INNSPUB

International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print) 2222-5234 (Online)
http://www.innspub.net
Vol. 21, No. 6, p. 7-17, 2022

RESEARCH PAPER OPEN ACCESS

In silico identification of *Shigella sonnei* hypothetical protein RUK71877.1 as interleukin receptor mimic Protein A and a potential drug target

Foeaz Ahmed[*1,2], Nadim Ahmed[2], Anindita Ash Prome[2], Tanjin Barketullah Robin[2], Nurul Amin Rani[2]

[1]*Department of Molecular Biology and Genetic Engineering, Sylhet Agricultural University, Sylhet, Bangladesh*

[2]*Faculty of Biotechnology and Genetic Engineering, Sylhet Agricultural University, Sylhet, Bangladesh*

Key words: *Shigella sonnei,* Docking, Hypothetical Proteins (HPs), CELLO, NCBI-CD, Jalview

http://dx.doi.org/10.12692/ijb/21.6.7-17 Article published on December 01, 2022

Abstract

Shigella spp. is strict human pathogens that cause shigellosis (bloody diarrhea) and are linked to a significant amount of morbidity and mortality worldwide. *Shigella sonnei* causes 90% of shigellosis cases and most of them became resistant to traditional antibiotics. The bacterial genome has been discovered, but there are some proteins whose function is not known. This *in silico* study was conducted to characterize the hypothetical protein RUK71877.1 of *S. sonnei*. Different bioinformatics web tools were utilized such as BLASTp, ProtParam, CELLO, Jalview etc. to determine the likely function of the hypothetical sequences by searching Sequence Databases for orthologous enzymatic conserved domains. Molecular modeling, energy minimization and docking analysis was evaluated to further validate our findings. In the study our target hypothetical protein RUK71877.1 showed highly similarity with IrmA Family protein. The protein is found to be outer-membrane and has an important role in *Shigella sonnei* pathogenicity. In NCBI-CD search the target protein was found to have functioned as interleukin receptor mimic protein A which also showed higher affinity with IL-4R in docking analysis. *In silico* drug development for the treatment of Shigellosis may use these newly predicted hypothetical proteins as potential drug targets in the future. It can also be utilized as target protein in vaccine construction. Our thorough investigation will contribute to identifying a vast range of therapeutic targets and a better knowledge of how to build unique possible treatment strategies to combat the *Shigella* infection.

[*] **Corresponding Author:** Foeaz Ahmed ✉ foeaz.mge@sau.ac.bd

Introduction

Hypothetical proteins (HPs), are proteins which have unknown function and can be predicted from solely nucleic acid sequences, make up a sizable percentage of the mammalian proteomes including many other species (Lubec *et al.*, 2005). Researchers can rapidly capture massive amounts of data by using next-generation sequencing (NGS) and also predict the sequences of a protein. More than 30% of proteins in many species have unknown molecular activities (Rabbi *et al.*, 2021) they can be predicted by computational analysis for the advancement of research activities. In order to predict protein function, scientists have devised a number of strategies with the help of various computational techniques. Sequence similarity, phylogenetic analysis, protein-protein interactions, protein-ligand interactions, similarity of active site residues, conserved domains, motifs, phosphorylation sites, and gene expression profiles have all contributed to this. However, the traditional approach to determining function relies on sequence similarity and tools like BLAST, FASTA, and PSI-BLAST (Pearson and Lipman, 1988). Characterization of hypothetical protein allows rapid development of novel therapeutic approaches including drug development and *in-silico* vaccination. This study is focused on characterizing a critical hypothetical protein of *S Sonnei* and it will open new opportunities to develop novel medication against this pathogenic organism.

Nucleic acid sequences in a hypothetical protein can be utilized to predict the structures and functions of the HPs. Moreover, hypothetical proteins do not have any experimental or biochemical proof of their existence. These proteins also exhibit a low degree of similarity to known, annotated proteins (Lubec *et al.*, 2005). Only a few numbers of HPs are preserved and are present in creatures from many evolutionary lineages. In sequenced microbial genomes, HPs make up a significant portion of the genes; yet, they have not yet been functionally defined and described at the protein chemistry level (Galperin and Koonin, 2004). Finding novel conformational orientations of 3-dimensional structures allows for the evaluation of new domains and motifs as well as the discovery of additional protein pathways and cascades, all of which are benefits of studying the function of proteins with unknown functions. Potential pharmaceutical targets might be found in these new domains.

Moreover, structural and functional characterization of HPs may reveal novel biomarkers and therapeutic targets. Several bioinformatics databases and methodologies were used to efficiently annotate the possible protein functions of various dangerous bacteria. Shigellosis, caused by *Shigella sonnei*, is a serious public health problem across the world, especially in impoverished nations.

The *Shigella* genus, which includes the *S. dysenteriae, S. boydii, S. flexneri, and S. sonnei* species, is a gram-negative, rod-shaped, facultative anaerobe, does not produce spores and is not mobile bacterial family. Although the bacterial genome is known, there are thousands of potential proteins (HPs) whose functions remain unknown. Every year, over 700,000 people die from Shigellosis, a digestive sickness caused by a family of bacteria (*Shigella* infection) (Rabbi *et al.*, 2021). *Shigella flexneri, Shigella boydii, Shigella sonnei,* and *Shigella dysenteriae* are the four pathogenic *Shigella* species recognized. One of these four species, *S. sonnei*, is often observed in impoverished countries and can cause deadly epidemics. It is a gram-negative, non-spore producing facultative anaerobe bacillus.

