Research article

Structural and Functional Annotation of Hypothetical Protein AVO28_00330 of *Yersinia pestis*: An *In Silico* Approach

Sheikh Sunzid Ahmed[1]*

[1]Department of Botany, University of Dhaka, Dhaka 1000, Bangladesh.

Corresponding author: E-mail: sunzid79@gmail.com

Abstract: *Yersinia pestis* is an infamous gram-negative, coccobacillus enterobacterium responsible for three devastating plague pandemics worldwide. The recent outbreak of this zoonotic disease demands *in silico* study of the hypothetical proteins for efficient drug and vaccine discovery. As hypothetical proteins constitute a substantial portion of the proteome, it's essential to annotate them structurally and functionally. The current study characterized physicochemical properties, predicted homology-based 3D structure and annotated functions of the hypothetical protein AVO28_00330 of *Y. pestis* using a range of bioinformatic tools and softwares. Swiss Model and Phyre2 server were utilized to predict the tertiary model which was minimized energetically using YASARA server. The quality assessment servers found the model as a good one. For future molecular docking analysis, active binding sites were predicted using CASTp. Protein-protein interaction analysis was performed in STRING server. For functional prediction InterPro, Pfam, Motif and other tools were used. The hypothetical protein revealed tricopeptide repeat domain and rubredoxin metal-binding domain which regulates lipopolysaccharide metabolic process in the outer cell membrane which contributes to virulence property of the protein. Therefore, this *in silico* analysis will improve the current understanding of the protein and aid in the future analysis regarding therapeutic drug and vaccine investigation.

Keywords: *Yersinia pestis*, homology modeling, rubredoxin, hypothetical protein, functional annotation

Introduction

Yersinia pestis, the etiologic agent of plague and a member of the family *Enterobacteriaceae,* is a gram-negative, non-spore forming, non-motile coccobacillus that grows within a temperature range of 4 to 400C and optimum pH range of 7.2 to 7.6 [1]. This beyond infamous bacterium is responsible for three devastating pandemics throughout history namely the Justinian's plague, the Black Death and the Modern plague [2]. The plague is zoonotic as it spreads from rodents as a natural reservoir to humans using fleas as the vector [3]. Bites of fleas during blood-meal to humans, direct contact with a mucous membrane or damaged skin, inhalation of aerosolized air droplets cause transmission of the pathogenic bacterium to human [4]. It causes the death of the individual within a week if left untreated for bubonic form and even less than a week for septicemic form and pneumonic form. The rapid development of bacterial biofilm inside the digestive tract of flea helps *Y. pestis* to adapt to a unique life stage for effective transmission [5]. For its high similarity in the genomic level with *Y. pseudotuberculosis*, *Y. pestis* is thought to be a recently emerged clone of it [6], [7]. The USA, the former Soviet Union and Japan developed *Y. pestis* as a biological weapon during the 20th century [2].

In the 21st century, the plague has been reported from Asia, Africa and America as the pathogenic strain is endemic to animal populations and the recent outbreak in Uganda, the Democratic Republic of Congo, China and Madagascar indicates the major health concern [6]. Though plague has a significant disease history, no highly efficient vaccine with long-lasting support has still been developed. Moreover, the recent emergence of antibiotic-resistant strains poses a serious threat to global public health and biodefense [6], [8], [9]. All these aspects trigger biotechnological interest among the scientists with an integrated *in silico* approach to study *Y. pestis* for new drug synthesis and vaccine development.

Hypothetical proteins are predicted or experimentally uncharacterized proteins and they constitute a substantial portion of the proteome of both eukaryotes and prokaryotes [10]. With the remarkable advancement in the field of Next Generation Sequencing (NGS), the number of hypothetical proteins is increasing rapidly and comparing to that experimental validation rate is not so high. This gap of structural and functional annotation can be reduced through *in silico* approach using modern bioinformatic tools which might pave the way for new drug synthesis and vaccine development. Thus, the current study focuses on annotating AVO28_00330 hypothetical protein of *Y. pestis*, both structurally and functionally for an improved understanding, which might help later at drug and vaccine development.

2. Materials and Methods

2.1 Sequence retrieval and similarity identification

The amino acid sequence of AVO28_00330 was retrieved in FASTA format from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) with the GenBank accession ID of KZC74892.1. A similarity search using the NCBI Blastp program [11] was performed initially against the non-redundant and UniProtKB/SwissProt [12] database to predict the function of the hypothetical protein.

2.2 Multiple sequence alignment and phylogeny analysis

Multiple sequence alignment (MSA) was performed using MUSCLE algorithm in MEGA 10 [13], [14] between the hypothetical protein and other similar proteins obtained from Blastp. MSA was crosschecked by Clustal Omega program of EMBL-EBI [15]. Then phylogeny analysis was done using NEXUS file generated by MEGA into Phylogeny.fr [16].

2.3 Physicochemical characterization

Different physical and chemical properties including molecular weight, amino acid composition, atomic position, extinction coefficient, estimated half-life, instability index, aliphatic index, grand average of hydropathicity (GRAVY), isoelectric point, total number of negatively charged residues (Asp + Glu), total number of positively charged residues (Arg + Lys) were predicted using ProtParum tool (http://web.expasy.org/protparam/) of ExPASy [17].

2.4 Subcellular localization

Subcellular localization was predicted using CELLO 2.5 [18]. Results were cross-checked with PSORTb [19], PSLpred [20], SOSUIGramN [21]. HMMTOP [22], TMHMM [23], SABLE [24] were utilized to predict the presence of transmembrane helices in the hypothetical protein. Protein solubility was predicted using Protein-Sol [25]. Signal peptide prediction was performed using PrediSi [26] and SignalP-5.0 server [27].

2.5 Secondary structure prediction

Self-optimized prediction method with alignment (SOPMA) [28] and PSI-blast based secondary structure prediction (PSIPRED 4.0) [29] servers were utilized for secondary structure prediction.

2.6 Tertiary structure modeling, visualization and quality assessment

Tertiary structure was modeled using Swiss Model [17] and Phyre2 server [30]. For higher accuracy, the best scoring template was selected for homology modeling. 3D model was visualized using UCSF Chimera [31]. For quality assessment of the obtained models, PROCHECK [32], Verify3D [33] and ERRAT [34] server were utilized. Finally, energy minimization was performed for the best predicted model using YASARA energy minimization server [35].

