Targeting Essential Hypothetical Proteins of *Pseudomonas aeruginosa* PAO1 for Mining of Novel Therapeutics: An *in silico* Approach

Atikur Rahman
Jashore University of Science and Technology
Md. Takim Sarker
Jashore University of Science and Technology
Md Ashiqul Islam
University of Windsor
Mohammad Uzzal Hossain
National Institute of Biotechnology
Mahmudul Hasan
Sylhet Agricultural University
Tasmina Ferdous Susmi (tsusmi7@gmail.com)
Jashore University of Science and Technology https://orcid.org/0000-0002-2628-2371

Research Article

Keywords: Pseudomonas aeruginosa, Functional Annotation, Protein-protein Interactions, Non-homology analysis, Therapeutics

Posted Date: May 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1650735/v1

Abstract

Pseudomonas aeruginosa PAO1, an omnipresent opportunistic bacterium responsible for acute and chronic infection in immunocompromised individuals, is currently on WHO's lists where new antibiotics are urgently required for those. Finding essential genes and essential hypothetical proteins(EHP) can be crucial in identifying novel druggable targets and therapeutics. This study aims to characterize these EHPs, analyze subcellular and physiochemical properties, PPIs network, non-homologous analysis against humans, virulence factor and novel drug target prediction, and finally structural analysis of the identified target employing around 42 robust bioinformatics tools/databases, the output of which was evaluated using ROC analysis. The study discieverd 18 EHPs from 336 essential genes, with domain and functional annotation revealing that 50% of these proteins belong to the enzyme category. The majority are cytoplasmic and cytoplasmic membrane proteins, with half of them being stable proteins which were subjected to PPIs network analysis. The network contains 261 nodes and 269 edges for 9 proteins of interest, with 11 hubs containing at least three nodes each. Finally, pipeline builder predicts 7 proteins with novel drug targets, 5 non-homologous proteins against human proteome, human anti-targets, human gut flora, and 3 virulent proteins. Among these, homology modeling of NP_249450 and NP_251676 were done and the Ramachandran plot analysis revealed that more than 94% of the residues were in the preferred region. By analyzing functional attributes and virulence characteristics, the findings of this study may facilitate the development of innovative antibacterial drug targets and drugs of *Pseudomonas aeruginosa* PAO1.

1. Introduction

Pseudomonas aeruginosa often termed as an opportunistic pathogen, is a rod-shaped, motile, gram-negative and non-fermenting bacteria found ubiquitously in soil and water as well as found in colonies on the animate part of plant and animal including humans [1, 2]. Isolates collected from diverse environments reported 272 species of the *Pseudomonas* genus in which *P. aeruginosa* PA01 is one of the most commonly used laboratory strains as well as employed to generate publicly accessible genomic resources [2, 3]. *P. aeruginosa* PA01 is the first-ever strain of its species having a completely sequenced genome from a chronic lesion isolate dated from the 1950s. The genome is 6.3 Mbp long that includes 5570 ORFs, roughly 89.4% coding regions, and 0.4% stable RNAs. This was the largest bacterial genome available during the year 2000 when sequenced. However, despite the same species, different genomic and phenotypic changes are found across isolates of *P. aeruginosa* PA01 strains stored in different laboratories worldwide [2, 4].

A broad spectrum of host targets including nematodes, insects, plants, and mammals are susceptible to infection by *P. aeruginosa* species [2]. It is found harmless in normal gut microflora but causes dangerous infection in critically ill ICU patients [5]. This trend in pathogenesis makes them an opportunistic pathogen [2]. It is regarded to be within the top three causative agents for infection caused by opportunistic pathogens annually in the community as well as related to (10–15) % of hospital-acquired infections [6]. In 2015, a report from the European Antimicrobial Resistance Surveillance Network (EARS-Net) on European regions revealed that around 13.7% strains of *P. aeruginosa* had acquired resistance to a minimum of three anti-microbial communities whereas about 5.5% of the strains were resistant against five anti-microbial groups. Every year in the USA alone, roughly 440 deaths and 51,000 infection cases are caused by *P. aeruginosa* of which over 13% results from multi-drug resistant *Pseudomonas* strains. As a consequence, *P. aeruginosa* has been announced as one of the greatest threats to public health amongst the 12 bacterial families from the antibiotic-resistance priority pathogens enlisted by WHO in 2017 [7]. It is also involved with some other nosocomial infections like bloodstream infection, gastrointestinal infection, and urinary tract infection [5]. This bacterium poses a devastating impact on lung disease patients with cystic fibrosis (CF). Apart from CF, it is equally deadly for individuals having compromised immune systems like AIDS, cancer, burn lesions, and eye injuries. The situation can get even worse despite having robust antibiotic medication since *P. aeruginosa* possess a wide spectrum of resistance against antibiotics including aminoglycosides, β-lactams, and fluoroquinolones. Therefore, disease stress subsequently results in organ failure and eventually death [2, 5].

P. aeruginosa adopts some survival strategy that helps them to resist environmental stressors and dodging host immune responses [8]. Some of these survival tools include biofilm formation, enzyme promiscuity, horizontal gene transfer, and quorum sensing [5]. It is one of the well-studied strains for investigating the bacterial biofilm formation process [9]. Three polysaccharides, alginate, Pel, and Psl, were discovered to be important for bacterial attachment and biofilm formation in *P. aeruginosa* PA01 [8]. Over 500 regulatory genes have been recorded from the *P. aeruginosa* PA01 genome investigation [10]. There is still a lot to discover for a better understanding of the intracellular signaling pathways and several other regulatory mechanisms involving many proteins that are still uncharacterized. Thus, domain analysis and functional annotation of essential hypothetical proteins (EHPs) can pave the way to identify new potential targets facilitating the drug repositioning development. Since these EHPs are needed for cellular, biological, and metabolic processes, their deletion or mutation can be fatal to the species. These prospective drug targets may be crucial in the development of antimicrobial drugs [11].

In this study, an *in-silico* based approach has adopted for the characterization of proteins with unknown functions via different algorithm-based tools and software. Besides, a network-based analysis was directed to find interaction with critically connected hub proteins that may control major molecular activities together. The pipeline builder was employed to analyze non-homologous proteins against humans, human anti-targets, and the proteome of the gut microbiota, as well as predict virulence factors and novel drug targets. Finally, using reliable software, the structural conformation of our protein of interest with potential druggability was predicted and assessed. Thus, our analysis mainly involves the identification of essential hypothetical proteins in *P. aeruginosa* PA01 and can further lead to the discovery of novel proteins of therapeutic targets.

2. Materials And Methods

2.1 *Sequence retrieval and analysis*

The full proteome of *P. aeruginosa* PAO1 (strain ATCC 15692) was retrieved from the NCBI genome database. The bacterial complete genome contains 6.3 million base pairs and 5564 proteins [4]. The essential genes database (DEG) is then subjected to find out the essential hypothetical proteins (EHPs) from this complete proteome list by employing a series of unique keywords [12]. To begin, we looked for similar hypothetical proteins where we found 2181 proteins

among these 5564 proteins. Following that, we searched for the exact matches of hypothetical proteins and exact matches of conserved hypothetical proteins and found 1540 and 625 hypothetical proteins, respectively. According to the DEG database, this bacterial proteome contains 336 essential proteins (EPs). Essential proteins are those that are inevitable and adequate for a living cell to survive under ideal circumstances. Consequently, we discovered 29 essential hypothetical proteins by manual curation whose genomes were entirely conserved among the 336 EPs. The status (reviewed or unreviewed), annotation score (1–5), structural and functional availability, and other factors were used to further validate these 29 EHPs from the NCBI and UniProt databases. Eventually, we excluded 11 proteins, leaving 18 essential hypothetical proteins whose FASTA sequences were used to facilitate further analysis throughout this study. The complete framework of our investigation is presented in Fig. 1 and all the databases/software used in this study are in Table 1.

Table 1
Bioinformatics resources used in the study

Serial no	*Server/ Database*	*Version*	*Using Reason*	*Link*	*References*
Functional Annotation					
1	DEG	15.2	Finding essential HPs	http://tubic.tju.edu.cn/deg/	[12]
2	GO FEAT	1.0	For functional annotation	http://computationalbiology.ufpa.br/gofeat/	[13]
3	CDART		Protein Homology search Domain Architecture	https://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi	[14]
4	SMART		Identification and annotation of protein domains	http://smart.embl-heidelberg.de/	[15]
5	SUPERFAMILY	1.75	For functional annotation	https://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	[16]
6	Pfam	34.0	Determine protein families	http://pfam.xfam.org/	[17]
7	SVMProt		Protein functional family prediction	http://bidd.group/cgi-bin/svmprot/svmprot.cgi	[18]
8	CATH	4.3	Protein domains into superfamily	http://www.cathdb.info/	[19]
9	InterPro	84.0	Classification of protein families	https://www.ebi.ac.uk/interpro/	[20]
10	HHPred		Sequence similarity searching, prediction of sequence features, and sequence classification.	https://toolkit.tuebingen.mpg.de/tools/hhpred	[21]
11	PANNZER		Functional annotation of uncharacterized proteins	http://ekhidna2.biocenter.helsinki.fi/sanspanz/	[22]
12	PFP		Automated protein function gene ontology prediction	https://kiharalab.org/web/pfp.php	[23]
13	ESG		Protein Function Prediction	https://kiharalab.org/web/esg.php	[24]
Subcellular Localization					
14	Psortb	3.0.2	Subcellular Localization	https://www.psort.org/psortb/	[25]
15	CELLO	v.2.5	Subcellular Localization	http://cello.life.nctu.edu.tw/	[26], [27]
16	TMHMM	v. 2.0	prediction of transmembrane helices in proteins	http://www.cbs.dtu.dk/services/TMHMM-2.0/	[28]
17	Phobius		prediction of transmembrane helices in proteins	https://phobius.sbc.su.se/index.html	[29]
18	HMMTOP	2.0	prediction of transmembrane helices in proteins	http://www.enzim.hu/hmmtop/index.php	[30]

Serial no	Server/ Database	Version	Using Reason	Link	References
19	CCTOP	1.00	prediction of transmembrane helices in proteins	http://cctop.enzim.ttk.mta.hu/	[31]
20	PROTTER	1.0	predicts the presence and location of signal peptide cleavage sites in amino acid sequences and prediction of transmembrane helices in proteins	https://wlab.ethz.ch/protter/start/	[32]
21	SignalP 4.1	4.1	predicts the presence and location of signal peptide cleavage sites in amino acid sequences	http://www.cbs.dtu.dk/services/SignalP-4.1/	[33]
22	PrediSi		Prediction of Signal peptides	http://www.predisi.de/	[34]
Physicochemical Properties					
23	ProtParam		computation of various physical and chemical parameters for a given protein	https://web.expasy.org/protparam/	[36]
Protein-Protein Interaction					
24	NetworkAnalyst	v3.0	PPI construction and visualization	https://www.networkanalyst.ca/NetworkAnalyst/uploads/ListUploadView.xhtml	[39]
Non-Homology Analysis					
25	PBIT		Pipeline building for non-homology analysis	http://www.pbit.bicnirrh.res.in/	[41]
Virulence Factor Analysis					
26	VICMpred		functional classification of proteins of bacteria into virulence factors	https://webs.iiitd.edu.in/raghava/vicmpred/index.html	[43]
27	VirulentPred		VirulentPred is a bacterial virulent protein prediction method	http://bioinfo.icgeb.res.in/virulent/	[44]
28	MP3		predict pathogenic proteins in both genomic and metagenomic datasets	http://metagenomics.iiserb.ac.in/mp3/tutorial.php	[45]
Druggability Analysis					
29	DrugBank	5.0	Identification of information on drugs and drug targets	https://go.drugbank.com/	[47]
Secondary Structure Analysis					
30	SOPMA		Secondary structure prediction	https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html	[48]
31	PSIPRED	4.0	Secondary structure prediction	http://bioinf.cs.ucl.ac.uk/psipred/	[49]
3D Structure Analysis					

Serial no	*Server/ Database*	*Version*	*Using Reason*	*Link*	*References*
32	SWISS-MODEL		Protein 3D structure determination	https://swissmodel.expasy.org/	[50]
33	Robetta		Protein 3D structure determination	https://robetta.bakerlab.org/	
34	Galaxy Refine		Refinement of protein structure	http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE	[52]
35	PyMOL Software	2.0	Structure visualization	https://pymol.org/2/	
Validation Check					
36	ERRAT		3D structure validation	https://saves.mbi.ucla.edu/	[53]
37	VARIFY 3D		3D structure validation	https://saves.mbi.ucla.edu/	[54], [55]
38	PROVE		3D structure validation	https://saves.mbi.ucla.edu/	[56]
39	WHATCHECK		3D structure validation	https://saves.mbi.ucla.edu/	[57]
40	PROCHECK		3D structure validation	https://saves.mbi.ucla.edu/	[58]
41	Ramachandran Plot		3D structure validation	http://services.mbi.ucla.edu/SAVES/Ramachandran/	[58]
42	ROC Analysis		This web page calculates a receiver operating characteristic (ROC) curve from data	http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html	

2.2 Segment I: Functional annotation and Properties Characterization

2.2.1 Functional annotation and domain analysis of EHPs:

The functional annotation of 18 *Pseudomonas aeruginosa* EHPs was unveiled by using numerous publicly accessible databases and tools. To gain more knowledge about the molecular functions and biological processes of the EHPs, we consider protein superfamily, family, conserved domain analysis, and Gene Ontology (GO) analysis. Using an online server GO FEAT, for the functional characterization by homology searching through multiple databases such as NCBI, Uniprot, and EMBL, a preliminary assessment was performed to see if any of the HPs were allocated a family and/or protein domain [13]. After preliminary evaluation, proteins conserved domains and protein functions based on domain architecture were determined by using CDART [14] from the conserved domain database (CDD) and SMART [15], respectively. For functional analysis, SUPERFAMILY 1.75 [16], Pfam 34.0 [17], SVMProt [18], CATH 4.3 [19], InterPro 84.0 [20], and HHPred [21] were used to identify the protein superfamily, functional family, domain, and essential sites based on similarity. PANNZER [22], PFP [23], and ESG [24] tools were used for high-throughput functional annotation of EHPs, which provided gene ontology information with z-scores as well as brief explanations of the annotated protein's functionality. These GO terms facilitate understanding a gene's molecular functions, physiological roles, and cellular mechanism, which refers to the location of the gene's product. We used default parameters for all databases.

