ISSN: 2577 - 8455

Research Article

Archives of Infectious Diseases & Therapy

Structural and Functional Prediction of the Hypothetical Proteins from Pseudomonas Aeruginosa PA7

Ummi Shahieda Lazaroo*, Zurrein Shah Lazaroo, Ummu Sakinah binti Fais and Suresh kumar

Faculty of Health and Life Sciences, Management & Science University, Shah Alam, Selangor, Malaysia

***Corresponding author**
Ummi Shahieda Lazaroo, Faculty of Health and Life Sciences, Management & Science University, Shah Alam, Selangor, Malaysia; E-mail: mishalazaroo@gmail.com

Submitted: 02 July 2019; **Accepted**: 20 July 2019; **Published**: 06 Aug 2019

Abstract

Background & Aim: *Pseudomonas aeruginosa is the most frequently isolated bacterium among those gram negative rods that are obligated aerobes. It is one of the important opportunistic human pathogens, causing severe chronic respiratory infection in patient with underlying conditions such as cystic fibrosis (CF) or bronchiectasis. The emergence of multi-drug resistance Pseudomonas aeruginosa strain in clinically isolated demands the development of better or new drugs against this pathogen. The study is to assign a precise function to hypothetical protein (HPs), whose functions are unknown.*

Materials and Methods: *With the help of various bioinformatics tools, the extensive functional analysis of these hypothetical proteins was performed. This study combines a number of bioinformatics tools including Blastp, Pfam, InterProScan, SMART, PSLPred, CELLO, Signal Peptide, Expasy's ProtParam tool, VirulenPred, VicmPred to gain information about the conserved regions, families, pathways, interactions, localizations and virulence related to particular protein.*

Result: *The hypothetical proteins present in Pseudomonas Aeruginosa PA7 genome was extensively analyzed and annotated, out of 1350 hypothetical proteins, 25 proteins are catalytic domains, 31 proteins are enzymes, 46 proteins are integral membrane proteins, 72 proteins are transporters, 104 proteins are binding proteins, 404 proteins sequences contain a domain of unknown function (DUF), and 540 proteins cannot be functionally determined by any of the tools.*

Conclusion: *Better understanding of the mechanism of pathogenesis and in finding novel therapeutic targets for Pseudomonas aeruginosa.*

Keywords: Pseudomonas Aeruginosa, Hypothetical Protein, Functional Annotation, Protein Sequences

Introduction

Pseudomonas aeruginosa is one of a gram-negative, human opportunistic pathogen and also a significant cause for acute and chronic infections especially for the patients with compromised host defenses mechanism [1]. P.aeruginosais one of the most common pathogen that is isolated from the patients who have been hospitalized which is not less than one week, and it is also a frequent cause of patient with nosocomial infections. Pseudomonal infections are very complicated and could be life threatening [2]. P.aeruginosa is a strictly aerobic, gram negative bacterium with relatively low virulence. P.aeruginosa is a ubiquitous organism; it can react to moist environments, primarily as waterborne and also soilborne organisms. Pseudomonal species could be found in water, soil, animals and plants. Pseudomonas aeruginosa colonization have been reportedly occurs in not less than 50% of humans, and P.aeruginosa is one of the most common pseudomonal species [3].

One of the top three causes of opportunistic human infections is by the P.aeruginosa which is defined as a ubiquitous environmental bacterium. Its intrinsic resistance to antibiotics and disinfectants is one of the major factors in its prominence as a pathogen [4]. The complete sequence of P.aeruginosa strain [5]. PA7 with 1350 base pairs and the sequence gives important components for the basis of the versatility as well as their intrinsic drug resistance of P.aeruginosa [6]. Consistent with its larger genome size and environmental adaptability, P. aeruginosa considered as the largest proportion of regulatory genes observed because of its consistency with larger genome size and also environmental adaptability [7]. It allows the involvement of the bacterial genome and a large number of genes in the catabolism, transport and also for the efflux of organic compounds. We propose that the size and complexity of the P. aeruginosa genome will reflect an evolutionary adaptation permitting it to thrive in diverse environments according to their size and complexity of it [2].

The respective genome shows that there are about 1350 proteins sequences with unknown function. These proteins are called as

hypothetical proteins (HPs) or also can be defined as putative conserved protein. This is because of the lacking correlation to study about the annotated proteins. The HPs have not yet been described in detail for their biological and physiological level[8]. Most of the proteins in the genomes are belong to HPs, and it is also one of the classification of the protein presumably that having their own importance for completing their proteomic and also for the genomic information.

One of the best and initial steps to explore and identify the homology shared between the proteins that may lead to the strong function prediction with the help of the advance bioinformatics tools for the sequences. This study have successfully characterized 1350 HPs of P.aeruginosa by using various computational bioinformatics tools [9].These analyses of sequences of all the HPS was carried out by using Basic Local Alignment Search Tool Protein (Blastp), Protein Family (pfam), Conserved Domain Database (CDD), The Integrative Protein Signature Database (interproScan) and also Simple Modular Architecture Research Tool and Database (SMART). After that the HPs was precisely defined as the subcellular localization, physiochemical properties as well as which family they belong to by using Expasy Protein Parameters Tools (Expasy's ProtParam), Signal Peptide Sequence Analysis (SignalP), Subcellular Localization Predictive System(CELLO), Prediction of Subcellular Localization of Bacterial Proteins (PSLpred), Predictions of Transmembrane Helices and Topology of Protein (HMMTOP) and last but not least is Prediction of Transmembrane Helices based on the Hidden Markov Model (TMHMM) [10].

Materials and Methods

Functional Annotation of Hypothetical Protein

One of the essential ways to understand the biological process at systems level and also to predict the displayed of the biological system to design a predictive disease model is by doing the functional annotation of the hypothetical proteins. The identification of novel drug target or their vaccine candidates for controlling the infections that caused by P.aeruginosa is by using functional annotation. It is to analyze and annotate the functions for the some of the pathogenic organism that caused by squeal of the diseases by effecting different sites for humans.

Hypothetical Protein Used For Drug Target

The identification of the families of hypothetical proteins is by similarity-based clustering or using more complex approaches. It is important to do the functional annotation of the hypothetical proteins which is involved in infection, drug resistance and essential biosynthetic for the development of the potent antibacterial between the infectious agents. To make them as the potential targets of antimicrobial drugs, the improvement of understanding of these protein may be necessary [10].

It is divided into four steps which are sequence comparison, protein categorization, sub-cellular localization and transmembrane prediction, and physiochemical characterization. The experiment is conducted using computation drug design. The experiments are gathered from multiple software with complete sequence of P.Aeruginosa PA7 [11].