This bacterium is commonly discovered in polluted water supplies as well as in the feces of sick people. It is important to characterize the hypothetical proteins of *S. sonnei* in order to understand the metabolic functionality and also to develop novel therapeutic approaches against this organism. There have been many previous reports of *in silico* structural and functional identification of hypothetical proteins in pathogenic bacteria including *Staphylococcus aureus*, *Vibrio cholerae*, and *Klebsiella pneumonia* (Varma *et al.*, 2015; Islam *et al.*, 2015; Singh *et al.*, 2022). Moreover, the *in silico* characterization of hypothetical protein was done successfully in *Shigella dysenteriae* species which revealed vital pathogenic protein's function and activities (Rabbi *et al.*, 2021).

Hence, we employed various bioinformatics tools to identify the important function of Hypothetical Protein RUK71877.1 of *S. Sonnei* which may open new possibilities to develop therapeutic means.

Materials and methods

Sequence retrieval

The NCBI database (http://www.ncbi.nlm.nih.gov/) contains 13,665 accessible *Shigella sonnei* genomes. The hypothetical protein ELP71_10070 from *Shigella sonnei,* accession number RUK71877.1, was used in this investigation. The amino acid sequences of the targeted protein were retrieved as FASTA format for further analysis.

Analysis of physicochemical properties

ProtParam is a widely used tool offered by ExPasy server to compute the physico-chemical properties of proteins from amino acid sequences (Gasteiger *et al.*, 2005). The physico-chemical properties of target protein, including its molecular weight, extinction coefficients, aliphatic index (AI), GRAVY (grand average of hydropathy), and isoelectric point (pI), were examined using ProtParam tool.

Subcellular localization and solubility prediction

CELLO (http://cello.life.nctu.edu.tw/) server was used to predict the subcellular localization of the targeted HP. CELLO is offered by The Molecular Bioinformatics Center and utilizes integrated use of molecular structures, gene and protein sequence information (Yu *et al.*, 2004). To forecast the solubility, the Protein-Sol web tool (https://protein-sol.manchester.ac.uk/) was used. In addition, To further investigate the result the solubility was examined through SOSUI (http://harrier.nagahama-i-bio.ac.jp/sosui/) that also calculates the average hydrophobicity (Hirokawa *et al.*, 1998). To further check our result the target protein was evaluated in Protein-sol server (https://protein-sol.manchester.ac.uk/) (Hebditch *et al.*, 2017).

Function prediction by domain and motif analysis

Conserved Domain (CD) Search can be used to detect the conserved domains present in a protein sequence. The NCBI Conserved Domain Search Service (CD Search) was employed for the domain analysis of our selected protein (https://www.ncbi.nlm.nih.gov.cgi). A query sequence is compared to position-specific score matrices produced from conserved domain alignments contained in the Conserved Domain Database (CDD) using RPS-BLAST (Reverse Position Specific BLAST) (Lu *et al.*, 2020). MOTIF server to search for the available motifs of protein (http://www.genome.jp/tools/motif/).

Multiple sequence alignment

To search for the homologues of the protein, NCBI's BLASTp (http://www.ncbi.nlm.nih.gov/) program was used as opposed to the non-redundant database. Jalview is a free, cross-platform tool for editing, visualizing, and analyzing multiple sequence alignment (Waterhouse *et al.*, 2009). Multiple sequence alignment and a phylogenetic tree were produced using Jalview.

Secondary structure determination

SOPMA server provides self-optimization prediction technique for the prediction of secondary structures of a protein (https://npsaprabi.ibcp.fr/cgibin/npsa automat.pl?page=/NPSA/npsa_sopma.html) (Combet *et al.*, 2000). Moreover, we used the PSIPRED server as supplementary tools for verifying the SOPMA results.

Molecular modeling and quality assessment

As the chosen protein structure was not available in the Protein Data Bank (RCSB PDB), we used I-TASSER server to build the three-dimensional structure of the protein (Zheng *et al.*, 2021). The GalaxyWEB server was then used to refine the predicted models (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE) (Ko *et al.*, 2012). Using Saves Server v6.0 (https://saves.mbi.ucla.edu), the models ERRAT quality scores and ramachandran plots were compared to determine the best model (Colovos and Yeates, 1993; Laskowski *et al.*, 1996). Additionally, the ProSA-web protein structure analysis tool was used to assess the projected model's quality (Wiederstein and Sippl, 2007).

Energy minimization of the model structure

Using the YASARA force field minimizer, the energy of the improved three-dimensional predicted model was reduced to a minimum (Land and Humble, 2018). The proposed protein has a more precise, stable, and energy-efficient ideal 3D structure.

Molecular Docking analysis

HDock server (HDOCK Server) (http://hdock.phys.hust.edu.cn/) was used to do the docking analysis. It facilitates the estimation of protein-protein docking interactions (Huang and Zou, 2014; Yan *et al.*, 2020). The IrmA Family protein from *Escherichia coli* WP 001614359.1 and the target putative protein's binding affinity to the receptor IL-4R were determined using the HDock server. PyMOL software was used to assess the docking data.

