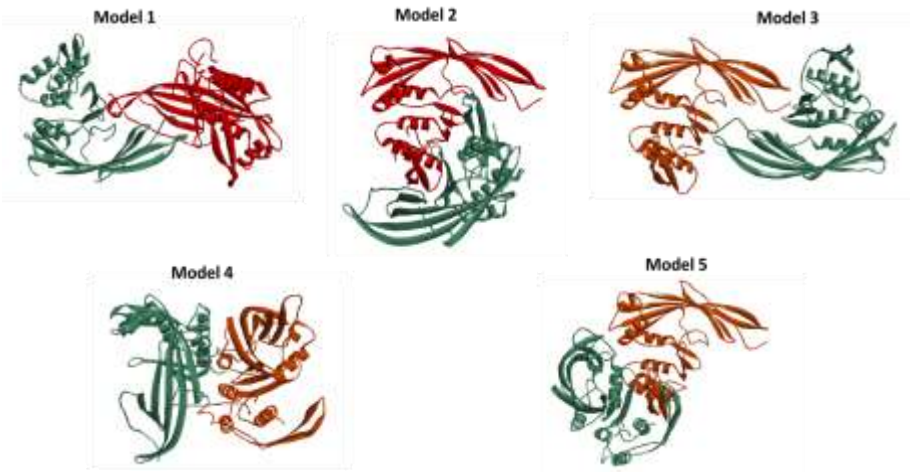

Figure 9. Homo-oligomer models of HP33 protein. Red color indicates Chain A and the Green color indicates Chain B.

the template protein are broken down by docking score (as determined by TM-align) and shown in Table 6. The docking score increases in direct proportion to the quality of the score. In a given sequence, the top five proteins with S >0.2 times the maximum S for the whole set and S >0.7 times the leading S for the oligomeric state in question are selected as templates. Suppose the number of discovered templates using this sequence-based technique is fewer than five for the given oligomeric state. Different templates are chosen using the monomer structure predicted by template-based modeling (Ko, Park, & Seok, 2012). To choose structure-based templates, S is ranked among those with monomer structures identical to the supplied monomer structure (TM-score derived using TM-align>0.5) and in the required oligomeric state (Y. Zhang & Skolnick, 2005). The total amount of residues in a homo-oligomer determines the response time. In this work, a template-based technique (red dots) (Figure 8) is applied, and the homo-oligomers structure prediction is completed in most cases in under 2 hours (Baek et al., 2017). Figure 9 shows five distinct homo-oligomer models of HP33 protein chains A and B.

Active Site Detection

Figure 10 (A) shows that the CASTp v.3.0 algorithm predicts that the modeled protein has 41 unique active sites. CASTp is a server-side database that can locate protein regions, delineate their boundaries, calculate their area, and determine their dimensions. Surface proteins that contain hidden cavities and cavities are also suspected. To define a pocket and volume spectrum or vacuum, one uses the surfaces of solvent-accessible molecules (Richard surface) and molecular surfaces (Connolly surface). Studying the functional regions and surface properties of proteins is possible using CASTp. CASTp is an interactive GUI that can instantly evaluate user-submitted building designs (Wei Tian, Chang Chen, Xue Lei, Jieling Zhao, & Jie Liang, 2018). Area 2602.319 and volume 4124.259 were used to rank the active sites of the model protein (Figure 10 (B)).

Conclusion

The results of this study show that the putative protein domain is crucial for biosynthesis and immunoglobin production. It was shown to be a nonpolar protein with a single surface-exposed domain. Novel antibacterial therapies may be developed because of the presence and widespread distribution of this putative protein domain in the *V. harveyi* strain. Protein-ligand docking studies and other investigation forms are being conducted to identify the critical amino acids in ligand binding. Because the ab-initio docking and homo-oligomer studies of the hypothetical protein HP33 provide immunological insights, they may be helpful to researchers working to create innovative treatments to treat SDMND.

