FENI UNIVERSITY JOURNAL, 2024, 3(1), PP. (1-30)
DOI: https://doi.org/10.3329/fuj.v3i1.86546
ISSN: 2518-3869

Structural Characterization and Functional Annotation of a Hypothetical Protein from *Salmonella Bongori*: An *In-Silico* Investigation

Md. Easin Mia[1]*, Partha Singha[2], Farzana Akter[3] and Nur Alam[4]

Abstract

Salmonella bongori, a gram-negative, rod-shaped bacterium, is responsible for causing salmonellosis, a gastrointestinal illness marked by symptoms such as sudden fever, nausea, vomiting, cramping diarrhea, and abdominal discomfort. Identifying the relevant protein could potentially facilitate the development of effective treatments for S. bongori infection. Currently, many proteins in S. bongori remain unidentified and are called hypothetical proteins (HPs). This study aimed to elucidate the structure and function of an uncharacterized HP (accession no. QXY84013.1) from S. bongori. Analysis of subcellular localization and various physicochemical properties suggested that this protein is cytoplasmic and exhibits stability. NCBI-CD Search, the functional annotation software, indicated that our selected protein would be categorized as a constituent of the formate-dependent nitrite reductase complex, known explicitly as NrfG. NrfG, composed of tetratricopeptide repeat (TPR) proteins, plays a pivotal role in bacterial virulence, aiding in virulence factor transfer into host cells and phagolysosome maturation inhibition. Additionally, it plays a crucial role in developing the heme lyase complex (NrfEFG), potentially impacting bacterial iron acquisition and pathogenesis, thus influencing the severity of human bacterial infections. The predominant secondary structure observed was the alpha helix. Utilizing homology modeling via the SWISS-MODEL server, the protein's three-dimensional (3D) structure was determined, employing a template protein with 100% sequence similarity (PDB ID: A0A5U3DZQ4.1.A). Several quality assessment tools including ERRAT, QMEAN, and PROCHECK were used to verify the 3D structure. Furthermore, the modeled structure's active site was predicted using the CASTp server. These findings hold promise as a potential foundation for the development of future antibacterial treatments.

1. *Corresponding author, Lecturer, Department of Biotechnology and Genetic Engineering, Noakhali Science and Technology University, Bangladesh. E-mail : easin.bge@nstu.edu.bd, Mobile : +8801751-46946

2. Student, B.Sc. in Biotechnology and Genetic Engineering Department, Noakhali Science and Technology University, Bangladesh. E-mail : partha1313@student.nstu.edu.bd

3. Lecturer, Department of Biotechnology and Genetic Engineering, Noakhali Science and Technology University, Bangladesh. E-mail : farzana.bge@nstu.edu.bd

4. Student, B.Sc. in Biotechnology and Genetic Engineering Department, Noakhali Science and Technology University, Bangladesh. E-mail : nuralam1313@student.nstu.edu.bd

Keywords: *Salmonella bongori; Hypothetical protein; In silico characterization; Three-dimensional structure.*

Introduction

Next-generation sequencing (NGS) permits researchers to collect enormous volumes of information quickly, but as more species are sequenced, it becomes increasingly difficult to assign genetic functions [1-2]. Over thirty percent of the proteins in many species are classified as "Hypothetical Proteins (HPs)" because they have unknown molecular functions [3]. Three-dimensional (3D) structures, novel domains, motifs, pathways, protein networks, and other important information are revealed by the *in-silico* characterization of HPs [4-6]. Moreover, functional and structural characterization of HPs can provide information on possible biomarkers and therapeutic targets [7]. The functions of HPs in several disease-causing bacteria have been successfully identified utilizing various computational programs and methods [8-11].

Salmonella, a prominent member of the Enterobacteriaceae family, is notorious for causing foodborne illnesses [12-13]. The *Salmonella* genus is divided into two harmful species, *Salmonella enterica,* and *Salmonella bongori*, according to variations in its 16S rRNA sequence [14]. One of these is *S. bongori*, a gram-negative, rod-like pathogenic organism that triggers salmonellosis, a gastrointestinal ailment characterized by cramps and diarrhea [15]. It was first identified in a 1-year-old kid with symptoms in Piedmont, northwest Italy [16]. *S. bongori* infections can produce fever, nausea, vomiting, diarrhea, stomach pain, and severe enteritis, among other symptoms [17]. The bacterium significantly contributes to foodborne outbreaks and has caused numerous human and animal diseases globally.

The computational analysis of HP from *S. bongori* is essential since knowing the genome of this organism may help build effective medications or vaccinations. This work focuses on a particular HP from *S. bongori*, designated by the accession number QXY84013.1. A detailed structural and functional investigation is carried out using various bioinformatics tools. A detailed list of all the software and tools utilized for the annotation of the functions of HP from *S. bongori* is given in Table 1.

Function	Tools/Server	URL
Sequence retrieval	NCBI	https://www.ncbi.nlm.nih.gov/
Determination of physiochemical properties	ExPASy ProtParam	https://web.expasy.org/protparam/
Subcellular localization determination	PSORT$_b$	https://www.psort.org/psortb/
	SOSUI server	https://harrier.nagahama-i-bio.ac.jp/sosui/
Functional annotation	Conserved Domain Database	https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
	MOTIF	https://www.genome.jp/tools/motif/
Sequence similarity search	BLASTp	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Prediction of secondary structure	PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/
	SOPMA	https://npsa.lyon.inserm.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html
Prediction of 3D structure	SWISS-MODEL	http://swissmodel.expasy.org/
Quality evaluation of 3D structure	PROCHECK, Verify3D & ERRAT	https://saves.mbi.ucla.edu/
	QMEAN4	https://swissmodel.expasy.org/qmean/
	ProSA-web (Z scores)	https://prosa.services.came.sbg.ac.at/prosa.php
Active site identification	CASTp	http://sts.bioe.uic.edu/castp/

Table 1: The bioinformatics tools used to characterize the hypothetical protein.

Materials and Methods

Sequence Retrieval

The amino acid sequence of an HP (accession no. QWW22972.1) from *S. bongori* was obtained in FASTA format from the database of proteins of the National Center for Biotechnology Information (NCBI) [18]. The corresponding protein sequence was then transmitted to multiple anticipation servers for analysis to perform *in-silico* annotation.

The NCBI-provided FASTA format is:

>QXY84013.1 hypothetical protein EWI73_08655 [*Salmonella bongori*]

MTLPANVVASRRPMPVKSLTAAAVLLGEYYLYRNAYDNALLAYRQALHLRG
DNAQLFAALATVLYYQAGQHMTPATREMIDKALALDAMEVTAKMLLAADA
FMQADYARAVSLWQALLDANSPRVNRAQLVEAINMAMLLQNRKK

Analysis of Physicochemical Characteristics

The physicochemical characteristics of the hypothetical protein (HP) were examined using the ExPASy ProtParam tool. These properties included amino acid composition, molecular weight, theoretical isoelectric point (pI), total number of positive and negative residues, instability index, estimated half-life, and aliphatic index (AI) [19].