2.2.2 Subcellular localization and Transmembrane Helices Analysis

Sub-cellular localization of a protein can help to infer much information about that protein’s function. In our study, we employed several databases to annotate the subcellular localization of the selected 9 EHPs which includes PSORTb [25], CELLO [26, 27], TMHMM [28], Phobius [29], HMMTOP [30], CCTOP [31], PROTTER [32], SignalP 4.1 [33], and PrediSi [34]. According to PSORTb and CELLO, the proteins were distinguished by 5 major cellular position: cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular. To predict transmembrane helices, TMHMM, Phobius, HMMTOP, CCTOP and PROTTER were employed. Information of transmembrane helices location is somehow beneficial for the conformation of possible 3D structure [35]. Besides, it is necessary to find out signal peptides which is the N-terminal part of a protein. Mainly, they are targeted to the endoplasmic reticulum to the secretory pathway and it is considered as the way of protein localization prediction [33]. Signal peptides were identified by using these SignalP 4.1, PROTTER and PrediSi databases.

2.2.3 Analysis of physicochemical properties

The Expasy’s ProtParam server [36] was utilized for the analysis of physicochemical properties of 18 selected essential hypothetical proteins (EHPs) which include molecular weight, theoretical pI (Isoelectric Point), Formula, the total number of positively and negatively charged residues, instability index, aliphatic index and grand average of hydropathicity (GRAVY).

2.3 Segment II: Protein-Protein Interaction Network of 9 EHPs

2.3.1 Protein-protein interaction network analysis

The function of a protein molecule often is modulated by its surrounding protein networks [37]. For this reason, it is important to discover the protein network to get an insight into the functional association of a particular protein [38]. In this study, we have used NetworkAnalyst v3.0 for network building [39]. We have inputted a list of genes containing 9 EHPs with their Uniprot IDs (Q9HXM8, Q9HWT5, Q9HVM2, Q9HVF5, Q9I5H0, Q9HZL8, Q9HYC8, Q9HXV5, and Q9HUH3) since all of these proteins were found stable through the physicochemical analysis. The Generic PPI option under Protein-protein Interactions (PPI) was checked for further processing. *P. aeruginosa* PA01 interactome database provided with robust computational prediction and experimentally validated data were adopted for network building. Next, the corresponding network was explored for further analysis in Cytoscape. Cytoscape is a standalone software that enables several topological parameter analyses like discovering the shortest possible path, node degree distribution, clustering hub genes of the network [40].

2.4 Segment III: Non-Homology Analysis, Virulence Factor Prediction and Druggability Identification

2.4.1 Non-homology analysis against human proteome and human anti-targets

Several features were needed for the identification of the drug target for any human diseases. For this reason, to analyze non-homology aspects, we tried Pipeline builder for identification of target (PBIT) server for the non-homology analysis against human proteome, against human anti-targets and human gut flora proteomes [41]. Using the pipeline builder, we first identified human homologous proteins that share high sequence similarity with human proteome. The sequence similarity of the inputted 9 sequences was figured using BLAST algorithm where E-value > 0.005 and % sequence identity < 50 was set. 8 of the 9 input sequences are non-homologous that were selected for further investigation. These homologous proteins were filtered to avoid the undesirable toxic-effects for these similarities. Filtered and selected 8 non-homologous proteins were further employed in the pipeline to recognize non-homologous proteins against human anti-targets. Proteins that contain harmful effects due to the impact of a drug named anti-targets [41]. To screen out the significant similar sequence with familiar human anti-targets, PBIT database uses BLAST algorithm where they utilize those human anti-targets proteins based on different literature [41]. Again E-value > 0.005 and % sequence identity < 50 was set and all non-homologous sequences were selected.

2.4.2 Non-homology analysis against human gut flora proteomes

PBIT also analysis human gut flora proteomes that make it easier to find out those highly similar sequences with human gut microbiota. It is known that gut microbiota plays an important role in human health that includes immune, metabolic and neurobehavioral characters [42]. That's why it is necessary to design such drugs whose target is non-homologous protein sequence of the gut microbiome. As result, such drugs could not be able to kill or hamper essential microbes found in human gut. For this, Pipeline builder for identification of target (PBIT) server was again used to identify non-homologous proteins against gut microbiota proteomes. As the third step of the pipeline builder, selected proteins were employed where E-value > 0.001and % sequence identity < 50 was set. Now non-homologous proteins were selected for the next investigation.

2.4.3 Analysis of virulence factor

Understanding the pathogenesis mechanism through the analysis of virulence factors can be a key to the discovery of new promising therapeutic targets [38]. Therefore, we have used VICMpred [43], VirulentPred [44] and MP3 [45] for the identification of the virulence property of the 9 EHPs. We have collected the results predicted combinedly by 2 out of the 3 tools. All the results were collected by using the provided default options by the servers.

2.4.4 Druggability analysis and New Target Identification

Identification of a new drug target can be a new window for the discovery and development of a new drug against infectious or serious disease [46]. Druggability analysis is the examination of a protein that has the possible capability or binding affinity towards a drug or drug-like molecules. This Druggability analysis can introduce a new drug target against a drug. Here we used DrugBank, a comprehensive, online database that contains information on drugs and drug targets [47]. Target identification segment was utilized for this purpose and amino acid sequences in FASTA Format was the searching index. All other BLAST Parameters and Filters were set as default where the Expectation value was set 0.00001.

2.5 Segment IV: Structure Prediction and Structure Validation

2.5.1 Secondary Structure Analysis

The interactions between neighboring polypeptides mainly design a protein's secondary structure. When the elements of the secondary structure have folded together among each other, the 3D structure of the protein is formed. The databases namely SOPMA [48] and PSIPRED [49] provide the secondary structure of a protein. These databases were used to predict the structure where protein sequence in FASTA format was the searching index for the websites and the rest of the parameters were set as default.

2.5.2 Essential hypothetical proteins 3D structure modelling

The protein 3D structure was determined based on two methods: template-based homology modelling, and trRosetta methods. The three-dimensional structure of the targeted protein was generated using the SWISS-MODEL server, which uses template search and then aligns the target sequence with the template structure to create the homology model [50]. To construct the model with an accuracy equal to low-resolution x-ray crystallography, we only consider templates with ≥ 30% sequence identity. Then the server, Robetta (https://robetta.bakerlab.org/) was employed to predict the 3D model by using trRosetta algorithm. It is a deep learning method based on direct energy minimizations that is the most accurate process of structure building provided by this server

[51]. Finally, the built structure was optimized using Galaxy Refiner, with the best-refined model based on the lowest MolProbity and highest GDT-HA value [52]. Consequently, PyMOL 2.0 visualization software is used to visualize all of the refined structure files, which are in .pdb format.

2.5.3 Protein structure validation assessment:

The reliability of a predicted 3D structure of a protein can be assessed by using various quality assessment tools. Here, in this study, we used SAVES version 6.0 (https://saves.mbi.ucla.edu/) which is a meta-server that runs six programs at once to check and validate protein structure during and after model refinement. This server validates the stereochemical consistency of a protein structure by performing residue by residue geometry and overall structure geometry. Furthermore, it also compares the results to good structures to see if an atomic model (3D) is compatible with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar, etc.). We run ERRAT [53], VARIFY 3D [54, 55], PROVE [56], WHATCHECK [57], PROCHECK [58], and Ramachandran Plot [58] from SAVES v6.0 to determine the consistency of the construct model.

2.6 Performance assessment of the Study

In our study, we have applied the receiver operating characteristic (ROC) analysis for validating the accuracy of our bioinformatics tools used for the functional annotation of EHPs from *P. aeruginosa* [59]. We have collected 100 arbitrary protein functions of *P. aeruginosa* along with their gene names using the same pipeline used prior to our study in the **Supplementary excel file**. Two integer values namely “1” as a truly positive and “0” as a truly negative were assigned to classify the prediction. The confidence rating was denoted by “2”, “3”, “4” and “5” respectively. The higher number denotes greater level of confidence. The input file consists of 2 columns where 1st column contains binary numbers like 1 (True positive) and 0 (True negative) and the 2nd column contains a rate of confidence ranging from 2 to 5. For the present study, six levels were considered for determining the diagnostic efficacy. The ROC analysis was used for 12 individual functional annotation tools. The data were submitted to an online-based ROC curve generating web server in format-1 [60]. The output result includes accuracy, sensitivity, specificity and the ROC area (**Supplementary File 1**). The accuracy of our adopted pipeline is 97.42% which indicates a very high and reliable result for the bioinformatics tools that we used in our study.

3. Results

3.1 Functional annotation and domain analysis of EHPs

The functional annotation of the 18 EHPs was examined using 12 reliable platforms that predict protein superfamily, family, conserved domains, and Gene Ontology terms (GO). Here, the functional annotation was assigned with high confidence as we considered only that function that was similar in three or more programs. Consequently, the functional characterization categorizes these proteins into 9 functional categories, namely enzymes (deaminases, dehydrogenases, helicases, transferases, DNases, oxidoreductases, kinases, etc), transporter protein, bacterial outer membrane protein, folate binding protein, peptidase inhibitor protein, electron transporter protein, chromosome partition protein, ribosome maturation protein, pathogenesis-related protein. Nine of the 18 EHPs are enzymes (NP_252456.1, NP_252782.1, NP_253095.1, NP_253326.1, NP_250846.1, NP_252375.1, NP_253678.1, NP_253679.1, NP_253685.1), two are transporter proteins (NP_253252.1, NP_251676.1), and the remaining seven proteins are in the seven groups. Table 2 enlists the 18 EHPs superfamily, functional family, molecular functions, biological functions, as well as their GO IDs and database IDs. Among these proteins, NP_249450.1 is a member of the folate-binding superfamily, with the aminomethyl transferase folate-binding domain as its functional family. Aminomethyl transferase and transaminase activity are the two molecular functions of this protein. Another protein sequence of NP_251676.1 was predicted belonging to the functional family that represents the periplasmic core domain found in a variety of ABC transporters. ATP binding, ATPase-coupled xenobiotic transmembrane transporter activity, efflux transmembrane transporter activity, and ATPase activity are some of the molecular functions of this protein. According to the GO annotation, there were 65 GO terminologies in total for the molecular function and biological process. These GO IDs can be used to retrieve Gene Ontology analysis of these 18 EHPs.