2.7 Active site detection

The active sites were determined using Computed Atlas of Surface Topography of Protein (CASTp) server which provides an online resource for locating, delineating, and measuring concave surface regions on three-dimensional structures of proteins [36].

2.8 Functional annotation

Conserved domain database (CDD, available at NCBI) [37] was searched for conserved domain. For protein motif and domain searching, Motif [38], Pfam [39], InterPro [40], ScanProsite [41], SMART [42], were utilized. Protein folding pattern was recognized using PFP-FunD SeqE server [43]. For virulence property analysis, VirulentPred [44] was utilized.

2.9 Protein-protein interaction analysis

STRING 11.5 [45] server was utilized to predict the possible protein-protein functional interaction network.

2.10 Submission of the model to protein model database

The suitable model generated for hypothetical protein AVO28_00330 of *Y. pestis* was successfully submitted to Protein Model Database (PMDB) [46].

3. Results and Discussion

3.1 Sequence similarity and phylogeny analysis

Blastp result against non-redundant and SwissProt database showed homology with lipopolysaccharide assembly proteins (Table 1). Multiple sequence alignment between the hypothetical protein and other homologous proteins generated NEXUS file in MEGA. To strengthen homology assessment between proteins, down to complex and subunit level, phylogenetic analysis was performed. The phylogenetic tree showed distances between branches and reveals close similarity of the hypothetical protein with WP_046596310.1 *Y. pestis* homolog while distantly related with NLU15143.1 *Serratia liquefaciens* (Fig. 1).

Table 1: Similar proteins obtained from non-redundant and SwissProt database

Non redundant database	Protein ID/Entry Name	Organism	Protein Name	Identity (%)	Score	Query Coverage (%)	e-value
	WP_011192449.1	*Yersinia pseudotuberculosis* complex	MULTISPECIES: Lipopolysaccharide assembly protein LapB	99.74	800	100	0.0
	WP_025383887.1	*Yersinia similis*	Lipopolysaccharide assembly protein LapB	98.71	796	100	0.0
	WP_174849589.1	*Yersinia enterocolitica*	Lipopolysaccharide assembly protein LapB	96.66	780	100	0.0
	WP_145537544.1	*Yersinia kristensenii*	Lipopolysaccharide assembly protein LapB	96.66	779	100	0.0
	WP_050125144.1	*Yersinia aleksiciae*	Lipopolysaccharide assembly protein LapB	95.63	775	100	0.0
	WP_038242561.1	*Yersinia ruckeri*	Lipopolysaccharide assembly protein LapB	90.26	720	97	0.0
	WP_037415793.1	*Serratia grimesii*	Lipopolysaccharide assembly protein LapB	87.89	706	97	0.0
	WP_115158869.1	*Serratia fonticola*	Lipopolysaccharide assembly protein LapB	87.66	701	97	0.0
	WP_126483050.1	*Serratia plymuthica*	Lipopolysaccharide assembly protein LapB	87.37	701	97	0.0
	NLU15143.1	*Serratia liquefaciens*	Lipopolysaccharide assembly protein LapB	87.37	698	97	0.0

UniProtKB/ SwissProt	P0AB58.1	*Escherichia coli* K-12	Lipopolysaccharide assembly protein B	77.57	631	97	0.0
	P44130.1	*Haemophilus influenzae* Rd KW20	Lipopolysaccharide assembly protein B	48.35	386	98	8e-132

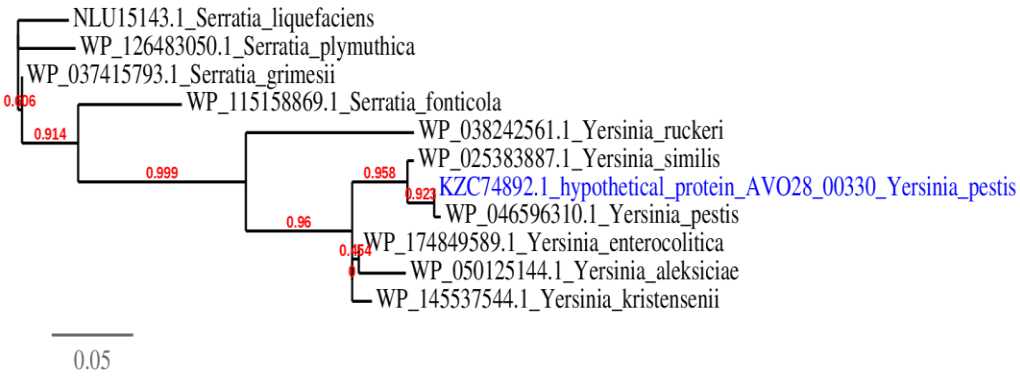

Fig. 1. Phylogenetic tree showing relationship of AVO28_00330 hypothetical protein with other similar proteins with true genetic distance.

3.2 Physicochemical characterization

The protein AVO28_00330 was predicted to contain 389 amino acids where Leu (51) and Trp (5) are most abundant and least abundant, respectively (Table 2). The molecular weight was calculated as 44336.69 Da and theoretical pI was 5.94, indicating the protein to be acidic and negatively charged. Total number of positively charged residues (Arg + Lys) and total number of negatively charged residues (Asp + Glu) were found 46 and 54, respectively. The instability index was 41.32, indicating the unstable nature of the protein [47]. Aliphatic index was 88.12 which gives an indication of proteins stability over a wide temperature range. The GRAVY was -0.366, which indicates the protein is non-polar and hydrophilic. This also indicates better interaction possibility with water [48]. High extinction coefficient value (47370) indicates the presence of Cys, Trp and Tyr residues [49]. The N-terminal of the sequence was considered M (Met). Protein half-life is an estimation of the period of time which is required for the radiolabeled focus protein density to be decreased by 50 percent compared to the amount at the onset of the chase [50]. Estimated half-life was found to be 30 h in mammalian reticulocytes (*in vitro*), >20 h in yeast (*in vivo*), >10 h in Escherichia coli (*in vivo*). Total number of atoms and molecular formula were 6189 and $C_{1942}H_{3082}N_{562}O_{579}S_{24}$, respectively.

