Research article

In Silico Characterization of Essential Hypothetical Proteins from *Francisella tularensis* Schu S4 Strain.

Sheikh Sunzid Ahmed[1]*

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

Corresponding author: E-mail: sheikhsunzid-2015118769@bot.du.ac.bd

Abstract

Francisella tularensis Schu S4 is the causal agent of a sporadic zoonotic disease known as Tularemia, which has shown epidemic outbreaks recently in certain parts of the world. This pathogen is a potential agent of biowarfare or bioterrorism and is classified as a category A pathogen by the National Institute of Allergy and Infectious Diseases. In this virulent strain, 453 genes have been identified as essential genes, indispensable for growth and survival of the pathogen. The functions of 44 proteins encoded by those essential genes were found to be hypothetical and thus defined as essential hypothetical proteins (EHPs). The current study used a wide range of *in silico* tools and servers to annotate the physicochemical, structural, and functional properties of these EHPs. Of all the EHPs, 24 were functionally annotated with a high degree of confidence and validated by Receiver Operating Characteristic curve analysis. Non-homology assessment revealed 20 pathogen-specific EHPs, which were further analyzed for protein-protein interactions and predicted for secondary and tertiary structure. All the 3D structures were checked on multiple quality assessment servers, and the best models were visualized. The outcome of the study could aid in enhancing current understanding of bacterial

Keywords: *Francisella tularensis,* essential hypothetical proteins, *in silico* annotation, ROC analysis, tularemia.

1. Introduction

Francisella tularensis is the causal agent of a sporadic zoonotic disease in humans called tularemia or "rabbit-fever". This gram-negative, facultative intracellular, pleomorphic coccobacillus was identified nearly 100 years ago, but the nature of its pathogenic interactions remained largely undefined until recent times (Chase et al., 2009). It is a highly virulent pathogenic organism that was given highest priority in the offensive biological weapons programs of the Soviet Union and the US during the Cold War (Dennis et al., 2001; Kingry and Petersen, 2014). The re-emergent concerns regarding the potential threat resulting from its misuse as a weapon of mass destruction by criminals and terrorists have emphasized the need for further research for vaccine development (Conlan and Oyston, 2007). Most of the virulent strains belong to the subspecies *tularensis* for which Schu S4 is the type strain (Kadzhaev et al., 2009). This Schu S4 strain can kill humans with a dose as low as 10 CFU (Saslaw and Carlisle, 1961). Clinical expression of the disease depends primarily on the route of transmission. Infection is acquired in the human body in various ways, such as skin contact with infected animals, ingestion of contaminated food and water, arthropod bites, and infective aerosol inhalation (Tarnvik and Berglund, 2003). In healthy individuals, fever and acute symptoms are hallmarks of the disease. Different forms of tularemia include ulceroglandular, oculoglandular, oropharyngeal, and pneumonic tularemia. Among these various forms, pneumonic tularemia is acquired through infective aerosol inhalation and is the most severe form that represents some particular challenges (Dennis et al., 2001). Most cases of natural acquisition can be successfully treated if the disease is diagnosed earlier. The Centers for Disease Control and Prevention (CDC) has listed this pathogen as a Class A biothreat agent (Twenhafel et al., 2009). Furthermore, the epidemic outbreak of tularemia disease in certain parts of the world, such as Finland, Russia, Sweden, and the south-central and western states of the USA, demands more research for successful drug and vaccine development (Keim et al., 2007; Twenhafel et al., 2009).

Essential genes are those that are indispensable for the survival and growth of an organism. The hypothetical or uncharacterized proteins encoded by the essential genes are referred to as essential hypothetical proteins (EHPs). All the essential genes/proteins theoretically are considered as putative drug targets as inactivation or deletion of such proteins/genes is lethal for the bacterium. Therefore, prediction of essential hypothetical proteins can play a significant role in shortlisting potential or putative drug targets (Prava et al., 2018). With the advancement of high-throughput sequencing technology, the number of sequenced genomes as well as essential hypothetical proteins is ever increasing (Raj et al., 2017). Bioinformatics-based analysis of these uncharacterized proteins can lead to robust predictions of structures, physicochemical properties, and functions that otherwise, if unknown, may lead to potential hindrance in the study of pathogenicity, vaccines, and drug discovery.

Therefore, this study aims to explore all the essential hypothetical proteins of the *Francisella tularensis* Schu S4 strain encoded by the essential genes available in the database of essential genes (DEG) to annotate physicochemical, structural, and functional properties of the EHPs using a wide range of bioinformatics servers and tools. The complete framework of this study has been elucidated in Fig. 1.

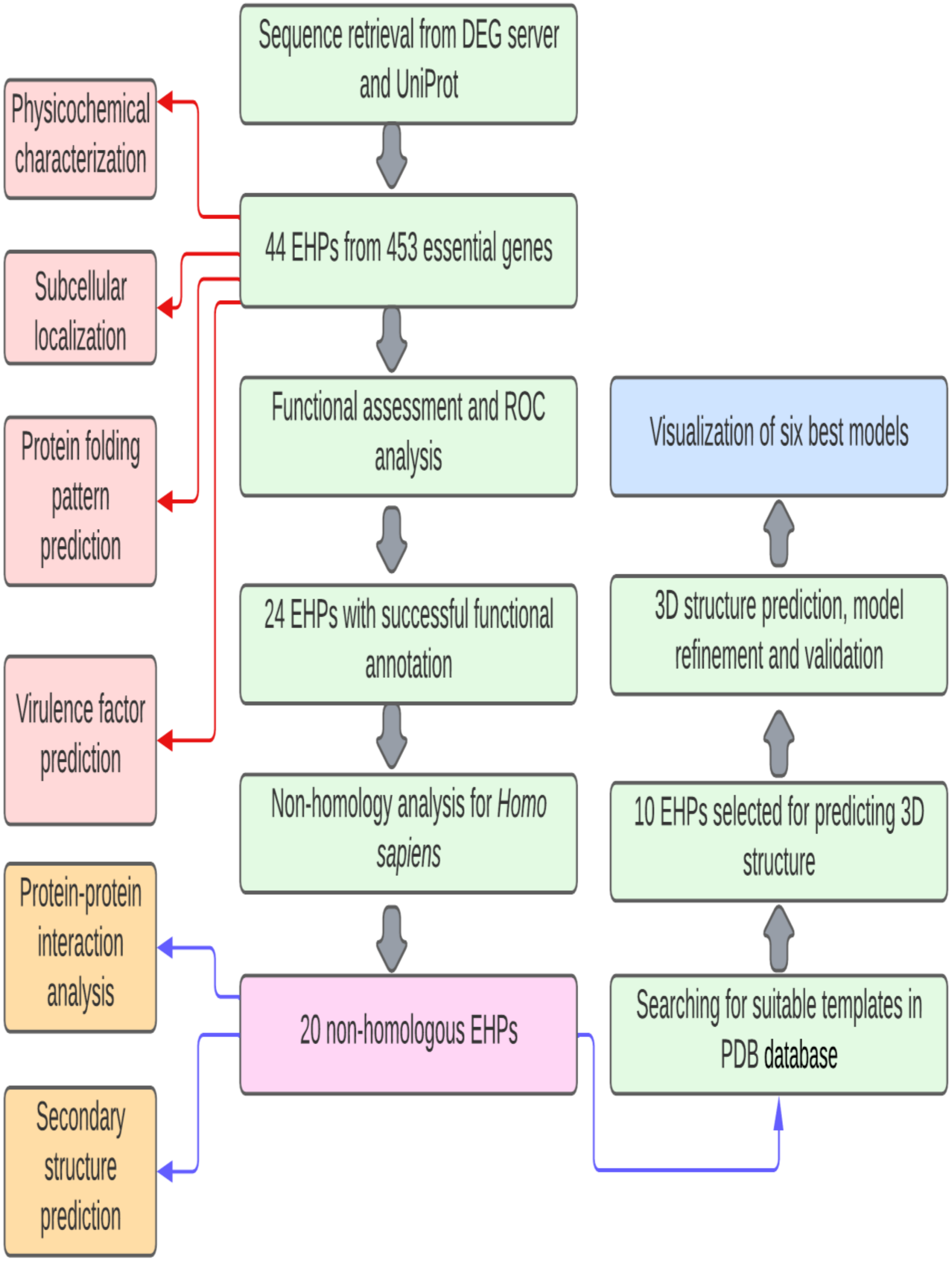
Sequence retrieval from DEG server and UniProt
44 EHPs from 453 essential genes
Physicochemical characterization
Subcellular localization
Protein folding pattern prediction
Virulence factor prediction
Functional assessment and ROC analysis
24 EHPs with successful functional annotation
Non-homology analysis for *Homo sapiens*
20 non-homologous EHPs
Protein-protein interaction analysis
Secondary structure prediction
Searching for suitable templates in PDB database
10 EHPs selected for predicting 3D structure
3D structure prediction, model refinement and validation
Visualization of six best models

Fig. 1. Complete framework used for *in silico* characterization of essential hypothetical proteins.

2. Materials and Methods

2.1 Sequence retrieval and physicochemical characterization

Sequence search was carried out in the Database of Essential Genes (DEG) (Luo et al., 2013). Out of 453 records, 44 were found to be unique hypothetical proteins encoded by these essential genes. These 44 hypothetical proteins were considered essential hypothetical proteins (EHPs) and retrieved from the UniProt (http://www.uniprot.org) database in the FASTA format along with their UniProt ID for further analysis. Physicochemical properties such as molecular weight (Mw), theoretical isoelectric point (pI), grand average of hydropathicity (GRAVY), aliphatic, and instability index were computed for the EHPs using ExPASy's ProtParam server (http://web.expasy.org/protparam) (Gasteiger et al., 2003).

2.2 Subcellular localization

PSORTb (Yu et al., 2010), PSLpred (Bhasin et al., 2005) and Cello 2.5 (Yu et al., 2006) were used to predict subcellular localization of the EHPs. Signal peptide was predicted using SignalP 4.1 (Nielsen, 2017) and transmembrane information was retrieved using TMHMM (Moller et al., 2001) and HMMTOP (Tusnady and Simon, 2001). SecretomeP (Bendtsen et al., 2005) was utilized to identify EHPs involved in non-classical secretory pathways.

2.3 Virulence factor prediction

The virulence nature of the EHPs was predicted using MP3, VICMpred and the VFDB server. The MP3 server uses an integrated SVM-HMM approach to provide improved efficiency and accuracy to analyze proteins of pathogenic sources for both metagenomic and genomic datasets (Gupta et al., 2014). VICMpred server with an overall accuracy of 70.75%, employs SVM-based methods having patterns, amino acid and dipeptide composition of bacterial protein sequences (Saha and Raghava, 2006). VFDB server performs an iterative and exhaustive sequence similarity searches among the hierarchical prebuilt datasets using VFanalyzer pipeline to accurately identify potential pathogenic strains (Liu et al., 2019).