Table 2
Functional annotations of 18 essential hypothetical proteins

Serial No.	*RefSeq*	*Superfamily*	*Family*	*Gene ontology*		*Go ID/D integrat*
				Biological process	**Molecular function**	
1	NP_252456.1	Cytidine deaminase-like	Deoxycytidylate deaminase-like	1. tRNA wobble adenosine to inosine editing	1.Hydrolase activity 2.Zinc ion binding 3.Catalytic activity 4.tRNA-specific adenosine 34-deaminase activity	(GO:000 (GO:001 (GO:000 (GO:000 (GO:005 Uniprot Interpro Interpro Interpro Interpro Interpro Pfam (P NCBI (5 EMBL (ATNK0
2	NP_252782.1	Hotdog Thioesterase / thiol ester dehydratase-isomerase	Thioesterase	1. Histidine biosynthetic process	1. Histidinol dehydrogenase activity 2. Zinc ion binding 3. NAD binding	(GO:000 (GO:000 (GO:000 (GO:005 (A0A44 Interpro Interpro Interpro Pfam (P EMBL (L
3	NP_253095.1	Uncharacterized protein	Dna[CI] antecedent, DciA	1. Protein dephosphorylation	1. Zinc ion binding 2. Protein tyrosine/serine/threonine phosphatase activity	(GO:000 (GO:000 (GO:000 Uniprot(Interpro KEGG(p KEGG GM(pae Interpro Pfam (P NCBI(48 EMBL (A

Serial No.	*RefSeq*	*Superfamily*	*Family*	*Gene ontology*		*Go ID/D integrat*
				Biological process	**Molecular function**	
4	NP_253252.1	MATE_like	Lipid II flippaseMurJ, Polysaccharide biosynthesis C-terminal domain	**1.** Cell wall organization **2.** Peptidoglycan biosynthetic process **3.** Regulation of cell shape	1. Lipid-linked peptidoglycan transporter activity	(GO:007 (GO:000 (GO:000 (GO:001 Uniprot Interpro Interpro Pfam (P NCBI (5 EMBL (ATNK0
5	NP_253326.1	Glycerol-3-phosphate (1)-acyltransferase	Glycerol-3-phosphate (1)-acyltransferase	**1.** D-galacturonate catabolic process **2.** D-glucuronate catabolic process	1. Transferase activity, transferring acyl groups	Uniprot Interpro KEGG (p KEGG GM(pae Interpro Pfam (P NCBI (48 EMBL (A
6	NP_253368.1	TonB-dependent receptor family	Energy transducer TonB	1. Viral process	1. GTP binding	(GO:001 (GO:000 (Q9HVB Interpro KEGG (p KEGG GM(pae NCBI (48 EMBL (A
7	NP_249450.1	Folate-binding	Aminomethyl transferase folate-binding domain	1. Iron-sulfur cluster assembly 2. Glycine decarboxylation via glycine cleavage system	1. Aminomethyl transferase activity 2. Transaminase activity	(GO:001 (GO:000 (GO:000 Superfa (GO:001 Uniprot Interpro KEGG (p KEGG GM(pae Interpro Interpro NCBI (48 EMBL (A

Serial No.	*RefSeq*	*Superfamily*	*Family*	*Gene ontology*		*Go ID/D integrat*
				Biological process	**Molecular function**	
8	NP_250659.1	Inhibitor_I78	Peptidase inhibitor I78 family	1. Cell adhesion 2. Homophilic cell adhesion via plasma membrane adhesion molecules	1. Calcium ion binding 2. Serine-type endopeptidase inhibitor activity	(GO:000 (GO:000 (GO:000 (GO:000 SMART Uniprot Interpro KEGG (p KEGG G (pae:PA Interpro Pfam (P NCBI (4 EMBL (A
9	NP_250846.1	DNase I-like	Endonuclease/Exonuclease/phosphatase	N/A	**1.** Endonuclease activity **2.** Exonuclease activity	SMART (GO:000 (GO:000 Uniprot Interpro Interpro Interpro Pfam (P EMBL ((
10	NP_251676.1	LolE	MacB-like periplasmic core domain, Lipoprotein-releasing ABC transporter permease	**1.** Lipoprotein localization to outer membrane **2.** Lipoprotein transport **3.** Protein localization to outer membrane	1. ATP binding 2. ATPase-coupled xenobiotic transmembrane transporter activity 3. Efflux transmembrane transporter activity 4. ATPase activity	SMART (GO:004 (GO:004 (GO:008 (GO:000 (GO:000 (GO:001 (GO:001 Uniprot Interpro KEGG (p KEGG G (pae:PA Interpro Interpro Interpro Pfam (P Pfam (P NCBI (4 EMBL (A

Serial No.	RefSeq	Superfamily	Family	Gene ontology		Go ID/D integrat
				Biological process	Molecular function	
11	NP_252171.1	Fe-S cluster assembly (FSCA) domain-like	Iron-sulfur cluster assembly protein	1. Iron-sulfur cluster assembly	1.ATPase activity 2.ATP binding 3.Iron-sulfur cluster binding 4.Metal ion binding	SMART (GO:001 (GO:001 (GO:000 (GO:005 (GO:004 Uniprot (A0A3S4 Interpro Interpro Interpro Interpro Interpro Interpro Interpro Pfam (P Pfam (P EMBL (L
12	NP_252375.1	Carbam_trans_N (Carbamoyltransferase N-terminus)	tRNA N6-adenosine threonyl carbamoyltransferase	1. tRNA threonyl carbamoyl adenosine modification	1. Metalloendopeptidase activity 2. Iron ion binding 3. N(6)-L-threonyl carbamoyl adenine synthase activity	SMART (GO:000 (GO:000 (GO:000 (GO:006 Uniprot Interpro KEGG (p KEGG G (pae:PA Interpro Interpro Interpro Pfam (P NCBI (88 EMBL (A

Serial No.	RefSeq	Superfamily	Family	Gene ontology		Go ID/D integrat
				Biological process	Molecular function	
13	NP_253374.1	MukE (MukE is part of the MukBEF condensin complex)	Bacterial condensin subunit MukE	1. Cell cycle 2. Cell division 3. DNA replication 4. Chromosome segregation 5. Chromosome condensation	1. GTP binding 2. GTPase activity 3. Translation elongation factor activity 4. ATP binding	(GO:000 (GO:005 (GO:000 (GO:000 (GO:003 (GO:000 (GO:000 (GO:000 (GO:000 Uniprot Interpro Interpro EMBL (L
14	NP_253434.1	1.RimP N-terminal domain 2.RimP C-terminal SH3 domain (also known as yhbC)	RimP N-terminal domain, RimP C-terminal SH3 domain	1.Ribosomal small subunit biogenesis	N/A	SMART (GO:004 Uniprot (A0A3S4 Interpro (A0A3S4 Interpro Interpro Interpro Interpro Interpro Pfam (P Pfam (P EMBL (L
15	NP_253455.1	Bet v1-like	Polyketide cyclase /dehydrase and lipid transport	1. Ubiquinone biosynthetic process 2. Cellular respiration	1. Ubiquinone binding	(GO:000 (GO:004 (GO:004 SUPERF SMART IPR005(Uniprot Interpro Interpro Interpro Pfam (P EMBL (L
16	NP_253678.1	FAD/NAD(P)-binding domain	FAD dependent oxidoreductase	1.Oxidation-reduction process	1.Oxidoreductase activity	SMART (GO:005 (GO:001

Serial No.	*RefSeq*	*Superfamily*	*Family*	*Gene ontology*		*Go ID/D integrat*
				Biological process	**Molecular function**	
17	NP_253679.1	NAD(P)-linked oxidoreductase/ Aldo-keto reductase (AKR) superfamily	Aldo/keto reductase family	1. Daunorubicin metabolic process 2. Doxorubicin metabolic process	1.Oxidoreductase activity 2.D-threo-aldose 1-dehydrogenase activity	SMART (GO:004 (GO:004 (GO:004 Uniprot Interpro Interpro Interpro Pfam (P EMBL (I
18	NP_253685.1	Protein kinase-like (PK-like)	Phosphotransferase enzyme family	1. Protein phosphorylation	1. ATP binding 2. Protein serine/threonine kinase activity	Pfam (GO:000 (GO:000 (GO:000 Uniprot Interpro Interpro EMBL (I

3.2 Subcellular Localizations of EHPs

To identify the cellular localization of our 18 EHPs, the websites PSORTb and CELLO were utilized. According to the data of PSORTb, among 18 essential hypothetical proteins, 6 proteins belong to cytoplasmic protein, 8 proteins belong to the location of the cytoplasmic membrane and the remaining 4 proteins are considered as unknown. The database, CELLO depicted that 14 proteins are cytoplasmic protein, 2 proteins are considered as inner membrane protein and the rest 2 are periplasmic protein. This is the generalized concept of the cellular location which is shown in Fig. 2 **and supplementary table 1**. The existence of the transmembrane helix was also figured out and this can help to carry out the function of a protein through transmembrane transportation. The amount of transmembrane helix was given in **supplementary table 1**. The presence of signal peptide was also investigated from the three websites SignalP 4.1, PROTTER and PrediSi. Among 18 proteins 14 proteins (NP_252456.1, NP_252782.1, NP_253095.1, NP_253326.1, NP_253368.1, NP_249450.1, NP_250846.1, NP_252171.1, NP_252375.1, NP_253374.1, NP_253455.1, NP_253678.1, NP_253679.1 and NP_253685.1) do not contain any signal peptide and one protein contain signal peptide unanimously. Whereas the remaining proteins (NP_253252.1, NP_251676.1 and NP_253434.1) are containing signal peptides from any of a website (**supplementary table 1**).

3.3 Physicochemical properties Analysis

We have searched for the physicochemical properties of 18 EHPs which is shown in Table 3. All the proteins had molecular weight ranging from 13335.11 to 56122.54. The highest molecular weight was observed to be 56122.54 for the NP_253252.1 protein, a probable lipid II flippaseMurJ [61]. The theoretical pI (Isoelectric Point) indicates the pH at which the charge of an amino acid of a protein remains neutral. Therefore, no movement occurs when placed in an electric field with a direct current. This parameter comes in handy as proteins are dense and stable at an isoelectric pH [62]. The theoretical pI ranged from 4.52 to 10.71. Both of these parameters (molecular weight and theoretical pI) help visualize the Two-dimensional gel electrophoresis or (2-DE). Hence contributes to the scientific examinations of these hypothetical proteins [63]. The aliphatic index can be an effective indicator for determining the thermostability of some protein molecules [64]. A protein molecule with a higher aliphatic index indicates its higher range of temperature at which it gains its thermostability [65]. The aliphatic index tabulated for our protein group ranged from 83.13 to 133.96. The NP_253252.1 protein showed the maximum thermostability and NP_252456.1 with the lowest. The parameter called Instability index determines a protein whether it's stable or unstable in a test tube [66]. For our analysis, we set the cutoff value to 40 where the value below 40 indicates a protein to be stable and above 40 predicts it as an unstable protein. Total 9 proteins (NP_252456.1, NP_252782.1, NP_253252.1, NP_253326.1, NP_249450.1, NP_251676.1, NP_252171.1, NP_252375.1, NP_253679.1) out of 18 proteins of interest found to be stable with Instability index values of 25.65, 37.46, 36.08, 38.61, 31.65, 38.17, 32.05, 29.03, 32.97 respectively. The grand average of hydropathy (GRAVY) determines the extent of protein-water interaction which is calculated by dividing the aggregate of all the amino acids hydropathy values with the total number of residues in the given sequence [65, 67]. the GRAVY values lied between − 0.427 to 0.857. The lower the GRAVY value, the more a protein interacts with water [65]. The NP_253095.1 protein was found to be most interactive among all these proteins having a GRAVY value of -0.427.

Table 3
Physicochemical properties of 18 essential hypothetical proteins

Serial No.	RefSeq.	Molecular weight	Theoretical pI	Formula	Total number of negatively charged residues (Asp + Glu)	Total number of positively charged residues (Arg + Lys)	Instability index (II)	Aliphatic index	Grand average of hydropathicity (GRAVY)
1	NP_252456.1	19937.91	9.12	$C_{869}H_{1409}N_{265}O_{255}S_9$	22	26	25.65 (Stable)	83.13	-0.257
2	NP_252782.1	14871.24	7.93	$C_{658}H_{1078}N_{188}O_{193}S_5$	15	16	37.46 (Stable)	100.29	0.162
3	NP_253095.1	15057.34	10.71	$C_{657}H_{1080}N_{210}O_{188}S_4$	12	21	57.65 (Unstable)	93.28	-0.427
4	NP_253252.1	56122.54	10.03	$C_{2643}H_{4201}N_{651}O_{651}S_{19}$	21	40	36.08 (Stable)	133.96	0.857
5	NP_253326.1	43779.82	6.85	$C_{1955}H_{3058}N_{554}O_{569}S_{11}$	51	50	38.61 (Stable)	87.02	-0.376
6	NP_253368.1	24873.64	5.30	$C_{1111}H_{1779}N_{317}O_{319}S_6$	28	25	73.37 (Unstable)	91.89	-0.119
7	NP_249450.1	33667.59	5.37	$C_{1492}H_{2415}N_{425}O_{446}S_7$	39	32	31.65 (Stable)	108.22	0.057
8	NP_250659.1	13335.11	8.98	$C_{567}H_{936}N_{178}O_{181}S_6$	12	15	53.34 (Unstable)	83.54	-0.085
9	NP_250846.1	27693.92	9.90	$C_{1245}H_{1962}N_{380}O_{330}S_5$	23	30	52.29 (Unstable)	100.69	-0.229
10	NP_251676.1	47387.94	9.69	$C_{2139}H_{3484}N_{582}O_{587}S_{20}$	34	44	38.17 (Stable)	114.85	0.365
11	NP_252171.1	38888.77	5.26	$C_{1711}H_{2780}N_{482}O_{517}S_{16}$	40	31	32.05 (Stable)	102.34	0.090
12	NP_252375.1	24180.71	5.02	$C_{1081}H_{1707}N_{303}O_{313}S_7$	27	19	29.03 (Stable)	102.48	0.166
13	NP_253374.1	26354.58	4.52	$C_{1166}H_{1811}N_{315}O_{366}S_8$	43	19	53.53 (Unstable)	89.96	-0.366
14	NP_253434.1	17171.46	4.59	$C_{763}H_{1208}N_{206}O_{236}S_4$	27	15	57.54 (Unstable)	105.07	-0.197
15	NP_253455.1	16000.46	6.72	$C_{720}H_{1130}N_{190}O_{208}S_7$	14	14	43.00 (Unstable)	88.12	-0.037
16	NP_253678.1	42109.34	7.73	$C_{1866}H_{3011}N_{551}O_{541}S_9$	47	48	48.61 (Unstable)	98.72	-0.130
17	NP_253679.1	29030.07	6.00	$C_{1281}H_{2067}N_{373}O_{386}S_5$	36	31	32.97 (Stable)	101.22	-0.101
18	NP_253685.1	24985.76	9.60	$C_{1112}H_{1791}N_{337}O_{311}S_4$	27	34	45.71 (Unstable)	104.77	-0.365

