MedCrave
Step into the World of Research

Journal of Microbiology & Experimentation

Research Article

Open Access

CrossMark

Characterization and functional analysis of the hypothetical protein in *Yersinia pestis*

Abstract

A hypothetical protein is a protein which exists but the function is not known. These are the proteins for which there is no experimental evidence that it is expressed *in vivo*. The usual scenario involving a hypothetical protein is gene identification during genome analysis. The function of a hypothetical protein can also be predicted by domain homology searches with various confidence levels.

The Human Genome project and the genome projects of many other organisms has been a source for the information on the sequences and their functional aspects. However, this resulted in identification of many proteins which are hypothetical. Our project is mainly focused on these proteins. One of the hypothetical proteins with accession number AAM86904 is analyzed. This protein is from the organism *Yersinia pestis* KIM. This is a bacteria and kim is the strain to which the bacteria belongs. This study revealed a good amount of information about this protein. The protein characteristics, properties and similarities with other proteins are identified. The main principle for this study is that similar sequences having structural similarity are functionally related.

Volume 10 Issue 5 - 2022

Vasudevan Ranganathan, Kainaat Munir Khoja
Aurora's Degree & PG College, Department of Microbiology, India

Correspondence: Vasudevan Ranganathan, Department of Microbiology, Aurora's Degree & PG College, Accredited by NAAC with B++ Grade, Chikadpally, Hyderabad, India, Tel 9100001639/8121119692, Email vasudeva123@gmail.com

Received: October 07, 2022 | **Published:** October 27, 2022

Introduction

The given hypothetical protein with the accession number AAM86904 is a bacteria by name *Yersinia pestis* kim. Kim is the strain to which the bacterium belongs. This is our sequence under study and hereby referred to as query. The length of the query sequence is 211aa (amino acids). This is a gram negative facultative anaerobic bacteria belonging to the family enterobacteriaceae. This is a coccobacilli exhibiting bipolar staining character. This organism was discovered in 1894 by a Swiss/French physician and bacteriologist from the Pasteur institute, Alexandre yersin during an epidemic of plaque in Hong Kong.[1]

Taxonomical classification of *Yersinia pestis* is as follows:

Kingdom: Eubacteria

Phylum: Proteobacteria

Class: Gamma Proteobacteria

Order: Enterobacteriales

Family: *Enterobacteriaceae*

Genus: Yersinia

Species: Y.pestis

Characters of Yersinia pestis

1. It is a gram negative bacterium.
2. Facultative anaerobe.
3. Coccobacilli exhibiting bipolar staining characters.
4. Belongs to family enterobacteriaceae.

Genome structure: The complete genome structure is available for two sub species. One belongs to the strain kim and other belongs to the strain CO92 (kim and CO92 are strains of *Yersinia pestis*). The chromosome of the strain kim (*Yersinia pestis)* is 4,600,755 base pairs long. The chromosome of strain CO92 (*Yersinia pestis*) is 4,653,728 base pairs long.[2,3]

Model organisms: Several scientific studies have been validated through animal models which provide comprehensive insights about the biological subject under study. A model organism elaborative study of the biological phenomena with the expectations of a cohesive understanding. Model organisms are used to explore potential causes and treatment for the human diseases as the involvement of humans in considered being unethical. As a matter of fact, these studies enhances the scope of exploring various facets affiliated facets like metabolic, developmental pathways and genetic material over the course of evolution including the biological significance of mutations.

Selection of the model organisms:

The query sequence is from a bacterium (Yersinia pestis) which is a prokaryote, so the selected model organisms include more number of prokaryotes with some eukaryotes.

Following are the model organisms selected:

1. *Yersinia pestis* Kim
2. *Yersenia pestis*
3. *Bacillus lichiniformis*
4. *Bacillus anthracis*
5. *Caenorhabditis elegans*
6. *Gallus gallus*
7. *Rattus norvegicus*
8. Homo sapiens

Principle

Hypothetical proteins are proteins whose existence is predicted, but function is unknown. The principle involved in the analysis of the function of protein is as follows:

Proteins having similar sequences share structural similarity and have similar function. This is because similar sequences have certain amino acids conserved which fold in the similar manner to form the 3D structure. As the folding is similar their functions are assumed to be similar.

Databases and tools used

S.no	Databases/servers	Tools	Purpose
1	NCBI	BLAST P	Pair sequence similarity search
2	SDSC BIOLOGY WORKBENCH	CLUSTAL W	Multiple sequence alignment
		TEX SHADE	Conservation of residues(local alignment)
		BOX SHADE	Conservation of residues(global alignment)
		CLUSTAL DISTANCE MATRIX	Evolutionary distance with clustal W
3	EXPASY	PROTPARAM	Computing protein parameters
		HNN	Computing secondary structure elements
		CPH MODEL	3D structure prediction
		PHYRE	3D structure prediction
4	PDB		3D structure data bank of proteins
5	RASMOL		3D structure visualization tool
6	PFAM		Protein family domains
7	PROSITE		Protein family based on patterns and profiles
8	PSORT		Cellular protein location
9	PRODOM		Proteins domains
10	SIGNAL P		Signaling peptides
11	TARGET P		Protein location
12	MOTIF		Locating conserved regions
13	SOSUI		Information on protein solubility
14	KEGG		Information on biological pathways

NCBI

National center for biotechnology information is public database (U.S, NLM) that serves as one of the largest reserve and resource of biological information and conducts research in computational biology to develop software tools for analyzing genome data, and disseminates biomedical information, for the better understanding of molecular process affecting human health and disease.