Comparative genomics approach

A BLASTp (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) search against the proteome of *Homo sapiens* was carried out to see whether our target hypothetical protein RUK71877.1 had any similarity to humans. The hits were filtered using a threshold E-value (expected value) of.005 and a minimum bit score of 100.

Results and discussion

Physicochemical properties and subcellular localization

The ProtParam tool evaluated many physicochemical characteristics of the putative protein RUK71877.1, which are displayed in Table 1. The protein was estimated to have 149 amino acids, a molecular weight of 16098.15, a theoretical pI of 5.01, and a GRAVY score of -0.244 for hydropathicity.

The target protein's estimated instability index (II) of 32.06 and categorized it as stable protein. A hypothetical protein's subcellular localization would be helpful to understand its function, as various cellular sites correspond to different activities.

The development of a drug that targets the target protein can also be achieved using this knowledge. The CELLO software projected that our target protein will have an "extracellular" subcellular location.

The protein was predicted to be soluble by the SOSUI server. The target protein's solubility is also confirmed by the Protein-sol server (Niwa *et al.*, 2009). Predicted scaled solubility was 0.654 which indicates the target protein to be soluble shown in Fig. 1.

Table 1. Physicochemical properties of the target protein.

No. of Amino acids.	Molecular Weight	Half life	pI	(Asp +Glu)	(Arg + Lys)	Aliphatic index (AI)	Instability index (II)	Grand average of hydropathicity (GRAVY)
149	16098.15	30hr	5.01	19	14	83.09	32.06 (Stable)	-0.244

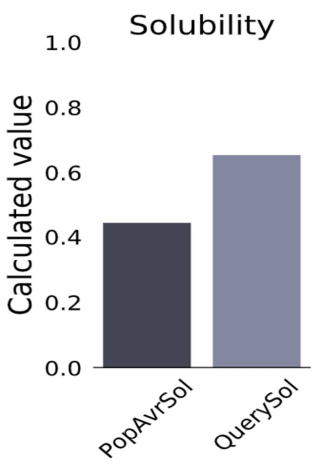

Fig. 1. Solubility chart analysis of the target protein.

Protein family and phylogeny analysis

We identified conserved domains and possible functions of our target protein using a variety of annotation methods. The target protein was predicted by NCBI-CD Search, Motif tool and found to contain the IrmA superfamily protein (cl39990) and the PRK10884 superfamily protein (cl32601), a protein with an SH3 domain; Provisional. The IrmA super family (cl39990) domain was predicted by the NCBI-CDD service with an E-value of 4.87e49 at amino acid residues 29–134. Additionally, MOTIF discovered the PRK10884 superfamily protein at position 95-135 with an E-value of 0.056 and the IrmA family protein at position 29-133

with an E-value of 2.4e-42. With other known IrmA family proteins from other types of organisms, the BLASTp search against the non-redundant database revealed 97% sequence homology (Table 2). Top 10 homologous proteins were retrieved for phylogeny analysis by utilizing the software Jalview. The phylogenetic analysis is given in Fig. 2, Fig. 3. The closest protein to our hypothetical one was WP 040220400.1.

Table 2. Similar proteins with hypothetical protein RUK71877.1.

Description	Scientific Name	Max Score	E value	Per. Identity	Accession
IrmA family protein [Enterobacteriaceae]	Enterobacteriaceae	100%	3e-106	100.00%	WP_040220400.1
hypothetical protein [*Escherichia coli*]	*Escherichia coli*	100%	1e-105	99.33%	EJL4301939.1
hypothetical protein [*Escherichia coli*]	*Escherichia coli*	100%	2e-105	99.33%	EJA9025142.1
IrmA family protein [Enterobacteriaceae]	*Enterobacteriaceae*	100%	2e-105	99.33%	WP_060615266.1
IrmA family protein [*Escherichia coli*]	*Escherichia coli*	100%	4e-105	99.33%	WP_001614359.1
IrmA family protein [Enterobacterales]	*Enterobacterales*	97%	7e-104	100.00%	WP_001061894.1
hypothetical protein [*Escherichia coli*]	*Escherichia coli*	97%	2e-103	99.32%	MBS8717841.1
IrmA family protein [Klebsiella pneumoniae]	*Klebsiella pneumoniae*	100%	4e-103	97.99%	WP_040207084.1
IrmA family protein [Enterobacterales]	*Enterobacterales*	97%	4e-103	99.32%	WP_020837115.1
IrmA family protein [Enterobacteriaceae]	*Enterobacteriaceae*	97%	4e-103	99.32%	WP_072650159.1
hypothetical protein [*Escherichia coli*]	*Escherichia coli*	97%	5e-103	99.32%	EHV0151273.1