3.4 Protein-protein interaction network analysis

The PPI represents the connection among the 9 stable EHPs and their corresponding functionally relative proteins from *P. aeruginosa* PA01. The network has 261 nodes and 269 edges for 9 proteins of interest. Here, the network is provided with 11 subnetworks (Hubs) with a minimum of 3 nodes each. The nodes with only 3 connections (Degree) are considered as Islands (ostA and PA1847) (Table 4) [68]. The node degree and Betweenness centrality range from 3 to 45

and 4750 to 17881.76, respectively. The interaction among the hub proteins can be seen in Fig. 3. The size and color gradient of the nodes determine the degree of a protein. A node degree reveals the extent of interaction of a particular node with other nodes. The nodes with lower degree values are colored green namely PA4992 (24), PA3481 (23), PA4093 (20), PA4636 (18). The color gradually turned into deep purple by the increase of node degree values. Nodes with enlarged size similarly denotes increased node degree values such as PA2986 (45), PA0759 (41), PA4562 (38), PA3685 (32), PA3767 (28) (Fig. 3) (Table 4). The nodes in cyan blue meaning 2 or more interactions with their corresponding subnetworks. Betweenness centrality is a topological measure that typically determines the number of shortest paths through nodes. The nodes with a higher degree and betweenness centrality values represent vital proteins for signal trafficking of the cellular system [68]. The function of all proteins in the network are collected from NCBI using their associated Entrez IDs and listed in the **supplementary table 2**.

Table 4
List of proteins with their Reference sequence, Uniprot ID, Protein name, Node degree value, and Betweenness centrality.

Serial No.	Ref seq.	UniProt ID	Protein name	Degree	Betweenness centrality
1	NP_251676.1	Q9HZL8	PA2986	45	9681.83
2	NP_249450.1	Q9I5H0	PA0759	41	17881.76
3	NP_253252.1	Q9HVM2	PA4562	38	9315.74
4	NP_252375.1	Q9HXV5	PA3685	32	8742.58
5	NP_252456.1	Q9HXM8	PA3767	28	11489.25
6	NP_253679.1	Q9HUH3	PA4992	24	10103.17
7	NP_252171.1	Q9HYC8	PA3481	23	5329.08
8	NP_252782.1	Q9HWT5	PA4093	20	4750.0
9	NP_253326.1	Q9HVF5	PA4636	18	6466.58
10	NP_249286.1	Q9I5U2	ostA	3	10848.52
11	NP_250538.1	Q9I2P8	PA1847	3	5698.26

3.5 Non-homology analysis against human proteome, human anti-targets and human gut flora proteomes

To introduce a novel target for a drug it must be non-homologous against human proteome, human anti-targets, human gut flora proteomes. Utilizing pipeline builder from the Pipeline builder for identification of target (PBIT) server, 9 protein sequences were inputted to find out the highly similar sequence with human proteome. Among the 9 EHPs sequences, one sequence was homologous with the human proteome. Filtering that one sequence, 8 non-homologous proteins were selected for the next pipeline analysis to find out the non-homologous proteins against human anti-targets. Among that 8 entered sequences, significant similar sequences of human anti-target proteins were screen out. This result depicted that 7 proteins are non-homologous and one protein is homologous to the human anti-target where this one homologous protein was omitted from the study. After the filtration, selected 7 proteins were further inputted onto the pipeline builder to analyze non-homologous proteins against human gut flora proteomes. This time 2 proteins were screen out because of containing high sequence similarity with the proteomes of the beneficiary microbes belong to the human gut. Then finally the sequences of 5 non-homologous EHPs were selected for the next parameter of finding virulence capability. The details of the non-homology analysis are given in Table 5.

Table 5
Aspects of the proteins like non-homology to human proteins and proteins of human gut flora, virulence of the pathogen, druggability for the 9 EHPs

Serial no	*Protein*	*Non-homology analysis against human proteome*	*Non-homology analysis against human anti-targets*	*Non-homology analysis against gut microbiota proteomes*	*Virulence analysis*	*Druggability analysis*
1	NP_252456.1	Non-homologous	Non-homologous	Non-homologous	Non-virulent	Old target
2	NP_252782.1	Non-homologous	Non-homologous	Non-homologous	Non-virulent	Novel target
3	NP_253252.1	Non-homologous	Non-homologous	Homologous	Non-virulent	Novel target
4	NP_253326.1	Non-homologous	Non-homologous	Non-homologous	Non-virulent	Novel target
5	NP_249450.1	Non-homologous	Non-homologous	Non-homologous	Virulent	Novel target
6	NP_251676.1	Non-homologous	Non-homologous	Non-homologous	Virulent	Novel target
7	NP_252171.1	Homologous	Non-homologous	Non-homologous	Non-virulent	Novel target
8	NP_252375.1	Non-homologous	Non-homologous	Homologous	Non-virulent	Novel target
9	NP_253679.1	Non-homologous	Non-homologous	Non-homologous	Non-virulent	Old target

3.6 Virulence factor

The virulent EHPs from *P. aeruginosa* PA01 are enlisted in Table 5. VICMpred is a Support Vector Machine (SVM) based webserver that predicted all of the 9 EHPs as non-virulent with 70.75% accuracy [43]. VirulentPred is also based on bi-layer cascade SVM with five-fold increased cross-validation methods that give 81.8% prediction accuracy [44]. Total 3 proteins namely NP_249450.1 (e-106), NP_251676.1 (e-171), NP_253679.1 (7e-77) were predicted as virulent by VirulentPred in *p. aeruginosa* PA01 strain utilizing the Similarity-Based search through PSI-BLAST. Another webserver called MP3 uses an integrated SVM-HMM approach which commonly predicted NP_251676.1 as a pathogenic protein.

3.7 A Possible New Drug Target Identification

Along with the two virulent EHPs, other 7 sequences of EHPs were employed to the DrugBank server for the identification of potentially new drug candidates. This server showed that NP_252456.1 contains one drug target against the drug Imidazole (E value: 5.62144e-18; Bit score: 75.485; Query length: 182; Alignment length: 77) and two drug targets were exhibited by the protein NP_253679.1 against the drug Nicotinamide adenine dinucleotide phosphate (E value: 3.79487e-15; Bit score: 72.4034; Query length: 270; Alignment length: 213) and Nicotinamide adenine dinucleotide phosphate (E value: 1.77261e-14; Bit score: 70.8626; Query length: 270; Alignment length: 217). The remaining 7 proteins (NP_252782.1, NP_253252.1, NP_253326.1, NP_249450.1, NP_251676.1, NP_252171.1 and NP_252375.1) was considered as a fresh or new drug target by the DrugBank database. This website also revealed that our targeted two proteins named NP_249450.1 and NP_251676.1 displayed zero matches for the drug target which means they are new potential drug candidates with druggability. The overall results are in Table 5.

3.8 Analyzing Secondary Structure

Based on the findings of segment III, we selected two proteins for the next level investigations that match all the criteria of segment III. As they are hypothetical proteins, they must lack some information. For this, to suggest them as a new drug target we explored their secondary structure. SOPMA and PSIPRED were the web tools that were used for the secondary structure analysis. According to the SOPMA server, the secondary structure of NP_249450.1 had Alpha helix (Hh): 126 (40.13%); Extended strand (Ee): 57 (18.15%); Beta turn (Tt): 20 (6.37%), and Random coil (Cc): 111 (35.35%) where the parameters were set as Window width: 17; Similarity threshold: 8 and Number of states: 4. The protein, NP_251676.1 had Alpha helix (Hh): 206 (47.58%); Extended strand (Ee): 83(19.17%); Beta turn (Tt): 23(5.31%) and Random coil (Cc): 121 (27.94%) with the same parameter as before. The results from SOPMA database for both proteins are given in **supplementary table 3**. The PSIPRED sequence plot and PSIPRED cartoon plot were provided as a result of the PSIPRED web servers. The sequence plot and cartoon plot structure described that goldenrod (semi-yellow) color is for the extracellular strand domain, pink color is for helix; grey color is for coil and blackish blue is for the confidence of the structure. According to this, NP_249450.1 showed more coil in its secondary structure whereas NP_251676.1 showed more helix. Figure 4 **(a)** is the secondary structure of NP_249450.1 from PSIPRED and Fig. 4 **(b)** is from SOPMA websites. Besides Fig. 5 **(a)** is the secondary structure of NP_251676.1 from PSIPRED and Fig. 5 **(b)** is from SOPMA database.

3.9 Essential hypothetical proteins 3D structure modelling

Only proteins that passed all of the above-mentioned pipeline analyses were assigned a three-dimensional structural conformation. Two proteins, NP_249450.1 and NP_251676.1, were subjected to a thorough pipeline review and thus have the potential to be used as new drug targets. As a result, these two proteins were subjected to 3D structural conformation determination using two methods: template-based homology modelling from SWISS-MODEL and *ab-initio* modeling using the trRosetta algorithm from the Robetta server. For template-based homology modelling, we searched for templates from SWISS-MODEL for these two proteins. 1vly.1 and 6f3z.2 were the best template for NP_249450.1 and NP_251676.1, respectively. The templates were chosen based on several parameters, including the Global Model Quality Estimation (GMQE), Qualitative Model Energy ANalysis (QMEAN), Z-score, sequence identity, sequence similarity, sequence coverage, oligo-state of the chosen templates, and so on. The template 1vly.1 was actually a 1.30 Å resolution x-ray diffraction

crystallography structure of a putative aminomethyltransferase (ygfz) from *E. coli*. This template shared 30.23% sequence identity with the 314 aa long NP_249450.1 protein, which spans from (4-307) aa. The template 6f3z.2, on the other hand, was a complex of *E. coli* LolA and the periplasmic domain of LolC that was also identified by x-ray diffraction crystallography at a resolution of 2.00 Å. The sequence identity was 30.73%, spanning (67–290) amino acids out of the 433 amino acids in the NP_251676.1 protein. Both of these templates had a monomer oligo-state. Finally, using 1vly.1 and 6f3z.2 templates, the structures of NP_249450.1 and NP_251676.1 EHPs were formed, as shown in Fig. 6a and 6c. Structure prediction by Robetta server illustrated that the provided model was build using trRefineRosetta modelling (*ab-initio* modeling using the trRosettaalgorithm). Structure of NP_249450.1 showed 0.79 score as confidence (Fig. 6b) while NP_251676.1 showed 0.81 score as confidence (Fig. 6d). Consequently, the structures from SWISS-MODEL were then refined from Galaxy Refiner where model 2 for NP_249450.1 and Model 5 for NP_251676.1 were downloaded after final refinement. For the NP_249450.1 and NP_251676.1 proteins, the lowest MolProbity was 1.738 (Model 2) and 1.729 (Model 5), respectively, while the initial score was 2.280 and 2.299. Also, the structures from Robbetta were refined from Galaxy Refiner.

3.10 Protein structure validation assessment

The predicted protein structure was validated by SAVES v6.0 server which runs six programs simultaneously to evaluate the quality of the build model. The ERRAT value served as the model's overall quality element. The overall quality factor for the NP_249450.1 protein structure from SWISS-MODEL and Robetta, respectively, was 87.6325% and 92.459%. It was 93.3649% and 98.063% for NP_251676.1 from these two servers, respectively. In **supplementary Fig. 1**, bar plots depict the overall quality factor from ERRAT. VARIFY3D conducts an analysis in which a structure passes if at least 80% of the amino acids in the 3D/1D profile have a score of > = 0.2. Three of the four structures passed this parameter (two from SWISS-MODEL and one from Robetta), while one structure failed for NP_251676.1 from Robetta. WHATCHECK included a color box with a number within it that reflects 46 different criteria, with the green, yellow, and maroon colors representing OK, warning, and error, respectively. The overall summary report is OK for all four structures. Table 6 included a comprehensive report on the consistency of the four structures that we retrieved from the SAVES v6.0 server. On the contrary, structures from SWISS-MODEL failed to pass the PROVE parameters, while structures from Robetta were placed in warning categories due to atomicB-factors, and the protein atoms having absolute Z-scores > 3. Ramachandran plot analysis from the PROCHECK program also demonstrated that more than 94% of residues were in the most favored region for all four structures from both SWISS-MODEL and Robetta. It was 97.3% for NP_251676.1 protein from Robetta, with 0.0% residues in the disallowed region. Ramachandran plot analysis unveiled that the generated structures of the proteins represent an excellent degree of validity and reliability, which is depicted in Fig. 7.