Table 2: Amino acid composition

Sl. No.	Amino Acids	Number of Residues	Percentage (%)
1	Ala (A)	44	11.3
2	Arg (R)	28	7.2
3	Asn (N)	10	2.6
4	Asp (D)	23	5.9
5	Cys (C)	8	2.1
6	Gln (Q)	30	7.7
7	Glu (E)	31	8.0
8	Gly (G)	22	5.7
9	His (H)	13	3.3
10	Ile (I)	10	2.6
11	Leu (L)	51	13.1
12	Lys (K)	18	4.6
13	Met (M)	16	4.1
14	Phe (F)	11	2.8
15	Pro (P)	7	1.8

16	Ser (S)	18	4.6
17	Thr (T)	10	2.6
18	Trp (W)	5	1.3
19	Tyr (Y)	13	3.3
20	Val (V)	21	5.4

3.3 Subcellular localization

Subcellular localization prediction helps to characterize a protein as a potential drug or vaccine candidate. Cytoplasmic matrix proteins are capable to be selected as potential drug targets and both inner and outer membrane proteins can act as potential vaccine targets [51]. CELLO 2.5 predicted the protein to be localized into the cytoplasm and the result was validated by PSORTb, PSLpred and SOSUIGramN (Table 3). The protein was predicted as soluble protein by Protein-Sol. Prediction of signal peptide is important to understand the transport system and cleavage sites of the hypothetical protein. Signal peptide was detected by both PrediSi and SignalP-5.0. However, no transmembrane helices were detected using HMMTOP, TMHMM and SABLE which further emphasizes the protein to be cytoplasmic.

Table 3: Assessment of subcellular localization

Predictions	**Servers**	**Results**
Prediction of subcellular localization	CELLO 2.5	Cytoplasmic
	PSORTb	Cytoplasmic
	PSLpred	Cytoplasmic
	SOSUIGramN	Cytoplasmic
Prediction of protein solubility	Protein-Sol	Soluble protein
Signal peptide prediction	PrediSi	Present (predicted for secretion)
	SignalP-5.0	Present
Prediction of transmembrane helices	HMMTOP	Absent
	TMHMM	Absent
	SABLE	Absent

3.4 Secondary structure prediction

Considering default parameters SOPMA server was utilized first. SOPMA predicted 25.71% residues as random coils in comparison to alpha-helix (68.12%), extended strand (2.31%) and beta turn (3.86%) (Table 4). PSIPRED also predicted similar result showing higher confidence (Fig. 2-3). Secondary structure helps to understand function of the protein better as strong correlation exists between protein structure and function.

Table 4: Secondary structure elements

Secondary Structure Elements	**Values (%)**
Alpha helix (Hh)	68.12
3_{10} helix (Gg)	0.00
Pi helix (Ii)	0.00
Beta bridge (Bb)	0.00
Extended strand (Ee)	2.31
Beta turn (Tt)	3.86
Bend region (Ss)	0.00
Random coil (Cc)	25.71
Ambiguous states	0.00
Other states	0.00

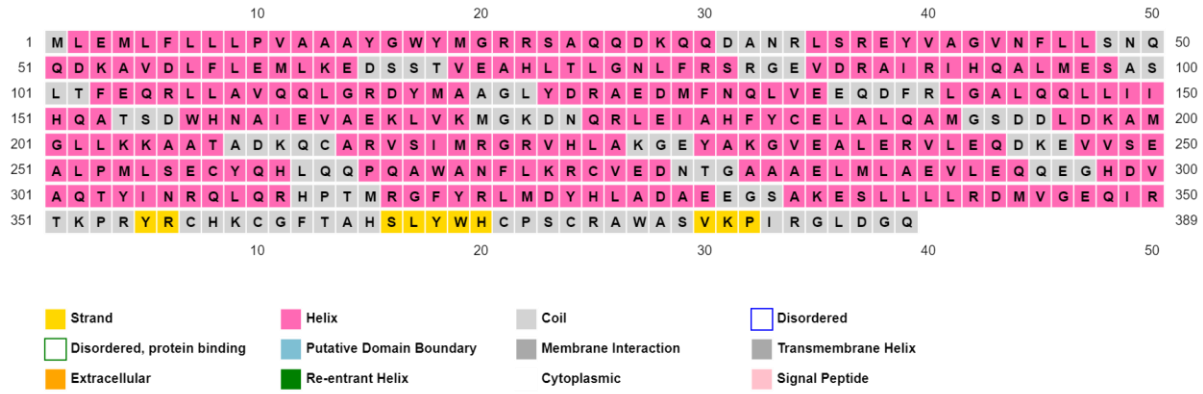

Fig. 2. Secondary structure compostion in annotation grid predicted by PSIPRED.