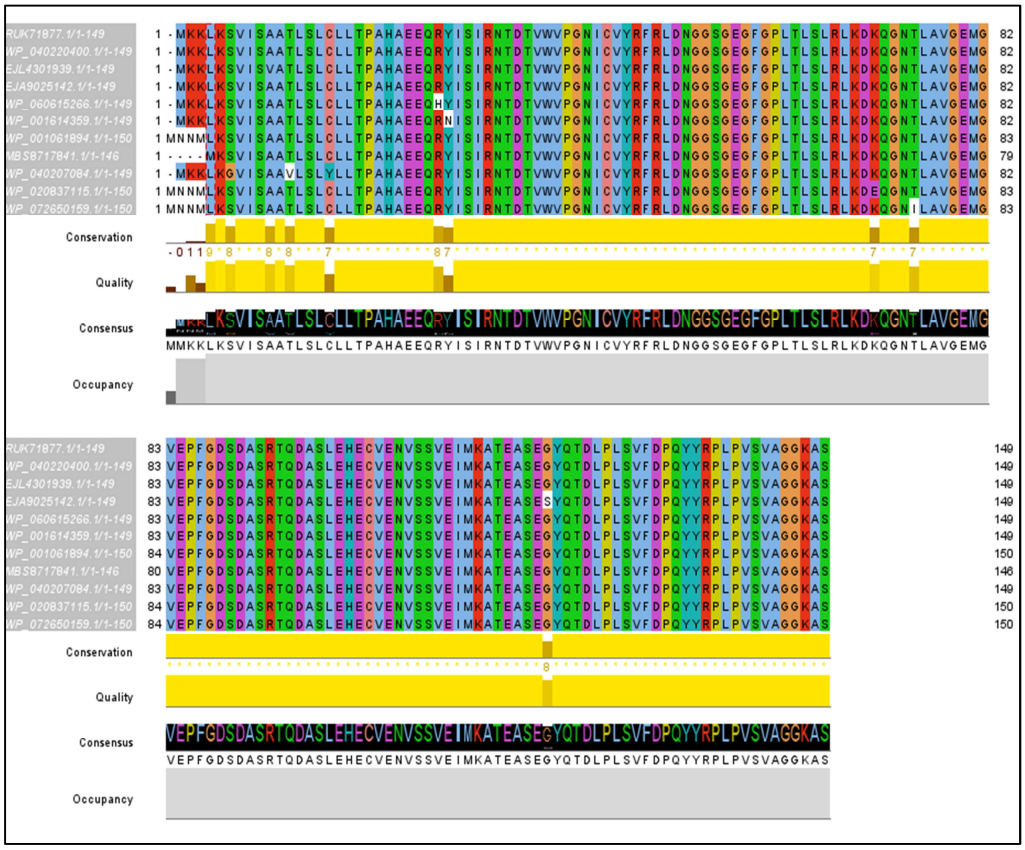

Fig. 2. Multiple sequence alignment with the target protein RUK71877.1.

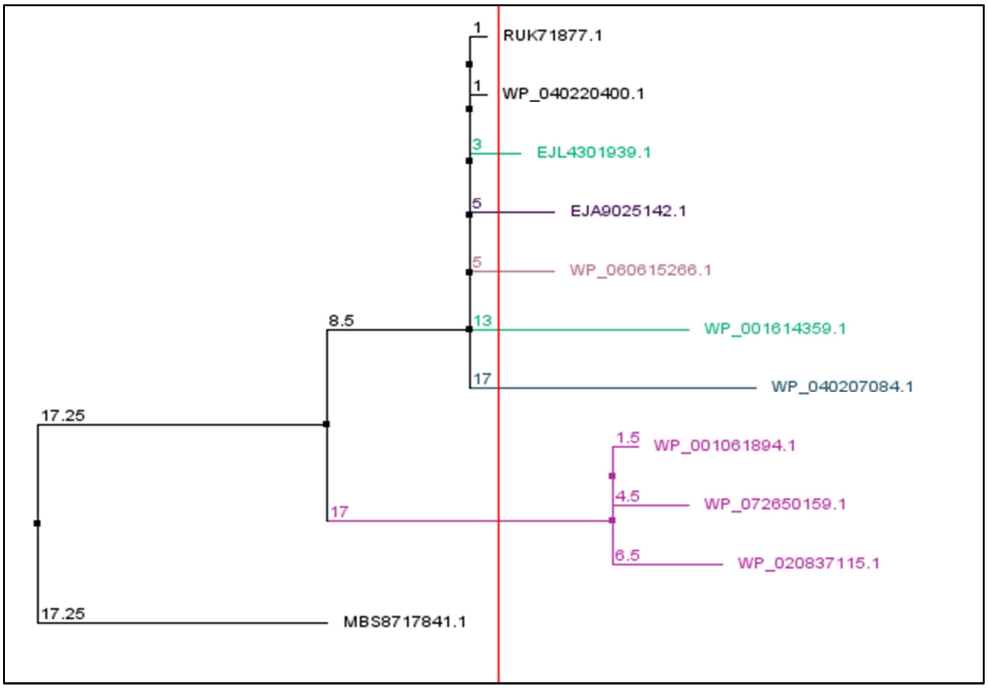

Fig. 3. Phylogenetic tree analysis of the target proteins.