Web-Tools

Blastp: Blastp is the tool that can be used to compare between an amino acid queries sequences with a protein sequence database.

Pfam: The protein families (Pfam) database is one of the large collection of profile hidden Markov models and the protein multiple sequence alignments. Pfam is currently a primarily based on the UniProtKB reference proteomes, with the counts of matched sequences and species reported on the website restricted to this smaller set. Pfam also deducted the number of sequences that showed on the website, whilst still giving the access to many useful model organisms.

CDD: The Conserved Domain Database (CDD) provides similarity searches of the NCBI Entre z Protein Database of domain architecture; it is defined as the sequential order of the conserved domains of a protein. CDD to be fast by relying on domain profiles, this is because it depends on annotated functional domains, informatively. Domain profiles are obtained from some collections of domain definitions which is include of functional annotation.

Interpro Scan: InterProScan is a Java-based architecture for a widely used protein function software package; it is used to characterize many millions of sequences [11]. The developments of InterproScan including improvements and also additions for the outputs of the software, that results in a flexible and stable system that allowed user to use both multiprocessor machines and conventional clusters to get the scalable distributed data analysis.

SMART: SMART is the most efficient and rapid meta-genomics classification algorithm that suitable to match between all the species as well as sequences that is present in the NCBI GenBank [12]. SMART also provides a single step for classification of microorganisms and large plant, invertebrate and also genomes from the meta-genomic sample that can be derived.

PSLPred: One example of hybrid approach-based method is PSLpred which can integrates PSIBLAST as well as three SVM modules that based on compositions of the residues, dipeptides and also physico-chemical properties.

CELLO: CELLO can predict not more than one sub cellular localization on the basis of the protein that found on any species [13]. The number of the terms that found with each of their categories, for example molecular function, biological process, cellular compartment were summed and showed as a pie charts that presenting as possible functional annotation for all the queried protein once the homologous for each query sequences has been identified [14]. Although the experiment of the sub cellular localization of a protein may not be known and yet not annotated, but CELLO still can because for a sub cellular localization [12].

HMMTOP: The prediction of both the localization of the helical trans-membrane segments with the topology of the trans-membrane protein can be done by using the HMMTOP trans membrane topology prediction server [2]. It allowed the user to submit additional data about the segment localization to increase the prediction power.

TMHMM: TMHMM is a tool that used to train using two-pass discriminative training approach that followed by decoding with the one-best algorithm.

SignalP: The tool that used to target and insertion of membrane form secretary and membranes proteins in both eukaryotes and prokaryotes is done by the signal peptide [14]. Signals sequences

will cleaved off from the membrane bound signal peptides once the membrane insertion is done.

ProtParam: The ProtParam tool of ExPASy was used to analyses the physiological as well as chemical properties of the targeted protein sequence. The properties which includes GRAVY (grand average of hydropathy), aliphatic index (AI), isoelectric point, extinction coefficients, and last but not least molecular weight were analyzed using this tool.

Results

It is divided into four steps which are sequence comparison, protein categorization, sub cellular localization and trans-membrane prediction, and physiochemical characterization.

Sequence Comparison: One thousand three hundred and fifty hypothetical human primary protein sequences retrieved from GenBank. GenBank is the National Institutes of Health (NIH) genetic sequence database, an annotated collection of all publicly available DNA sequences. Hypothetical protein is a protein whose existence has been predicted, but for which there is a lack of experimental evidence that it is expressed in vivo. Sequencing of several genomes has resulted in numerous predicted open reading frames to which functions cannot be readily assigned. Blastp was used to find the homologous proteins to the sequence. BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequences databases and calculates the statistical significance. Paste the sequence and click submit to get the result of each sequence. In the result, the first five relevant descriptions are considered [9].

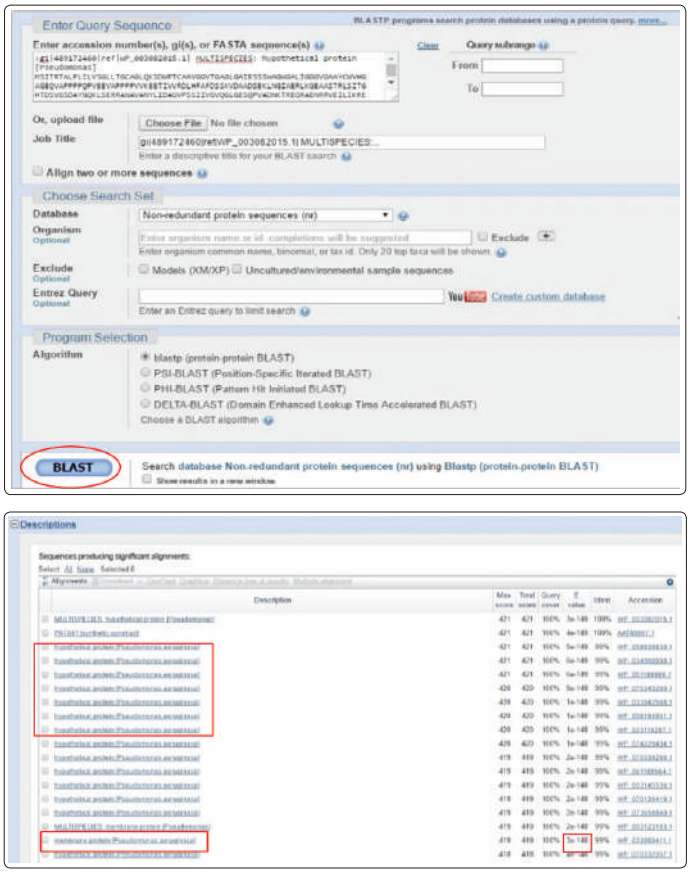


Figure A: Shows the output of BlastP

Blastp was used to explore the protein function as well as to find the homologous protein of the sequences as shown in the (Table 1.0) below and the pie chart was plotted as shown in (Figure 1.1.0). The result showed that Non-hypothetical protein is more than the hypothetical protein which is 864 of it are non-hypothetical while only 486 are hypothetical.