Prediction of Subcellular Localization

Determining subcellular localization is crucial to comprehensively analyzing the genome and elucidating protein function. The hypothetical protein (HP) for the subcellular location was estimated utilizing both the PSORTb3.0 [20] and SOSUIgramN Server [21].

Functional Annotation

Proteins, complex molecules with diverse functions in living organisms, are categorized into various families and superfamilies according to similarities in their domains, motifs, sequence features, and functional characteristics [22]. The NCBI Conserved Domain Search Service (CD Search) was employed to identify potential functions and conserved domains of the target protein [23].

Multiple Sequences Alignment and Analysis of Phylogenetic Tree

The non-redundant database on the NCBI website was searched using BLASTp with the default parameters to find protein homologs [24]. Multiple sequence alignments and a tree of phylogenetic links were produced using Clustal Omega 2.6.1 [25].

Secondary Structure Prediction

Through the use of the self-optimized prediction method with alignment (SOPMA) [26], the two-dimensional structure of the HP was anticipated. PSIPRED [27] was employed as an additional method to confirm the SOPMA results.

Tertiary Structure Prediction

The selected protein's three-dimensional structure was determined using a SWISS-MODEL server employing homology modeling techniques [28]. The server performs a BLASTp search on every protein sequence to find appropriate templates. Among the search outcomes, the protein A0A5U3DZQ4.1 was chosen for homology modeling as it shares 100% sequence similarity. This selected template, depicting the X-ray diffraction model of a Heme lyase NrfEFG subunit NrfG protein, served as a dependable basis for the modeling process. The visualization of the resulting 3D model structure was achieved using PyMOL v2.0.

Quality Assessment

The SAVES servers PROCHECK [29] and ERRAT [30] modules were utilized to validate the anticipated 3D structure. In addition, the QMEAN Z-score and model quality assessment were performed using the QMEAN programs [31] accessible via the ExPASy server of the SWISS-MODEL Workspace. The ProSA-web server was utilized to estimate the Z-scores for the target protein [32].

Active Ste Determination

The protein’s active region was determined using the Computed Atlas of Surface Topography of Proteins (CASTp) server. CASTp provides an accurate, exhaustive, and numerical description of a protein’s topographical characteristics. Active pockets can be specifically identified and quantified on the surfaces of proteins and the core of three-dimensional structures. Additionally, the CASTp output was shown using PyMOL software [33].

Results

Analysis of Physicochemical Properties

The ProtParam program was utilized to determine various physicochemical characteristics of the hypothetical protein (accession no. QXY84013.1), as outlined in Table 2. The protein comprises 145 amino acids with a grand hydropathicity (GRAVY) average of 0.095. It has a molecular weight (MW) of 16073.77 and a theoretical isoelectric point (pI) of 9.47. A total of 11 negatively charged residues (Asp + Glu) and 15 positively charged residues (Arg + Lys) were identified. The target protein has a predicted half-life (HL) of 30 hours and an instability index (II) of 35.43. Moreover, the protein demonstrates stability across a wide temperature range, with an aliphatic index (AI) of 103.24.

Characteristics	Value
Amino acids number	145
Molecular weight (Da)	16073.77
Theoretical pI	9.47
Instability index	35.43
Positively charged residues	15
Negatively charged residues	11
Grand average of hydropathicity (GRAVY)	0.095
Aliphatic index	103.24

Table 2. Estimation of physiochemical characteristics by ProtParam tool.

Subcellular Localization Prediction

Determining the subcellular localization of a hypothetical protein is vital for understanding its function, as different cellular locations correspond to distinct activities. This knowledge can also inform the development of drugs against the target protein. In our case, PSORTb and the SOSUIgramN server determined the target protein to be cytoplasmic.

Function Annotation

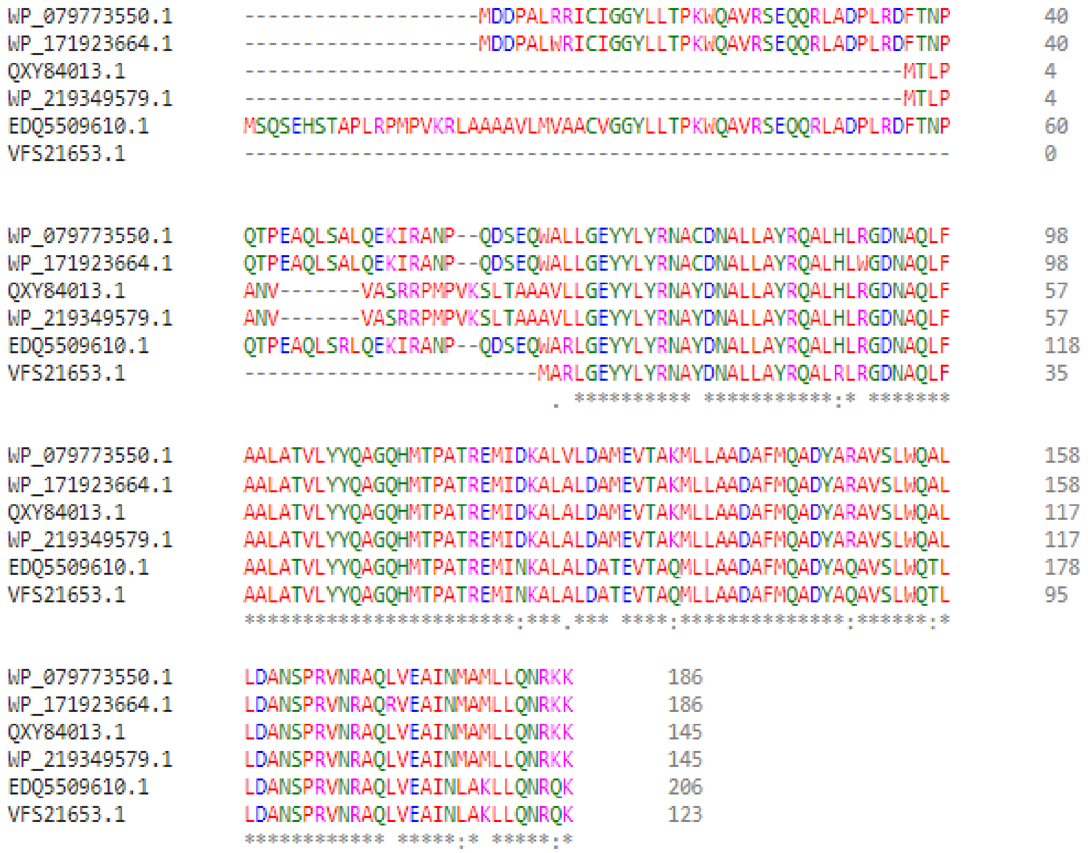
We delineated conserved domains and possible functionalities of our desired protein through diverse annotation methods. We used the NCBI CD Search tool to identify conserved domains within the protein sequence. This process entailed comparing the query sequence with position-specific score matrices derived from alignments of conserved domains. Our analysis revealed that the target protein harbors a domain affiliated with the PRK10370 superfamily. This domain was identified as the formate-dependent nitrite reductase complex component NrfG, as predicted by the NCBI-CD Search. Nitrite reductase facilitates the conversion of nitrite to ammonia in anaerobic conditions, which is crucial for microbial metabolism where oxygen is scarce. NrfG is composed of tetratricopeptide repeat (TPR) proteins, which are ubiquitous in various organisms. Within bacterial pathogens, TPR-containing proteins play a direct role in modulating functions associated with virulence. The NCBI-CD server estimated the location of the PRK10370 superfamily domain with amino acids to be 12-145 and an E-value of 1.47e-67.