Secondary structure analysis

Secondary structure analysis by SOPMA revealed that the random coil was predominant (36.24%) followed by alpha helix (30.20%) extended strand (25.50%), and beta turn (8.05%). In the case of 3 conformational states prediction by SOPMA, the results were found to be random coil (43.62%), alpha helix (31.54%), and extended strand (24.83%). Secondary structure of the protein predicted by PSIPRED is shown in Fig. 4

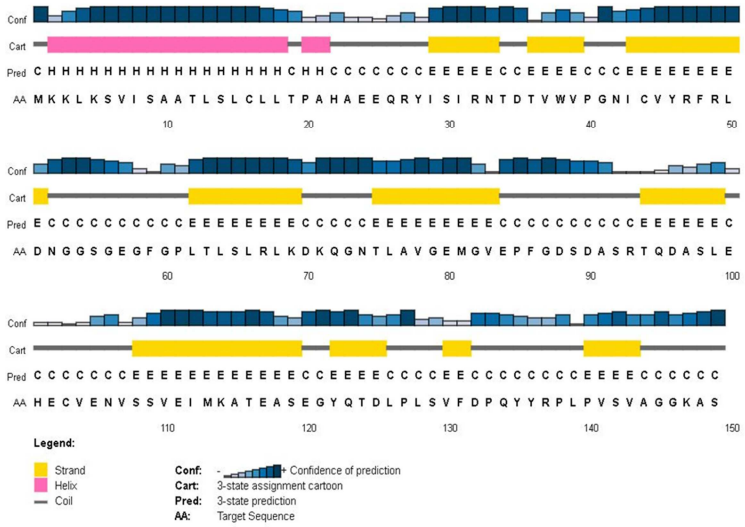

Fig. 4. Secondary structure of the target protein RUK71877.1.

Tertiary model analysis, quality assessment

The I-TASSER server for tertiary structure prediction revealed 5 models for our protein of interest. The best structures were picked up based on their ERRAT quality score and Ramachandran plot analysis and the best model was improved using the GAlaxyWEB server. This server revealed 5 better models for our protein. We eventually picked the best model which is based on their ERRAT quality score and Ramachandan plot analysis.

The ERRAT value and Ramachandan plot results are described in Fig. 5. The model contained residues in most favored regions at 88.9% and residues in disallowed regions at 1.6% only and the ERRAT value of overall quality factor remained at 97.761. The predicted model was then tested using the ProSA-web tool, and it was predicted to be good with a low error rate and Z score of -4.34 (Fig. 6).

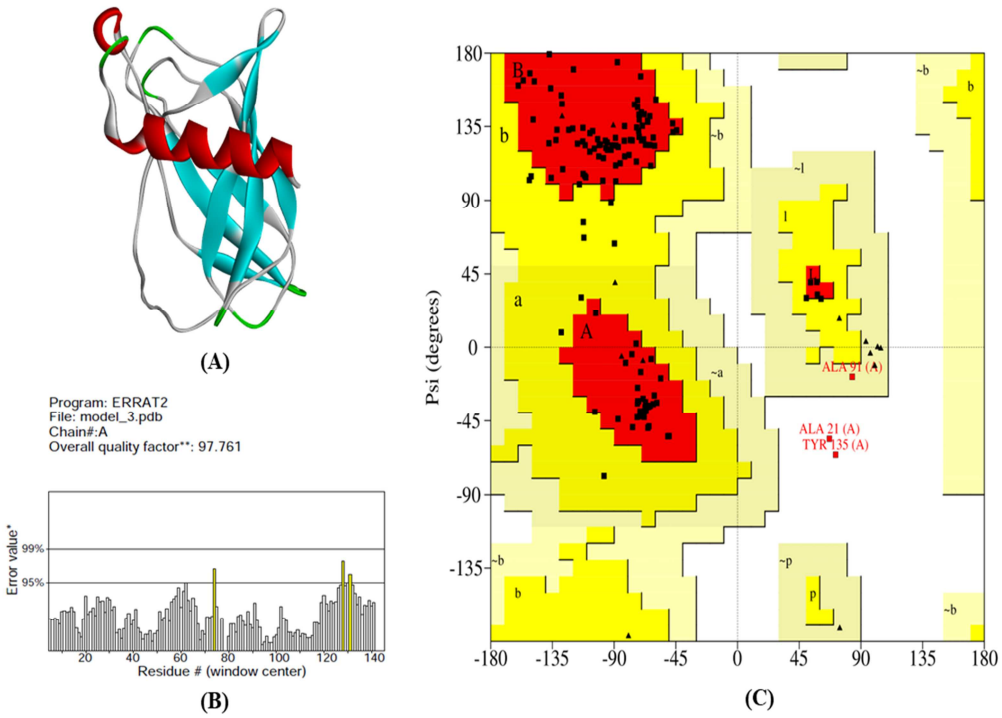

Fig. 5. Prediction of the A) 3D model, B) Errat value, C) ramachadan plot of the target protein.