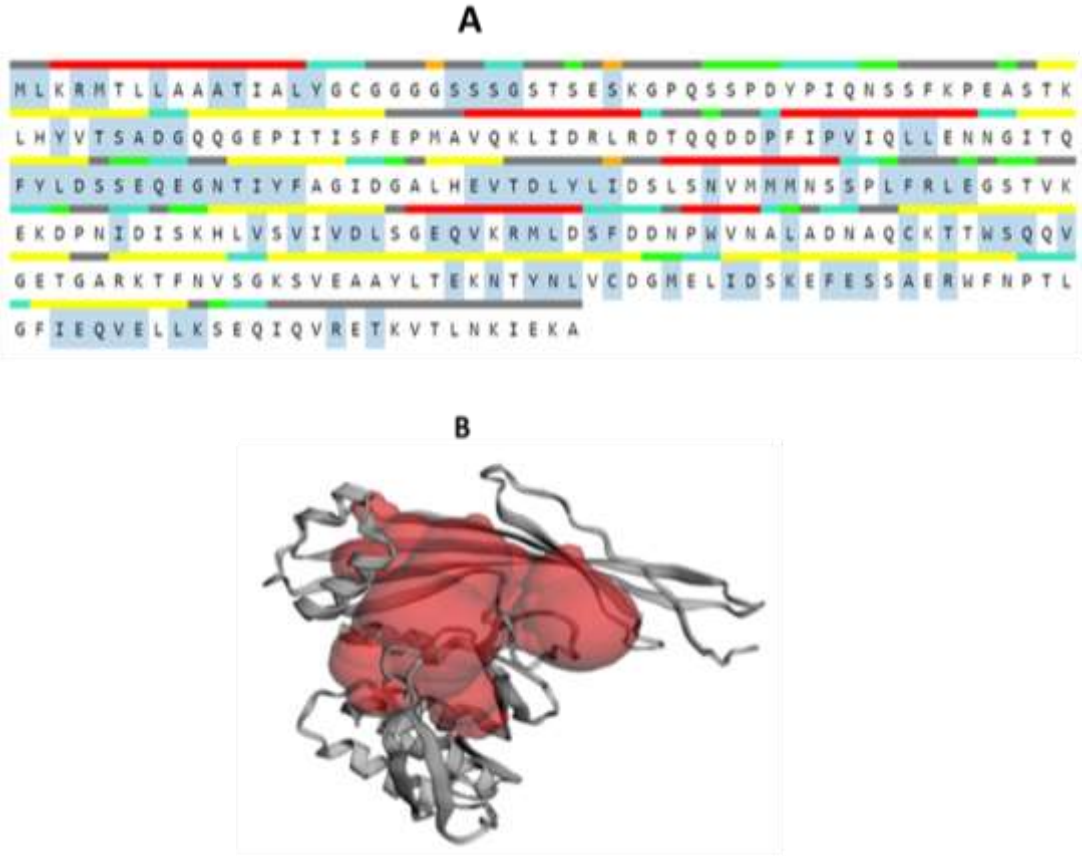

Figure 10: Active location of the hypothetical protein. (A) The active site of amino acid residues (Blue color) (B) Red color indicates the total volume and area of the active site of the protein

Ethical Statement

Not applicable.

Funding Information

No funding available for this research.

Author Contribution

Sk Injamamul Islam: Conceptualization, Writing - review and editing; Saloa Sanjida: Data Curation, Formal Analysis, Investigation, Methodology, Visualization and Writing -original draft; Md. Akib Ferdous: Funding Acquisition, Project Administration, Resources, Writing - review and editing; and Nasim Habib: Supervision, Writing - review and editing.

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 author thanks Dr. Foysal Ahmed Sagore and Dr. Kazi Abdus Samad for helpful comments.

References

Ahmed, S. (2022). *In Silico Characterization of Essential Hypothetical Proteins from Francisella tularensis Schu S4 Strain*.

Alemayehu, D., & Zou, K.H. (2012). Applications of ROC Analysis in Medical Research: Recent Developments and Future Directions. *Academic Radiology, 19*(12), 1457-1464. https://doi.org/10.1016/j.acra.2012.09.006

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

Arakawa, M., Yanamala, N., Upadhyaya, J., Halayko, A., Klein-Seetharaman, J., & Chelikani, P. (2010). The importance of valine 114 in ligand binding in beta(2)-adrenergic receptor. *Protein science: a publication of the Protein Society, 19*(1), 85-93. https://doi.org/10.1002/pro.285

Baek, M., Park, T., Heo, L., Park, C., & Seok, C. (2017). GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Research, 45*(W1), W320-W324. https://doi.org/10.1093/nar/gkx246

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., ... Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research, 37*(suppl_2), W202-W208. https://doi.org/10.1093/nar/gkp335

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., . . . Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol, 112*(3), 535-542. https://doi.org/10.1016/s0022-2836(77)80200-3

