TLS
Trends in Life Sciences
An International Peer-reviewed Journal
www.sciencejournal.in

ENZYME FUNCTION PREDICTION IN THE HYPOTHETICAL PROTEINS OF *YERSINIA PSEUDOTUBERCULOSIS* -WAY TO LINK NEW PATHWAY

A.S. Kulkarni
Department of Microbiology : Dharampeth M. P. Deo Memorial Science College, , Nagpur.
Corresponding author: archanakulkarni6212@gmail.com M: 9823374091

ABSTRACT
The pathogenicity of *Yersinia pseudotuberculosis* is increasing not only in animals but also among human. The genome sequence of it gives us detail insight about protein-coding ability and molecular analysis is also possible with many uncharacterized proteins marked on the genome. In the present study, hypothetical proteins encoded by the *Y. pseudotuberculosis* searched for the available conserved domain capable of encoding enzyme function once searched by servers like CDDBLAST, Interproscan, PFAM and CATH. The structure-function relation of enzyme coding hypothetical proteins determined by homology modelling to decipher the tertiary structure of a hypothetical protein using close sequence template available with RCSB PDB. In a result, 34 hypothetical proteins out of 759 proteins (Hypothetical) linked with enzyme function successfully with 100% confidence level. Among them, 15 hypothetical proteins structurally modelled that showcase structural homolog also. In a conclusive remark, hypothetical proteins of *Y. pseudotuberculosis* predicted to function like enzyme and demanded a further investigation by cloning and expression studies with ideal host as *E. coli* to confirm its metabolic function in *Y. pseudotuberculosis*.

KEYWORDS: Hypothetical protein, Conserved Domain, Bioinformatics, Homology Modelling

INTRODUCTION

As per the pathogenic link, the bacterium -*Yersinia pseudotuberculosis* reported being foodborne pathogen bringing about acute gastrointestinal illness (Kim *et al.*, 2018). The resultant gastrointestinal infection by the *Y. pseudotuberculosis* remains persistent and almost complicated that bring about relapsing enteritis and sometimes severe autoimmune disorders (Heine *et al.*, 2018). It is important to learn about the new protein-coding organism like *Y. pseudotuberculosis* expresses Rfalt that enhances transcription of the number of operons involved in lipopolysaccharide formation and that results in resistance towards antimicrobial chemokines and assures an increase in virulence (Hoffman *et al.*, 2017). Researchers also investigated in detail about genome arrangement of *Y. pseudotuberculosis*.

One such study carried out genome analysis of 134 strains of *Y. pseudotuberculosis* and used CRISPER in understanding evolutionary trajectory and protein-based functions (Seecharran *et al.*, 2017). Willcocks *et al.*, (2018) reported *Y. pseudotuberculosis* as the zoonotic pathogen, that can bring about gastrointestinal infection in human. Here they genome marked the gene ypt 3665 involved in peptide deformylase, that makes the organism sensitive towards actinonin, a deformylase inhibitor. This finding is put forward by close homolog study of other *Yersinia spp.* related successfully with divergence and homology of the species. (Will cocks *et al.*, 2018). Researcher An *et al.*, (2009) related one gene Ker V able to encode hypothetical methyltransferase and found to be highly conserved among the other genera such as *Burkholderia*, *Escherichia*, *Shigella*, *Vibrio*, and *Yersinia*. Garborm *et al.*, (2004) advocated linking novel virulence-associated genes once as hypothetical protein in *Yersinia* sp., *Helicobacter* sp., *Borrelia* sp., and *Streptococcus* sp.

Lastly, Schrimpe-Rutledge *et al.*, (2012) strongly recommended adopting the methodology for genome annotations especially while studying *Yersinia* species. The emphases on use of omics-based annotation methodology to link un-annotated genome of *Yersinia* species once taking the assistance of computational biology. The searching for function in hypothetical proteins along with virulence genes and likewise is strongly recommended.

In the present study, an attempt has been made to search enzyme function in the hypothetical proteins of *Y. pseudotuberculosis* by involving the bioinformatic approach. The structure-function relationship has also been established with several hypothetical proteins specially to get engage in enzyme activity.

MATERIALS AND METHODS

Methodology

Data collection of protein sequence

The pathogen *Yersinia pseudotuberculosis* has been sequenced for the genome to encode the number of proteins expressed by the genes. The server available at www.genome.JP/kgg used to retrieve the protein sequence using the code 'ypf' assigned for the organism. These sequences saved in 'FASTA' format and used further for the screening of hypothetical proteins encoded by the genome as per record.

Search for conserved domain

Once the number of hypothetical proteins marked on the *Y. pseudotuberculosis* genome, those proteins with hypothetical features searched for the enzyme coding ability by locating signatures of conserved domain assigned for enzyme function. The analysis made realistic by involving conserved domain search engines such as-

a) **Conserved domain BLAST**

The server is provided by the NCBI with website extension www.ncbi.nlm.nih.gov/BLAST. The server searched CDD 27036 PSSM's database having the detailed entries of protein conserved domain families (Altschul *et al.*, 1997; Schaffer *et al.*, 2001).

b) **Interproscan:**

The server used the number of sequence database to determine conserved domain feature to query protein such as Blastprodom, FPrintscan, HMMPIR, HMMPfam and others.

c) **PFAM**

The server is available at www.pfam.sanger.ac.uk/ able to find out a conserved domain in query protein once E-value set at 1 as recorded in the present study.

d) **CATH**

The server class, Architecture, topology and Homology (CATH) involves functional family (Funfams) sub-classification method to give better search facility of conserved domains available in query protein.