Table 1.0: Example of the listed of sequence similarities by using BlastP

S. No.	Uniport ID	Blastp	E-value
1	489172460	PA1041	4e-149
2	489173103	Pseudomonas Aeruginosa	7e-64
3	489173110	Pseudomonas Aeruginosa	8e-56
4	489173277	Pseudomonas Aeruginosa	4e-39
5	489174814	Pseudomonas Aeruginosa	6e-100
6	489174996	Gammaproteobacteria Bacterium SG8_47	3.3
7	489174997	Pseudomonas Aeruginosa	3e-69
8	489175353	Pseudomonas Aeruginosa	5e-48
9	489175448	Pseudomonas Aeruginosa	3e-79
10	489175454	Kingella Kingae ATCC 23330	2e-43
11	489175516	Pseudomonas Aeruginosa	3e-38
12	489176351	Pseudomonas Aeruginosa	4e-97
13	489176009	Pseudomonas Coronafaciens pv. Garcae	3e-38
14	489176122	Pseudomonas Aeruginosa	8e-50
15	489176351	Pseudomonas Aeruginosa	8e-31
16	489176637	Pseudomonas sp. CF150	9e-30
17	489177009	Pseudomonas Aeruginosa	0.0
18	489177317	Pseudomonas sp. TCU-HL1	4e-56

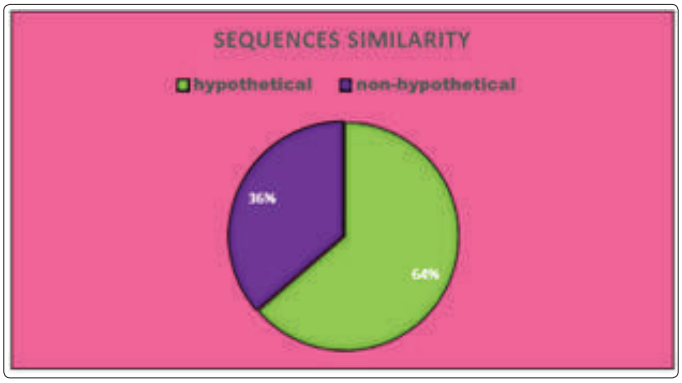
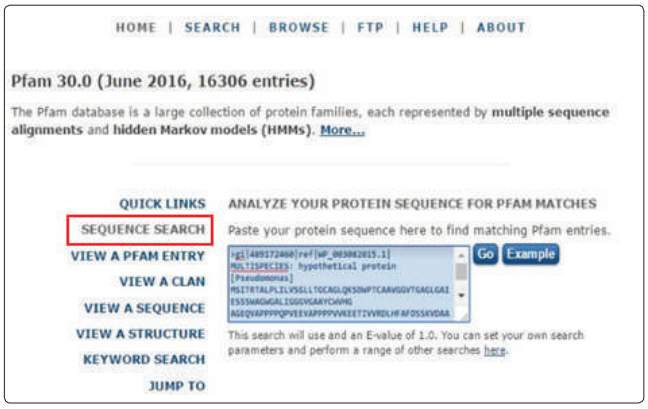


Figure 1.1.0: Statistic result of the sequences similarity by using percentage

Functional Annotation

The moderate domain analysis and function prediction of the hypothetical protein is done by using various tools such as pfam, conserved domain databases or CDART, interproScan and also SMART for each uniprot sequence to do the prediction with these specific tools. And all the description are listed in the table.



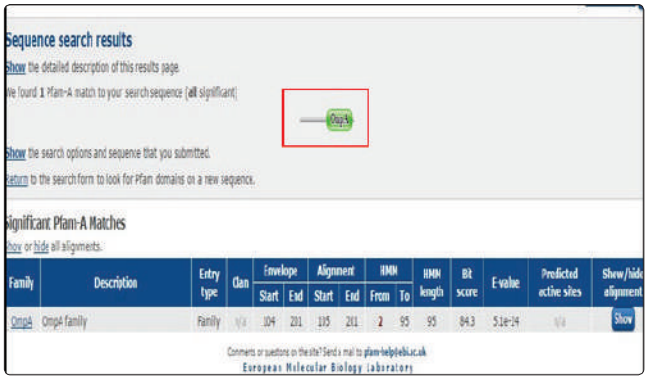


Figure B: Shows the output of pfam

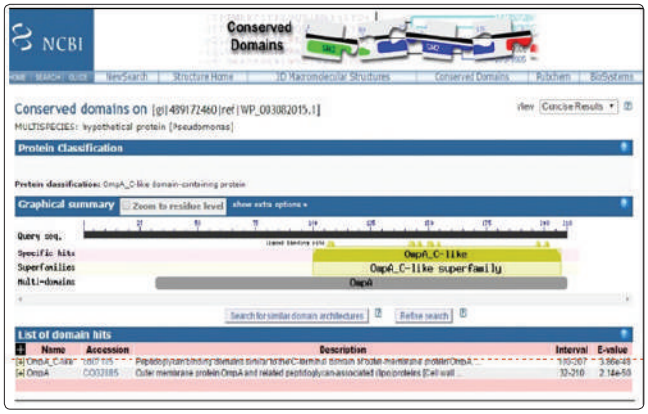


Figure C: Show the output of CDD

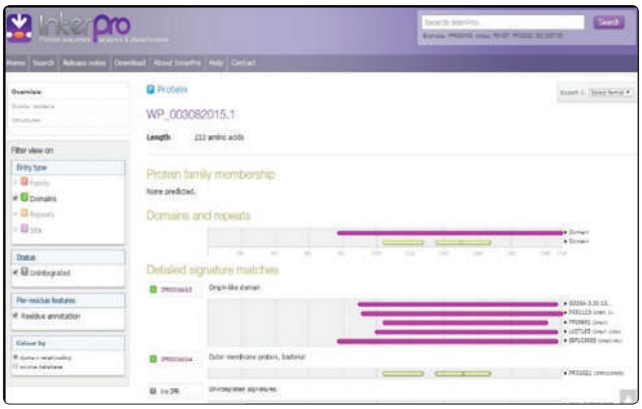


Figure D: Show the output of SMART

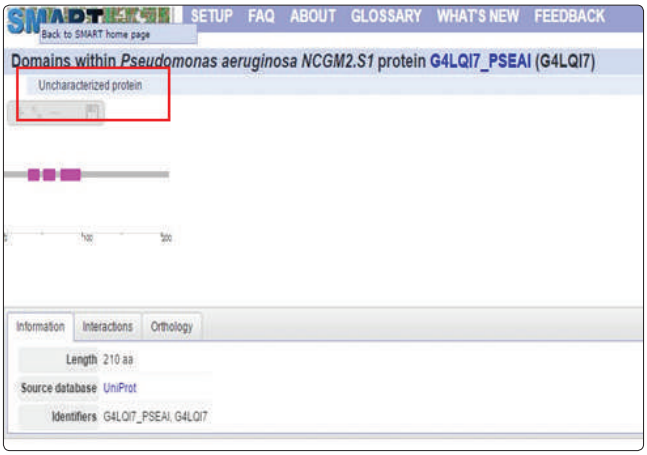


Figure E: Show the output InterProScan

Domain analysis is used to categorize 100% confidence sequences with their function. The Table 1.1 below shows the domain analysis on each web-tools site, while Table 1.2 below shows the sequences with their function, and pie chart was plotted in the Figure 1.1.1.