2.4 Functional assignment and domain analysis

Functions of all the essential hypothetical proteins (EHPs) were predicted using different available functional databases and tools including CDD, Pfam, InterProScan, SMART, PANTHER, MOTIF and CATH. The CDD offers live search services and an archive of pre-computed domain annotations for both single protein and nucleotide queries and larger sets of protein query sequences (Lu et al., 2020). A large collection of protein families are deposited into the Pfam database, each represented with multiple sequence alignments and Hidden Markov Models (HMMs) (Finn et al., 2016). InterPro detects input sequences to search for similarity against InterPro protein signature databases using the InterProScan tool (Jones et al., 2014). SMART correlates input sequences with the database and searches for sequences with similar domains based on domain architecture and profiles (Letunic et al., 2012). PANTHER uses a library system and an indexing system, both of which are based on HMMs, multiple sequence alignments, and gene ontology searching to identify protein superfamilies (Thomas et al., 2003). Available motifs in the EHPs were identified with the MOTIF search, which uses PROSITE, CDD, and Pfam databases as libraries to search for similarity (Kanehisa et al., 2016). CATH identifies protein superfamilies and detects structurally related proteins even with lower sequence identity (Orengo et al., 1997). PFP-FunDSeqE server was used to explore protein folding patterns in the EHPs, which combines information on functional domains and evolution to predict patterns of protein folding in the protein structures (Shen and Chou, 2009).

2.5 Evaluation of performance

ROC curve analysis was performed for all the 44 EHPs to evaluate the accuracy of the *in silico* tools used for functional prediction and domain analysis (Bradley, 1997). For each one out of seven tools, five levels were considered to estimate efficiency. The input data had two columns. The first column was assigned by binary 0 as for true negative prediction and binary 1 for true positive prediction. In the second column, integer values ranging from one to five were assigned, where a higher value indicated higher confidence. The input data was submitted to the ROC Analysis server (http://www.jrocfit.org) (Eng, 2014) following the format 1. Upon executing the online ROC program, the measures of the ROC curve, such as accuracy, sensitivity, specificity, and area under the curve (AUC) were obtained.

2.6 Host non-homology analysis

Pathogen-specific proteins were identified by performing host non-homology analysis for all the functionally annotated EHPs. For that, functional EHPs were subjected to a BLASTp (Altschul et al., 1990) search against the non-redundant database of the human proteome with an e-value threshold of 0.0001. Non-homologous proteins were selected for secondary and tertiary structure prediction.

2.7 Secondary structure prediction

Secondary structure was predicted using PSIPRED server (Buchan and Jones, 2019) for all the functionally annotated EHPs. PSIPRED uses two-feed forward neural networks to perform analysis on the output obtained from PSI-BLAST (Altschul et al., 1990).

2.8 Tertiary structure prediction and quality assessment

Non-homologous proteins obtained after non-homology analysis were subjected to a BLASTp search against the PDB database to get suitable templates for homology modeling. Non-homologous proteins showing a sequence identity greater than 30% with the PDB templates were considered for homology modeling. The SWISS-MODEL (Waterhouse et al., 2018) and Phyre2 (Kelley et al., 2015) servers were used to predict the 3D structure of the EHPs based on best scoring templates. For quality assessment, the models were initially checked using PROCHECK (Laskowski et al., 1993), Verify3D (Eisenberg et al., 1997) and ERRAT (Colovos and Yeates, 1993) servers. Then the best scoring models were refined using the GalaxyRefine (Heo et al., 2013) server and rechecked with quality assessment tools. Based on the score selected 3D models were visualized using BIOVIA Discovery Studio v21.1.0.20298 (BIOVIA, 2020).

2.9 Protein-protein interaction network analysis

STRING server currently contains approximately 24.6 million protein sequences from 5090 organisms (http://string-db.org/). STRING 11.5 (Szklarczyk et al., 2021) was used to predict possible functional partners of the non-homologous EHPs.

3. Results and Discussion

3.1 Physicochemical characterization

All the 44 EHPs retrieved from the DEG database were considered for physicochemical characterization (Table 1). Molecular weight plays an important role in the functional characterization of proteins. Protein Q5NFM6 and protein Q5NG09 showed the highest and lowest molecular weights of 52390.21 Da and 5897.62 Da, respectively. Isoelectric point (pI) prediction helps in the development of buffer systems and subsequent purification processes. The predicted pI value ranged from 4.21 to 9.77. Out of 44 EHPs, 23 were found to be acidic, having pI values ranging from 4.21 to 6.96 and 21 were found to be basic, having pI values ranging from 7.63 to 9.77. The extinction coefficient of the EHPs was estimated in water at 280 nm based on the concentration of cysteine, tryptophan, and tyrosine residues in the protein sequences (Gasteiger et al., 2003). A higher percentage of these residues is responsible for higher extinction coefficient values. Calculation of the extinction coefficient helps in the quantitative analysis of protein-ligand and protein-protein interactions for drug discovery investigations. Protein Q5NFU0 and protein Q5NEJ2 were found to have the highest and lowest extinction coefficient values, respectively, whereas Q5NIM1 did not show the value due to the absence of cysteine, tryptophan, and tyrosine amino acid residues. The instability index indicates the stability of protein in the test tube environment (Guruprasad et al., 1990). 16 proteins scored greater than 40 with a maximum score of 50.89 (Q5NFJ0) in the instability index and thus classified as unstable, whereas 28 proteins were confirmed to be stable, having the lowest score of 12.95 (Q5NHH0). Proteins with a higher aliphatic index usually show higher thermal stability (Prabhu et al., 2020). The aliphatic index values for the EHPs ranged from 48.89 (Q5NG09) to 156.01 (Q5NIL0). Grand average of hydropathicity (GRAVY) value elucidates protein-water interactions (Uddin et al., 2014). Estimated GRAVY value ranged from -0.951 to 1.169 and showed 9 proteins to be hydrophobic and 35 proteins to be hydrophilic.

Table 1. Physicochemical characterization of EHPs predicted by ProtParam server.

DEG Accession ID	UniProt ID	Molecular Weight (Da)	Theoretical pI	Extinction Coefficient ($M^{-1}cm^{-1}$)	Instability Index		Aliphatic Index	Grand Average of Hydropathicity
					Computed	Class		
DEG10520022	Q5NIM1	7459.89	9.60	N/A	29.12	stable	115.62	-0.659
DEG10520024	Q5NIL9	16806.32	4.49	9970	32.10	stable	112.70	0.147
DEG10520029	Q5NIL0	16684.54	9.33	16055	39.65	stable	156.01	1.343
DEG10520054	Q5NIF3	49317.86	4.64	30370	48.37	unstable	83.68	-0.556
DEG10520093	Q5NI67	25388.68	5.84	10430	44.70	unstable	114.40	-0.143

DEG10520101	Q5NI47	21671.51	9.36	20650	24.44	stable	99.14	-0.099
DEG10520164	Q5NHQ3	32986.07	8.99	27515	45.45	unstable	96.77	-0.517
DEG10520190	Q5NHH0	25682.90	8.49	36120	12.95	stable	63.31	-0.411
DEG10520200	Q5NH98	37785.11	7.84	40340	33.29	stable	137.94	0.916
DEG10520209	Q5NH55	19425.83	4.82	8480	48.55	unstable	101.61	-0.621
DEG10520213	Q5NH28	48196.31	9.43	47135	33.52	stable	153.08	1.169
DEG10520222	Q5NGY0	13662.98	8.96	8480	27.83	stable	106.72	-0.120
DEG10520226	Q5NGX6	12658.13	4.21	7575	32.65	stable	81.38	-0.066
DEG10520231	Q5NGT8	30798.68	8.39	32025	45.54	unstable	93.21	-0.296
DEG10520253	Q5NGI1	12702.64	9.77	1490	49.23	unstable	86.54	-0.882
DEG10520257	Q5NGE5	13890.23	5.32	6085	37.46	stable	118.13	0.047
DEG10520259	Q5NGD2	17919.60	8.40	16515	36.40	stable	91.23	-0.331
DEG10520264	Q5NGC2	15059.12	6.82	13535	29.41	stable	84.92	-0.336
DEG10520278	Q5NG80	43500.21	5.21	27515	34.60	stable	101.38	-0.294
DEG10520279	Q5NG68	15472.80	6.73	28085	34.04	stable	93.97	-0.279
DEG10520289	Q5NG32	31664.69	6.96	26610	40.87	unstable	84.44	-0.630
DEG10520290	Q5NG31	23873.96	9.19	27850	38.25	stable	70.38	-0.600
DEG10520293	Q5NG27	10958.36	5.06	4470	43.67	unstable	70.53	-0.499
DEG10520300	Q5NG09	5897.62	6.54	5500	28.68	stable	48.89	-0.854
DEG10520301	Q5NG06	16219.36	5.13	14440	40.40	unstable	101.70	-0.325
DEG10520307	Q5NFX5	28230.58	5.24	17670	24.25	stable	95.89	-0.102
DEG10520310	Q5NFV5	20425.28	8.62	26025	44.55	unstable	75.91	-0.706
DEG10520316	Q5NFU0	63041.21	5.46	67770	25.72	stable	95.78	-0.201
DEG10520319	Q5NFS3	14723.60	4.97	11460	40.68	unstable	101.42	-0.665
DEG10520326	Q5NFP5	16455.99	5.64	13075	42.26	unstable	72.24	-0.310
DEG10520330	Q5NFM6	52390.21	9.32	50115	37.55	stable	135.55	0.924
DEG10520334	Q5NFL5	19981.74	4.42	21555	25.86	stable	88.31	-0.140
DEG10520335	Q5NFK9	12606.75	5.37	15930	35.72	stable	127.06	0.108
DEG10520342	Q5NFJ0	32068.80	9.01	59710	50.89	unstable	77.66	-0.361
DEG10520346	Q5NFG4	45552.35	9.21	42330	33.43	stable	87.13	-0.436
DEG10520356	Q5NFE0	22900.73	5.41	8940	39.50	stable	124.73	-0.069
DEG10520358	Q5NFD6	32458.09	7.77	31860	30.78	stable	92.61	-0.432
DEG10520385	Q5NF39	15679.65	4.49	18450	43.49	unstable	88.33	-0.375
DEG10520406	Q5NEX1	43776.90	5.16	42080	38.03	stable	93.92	-0.188
DEG10520409	Q5NEW8	27805.66	4.91	41285	41.24	unstable	91.12	-0.236
DEG10520430	Q5NEJ5	25608.62	9.53	45295	25.90	stable	100.09	0.751
DEG10520433	Q5NEJ2	7992.21	9.57	1490	46.41	unstable	74.33	-0.951
DEG10520446	Q5NEB7	24862.46	7.63	30160	39.26	stable	92.52	-0.152
DEG10520453	Q5NE86	24579.47	9.28	45170	25.06	stable	126.68	0.803

3.2 Subcellular localization

Prediction of the subcellular localization of proteins is crucial for understanding not only the function of the proteins but also the organization of the cell, especially when experimental methods become unable to provide the full coverage of localization. Cytoplasmic matrix proteins are often considered as potential drug targets, whereas inner and outer membrane proteins are regarded as potential vaccine targets. Subcellular localization was confidently predicted for 33 EHPs out of 44, after comparing the results of 3 different tools (Table 2). Among them, 66.66% (22) proteins were found to be present in the cytoplasm, whereas 21.21% (7) were for the inner membrane, 6.1% (2) were for extracellular and 6.1% (2) were for the periplasm. Signal peptides play a significant role in the transport of proteins to their target locations and provide information regarding cleavage sites. Four EHPs were predicted to have signal peptides, and eight secretomes were found among all the EHPs. Cell secretomes are proteins secreted outside of the cells that help to regulate cell proliferation, cell-to-cell communications, and pathogenesis (Prabhu et al., 2020). Both HMMTOP and TMHMM servers predicted the presence of transmembrane helices in 11 EHPs for each. These transmembrane helices are important for membrane proteins which play crucial parts in the regulation of energy transduction, signaling, and transmembrane transport in cells as well as for drug development as nearly half of the targets used for developing new drugs are membrane proteins (Cuthbertson et al., 2005).