Functional Categorization

Since in study four specific server searching enzyme domain in hypothetical proteins, the search performance grouped as 100%, 75%, 50%, 25% and 0% once 4 programs given same enzyme function, similarly 3, 2, 1 and 0, respectively. Here only those hypothetical proteins showcasing 100% confidence reported being promising to investigate further.

Function prediction via protein three-dimensional analysis

Once the proteins recorded with enzyme activity by conserved domain sequence homology, these proteins were analyzed further for their three-dimensional structure homolog indicating their real structure function-based evidences. In the study, homology modelling concept utilized to ascertain the three-dimensional structure of enzyme coding hypothetical proteins once input of it given to the PS^2 protein structure prediction server. Their server utilizes a consensus strategy to use PSI-BLAST, IMPALA and 7 Coffee to select best-scored template and target template alignment. The server then engages Modeler to build a three-dimensional structure. The mentioned server available at www.ps2.life.nctu.edu.tw/ (Chih-Chieh Chen *et al.*, 2006). Once the modeller template remains specific to the earlier result of conserved domain derived from four programs then and then only result considered positive and otherwise rejected.

RESULT

Presence of hypothetical proteins

As per genome sequencing followed by marking of hypothetical proteins presence, *Y. pseudotuberculosis* found to be having 759 hypothetical proteins. These all proteins tested successfully for determining the presence of enzymatic conserved domain.

Confirmation of enzyme domain

The *Y. pseudotuberculosis* hypothetical proteins once analysed for conserved domain using CDD-BLAST, Interproscan, CATH and PFAM, its enzyme coding ability detected in at least 358 hypothetical proteins; while 401 remained either uncharacterized or having a non-enzymatic function. These 358 proteins grouped in variable confidence level as 100% for 34 proteins, 75% for 29, 50% for 27, 25% for 268 proteins as given in Table 1. The hypothetical protein predicted to encode enzyme function with 100% confidence showcased in Table 2 with the exact enzyme function those predicted with.

Trends in Life Sciences
An International Peer-reviewed Journal
www.sciencejournal.in

Table 01: Functional annotation details for the Hypothetical proteins of *Y. pseudotuberculosis*

Total Hypothetical proteins	759
Enzyme domain Hypothetical proteins	358
25%	268
50%	27
75%	29
100%	34

Table No 2: Enzymatic conserved domains detected in *Y. pseudotuberculosis* proteome of hypothetical proteins