Based on the Table 1.2 below shows that out of 1350 proteins, there are only 678 protein that have a stable instability index.

Table 1.1 Example of listed domain analysis

S. No	Uniprot ID	Pfam (family/Domain)	CDD	Interpor Scan	Smart
1	489172460	Family: OmpA Family Domain:-	Peptidoglycan Binding Domains Similar to the C-Terminal Domain of Outer-Membrane Protein OmpA	Family: None Predicted. Domain: Outer Membrane Protein, Bacterial	Uncharacterized Protein
2	489173103	Family:- Domian: Ycgl Domain	Uncharacterized Conserved Protien YcgL, UPF0745 Family	Family: None Predicted. Domain: YcgL Domain	Ycgl Domain-Containing Protein PLES 38871
3	489173110	Family Metal-Sensitive Transcriptional Repressor Domain:-	Transcriptional Regulators RcnR and FrmR, and Related Domains; this Domains Family was Previously Known as part of DUF156	Family: Metal-Sensitive Transcriptional Repressor Domain: None Predicated	Uncharacterized Protein
4	489173277	Family: Protein of Unknown Function (DUF1656) Domain:-	Protein of Unknown Function (DUF1656)	Family: Protein of Unknown function DUF1656 Domain: None Predicted	Conserved Domain Protein
5	489174814	Family:- Domain: Domain of Unknown Function (DUF4399)	Domain of Unknown Function (DUF4399)	Family: None Predicted Domain: Domain of unknown function DUF4399	Uncharacterized Protein
6	489174996	Family:- Domain:-	-	Family: None Predicted. Domain None Predicted	Uncharacterized Protein
7	489174997	Family:- Protein of Unknown Function (DUF3392) Domain:-	Protein of Unknown Function (DUF3392)	Family: Protein of Unknown function DUF3392 Domain: None Predicted	Uncharacterized Protein
8	489175353	Family: Prokaryotic Cytochrome C oxidase Subunit IV Domain:-	-	Family: Cytochrome C oxidase Subunit IV, Prokaryotes Domain: None Predicted	Uncharacterized Protein
9	489175448	Family: Protein of Unknown Function (DUF805) Domain:-	Uncharacterized Membrane Protein YhaH, DUF805 Family	Family: Protein of unknown Function DUF805 Domain: None Predicted	Uncharacterized Protein
10	489175454	Family: Carboxymuconolactone Decarboxylase Family Domain:-	Uncharacterized Conserved Protein YurZ, Alkyl hydroperoxide/ Carboxymuconolatone Decarboxylase Family	Family: None Predicted Domain: AhpD-like, Carboxymuconolactone Decarboxylase-Like, Alkhydroperoxidease AhpD core	Uncharacterized Protein
11	489175516	Family:- Domain: Yqey-like Protein	Yqey-like Protein	Family: None Predicted Domain: Aspartyl/ Glutamyl-tRNA Amidotransferase Subunit B-Related, Aspartyl/Glutamyl-tRNA (Asn/Gln) Amidotransferase Subunit B, C-Terminal	Uncharacterized Protein
12	489176351	Family: Protein of Unknown Function (DUF1427) Domain:-	Protein of unknown Function (DUF1427)	Family: Protein of Unknown Function DUF1427 Domain: XapX Domain	Uncharacterized Protein
13	489176009	Family: Domain: Flavinator of Succinate Dehydrogenase	Succinate Dehydrogenase Flavin-Adding Protein, Antitoxin Component of the CptAB Toxin-Antitoxin Module	Family Flavinator of Succinate Dehydrogenase Domain: None Predicted	Uncharacterized Protein PA0706
14	489176122	Family:- Domain:-	-	Family: None Predicted Domain: None Predicted	Uncharacterized Protein
15	489176351	Family: Protein of Unknown Function (DUF1427) Domain:-	Protein of Unknown Function (DUF1427)	Family: Protein of Unknown Function DUF1427 Domain: XpaX Domain	Uncharacterized Protein
16	489176637	Family:- Domain:-	Type 2 Periplasmic Binding Fold Super Family	Family: None Predicted Domain: Ku70/ Ku80 C-Terminal arm	Uncharacterized Protein

17	489177009	Family: Domain: Cupin Domain	Conserved Domain Found in Cupin and Related Proteins	Family: Conserved Hypothetical Protein CHP03214, Putative Allantoin-urate Catabolism Protein Domain: RmIC- Like Jelly Roll Fold, RmIC- Like Cupin Domain Cupin 2, Conserved Barrel	Uncharacterized Protein
18	489177317	Family:- Domain:-	Membrane-Bound PQQ-Dependent Dehydrogenase, Glucose/Quinate/ Shikimate Family	Family: None Predicted Domain: Peptidase M14, Carboxypeptidase A	Uncharacterized Protein
19	489177722	Family: YbaB/EbfC DNA-Binding Family Domain:-	Hypothetical Protein	Family: None Predicted Domain: Non Predicted	Nucleoid-Associated Protein PA1533
20	489177881	Family: Putative Member of DMT Super Family (DUF486) Domain:	Putative Member of DMT Super Family	Family: None Predicted Domain: Cupredoxin Blue (type 1) copper domain	Uncharacterized Protein

After the domain analysis, the functions of 678 proteins with high confidence were assigned. The result in Figure 1.1.1 below shows that there are 25 proteins that are catalytic domains, 31 proteins are enzymes, 46 proteins are integral membrane proteins, 72 proteins are transporters, 104 are binding proteins, 404 of protein sequence contains of a domain of unknown functions (DUF), and last but not least, there are total 540 proteins that cannot be functionally determined by any other tools.