European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ), NCBI works for world wide data exchange. Various databases and tools such as nucleotide, protein, genome, cancer, chromosome, 3D domain structure taxonomy and journals and maintain by NCBI. Entrez is a retrieval system of NCBI designed for searching several linked databases.[4]

BLAST

BLAST stands for Basic Local Alignment Search Tool is a NCBI tool, conducts sequence similarity search, for identifying genes and genetic features. BLAST can execute sequence searches against entire DNA database in less than 15 seconds .Different types of BLAST such as protein blast (BlastP), nucleotide blast, BlastX, TblastN,TblastX are used for nucleotide and protein search.[5]

SDSC biology workbench (San Diego supercomputer center)

SDSC stands for San Diego Super computer is a Biological Workbench. SDSC is a web-based tool, which allows biologists to search many popular protein and nucleotide sequence databases and modeling tools, all within a point and click interface that elements file format compatibility problems. Clustal W is a workbench protein tool used for multiple sequence analysis. Tex shade as Box shade tools are used to see the conservation within the protein useful to predict domins**.** Clustal distance matrix calculates the evalution time with Clustal W. All these tools are useful for evolution studies.[6]

EXPASY

Expasy (Expert protein Analysis System) is a proteomics server maintained by the Swiss Institute of Bioinformatics (SIB) performs the analysis of protein sequences and structures as well as 2D-PAGE. Databases like Unipart, Prosite, Swiss-2D and Enzyme page and, tools for protein search, patter Topology prediction, Structure analysis and Pylogenitical analysis are included in Expasy.[7]

PDB (Protein Data Bank): PDB is a protein structure database for studying the structures of biological macromolecules to know their function.[8]

RASMOL: It is a Molecular Visualization Freeware used to visualize 3D-Structure of protein.DNA and macromolecules to know their function.[9]

Pfam-A: The database is a large collection of protein families, represented by multiple sequence alignments and hidden Markov models (HMMs).[10, 12]

PRODOM: Prodom is a protein domain database.[11]

PROSITE: PROSITE is an online database that maintains the entries of proteins which includes their description, families and functions.

TARGETP: predicts the sub-cellular location and cleavage sites of macromolecules.

SOSUI: This allows the classification and secondary structure prediction of membrane proteins.

KEGG (Kyoto Encyclopedia of genes and Genomes): This is a database resource for comprehending the functional abilities of the biological system

Procedure

The query sequence given is a hypothetical protein with accession number AAM86904. The sequence is retrieved and is identified that it is from a bacteria by name *Yersinia pestis* kim. The steps are as follows.

Step 1: Pair wise sequence similarity search using blast

Blast tool from NCBI is used for identifying similar and homologous sequences. Similarity search is performed using blast with different model organisms as mentioned above, to identify

Citation: Ranganathan V, Khoja KM. Characterization and functional analysis of the hypothetical protein in *Yersinia pestis*. *J Microbiol Exp*. 2022;10(5):170–179. DOI: 10.15406/jmen.2022.10.00370

sequences having functional similarity with good sequence identity. Parameters like expect threshold and blossom were implemented to restrict the search in identifying homologous sequences with a better similarity.

Step 2: Phylogenetic analysis

The similar sequences retrieved from the similarity search are subjected to phylogenetic analysis. The multiple sequence alignment is constructed to find locally conserved residues. As an evidence of the similarity search, clustal distance and dendrogram is drawn to identify the evolutionary relationship, Texshade and boxshade is performed to identify the residues conserved both locally and globally.

Step 3: structural analysis

The structural analysis is performed to identify the physic-chemical parameters of the protein, the secondary structure elements and the fold of the protein. Various tools from the proteomics server expasy are used for the same. The 3d structure analysis also emphasizes on the function of the protein as it identifies all the possible folds the hypothetical protein may acquire. Different principles like homology modeling and threading were used to identify the similar structure for the hypothetical protein.

Step 4: Other biological tools for domain, pattern and family identification

As a hypothesis about the function of the protein is made, several other tools were used to identify the family, domain, pattern other properties of the protein. All these tools were used to support the hypothesis being made.

Results

Results of the similarity search

Database used: NCBI (national centre of biotechnology information)

Tool used: BLAST

Hypothetical protein search

A) Query sequence: *Yersinia pestis* KIM

Accession number: AAM86904

Sequence length: 211aa

Common name: bacteria

Fasta format:

>gi|21960299|gb|AAM86904.1|AE013936_8 hypothetical [Yersinia pestis KIM]

M Q P A I S L L K S A Q E Q M E A I S A D A Q T A T A S P A D L Q A Q I S L L Q Q N L T E L K Q A V L L L S A P K G I A L S S G E H L Q M S A S E N L I A T A G K N A D V S V G K N F F I G V G N T L S V F V R K L G I K L I A N Q G P I T V Q A Q N D L M E L L A R K A I T I T S T E D E I K I T A K K K I T L N A G G S Y I T L D E N R I E S G T A G EYLTKAGYYGRLDKAKLPT EFPALAAKTEDPIKRWLFS

1) Organism: *Yersinia pestis*

Accession number: gb|AAS63631.1

Length: 211

Blast parameters:

E value: 1

Blossom: 80

Blast result:

gb|AAS63631.1| ABC-type multidrug/protein/lipid transport system, ATPase component [Yersinia pestis biovar Microtus str. 91001]

Length=211

GENE ID: 1173804 YPO0967 | hypothetical protein [*Yersinia pestis* CO92]

Score = 441 bits (1014), Expect = 8e-125, Method: Compositional matrix adjust.