Table 6: Three-dimensional structure validation of the predicted two hypothetical proteins from SAVES v6.0 server

Saves Result *Protein Name*	*ERRAT*	*VARIFY 3D*	*PROVE*	*WHATCHECK*	*PROCHECK*	*Ramachandran Plot* *(% residue in the most favored region)*
NP_249450.1 (Swiss Model)	Overall Quality Factor 87.6325	97.70% of the residues have averaged 3D-1D score >= 0.2 Pass	Buried outlier protein atoms total from 1 Model: 6.1% fail	1234567891011 12131415161718 19202122232425 26272829303132 33343536373839 40414243444546	Out of 8 evaluations Errors: 3 Warning: 2 Pass: 3	94.6%
NP_249450.1 (Robetta)	Overall Quality Factor 92.459	94.90% of the residues have averaged 3D-1D score >= 0.2 Pass	Buried outlier protein atoms total from 1 Model: 4.1% warning	1234567891011 12131415161718 19202122232425 26272829303132 33343536373839 40414243444546	Out of 8 evaluations Errors: 3 Warning: 2 Pass: 3	94.0%
NP_251676.1 (Swiss Model)	Overall Quality Factor 93.3649	82.59% of the residues have averaged 3D-1D score >= 0.2 Pass	Buried outlier protein atoms total from 1 Model: 5.4% fail	1234567891011 12131415161718 19202122232425 26272829303132 33343536373839 40414243444546	Out of 8 evaluations Errors: 2 Warning: 4 Pass: 2	95.3 %
NP_251676.1 (Robetta)	Overall Quality Factor 98.063	66.97% of the residues have averaged 3D-1D score >= 0.2 Fail	Buried outlier protein atoms total from 1 Model: 4.2% warning	1234567891011 12131415161718 19202122232425 26272829303132 33343536373839 40414243444546	Out of 8 evaluations Errors: 2 Warning: 1 Pass: 5	97.3%

4. Discussion

P. aeruginosa PA01 is an omnipresent pathogenic bacterium that can cause acute and chronic infection to humans by contaminating environmental water and food, daily food spoilage, and infections. It is a rising concern for its increasing resistance against a broad range of antimicrobials. The biofilm-forming ability and evolution of antibiotic tolerance shapes *pseudomonas* isolate highly resistant against imipenem (95.3%), trimethoprim-sulfamethoxazole (69.8%), aztreonam (60.5%), chloramphenicol (45.3%), and meropenem (27.9%) [69]. Factors like chromosomal mutations and transferring of resistant genes through horizontal gene transfer contribute to its broad-spectrum drug resistance property [70]. Thus, it is necessary to introduce a new drug target when there will be noticed multi-drug resistance for any diseases or problem. *In silico* process has a great advantage for the identification of new drug targets in that situation within a very short time. Consequently, to combat the ever-increasing danger of antibiotic resistance, identifying novel drug targets is a dire necessity. A drug target should have some properties before it is considered as a new target which includes being non-homologous to human proteome, human anti-targets, human gut microbiota, having virulence capability, having druggability, and so on. For this reason, we scrutinized the properties of our targeted essential hypothetical proteins where analyzing these EHPs from multidrug resistance bacteria can lead to the identification of new potential therapeutic solutions.

We searched for essential hypothetical proteins (EHPs) among the 336 essential proteins of this bacterial strain to meet this need. Essential genes/proteins are those that are vital for a pathogen's survival and thereby analyzing their functions and metabolic pathways, crucial information that may be central to life can be retrieved. In this research, we discovered 18 EHPs for the first time that may provide valuable information about the pathogenesis, molecular mechanisms, and functions of this bacteria. Functional annotation is a prerequisite in understanding the pathogen metabolic pathways and the products that they synthesize for their survival in adverse conditions. Moreover, domain analysis, which is a basic, distinctive, and stable unit of a protein structure that is fiercely conserved during the evolutionary process, is crucial for further investigation [71]. Moreover, the function of a protein is directly or indirectly related to

the subcellular localization [72]. The physicochemical properties of a protein depict a chemical assessment that shows the identity of chemical nature, physical hazards and to understand or predict molecular attributes. The combined analysis of the physicochemical properties helps to characterize the proteins annotated as hypothetical proteins from the genome of an opportunistic pathogen like *P. aeruginosa* PAO1(Table 3) [73].

In this study, the PPI network has provided congruent meaningful insights into the protein's function. Here, we have looked for potential relativity to our predicted function of EHPs and their connectivity with proteins involved with functionally important activities. The protein PA2986 (NP_251676.1) related to the MacB-like periplasmic core domain, represents a connection with 45 proteins of which 8 are hypothetical proteins (HP). A notable number of proteins grouped with PA2986 are involved in protein translocation activities such as translocation protein TolQ, TolR and TolB. Tol proteins show activity in gram-negative bacteria by providing stability to the outer membrane [74]. Moreover, ABC transporter ATP-binding protein (PA0073) is related to it as it functions by utilizing TolC exit duct by shifting substrates to extracellular space from the periplasm [75]. This finding supports the idea of PA2986 being a member of the MacB-like periplasmic core domain. Another important protein for bacterial survival lysS, a lysine tRNA ligase was found to interact with PA2986 which is a mutant in some gram-negative bacteria conferring resistance against the OP0595, diazabicyclooctane b-lactamase inhibitor (an antibiotic) [76]. We have found Penicillin-Binding Protein 1 called ponA protein in this group of networks. Alteration in the ponA protein (penicillin-binding protein 1A) has a significant role in harnessing Chromosomally mediated resistance against penicillin in *N. gonorrhoeae* [77]. Similarly, other considerable proteins like outer-membrane lipoprotein carrier protein lolA, transporter ExbB, penicillin-binding protein 1A (ponA) interacted with PA2986 Fig. 3.

PA0759 (NP_249450.1) has got the 2nd largest degree value having 41 nodes (Table 4) in connection of which 17 are HPs. The highest betweenness centrality value of 17881.76 determines its significance towards cell signaling pathways as in the case of directed or regulated networks, Betweenness centrality is considered to be a much robust essentiality indicator than degree value [78]. Genes in this hub include proteins having prime roles in translational regulation and cellular metabolic activities such as glycine cleavage system protein T2 [79], translation elongation factor (tsf) [80], and ribosomal large subunit pseudouridine synthase C (rluC) [81], respectively. Moreover, RecO protein in this network is a replication repairing protein from the RecF recombination repair pathway that facilitates both DNA strand annealing and DNA recombination in complex with RecA protein found in high radiation tolerant bacteria *Deinococcus radiodurans* [82]. This property may also contribute to better survival efficacy for *P. aeruginosa* PA01 in extreme conditions.

The PA4562 (NP_253252.1) protein is a probable member of the Lipid II flippase MurJ family which is used for the genesis of lipid II on both inner and outer leaflets and that ultimately produces peptidoglycan in almost every bacterial species. Peptidoglycan is the primary protective foundation for shielding against environmental hazards and is involved in cell wall organizations [83]. Proteins related with morphological importance in bacteria such as flagellar basal body rod protein (FlgC) [84], rod shape-determining protein (rodA) [85], type 4 fimbrial biogenesis outer membrane protein (PilQ) [86] are present in a connection with the PA4562 proteins that strengthen our prediction regarding this protein function.

The proteins PA3685 (NP_252375.1) and PA3767 (NP_252456.1) are adjoined with 12 and 6 HPs respectively. Both of these proteins are largely involved with enzymes of different molecular functions-tRNA N6-adenosine threonyl carbamoyl transferase (gcp) is a universal structural modifier found at position 37 of tRNAs that provides the anticodon loop with greater binding efficiency to ribosomes invitro in *E. coli* [87]. The protein UDP-2,3-diacyl glucosamine hydrolase (PA1792) is hypothesized to be catalyzing lipid-A biogenesis in *E. coli* bacteria [88]. Lipid-A is a saccharolipid that modulates lipopolysaccharide (LPS) anchorage on the outer leaflet of the outer membrane in gram-negative bacteria which is an essential component for the bacteria shielding from antibiotics and sustain its viability [89]. Besides other proteins having enzymatic properties include thiamine monophosphate kinase (thiL), ATP-dependent DNA helicase DinG (PA1045), riboflavin-specific deaminase/reductase (ribD), amidotransferase(PA1742), acetyltransferase (PA2631) Fig. 3.

The majority of the proteins connected with PA4992 (NP_253679.1) from aldo/keto reductase family are uncharacterized proteins. The PA4167 protein has a contributing role as a source of carbon and energy for a large number of bacteria [90]. The hub proteins PA4093 (NP_252782.1) and PA4992 (NP_253679.1) are interconnected through the intermediate protein PA4098, a probable short-chain dehydrogenase enzyme [91].

The mgtE is an Mg transporter that represents a connection with the PA3481 (NP_252171.1) which can be an opportunistic inhibitor for the type III secretion system (T3SS). T3SS is a formidable toxin injected by *P. aeruginosa* that can ultimately cause cell death into its host. The mgtE interrupts with the T3SS transcription regulation system by provoking rsmYZ gene transcription hence inhibits T3SS protein expression [92]. Interestingly another analogous protein, DNA polymerase II (polB) is associated with this same hub protein PA3481. PolB functions as a crucial candidate for repressing the translation process of master T3SS regulator ExsA. ExsA operates a major role in maintaining the regulatory cascade of T3SS. Thus affecting ExsA expression can prohibit T3SS toxin secretion process. Furthermore, S Chakravarty *et al*., found that T3SS transcription is attenuated when polB is overexpressed. Therefore, polB may act as a promising target for therapeutic interventions [93]. Besides, proteins responsible for exopolysaccharide biosynthesis and biofilm formation namely pslA [94] and pslD [95] are both members of the psl operon. The presence of such virulent protein types in this protein hub suggests PA3481 as a crucial protein involved in multiple virulence pathways in *P. aeruginosa*. The PA4636 (NP_253326.1) protein harbors some of the virulent proteins like lptA and algQ that is required for the biogenesis of lipid bilayer in the outer membrane in *P. aeruginosa* [96] and facilitates in developing a chronic infection in cystic fibrosis [97].

Some notable mutual interactions also have been observed between two hub proteins like PA2986 and PA4562 where interrelated proteins include – mraY, a potential target for antibiotic development is a crucial element for the bacterial cell wall synthesis [98]; opr86, an outer membrane protein found previously in all gram-negative bacteria. Likewise suggested as a potential drug target with significant therapeutic potential against *P. aeruginosa* in earlier studies [99]; rpoH, a 32-kDa heat shock protein in *E. coli* can also take part as a complementary for sigma factor during the increasing temperature in the environment as well as while starving [100]; PA5568 possess an inner membrane translocation subunit protein YidC which facilitates proteins to be passed onto inner membranes without the help of Sec translocase complex proteins [101]; ComL is a lipoprotein that facilitates the DNA transformation process in *N. gonorrhoeae* [102]. Lastly, organic solvent tolerance protein OstA holds interaction simultaneously with the top 3 hub genes of maximum node connection. Concurrently, OstA is a protein of high molecular significance as it is found in almost all gram-negative bacteria and is involved in the bacterial envelope biogenesis process. A study by HC Chiu *et al*., found that OstA deficiency in *Helicobacter pylori* causes sensitivity to organic solvent, impaired membrane

permeability, and vulnerability to antibiotics [103]. The function of the proteins in this network shows relational integrity with our predicted HPs. Knowing the protein's function in a protein-protein interaction network can facilitate the process of discovering the proteins with unknown functions [104].

Herein, we analyzed these properties of our targeted essential hypothetical proteins where PBIT servers direct categorized all the non- homology features (Table 5). We have selected NP_249450.1 and NP_251676.1 respectively for being virulent determined by two of our tools with strong confidence scores. Targeting these virulent factors can limit the pathogenicity of *P. aeruginosa*. Even antivirulence drugs insist a pathogen towards a weaker selection for resistance in them compared to antibiotics [11]. Therefore, understanding the virulence factors and their role in pathogenesis can lead us to a new potential therapeutic solution. Besides, druggability analysis also confirmed that NP_249450.1 and NP_251676.1 can be a new and potential drug targets.