Table 1.2 Example listed of the sequences with their function

No	Uniport ID	Description	Function
1	489172460	OmpA Family	Integral Membrane Protein
2	489173103	YcgL Domain	DUF
3	489173110	Metal-Sensitive Transcriptional Repressor	Binding Protein
4	489173277	Protein of Unknown Function	DUF
5	489174814	Domain of Unknown Dunction	DUF
6	489175353	Prokaryotic Cytochrome C oxidase Subunit IV	Transporter
7	489175448	Protein unknown Function	Catalytic Domain
8	489175454	Carboxymuconolactone Decarboxylase Family	Catalytic Domain
9	489175516	Yqey-like Protein	Integral Membrane Protein
10	489176351	Protein of unknown Function	DUF
11	489176009	Flavinator of Succinate Dehydrogenase	Transporter
12	489176351	Protein unknown Function	DUF
13	489176637	Ku70/Ku80 C- Terminal arm	Binding Protein
14	489177009	Cupin Domain	Integral Membrane Protein
15	489177317	Peptidase M14, Carboxypeptidase A	Catalytic Domain
16	489177722	YbaB/EbfC DNA-Binding Family	Binding Protein
17	489177881	DNA-Binding Proteins	Binding Protein
18	489177979	Thioesterase Superfamily	Catalytic Domain
19	489178112	Type VI Secretion Syatem Protein DotU	DUF
20	489178485	SCP-2 Sterol Transfer Family	Binding Protein
21	489178499	P-loop Containing Nucleoside Triphoaphate Hydrolase	Integral Membrane Protein
22	489179616	Ribbon-Helix-Helix Domain	Binding Protein
23	489180008	Type VI Protein Secretion System Component Hcp	Integral Membrane Protein
24	489180152	MbtH-Like Protein	DUF
25	489180332	Putative Bacterial Toxin ydaT	Enzyme
26	489181024	Peptidase Propeptide and YPEB Domain	Integral Membrane Protein
27	489181184	MerR HTH Family Regulatory Protein	Binding Protein

28	489181415	Domain of unknown Function	DUF
29	489181482	Alanine-Zipper, major Outer Membrane Lipoprotein	Binding Protein
30	489181790	Bacterial Protein of unknown Function	DUF
31	489181806	Bacterial Protein of unknown Function	DUF
32	489182036	YCII-Related Domain	Enzyme
33	489182039	LTXXQ motif Family Protein	Integral Membrane Protein

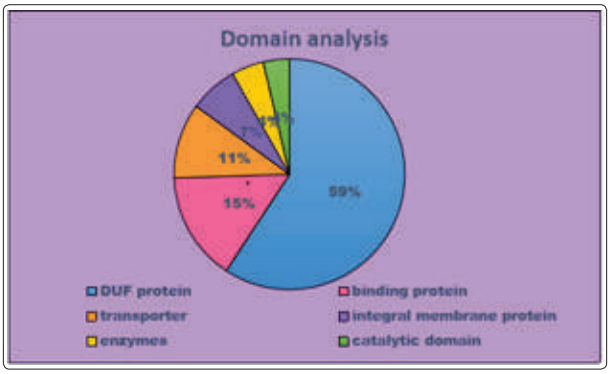


Figure 1.1.1: Statistic result of domain analysis and functional annotation by using percentage

Sub-Cellular Localization

Sub-cellular localization is used to find the organism whether it is Gram-positive or Gram negative. The cytoplasms are sub-cellular localization. The periplasmic and outer membrane are Gram-positive while the extracellular and inner-membrane are gram negative. To get these results some various tools had been used such as PSLpred, CELLO, signal peptide, HMMTOP and also TMHMM.

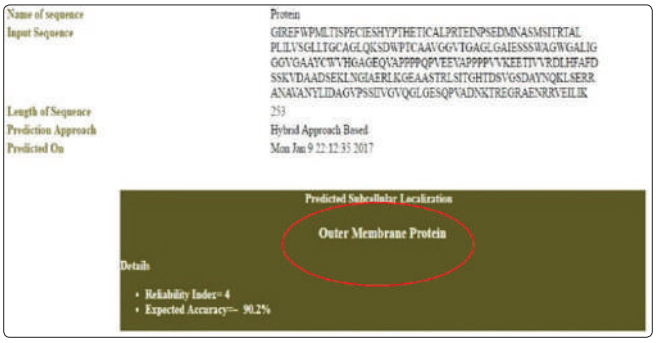
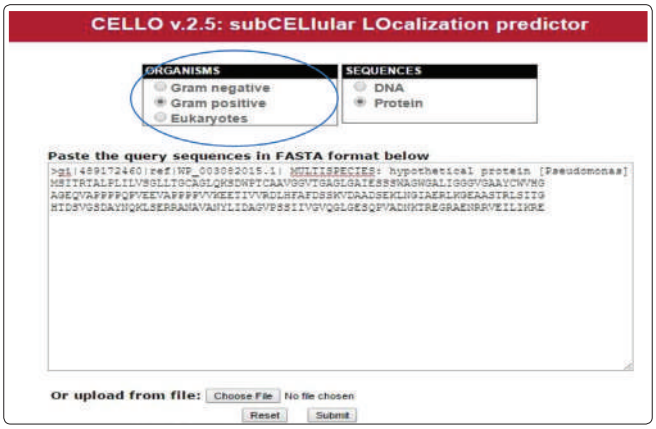