Table 2. Subcellular localization of EHPs predicted by different servers.

DEG Accession No	Sub-Cellular Localization			Signal Peptide (SignalP 4.1)	Secretory protein (SecretomeP)	Transmembrane helices prediction	
	PSORTb	PSLpred	CELLO			HMMTOP	TMHMM
DEG10520022	unknown	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520024	cytoplasmic	inner-membrane	cytoplasmic	No	No	0	0
DEG10520029	cytoplasmic membrane	inner-membrane	inner-membrane	No	No	5	4
DEG10520054	unknown	extracellular	extracellular	No	No	1	1
DEG10520093	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520101	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520164	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520190	periplasmic	periplasmic	periplasmic	Yes	No	0	1
DEG10520200	cytoplasmic membrane	inner-membrane	inner-membrane	No	No	9	10
DEG10520209	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520213	cytoplasmic membrane	inner-membrane	inner-membrane	No	No	12	13
DEG10520222	cytoplasmic membrane	inner-membrane	inner-membrane	No	Yes	1	1
DEG10520226	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520231	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520253	unknown	extracellular	cytoplasmic	No	No	0	0
DEG10520257	unknown	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520259	unknown	cytoplasmic	periplasmic	Yes	No	0	0
DEG10520264	unknown	cytoplasmic	periplasmic	No	No	1	1
DEG10520278	cytoplasmic	extracellular	cytoplasmic	No	No	0	0
DEG10520279	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520289	unknown	extracellular	outer-membrane	Yes	No	0	0
DEG10520290	unknown	periplasmic	extracellular	No	No	2	1
DEG10520293	unknown	extracellular	cytoplasmic	No	Yes	0	0
DEG10520300	unknown	unknown	periplasmic	No	No	0	0
DEG10520301	extracellular	extracellular	cytoplasmic	No	No	0	0
DEG10520307	cytoplasmic	cytoplasmic	cytoplasmic	No	Yes	0	0
DEG10520310	unknown	periplasmic	periplasmic	Yes	No	0	0
DEG10520316	cytoplasmic	cytoplasmic	cytoplasmic	No	Yes	0	0
DEG10520319	unknown	extracellular	cytoplasmic	No	Yes	0	0
DEG10520326	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520330	cytoplasmic membrane	inner-membrane	inner-membrane	No	Yes	11	12
DEG10520334	unknown	periplasmic	cytoplasmic	No	No	0	0
DEG10520335	unknown	inner-membrane	cytoplasmic	No	No	0	0
DEG10520342	outer-membrane	inner-membrane	periplasmic	No	No	2	0
DEG10520346	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520356	unknown	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520358	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520385	cytoplasmic	cytoplasmic	cytoplasmic	No	Yes	0	0
DEG10520406	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520409	cytoplasmic	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520430	cytoplasmic membrane	inner-membrane	inner-membrane	No	No	6	6
DEG10520433	unknown	cytoplasmic	cytoplasmic	No	Yes	0	0
DEG10520446	unknown	cytoplasmic	cytoplasmic	No	No	0	0
DEG10520453	cytoplasmic membrane	inner-membrane	inner-membrane	No	No	6	6

3.3 Virulence factor prediction

Virulence factors generated by pathogenic organisms are considered indispensable for causing diseases in hosts, as these factors help pathogens to evade the defense mechanisms of hosts. Understanding the molecular mechanisms of virulence, therefore, is of great significance in vaccine development and to initiate reverse vaccinology (Chaudhuri and Ramachandran, 2014). By using MP3, VFDB, and VICMpred servers, a total of 15 EHPs have been confidently predicted as virulence factors (marked by '*') out of the 44 EHPs (Table 3).

Table 3. Virulence factor analysis of the EHPs using various *in silico* tools.

DEG Accession ID	UniProt ID	MP3	VFDB	VICMpred
DEG10520022	Q5NIM1	Pathogenic	Non-pathogenic	Cellular process
DEG10520024	Q5NIL9	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520029	Q5NIL0	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520054	**Q5NIF3***	Pathogenic	Pathogenic	Virulence factors
DEG10520093	Q5NI67	Non-pathogenic	Non-pathogenic	Information and storage
DEG10520101	Q5NI47	Non-pathogenic	Non-pathogenic	Metabolism molecule
DEG10520164	**Q5NHQ3***	Pathogenic	Non-pathogenic	Cellular process
DEG10520190	**Q5NHH0***	Pathogenic	Non-pathogenic	Virulence factors
DEG10520200	Q5NH98	Non-pathogenic	Non-pathogenic	Metabolism molecule
DEG10520209	Q5NH55	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520213	**Q5NH28***	Pathogenic	Non-pathogenic	Metabolism molecule
DEG10520222	Q5NGY0	Non-pathogenic	Non-pathogenic	Information and storage
DEG10520226	Q5NGX6	Non-pathogenic	Non-pathogenic	Information and storage
DEG10520231	Q5NGT8	Non-pathogenic	Non-pathogenic	Information and storage
DEG10520253	Q5NGI1	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520257	Q5NGE5	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520259	**Q5NGD2***	Pathogenic	Non-pathogenic	Cellular process
DEG10520264	**Q5NGC2***	Pathogenic	Non-pathogenic	Cellular process
DEG10520278	Q5NG80	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520279	Q5NG68	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520289	**Q5NG32***	Pathogenic	Non-pathogenic	Cellular process

DEG10520290	**Q5NG31***	Pathogenic	Non-pathogenic	Metabolism molecule
DEG10520293	**Q5NG27***	Pathogenic	Non-pathogenic	Information and storage
DEG10520300	**Q5NG09***	Pathogenic	Non-pathogenic	Unknown
DEG10520301	**Q5NG06***	Pathogenic	Non-pathogenic	Virulence factors
DEG10520307	Q5NFX5	Non-pathogenic	Non-pathogenic	Information and storage
DEG10520310	**Q5NFV5***	Pathogenic	Non-pathogenic	Cellular process
DEG10520316	Q5NFU0	Non-pathogenic	Non-pathogenic	Information and storage
DEG10520319	**Q5NFS3***	Pathogenic	Non-pathogenic	Information and storage
DEG10520326	Q5NFP5	Non-pathogenic	Non-pathogenic	Information and storage
DEG10520330	Q5NFM6	Non-pathogenic	Non-pathogenic	Metabolism molecule
DEG10520334	Q5NFL5	Non-pathogenic	Non-pathogenic	Metabolism molecule
DEG10520335	Q5NFK9	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520342	Q5NFJ0	Non-pathogenic	Pathogenic	Metabolism molecule
DEG10520346	Q5NFG4	Non-pathogenic	Non-pathogenic	Metabolism molecule
DEG10520356	Q5NFE0	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520358	**Q5NFD6***	Pathogenic	Non-pathogenic	Information and storage
DEG10520385	Q5NF39	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520406	Q5NEX1	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520409	**Q5NEW8***	Pathogenic	Non-pathogenic	Metabolism molecule
DEG10520430	Q5NEJ5	Non-pathogenic	Non-pathogenic	Metabolism molecule
DEG10520433	Q5NEJ2	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520446	Q5NEB7	Non-pathogenic	Non-pathogenic	Cellular process
DEG10520453	Q5NE86	Non-pathogenic	Non-pathogenic	Virulence factors

3.4 Functional assignment and domain analysis

Functional annotation of proteins helps greatly to strengthen our perception of life at the molecular level, which has tremendous pharmaceutical and biomedical significance. All the 44 EHPs were undertaken for domain analysis to predict possible functions using a variety of *in silico* tools (Supplementary file 1). However, successful annotation was confirmed for 24 EHPs with confidence by comparing results of the various tools used. The annotated proteins involved in diverse biological processes were classified into five categories, such as, enzymes, binding proteins, accessory proteins, cell division proteins, and transporters (Table 4, 5 and Fig. 2).

3.4.1 Enzymes

Bacterial enzymes serve as biocatalysts in an ecofriendly and economical way to regulate different metabolic and cellular activities for growth and pathogenesis of microbes (Nigam, 2013). Of the 22 EHPs, 10 proteins were annotated as enzymes, and four proteins were characterized as transferases (Q5NFX5, Q5NFD6, Q5NEX1, and Q5NEB7). Transferases facilitate or catalyze the transfer of a functional group, except hydrogen, from a donor to an acceptor molecule. Q5NFX5 is predicted to contain a glycine cleavage system (GCS) T which is triggered when glycine concentrations are high. Their subsequent unstable complexes are loosely attached to the inner membrane of the mitochondria, and mutations in this system are connected with glycine encephalopathy (Kikuchi et al., 2008). Q5NFD6 and Q5NEX1 proteins represent S-adenosyl-L-methionine-dependent methyltransferase (SAM Mtase) superfamily. All SAM Mtase possess a structurally conserved SAM-binding domain consisting of a central seven-stranded beta sheet that is flanked by three alpha helices per side of the sheet (Martin and McMillan, 2002). Q5NEB7 is predicted to have a 4'-phosphopantetheinyl transferase superfamily that transfers the 4'-phosphopantetheine (4'-PP) moiety from coenzyme A (CoA) to the invariant serine of pp-binding. This post-translational modification renders holo-ACP capable of acyl group activation via thioesterification of the cysteamine thiol of 4'-PP (Lambalot and Walsh, 1995).