Identities = 211/211 (100%), Positives = 211/211 (100%), Gaps = 0/211 (0%)

Fasta format:

>gi|45438083|gb|AAS63631.1| ABC-type multidrug/protein/lipid transport system, ATPase component [Yersinia pestis biovar Microtus str. 91001]

M Q P A I S L L K S A Q E Q M E A I S A D A Q T A T A S P A D L Q A Q I S L L Q Q N L T E L K Q A V L L L S A P K G I A L S S G E H L Q M S A S E N L I A T A G K N A D V S V G K N F F I G V G N T L S V F V R K L G I K L I A N Q G P I T V Q A Q N D L M E L L A R K A I T I T S T E D E I K I T A K K K I T L N A G G S Y I T L D E N R I E S G T A G E Y L T K A G Y Y G R L DKAKLPTEFPALAAKTEDPIKRWLFS

Function:

KATP channels are large hetero-multimeric complexes consisting of four subunits and belong to K+ channel family.

2) Organism: *Bacillus licheniformis*

Sequence length: 328aa

Accession number: gb|ABY43187.1

Blast parameters:

E value: 10

Blossom: 62

Blast result:

gb|ABY43187.1 Alcohol dehydrogenase zinc-binding domain protein [Bacillus lichniformis KBAB4]

Length=328

GENE ID: 5842170 BcerKBAB4_1957 | Alcohol dehydrogenase zinc-binding domain

protein [Bacillus licheniformis KBAB4]

Score = 28.1 bits (61), Expect = 9.0, Method: Compositional matrix adjust.

Identities = 26/79 (32%), Positives = 33/79 (41%), Gaps = 27/79 (34%)

Fasta format:

>gi|163862128|gb|ABY43187.1| Alcohol dehydrogenase zinc-binding domain protein [Bacillus lichniformis KBAB4]

M K A I V V T S F G G P E V L K Y T D M D I P T I S D N Q V L I R V V A T S V N F A D I K S R Y G K K G N K S L P F I P G I D A A G I V E H V G S H V K N I H P G Q R V I T F P Q N G S Y A E Y V V A N E N L T F V L P D E V N F Q T A A A C P I V S F T S Y N L L A N V A R I Q Q G E S V L I H A V A G G I G T T A I Q L A K L L G A K K V I G T V G S E A K R K I A L D A G A D Y V I C H Q D E D F V E R V N Q L T H G E G V N I V L D S I S G T V S E R S L N C L A Y Y G R L V H F G N A S G E V G S F Q T K D L H A S C R S I L G F S F G T T R K K R P E L L Q E T A N E V F R Y L R D G S L QIKATKSFPLQDAGKAHEWVESR QSTGKVILHVQTAP

Function:

87467 Alcohol dehydrogenase GroES-like domain. This is the catalytic domain of alcohol dehydrogenases. Many of them contain an inserted zinc binding domain. This domain has a GroES-like structure.

3) Organism name: *Bacillus anthracis*

Accession number: NP_844511.1

Length: 328aa

Blast parameters:

E value: 10

Blossom: 62

Blast result:

ref|NP_844511.1 quinone oxidoreductase [Bacillus anthracis str. Ames]

Length=328

GENE ID: 1085801 qor-1 quinone oxidoreductase [Bacillus anthracis str. Ames]

Score = 25.4 bits (54), Expect = 5.4, Method: Compositional matrix adjust.

Identities = 21/60 (35%), Positives = 26/60 (43%), Gaps = 21/60 (35%)

Fasta format:

>gi|30262134|ref|NP_844511.1| quinone oxidoreductase [Bacillus anthracis str. Ames]

M K A I V V T S F G G S E V M K Y T D V D I P A I S E D Q V L I R V V A T S V N F A D I K S R Y G K K G N K A L P F I L G I D A A G I V E R V G S H V K N I Y P G Q R V I A F P Q N G S Y A E Y V V A N E N L T F V L P D E V D F Q T A A A C P I V S F T S Y N L L A N V A R L Q Q G E S V L I H A A A G G I G T T A I Q L A K L L G A G T V I G T V G S E A K K E I A L D A G A D Y V I G H Q D E D F V E K V N E L T N G E G V D V I L D S I S G T V S E R S L K C L A Y Y G R L I H F G N A S G E I G N F Q T K D L H A S C R S I L G F S F G T T R K K R P ELLQETANEVFRYLRDGHLQIK ATKSFPLQDAGKAHEW VESRKSTGKVILTVQSSS

Function:

This is the catalytic domain of alcohol dehydrogenases. Many of them contain an inserted zinc binding domain. This domain has a GroES-like structure.

4) Organism: *Caenorharbditis elegans*

Accession number: NP_505795.1

Length: 204 aa

Blast parameters:

E value: 10

Blossom: 62

Blast result:

ref|NP_505795.1| UG F17C11.5 [Caenorhabditis elegans]

Length=204

GENE ID: 184615 F17C11.5 | F17C11.5 [Caenorhabditis elegans]

Score = 27.7 bits (60), Expect = 5.6, Method: Compositional matrix adjust.

Identities = 15/68 (22%), Positives = 32/68 (47%), Gaps = 3/68 (4%)

Fasta format:

>gi|17559816|ref|NP_505795.1| F17C11.5 [Caenorhabditis elegans]

M V H S F S V L L V L G T L F Y Y S Q A Q N A A Q E V K T Q V P K C N L W Q G K C P P G W K T F I R P P G V R I C L K I I N Q N ATYLEAQNICRKQYNSRIHGIE NKEEHNWVKMKSRLVIKQGA S F M W V G A R R K S W C Y K T E A Q L K T D D K C K S G L N S F E W D D Q L T D G S Y L F T Q W D K T I S A P E G I M S N N APEDCVVMNVHEGDAPIDDKD CSLPQAGVLCGMTA

Function:

CLECT: C-type lectin (CTL)/C-type lectin-like (CTLD) domain are protein domains homologous to the carbohydrate-recognition domains (CRDs) of the C-type lectins. They are calcium-dependent carbohydrate binding modules and bind to a variety of carbohydrate ligands including mannose, N-acetylglucosamine, galactose, N-acetylgalactosamine, and fucose.