Bhasin, M., Garg, A., & Raghava, G.P.S. (2005). PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics, 21*(10), 2522-2524. https://doi.org/10.1093/bioinformatics/bti309

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., ... Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research, 31*(1), 365-370. https://doi.org/10.1093/nar/gkg095

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Chaudhuri, R., & Ramachandran, S. (2014). Prediction of virulence factors using bioinformatics approaches. *Methods in molecular biology (Clifton, N.J.), 1184*, 389-400. https://doi.org/10.1007/978-1-4939-1115-8_22

Chen, D., Oezguen, N., Urvil, P., Ferguson, C., Dann, S.M., & Savidge, T.C. (2016). Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Science advances, 2*(3), e1501240-e1501240. https://doi.org/10.1126/sciadv.1501240

Crane, M., & Hyatt, A. (2011). Viruses of fish: an overview of significant pathogens. *Viruses, 3*(11), 2025-2046. https://doi.org/10.3390/v3112025

Dobson, L., Reményi, I., & Tusnády, G.E. (2015). CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Research, 43*(W1), W408-W412. https://doi.org/10.1093/nar/gkv451

Dong, H.T., Taengphu, S., Sangsuriya, P., Charoensapsri, W., Phiwsaiya, K., Sornwatana, T., ... Senapin, S. (2017). Recovery of Vibrio harveyi from scale drop and muscle necrosis disease in farmed barramundi, Lates calcarifer in Vietnam. *Aquaculture, 473*, 89-96. https://doi.org/10.1016/j.aquaculture.2017.02.005

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 Research, 34*(suppl_2), W116-W118. https://doi.org/10.1093/nar/gkl282

Ferrè, F., & Clote, P. (2005). DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res, 33*(Web Server issue), W230-232. https://doi.org/10.1093/nar/gki412

Finn, R.D. (2005). Pfam: the protein families database. In *Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics*.

Front-matter. (2015). In Y.-W. Tang, M. Sussman, D. Liu, I. Poxton, & J. Schwartzman (Eds.), *Molecular Medical Microbiology (Second Edition)* (pp. i-iii). Boston: Academic Press.

Gamage, D.G., Gunaratne, A., Periyannan, G.R., & Russell, T.G. (2019). Applicability of Instability Index for in vitro Protein Stability Prediction. *Protein Pept Lett, 26*(5), 339-347. https://doi.org/10.2174/0929866526666190228144219

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research, 31*(13), 3784-3788. https://doi.org/10.1093/nar/gkg563

Grützner, A., Garcia-Manyes, S., Kötter, S., Badilla, C.L., Fernandez, J.M., & Linke, W.A. (2009). Modulation of titin-based stiffness by disulfide bonding in the cardiac

titin N2-B unique sequence. *Biophysical journal, 97*(3), 825-834. https://doi.org/10.1016/j.bpj.2009.05.037

Gupta, A., Kapil, R., Dhakan, D.B., & Sharma, V.K. (2014). MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS One, 9*(4), e93907-e93907. https://doi.org/10.1371/journal.pone.0093907

Han, L., Zhang, K., Ishida, H., & Froimowicz, P. (2017). Study of the Effects of Intramolecular and Intermolecular Hydrogen-Bonding Systems on the Polymerization of Amide-Containing Benzoxazines. *Macromolecular Chemistry and Physics, 218*(18), 1600562. https://doi.org/10.1002/macp.201600562

Heo, L., Shin, W.H., Lee, M.S., & Seok, C. (2014). GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Res, 42*(Web Server issue), W210-214. https://doi.org/10.1093/nar/gku321

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., ... Yeats, C. (2008). InterPro: the integrative protein signature database. *Nucleic Acids Research, 37*(suppl_1), D211-D215. https://doi.org/10.1093/nar/gkn785

Ijaq, J., Chandra, D., Ray, M.K., & Jagannadham, M.V. (2022). Investigating the Functional Role of Hypothetical Proteins from an Antarctic Bacterium Pseudomonas sp. Lz4W: Emphasis on Identifying Proteins Involved in Cold Adaptation. *Frontiers in Genetics, 13*. https://doi.org/10.3389/fgene.2022.825269

Islam, S., & Mou, M. (2022). Functional Annotation of Uncharacterized Protein from Photobacterium damselae subsp. piscicida (Pasteurella piscicida) and Comparison of Drug Target Between Conventional Medicine and Phytochemical Compound Against Disease Treatment in Fish: An In-silico Approach. *Genetics of Aquatic Organisms, 6*, 453. https://doi.org/10.4194/GA453