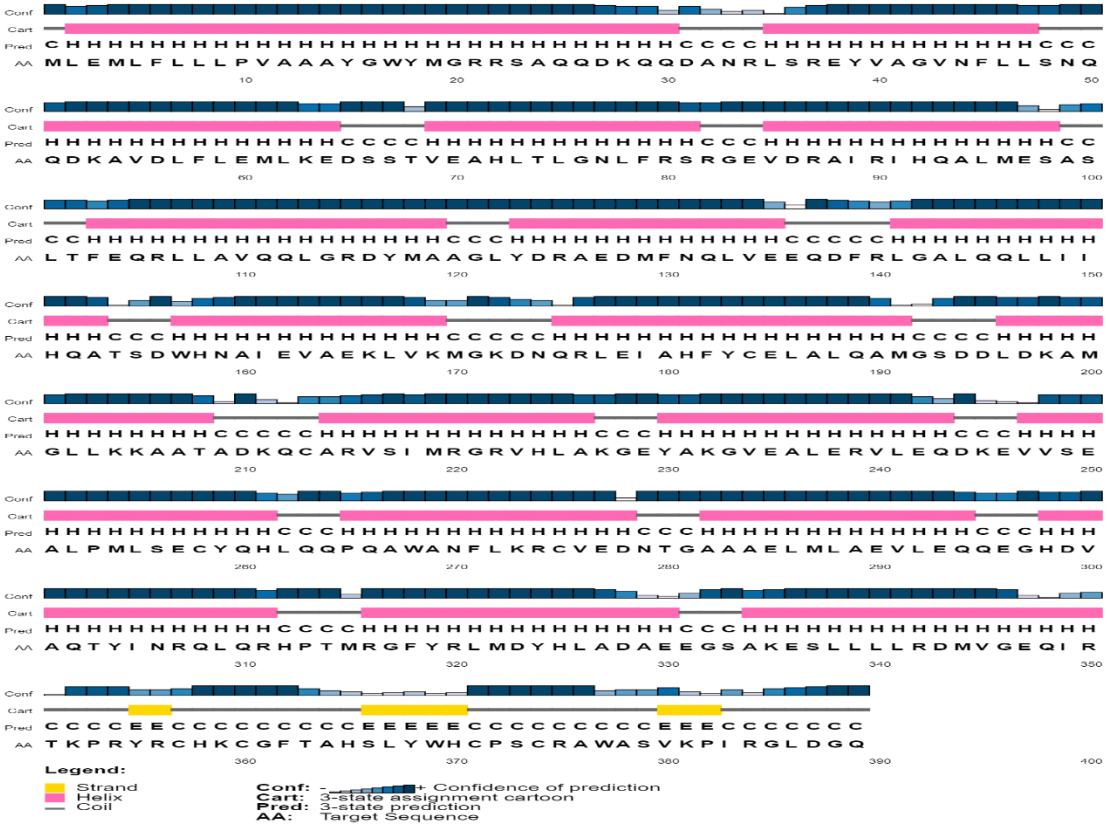

Fig. 3. Predicted secondary structure in PSIPRED chart showing level of confidence.

3.5 Tertiary structure modeling, visualization and quality assessment

Homology modeling approach was taken for determining the tertiary structure of the hypothetical protein. Swiss Model server predicted the 3D structure (Fig. 4) based on the most favored template 4zlh.1.B (PDB ID: 4ZLH_B). 4ZLH is the crystal structure of *Escherichia coli* protein with lipopolysaccharide assembly protein B (LapB) cytoplasmic domain. This template protein is a homo-dimer which has two chains (Chain A and Chain B) and chain B was used to build the model by Swiss

Model server. For this template, values of Global Model Quality Estimation (GMQE), Quaternary Structure Quality Estimation (QSQE) and identity score were 0.83, 0.51 and 76.04, respectively. The quality of the model was assessed by PROCHECK through Ramachandran plot analysis, where the distribution of ψ angle and the φ angle in the model within the limits are shown (Fig. 5, Table 5). Residues in the most favored regions covered 93.1% which indicated good quality and validity of the model. Verify3D showed 94.69% of the residues have averaged 3D-1D score ≥ 0.2 (Fig. 6), which indicates good quality of the environmental profile for the predicted model. The overall quality factor predicted by ERRAT server was 98.752, which validates the model as a good one.

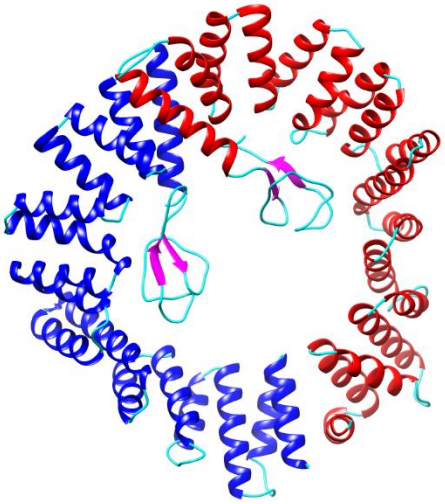

Fig. 4. Energy minimized 3D model of the hypothetical protein predicted by Swiss Model.

Similarly, Phyre2 server predicted the 3D model with 100% confidence and 87% coverage. 337 residues out of 389 were modeled with 100% confidence by the single highest scoring template which was c4zlhB (PDB ID: 4ZLH_B). PROCHECK predicted 91.0% residues in the most favored regions, which indicates good confidence for the predicted model (Table 5). 87.33% of the residues had averaged 3D-1D score ≥ 0.2, according to Verify3D which validates the predicted model. ERRAT server quality factor score was 89.6024, which is suggestive of a good valid model.

Table 5: Ramachandran Plot Statistics

Ramachandran Plot Statistics	**Swiss Model**		**Phyre2**	
	Number	**Values (%)**	**Number**	**Values (%)**
Residues in most favored regions [A,B,L]	565	93.1	283	91.0
Residues in additional allowed regions [a,b,l,p]	40	6.6	26	8.4
Residues in generously allowed regions [a, b, l, p]	1	0.2	1	0.3
Residues in disallowed regions	1	0.2	1	0.3
Number of non-glycine and non-proline residues	607	100	311	100
Number of end-residues (excl. Gly and Pro)	4	--	1	--
Number of glycine residues (shown as triangles)	38	--	19	--
Number of proline residues	12	--	6	--
Total number of residues	661	--	337	--

The tertiary structure modeled by Swiss Model was more preferable than the model predicted by Phyre2 server considering Ramachandran map analysis, Verify3D results and ERRAT server results. Therefore, energy minimization was performed using YASARA server for the Swiss Model 3D structure and scene file (.sce) was visualized in YASARA scene. The energy calculated before energy minimization was -338180.6 kJ/mol and that was changed to a far less value of -431656.0 kJ/mol after energy minimization which makes the predicted model more stable.

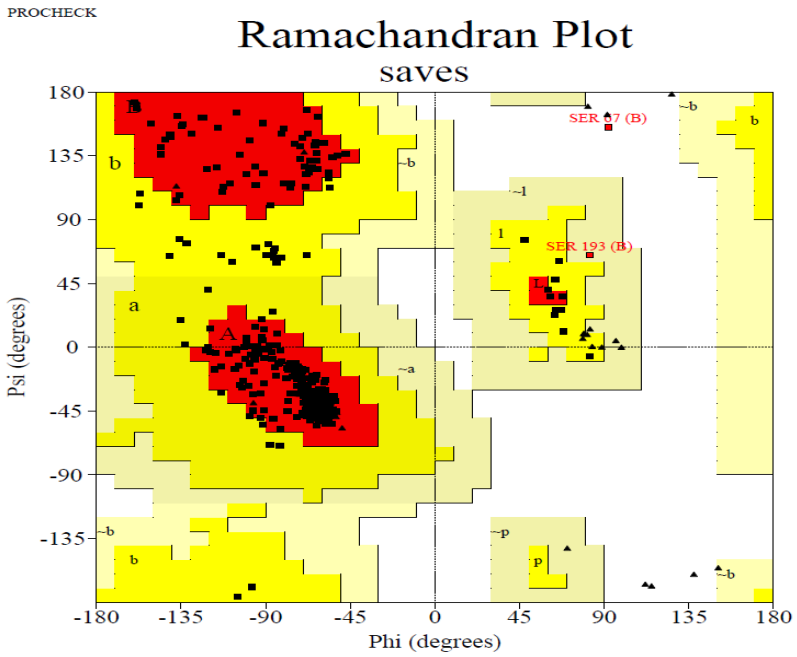

Fig. 5. Ramachandran plot for 3D model predicted by Swiss Model.