Structure prediction and quality assessment of the predicted structure are also parallelly important to evaluate the molecular and biological functions of a protein in cells for in-depth analysis and drug target identification [105]. Before structure prediction, the information of Alpha helix (Hh), Extended strand (Ee), Beta turn (Tt), or Random coil (Cc) helps to establish the secondary structure. That's why the secondary structure was annotated to complete the structure related to all the information of our selected two proteins (Fig. 4, Fig. 5, **and supplementary table 3**). Thus, we further predicted the 3D structure of two EHPs (NP_249450.1 and NP_251676.1) and assessed the quality of these structures to decipher their unique conformation. The 3D structure and Ramachandran plot are depicted in Fig. 6 and Fig. 7, respectively. The predicted structure's quality assessment parameters are listed in Table 6, and the overall quality factor is shown in **Supplementary Fig. 1**. Our predicted structure is accurate and reliable, according to Ramachandran plot analysis, since more than 90% of residues are considered the cutoff value and our findings surpass that range by more than **94%.** Therefore, this structural and functional information will open a new window for further identification of potential drug candidates that can halt the surge of this pathogenic bacterium from becoming resistant.

5. Conclusion

Unveiling the functional characterization of pathogenic microorganisms is of great importance in biological processes and medical science. Essential proteins and essential hypothetical proteins are versatile macromolecules that can be crucial in inferring new treatment strategies towards these pathogenic bacteria. For functional characterization of EHPs, we used an *in-silico* approach in combination with different bioinformatics databases/tools, with ROC analysis indicating that these tools are highly reliable for functional characterization of *P. aeruginosa* PA01. We attributed function to 18 EHPs and analyzed subcellular localization and physiochemical properties of these proteins. Afterward, a PPIs network analysis was carried out on 9 stable EHPs and their functionally related proteins from this bacterium. Further, host non-homologous analysis predicts 5 pathogen-specific proteins, three of which have virulent factors that could be used as novel therapeutic targets. Finally, the structural conformation of two EHPs was determined, and the accuracy of the predicted model was evaluated, indicating that this model is highly accurate. Our findings will pave the way for new antibacterial drugs and treatment strategies to be developed by focusing on these novel drug targets.

Declarations

Funding

There was no significant funding support for this investigation.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Klockgether J, Tümmler B (2017) Recent advances in understanding Pseudomonas aeruginosa as a pathogen. F1000Research, p 6
2. Diggle SP, Whiteley M, Profile M (2020) Pseudomonas aeruginosa: opportunistic pathogen and lab rat. Microbiology 166(1):30
3. Nikolaidis M, Mossialos D, Oliver SG et al (2020) Comparative Analysis of the Core Proteomes among the Pseudomonas Major Evolutionary Groups Reveals Species-Specific Adaptations for Pseudomonas aeruginosa and Pseudomonas chlororaphis. Diversity 12(8):289
4. Stover CK, Pham XQ, Erwin A et al (2000) Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature 406(6799):959–964
5. Pachori P, Gothalwal R, Gandhi P (2019) Emergence of antibiotic resistance Pseudomonas aeruginosa in intensive care unit; a critical review, vol 6. Genes & diseases, pp 109–119. 2
6. Dubern JF, Cigana C, De Simone M et al (2015) Integrated whole-genome screening for P seudomonas aeruginosa virulence genes using multiple disease models reveals that pathogenicity is host specific. Environ Microbiol 17(11):4379–4393
7. Azam MW, Khan AU (2019) Updates on the pathogenicity status of Pseudomonas aeruginosa. Drug discovery today. 24:350–3591
8. Periasamy S, Nair HA, Lee KW et al (2015) Pseudomonas aeruginosa PAO1 exopolysaccharides are important for mixed species biofilm community development and stress tolerance. Front Microbiol 6:851
9. Schleheck D, Barraud N, Klebensberger J et al (2009) Pseudomonas aeruginosa PAO1 preferentially grows as aggregates in liquid batch cultures and disperses upon starvation. PLoS ONE 4(5):e5513
10. Klockgether J, Cramer N, Wiehlmann L et al (2011) Pseudomonas aeruginosa genomic structure and diversity. Front Microbiol 2:150
11. Prava J, Pranavathiyani G, Pan A (2018) Functional assignment for essential hypothetical proteins of Staphylococcus aureus N315. Int J Biol Macromol 108:765–774

12. Luo H, Lin Y, Gao F et al (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res 42(D1):D574–D580
13. Araujo FA, Barh D, Silva A et al (2018) GO FEAT: a rapid web-based functional annotation tool for genomic and transcriptomic data. Scientific reports. 8:1–41
14. Geer LY, Domrachev M, Lipman DJ et al (2002) CDART: protein homology by domain architecture. Genome Res 12(10):1619–1623
15. Letunic I, Khedkar S, Bork P (2021) SMART: recent updates, new developments and status in 2020. Nucleic acids research. 49:D458–D460D1
16. Gough J, Karplus K, Hughey R et al (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313(4):903–919
17. Mistry J, Chuguransky S, Williams L et al (2021) Pfam: The protein families database in 2021. Nucleic acids research. 49:D412–D419D1
18. Li YH, Xu JY, Tao L et al (2016) SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. PLoS ONE 11(8):e0155290
19. Sillitoe I, Bordin N, Dawson N et al (2021) CATH: increased structural coverage of functional space. Nucleic Acids Res 49(D1):D266–D273
20. Blum M, Chang H-Y, Chuguransky S et al (2021) The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49(D1):D344–D354
21. Gabler F, Nam SZ, Till S et al (2020) Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protocols Bioinf 72(1):e108
22. Koskinen P, Törönen P, Nokso-Koivisto J et al (2015) PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics 31(10):1544–1552
23. Hawkins T, Chitale M, Luban S et al (2009) PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data, vol 74. Structure, Function, and Bioinformatics, Proteins, pp 566–582. 3
24. Chitale M, Hawkins T, Park C et al (2009) ESG: extended similarity group method for automated protein function prediction. Bioinformatics 25(14):1739–1745
25. Yu NY, Wagner JR, Laird MR et al (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26(13):1608–1615
26. Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein science. 13:1402–14065
27. Yu CS, Chen YC, Lu CH et al (2006) Prediction of protein subcellular localization. Proteins Struct Funct Bioinform 64(3):643–651
28. Krogh A, Larsson B, Von Heijne G et al (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580
29. Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338(5):1027–1036
30. Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17(9):849–850
31. Dobson L, Reményi I, Tusnády GE (2015) CCTOP: a Consensus Constrained TOPology prediction web server. Nucleic Acids Res 43(W1):W408–W412
32. Omasits U, Ahrens CH, Müller S et al (2014) Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics 30(6):884–886
33. Nielsen H *Predicting secretory proteins with SignalP*, in *Protein function prediction*2017,Springer. p.59–73
34. Hiller K, Grote A, Scheer M et al (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 32(suppl2):W375–W379
35. Ganapathiraju M, Balakrishnan N, Reddy R et al (2008) Transmembrane helix prediction using amino acid property features and latent semantic analysis. Bmc Bioinformatics. Springer
36. Gasteiger E, Gattiker A, Hoogland C et al (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31(13):3784–3788
37. Shahbaaz M, ImtaiyazHassan M, Ahmad F (2013) Functional annotation of conserved hypothetical proteins from Haemophilus influenzae Rd KW20. PloS one. 8:e8426312
38. Naqvi AAT, Shahbaaz M, Ahmad F et al (2015) Identification of functional candidates amongst hypothetical proteins of Treponema pallidum ssp. pallidum PloS one 10(4):e0124177
39. Zhou G, Soufan O, Ewald J et al (2019) NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res 47(W1):W234–W241
40. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
41. Shende G, Haldankar H, Barai RS et al (2017) Pipeline Builder for Identification of drug Targets for infectious diseases. Bioinformatics 33(6):929–931
42. Valdes AM, Walter J, Segal E et al (2018) Role of the gut microbiota in nutrition and health.Bmj,361
43. Saha S, Raghava G (2006) VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition, vol 4. Genomics, proteomics & bioinformatics, pp 42–47. 1
44. Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 9(1):1–12
45. Gupta A, Kapil R, Dhakan DB et al (2014) MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLoS ONE 9(4):e93907

46. Emmerich CH, Gamboa LM, Hofmann MC et al (2020) Improving target assessment in biomedical research: the GOT-IT recommendations.Nature Reviews Drug Discovery, : p.1–18
47. Wishart DS, Feunang YD, Guo AC et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46(D1):D1074–D1082
48. Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics 11(6):681–684
49. Buchan DW, Jones DT (2019) The PSIPRED protein analysis workbench: 20 years on. Nucleic Acids Res 47(W1):W402–W407
50. Waterhouse A, Bertoni M, Bienert S et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46(W1):W296–W303
51. Yang J, Anishchenko I, Park H et al (2020) Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117(3): p. 1496–1503
52. HeeShin W (2014) Prediction of protein structure and interaction by GALAXY protein modeling programs. Biodesign 2:1–11
53. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2(9):1511–1519
54. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164–170
55. Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356(6364):83–85
56. Pontius J, Richelle J, Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264(1):121–136
57. Hooft RW, Vriend G, Sander C et al (1996) Errors in protein structures. Nature 381(6580):272–272
58. Laskowski RA, MacArthur MW, Moss DS et al (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26(2):283–291
59. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition. 30:1145–11597
60. Eng J (2017) *ROC analysis: web-based calculator for ROC curves. Baltimore: Johns Hopkins University [updated 2014 March 19]*, Accessed 25/05/17. Available from: http://www. jrocfit. org
61. Kuk AC, Mashalidis EH, Lee S-Y (2017) Crystal structure of the MOP flippase MurJ in an inward-facing conformation. Nat Struct Mol Biol 24(2):171–176
62. Islam MS, Shahik SM, Sohel M et al (2015) In silico structural and functional annotation of hypothetical proteins of Vibrio cholerae O139, vol 13. Genomics & informatics, p 53. 2
63. Malhotra H, Kaur H (2016) A bioinformatics approach for functional and structural analysis of hypothetical proteins of Clostridium difficile. Imp J Interdiscip Res 2:1601–1609
64. Ikai A (1980) Thermostability and aliphatic index of globular proteins. J Biochem 88(6):1895–1898
65. da Costa WLO, Araújo CLdA, Dias LM et al (2018) Functional annotation of hypothetical proteins from the Exiguobacterium antarcticum strain B7 reveals proteins involved in adaptation to extreme environments, including high arsenic resistance. PLoS ONE 13(6):e0198965
66. Guruprasad K, Reddy BB, Pandit MW (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 Eng Des Selection 4(2):155–161
67. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
68. Xia J, Benner MJ, Hancock RE (2014) NetworkAnalyst-integrative approaches for protein–protein interaction network analysis and visual exploration. Nucleic Acids Res 42(W1):W167–W174
69. Meng L, Liu H, Lan T et al (2020) Antibiotic Resistance Patterns of Pseudomonas spp. Isolated From Raw Milk Revealed by Whole Genome Sequencing. Front Microbiol 11:1005
70. Poole K (2011) Pseudomonas aeruginosa: resistance to the max. Front Microbiol 2:65
71. Rahman A, Susmi TF, Yasmin F et al (2020) Functional annotation of an ecologically important protein from Chloroflexus aurantiacus involved in polyhydroxyalkanoates (PHA) biosynthetic pathway. SN Appl Sci 2(11):1–13
72. Itzhak DN, Tyanova S, Cox J et al (2016) Global, quantitative and dynamic mapping of protein subcellular localization. elife. 5:e16950
73. Samsonov GV (2012) Handbook of the Physicochemical Properties of the Elements. Springer Science & Business Media
74. Lazzaroni J-C, Dubuisson J-F, Vianney A (2002) The Tol proteins of Escherichia coli and their involvement in the translocation of group A colicins, vol 84. Biochimie, pp 391–397. 5–6
75. Crow A, Greene NP, Kaplan E et al (2017) Structure and mechanotransmission mechanism of the MacB ABC transporter superfamily. Proceedings of the National Academy of Sciences, 114(47): p. 12572–12577
76. Doumith M, Mushtaq S, Livermore D et al (2016) New insights into the regulatory pathways associated with the activation of the stringent response in bacterial resistance to the PBP2-targeted antibiotics, mecillinam and OP0595/RG6080. J Antimicrob Chemother 71(10):2810–2814
77. Ropp PA, Hu M, Olesky M et al (2002) Mutations in ponA, the gene encoding penicillin-binding protein 1, and a novel locus, penC, are required for high-level chromosomally mediated penicillin resistance in Neisseria gonorrhoeae. Antimicrob Agents Chemother 46(3):769–777
78. Yu H, Kim PM, Sprecher E et al (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3(4):e59
79. Müller M, Papadopoulou B (2010) Stage-specific expression of the glycine cleavage complex subunits in Leishmania infantum. Molecular and biochemical parasitology. 170:17–271