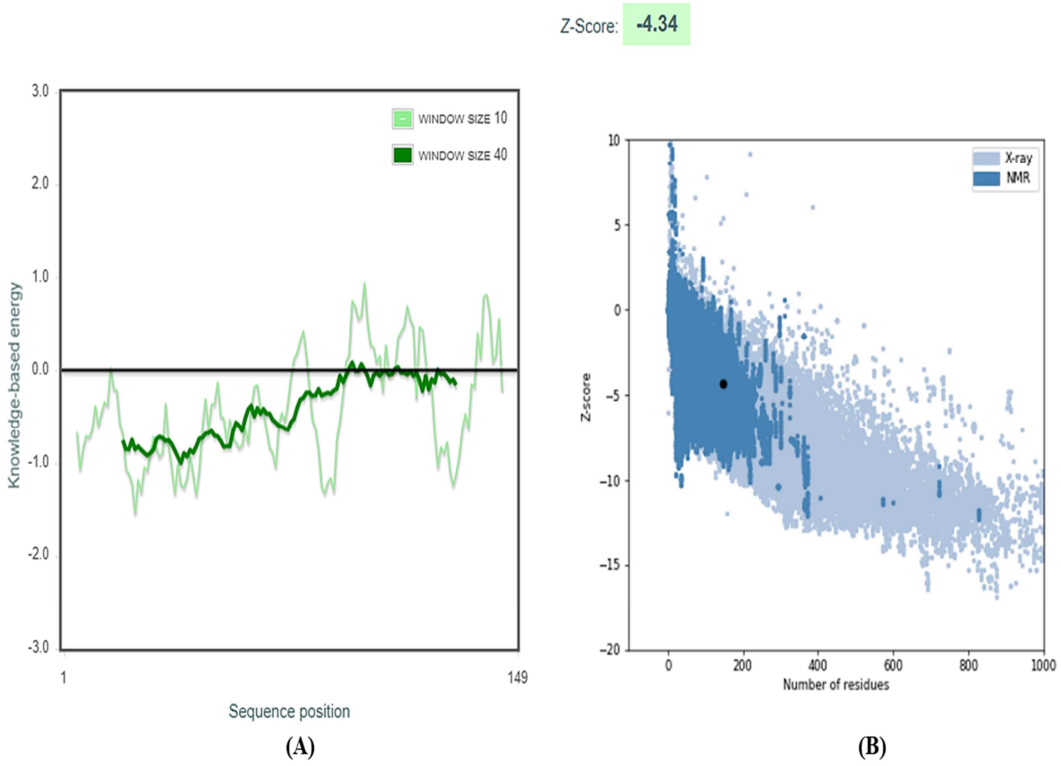

Fig. 6. Quality assessment of the target protein evaluation A) Overall quality and B) Z score.

Energy minimization

Before docking, these refined models were further used to reduce its energy using YASARA software. The energy was minimized into -70196.7 kj/mol and the score was -1.18 reduced from -4849.2kj/mol and a score of -1.71. Utilizing the Pymol software The RMSD of the energy minimized protein (Fig. 7) was found to be 0.472.

Molecular docking analysis

IrmA Family protein of *Escherichia coli* with accession id WP_001614359.1 is a well characterized protein which was compared with our hypothetical protein RUK71877.1. Both of them showed good binding affinity against IL-4R.

In HDock server RUK71877.1 with IL-4R showed lowest score with -336.19 (Table 3). Further assessment was done using ClusPro server which again proved the high affinity of our hypothetical protein with IL-4R (Table 4, Fig. 8).

Human homologous analysis

In order to determine whether the target protein has any known human homologues, a BLASTp search against the human proteome was conducted. The bacterial proteins that are non-homologous to the human proteins might be a potential therapeutic candidate without any potential side effects. According to the results, as our target protein showed no similarity with any of the known human proteins, it could be a potential drug target to treat *Shigella sonnei.*

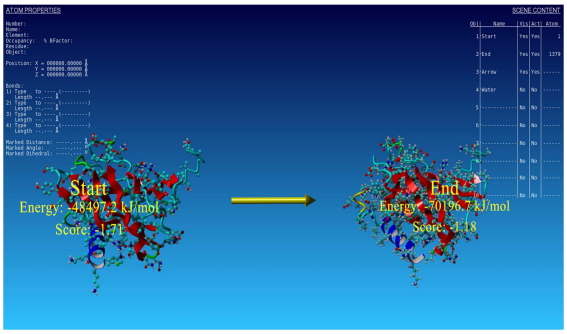

Fig. 7. Energy minimization of the predicted protein.