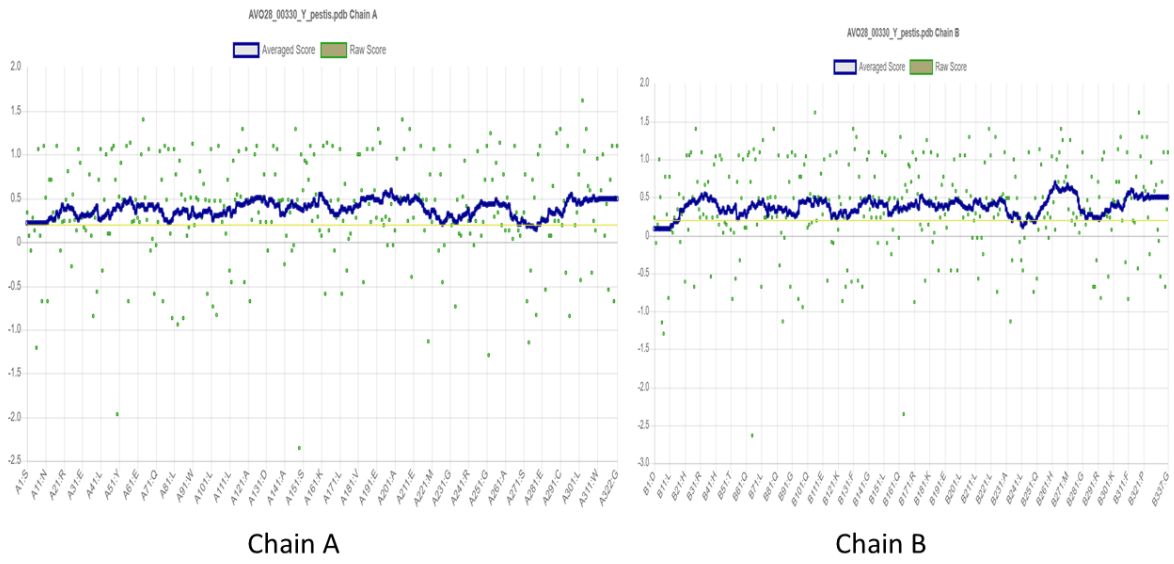

Fig. 6. Verify3D score for Swiss Model predicted tertiary model.

3.6 Active site determination

CASTp 3.0 predicted 52 amino acids to be involved in the potent active sites (Fig. 7). The best active site was found in areas with 968.799 and a volume of 3258.076 amino acids (Fig. 8). Chain A

contains 32 active sites with 15 different amino acids (R, S, G, M, A, Y, Q, I, V, T, K, P, C, F, D) and chain B contains 20 active sites with 14 different amino acids (N, R, S, Q, I, K, P, C, G, F, T, L, D, M). Determined active sites would be helpful for further analysis during the study of pathogenesis, drug and vaccine development.

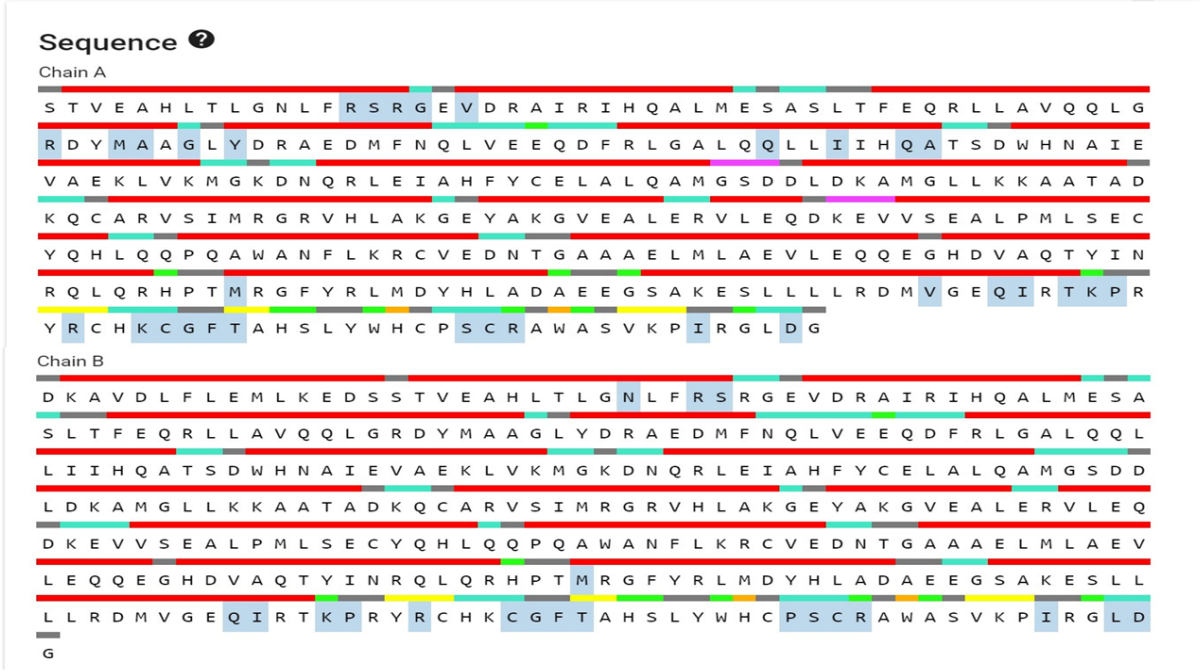

Fig. 7. Sequential representation of active sites (blue block) in both chains of the hypothetical protein.