5) Organism: *Gallus gallus*

Accession number: gb|AAB94071.1

Length: 238 aa

Blast parameters:

E value: 10

Blossom: 80

Blast result:

gb|AAB94071.1| mannan-binding lectin; collectin [Gallus gallus]

Length=238

GENE ID: 374267 MBL2 | mannose-binding lectin (protein C) 2, soluble

[Gallus gallus] (10 or fewer PubMed links)

Score = 28.0 bits (56), Expect = 5.7, Method: Compositional matrix adjust.

Identities = 16/53 (30%), Positives = 32/53 (60%), Gaps = 3/53 (5%)

Fasta format:

>gi|2736145|gb|AAB94071.1| mannan-binding lectin; collectin [Gallus gallus]

M M A T S L L T T D K P E E K M Y S C P I I Q C S A P A V N G L P G R D G R D G P K G E K G D P G E G L R G L Q G L P G K A G P Q G L K G E V G P Q G E K G Q K G E R G I V V T D D L H R Q I T D L E A K I R V L E D D L S R Y K K A L S L K D V V N I G K K M F V S T G K K Y N F E K G K S L C A K A G S V L A S P R N E A E N T A L K D L I D P S S Q A Y I G I S D A Q T E G R F M Y L S G G P L T YSNWKPGEPNNHKNEDCAVIED SGKWNDLDCSNSNIFIICEL

Function:

C-type lectin-like domain (CTLD) of the type found in humans including lung surfactant proteins A and D, mannose- or mannan binding lectin (MBL), and CL-L1. These are carbohydrate recognizing domains.

6) Organism: *Rattus norvegicus*

Length: 211aa

Accession number: NP_001099752.1

Blast parameters:

E value: 20

Blossom: 45

Blast result:

ref|NP_001099752.1| mitochondrial ribosomal protein L48 [Rattus norvegicus]

Length=211

GENE ID: 293149 Mrpl48_predicted| mitochondrial ribosomal protein L48

(predicted) [Rattus norvegicus]

Score = 26.6 bits (75), Expect = 16, Method: Compositional matrix adjust.

Identities = 12/33 (36%), Positives = 18/33 (54%), Gaps = 0/33 (0%)

Fasta format:

>gi|157822477|ref|NP_001099752.1| mitochondrial ribosomal protein L48 [Rattus norvegicus]

MSGTLGKVLGLWTNTVSK QGFSLQRFRILGENPIYSAGG IVRTSRHYKTKPTHGIGRYRH LVKVLEPKKKKAKVELRAIN VGTDYEYGVLNIHLTAYDMT LAESYARYVHRLCNQLSIKVE ESYAMPTKTMEVMRLPDQGN KMVLDSVLTTHERVVQISGL SATFAEIFLEILQINLPEGVRL SVREHTEEDFKGRFKARPELE ELLAKLN

Function:

Studies have claimed the significance of mitochondrial ribosomal protein L41 (MRPL41) in stabilizing p53 leading to p53-induced apoptosis in response to growth-inhibitory conditions such as actinomycin D treatment and serum starvation.

7) Organism: Homo sapiens

Accession number: gb|AAH57760.1|

Length: 240

Blast parameters:

E value: 20

Blossom: 45

Blast result:

gb|AAH57760.1| MORN repeat containing 3 [Homo sapiens]

Length=240

GENE ID: 283385 MORN3 | MORN repeat containing 3 [Homo sapiens]

(10 or fewer PubMed links)

Score = 27.7 bits (79), Expect = 18, Method: Compositional matrix adjust.

Identities = 18/54 (33%), Positives = 26/54 (48%), Gaps = 4/54 (7%)

Fasta format:

>gi|34785506|gb|AAH57760.1| MORN repeat containing 3 [Homo sapiens]

MPVSKCPKKSESLWKGWD RKAQRNGLRSQVYAVNGDY YVGEWEDNVKHGKGTQVW KKKGAIYEGDWKFGKRDGY GTLSLPDQQTGKCRRVYSGW WKGDKKSGYGIQFFGPKEYY EGDWCGSQRSGWGRMYYSN GDIYEGQWENDKPNGEGMLR LKNGNRYEGCWERGMKNGA GRFFHLDHGQLFEGFWVDNM AKCGTMIDFGRDEAPEPTQ FPIPEVKILDPDGVLAEALA MFRKTEEGD

Function:

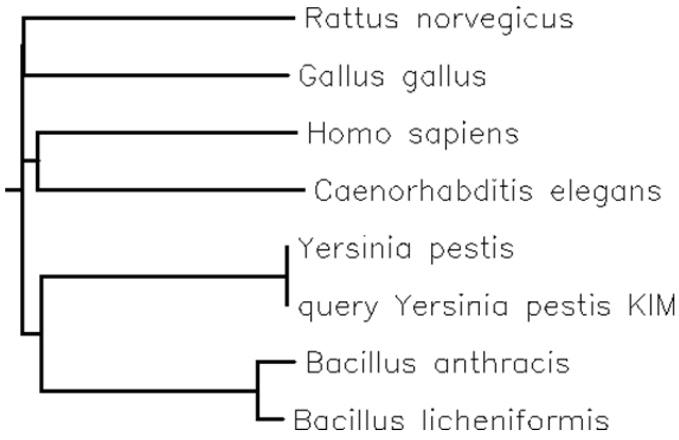
The retinophilins code for macromolecules associated with the expression of the phyto receptor cells which is believed to be highly conserved.