Multiple Sequence Alignment and Analysis of Phylogenetic Tree

A BLASTp search against the non-redundant database identified equivalents with other known various Heme lyase NrfEFG component NrfG proteins from diverse bacteria (Table 3). After this, multiple sequence alignments were performed on a chosen subset of proteins obtained from BLASTp outcomes to scrutinize conserved and distinct residues among homologs (Figure 1). Furthermore, a phylogenetic tree was created utilizing a similar dataset (Figure 2). Interestingly, there seems to be a common ancestor between our target protein and *Salmonella bongori* (WP_219349579.1).

Accession no.	Name of organism	Name of protein	No. of score	Identity (%)	E-value
WP_219349579.1	*Salmonella bongori*	Hypothetical protein	286	100.00%	3e-97
WP_079773550.1	*Salmonella bongori*	Heme lyase NrfEFG subunit NrfG	236	98.35%	2e-76
WP_171923664.1	*Salmonella bongori*	Heme lyase NrfEFG subunit NrfG	231	97.52%	7e-75
EDQ5509610.1	*Salmonella enterica*	Heme lyase NrfEFG subunit NrfG	225	93.33%	4e-72
VFS21653.1	*Salmonella enterica* subsp.	Formate-dependent nitrite reductase complex subunit NrfG	221	92.50%	1e-71

Table 3. BLASTp result shows similarities between proteins.



```
WP_079773550.1      --------------------MDDPALRRICIGGYLLTPKWQAVRSEQQRLADPLRDFTNP      40
WP_171923664.1      --------------------MDDPALWRICIGGYLLTPKWQAVRSEQQRLADPLRDFTNP      40
QXY84013.1          --------------------------------------------------------MTLP      4
WP_219349579.1      --------------------------------------------------------MTLP      4
EDQ5509610.1        MSQSEHSTAPLRPMPVKRLAAAAVLMVAACVGGYLLTPKWQAVRSEQQRLADPLRDFTNP      60
VFS21653.1          ------------------------------------------------------------      0


WP_079773550.1      QTPEAQLSALQEKIRANP--QDSEQWALLGEYYLYRNACDNALLAYRQALHLRGDNAQLF      98
WP_171923664.1      QTPEAQLSALQEKIRANP--QDSEQWALLGEYYLYRNACDNALLAYRQALHLWGDNAQLF      98
QXY84013.1          ANV-------VASRRPMPVKSLTAAAVLLGEYYLYRNAYDNALLAYRQALHLRGDNAQLF      57
WP_219349579.1      ANV-------VASRRPMPVKSLTAAAVLLGEYYLYRNAYDNALLAYRQALHLRGDNAQLF      57
EDQ5509610.1        QTPEAQLSRLQEKIRANP--QDSEQWARLGEYYLYRNAYDNALLAYRQALHLRGDNAQLF      118
VFS21653.1          -------------------------MARLGEYYLYRNAYDNALLAYRQALRLRGDNAQLF      35
                                              . ********** ***********:* *******

WP_079773550.1      AALATVLYYQAGQHMTPATREMIDKALVLDAMEVTAKMLLAADAFMQADYARAVSLWQAL      158
WP_171923664.1      AALATVLYYQAGQHMTPATREMIDKALALDAMEVTAKMLLAADAFMQADYARAVSLWQAL      158
QXY84013.1          AALATVLYYQAGQHMTPATREMIDKALALDAMEVTAKMLLAADAFMQADYARAVSLWQAL      117
WP_219349579.1      AALATVLYYQAGQHMTPATREMIDKALALDAMEVTAKMLLAADAFMQADYARAVSLWQAL      117
EDQ5509610.1        AALATVLYYQAGQHMTPATREMINKALALDATEVTAQMLLAADAFMQADYAQAVSLWQTL      178
VFS21653.1          AALATVLYYQAGQHMTPATREMINKALALDATEVTAQMLLAADAFMQADYAQAVSLWQTL      95
                    ***********************:***.*** ****:**************:******:*

WP_079773550.1      LDANSPRVNRAQLVEAINMAMLLQNRKK       186
WP_171923664.1      LDANSPRVNRAQRVEAINMAMLLQNRKK       186
QXY84013.1          LDANSPRVNRAQLVEAINMAMLLQNRKK       145
WP_219349579.1      LDANSPRVNRAQLVEAINMAMLLQNRKK       145
EDQ5509610.1        LDANSPRVNRAQLVEAINLAKLLQNRQK       206
VFS21653.1          LDANSPRVNRAQLVEAINLAKLLQNRQK       123
                    ************ *****:* *****:*
```

Figure 1. Multiple sequence alignments among different heme lyase (NrfEFG) formate-dependent nitrite reductase complex subunit NrfG protein using Clustal Omega 2.6.1.

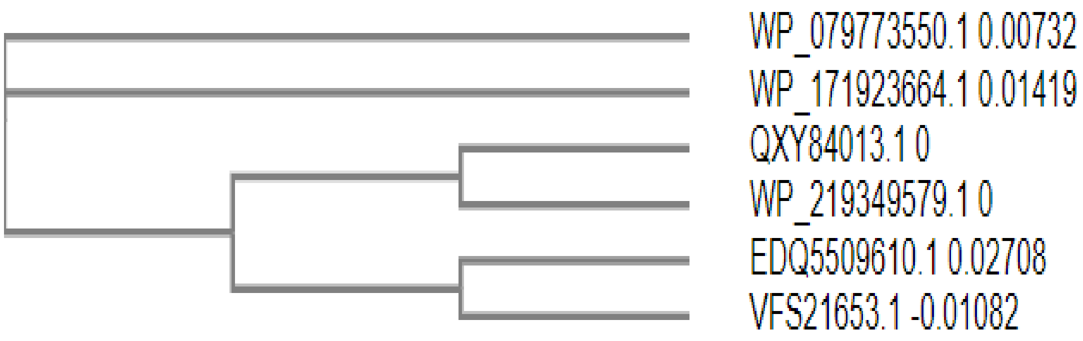


Figure 2. Phylogenetic tree depicting the evolutionary relationship between the target protein and other NrfG subunits of the heme lyase (NrfEFG) formate-dependent nitrite reductase complex.

Determination of Secondary Structure

The secondary structure of the protein was analyzed using PSIPRED and the SOPMA server.

As per the SOPMA estimation, the most prevalent secondary structure was identified to be the alpha helix, comprising 71.03% of the protein's structure, followed by the random coil at 23.45%, the extended strand at 2.07%, and the beta-turn at 3.45% (Figure 3A). The findings obtained from the PSIPRED server were utilized to validate the outcomes derived from SOPMA (Figure 3B).