Islam, S., Mou, M., Sanjida, S., & Mahfuj, M.s.E. (2022a). Functional Annotation and Characterization of a Hypothetical Protein from Pseudoalteromonas spp. Identify Potential Biomarker: An In-silico Approach. *Aquatic Food Studies, 2*, 57. https://doi.org/10.4194/AFS57

Islam, S., Mou, M., Sanjida, S., & Mahfuj, M.s.E. (2022b). An In-silico Approach for Identifying Phytochemical Inhibitors Against Nervous Necrosis Virus (NNV) in Asian Sea Bass by Targeting Capsid Protein. *Genetics of Aquatic Organisms, 6*, 487. https://doi.org/10.4194/GA487

Islam, S., Sanjida, S., Mahfuj, M.s.E., Islam, M.J., & Mou, M. (2022). Computer-aided drug design of Azadirachta indica compounds against nervous necrosis virus by targeting grouper heat shock cognate protein 70 (GHSC70): quantum mechanics calculations and molecular dynamic simulation approaches. *Genomics & Informatics, 20*. https://doi.org/10.5808/gi.21063

Islam, S., Sanjida, S., Mou, M., Mahfuj, M.s.E., & Nasir, S. (2022). In-silico functional annotation of a hypothetical protein from Edwardsiella tarda revealed Proline metabolism and apoptosis in fish. *International Journal of Life Sciences and Biotechnology, 5*, 78-96. https://doi.org/10.38001/ijlsb.1032171

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T.L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research, 36*(suppl_2), W5-W9. https://doi.org/10.1093/nar/gkn201

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., ... Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240. https://doi.org/10.1093/bioinformatics/btu031

Källberg, M., Margaryan, G., Wang, S., Ma, J., & Xu, J. (2014). RaptorX server: a resource for template-based protein structure modeling. *Methods in molecular biology (Clifton, N.J.), 1137*, 17-27. https://doi.org/10.1007/978-1-4939-0366-5_2

Kayansamruaj, P., Soontara, C., Unajak, S., Dong, H.T., Rodkhum, C., Kondo, H., ... Areechon, N. (2019). Comparative genomics inferred two distinct populations of piscine pathogenic Streptococcus agalactiae, serotype Ia ST7 and serotype III ST283, in Thailand and Vietnam. *Genomics, 111*(6), 1657-1667.

Khang, P.V., Phuong, T.H., Dat, N.K., Knibb, W., & Nguyen, N.H. (2018). An 8-Year Breeding Program for Asian Seabass Lates calcarifer: Genetic Evaluation, Experiences, and Challenges. *Frontiers in Genetics, 9*. https://doi.org/10.3389/fgene.2018.00191

Ko, J., Park, H., & Seok, C. (2012). GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics, 13*, 198. https://doi.org/10.1186/1471-2105-13-198

Kumar, A., Maan, P., Singh, G., & Kaur, J. (2017). In-Silico Characterization of a Hypothetical Protein, Rv1288 of Mycobacterium tuberculosis Containing an Esterase Signature and an Uncommon LytE Domain. *Curr Comput Aided Drug Des, 13*(2), 101-111. https://doi.org/10.2174/1573409912666161124144725

Kwankijudomkul, A., Dong, H.T., Longyant, S., Sithigorngul, P., Khunrae, P., Rattanarojpong, T., & Senapin, S. (2021). Antigenicity of hypothetical protein HP33 of Vibrio harveyi Y6 causing scale drop and muscle necrosis disease in Asian sea bass. *Fish Shellfish Immunol, 108*, 73-79. https://doi.org/10.1016/j.fsi.2020.11.034

Letunic, I., Doerks, T., & Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res, 40*(Database issue), D302-305. https://doi.org/10.1093/nar/gkr931

Liu, B., Zheng, D., Jin, Q., Chen, L., & Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res, 47*(D1), D687-d692. https://doi.org/10.1093/nar/gky1080

Lubec, G., Afjehi-Sadat, L., Yang, J.W., & John, J.P. (2005). Searching for hypothetical proteins: theory and practice based upon original data and literature. *Prog Neurobiol, 77*(1-2), 90-127. https://doi.org/10.1016/j.pneurobio.2005.10.001