Phylogenetic analysis

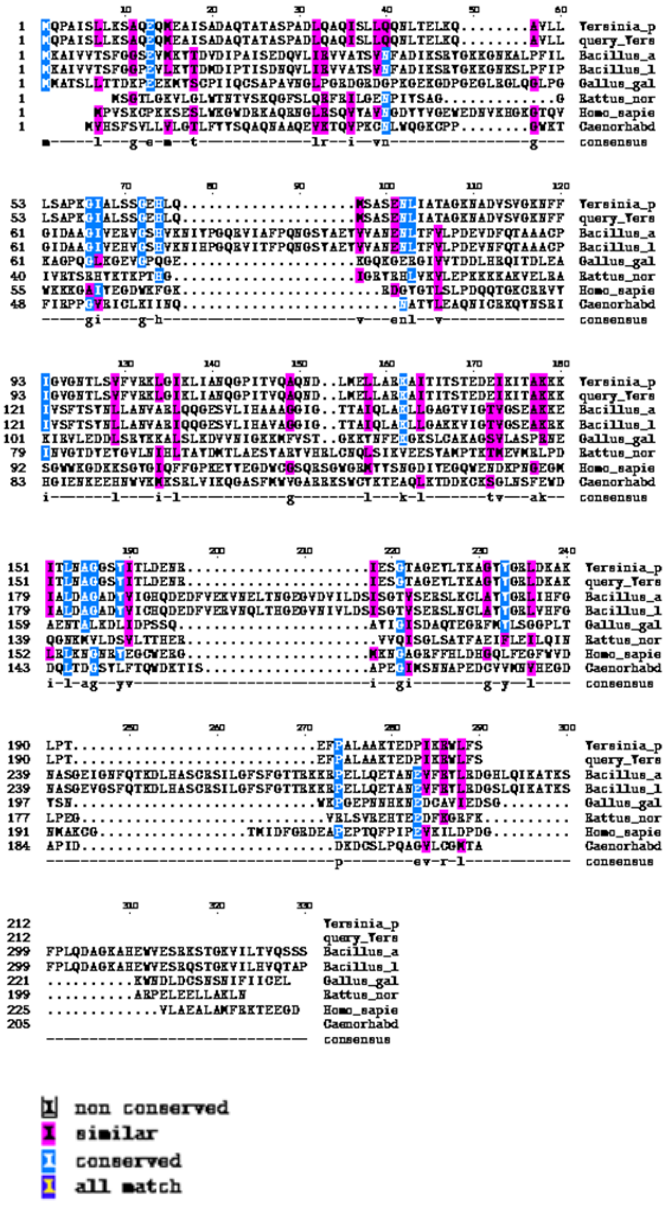
The software used for this analysis is SDSC biology work bench. SDSC is San Diego super computer center. Tools used in this software are as follows.

1) Clustal W
2) Tex shade
3) Box shade
4) Clustal distance

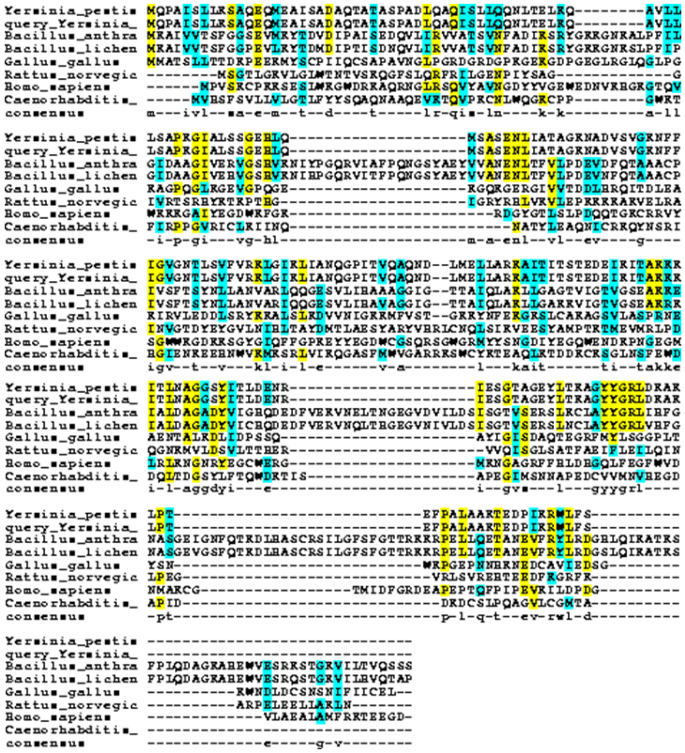
Clustal W results: Multiple sequence alignment



Tex shade:



Box Shade:



Green: completely conserved residues

Yellow: identical redidues

Cyan: similar residues

White: different residues

Clustal distance matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Yersinia_pestis (1)	0.000	0.000	0.829	0.829	0.924	0.918	0.921	0.910
query_Yersinia_pesti (2)	0.000	0.000	0.829	0.829	0.924	0.918	0.921	0.910
Bacillus_anthracis (3)	0.829	0.829	0.000	0.107	0.870	0.880	0.916	0.941
Bacillus_licheniform (4)	0.829	0.829	0.107	0.000	0.874	0.885	0.916	0.941
Gallus_gallus (5)	0.924	0.924	0.870	0.874	0.000	0.889	0.898	0.899
Rattus_norvegicus (6)	0.918	0.918	0.880	0.885	0.889	0.000	0.893	0.927
Homo_sapiens (7)	0.921	0.921	0.916	0.916	0.898	0.893	0.000	0.920
Caenorhabditis_elega (8)	0.910	0.910	0.941	0.941	0.899	0.927	0.920	0.000

Structural analysis

The software used is EXPASY. The tools used are as follows:

- Primary structure analysis: Protparam
- Secondary structure analysis: HNN (hierarchical neural network)
- Tertiary structure analysis: Tools used are
  - CPH model: Comparative protein homology model
  - Phyre: Protein homology recognition engine
  - HH Pred: Homology prediction

Citation: Ranganathan V, Khoja KM. Characterization and functional analysis of the hypothetical protein in *Yersinia pestis*. *J Microbiol Exp*. 2022;10(5):170–179. DOI: 10.15406/jmen.2022.10.00370