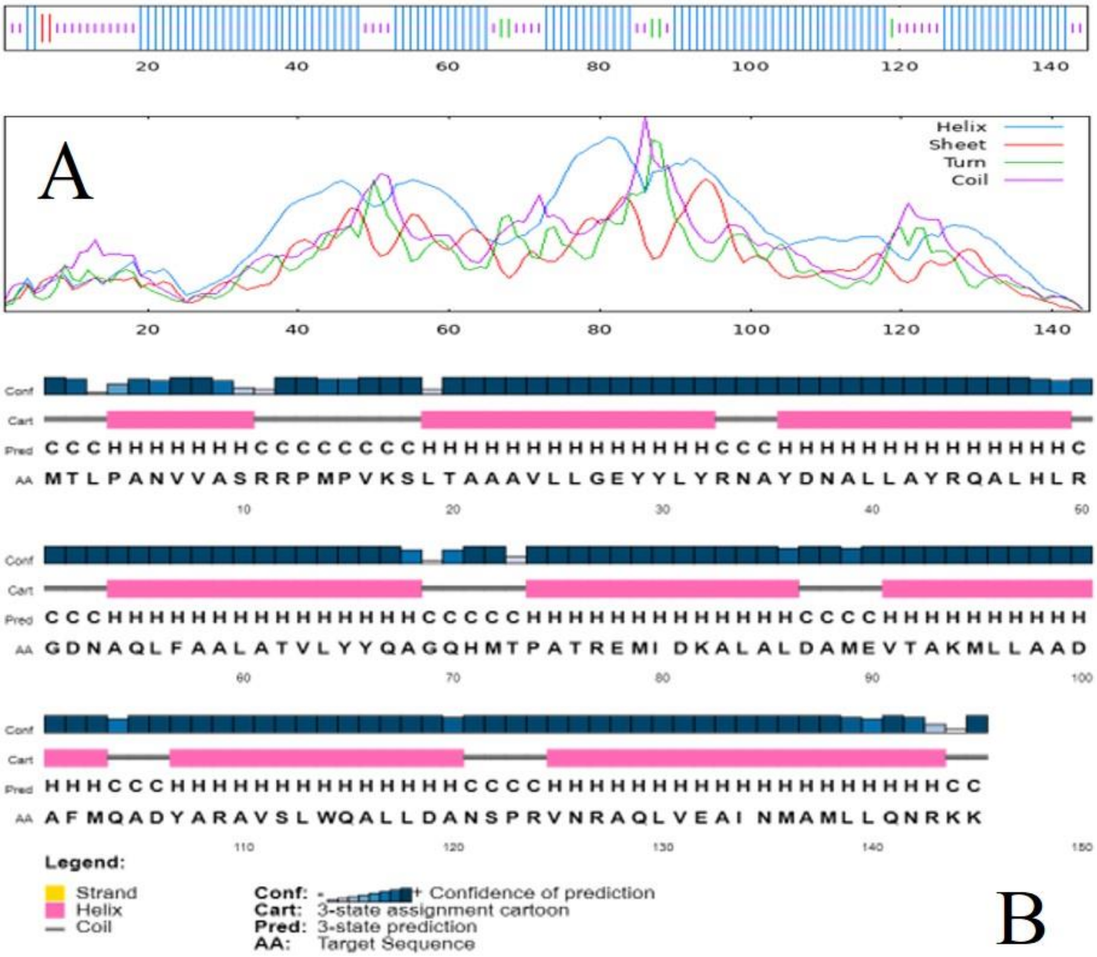


Figure 3. Projected secondary structure of the target protein: **(A)** SOPMA, **(B)** PSIPRED server.

Tertiary Structure Prediction

The tertiary structure of the target protein was established utilizing template A0A5U3DZQ4.1.A source from the SWISS-MODEL database exhibits complete sequence identity with the target protein. The 3D structure of the protein was visualized utilizing PyMOL 2.0 [Figure 4].

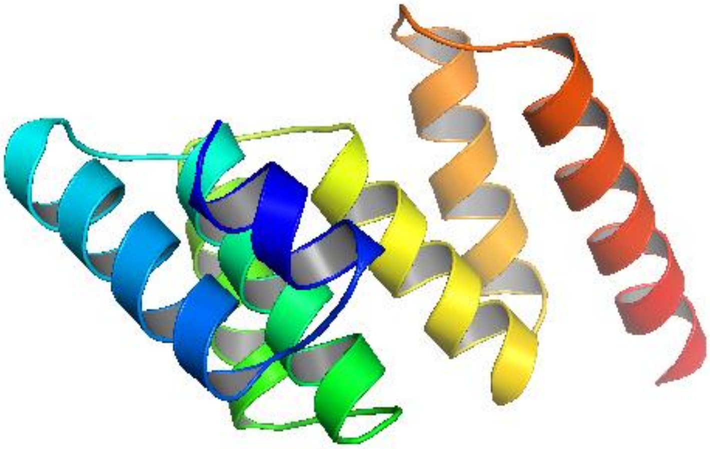


Figure 4. Anticipated 3D structure of the target protein through SWISS-MODEL (displayed by PyMOL software).

Model Quality Assessment

The three-dimensional structural model was evaluated using QMEAN, ERRAT, PROCHECK, and Verify 3D. PROCHECK study revealed that 93.9% of amino acid residues were located in the most favored zone of the “Ramachandran plot” (Table 4 and Figure 5). A high-caliber protein structure is indicated by the quality rating of 100 given by ERRAT (Figure 6). With a QMEAN4 score of 0.65, the QMEAN tool placed the model into the allowed black zone (Figure 7). The Z-scores for the model’s target protein were made available by the ProSA web server. A Z-score of -5.38 indicates that the input structure is within the typical range of scores for native proteins of comparable size (Figure 8).

Statistics	AA residues	Percentage (%)
Most favored regions [A, B, L]	107	93.9%
Additional allowed regions [a, b, l, p]	7	6.1%
Generously allowed regions [~a, ~b, ~l, ~p]	0.0	0.00%
Disallowed regions	0.0	0.00% Total= 100%
Non-glycine and non-proline number	114	
End-residues number (excl. Gly and Pro)	2	
Glycine residues	3	
Proline residues	2	
Total number of residues	121	

Table 4. Statistics of Ramachandran plots for the target protein.

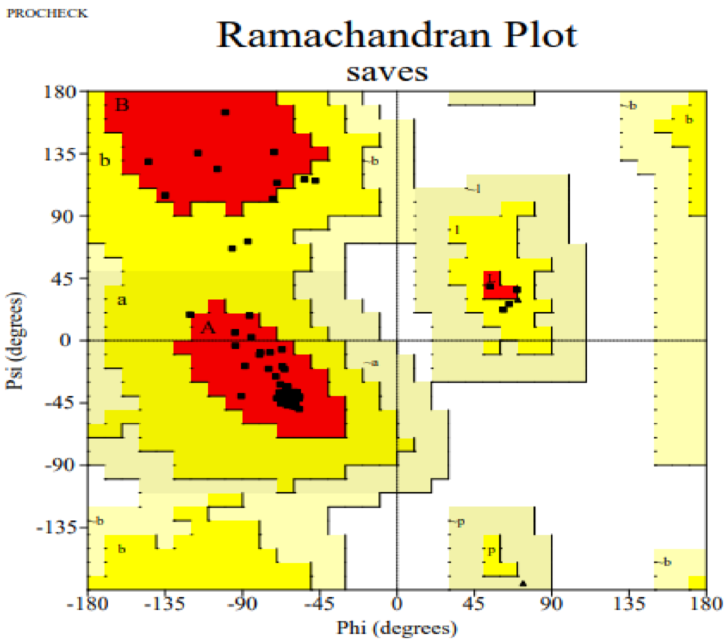


Figure 5. Validation of Ramachandran plot of model structure through PROCHECK program.