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., ... Bryant, S.H. (2005). CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Research, 33*(suppl_1), D192-D196. https://doi.org/10.1093/nar/gki069

Minion, F.C., Lefkowitz, E.J., Madsen, M.L., Cleary, B.J., Swartzell, S.M., & Mahairas, G.G. (2004). The genome sequence of Mycoplasma hyopneumoniae strain 232, the agent of swine mycoplasmosis. *J Bacteriol, 186*(21), 7123-7133. https://doi.org/10.1128/jb.186.21.7123-7133.2004

Möller, S., Croning, M.D.R., & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics, 17*(7), 646-653. https://doi.org/10.1093/bioinformatics/17.7.646

Montgomerie, S., Sundararaj, S., Gallin, W.J., & Wishart, D.S.

(2006). Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC bioinformatics, 7*(1), 301. https://doi.org/10.1186/1471-2105-7-301

Mou, M., Islam, S., & Mahfuj, M.s.E. (2021). In Silico Functional Annotation of VP 128 Hypothetical Protein from Vibrio parahaemolyticus. https://doi.org/10.4194/AFS37

Naveed, M., Kazmi, K., Anwar, F., Arshad, F., Dar, T., & Zafar, M. (2016). Computational Analysis and Polymorphism study of Tumor Suppressor Candidate Gene-3 for Non Syndromic Autosomal Recessive Mental Retardation. *Journal of Applied Bioinformatics & Computational Biology, 5*. https://doi.org/10.4172/2329-9533.1000127

Nehete, J.Y., Bhambar, R.S., Narkhede, M.R., & Gawali, S.R. (2013). Natural proteins: Sources, isolation, characterization and applications. *Pharmacognosy reviews, 7*(14), 107-116. https://doi.org/10.4103/0973-7847.120508

Pace, C.N., Grimsley, G.R., & Scholtz, J.M. (2009). Protein ionizable groups: pK values and their contribution to protein stability and solubility. *The Journal of biological chemistry, 284*(20), 13285-13289. https://doi.org/10.1074/jbc.R800080200

Saha, S., & Raghava, G.P. (2006). VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics, 4*(1), 42-47. https://doi.org/10.1016/s1672-0229(06)60015-6

Sirikharin, R., Taengchaiyaphum, S., Sanguanrut, P., Chi, T.D., Mavichak, R., Proespraiwong, P., ... Sritunyalucksana, K. (2015). Characterization and PCR Detection of Binary, Pir-Like Toxins from Vibrio parahaemolyticus Isolates that Cause Acute Hepatopancreatic Necrosis Disease (AHPND) in Shrimp. *PLoS One, 10*(5), e0126987. https://doi.org/10.1371/journal.pone.0126987

Sivashankari, S., & Shanmughavel, P. (2006). Functional annotation of hypothetical proteins - A review. *Bioinformation, 1*(8), 335-338. https://doi.org/10.6026/97320630001335

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res, 43*(Database issue), D447-452. https://doi.org/10.1093/nar/gku1003

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

Tian, W., Chen, C., Lei, X., Zhao, J., & Liang, J. (2018). CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res, 46*(W1), W363-w367. https://doi.org/10.1093/nar/gky473

Turab Naqvi, A.A., Rahman, S., Rubi, Zeya, F., Kumar, K., Choudhary, H., ... Hassan, M.I. (2017). Genome analysis of Chlamydia trachomatis for functional characterization of hypothetical proteins to discover novel drug targets. *Int J Biol Macromol, 96*, 234-240. https://doi.org/10.1016/j.ijbiomac.2016.12.045

Tusnády, G.E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics, 17*(9), 849-850. https://doi.org/10.1093/bioinformatics/17.9.849

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res, 31*(1), 258-261. https://doi.org/10.1093/nar/gkg034

Wang, J., Sung, W.K., Krishnan, A., & Li, K.B. (2005). Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinformatics, 6*, 174. https://doi.org/10.1186/1471-2105-6-174

Wang, Y.M., Khederzadeh, S., Li, S.R., Otecko, N.O., Irwin, D.M., Thakur, M., ... Zhang, Y.P. (2021). Integrating Genomic and Transcriptomic Data to Reveal Genetic Mechanisms Underlying Piao Chicken Rumpless Trait. *Genomics Proteomics Bioinformatics, 19*(5), 787-799. https://doi.org/10.1016/j.gpb.2020.06.019