80. An G, Bendiak DS, Mamelak LA et al (1981) Organization and nucleotide sequence of a new ribosomal operon in Escherichia coli containing the genes for ribosomal protein S2 and elongation factor Ts, vol 9. Nucleic acids research, pp 4163–4172. 16
81. Conrad J, Sun D, Englund N et al (1998) The rluC gene of Escherichia coli codes for a pseudouridine synthase that is solely responsible for synthesis of pseudouridine at positions 955, 2504, and 2580 in 23 S ribosomal RNA. J Biol Chem 273(29):18562–18566
82. Makharashvili N, Koroleva O, Bera S et al (2004) A novel structure of DNA repair protein RecO from Deinococcus radiodurans. Structure 12(10):1881–1889
83. Zheng S, Sham L-T, Rubino FA et al (2018) Structure and mutagenic analysis of the lipid II flippase MurJ from Escherichia coli. Proceedings of the National Academy of Sciences, 115(26): p. 6709–6714
84. Zuberi AR, Ying C, Bischoff DS et al (1991) Gene-protein relationships in the flagellar hook-basal body complex of Bacillus subtilis: sequences of the flgB, flgC, flgG, fliE and fliF genes. Gene 101(1):23–31
85. Matsuzawa H, Asoh S, Kunai K et al (1989) Nucleotide sequence of the rodA gene, responsible for the rod shape of Escherichia coli: rodA and the pbpA gene, encoding penicillin-binding protein 2, constitute the rodA operon. J Bacteriol 171(1):558–560
86. Martin PR, Hobbs M, Free PD et al (1993) Characterization of pilQ, a new gene required for the biogenesis of type 4 fimbriae in Pseudomonas aeruginosa. Molecular microbiology. 9:857–8684
87. El Yacoubi B, Lyons B, Cruz Y et al (2009) The universal YrdC/Sua5 family is required for the formation of threonylcarbamoyladenosine in tRNA. Nucleic Acids Res 37(9):2894–2909
88. Babinski KJ (2004) Genetic and biochemical characterization of the specific UDP-2, 3-diacylglucosamine hydrolase of lipid A biosynthesis.
89. Metzger LE, Lee JK, Stroud RM et al (2010) Discovery, characterization, and structural determination of a novel UDP-2, 3-diacylglucosamine hydrolase. Wiley Online Library
90. Yum D-Y, Lee B-Y, Pan J-G (1999) Identification of the yqhE andyafB Genes Encoding Two 2, 5-Diketo-d-Gluconate Reductases inEscherichia coli. Appl Environ Microbiol 65(8):3341–3346
91. Moynie L, Schnell R, McMahon SA et al (2013) The AEROPATH project targeting Pseudomonas aeruginosa: crystallographic studies for assessment of potential targets in early-stage drug discovery. Acta Crystallographica Section F: Structural Biology and Crystallization Communications, 69(1): p. 25–34
92. Chakravarty S, Melton CN, Bailin A et al (2017) Pseudomonas aeruginosa magnesium transporter MgtE inhibits type III secretion system gene expression by stimulating rsmYZ transcription.Journal of bacteriology,199(23)
93. Chakravarty S, Ramos-Hegazy L, Gasparovic A et al (2020) DNA alternate polymerase PolB mediates inhibition of type III secretion in Pseudomonas aeruginosa.Microbes and Infection, : p.104777
94. Overhage J, Schemionek M, Webb JS et al (2005) Expression of the psl operon in Pseudomonas aeruginosa PAO1 biofilms: PslA performs an essential function in biofilm formation. Appl Environ Microbiol 71(8):4407–4413
95. Campisano A, Schroeder C, Schemionek M et al (2006) PslD is a secreted protein required for biofilm formation by Pseudomonas aeruginosa. Appl Environ Microbiol 72(4):3066–3068
96. Sperandeo P, Cescutti R, Villa R et al (2007) Characterization of lptA and lptB, two essential genes implicated in lipopolysaccharide transport to the outer membrane of Escherichia coli. J Bacteriol 189(1):244–253
97. Konyecsni W, Deretic V (1990) DNA sequence and expression analysis of algP and algQ, components of the multigene system transcriptionally regulating mucoidy in Pseudomonas aeruginosa: algP contains multiple direct repeats. J Bacteriol 172(5):2511–2520
98. Chung BC, Zhao J, Gillespie RA et al (2013) Crystal structure of MraY, an essential membrane enzyme for bacterial cell wall synthesis. Science 341(6149):1012–1016
99. Tashiro Y, Nomura N, Nakao R et al (2008) Opr86 is essential for viability and is a potential candidate for a protective antigen against biofilm formation by Pseudomonas aeruginosa. J Bacteriol 190(11):3969–3978
100. Jenkins DE, Auger EA, Matin A (1991) Role of RpoH, a heat shock regulator protein, in Escherichia coli carbon starvation protein synthesis and survival. J Bacteriol 173(6):1992–1996
101. Samuelson JC, Chen M, Jiang F et al (2000) YidC mediates membrane protein insertion in bacteria. Nature 406(6796):637–641
102. Fussenegger M, Facius D, Meier J et al (1996) A novel peptidoglycan-linked lipoprotein (ComL) that functions in natural transformation competence of Neisseria gonorrhoeae. Molecular microbiology. 19:1095–11055
103. Chiu HC, Lin TL, Wang JT (2007) Identification and characterization of an organic solvent tolerance gene in Helicobacter pylori. Helicobacter 12(1):74–81
104. Rao VS, Srinivas K, Sujini G et al (2014) Protein-protein interaction detection: methods and analysis. International journal of proteomics, 2014
105. Imam N, Alam A, Ali R et al (2019) In silico characterization of hypothetical proteins from Orientia tsutsugamushi str. Karp uncovers virulence genes. Heliyon 5(10):e02734

Figures

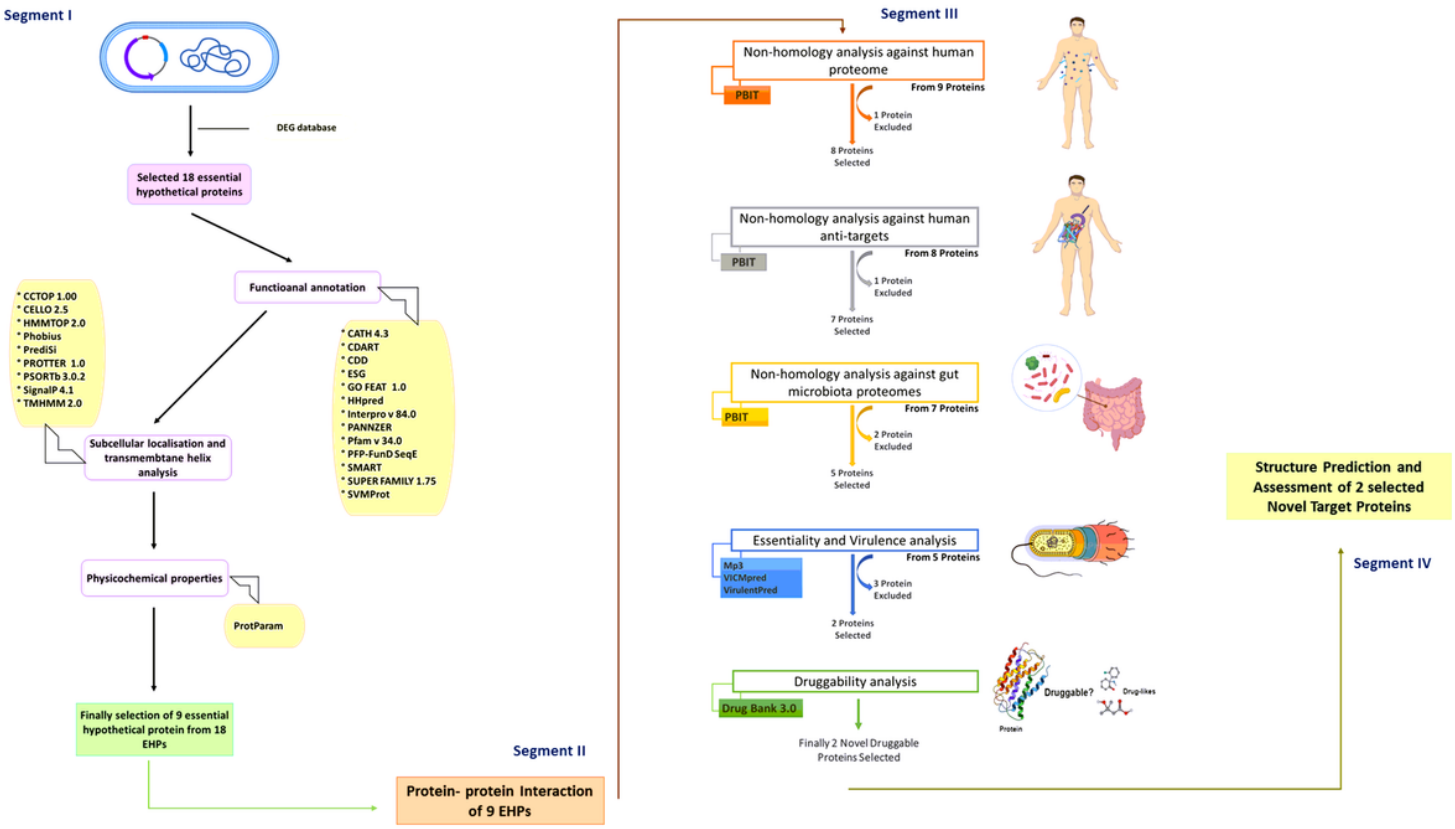

Figure 1

Schematic representation of the whole methodology used in our investigation. There are four segments, **Segment I:** Functional annotation and properties characterization; **Segment II:** Protein-protein interaction network; **Segment III:** Non-homology analysis, virulence factor prediction and druggability identification; **Segment IV:** Structure prediction and structure validation.

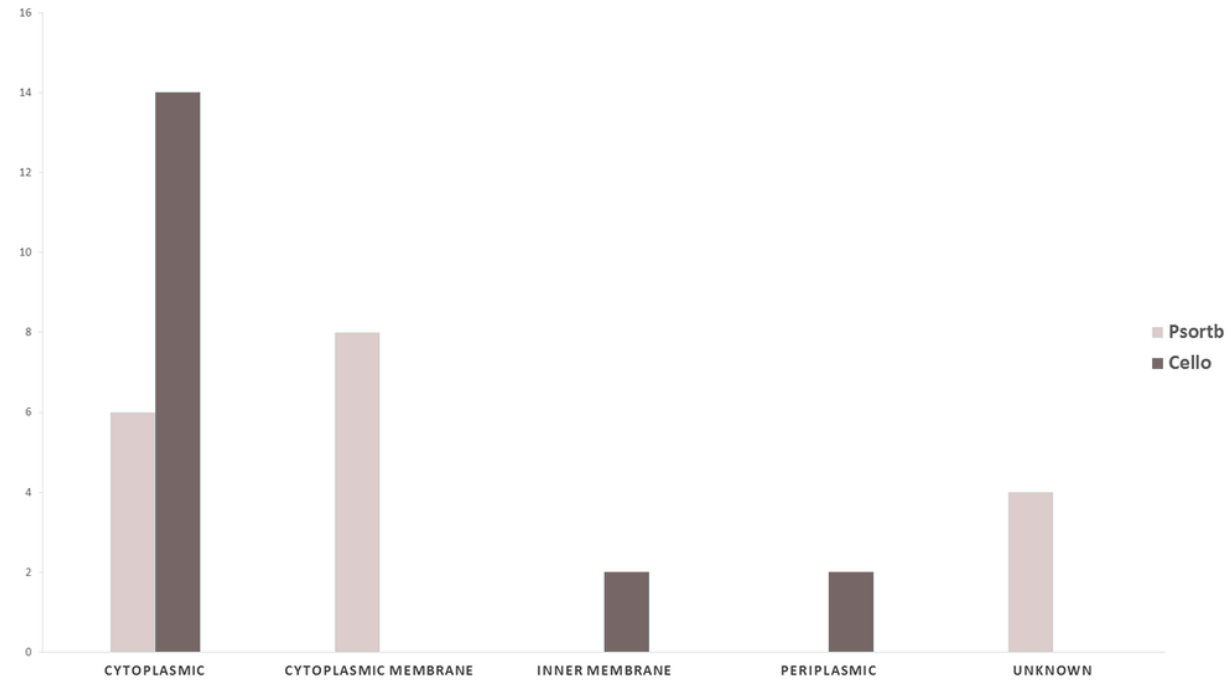

Figure 2

The subcellular localization of 18 EHPs is displayed by the column plot. Five categories of columns are for 5 types of subcellular localization (cytoplasmic, cytoplasmic membrane, inner membrane, periplasmic and unknown). Here lighter black represents the data from CELLO database and the color grey is for the database PSORTb.