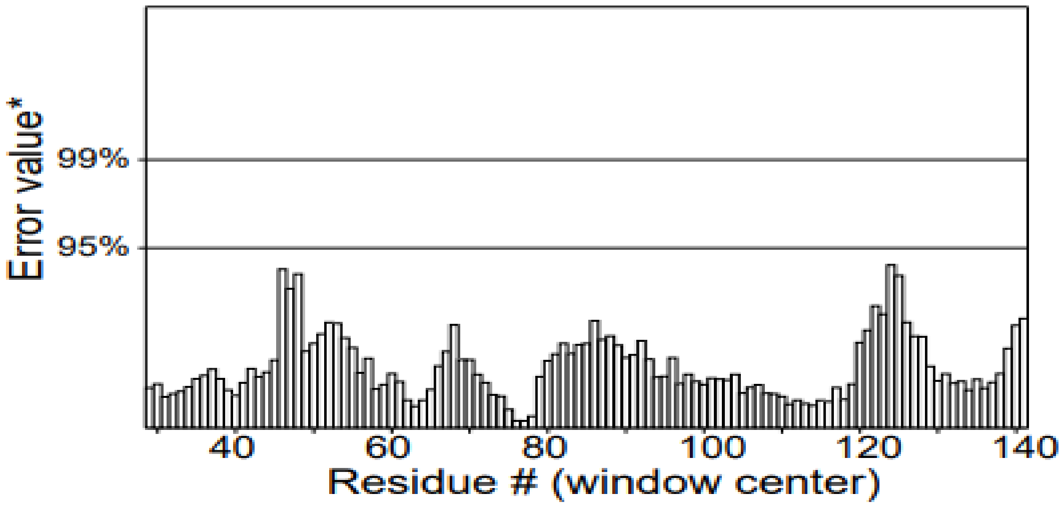


Figure 6. ERRAT result: The two lines on the error axis indicate the confidence level needed to reject regions exceeding the error value.

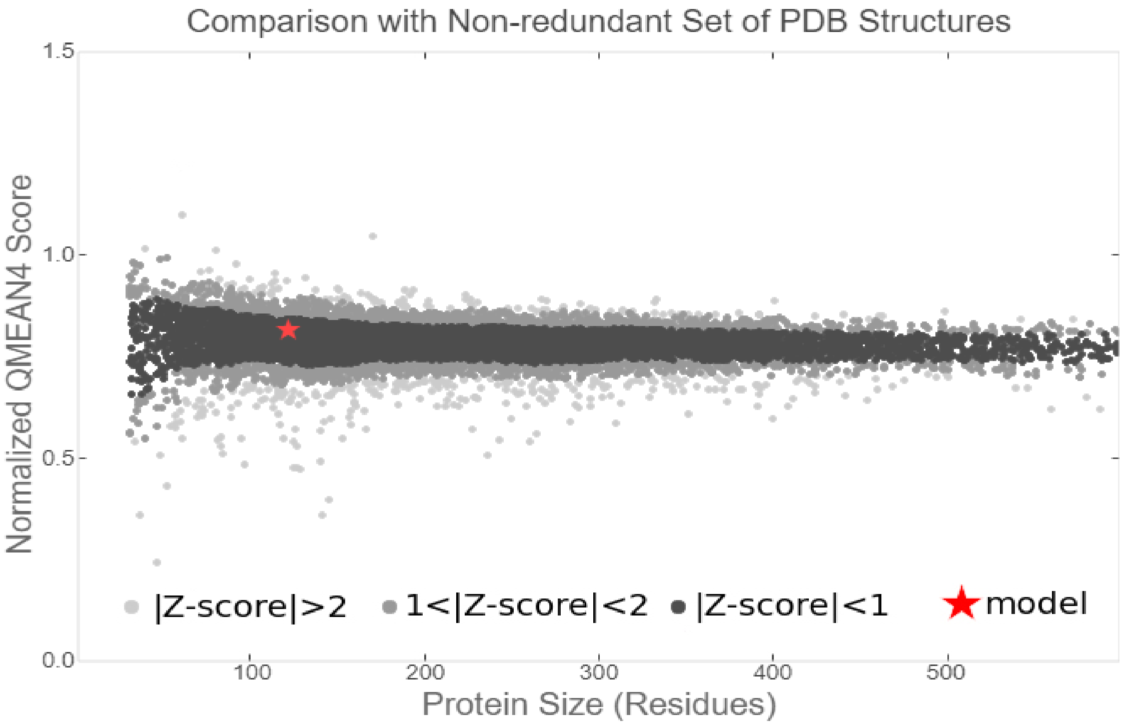


Figure 7. Visualization of QMEAN Results for the Model Structure.

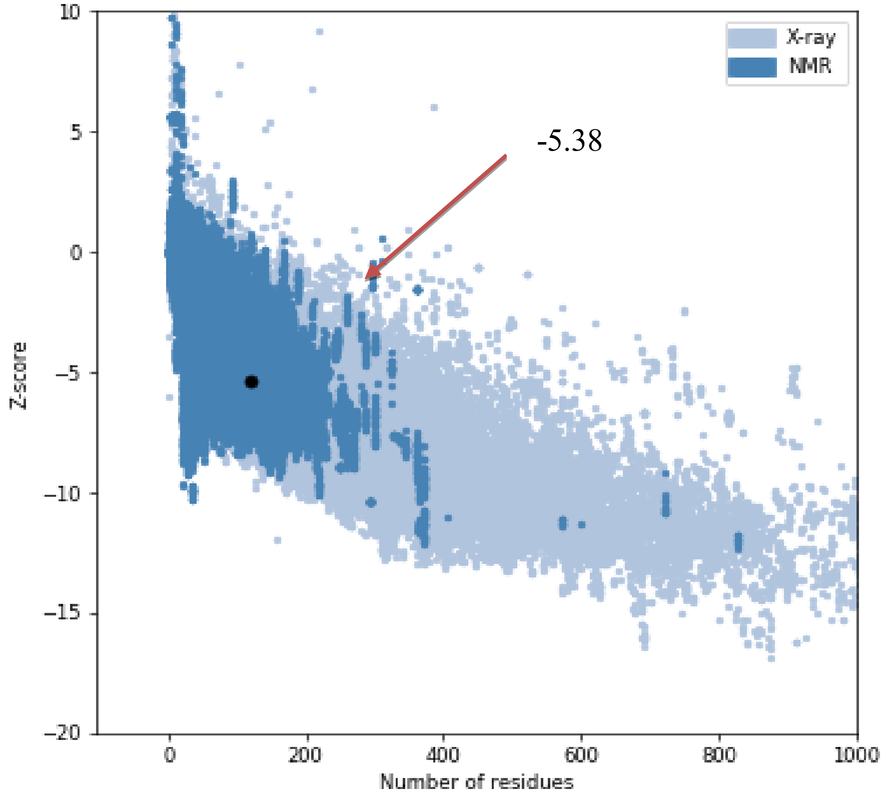


Figure 8. Z score of the target protein utilizing the ProSA server.