Wenzel, S.C., Hoffmann, H., Zhang, J., Debussche, L., Haag-Richter, S., Kurz, M., ... Brönstrup, M. (2015). Production of the Bengamide Class of Marine Natural Products in Myxobacteria: Biosynthesis and Structure-Activity Relationships. *Angew Chem Int Ed Engl, 54*(51), 15560-15564. https://doi.org/10.1002/anie.201508277

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.-C., Williams, K., Appel, R., & Hochstrasser, D.F. (1999). Protein Identification and Analysis Tools in the ExPASy Server. *Methods in molecular biology (Clifton, N.J.), 112*, 531-552. https://doi.org/10.1385/1.59259-584-7:531

Wilson, D., Madera, M., Vogel, C., Chothia, C., & Gough, J. (2006). The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Research, 35*(suppl_1), D308-D313. https://doi.org/10.1093/nar/gkl910

Xu, J., McPartlon, M., & Li, J. (2021). Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat Mach Intell, 3*, 601-609. https://doi.org/10.1038/s42256-021-00348-5

Yu, C., & Hwang, J. (2008, 26-28 Nov. 2008). *Prediction of Protein Subcellular Localizations.* Paper presented at the 2008 Eighth International Conference on Intelligent Systems Design and Applications.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., ... 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. https://doi.org/10.1093/bioinformatics/btq249

Zhang, J., Jia, H., Li, J., Li, Y., Lu, M., & Hu, J. (2016). Molecular evolution and expression divergence of the Populus euphratica Hsf genes provide insight into the stress acclimation of desert poplar. *Scientific Reports, 6*(1), 30050. https://doi.org/10.1038/srep30050

Zhang, L., Chou, C., & Moo-Young, M. (2011). Disulfide bond formation and its impact on the biological activity and stability of recombinant therapeutic proteins produced by Escherichia coli expression system. *Biotechnology advances, 29*, 923-929. https://doi.org/10.1016/j.biotechadv.2011.07.013

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res, 33*(7), 2302-2309. https://doi.org/10.1093/nar/gki524

Supplementary Table 1. String Server result of Protein-protein interaction of HP33 protein

node1	node2	node1 annotation	node2 annotation	score
BAOA01000003_gene3725	BAOA01000003_gene3726	*annotation not available*	*annotation not available*	0.516
BAOA01000003_gene3725	BAOA01000049_gene1318	*annotation not available*	*annotation not available*	0.690
BAOA01000003_gene3725	BAOA01000049_gene1319	*annotation not available*	*annotation not available*	0.750
BAOA01000003_gene3725	BAOA01000049_gene1322	*annotation not available*	*annotation not available*	0.691
BAOA01000003_gene3725	BAOA01000049_gene1323	*annotation not available*	*annotation not available*	0.762
BAOA01000003_gene3725	BAOA01000074_gene3453	*annotation not available*	*annotation not available*	0.637
BAOA01000003_gene3725	BAOA01000092_gene3527	*annotation not available*	*annotation not available*	0.689
BAOA01000003_gene3725	BAOA01000092_gene3529	*annotation not available*	*annotation not available*	0.704
BAOA01000003_gene3725	BAOA01000105_gene4706	*annotation not available*	*annotation not available*	0.757
BAOA01000003_gene3725	BAOA01000126_gene1148	*annotation not available*	*annotation not available*	0.784
BAOA01000003_gene3725	BAOA01000157_gene4749	*annotation not available*	*annotation not available*	0.724
BAOA01000003_gene3726	BAOA01000003_gene3725	*annotation not available*	*annotation not available*	0.516
BAOA01000049_gene1318	BAOA01000003_gene3725	*annotation not available*	*annotation not available*	0.690
BAOA01000049_gene1318	BAOA01000049_gene1319	*annotation not available*	*annotation not available*	0.909
BAOA01000049_gene1318	BAOA01000049_gene1323	*annotation not available*	*annotation not available*	0.826
BAOA01000049_gene1318	BAOA01000092_gene3527	*annotation not available*	*annotation not available*	0.448
BAOA01000049_gene1318	BAOA01000092_gene3529	*annotation not available*	*annotation not available*	0.457
BAOA01000049_gene1318	BAOA01000105_gene4706	*annotation not available*	*annotation not available*	0.443
BAOA01000049_gene1318	BAOA01000126_gene1148	*annotation not available*	*annotation not available*	0.706
BAOA01000049_gene1318	BAOA01000157_gene4749	*annotation not available*	*annotation not available*	0.485