Q5NG68 is found to be a nuclease and to contain YqgF or RNase H-like superfamily which is found primarily in the low-GC gram-positive bacteria holliday junction resolvases (HJRs) and in eukaryote orthologs whereas the function of eukaryotic protein having this domain is less well described (Mahdi et al., 1996). Q5NH55 is an endoribonuclease and predicted as a metalloprotease catalytic domain superfamily, which is mainly a protease enzyme that uses metal for its catalytic activity and also shows metalloendopeptidase activity (Oganesyan et al., 2003). Q5NIL0 is a hydrolase and is predicted as an ATP synthase protein I which utilizes ATP hydrolysis to drive the transport of protons across a membrane. There are several different types of transmembrane ATPases that can differ on the basis of function, structure, and transport of ions (Cross and Muller, 2004). Q5NI47 is a phosphatase which showed haloacid dehydrogenase (HAD) superfamily includes phosphatases, phosphonatases, P-type ATPases, beta-phosphoglucomutases, phosphomannomutases, and dehalogenases, which are involved in a variety of cellular processes ranging from amino acid biosynthesis to detoxification (Koonin and Tatusov, 1994). Q5NFG4 is an oxidoreductase and is found to contain an FAD binding domain that characterized the presence of a nested NADH binding domain which is found in both class I and class II oxidoreductases (Hanukoglu and Gutfinger, 1989). Q5NEJ5 is a haloperoxidase and is predicted to contain phosphatidic acid phosphatase type 2 superfamily (PAP2). The dephosphorylation of phosphatidate is catalyzed by PAP2 enzymes and produces diacylglycerol with inorganic phosphate. In eukaryotic cells, especially in the synthesis of phospholipids and triacylglycerol, PAP plays a key role through its product diacylglycerol, and it also produces and/or degrades lipid-signalling molecules that are related to phosphatidate (Littlechild et al., 2002).

Table 4. List of functionally annotated EHPs.

DEG Accession ID	UniProt ID	Protein function	Class
DEG10520022	Q5NIM1	Efficient ubiquinone biosynthesis in aerobic conditions, form complex with UbiJ	Binding protein
DEG10520024	Q5NIL9	Ribosomal small subunit biogenesis, efficient production of translationally competent ribosomes	Accessory protein
DEG10520029	Q5NIL0	Driving transport of protons across a membrane by ATP hydrolysis	Enzyme (Hydrolase)
DEG10520054	Q5NIF3	Binding peptidoglycan in bacteria and chitin in eukaryotes	Binding protein
DEG10520101	Q5NI47	Dephosphorylates phosphatidylglycerolphosphate in cardiolipin biosynthesis	Enzyme (Phosphatase)
DEG10520209	Q5NH55	rRNA processing, metalloendopeptidase activity	Enzyme (Endoribonuclease)
DEG10520222	Q5NGY0	Cell division septum formation as integral component of cell membrane	Cell division protein
DEG10520226	Q5NGX6	Cellular metabolism, iron-sulfur cluster binding	Binding protein
DEG10520257	Q5NGE5	Assembly of mitochondrial NADH:ubiquinone oxidoreductase complex.	Accessory protein
DEG10520278	Q5NG80	Iron-sulfur cluster assembly	Binding protein
DEG10520279	Q5NG68	Nucleobase-containing compound metabolic process, rRNA processing	Enzyme (Nuclease)
DEG10520289	Q5NG32	Organic solvent tolerance factor	Accessory protein
DEG10520290	Q5NG31	Lipopolysaccharide transporting to outer membrane	Transporter
DEG10520307	Q5NFX5	Catabolism of glycine in eukaryotes	Enzyme (Aminomethyltransferase)
DEG10520316	Q5NFU0	ATP binding activity in biosynthesis of peptidoglycan	Binding protein
DEG10520326	Q5NFP5	Coenzyme Q biosynthetic process, cellular respiration	Binding protein
DEG10520342	Q5NFJ0	Outer membrane protein assembly activity	Binding protein
DEG10520346	Q5NFG4	FAD-dependent pyridine nucleotide reductase activity	Enzyme (Oxidoreductase)
DEG10520356	Q5NFE0	Ubiquinone biosynthetic process from chorismate	Binding protein
DEG10520358	Q5NFD6	Transferring methyl group from donor to acceptor	Enzyme (Methyltransferase)

DEG10520385	Q5NF39	Iron-sulfur cluster assembly	Binding protein
DEG10520406	Q5NEX1	Mitochondrial complex I activity	Enzyme (Methyltransferase)
DEG10520430	Q5NEJ5	Synthesis of phospholipids and triacylglycerol	Enzyme (Haloperoxidase)
DEG10520446	Q5NEB7	Magnesium ion binding activity	Enzyme (Transferase)

3.4.2 Binding proteins

Nine EHPs were predicted as binding proteins. Among them, Q5NIM1 is predicted as an Ubiquinone biosynthesis accessory factor (UbiK), which is required for effective biosynthesis of ubiquinone under aerobic conditions as it forms a complex with UQ biogenesis factor UbiJ (Loiseau et al., 2017). Q5NIF3 is found to contain a lysine motif domain that is engaged with the binding process of peptidoglycan in bacteria and chitin in eukaryotes (Joris et al., 1992). This domain drives the signaling for distinct plant-bacteria recognition in bacterial pathogenesis (Spaink, 2004). Q5NGX6 was found to be involved in iron-sulfur cluster biogenesis and to be a part of the HesB superfamily. The HesB gene is expressed only under the condition of nitrogen fixation (Huang et al., 1999) and is found in a variety of species ranging from *Haemophilus influenzae* to *Homo sapiens*, which suggests their diversity and participation in fundamental cellular processes (Hwang et al., 1996). Q5NG80 is predicted to contain a domain of SUF machinery involved in the biogenesis of iron-sulfur clusters. This SUF system acts as an alternative pathway to the ISC system that functions under oxidative stress and iron starvation conditions (Pérard and Ollagnier, 2018).

Q5NFU0 is found to contain a Mur-like catalytic domain superfamily. Mur ligases play a critical role in the intercellular biogenesis of peptidoglycan in bacteria (Sink et al., 2016). The superfamily represents the central domain of all four Mur enzymes. Q5NFP5 is predicted to have a START like superfamily domain, which is involved in lipid binding in StAR, HD-ZIP and signaling proteins (Ponting and Aravind, 1999). StAR proteins are required for acute regulation of steroidogenesis and are expressed in the absence of hormone stimulation to drive the steroid production process (Clark et al., 1994). Q5NFJ0 was found to have a BamD-like outer membrane lipoprotein domain. BamD/YfiO is part of the beta-barrel assembly machinery which is required for the insertion and folding of outer membrane proteins into the outer membrane of gram-negative bacteria (Kim et al., 2011). As the only BAM lipoprotein required for viability, BamD contains five tetratricopeptide repeats which are suggested to be involved in the binding with other BAM components (Dong et al., 2012). Q5NFE0 is predicted to contain ubiquinone biosynthesis accessory factor (UbiJ), which is associated with the biosynthesis of ubiquinone under aerobic conditions (Aussel et al., 2014). This promotes binding of hydrophobic ubiquinone biosynthetic intermediates via the SCP2 domain, which is essential for the Ubi complex stability (Chehade et al., 2019). Q5NF39 was predicted to have a SufE-like domain, which is associated with iron-sulfur metabolism. The domain has a strong structural similarity to IscU and the sulfur-acceptor site incorporates cysteine residues to mediate iron-sulfur cluster assembly in IscU (Goldsmith-Fischman et al., 2004).

3.4.3 Accessory proteins

Three accessory proteins were predicted, of which Q5NIL9 is found to contain ribosome maturation factor (RimP), which induces maturation of the 30S ribosomal subunit and is essential for the effective development of translationally competent ribosomes (Nord et al., 2009). Q5NGE5 is predicted to contain an MTH-938-like superfamily domain. MTH938 is a hypothetical protein encoded by *Methanobacterium thermoautotrophicum* (Das et al., 2001) and the superfamily contains NDUFAF3 essential factor for the assembly of mitochondrial NADH: ubiquinone oxidoreductase complex (Saada et al., 2009). Q5NG32 is found to have an organic solvent tolerance-like domain, which is found in a number of bacterial proteins, including the lipopolysaccharide assembly protein lptD (Aono et al., 1994).

3.4.4 Cell division protein

Knowledge of the mechanism and function of cell division proteins is essential to know about the novel targets for durg discovery. Of all the EHPs, only one is predicted to be involved in the cell division process. Q5NGY0 is predicted to have cell devision septum formation superfamily. It acts as an integral component of the membrane and helps in the cell division process. The protein is small in size, highly divergent and low in complexity (Sievers and Errington, 2000).

3.4.5 Transporters

Transporters constitute approximately 10% of most proteomes and play a crucial role in the translocation of solutes across membranes. Their function and dysfunction have profound implication in the import and export of substances such as metabolites, nutrients, amino acids etc. and for that they are profusely utilized in the pharmacotherapy (Quick and Javitch, 2007). Q5NG31 is found to contain an LptC-like lipopolysaccharide assembly protein superfamily. LptC is involved in the assembly of lipopolysaccharides on the outer membrane of gram-negative organisms. The lipopolysaccharide is transported from its source of origin to the outer membrane through a transport machinery consisting of LptA, LptB, LptC, LptD, and LptE. The LptC is situated in the inner membrane portion of the intermembrane space (Sperandeo et al., 2008).

Table 5. List of conserved domains identified from the EHPs.