Figure F: Shows the output of PSLpred



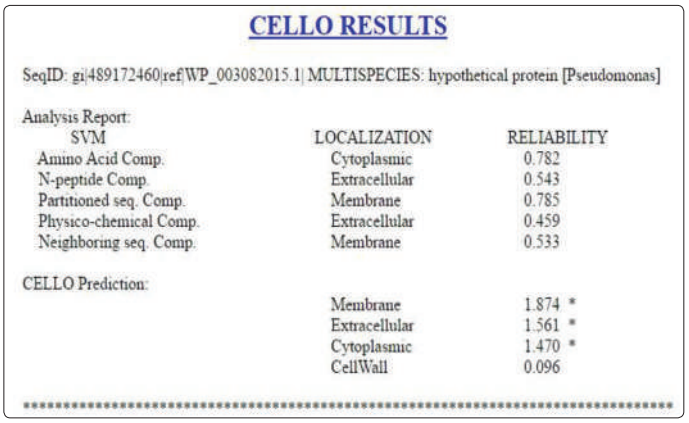


Figure G: Shows the output of CELLO

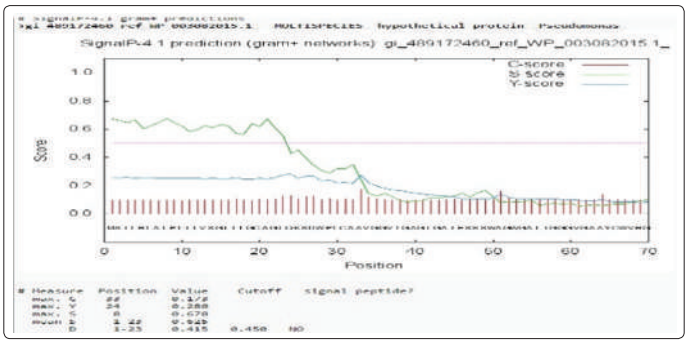


Figure H: Shows the output of signal peptide

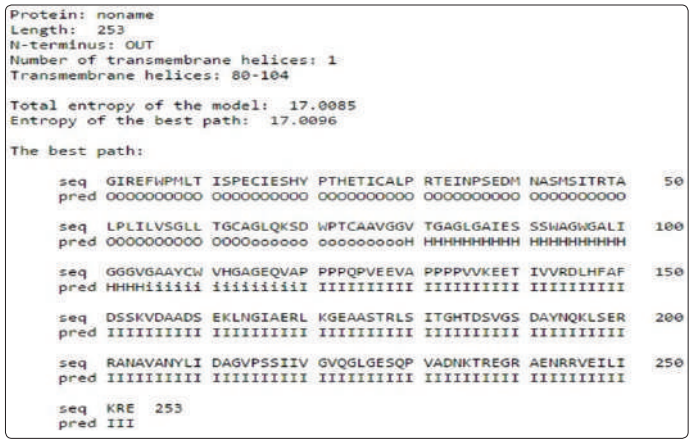


Figure I: Shows the output of HMMTOP

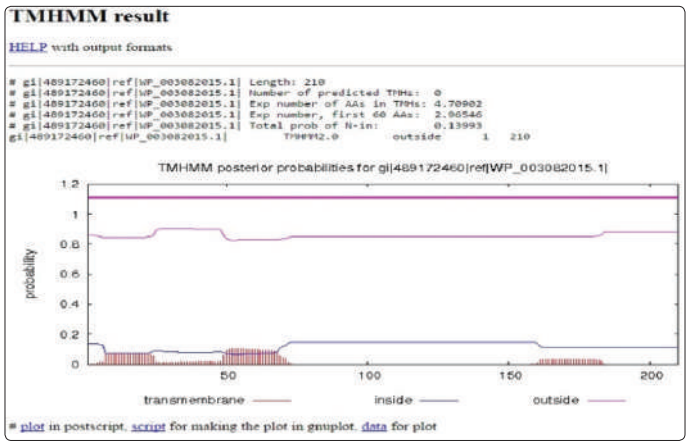


Figure J: Shows the output of TMHMM

Table 1.3: Below show the result of the sub-cellular location, signal peptide and transmembrane helicase of the hypothetical proteins selected

Table 1.3: Example of listed sub-cellular localization

S. No	UNIPORT ID	PSLpred	CELLO	SIGNAL PEPTIDE	HMMTOP		TMHMM	TMHMM
1	489172460	Outer Membrane Protein	Membrane Protein	NO	YES-1	NO	85.91	-0.127
2	489173103	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	82.81	-0.459
3	489173110	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	90.71	-0.292
4	489173277	Inner-Membrane Protein	Membrane Protein	NO	YES-2	YES-2	94.43	0.361
5	489174814	Cytoplasmic Protein	Cytoplasmic Protein	NO	YES-1	NO	83.20	-0.118
6	489174966	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	89.55	-0.387
7	489174997	Inner-Membrane Protein	Membrane Protein	NO	YES-3	YES-3	111.92	0.610
8	489175353	Inner-Membrane Protein	Membrane Protein	NO	YES-3	YES-3	113.81	0.581
9	489175448	Inner-Membrane Protein	Membrane Protein	NO	YES-3	YES-3	108.13	0.190
10	489175454	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	83.91	-0.011
11	489175516	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	92.68	-0.398
12	489176351	Inner-Membrane Protein	Extracellular Protein	NO	YES-2	YES-1	108.58	0.294
13	489176009	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	73.31	-0.514
14	489176122	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	63.86	-0.532

15	489176351	Inner-Membrane Protein	Extracellular Protein	NO	YES-2	YES-1	108.58	0.294
16	489176637	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	100.35	-0.405
17	489177009	Cytoplasmic Protein	Cytoplasmic Protein	NO	NO	NO	69.76	-0.351
18	489177317	Inner-Membrane Protein	Membrane Protein	NO	YES-3	YES-2	95.24	-0.063
19	489177722	Periplasmic Protein	Cytoplasmic Protein	NO	NO	NO	66.50	-0.419
20	489177881	Inner-Membrane Protein	Membrane Protein	NO	YES-4	YES-4		

Based on the (Table 1.3) above shows that out of 1350 hypothetical protein, there are only 945 cytoplasm are found, while 90 are inner-membrane proteins, 50 are outer-membrane proteins, 122 are periplasmic proteins, and 143 are extracellular proteins. The pie chart result was plotted in (Figure 1.1.2) below.

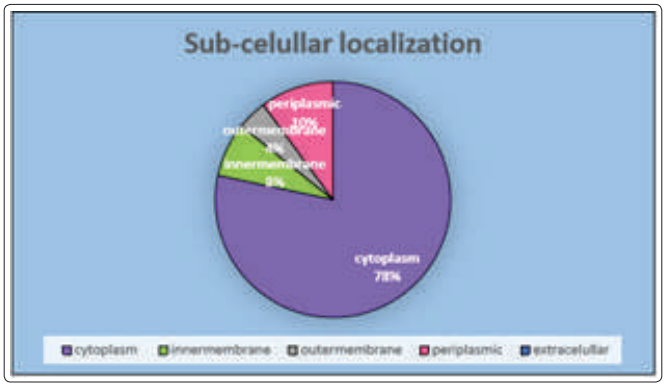


Figure 1.1.2: Statistic result of sub-cellular localization by using percentage

Physiochemical Characterization

Physiochemical characterization is used to understand the physiochemical properties of the compound such as stability, solubility and also the solid state properties. Expasy's ProtParam was used to get the result of the computed parameters such as the molecular weight, theoretical PI, extinction coefficient, instability index computed and classification, Aliphatic index, Grand average of hydopathicity (GRAVY).

ProtParam

User-provided sequence:

```
        10         20         30         40         50         60
GIREFWPMUL TISPECIESH YPOTHETICA LPROTEINPS EUDOMONASM SITRTALPLI

        70         80         90        100        110        120
LVSGLLTGCA GLQKSDWPTC AAVGGVTGAG LGAIESSSWA GWGALIGGGV GAAYCWVHGA

       130        140        150        160        170        180
GEQVAPPPPQ PVEEVAPPPP VVKEETIVVR DLHFAFDSSK VDAADSEKLN GIAERLKGEA

       190        200        210        220        230        240
ASTRLSITGH TDSVGSDAYN QKLSERRANA VANYLIDAGV PSSIIVGVQG LGESQPVADN

       250
KTREGRAENR RVEILIKRE
```

Number of amino acids: 259

Molecular weight: 27896.33

Theoretical pI: 5.02

Amino acid composition: CSV format

Ala (A)	28	10.8%
Arg (R)	13	5.0%
Asn (N)	8	3.1%
Asp (D)	10	3.9%
Cys (C)	5	1.9%
Gln (Q)	6	2.3%
Glu (E)	21	8.1%
Gly (G)	26	10.0%
His (H)	5	1.9%
Ile (I)	17	6.6%
Leu (L)	18	6.9%
Lys (K)	8	3.1%
Met (M)	3	1.2%
Phe (F)	3	1.2%
Pro (P)	18	6.9%
Ser (S)	21	8.1%
Thr (T)	14	5.4%
Trp (W)	5	1.9%
Tyr (Y)	4	1.5%
Val (V)	20	7.7%
Pyl (O)	4	1.5%
Sec (U)	2	0.8%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Atomic composition:

Carbon	C	1220
Hydrogen	H	1945
Nitrogen	N	343
Oxygen	O	379
Sulfur	S	8
Selenium	Se	2

Formula: $C_{1220}H_{1945}N_{343}O_{379}S_8Se_2$

Total number of atoms: 3897

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 33710
Abs 0.1% (=1 g/l) 1.208, assuming all pairs of Cys residues form cystines

Ext. coefficient 33460
Abs 0.1% (=1 g/l) 1.199, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is G (Gly).