Primary structure analysis

Software used: Expasy

Tool used: Protparam

Results:

ProtParam

Number of amino acids: 211

Molecular weight: 22545.9

Theoretical pI: 7.84

Amino acid composition:

Ala (A) 28	13.30%
Arg (R) 5	2.40%
Asn (N) 9	4.30%
Asp (D) 8	3.80%
Cys (C) 0	0.00%
Gln (Q) 13	6.20%
Glu (E) 13	6.20%
Gly (G) 14	6.60%
His (H) 1	0.50%
Ile (I) 17	8.10%
Leu (L) 26	12.30%
Lys (K) 17	8.10%
Met (M) 4	1.90%
Phe (F) 5	2.40%
Pro (P) 7	3.30%
Ser (S) 16	7.60%
Thr (T) 16	7.60%
Trp (W) 1	0.50%
Tyr (Y) 4	1.90%
Val (V) 7	3.30%
Pyl (O) 0	0.00%
Sec (U) 0	0.00%

Total number of negatively charged residues (Asp + Glu): 21

Total number of positively charged residues (Arg + Lys): 22

Atomic composition:

Carbon C	1000
Hydrogen H	1648
Nitrogen N	268
Oxygen O	312
Sulfur S	4

Formula: $C_{1000}H_{1648}N_{268}O_{312}S_4$

Total number of atoms: 3232

Extinction coefficients

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

Ext. coefficient 11460

Abs 0.1% (=1 g/l) 0.508, assuming ALL Cys residues appear as half cysetines

Estimated half-life:

The N-terminal of the sequence considered is M (Met).

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 33.27

This classifies the protein as stable.

Aliphatic index: 102.37

Grand average of hydropathicity (GRAVY): -0.060

Secondary structure analysis:

Tool used: HNN (Hierarchical neural network)

Hierarchical Neural Network result

10 20 30 40 50 60 70

| | | | | | |

M Q P A I S L L K S A Q E Q M E A I S A D A Q T A T A S P A D L Q A Q I S L L Q Q N L T E L K QAVLLLSAPKGIALSSGEHLQMS

c c c h h h h h h h h h h h h h h h h c c c c c c c c c h h h h h h h h h h h hhhhhhhhhehhccccceeeccccchhc

A S E N L I A T A G K N A D V S V G K N FFIGVGNTLSVFVRKLGIKLIANQGPITVQAQNDLMEL LARKAITITSTE

h h h h h h h c c c c c c e e e c c c e e e e e c c h h h h h h h h h c e e e e eeccccceeeecchhhhhhhhhhe eecccc

D E I K I T A K K K I T L N A G G S Y I T L D E N R I E S G T A G E YLTKAGYYGRLDKAKLPTEFPALAA KTEDPIKRWLF

h h h h h h c c e e e e c c c c c e e e e e c c c c e c c c c c c h h h h h c c c c c c cccccccccccchhhhhhcchhhhec

S

c

Sequence length: 211

HNN

Alpha helix (Hh)	:	92 is	43.60%
310 helix (Gg)	:	0 is	0.00%
Pi helix (Ii)	:	0 is	0.00%
Beta bridge (Bb)	:	0 is	0.00%
Extended strand(Ee)	:	35 is	16.59%
Beta turn (Tt)	:	0 is	0.00%
Bend region (Ss)	:	0 is	0.00%
Random coil (Cc)	:	84 is	39.81%
Ambigous states (?)	:	0 is	0.00%
Other states	:	0 is	0.00%

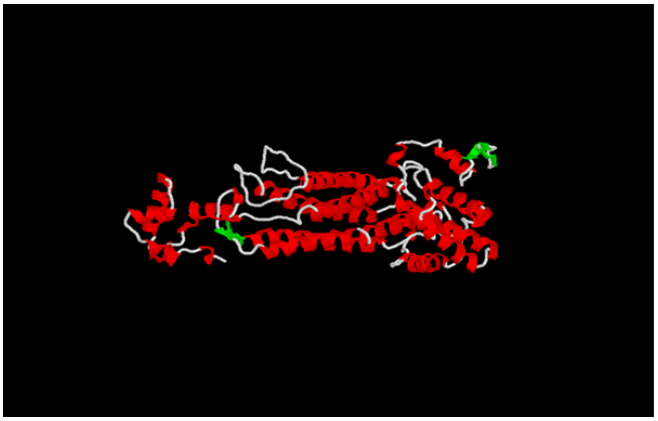
Tertiary structure analysis

CPH model result

entry: 1YFM chain: A score: 29 E: 3.5

SCOP Classification (version 1.71)	Domain Info	Class	Fold	Superfamily	Family	Domain	Species
	d1yfm__	All alpha proteins	L-aspartase-like	L-aspartase-like	L-aspartase/ fumarase	Fumarase	Baker'syeast (Saccharomyces cerevisiae)

Citation: Ranganathan V, Khoja KM. Characterization and functional analysis of the hypothetical protein in *Yersinia pestis*. *J Microbiol Exp.* 2022;10(5):170‒179. DOI: 10.15406/jmen.2022.10.00370



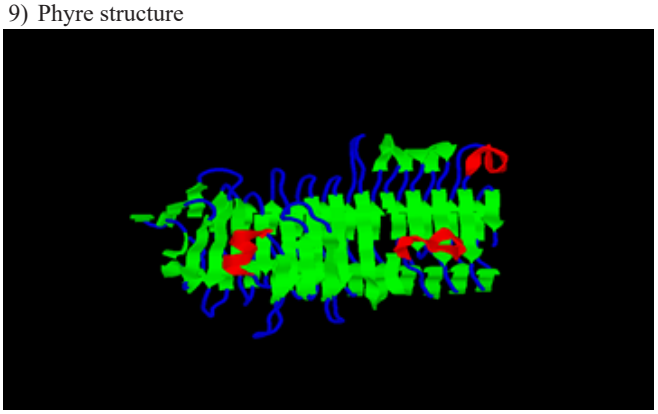
Red colour: helices

Green colour: sheets

White colour: co

Phyre result:

1) SCOP code: C2OdlA
2) E value: 8.6
3) Estimated precision: 25%
4) Bio text: n/a
5) PDB header: Cell adhesion
6) Super family: Chain APDB molecule adhesion
7) Family:
8) PDB title: Crystal structure of HMW1 secretion domain from 2 haemophilus influenza.
9) Phyre structure



Red colour: helices

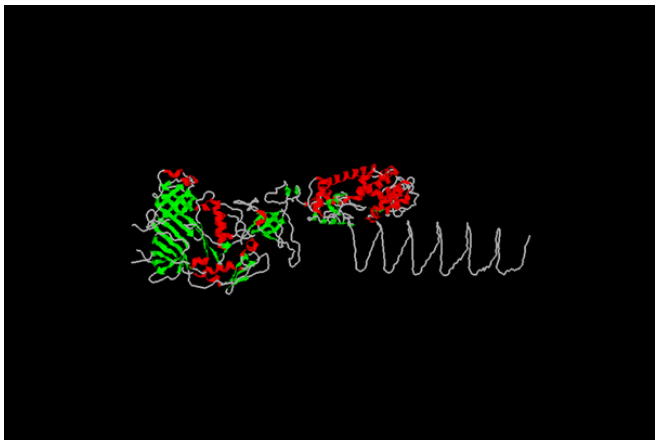
Green colour: sheets

Blue colour: coils

HH pred result

1wth_A Protein GP5, tail-associated lysozyme; triple-stranded beta-helix, OB fold, pseudohexamer, T4 tail

- lysozyme, GP5-GP27; 2.80A {Bacteriophage T4}
- Probab=37.15
- E-value=0.31
- Score=24.37
- Aligned_cols=102
- Identities=17%
- Similarity=0.169
- Sum_probs=0.0



Red colour: helices

Green colour: sheets

White colour: coils

Specialized database results

1) Psort result:

Query Protein details

cyto: 22.0,

pero: 4.0,

nucl: 3.0

2) Prodom result:

>PD494095 (Closest domain: Q8ZHD3_YERPE 1-125)

Number of domains in family: 34

Commentary (automatic):

VGR-RELATED RELATED PLASMID PROBABLE SIMILAR VGR VGRG YPO1472 YPO0962 C3393

Length = 125

Score = 519 (204.5 bits), Expect = 2e-52

Identities = 125/125 (100%), Positives = 125/125 (100%)

3) Target P result:

Name Len mTP SP other Loc RC

---

gi_21960299_gb_AAM86 211 0.092 0.077 0.868 _ 2

---

cutoff 0.000 0.000 0.000

4) Sosui result:

This amino acid sequence has no signal peptide.

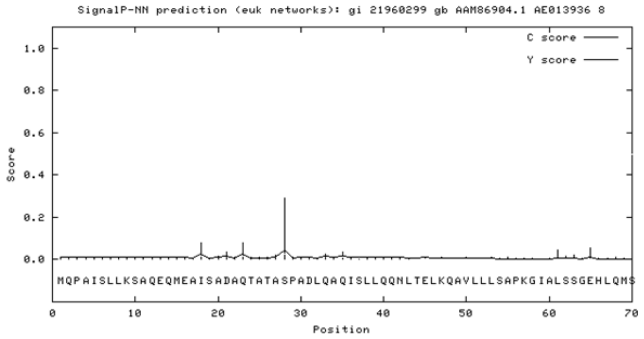
This amino acid sequence is of a SOLUBLE PROTEIN

Citation: Ranganathan V, Khoja KM. Characterization and functional analysis of the hypothetical protein in *Yersinia pestis*. *J Microbiol Exp*. 2022;10(5):170–179. DOI: 10.15406/jmen.2022.10.00370

5) Signal P result:

>gi_21960299_gb_AAM86904.1_AE013936_8 hypothetical _ *Yersinia pestis* KIM_

Signal P-NN result:

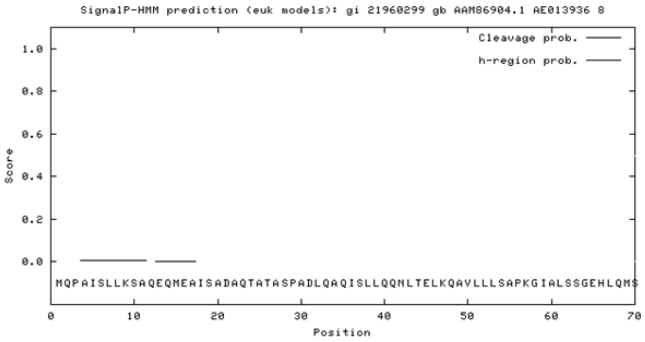


data

>gi_21960299_gb_AAM86 length = 70

Measure Position	Value	Cutoff	Signal	Peptide?
max. C	28	0.292	0.32	NO
max.Y	28	0.046	0.33	NO
max. S	2	0.056	0.87	NO
mean S	1-27	0.028	0.48	NO
D	1-27	0.037	0.43	NO