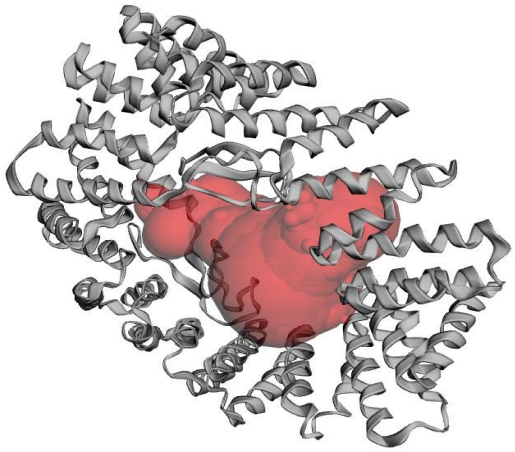

Fig. 8. Active sites (red sphere) of the hypothetical protein in 3D model.

3.7 Functional annotation

Conserved domain search tool predicted the presence of lipopolysaccharide biosynthesis regulator domain (Accession ID: COG2956). The result was cross-checked with other tools. Pfam server predicted significant matches for tricopeptide repeat domain with an e-value of 8.4e-06 and rubredoxin metal-binding domain with an e-value of 1.5e-11 at 190-255 amino acid residues and 355-382 amino acid residues, respectively. InterPro server predicted tricopeptide repeat domain at 69-315 amino acid residues and rubredoxin metal-binding domain at 355-382 amino acid residues. SMART predicted tricopeptide repeat domains at 69-102, 107-140, 180-213, 214-247, 282-315 positions.

ScanProsite and Motif also showed the presence of tricopeptide repeat domain and rubredoxin metal-binding domain. Rubredoxin helps to form small non-heme iron-binding sites that use four cysteine residues to coordinate a single metal ion in a tetrahedral environment. The main feature of this domain is the extended loop or knuckles. Rubredoxin domain binds intimately with tricopeptide motif and this association is essential for lipopolysaccharide regulation and growth into bacterial cells [52]. Lipopolysaccharide at the outer membrane of the cell wall contributes significantly to the pathogenicity of Y. pestis as it enables the bacterium with unique ability to overcome the defense mechanism of both mammalian and insect hosts as well as antibiotics by using lipid A as an anchor to keep the LPS bounded to the membrane whereas orienting its carbohydrate chain towards the environment [4]. After amino acid composition based analysis, VirulentPred suggested this protein as virulent. The globin-like folding pattern was predicted by PFP-FunDSeqE. InterPro server predicted TPR (tricopeptide repeat)-like superfamily for the hypothetical protein. All these results confirm the role of the protein in the metabolic process of lipopolysaccharides, a group of related, structurally complex components of the outer membrane of gram-negative bacteria.

3.8 Protein-protein interaction analysis

Protein-protein interactions (PPI) play a crucial role in basic processing of living cells. PPI data can provide deep insights to reveal molecular machinery for our better understanding of the mechanism of diseases [53]. STRING 11.5 server was used to search for the possible functional fellows of the hypothetical protein in the PPI network. The identified functional partners with scores were- lapA (0.973), cutC (0.634), pgpB (0.616), asmB (0.573), rlpB (0.548), hemX (0.546), lpxH (0.536), ftsH (0.527), YPO3362 (0.520), yfiO (0.515). Of them, YPO3362 is essential cell division protein, yfiO is a part of the outer membrane protein assembly complex, ftsH is a processive protein in the quality control of integral membrane proteins, lpxH is lipid A biosynthesizer, hemX is a methyltransferase, rlpB is a lipopolysaccharide assembler, asmB is involved in lipid A biosynthesis, pgpB is phosphatidyglycerophosphatase B like protein, cutC is involved in the control of copper homeostasis and lapA is involved in the assembly of lipopolysaccharide (Fig. 9)

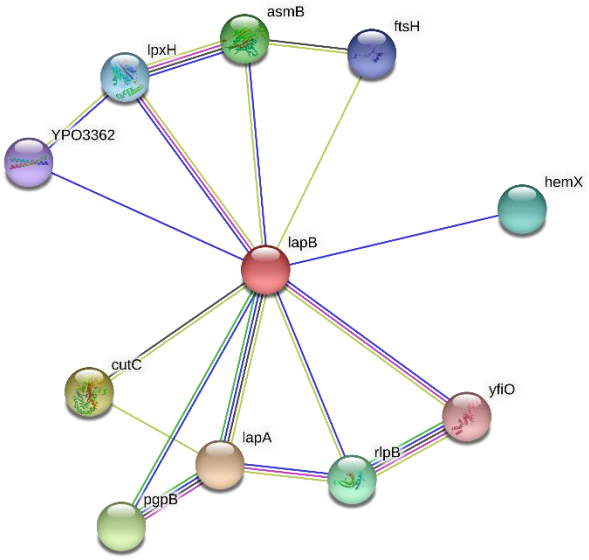

Fig. 9. String network protein-protein interaction analysis showing the functional partners of lapB.

3.9 Submission to Protein Model Database (PMDB)

The predicted 3D model of AVO28_00330 hypothetical protein of *Yersinia pestis* was successfully submitted to the PMDB database. The model can be found searching PMDB ID: PM0084191.

4. Conclusion

The current study was directed to create the first 3D structure and propose probable functions of the *Yersinia pestis* hypothetical protein AVO28_00330. It was submitted as the new record to the protein model database. The identified protein revealed its essential role in the regulation of the lipopolysaccharide metabolic process of the bacterium cell using globin-like folding pattern. Predicted active binding sites of the homology modeled protein would be helpful for further investigation of therapeutic drug designing against the protein using the molecular docking approach. The physicochemical, structural and functional annotation would provide a better understanding of the protein's activity. This sort of methodology would be helpful in the structural and functional elucidation of other uncharacterized proteins. Finally, *in vitro* experimentation should be conducted to validate the predicted results that are shown here and to annotate the protein's role in biotechnology.