The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:

The instability index (II) is computed to be 50.73
This classifies the protein as unstable.

Aliphatic index: 85.91

Grand average of hydropathicity (GRAVY): -0.127

Figure K: Show the output of Expasy's Prot Param

Table 1.4 below shows the result of the molecular weight, Theoretical PI, extinction coefficient, instability index, aliphatic index and Grand average of hydropath city (GRAVY) of each protein.

Table 1.4 Example of listed Expasy'sProtParam

S. No	UNIPORT ID	Molecular Weight, M (Da)	Theoretical PI	Extinction Coefficient (M-1 cm-1)	Instability Index		Aliphatic Index	Grand Average of Hydropathicity (GRAVY)
1	489172460	27896.33	5.02	33710	50.73	Unstable	85.91	-0.127
2	489173103	17537.01	5.03	18012	77.28	Unstable	82.81	-0.459
3	489173110	16172.34	5.87	8605	58.33	Unstable	90.71	-0.292
4	489173277	14469.58	5.17	22585	47.44	Unstable	94.43	0.361
5	489174814	21973.02	6.06	8750	42.22	Unstable	83.20	-0.118
6	489174966	15485.90	4.95	11585	52.38	Unstable	89.55	-0.387
7	489174997	18037.05	5.47	29575	43.40	Unstable	111.92	0.610
8	489175353	15215.89	5.87	36230	53.12	Unstable	113.81	0.581
9	489175448	19442.61	6.47	38055	43.35	Unstable	108.13	0.190
10	489175454	17908.29	5.20	18700	37.57	Stable	83.91	-0.011
11	489175516	22991.24	5.21	7115	54.48	Unstable	92.68	-0.398
12	489176351	12351.26	4.70	15595	62.56	Unstable	108.58	0.294
13	489176009	16295.42	4.57	27095	65.22	Unstable	73.31	-0.514
14	489176122	14951.16	5.53	8750	66.12	Unstable	63.86	-0.532
15	489176351	12351.26	4.70	15585	62.56	Unstable	108.58	0.294
16	489176637	17047.26	4.61	12615	78.39	Unstable	100.35	-0.405
17	489177009	37762.72	5.56	58120	43.19	Unstable	69.76	-0.351
18	489177317	24379.12	4.99	57075	56.40	Unstable	95.24	-0.063
19	489177722	18025.53	4.85	7115	66.95	Unstable	66.50	-0.419

Based on the result from (Table 1.4) above shows that out of 1350 hypothetical proteins, there are 729 of the protein are considered as unstable while 621 proteins are considered as stable based on their instability index. The result was also plotted in the (Figure 1.1.3) below.

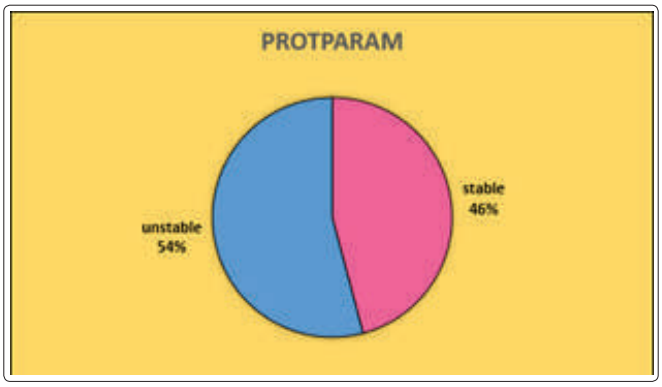


Figure 1.1.3: statistic of instability index by using percentage

Discussion

It is divided into four steps which are sequence comparison, protein categorization, sub-cellular localization and trans-membrane prediction, and physiochemical characterization. For the sequence comparison, it is about to find the homologous protein of the protein sequences. After that is the protein domain by functional annotation [10]. Sub-cellular localization is to find the organism whether it is Gram positive or Gram negative, and it also where the protein sub cellular localization and the number of the trans-membrane helicase. While for the physiochemical characterization, it is to understand the physiochemical properties of each of the compound such as their stability, solubility as well as their state properties.

Based on the model organism of Blastp, it shows the results of NCBI Blastp analyses and their default parameters, from each of the P.Aeruginosa protein sequences [15]. Each of the organisms shows the best Blastp hit with their E-value of <0.01 or E-values identification for the organisms with more than one hit. Out of 1350 protein sequences, the results displays that non-hypothetical is more than hypothetical protein which is 864 are non-hypothetical and 486 are hypothetical.

Conserved domain analysis of the hypothetical protein is done by using the sequence similarity to search for the orthologous close family member, some of the specific tools have been used such as CDD, Pfam, InterproScan and SMART. The functions of 1350 protein sequences with high confidence were assigned. The result in (Figure 1.1.3) shows that only 678 proteins have a stable instability index. It displays that 25 proteins are catalytic domains, 31 proteins are enzymes, 46 proteins are integral membrane proteins, 72 proteins are transporters, 104 proteins are binding proteins, 404 proteins sequences contain a domain of unknown function (DUF), and 540 proteins cannot be functionally determined by any of the tools. Many of the HPs contain domains with enzymatic activities and predicted as transferases, phosphatases, isomerases, hydrolases,

oxidoreductases, lyases, kinase and also ligase [9].

One of the biological process that occurred in all of the living is the selective transport of the molecules across the membrane pores. It includes the protein chaperone and mRNA in and out of the cell nucleus. Mediated with the help of complex of nuclear pore and also transportation of the specific molecules that across the biological membrane which is mediated from the passive and active transport.