Active Site Determination

Utilizing the CASTp server, the active residues within the generated 3D model structure was examined, and its constituent amino acid residues were identified (Figure 9). A total of 21 amino acids were discovered to compose the putative active site of the protein. Among the identified pockets, one of the largest displayed a total volume of 425.778 amino acids and a solvent-accessible (SA) surface area of 295.614. Notably, amino acids GLU28, LEU31, TYR32, TYR43, GLN55, ALA58, ALA59, THR62, TYR65, TYR66, GLU90, VAL91, THR92, MET95, LEU96, ALA99, ARG124, VAL125, ASN126, GLY129, and LEU130 were identified as critical active residues within this pocket.

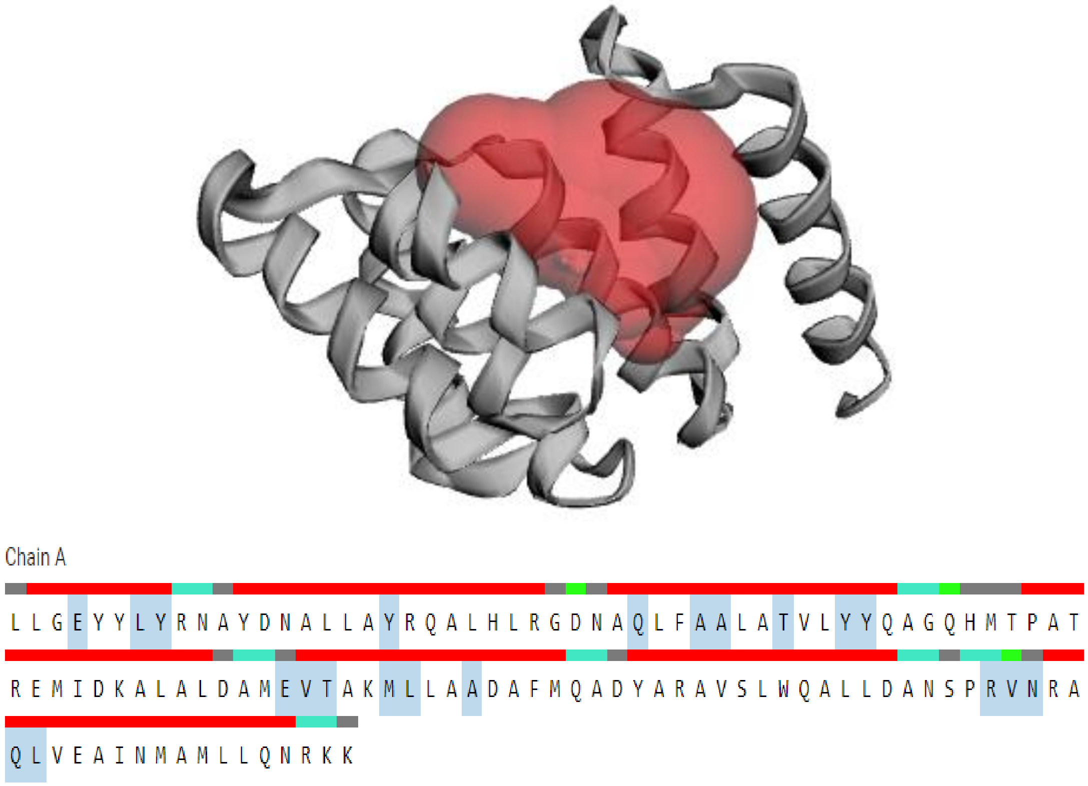

Chain A
L L G E Y Y L Y R N A Y D N A L L A Y R Q A L H L R G D N A Q L F A A L A T V L Y Y Q A G Q H M T P A T
R E M I D K A L A L D A M E V T A K M L L A A D A F M Q A D Y A R A V S L W Q A L L D A N S P R V N R A
Q L V E A I N M A M L L Q N R K K

Figure 9. Identification of the Active Site Utilizing the CASTp Server (Highlighted in Red). The most extensive active site was detected in regions spanning 295.614 and possessing a volume of 425.778 amino acids. Active amino acid residues are accentuated in the subsequent figure.

Discussion

Numerous ongoing studies focus on elucidating the structures and functions of hypothetical proteins (HPs), with some investigating their potential roles in various diseases [34]. Characterizing hypothetical proteins (HPs) can advance our knowledge of bacterial metabolic processes, facilitate medication development, and improve disease management strategies [35]. In this study, the HP protein (accession no. QXY84013.1) from the *S. bongori* strain underwent structural and functional characterization utilizing diverse computational tools. Its theoretical isoelectric point (pI) of 9.47, in particular, suggests that it is acidic ($pH > 7$), while a grand average of hydropathicity (GRAVY) value of 0.095 indicates hydrophobicity, implying insolubility in water.

Additionally, with an instability index (II) of 35.43, the HP was identified as a stable protein (Table 2). Predictions from the PSORTb and SOSUIgramN servers suggested cytoplasmic localization for this protein. The formate-dependent nitrite reductase complex member NrfG was identified as our target hypothetical protein by domain and motif analysis, and this prediction was made initially with high confidence using all annotation techniques (Table 3). NrfG, a 21 kDa protein of 198 amino acids, is one of the three subunits that form the heme lyase complex called NrfEFG. It is commonly understood that NrfG is composed of tetratricopeptide repeat (TPR) proteins, a class of proteins found across various life forms, including bacteria and humans [36]. In bacterial pathogens, proteins containing TPR motifs play a direct role in functions associated with virulence. These functions include transferring virulence factors into host cells, attaching to host cells, and inhibiting the maturation of phagolysosomes [37-38]. The significance of TPR motifs in the functionality of class II chaperones within the type III secretion system (TTSS) is underscored in the pathogenesis of *Yersinia*, *Pseudomonas*, and *Shigella*, underscoring their importance in bacterial virulence mechanisms [39]. NrfG, a bitopic inner membrane protein, is present in various *Escherichia coli* strains. It is essential for assembling the heme lyase complex (NrfEFG), enabling it to interact with the formate-dependent nitrite reductase [40-41]. Bacterial heme lyases potentially break down heme derived from hemoglobin, assisting in bacterial iron acquisition and contributing to pathogenesis. By facilitating iron uptake and adaptation to the host environment, heme lyases may influence the severity and outcome of bacterial infections in humans [42]. The BLASTp analysis conducted against the non-redundant database uncovered resemblances with other established heme lyase (NrfEFG) formate-dependent nitrite reductase complex subunit NrfG proteins, thereby affirming the initial prediction (Table 3).