Acknowledgments

The author is grateful to Kaiser M Mohaimen for his cordial support during data analysis.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

References

[1] R. D. Perry and J. D. Fetherston, "Yersinia pestis - Etiologic agent of plague," *Clin. Microbiol. Rev.*, vol. 10, no. 1, pp. 35–66, 1997, doi: 10.1128/cmr.10.1.35.

[2] R. K. Auerbach *et al.*, "Yersinia pestis evolution on a small timescale: Comparison of whole genome sequences from North America," *PLoS One*, vol. 2, no. 8, 2007, doi: 10.1371/journal.pone.0000770.

[3] M. Drancourt, L. Houhamdi, and D. Raoult, "Yersinia pestis as a telluric, human ectoparasite-borne organism," *Lancet Infect. Dis.*, vol. 6, no. 4, pp. 234–241, 2006, doi: 10.1016/S1473-3099(06)70438-8.

[4] Y. A. Knirel and A. P. Anisimov, "Lipopolysaccharide of Yersinia Pestis, the Cause of Plague: Structure, Genetics, Biological Properties," *Acta Naturae*, vol. 4, no. 3, pp. 46–58, 2012, doi: 10.32607/20758251-2012-4-3-46-58.

[5] B. J. Hinnebusch, C. O. Jarrett, and D. M. Bland, "'fleaing' the Plague: Adaptations of Yersinia pestis to Its Insect Vector That Lead to Transmission," *Annu. Rev. Microbiol.*, vol. 71, pp. 215–232, 2017, doi: 10.1146/annurev-micro-090816-093521.

[6] C. Demeure, O. Dussurget, G. M. Fiol, A. S. Le Guern, C. Savin, and J. Pizarro-Cerdá, "Yersinia pestis and plague: an updated view on evolution, virulence determinants, immune subversion, vaccination and diagnostics," *Microbes Infect.*, vol. 21, no. 5–6, pp. 202–212, 2019, doi: 10.1016/j.micinf.2019.06.007.

[7] M. Achtman *et al.*, "Microevolution and history of the plague bacillus, Yersinia pestis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 51, pp. 17837–17842, 2004, doi: 10.1073/pnas.0408026101.

[8] T. J. Welch *et al.*, "Multiple antimicrobial resistance in plague: An emerging public health risk," *PLoS One*, vol. 2, no. 3, 2007, doi: 10.1371/journal.pone.0000309.

[9] R. Randremanana *et al.*, "Epidemiological characteristics of an urban plague epidemic in Madagascar, August–November, 2017: an outbreak report," *Lancet Infect. Dis.*, vol. 19, no. 5, pp. 537–545, 2019, doi: 10.1016/S1473-3099(18)30730-8.

[10] J. Ijaq, M. Chandrasekharan, R. Poddar, N. Bethi, and V. S. Sundararajan, "Annotation and

curation of uncharacterized proteins- challenges," *Front. Genet.*, vol. 6, no. MAR, pp. 1–7, 2015, doi: 10.3389/fgene.2015.00119.

[11] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, "NCBI BLAST: a better web interface.," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. 5–9, 2008, doi: 10.1093/nar/gkn201.

[12] B. Boeckmann *et al.*, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 365–370, 2003, doi: 10.1093/nar/gkg095.

[13] R. C. Edgar, "MUSCLE: A multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, pp. 1–19, 2004, doi: 10.1186/1471-2105-5-113.

[14] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, "MEGA X: Molecular evolutionary genetics analysis across computing platforms," *Mol. Biol. Evol.*, vol. 35, no. 6, pp. 1547–1549, 2018, doi: 10.1093/molbev/msy096.

[15] F. Madeira *et al.*, "The EMBL-EBI search and sequence analysis tools APIs in 2019," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W636–W641, 2019, doi: 10.1093/nar/gkz268.

[16] A. Dereeper *et al.*, "Phylogeny.fr: robust phylogenetic analysis for the non-specialist.," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. 465–469, 2008, doi: 10.1093/nar/gkn180.

[17] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch, "ExPASy: The proteomics server for in-depth protein knowledge and analysis," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3784–3788, 2003, doi: 10.1093/nar/gkg563.

[18] O. A. Iyiola *et al.*, "DNA barcoding of economically important freshwater fish species from north-central Nigeria uncovers cryptic diversity," *Ecol. Evol.*, vol. 8, no. 14, pp. 6932–6951, 2018, doi: 10.1002/ece3.4210.

[19] N. Y. Yu *et al.*, "PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics*, vol. 26, no. 13, pp. 1608–1615, 2010, doi: 10.1093/bioinformatics/btq249.

[20] M. Bhasin, A. Garg, and G. P. S. Raghava, "PSLpred: Prediction of subcellular localization of bacterial proteins," *Bioinformatics*, vol. 21, no. 10, pp. 2522–2524, 2005, doi: 10.1093/bioinformatics/bti309.

[21] K. Imai *et al.*, "SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria," *Bioinformation*, vol. 2, no. 9, pp. 417–421, 2008, doi: 10.6026/97320630002417.

[22] G. E. Tusnády and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, no. 9, pp. 849–850, 2001, doi: 10.1093/bioinformatics/17.9.849.

[23] S. Möller, M. D. R. Croning, and R. Apweiler, "Evaluation of methods for the prediction of membrane spanning regions," *Bioinformatics*, vol. 17, no. 7, pp. 646–653, 2001, doi: 10.1093/bioinformatics/17.7.646.

[24] M. Wagner, R. Adamczak, A. Porollo, and J. Meller, "Linear regression models for solvent accessibility prediction in proteins," *J. Comput. Biol.*, vol. 12, no. 3, pp. 355–369, 2005, doi: 10.1089/cmb.2005.12.355.

[25] M. Hebditch, M. A. Carballo-Amador, S. Charonis, R. Curtis, and J. Warwicker, "Protein-Sol: A web tool for predicting protein solubility from sequence," *Bioinformatics*, vol. 33, no. 19, pp. 3098–3100, 2017, doi: 10.1093/bioinformatics/btx345.

[26] K. Hiller, A. Grote, M. Scheer, R. Münch, and D. Jahn, "PrediSi: Prediction of signal peptides and their cleavage positions," *Nucleic Acids Res.*, vol. 32, no. WEB SERVER ISS., pp. 375–379, 2004, doi: 10.1093/nar/gkh378.