UniProt ID	Conserved Domain
Q5NIM1	Ubiquinone biosynthesis accessory factor (UbiK)
Q5NIL9	Ribosome maturation factor (RimP) N terminal domain
Q5NIL0	ATP synthase protein I
Q5NIF3	Lysine motif domain

Q5NI47	Haloacid dehydrogenase (HAD) like superfamily
Q5NH55	Metalloprotease catalytic domain superfamily
Q5NGY0	Cell division septum formation superfamily (FtsL)
Q5NGX6	HesB like superfamily
Q5NGE5	MTH938 (Methanobacterium thermoautotrophicum) like superfamily
Q5NG80	SUF system FeS cluster assembly (SUFBD) superfamily
Q5NG68	YqgF or RNase H like superfamily
Q5NG32	Organic solvent tolerance like N terminal domain
Q5NG31	Lipopolysaccharide assembly protein (LptC) superfamily
Q5NFX5	Glycine cleavage system T protein domain
Q5NFU0	Mur like catalytic domain superfamily
Q5NFP5	START like domain superfamily
Q5NFJ0	Tricopeptide like helical domain superfamily
Q5NFG4	FAD or NAD(P) binding domain superfamily
Q5NFE0	Ubiquinone biosynthesis accessory factor (UbiJ)
Q5NFD6	S-adenosyl-L-methionine-dependent methyltransferase superfamily
Q5NF39	Fe-S metabolism associated domain
Q5NEX1	S-adenosyl-L-methionine-dependent methyltransferase superfamily
Q5NEJ5	Phosphatidic acid phosphatase type 2 superfamily
Q5NEB7	4'-phosphopantetheinyl transferase domain superfamily

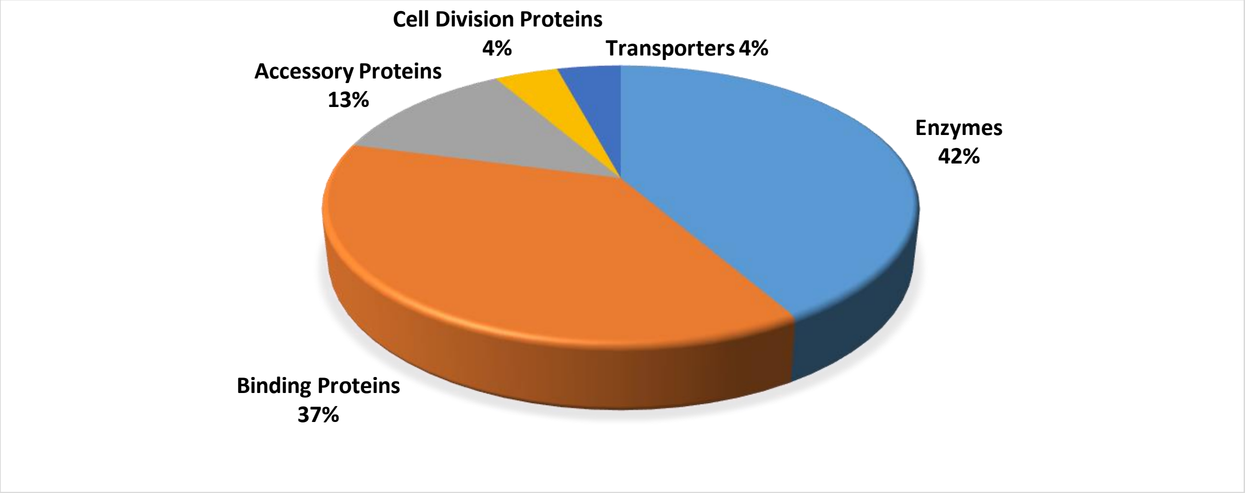

Fig. 2. Different classes of functionally annotated EHPs.

3.4.6 Protein folding pattern recognition

The folding of proteins is a vital cellular process that affects the functionality of the protein significantly. Different types of folding patterns are responsible for variations in the functionality of proteins. Improper folding of proteins causes deviation from their regular functions, leading to inactive or toxic proteins and products formation that malfunction and contribute to pathogenicity. Using the PFP-FunDSeqE server, folding patterns were predicted for all the 44 EHPs, and 13 different types of foldings were recorded (Table 6 and Fig. 3).

Table 6. Folding patterns of the EHPs.

Fold type	UniProt ID
OB-fold	Q5NIM1, Q5NFE0, Q5NI67, Q5NHQ3, Q5NFV5, Q5NFK9, Q5NE86
TIM - Barrel	Q5NIL9, Q5NH55, Q5NG68, Q5NFX5, Q5NEB7, Q5NGC2
Cupredoxins	Q5NIL0, Q5NH28
Immunoglobulin like	Q5NIF3, Q5NG80, Q5NG32, Q5NG31, Q5NHH0, Q5NGT8, Q5NGD2, Q5NG27, Q5NG09, Q5NG06
Ferredoxin like	Q5NI47, Q5NGE5, Q5NF39
DNA binding 3-helical bundle	Q5NGY0, Q5NEJ2
Thioredoxin like	Q5NGX6, Q5NFJ0
NAD(P) binding Rossmann fold	Q5NFU0
Beta-trefoil	Q5NFP5, Q5NFD6, Q5NEX1, Q5NFM6
FAD (also NAD) –binding motif	Q5NFG4
Globin like	Q5NEJ5
4- helical cytokines	Q5NH98, Q5NGI1, Q5NFS3, Q5NEW8
EF-hand	Q5NFL5

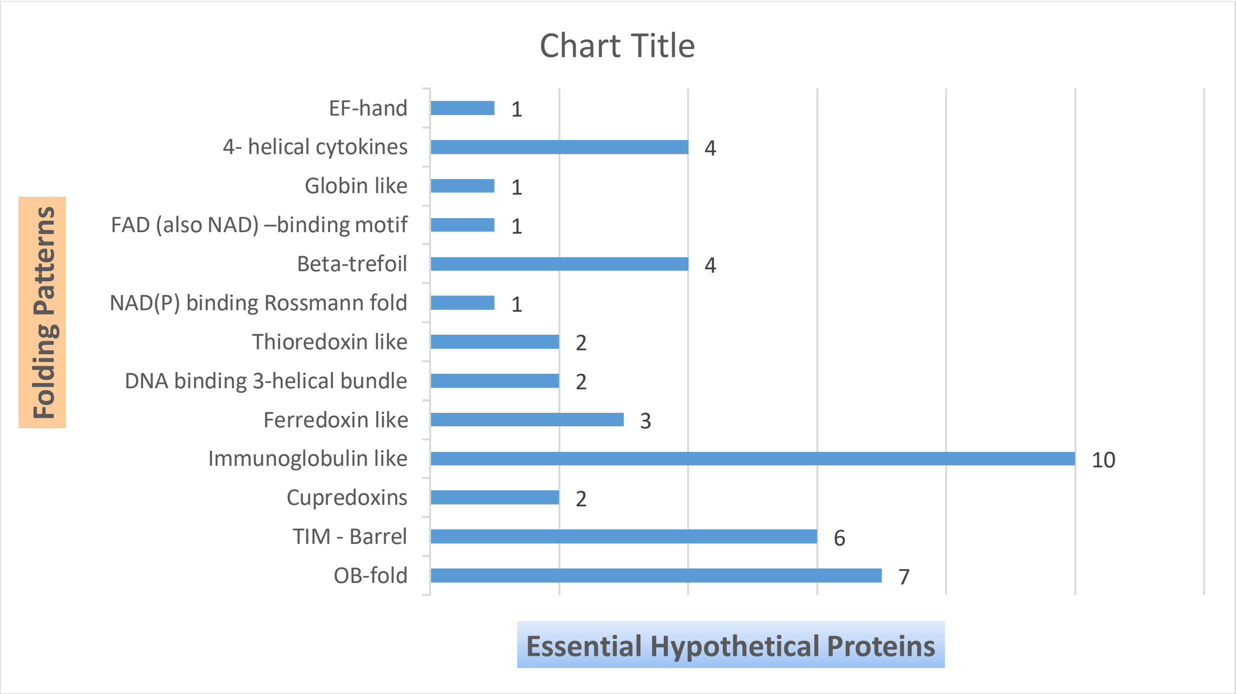

Fig. 3. Folding patterns of the EHPs.

3.5 Evaluation of performance

ROC analysis showed a high degree of reliability and credibility for the seven *in silico* tools and servers (Raj et al., 2017). The confidence of prediction for each EHP was considered high when the same result was predicted by three or more tools. For all the tools used in functional annotation, average values were recorded as 99%, 98.66%, and 100% for accuracy, sensitivity, and specificity, respectively (Table 7 and Fig. 4).

Table 7. ROC curve assessment analysis.

SL. No.	Tools/Servers	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC area
1.	CDD	100	100	100	1
2.	Pfam	100	100	100	1
3.	InterProScan	100	100	100	1
4.	SMART	100	100	100	1
5.	PANTHER	97.7	96.7	100	0.98
6.	MOTIF	95.3	93.9	100	0.97
7.	CATH	100	100	100	1
	Average	99	98.66	100	0.993

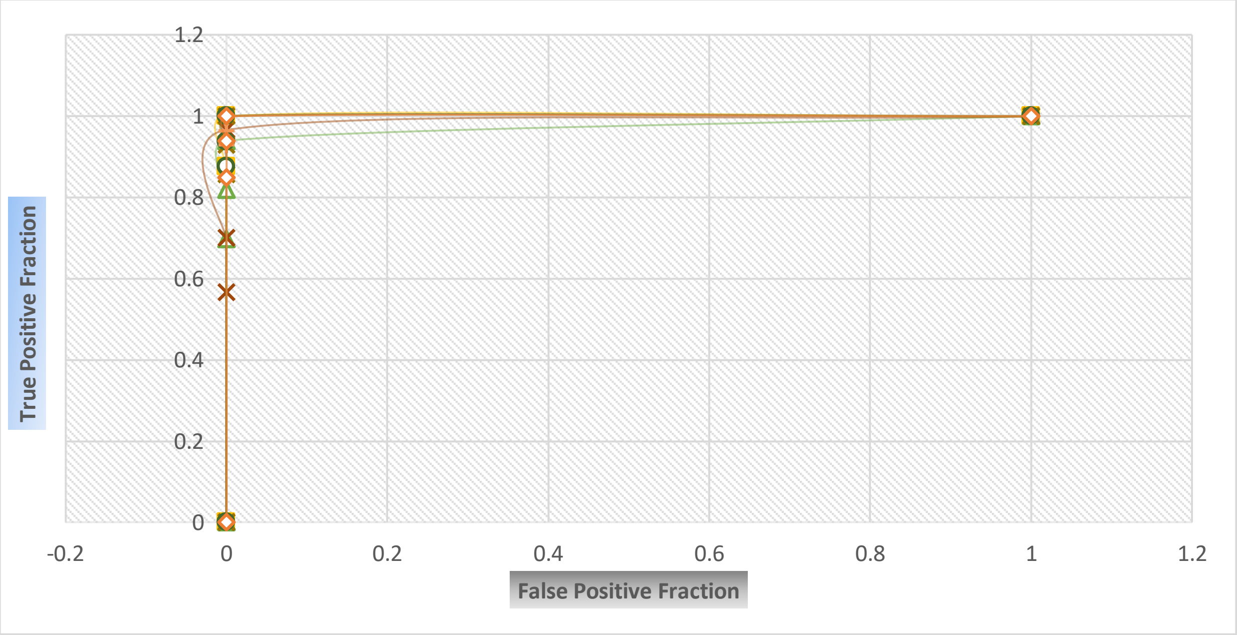

Fig. 4. Statistical analysis of the bioinformatics tools used for functional annotations of EHPs.

3.6 Host non-homology analysis

The BLASTp search against the non-redundant database of the human proteome, setting an e-value threshold of 0.0001, revealed 20 non-homologous EHPs out of 24 functionally annotated EHPs. For an ideal drug target, it should not have any close homologs in the human proteome so that it can minimize the unwanted cross reactivity of a potential drug with the host proteins (Prava et al., 2018). Therefore, these non-homologous proteins are pathogen specific, which means solely present in the pathogen, and can be further studied to find an ideal drug target.