Regarding secondary structure, the protein predominantly comprises alpha helix, complemented by elements including random coils, beta turns, and extended strands. The 3D structure of the target protein was acquired from the SWISS-MODEL server, employing template A0A5U3DZQ4.1.A, which exhibited complete sequence identity with the target protein (Figure 4). Comprehensive assessments using various model quality evaluation tools, including PROCHECK, QMEAN, and ERRAT, were conducted on the three-dimensional (3D) structure generated by the SWISS-MODEL server. Additionally, the model attained a QMEAN4 score of 0.65, positioning it within the desirable dark grey zone (Figure 7) compared to other experimental structures of similar size. Based on these findings, the predicted protein model demonstrates good quality and is suitable for further detailed investigation and analysis. The ProSA web server was employed to ascertain the Z-score, a measure used to evaluate the quality of the predicted protein model. This score is commonly used to assess whether the input structure lies within the range observed for native proteins of comparable size. The Z-score for the query model of the target protein was documented as -5.38 (Figure 8). A fundamental drug or inhibitor development aspect involves identifying and characterizing active site residues. In this investigation, the CASTp server was employed to explore the active site of the modeled protein structure and pinpoint the specific amino acid residues involved. Analysis by CASTp unveiled the existence of 12 active sites and their corresponding binding pockets within the protein (Figure 9). The optimal model was chosen based on pocket size, volume, and the presence of conserved residues. A total of 21 amino acids were found to constitute the putative active site of the protein. Remarkably, among the identified pockets, one of the largest, boasting a total volume of 425.778 cubic angstroms and a solvent-accessible (SA) surface area of 295.614 square angstroms, contained the most active site. Investigating these active site residues offers valuable insights for developing therapeutics or inhibitors targeting the specific protein. These results indicate the potential usefulness of the protein structure model for subsequent research and drug development endeavors.

Conclusions

This study utilized various bioinformatics methods to scrutinize a hypothetical protein originating from *Salmonella bongori.* The findings of our study have implications for improving the functionality of the target protein and optimizing resource utilization. Future investigations aim to validate our findings experimentally and devise novel ligands for medication development, leveraging structural and functional insights. Continued exploration of target proteins and their effectors in *S. bongori* and other species is essential for refining future treatment strategies. Additionally, this study contributes to understanding structural and functional studies of proteins with unidentified activities. The findings from this study could lay the groundwork for future *in-silico* research endeavors undertaken by other researchers.

Acknowledgments: We would like to thank the Department of Biotechnology and Genetic Engineering at Noakhali Science and Technology University in Noakhali, Bangladesh (3814), for their support in providing the resources that enabled our research efforts.

Funding: The author(s) did not receive any financial support for their research, writing, and/or publication of this work.

Declaration of Conflicting Interests: The author(s) have declared no potential conflicts of interest associated with this paper's research, writing, or publication.

Author Contributions: MEM conceived and designed the experiments, contributed to critical revisions, and approved the final manuscript; PS and NA conducted data analysis and drafted the initial manuscript; MEM, FA, and PS prepared the tables and figures; All authors strategically reviewed and confirmed the final version of the manuscript.

References

1. H. P. Choi, S. Juarez, S. Ciordia, M. Fernandez, R. Bargiela, J. P. Albar, V. Mazumdar, B. P. Anton, S. Kasif, M. Ferrer and M. Steffen, "Biochemical Characterization of Hypothetical Proteins from Helicobacter pylori." *PloS one*, **2013**, *8*(6), e66605. https://doi.org/10.1371/journal.pone.0066605

2. O. Morozova and M. A. Marra, "Applications of next-generation sequencing technologies in functional genomics." *Genomics*, **2008**, *92*(5), 255-264. https://doi.org/10.1016/j.ygeno.2008.07.001

3. M. Shahbaaz, K. Bisetty, F. Ahmad and M. I. Hassan, "Current Advances in the Identification and Characterization of Putative Drug and Vaccine Targets in the Bacterial Genomes." *Current topics in medicinal chemistry*, **2016**, *16*(9), 1040-1069. https://doi.org/10.2174/1568026615666150825143307

4. G. Nimrod, M. Schushan, D. M. Steinberg and N. Ben-Tal, "Detection of functionally important regions in "hypothetical proteins" of known structure." *Structure,* **2008**, *16*(12), 1755-1763. https://doi.org/10.1016/j.str.2008.10.017

5. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." *Proceedings of the National Academy of Sciences of the United States of America*, **1999**, *96*(8), 4285-4288. https://doi.org/10.1073/pnas.96.8.4285

6. S. Idrees, S. Nadeem, S. Kanwal, B. Ehsan, A. Yousaf, S. Nadeem and M. I. Rajoka, "In silico sequence analysis, homology modeling and function annotation of Ocimum basilicum hypothetical protein G1CT28_OCIBA." *International Journal Bioautomation*, **2012**, *16*(2), 111-118.

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

8. A. A. Turab Naqvi, S. Rahman, F. Rubi, Zeya, K. Kumar, H. Choudhary, M. S. Jamal, J. Kim and M. I. Hassan, “Genome analysis of Chlamydia trachomatis for functional characterization of hypothetical proteins to discover novel drug targets.” *International journal of biological macromolecules*, **2017**, *96*, 234-240. https://doi.org/10.1016/j.ijbiomac.2016.12.045

9. A. A. Naqvi, F. Anjum, F. I. Khan, A. Islam, F. Ahmad & M. I. Hassan, “Sequence Analysis of Hypothetical Proteins from *Helicobacter pylori* 26695 to Identify Potential Virulence Factors.” *Genomics & informatics*, **2016**, *14*(3), 125-135. https://doi.org/10.5808/GI.2016.14.3.125

10. Z. Yang, X. Zeng and S. K. Tsui, “Investigating function roles of hypothetical proteins encoded by the Mycobacterium tuberculosis H37Rv genome.” *BMC genomics*, **2019**, *20*(1), 394. https://doi.org/10.1186/s12864-019-5746-6

11. M. S. Islam, S. M. Shahik, M. Sohel, N. I. Patwary and M. A. Hasan, “In Silico Structural and Functional Annotation of Hypothetical Proteins of Vibrio cholerae O139.” *Genomics & informatics*, **2015**, *13*(2), 53–59. https://doi.org/10.5808/GI.2015.13.2.53

12. R. F. Doolittle, D. F. Feng, S. Tsang, G. Cho & E. Little, “Determining divergence times of the major kingdoms of living organisms with a protein clock.” *Science (New York, N.Y.)*, **1996**, *271*(5248), 470-477. https://doi.org/10.1126/science.271.5248.470

13. S. K. Eng, P. Pusparajah, N. S. Ab Mutalib, H. L. Ser, K.G. Chan and L. H. Lee, “Salmonella: A review on pathogenesis, epidemiology and antibiotic resistance.” *Frontiers in Life Science*, **2015**, *8*(3), 284-293. https://doi.org/10.1080/21553769.2015.1051243

14. M. W. Reeves, G. M. Evins, A. A. Heiba, B. D. Plikaytis and J. J. Farmer, “Clonal nature of Salmonella typhi and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of Salmonella bongori comb. nov.” *Journal of clinical microbiology*, **1989**, *27*(2), 313-320. https://doi.org/10.1128/jcm.27.2.313-320.1989