[27] J. J. Almagro Armenteros *et al.*, “SignalP 5.0 improves signal peptide predictions using deep neural networks,” *Nat. Biotechnol.*, vol. 37, no. 4, pp. 420–423, 2019, doi: 10.1038/s41587-019-0036-z.

[28] C. Geourjon and G. Deléage, “Sopma: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments,” *Bioinformatics*, vol. 11, no. 6, pp. 681–684, 1995, doi: 10.1093/bioinformatics/11.6.681.

[29] D. W. A. Buchan and D. T. Jones, “The PSIPRED Protein Analysis Workbench: 20 years on,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W402–W407, 2019, doi: 10.1093/nar/gkz297.

[30] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. Sternberg, “Trabajo práctico Nº 13 . Varianzas en función de variable independiente categórica,” *Nat. Protoc.*, vol. 10, no. 6, pp. 845–858, 2016, doi: 10.1038/nprot.2015-053.

[31] E. F. Pettersen *et al.*, “UCSF Chimera - A visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.

[32] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, “PROCHECK: a program to check the stereochemical quality of protein structures,” *J. Appl. Crystallogr.*, vol. 26, no. 2, pp. 283–291, 1993, doi: 10.1107/s0021889892009944.

[33] D. Eisenberg, R. Luthy, and J. Bowei, “Verify3D: Assessment of protein models with three-dimensional profiles,” *Methods Enzymol.*, vol. 277, pp. 396–404, 1997, doi: https://doi.org/10.1016/S0076-6879(97)77022-8.

[34] C. Colovos and T. O. Yeates, “Verification of protein structures: Patterns of nonbonded atomic interactions,” *Protein Sci.*, vol. 2, no. 9, pp. 1511–1519, 1993, doi: 10.1002/pro.5560020916.

[35] E. Krieger *et al.*, “Improving physical realism, stereochemistry and side-chain accuracy in homology modeling: four approaches that performed well in CASP8 Elmar,” *Proteins*, vol. 77, no. Suppl 9, pp. 114–122, 2009, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf.

[36] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang, “CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues,” *Nucleic Acids Res.*, vol. 34, no. WEB. SERV. ISS., pp. 116–118, 2006, doi: 10.1093/nar/gkl282.

[37] A. Marchler-Bauer *et al.*, “CDD: A Conserved Domain Database for protein classification,” *Nucleic Acids Res.*, vol. 33, no. DATABASE ISS., pp. 192–196, 2005, doi: 10.1093/nar/gki069.

[38] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D457–D462, 2016, doi: 10.1093/nar/gkv1070.

[39] R. D. Finn *et al.*, “Pfam: The protein families database,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. 1–9, 2014, doi: 10.1093/nar/gkt1223.

[40] S. Hunter *et al.*, “InterPro: The integrative protein signature database,” *Nucleic Acids Res.*, vol. 37, no. SUPPL. 1, pp. 211–215, 2009, doi: 10.1093/nar/gkn785.

[41] E. de Castro *et al.*, “ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins,” *Nucleic Acids Res.*, vol. 34, no. WEB. SERV. ISS., pp. 362–365, 2006, doi: 10.1093/nar/gkl124.

[42] I. Letunic and P. Bork, “20 years of the SMART protein domain annotation resource,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D493–D496, 2018, doi: 10.1093/nar/gkx922.

[43] H. Bin Shen and K. C. Chou, “Predicting protein fold pattern with functional domain and sequential evolution information,” *J. Theor. Biol.*, vol. 256, no. 3, pp. 441–446, 2009, doi:

10.1016/j.jtbi.2008.10.007.

[44] A. Garg and D. Gupta, “VirulentPred: A SVM based prediction method for virulent proteins in bacterial pathogens,” *BMC Bioinformatics*, vol. 9, no. 62, pp. 1–12, 2008, doi: 10.1186/1471-2105-9-62.

[45] D. Szklarczyk *et al.*, “STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, 2019, doi: 10.1093/nar/gky1131.

[46] T. Castrignanò, P. D. O. De Meo, D. Cozzetto, I. G. Talamo, and A. Tramontano, “The PMDB Protein Model Database.,” *Nucleic Acids Res.*, vol. 34, no. Database issue, 2006, doi: 10.1093/nar/gkj105.

[47] K. Guruprasad, B. V. B. Reddy, and M. W. Pandit, “Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence,” *Protein Eng. Des. Sel.*, vol. 4, no. 2, pp. 155–161, 1990, doi: 10.1093/protein/4.2.155.

[48] M. E. Uddin, P. Maitra, M. F. Hossain, and M. F. Alam, “Isolation and Characterization of Proteases Enzyme from Locally Isolated Bacillus sp.,” *Am. J. Life Sci.*, vol. 2, no. 6, pp. 338–344, 2014, doi: 10.11648/j.ajls.20140206.12.

[49] S. C. Gill and P. H. von Hippel, “Calculation of protein extinction coefficients from amino acid sequence data [published erratum appears in Anal Biochem 1990 Sep;189(2):283],” *Anal Biochem*, vol. 182, no. 2, pp. 319–326, 1989.

[50] P. Zhou, “Determining protein half-lives.,” *Signal Transduct. Protoc.*, vol. 284, pp. 67–77, 2004, doi: 10.1385/1-59259-816-1:067.

[51] D. Prabhu, S. Rajamanikandan, S. B. Anusha, M. S. Chowdary, M. Veerapandiyan, and J. Jeyakanthan, “In silico Functional Annotation and Characterization of Hypothetical Proteins from Serratia marcescens FGI94,” *Biol. Bull.*, vol. 47, no. 4, pp. 319–331, 2020, doi: 10.1134/S1062359020300019.

[52] C. Prince and Z. Jia, “An Unexpected Duo: Rubredoxin Binds Nine TPR Motifs to Form LapB, an Essential Regulator of Lipopolysaccharide Synthesis,” *Structure*, vol. 23, no. 8, pp. 1500–1506, 2015, doi: 10.1016/j.str.2015.06.011.

[53] X. Peng, J. Wang, W. Peng, F. X. Wu, and Y. Pan, “Protein-protein interactions: detection, reliability assessment and applications,” *Brief. Bioinform.*, vol. 18, no. 5, pp. 798–819, 2017, doi: 10.1093/bib/bbw066.