Table 8. Non-homology analysis for the 24 shortlisted, functionally annotated EHPs.

DEG Accession ID	UniProt ID	BLASTp against *H. sapiens*	Pathogen Specificity
DEG10520022	Q5NIM1	Non-homologous	Yes
DEG10520024	Q5NIL9	Non-homologous	Yes
DEG10520029	Q5NIL0	Non-homologous	Yes
DEG10520054	Q5NIF3	Non-homologous	Yes
DEG10520101	Q5NI47	Non-homologous	Yes
DEG10520209	**Q5NH55***	Homologous	No*
DEG10520222	Q5NGY0	Non-homologous	Yes
DEG10520226	**Q5NGX6***	Homologous	No*
DEG10520257	**Q5NGE5***	Homologous	No*
DEG10520278	Q5NG80	Non-homologous	Yes
DEG10520279	Q5NG68	Non-homologous	Yes
DEG10520289	Q5NG32	Non-homologous	Yes
DEG10520290	Q5NG31	Non-homologous	Yes
DEG10520307	Q5NFX5	Non-homologous	Yes
DEG10520316	Q5NFU0	Non-homologous	Yes
DEG10520326	**Q5NFP5***	Homologous	No*
DEG10520342	Q5NFJ0	Non-homologous	Yes
DEG10520346	Q5NFG4	Non-homologous	Yes
DEG10520356	Q5NFE0	Non-homologous	Yes
DEG10520358	Q5NFD6	Non-homologous	Yes
DEG10520385	Q5NF39	Non-homologous	Yes
DEG10520406	Q5NEX1	Non-homologous	Yes
DEG10520430	Q5NEJ5	Non-homologous	Yes
DEG10520446	Q5NEB7	Non-homologous	Yes

3.7 Secondary structure prediction

The secondary structure of proteins can help to predict the tertiary structure and also play an important role in determining the folding pattern and function of proteins. The PSIPRED server predicted secondary structures for all the 20 non-homologous, pathogen specific proteins using different parameters such as alpha helix, beta-strand, coiled structure, etc., which are given in the supplementary file 2.

3.8 Tertiary structure prediction and quality assessment

Homology modeling is a very useful and essential technique for determining protein tertiary structure using known protein or template structure and amino acid sequence data. However, if the sequence identity is below 30%, the model will not provide suitable efficiency in structure determination (Xiang, 2006; Gromiha et al., 2018). After searching for suitable templates in the PDB database using BLASTp, 10 non-homologous proteins out of 20 EHPs showed sequence identity greater than 30% (Table 9). Therefore, those 10 EHPs were used to build tertiary structures using two different servers, such as SWISS-MODEL and Phyre2.

Table 9. Tertiary model assessment for the 10 pathogen specific proteins before model refinement.

UniProt ID	Seq. Iden. (PDB database) %	Query Cov. (PDB database) %	SWISS-MODEL			Phyre2		
			Rama favored %	Verify3D % (Fail-F, Pass-P)	ERRAT quality factor	Rama favored %	Verify3D % (Fail-F, Pass-P)	ERRAT quality factor
Q5NEB7	33.62	52	90.4	74.50 (F)	85.64	81.6	56.07 (F)	40.5
Q5NFU0	31.84	66	85.1	87.92 (P)	82.89	83.6	70.94 (F)	35.44
Q5NEX1	33.95	98	89.7	94.41 (P)	82.69	88.6	78.84 (F)	52.16
Q5NF39	37.61	83	94.3	91.85 (P)	93.6	90.5	86.96 (P)	70.76
Q5NFG4	51.35	9	86.4	76.04 (F)	82.75	83.2	71.57 (F)	50
Q5NFJ0	30.08	89	91.1	82.30 (P)	84.86	89.8	62.04 (F)	69.32
Q5NEJ5	31.17	33	91.0	19.49 (F)	96.36	90.2	32.88 (F)	51.40
Q5NFX5	43.18	17	84.3	68.91 (F)	78.34	84.1	50.40 (F)	62.5
Q5NG68	37.23	100	91.9	69.85 (F)	100	92.7	72.06 (F)	78.125
Q5NIL9	38.16	100	86.4	62.25 (F)	86.71	85.0	77.63 (F)	71.53

After generating the models, PROCHECK Ramachandran plot showed residues in the most favored regions ≥ 90% for only five proteins generated by SWISS-MODEL and for only three proteins generated by Phyre2. In the Verify3D test, four proteins from SWISS-MODEL passed the test whereas only one protein from Phyre2 passed the test. Similar trend was observed in the ERRAT server test. Therefore, SWISS-MODEL was found to be more efficient than Phyre2 according to the quality assessment results. Therefore, the PDB files for all the proteins generated by SWISS-MODEL were further submitted to the GalaxyRefine server for model refinement.

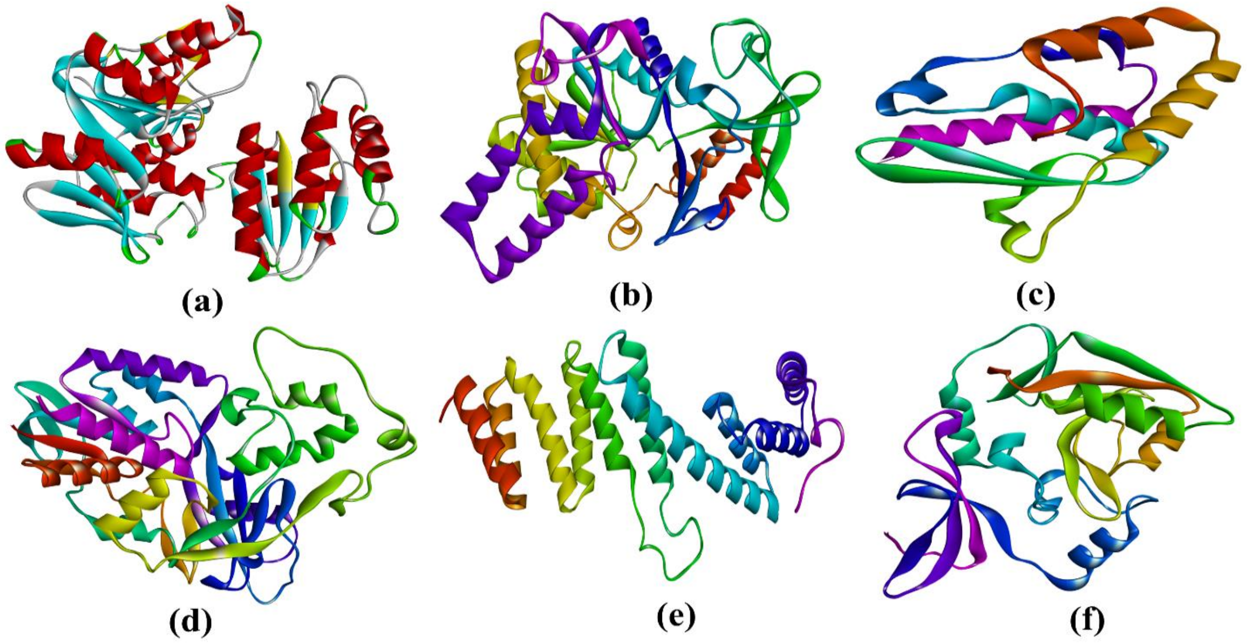

Fig. 5. Three dimensional structures of the best scoring EHPs-Q5NFU0 (a), Q5NEX1 (b), Q5NF39 (c), Q5NFG4 (d), Q5NFJ0 (e) and Q5NFX5 (f).

GalaxyRefine performs repeated structure perturbation and subsequent overall structural relaxation by molecular dynamics simulation (Heo et al., 2013). After energy minimization and the necessary refinement processes, all 10 models were reevaluated in the quality assessment parameters (Table 10). Significant improvements were observed in the quality of the refined models as the scores increased for all the quality assessment servers for all the proteins. All the models showed residues in the most favored regions ≥ 90% except Q5NFU0 and Q5NFX5, which scored farily good scores of 88.90% and 89.20%, respectively. Q5NFU0, Q5NEX1, Q5NF39, Q5NFG4, Q5NFJ0 and Q5NFX5 showed best scores than others in all the three quality assessment servers and were visualized (Fig. 5).

Table 10. Tertiary model assessment for the 10 pathogen specific proteins after model refinement.

UniProt ID	**Swiss Model**		
	Rama favored %	**Verify3D % (Fail-F, Pass-P)**	**ERRAT quality factor**
Q5NEB7	95.7	70.00 (F)	84.26
Q5NFU0*	88.90	84.83 (P)	95.78
Q5NEX1*	92.30	94.15 (P)	90.66
Q5NF39*	96.70	90.37 (P)	98.36
Q5NFG4*	91.50	85.16 (P)	85.79
Q5NFJ0*	94.10	84.96 (P)	94.03
Q5NEJ5	95.00	33.90 (F)	96.29
Q5NFX5*	89.20	86.13 (P)	84.97
Q5NG68	92.70	79.41 (F)	96.09
Q5NIL9	91.70	65.56 (F)	90.21

3.9 Protein-protein interaction network analysis

Protein-protein interactions (PPIs) have significant impacts on biological activities, cellular functions, drug repurposing, and drug target discovery. The STRING server integrates all known and predicted associations between proteins, including both functional associations and physical interactions (Szklarczyk et al., 2021). Protein-protein interactions were predicted for 20 pathogen specific EHPs, of which the best 2 were selected based on their STRING score (Table 11). The STRING scores ranged from 0.588 to 0.999. Three EHPs namely, Q5NG80, Q5NG31, and Q5NFJ0, showed the highest scores, whereas Q5NEJ5 showed the lowest scores (Fig. 6).