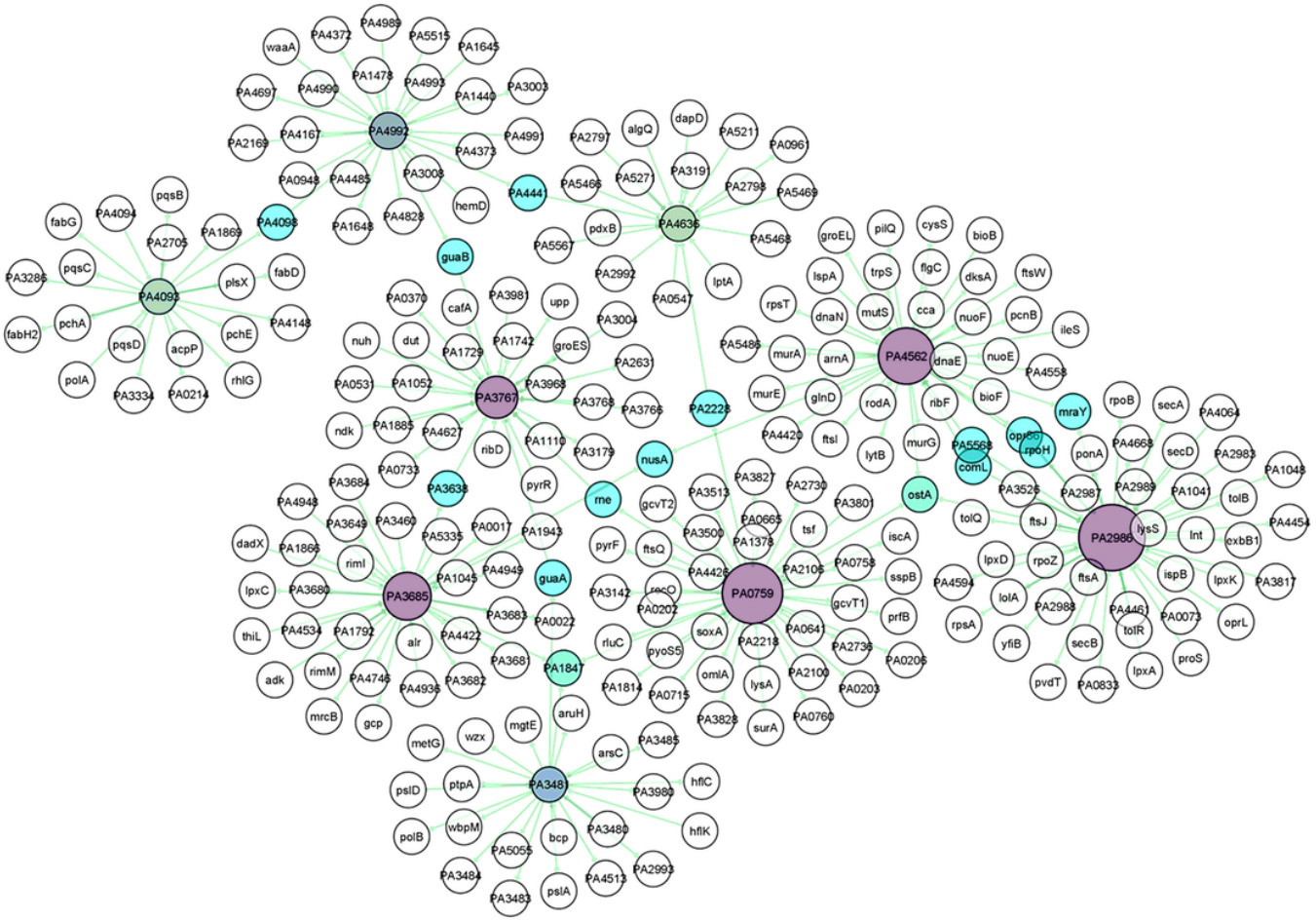

Figure 3

Protein-protein interaction network of 9 stable EHPs collected from P. aeruginosa PAO1 Protein Interactome database. The network has 261 nodes and 269 edges provided with 11 subnetworks (Hubs) with a minimum of 3 nodes each. The nodes with lower degree values are colored green namely PA4992, PA3481, PA4093, PA4636. The color gradually turned into deep purple by the increase of node degree values. The nodes in cyan blue meaning 2 or more interactions with their corresponding subnetworks.

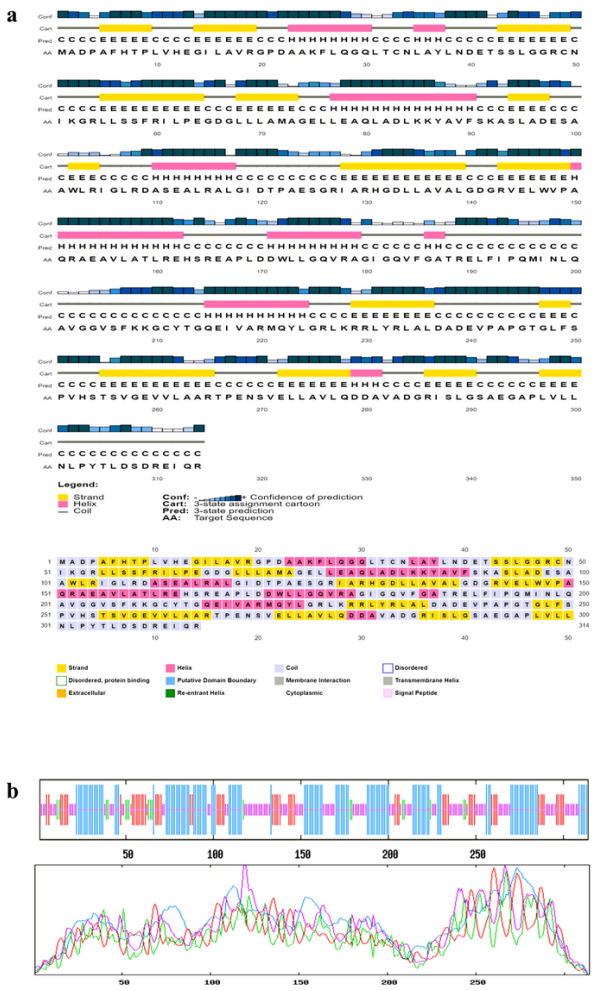

Figure 4

The secondary structure of NP_249450.1 from (a) PSIPRED website, and (b) SOPMA websites.

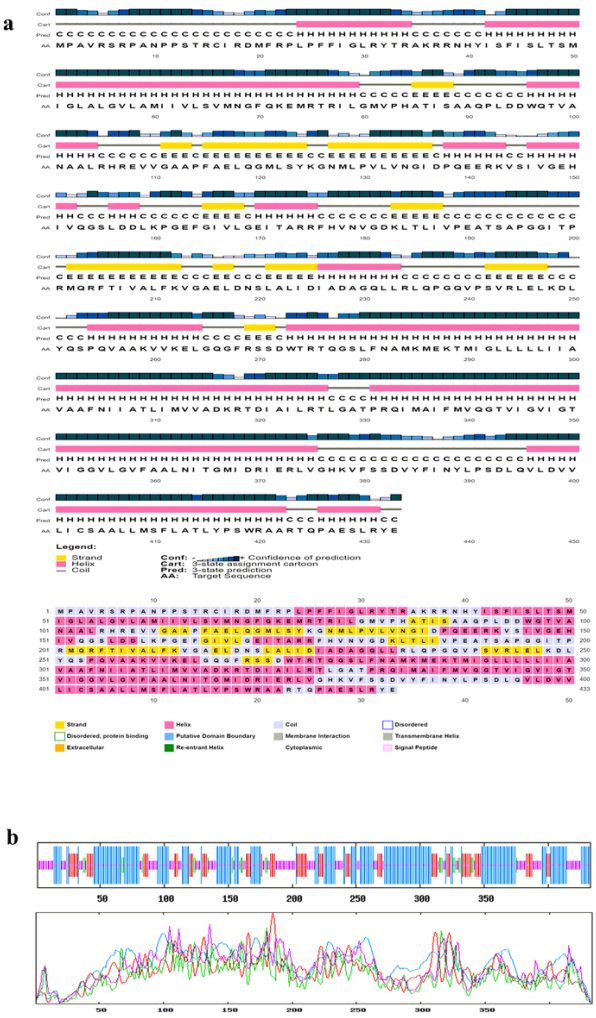

Figure 5

The secondary structure of NP_251676.1 from (a) PSIPRED server, and (b) SOPMA database.

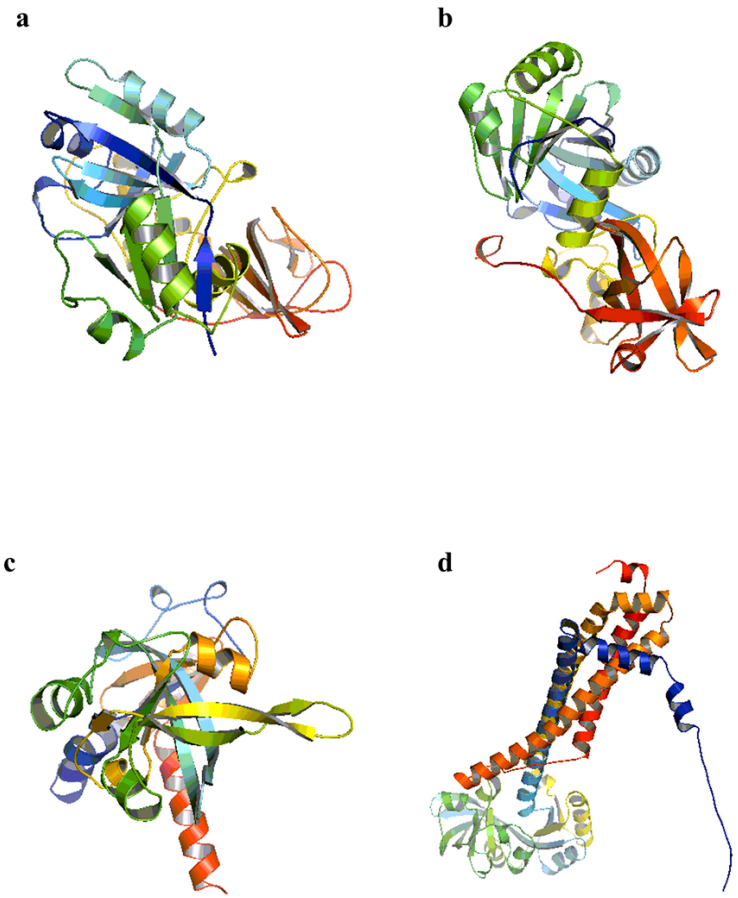

Figure 6

Three-dimen*sional *homolopgy modeling. (a) template-based homology modelling structure of NP_249450.1 from SWISS-MODEL, (b) *ab-initio* modelling structure of NP_249450.1 from the Robetta server, (c) template-based homology modelling structure of NP_251676.1 from SWISS-MODEL, and (d) *ab-initio* modelling structure of NP_251676.1 from the Robetta server.

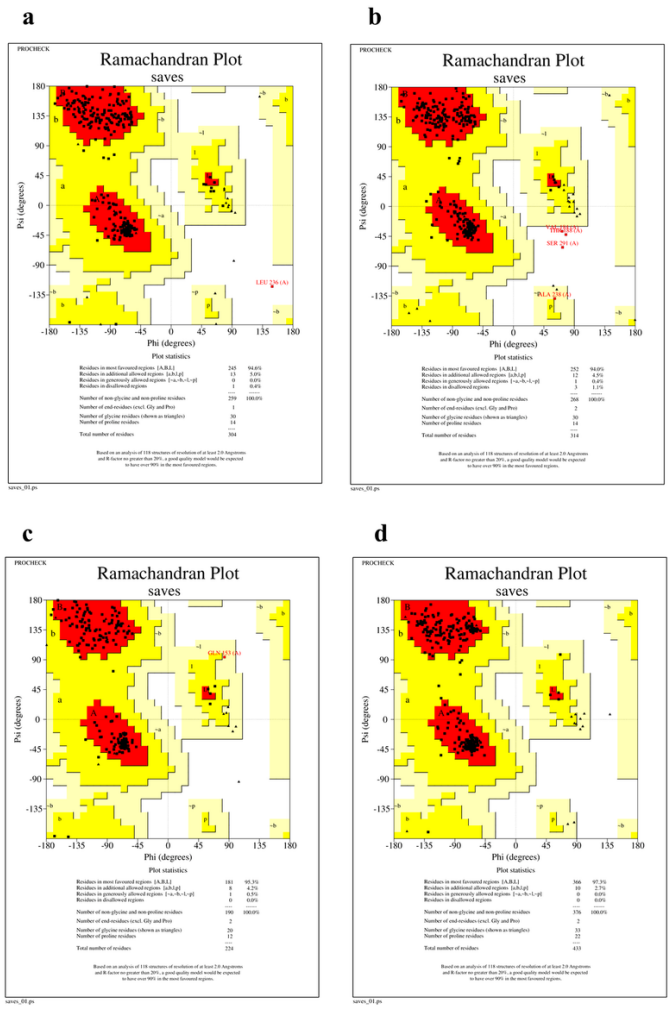

Figure 7

Three-dimensional structure assessment by Ramachandran plot analysis by PROCHECK. (a) Ramachandran plot of NP_249450.1 protein from SWISS-MODEL structure; (b) Ramachandran plot of NP_249450.1 protein from Robetta model; (c) Ramachandran plot of NP_251676.1 protein from SWISS-MODEL structure; (d) Ramachandran plot of NP_251676.1 protein from Robetta model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.xlsx
- SupplementaryFigure1.tif
- SupplementaryTable1.docx
- SupplementaryTable2.docx
- SupplementaryTable3.docx