The DNA binding proteins control the chromosomal maintenance and also global gene expression which related to the chromosomal locations that the proteins function in vivo. Monitoring binding of the gene specific transcription activators by using microarray method which could reveal the genome-wide location of the DNA bound proteins. Facilitating investigation of gene regulatory network, gene function and genome maintenance by using genome-wide location analysis Catalytic domains also known as kinase domain are related by virtue the homologous protein that make up a large superfamily by the eukaryotic protein kinases. It consists of around250-300 amino acid residues. A classification of the scheme can be seen on the kinase domain phylogeny which have related substrate specificities and modes of regulation that reveals from the families of enzymes [1].

The combination of the integral membrane protein are important because they are both structurally and functionally an integral component of a membrane. As for the functionally, the membrane specific functions are imparts with the presence of these proteins. While for the structurally, the hydrophobic regions of the phospholipid bilayer is penetrate by the regions of the integral membrane proteins [17]. The use of the detergents that disrupt the hydrophobic interaction of the bilayer removing the integral membrane proteins from the membrane during the interaction. When a membrane contain this protein it will automated can be function for the photosynthesis process.

A substantial part of the protein domain database like Pfam is formed by DUF. Most of the DUFs are shared between the bacteria, archea and eukaryote kingdom but they have not been characterized yet due to the difficulty of figuring out their function. The possibility of the space could be just too big to explore if the knockout of the corresponding gene does not result in an easily detectable phenotype. Knowing the DUF somehow can involve in the utilization of a carbon-containing compound and there are simply too many to test them all [15].

Sub-cellular localization is derived from the database and is automated text mining of biomedical literature and sequence-based prediction, as the drug target for P.aeruginosa, sub-cellular localization was found in the cytoplasm, especially in the endosome membrane.

Out of these 1350 hypothetical protein sequences, there are 945 HPs are predicted to be in cytoplasm, 90 inner-membrane proteins, 50 are outer-membrane proteins, 122 are periplasmic proteins and 143 are extracellular proteins. The prediction of membrane proteins is important as component in cell-cell signaling, ion and solute transportation and self-recognition. In pharmaceutical industry, the membrane bound receptors are very important.

The physiochemical properties acquired through Expasy's ProtParam displays that out of 1350 hypothetical protein sequences, 619 HPs are considered as stable while 731 HPs are considered as unstable. Molecular weight is an important parameter which is used to predict the localization and also the protein interaction. While theoretical point or also known as isoelectric point which the positive charges are coordinated to the negative charges nulling presence of development in an electrical field speaks to the pH where the theoretical proteins could not demonstrate at least dissolvability in trial examines encouraging its segregation in an electric field amid the same. Extinction coefficient displays the total amount of light a protein can retain a certain wavelengths.

The hypothetical protein is a protein that has been predicted but there is still lack of experimental evidence in. the function in unknown but it still can be modified and could be responsible for nuclear transportation, apoptosis stability protein and also degeneration for it to have the important role of recognition [16-19].

Based on the results that obtained from all the methods, from 1350 hypothetical proteins that have been analyzed in this study, there are only 278 proteins are considered stable In stability index, less molecular weight, and also high alpha-helix, sub-cellular localization in cytoplasm which can be used as possible drug target for Pseudomonas aeruginosa.

Acknowledgments

We thank Dr.Suresh Kumar and UmmuSakinah for sharing the uses of bioinformatics tools and software.

References

1. Varga JJ, Barbier M, Mulet X, Bielecki P, Bartell J A, et al. (2015) Genotypic and phenotypic analyses of a Pseudomonas aeruginosa chronic bronchiectasis isolate reveal differences from cystic fibrosis and laboratory strains. BMC Genomics 16: 01-27.
2. Menichelli C, Gascuel O, Bréhélin L (2018) Improving pairwise comparison of protein sequences with domain co-occurrence. PLOS Computational Biology 14: e1005889.
3. Pseudomonas aeruginosa (2019) ScienceDirect https://www.sciencedirect.com/topics/immunology-and-microbiology/pseudomonas-aeruginosa
4. Pseudomonas aeruginosa (2019) ScienceDirect https://www.sciencedirect.com/topics/immunology-and-microbiology/pseudomonas-aeruginosa
5. Balasubramanian D, Mathee K (2009) Comparative transcriptome analyses of Pseudomonas aeruginosa. Human Genomics 3: 349-361.
6. Genome Annotation (2019) ScienceDirect. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/genome-annotation
7. Understanding an Unknown Protein Sequence (2019) https://www.ukessays.com/essays/biology/understanding-unknown-protein-sequence-5465.php
8. Pseudomonas Infection: Background, Alhazmi A (2015) Pseudomonas aeruginosa – Pathogenesis and Pathogenic Mechanisms. International Journal of Biology 7: 44-67.
9. Offin M, Chan JM, Tenet M, Rizvi HA, Shen R, et al. (2019) Concurrent RB1 and TP53 alterations define a subset of EGFR-mutant lung cancers at risk for histologic transformation and inferior clinical outcomes. Journal of Thoracic Oncology doi.org/10.1016/j.jtho.2019.06.002.
10. Holman AG, Davis PJ, Foster JM, Carlow CK, Kumar S (2009)

Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of Brugia malayi. BMC Microbiology 9: 243.
11. Paridah M, Moradbak A, Mohamed A, Owolabi F, Abdulwahab taiwo, et al. (2016) We are Intech Open, the world ' s leading publisher of Open Access books Built by scientists, for scientists TOP 1 %. Intech, I (tourism), 13 doi.org/http://dx.doi.org/10.5772/57353.
12. Rosa RG, Goldani LZ (2014) Cohort Study of the Impact of Time to Antibiotic Administration on Mortality in Patients with Febrile Neutropenia. Antimicrobial Agents and Chemotherapy 58: 3799-3803.
13. Tipton KD, Hamilton DL, Gallagher IJ (2018) Assessing the Role of Muscle Protein Breakdown in Response to Nutrition and Exercise in Humans. Sports Medicine 48: 53-64.
14. Everett J, Turner K, Cai Q, Gordon V, Whiteley M, et al. (2017) Arginine Is a Critical Substrate for the Pathogenesis of Pseudomonas aeruginosa in Burn Wound Infections. MBio 8: 01-10.
15. Pseudomonas Infection: Background, Pathophysiology, Epidemiology (n.d.). (2019) https://emedicine.medscape.com/article/970904-overview
16. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLoS Computational Biology 8: e1002514.
17. Lhazmi A (2015) Pseudomonas aeruginosa – Pathogenesis and Pathogenic Mechanisms. International Journal of Biology 7: 44-67.
18. Shigeki Fujitani, Kathryn S Moffett, Victor L Yu (2019) Pseudomonas aeruginosa. http://www.antimicrobe.org/b112.asp
19. Pseudomonas aeruginosa (2019) http://www.antimicrobe.org/b112.asp