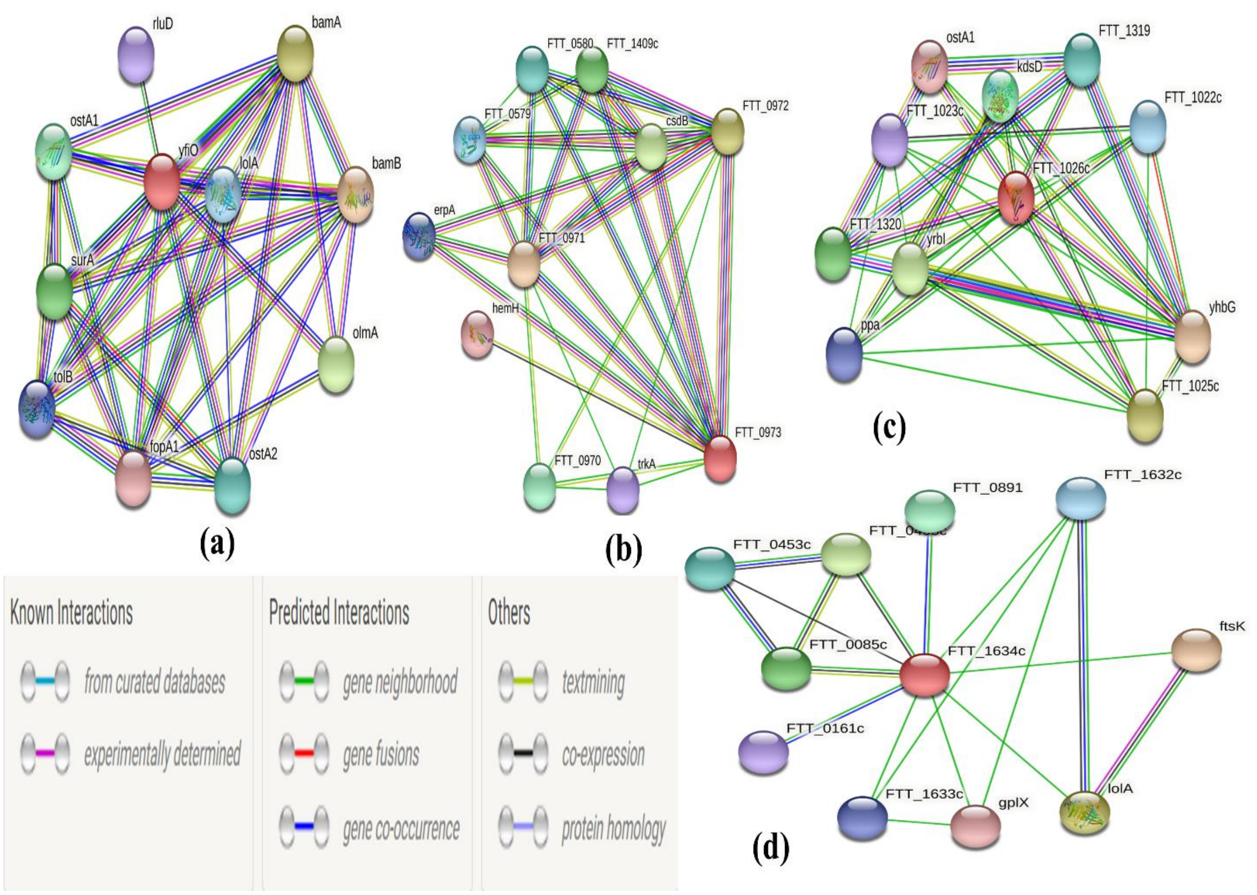

Fig. 6. Protein-protein interaction network for some selected EHPs (Q5NFJ0, Q5NG80, Q5NG31 and Q5NEJ5 for a, b, c and d) based on STRING scores.

Table. 11. Protein-protein interactions predicted by STRING server.

DEG Accession ID	UniProt ID	Interacting Protein	Protein Function	STRING SCORE
DEG10520022	Q5NIM1	nuoN	Proton antiporter, cellular process	0.624
		nuoM	Proton antiporter, cellular process	0.610
DEG10520024	Q5NIL9	nusA	Transcription termination and antitermination	0.998
		infB	Initiation of protein synthesis, hydrolysis of GTP	0.938
DEG10520029	Q5NIL0	atpB	Translocation of protons across the membrane	0.674
		FTT0056_c	Major facilitator transporter superfamily of membrane proteins	0.606
DEG10520054	Q5NIF3	ftsY	Insertion of membrane protein into cytoplasmic membrane	0.852
		prfC	Facilitates formation of ribosomal termination complexes	0.618
DEG10520101	Q5NI47	FTT_0244	Similar to DNA/RNA helicase	0.690
		FTT_0243	Regulation of chromosome condensation	0.642
DEG10520222	Q5NGY0	ftsB	Essential cell division protein	0.982
		ftsQ	Similar to cell division protein	0.973
DEG10520278	**Q5NG80***	FTT_0971	Iron-Sulfur assembly, binding protein	0.999
		FTT_0972	ATP binding cassette (ABC) transporter protein	0.999
DEG10520279	Q5NG68	FTT_0985	Containing domain of unknown function	0.954
		mutT	Hydrolase enzyme, belongs to Nudix superfamily	0.700

DEG10520289	Q5NG32	FTT_1026c	Lipopolysaccharide assembly in the outer membrane	0.974
		yhbG	Similar to ABC transporter ATP binding protein	0.892
DEG10520290	**Q5NG31***	yhbG	Similar to ABC transporter ATP binding protein	0.999
		FTT_1025c	OstA like family protein	0.974
DEG10520307	Q5NFX5	rep	DNA dependent ATPase involved in DNA replication	0.845
		FTT_1086c	Nucleic acid binding with OB-fold like structure	0.668
DEG10520316	Q5NFU0	cphA	Pyrimidine and arginine biosynthesis by catalyzing the synthesis of carbamoyl phosphate	0.870
		cphB	Serine peptidase enzyme with proteolytic activity	0.862
DEG10520342	**Q5NFJ0***	bamB	Assembly and insertion of beta-barrel proteins into the outer membrane	0.999
		bamA	Assembly and insertion of beta-barrel proteins into the outer membrane	0.999
DEG10520346	Q5NFG4	mltA	Murein degrading enzyme, cell division process	0.880
		hemD	Uroporphyrinogen III and heme biosynthesis	0.710
DEG10520356	Q5NFE0	ubiB	Kinase regulator in ubiquinone biosynthesis	0.871
		ubiE	Methyltransferase involved in the biosynthesis of ubiquinone	0.853
DEG10520358	Q5NFD6	bioD	Biosynthesis of cobalamin from uroporphyrinogen III	0.969
		bioB	Biotin and thiamin biosynthesis	0.909
DEG10520385	Q5NF39	csdB	Aminotransferase enzyme catalyzes the removal of sulfur and selenium atoms	0.975
		FTT_0971	Iron-sulfur assembly protein shows binding activity	0.878
DEG10520406	Q5NEX1	nuoI	Proton translocation acitivity across the membrane	0.912
		nuoC	Proton translocation activity across the membrane	0.902
DEG10520430	**Q5NEJ5***	ftsK	DNA translocase enzyme	0.588
		lolA	Translocation of lipoprotein from inner to outer membrane	0.588
DEG10520446	Q5NEB7	coaE	Aminotransferase, catalyzes phosphorylation to form coenzyme A	0.925
		tolC	Outer membrane efflux protein, export of antibiotics and toxic compounds from the cell	0.851

4. Conclusions

Essential hypothetical proteins (EHPs) derived from essential genes of the pathogenic *Francisella tularensis* Schu S4 strain have great implications in comparative and functional genomics investigations as these uncharacterized hypothetical proteins impede the search for potential drug targets. The present *in silico* study emphasized on the physicochemical, functional and structural annotations of the EHPs. 24 out of 44 EHPs were functionally annotated successfully and supported by statistical quality assessment parameters whereas functionality of rest of the EHPs could not be predicted effectively due to the insufficient sequence resemblance in the database. Subcellular localization, virulence factor prediction and non-homology analysis will be effective for host-pathogen interaction study as well as drug/vaccine development. The secondary and tertiary structure prediction provides valuable insights regarding the spatial positioning of amino acids in the proteins to find the potential binding sites for drugs. Therefore *in vitro* and *in vivo* experimental studies of these EHPs should be carried out to draw effective conclusions in the rapidly advancing field of drug discovery.

Acknowledgments

The author is grateful to Md. Arif Khan for his cordial support during data analysis.

Conflicts of Interest

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

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Aono, R., Negishi, T., & Nakajima, H. (1994). Cloning of organic solvent tolerance gene ostA that determines n-hexane tolerance level in Escherichia coli. *Applied and environmental microbiology*, *60*(12), 4624-4626.

Aussel, L., Loiseau, L., Hajj Chehade, M., Pocachard, B., Fontecave, M., Pierrel, F., & Barras, F. (2014). ubiJ, a new gene required for aerobic growth and proliferation in macrophage, is involved in coenzyme Q biosynthesis in Escherichia coli and Salmonella enterica serovar Typhimurium. *Journal of bacteriology*, *196*(1), 70-79.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159.

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

Bendtsen, J. D., Kiemer, L., Fausbøll, A., & Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC microbiology*, *5*(1), 1-13.

Buchan, D. W., & Jones, D. T. (2019). The PSIPRED protein analysis workbench: 20 years on. *Nucleic acids research*, *47*(W1), W402-W407.

BIOVIA (2021). Dassault Systèmes. in Discovery Studio Visualizer. v.21.10.20298.San Diego, CA: Dassault Systèmes.

Colovos, C., & Yeates, T. (1993). ERRAT: an empirical atom-based method for validating protein structures. *Protein Sci*, *2*(9), 1511-1519.

Clark, B. J., Wells, J., King, S. R., & Stocco, D. M. (1994). The purification, cloning, and expression of a novel luteinizing hormone-induced mitochondrial protein in MA-10 mouse Leydig tumor cells. Characterization of the steroidogenic acute regulatory protein (StAR). *Journal of Biological Chemistry*, *269*(45), 28314-28322.

Cross, R. L., & Müller, V. (2004). The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio. *FEBS letters*, *576*(1-2), 1-4.

Cuthbertson, J. M., Doyle, D. A., & Sansom, M. S. (2005). Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Engineering Design and Selection*, *18*(6), 295-308.

Conlan J.W., & Oyston, P.C. (2007). Vaccines against *Francisella tularensis*. *Annals of the New York Academy of Sciences* 1105: 325–350.

Chase, J. C., Celli, J., & Bosio, C. M. (2009). Direct and indirect impairment of human dendritic cell function by virulent *Francisella tularensis* Schu S4. *Infection and immunity*, *77*(1), 180-195.

Chaudhuri, R., & Ramachandran, S. (2014). Prediction of virulence factors using bioinformatics approaches. *Methods in molecular biology*, *1184*, 389–400.

Chehade, M. H., Pelosi, L., Fyfe, C. D., Loiseau, L., Rascalou, B., Brugière, S., ... & Pierrel, F. (2019). A soluble metabolon synthesizes the isoprenoid lipid ubiquinone. *Cell chemical biology*, *26*(4), 482-492.

Dennis, D., Inglesby, T., Henderson, D., Bartlett, J., Ascher, M., & Eitzen, E. et al. (2001). Tularemia as a Biological Weapon. *JAMA*, *285*(21), 2763. doi: 10.1001/jama.285.21.2763

Das, K., Xiao, R., Wahlberg, E., Hsu, F., Arrowsmith, C. H., Montelione, G. T., & Arnold, E. (2001). X‐ray crystal structure of MTH938 from Methanobacterium thermoautotrophicum at 2.2 Å resolution reveals a novel tertiary protein fold. *Proteins: Structure, Function, and Bioinformatics*, *45*(4), 486-488.