Signal P-HMM result



data

>gi_21960299_gb_AAM86904.1_AE013936_8

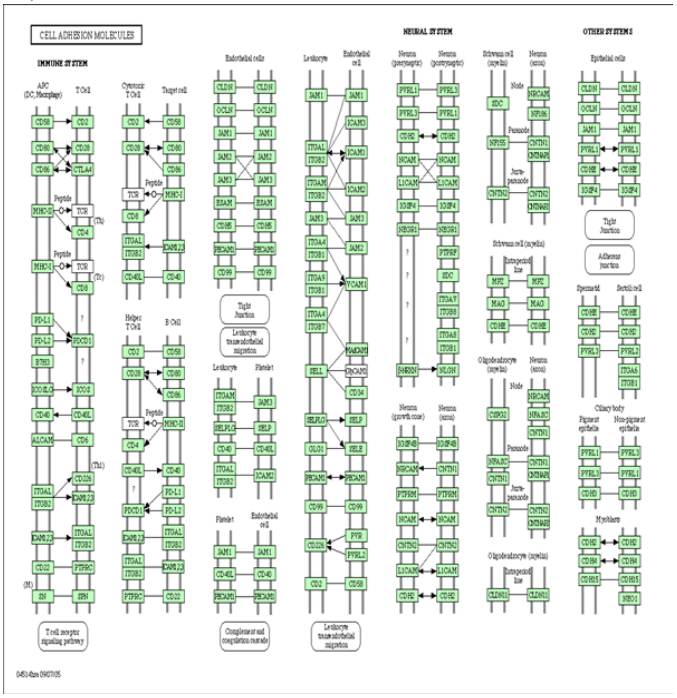
Prediction: Non-secretory protein

Signal peptide probability: 0.006

Signal anchor probability: 0.000

Max cleavage site probability: 0.001 between pos. 27 and 28

6) KEGG result



The primary structure analysis done by tool protparam identifies the protein to be a hydrophilic protein. The secondary structure analysis by HNN shows that the protein has high percentage of helices which indicates that the protein is hydrophilic. The tertiary structure analysis by CPH model identified a similar structure to that of query in the protein data bank (PDB). The PDB id is 1YFM. The tertiary structure of the protein was observed using RASMOL. This tool is used to visualize the 3D structure of the biological molecules. The 3D structure also showed more number of helices when compared to sheets which indicates that the protein is hydrophilic in nature.

Psort gives the information about the protein location. Signal P identified that the protein is not a signal peptide. Sosui identified the protein to be a soluble protein.

Inferences

The results of the similarity search using BLAST shows that the protein functions similarly in all the model organisms taken. The hits are selected based on their identities, e value, score, gaps and positives.

Phylogeny is the ancestral study. The phylogenetic analysis identifies the evolutionary relationship to the query sequence with the similar sequences obtained from the model organisms, providing an evidence for the hypothesis made from the similarity search results. The dendogram and clustal distance matrix shows that the protein is very much related to the lower organisms where as the higher organisms were branched out separately. The Texshade and Box shade results give a clear picture of the conservation observed between the residues of the query with that of the sequences of the model organisms taken from the similarity search. Pink and blue colour indicates the similar and conserved residues respectively in Texshade. In box shade yellow and cyan colours indicate the similar and conserved residues respectively.

Discussion

After performing the analysis for the hypothetical protein using different tools it is inferred that the hypothetical protein is the protein belonging to the class of binding proteins and plays a major role in binding mechanism. The protein is related to the proteins belonging to K+ channel family which are the trans membrane proteins. The nucleotide binding domains of the SUR1 interact with proteins in this family.

The protein is also associated with related Zn-dependent oxidoreductases and Alcohol dehydrogenase GroES-like domain. The protein is mainly involved in binding function preferable at the zinc binding domain. The protein also performs function similar to CTLDs. These are calcium-dependent carbohydrate binding modules.

Citation: Ranganathan V, Khoja KM. Characterization and functional analysis of the hypothetical protein in *Yersinia pestis*. *J Microbiol Exp*. 2022;10(5):170–179. DOI: 10.15406/jmen.2022.10.00370

Apart from the above function the protein also acts as a pulmonary surfactant and reduces the surface tension within the lungs.

The CTLDs of these collectins bind carbohydrates on surfaces (e.g. pathogens, allergens, necrotic, or apoptotic cells) and mediate functions associated with killing and phagocytosis. SP-A and SP-D in addition to functioning as host defense components, are components of pulmonary surfactant which play a role in surfactant homeostasis. Pulmonary surfactant is a phospholipid-protein complex which reduces the surface tension within the lungs.

Conclusion

The given hypothetical protein with accession number AAM86904 is from a bacteria named *Yersinia pestis* kim. Kim is the strain to which the bacterium belongs. From the above analysis it is found that the protein is a hydrophilic protein with more number of helices than sheets and is involved in binding mechanism preferably at the zinc binding domain and also plays a minor role as pulmonary surfactants.

Acknowledgments

None.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

1. http://en.wikipedia.org/wiki/Yersinia_pestis
2. Deng W, Burland V, Plunkett G, et al. Genome sequence of *Yersinia pestis*. *J Bacteriol.* 2002;184(16):460–4611.
3. http://www.systems-biology.org/001/kegg/ypk.html
4. Basic local alignment search tool; 2022.
5. http://workbench.sdsc.edu/
6. http://expasy.org/prosite/
7. http://www.rcsb.org/pdb/home/home.do
8. http://www.umass.edu/microbio/rasmol/
9. http://pfam.sanger.ac.uk/
10. http://prodom.prabi.fr/prodom/current/html/home.php
11. http://www.sanger.ac.uk/Software/Pfam/
12. Gouet P, Courcelle E, Stuart, et al. Multiple sequence alignment in post script. *Bioinformatics*. 1999;15(4):305–308.
13. Nakai K, Horton P. Psort: a program for detecting the sorting signals of the proteins and predicting their sub cellular localization. *Trends Biochem Sci*. 1999;24(1):34–36.
14. Emanuelson O, Brunek S, Nielson H. Locating proteins in cell using target P, signal P and related tools. *Nat protoc*. 2007;2(4):953–971.