15. V. R. Bhagwat, “Safety of Water Used in Food Production.” *Food Safety and Human Health*, **2018**, 219-247. https://doi.org/10.1016/B978-0-12-816333-7.00009-6

16. A. Romano, A. Bellio, G. Macori, P. D. Cotter, D. M. Bianchi, S. Gallina and L. Decastelli, “Whole-Genome Shotgun Sequence of *Salmonella bongori*, First Isolated in Northwestern Italy.” *Genome announcements*, **2017**, *5*(27), e00560-17. https://doi.org/10.1128/genomeA.00560-17

17. D. L. Woodward, R. Khakhria and W. M. Johnson, “Human salmonellosis associated with exotic pets.” *Journal of clinical microbiology*, **1997**, *35*(11), 2786-2790. https://doi.org/10.1128/jcm.35.11.2786-2790.1997

18. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler, “GenBank.” *Nucleic acids research*, **2003**, *31*(1), 23-27. https://doi.org/10.1093/nar/gkg057

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

20. J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnády, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai and F. S. Brinkman, “PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria.” *Nucleic acids research*, **2003**, *31*(13), 3613-3617. https://doi.org/10.1093/nar/gkg602

21. K. Imai, N. Asakawa, T. Tsuji, F. Akazawa, A. Ino, M. Sonoyama and S. Mitaku, “SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in gram-negative bacteria.” *Bioinformation*, **2008**, *2*(9), 417-421. https://doi.org/10.6026/97320630002417

22. C. H. Wu, H. Huang, L. S. Yeh, and W. C. Barker, “Protein family classification and functional annotation.” *Computational biology and chemistry*, **2003**, *27*(1), 37-47. https://doi.org/10.1016/s1476-9271(02)00098-1

23. S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki and A. Marchler-Bauer, “CDD/SPARCLE: the conserved domain database in 2020.” *Nucleic acids research*, **2020**, *48*(D1), D265-D268. https://doi.org/10.1093/nar/gkz991

24. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, “Basic local alignment search tool.” *Journal of molecular biology*, **1990**, *215*(3), 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

25. P. Hogeweg and B. Hesper, “The alignment of sets of sequences and the construction of phyletic trees: an integrated method.” *Journal of molecular evolution*, **1984**, *20*(2), 175-186. https://doi.org/10.1007/BF02257378

26. C. Combet, C. Blanchet, C. Geourjon and G. Deléage, “NPS@: network protein sequence analysis.” *Trends in biochemical sciences*, **2000**, *25*(3), 147-150. https://doi.org/10.1016/s0968-0004(99)01540-6

27. D. W. A. Buchan and D. T. Jones, “The PSIPRED Protein Analysis Workbench: 20 years on.” *Nucleic acids research*, **2019**, *47*(1), 402-407. https://doi.org/10.1093/nar/gkz297

28. A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore and T. Schwede, “SWISS-MODEL: homology modelling of protein structures and complexes.” *Nucleic acids research*, **2018**, *46*(1), 296-303. https://doi.org/10.1093/nar/gky427

29. R. A. Laskowski, M. W. Macarthur, D. S. Moss and J. M. Thornton, “ProCheck: A Program to Check the Stereochemical Quality of Protein Structures.” *Journal of Applied Crystallography,* **1993**, *26*, 283-291. http://dx.doi.org/10.1107/S0021889892009944

30. C. Colovos and T. O. Yeates, “Verification of protein structures: patterns of nonbonded atomic interactions.” *Protein science : a publication of the Protein Society*, **1993**, *2*(9), 1511-1519. https://doi.org/10.1002/pro.5560020916

31. P. Benkert, M. Biasini and T. Schwede, “Toward the estimation of the absolute quality of individual protein structure models.” *Bioinformatics (Oxford, England)*, **2011**, *27*(3), 343-350. https://doi.org/10.1093/bioinformatics/btq662

32. M. Wiederstein and M. J. Sippl, “ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins.” *Nucleic acids research*, **2007**, *35*, 407-410. https://doi.org/10.1093/nar/gkm290

33. W. Tian, C. Chen, X. Lei, J. Zhao and J. Liang, "CASTp 3.0: computed atlas of surface topography of proteins." *Nucleic Acids Research*, **2018**, *46*(1), 363-367. https://doi.org/10.1093/nar/gky473

34. L. Barragan-Osorio, G. Giraldo, C. J. Almeciga-Diaz, G. Aliev, G. E. Barreto and J. Gonzalez, "Computational Analysis and Functional Prediction of Ubiquitin Hypothetical Protein: A Possible Target in Parkinson Disease." *Central nervous system agents in medicinal chemistry*, **2015**, *16*(1), 4–11. https://doi.org/10.2174/1871524915666150722120605

35. T. Sen and N. K. Verma, "Functional Annotation and Curation of Hypothetical Proteins Present in A Newly Emerged Serotype 1c of Shigella flexneri: Emphasis on Selecting Targets for Virulence and Vaccine Design Studies." *Genes*, **2020**, *11*(3), 340. https://doi.org/10.3390/genes11030340

36. L. D. D'Andrea and L. Regan, "TPR proteins: the versatile helix." *Trends in biochemical sciences*, **2003**, *28*(12), 655-662. https://doi.org/10.1016/j.tibs.2003.10.007

37. P. J. Edqvist, J. E. Bröms, H. J. Betts, A. Forsberg, M. J. Pallen and M. S. Francis, "Tetratricopeptide repeats in the type III secretion chaperone, LcrH: their role in substrate binding and secretion." *Molecular microbiology*, **2006**, *59*(1), 31-44. https://doi.org/10.1111/j.1365-2958.2005.04923.x

38. S. Chakraborty, M. Monfett, T. M. Maier, J. L. Benach, D. W. Frank & D. G. Thanassi, "Type IV pili in Francisella tularensis: roles of pilF and pilT in fiber assembly, host cell adherence, and virulence." *Infection and immunity*, **2008**, *76*(7), 2852-2861. https://doi.org/10.1128/IAI.01726-07

39. L. Cerveny, A. Straskova, V. Dankova, A. Hartlova, M. Ceckova, F. Staud and J. Stulik, "Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms." *Infection and immunity*, **2013**, *81*(3), 629-635. https://doi.org/10.1128/IAI.01035-12

40. D. Han, K. Kim, J. Oh, J. Park and Y. Kim, "TPR domain of NrfG mediates complex formation between heme lyase and formate-dependent nitrite reductase in Escherichia coli O157:H7." *Proteins*, **2008**, *70*(3), 900-914. https://doi.org/10.1002/prot.21597

41. A. L. Lomize, M. A. Lomize, S. R. Krolicki & I. DS. Pogozheva, "Membranome: a database for proteome-wide analysis of single-pass membrane proteins." *Nucleic acids research*, **2017**, *45*(1), 250-255. https://doi.org/10.1093/nar/gkw712

42. J. E. Choby and E. P. Skaar, “Heme Synthesis and Acquisition in Bacterial Pathogens.” *Journal of molecular biology*, **2016**, *428*(17), 3408-3428. https://doi.org/10.1016/j.jmb.2016.03.018