Dong, C., Hou, H. F., Yang, X., Shen, Y. Q., & Dong, Y. H. (2012). Structure of Escherichia coli BamD and its functional implications in outer membrane protein assembly. *Acta Crystallographica Section D: Biological Crystallography*, *68*(2), 95-101.

Eng., J., ROC Analysis: Web-based Calculator for ROC Curves, (n.d.). from http://www.jrocfit.org

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

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., ... & Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, *44*(D1), D279-D285.

Guruprasad, K., Reddy, B. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection*, *4*(2), 155-161.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, *31*(13), 3784-3788.

Goldsmith-Fischman, S., Kuzin, A., Edstrom, W. C., Benach, J., Shastry, R., Xiao, R., ... & Hunt, J. F. (2004). The SufE sulfur-acceptor protein contains a conserved core structure that mediates interdomain interactions in a variety of redox protein complexes. *Journal of molecular biology*, *344*(2), 549-565.

Gupta, A., Kapil, R., Dhakan, D. B., & Sharma, V. K. (2014). MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PloS one*, *9*(4), e93907.

Gromiha, M., Nagarajan, R. & Selvaraj, S. (2018). Protein structural bioinformatics: An overview. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1-3: 445-459.

Hanukoglu, I., & Gutfinger, T. (1989). cDNA sequence of adrenodoxin reductase: Identification of NADP-binding sites in oxidoreductases. *European journal of biochemistry*, *180*(2), 479-484.

Hwang, D. M., Dempsey, A., Tan, K. T., & Liew, C. C. (1996). A modular domain of NifU, a nitrogen fixation cluster protein, is highly conserved in evolution. *Journal of molecular evolution*, *43*(5), 536-540.

Huang, T. C., Lin, R. F., Chu, M. K., & Chen, H. M. (1999). Organization and expression of nitrogen-fixation genes in the aerobic nitrogen-fixing unicellular cyanobacterium *Synechococcus* sp. strain RF-1. *Microbiology*, *145*(3), 743-753.

Heo, L., Park, H., & Seok, C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic acids research*, *41*(W1), W384-W388.

Joris, B., Englebert, S., Chu, C. P., Kariyama, R., Daneo-Moore, L., Shockman, G. D., & Ghuysen, J. M. (1992). Modular design of the Enterococcus hirae muramidase-2 and Streptococcus faecalis autolysin. *FEMS microbiology letters*, *91*(3), 257-264.

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

Koonin, E. V., & Tatusov, R. L. (1994). Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity: application of an iterative approach to database search. *Journal of molecular biology*, *244*(1), 125-132.

Keim, P., Johansson, A., & Wagner, D. M. (2007). Molecular epidemiology, evolution, and ecology of Francisella. *Annals of the New York Academy of Sciences*, *1105*(1), 30-66.

Kikuchi, G., Motokawa, Y., Yoshida, T., & Hiraga, K. (2008). Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proceedings of the Japan Academy. Series B, Physical and biological sciences*, *84*(7), 246–263. https://doi.org/10.2183/pjab.84.246

Kadzhaev, K., Zingmark, C., Golovliov, I., Bolanowski, M., Shen, H., Conlan, W., & Sjöstedt, A. (2009). Identification of genes contributing to the virulence of *Francisella tularensis* SCHU S4 in a mouse intradermal infection model. *PLoS One*, *4*(5), e5463.

Kim, K. H., Aulakh, S., & Paetzel, M. (2011). Crystal structure of β-barrel assembly machinery BamCD protein complex. *Journal of Biological Chemistry*, *286*(45), 39116-39121.

Kingry, L. C., & Petersen, J. M. (2014). Comparative review of *Francisella tularensis* and *Francisella novicida*. *Frontiers in cellular and infection microbiology*, *4*, 35.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, *10*(6), 845-858.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic acids research*, *47*(D1), D590-D595.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, *26*(2), 283-291.

Lambalot, R. H., & Walsh, C. T. (1995). Cloning, Overproduction, and Characterization of the Escherichia coli Holo-acyl Carrier Protein Synthase. *Journal of Biological Chemistry*, *270*(42), 24658-24661.

Littlechild, J., Garcia-Rodriguez, E., Dalby, A., & Isupov, M. (2002). Structural and functional comparisons between vanadium haloperoxidase and acid phosphatase enzymes. *Journal of Molecular Recognition*, *15*(5), 291-296.

Letunic, I., Doerks, T., & Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic acids research*, *40*(D1), D302-D305.

Luo, H., Lin, Y., Gao, F., Zhang, C. T., & Zhang, R. (2013). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic acids research*, *42*(D1), D574-D580.

Loiseau, L., Fyfe, C., Aussel, L., Chehade, M. H., Hernández, S. B., Faivre, B., ... & Barras, F. (2017). The UbiK protein is an accessory factor necessary for bacterial ubiquinone (UQ) biosynthesis and forms a complex with the UQ biogenesis factor UbiJ. *Journal of Biological Chemistry*, *292*(28), 11937-11950.

Liu, B., Zheng, D., Jin, Q., Chen, L., & Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic acids research*, *47*(D1), D687-D692.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., ... & Marchler-Bauer, A. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic acids research*, *48*(D1), D265-D268.

Mahdi, A. A., Sharples, G. J., Mandal, T. N., & Lloyd, R. G. (1996). Holliday Junction Resolvases Encoded by HomologousrusAGenes inEscherichia coliK-12 and Phage 82. *Journal of molecular biology*, *257*(3), 561-573.

Möller, S., Croning, M. D., & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, *17*(7), 646-653.

Martin, J. L., & McMillan, F. M. (2002). SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. *Current opinion in structural biology*, *12*(6), 783-793.

Nord, S., Bylund, G. O., Lövgren, J. M., & Wikström, P. M. (2009). The RimP protein is important for maturation of the 30S ribosomal subunit. *Journal of molecular biology*, *386*(3), 742-753.

Nigam P. S. (2013). Microbial enzymes with special characteristics for biotechnological applications. *Biomolecules*, *3*(3), 597–611.

Nielsen, H., in Kihara, D. (Ed.). (2017). *Protein Function Prediction: Methods and Protocols*. Totowa: Humana Press.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093-1109.

Oganesyan, V., Busso, D., Brandsen, J., Chen, S., Jancarik, J., Kim, R., & Kim, S. H. (2003). Structure of the hypothetical protein AQ_1354 from *Aquifex aeolicus*. *Acta Crystallographica Section D: Biological Crystallography*, *59*(7), 1219-1223.

Ponting, C. P., & Aravind, L. (1999). START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. *Trends in biochemical sciences*, *24*(4), 130-132.

Perard, J. Ollagnier de Choudens S. 2018. Iron-sulfur clusters biogenesis by the SUF machinery: close to the molecular mechanism understanding. *J Biol Inorg Chem*, *23*, 581-596.

Prava, J., Pranavathiyani, G., & Pan, A. (2018). Functional assignment for essential hypothetical proteins of *Staphylococcus aureus* N315. *International journal of biological macromolecules, 108*, 765-774.

Prabhu, D., Rajamanikandan, S., Anusha, S., Chowdary, M. S., Veerapandiyan, M., & Jeyakanthan, J. (2020). In silico functional annotation and characterization of hypothetical proteins from *Serratia marcescens* FGI94. *Biology Bulletin*, *47*(4), 319-331.

Quick, M., & Javitch, J. A. (2007). Monitoring the function of membrane transport proteins in detergent-solubilized form. *Proceedings of the National Academy of Sciences*, *104*(9), 3603-3608.

Raj, U., Sharma, A. K., Aier, I., & Varadwaj, P. K. (2017). In silico characterization of hypothetical proteins obtained from *Mycobacterium tuberculosis* H37Rv. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *6*(1), 1-9.

Saslaw, S., & Carlisle, H. N. (1961). Studies with tularemia vaccines in volunteers. IV. *Brucella* aggiutinins in vaccinated and nonvaccinated volunteers challenged with *Pasteurella tularensis*. *The American journal of the medical sciences*, *242*, 166-172.

Sievers, J., & Errington, J. (2000). Analysis of the essential cell division gene ftsL of *Bacillus subtilis* by mutagenesis and heterologous complementation. *Journal of bacteriology*, *182*(19), 5572-5579.

Spaink, H. P. (2004). Specific recognition of bacteria by plant LysM domain receptor kinases. *Trends in microbiology*, *12*(5), 201-204.

Saha, S., & Raghava, G. P. S. (2006). VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics, proteomics & bioinformatics*, *4*(1), 42-47.

Sperandeo, P., Lau, F. K., Carpentieri, A., De Castro, C., Molinaro, A., Deho, G., ... & Polissi, A. (2008). Functional analysis of the protein machinery required for transport of lipopolysaccharide to the outer membrane of *Escherichia coli*. *Journal of bacteriology*, *190*(13), 4460-4469.

Saada, A., Vogel, R. O., Hoefs, S. J., van den Brand, M. A., Wessels, H. J., Willems, P. H., ... & Nijtmans, L. G. (2009). Mutations in NDUFAF3 (C3ORF60), encoding an NDUFAF4 (C6ORF66)-interacting complex I assembly protein, cause fatal neonatal mitochondrial disease. *The American Journal of Human Genetics*, *84*(6), 718-727

Shen, H. B., & Chou, K. C. (2009). Predicting protein fold pattern with functional domain and sequential evolution information. *Journal of Theoretical Biology*, *256*(3), 441-446.

Šink, R., Kotnik, M., Zega, A., Barreteau, H., Gobec, S., Blanot, D., ... & Contreras-Martel, C. (2016). Crystallographic study of peptidoglycan biosynthesis enzyme MurD: domain movement revisited. *PloS one*, *11*(3), e0152075.

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., ... & von Mering, C. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, *49*(D1), D605-D612.

Tusnady, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics*, *17*(9), 849-850.

Tärnvik, A., & Berglund, L. (2003). Tularaemia. *European Respiratory Journal*, *21*(2), 361-373.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., ... & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research*, *13*(9), 2129-2141.

Twenhafel, N. A., Alves, D. A., & Purcell, B. K. (2009). Pathology of inhalational *Francisella tularensis* spp. *tularensis* SCHU S4 infection in African green monkeys (*Chlorocebus aethiops*). *Veterinary pathology*, *46*(4), 698-706.

Uddin, M. E., Maitra, P., Faruquee, H. M., & Alam, M. F. (2014). Isolation and characterization of proteases enzyme from locally isolated *Bacillus* sp. *American Journal of Life Sciences*, *2*(6), 338-344.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., ... & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*(W1), W296-W303.

Xiang, Z. (2006). Advances in homology protein structure modeling. *Current Protein and Peptide Science*, *7*(3), 217-227.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., ... & Brinkman, F. S. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), 1608-1615.

Yu, C. S., Chen, Y. C., Lu, C. H., & Hwang, J. K. (2006). Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, *64*(3), 643-651.

.