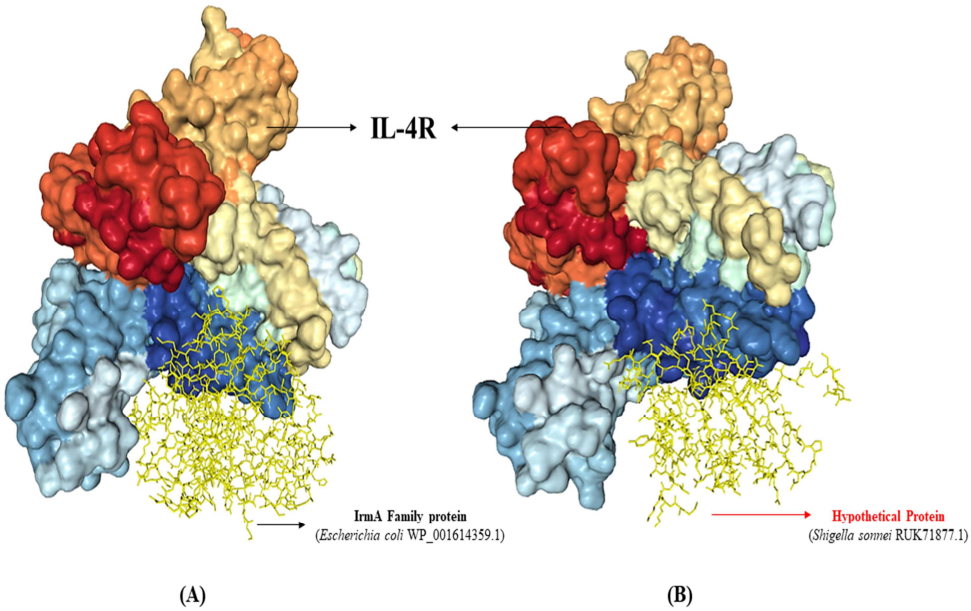

Fig. 8. Docking analysis of A) IL-4R with WP_001614359.1 complex and B)IL-4R with RUK71877.1 target protein.

Table 3. Docking score analysis using HDock server.

Receptor	Protein	Docking Score	rmsd (Å)
IL-4R	>WP_001614359.1 IrmA Family protein [*Escherichia coli*]	-306.97	56.69
	>RUK71877.1 hypothetical protein ELP71_10070 [*Shigella sonnei*]	-336.19	83.52

Table 4. Docking Analysis using ClusPro.

Receptor	Protein	Members	Center
IL-4R	>WP_001614359.1 IrmA Family protein [*Escherichia coli*]	101	-860.2
	>RUK71877.1 hypothetical protein ELP71_10070 [*Shigella sonnei*]	65	-948.2

Discussion

Shigella sonnei is an increasingly prominent pathogen globally. Since it is the most prevalent infectious species of shigellosis (bloody diarrhea) in affluent nations and the second most common in underdeveloped and developing countries (LMICs) (Thompson *et al.*, 2015).

It is a strict human pathogen linked to a significant amount of morbidity and mortality worldwide (Kotloff *et al.*, 2018). *S. sonnei* has been able to equip and adapt with antimicrobial resistance rapidly (Muthuirulandi Sethuvel *et al.*, 2017). Therefore, it is important to identify and characterize new proteins of *S. sonnei* that can be utilized as drug targets. There are currently no licensed vaccinations for *Shigella*, researchers are attempting to develop one. Therefore, identifying and characterizing a vast range of hypothetical proteins is vitally important in order to utilize them for vaccination or drug development. However, studies on hypothetical proteins have not yet kept up. Substantial quantities of genomic and proteomic data have been produced due to the low-cost sequencing technologies' quick development. Understanding bacterial metabolic pathways, illness progression, therapeutic development, and disease control techniques can all be made better with the help of hypothetical protein characterization.

In this study, we employed various bioinformatics resources to structurally and functionally characterize the Hypothetical Protein RUK71877.1 of *Shigella sonnei*. The protein sequence was retrieved from the NCBI database. By analyzing physicochemical properties, the protein was estimated to contain 149 amino acids with a molecular weight of 16098.15, theoretical pI of 5.01, and grand average of hydropathicity (GRAVY) of −0.224 (Table 1).

CELLO server predicted this soluble protein to be the outer-membrane, which means it could be utilized for vaccine targets. Phylogenetic analysis was done to identify the resembled family of the hypothetical protein. Domain and motif analysis revealed that our target hypothetical protein is IrmA family protein which is an interleukin receptor that mimics protein A. The findings were verified by all the annotation tools with high confidence. The IrmA domain was predicted by the NCBI-CDD at amino acid residues 29–134 also found the PRK10884 superfamily protein at position 95-135. The BLASTp conducted against the non-redundant database also showed maximum similarity with IrmA family proteins. Secondary structure of the protein consists of random coil, alpha helix, beta turn, and extended strand with random coil being the predominant one.

The predicted 3D structure of selected protein was valid by ERRAT value and Ramachandran plot analysis. Among all the refined models, the model 3 was found to be best. The Errat value of the model was found to be 97.761. The projected model was then assessed using the ProSA-web tool, and it was shown to be effective with a Z score value of -4.34. The energy of the refined model was minimized using YASARA. It was minimized to -70196.7 kj/mol and the score was -1.18. which makes the complex more stable. Then the model was docked against IL-4R. We choose Interleukin 4 because the other protein of the IrmA family binds well with this receptor (Moriel *et al.*, 2016). The IL-4R also contributes to IgE production in B cells in humans. In the docking study, the selected hypothetical protein showed greater binding energy than IrmA family proteins. IrmA family proteins showed a docking score of -306.97, while Putative Protein RUK71877.1 showed docking score of -336.19. Which means this protein has a very high binding efficiency. Finally, Human homologous analysis was conducted that revealed the selected protein as non-homologous to humans.

This eradicates any chances of unwanted effect and indicates that this protein could be an ideal therapeutic target. The overall bioinformatics study characterized the Hypothetical Protein RUK71877.1 and revealed that it is an interleukin receptor mimic protein A. The result of the study also indicates that this protein could be important for understanding the bacterial proteomes, disease prognosis, drug development, vaccine design and disease control strategies causes. The genome of this bacterium is known, however, numerous hypothetical proteins with unidentified activities still exist. Hypothetical protein characterization can help us learn more about the anabolic and catabolic pathways of this organism.

Recommendations

The findings of this study provide a valuable foundation for future antibacterial medications. This protein can be targeted for *in silico* vaccination and drug development. With appropriate, *in vivo* and *in vitro* approach this protein can be successfully used in therapeutic processes. To get an experimental validation *in vitro* trial is highly recommended.

Acknowledgement

The authors are thankful to the Bioinformatics Laboratory of the Department of Molecular Biology and Genetic Engineering, Sylhet Agricultural University where the research is carried out. The authors also want to thank Md. Shariful Islam, Assistant Professor, Department of Molecular Biology and Genetic Engineering for his support and cooperation in this research.

Conflict of interest

Authors have declared that no competing interests exist.

References

Colovos C, Yeates TO. 1993. Verification of protein structures: Patterns of non-bonded atomic interactions. Protein Science **2(9),** 1511-1519.

Combet C, Blanchet C, Geourjon C, Deleage G. 2000. NPS@: Network protein sequence analysis. Trends in biochemical sciences **25(3),** 147-150.

Galperin MY, Koonin EV. 2004. 'Conserved hypothetical proteins: Prioritization of targets for experimental study. Nucleic Acids Research **32(18),** 5452-5463.

Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server. The Proteomics Protocols Handbook 571-607.

Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. 2017. Protein- Sol: A web tool for predicting protein solubility from sequence. Bioinformatics **33(19),** 3098-3100.

Hirokawa T, Boon-Chieng S, Mitaku S. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics (Oxford, England) **14(4),** 378-379.

Huang SY, Zou X. 2014. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucleic Acids Research **42(7),** e55-e55.

Islam MS, Shahik SM, Sohel M, Patwary NI, Hasan M A. 2015. *In silico* structural and functional annotation of hypothetical proteins of *Vibrio cholerae* O139. Genomics & informatics **13(2),** 53.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology **292(2),** 195-202.

Ko J, Park H, Heo L, Seok C. 2012. GalaxyWEB server for protein structure prediction and refinement. Nucleic Acids Research **40(W1),** W294-W297.

Kotloff KL, Riddle MS, Platts-Mills JA, Pavlinac P, Zaidi AK. 2018. Shigellosis. The Lancet **391(10122),** 801-812.

Land H, Humble MS. 2018. YASARA: a tool to obtain structural guidance in biocatalytic investigations. In Protein Engineering p. 43-67.

Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton JM. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. Journal of Biomolecular NMR **8(4),** 477-486.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS. 2020. CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Research **48(D1),** D265-D268.

Lubec G, Afjehi-Sadat L, Yang J-W, John JPP. 2005. Searching for hypothetical proteins: Theory and practice based upon original data and literature. Progress in Neurobiology **77(1-2),** 90-127.

Moriel DG, Heras B, Paxman JJ, Lo AW, Tan L, Sullivan MJ, Dando SJ, Beatson SA, Ulett GC, Schembri MA. 2016. Molecular and Structural Characterization of a Novel Escherichia coli Interleukin Receptor Mimic Protein. mBio **7(2),** e02046.

Muthuirulandi Sethuvel D, Devanga Ragupathi N, Anandan S, Veeraraghavan B. 2017. Update on: *Shigella* new serogroups/serotypes and their antimicrobial resistance. Letters in Applied Microbiology **64(1),** 8-18.

Niwa T, Ying B-W, Saito K, Jin W, Takada S, Ueda T, Taguchi H. 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. Proceedings of the National Academy of Sciences **106(11),** 4201-4206.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences **85(8),** 2444-2448.

Rabbi MF, Akter SA, Hasan MJ, Amin A. 2021. *In silico* characterization of a hypothetical protein from *Shigella dysenteriae* ATCC 12039 Reveals a Pathogenesis-Related Protein of the Type-VI Secretion System. Bioinformatics and Biology Insights **15**, 11779322211011140.

Singh V, Dhankhar P, Dalal V, Tomar S, Kumar P. 2022. *In-silico* functional and structural annotation of hypothetical protein from *Klebsiella pneumonia*: A potential drug target. Journal of Molecular Graphics and Modelling **116,** 108262.

Thompson CN, Duy PT, Baker S. 2015. The rising dominance of *Shigella sonnei*: an intercontinental shift in the etiology of bacillary dysentery. PLoS neglected tropical diseases **9(6),** e0003708.

Varma PBS, Adimulam YB, Kodukula S. 2015. *In silico* functional annotation of a hypothetical protein from *Staphylococcus aureus*. Journal of Infection and Public Health **8(6),** 526-532.

Waterhouse A, Procter J, Martin D. 2009. a, Clamp M, Barton GJ. 2009. Jalview Version 680 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics **25,** 1189-1681.

Wiederstein M, Sippl MJ. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic acids research **35(suppl_2),** W407-W410.

Yan Y, Tao H, He J, Huang S-Y. 2020. The HDOCK server for integrated protein–protein docking. Nature protocols **15(5),** 1829-1852.

Yu CS, Lin CJ, Hwang JK. 2004. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science **13(5),** 1402-1406.

Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang Y. 2021. Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell reports methods **1(3),** 100014.