Sr.No.	CDD BLAST	Interproscan	PFAM	CATH	%
BZ19_241	Acyl-coa dehydrogenase	Acyl-coa dehydrogenase-like, C-terminal & c-terminal	Acyl-coa dehydrogenase	Butyryl-Coa Dehydrogenase, subunit A, domain 1	100
BZ19_279	Maltodextrin glucosidase	Maltodextrin glucosidase.	alpha amylase.	Glycosidases	100
BZ19_404	Diacylglycerol kinase	Inositol phosphor transferase.	Phosphatase, catalytic domain	Protein tyrosine phosphatase	100
BZ19_435	Predicted dehydrogenase	NAD(P)-binding domain superfamily. Oxidoreductase, C-terminal	Oxidoreductase family, NAD-binding Rossmann fold.	Dihydrodipicolinate Reductase; domain 2	100
BZ19_612	ATP-dependent RNA helicase rhle	ATP-dependent RNA helicase rhle.	Helicase conserved C-terminal domain	P-loop containing nucleotide triphosphate hydrolases	100
BZ19_632	2OG-Fe(II) oxygenase	Iron-dependent dioxygenase	2OG-Fe(II) oxygenase	Phosphodiesterase	100
BZ19_676	Aldolase	Class ii aldolase/adducin n-terminal domain superfamily	Class ii aldolase and adducin n-terminal domain	l-fuculose-1-phosphate aldolase	100
BZ19_868	DNA helicase IV	DNA helicase.	DNA helicase IV / RNA helicase N terminal.	P-loop containing nucleotide triphosphate hydrolases	100
BZ19_1178	Dnab helicase C terminal domain.	DNA helicase dnab, N-terminal.	Dnab-like helicase N terminal domain.	P-loop containing nucleotide triphosphate hydrolases	100
BZ19_1245	Ribonuclease toxin, brnt, of type II toxin-antitoxin system.	Ribonuclease toxin, brnt, of type II toxin-antitoxin system	Ribonuclease toxin, brnt, of type II toxin-antitoxin system.	Aldolase class I.	100
BZ19_1294	Putative endopeptidase	Creatinase/aminopeptidase	creatinase/prolidase n-terminal domain.	Creatinase/methionine aminopeptidase superfamily	100
BZ19_1436	Prka family serine protein kinase	Serine-protein kinase.	Prka serine protein kinase C-terminal domain	P-loop containing nucleotide triphosphate hydrolases	100
BZ19_1618	L-Ala-D/L-Glu epimerase	Enolase-like, N-terminal & C-terminal.	Enolase C-terminal domain-like	Enolase superfamily	100
BZ19_1645	Anhydro-N-acetylmuramic acid kinase	Anhydro-N-acetylmuramic acid kinase	Anhydro-N-acetylmuramic acid kinase	Aldolase class I	100
BZ19_1651	Putative metal dependent hydrolase	Alkaline phosphatase-like, alpha/beta/alpha.	sulfatase	Alkaline phosphatase.	100
BZ19_1682	FAD/FMN-containing dehydrogenase. Fe-S oxidoreductase	CO dehydrogenase flavoprotein-like, FAD-binding.	FAD binding domain. FAD linked oxidases, C-terminal domain.	NADP-dependent oxidoreductase	100
BZ19_1877	Predicted glycosyl hydrolase.	Mannoside phosphorylase.	Beta-1,4-mannooligosaccharide phosphorylase	Glycosyl hydrolase domain	100
BZ19_2312	Urease accessory protein uree	Uree urease accessory, N-terminal & C-terminal	Uree urease accessory protein, C-terminal domain	Urease metallochaperone	100
BZ19_2392	Flap endonuclease-like protein	5'-3' exonuclease, alpha-helical arch, n-terminal.	5'-3' exonuclease, n-terminal resolvase-like domain.	5'-nuclease	100
BZ19_2531	Proline aminopeptidase P II	Peptidase M24, methionine aminopeptidase.	Aminopeptidase P, N-terminal domain.	Creatinase/methionine aminopeptidase superfamily	100
BZ19_2555	Holliday junction resolvase-like protein	Putative pre-16S rrna nuclease.	Holliday junction resolvase	Ribonuclease H-like	100
BZ19_2572	Murein transglycosylase C	Murein transglycosylase-C Lysozyme-like domain superfamily.	Transglycosylase SLT domain	Glycosidases	100
BZ19_2582	Superfamily I DNA and RNA helicases.	DNA helicase, uvrd/REP type.	Uvrd-like helicase C-terminal domain	P-loop containing nucleotide triphosphate hydrolases	100
BZ19_2616	Serine Recombinase family, catalytic domain.	DNA-binding recombinase domain	Resolvase, N terminal domain.	Arylamine N-acetyltransferase	100
BZ19_2625	Predicted gtpase	P-loop containing nucleoside triphosphate hydrolase.	50S ribosome-binding gtpase	P-loop containing nucleotide triphosphate hydrolases	100
BZ19_2935	D-galactarate dehydratase	D-galactarate/Altronate dehydratase, C-terminal	SAF domain. D-galactarate dehydratase	Phenylalanine Hydroxylase	100
BZ19_3217	Glycoside hydrolases family 4	Glycoside hydrolase, family 4.	family 4 glycosyl hydrofase.	L-2-hydroxyisocaproate dehydrogenase.	100

Table 2: Continued....

BZ19_3339	Phosphoethanolamine transferase	Phosphoethanolamine transferase.	sulfatase	Prolyl oligopeptidase	100
BZ19_3533	Predicted transposase	Recombination-promoting nuclease	Putative transposase.	Sulfate adenylyltransferase	100
BZ19_3909	Isovaleryl coa dehydrogenase	Acyl-coa dehydrogenase	Acyl-coa dehydrogenase	Butyryl-coa Dehydrogenase	100
BZ19_3972	Collagenase	Thiamin phosphate synthase superfamily	peptidase family u32	glycosidases	100
BZ19_4009	Multifunctional aminopeptidase A	Leucine aminopeptidase	Cytosol aminopeptidase family	Zn peptidases	100
BZ19_4017	DEAD-like helicases superfamily-	P-loop containing nucleoside triphosphate hydrolase.	Helicase conserved C-terminal domain	Pectin lyase-like	100
BZ19_4055	Autoinducer 2 aldolase	3-hydroxy-5-phosphonooxypentane-2,4-dione thiolase.	Deoc/lacd family aldolase	Aldolase class i	100

Table 3: Protein structure prediction of the hypothetical protein encoding enzyme domains in *Y. pseudotuberculosis* by using templates from RCSB PDB

Sr.No.	Template	Seq-len	Aligned (%)	Identity (%)	Bit-score	E-value	Template name
BZ19_241	2i46A	152	87.85	17.61	120.3	0.014	Human TPP1
BZ19_279	1j0hA	588	98.69	30.41	624.7	1.10E-28	Neopullulanase
BZ19_435	3e18A	348	98.67	21.16	288.7	5.90E-10	Nad-binding protein
BZ19_632	3bvcA	203	63.51	13.71	145.6	0.055	Uncharacterized protein lsm_01780
BZ19_676	1e4cP	206	86.01	25.47	380.7	4.50E-15	L-Fuculose 1-Phosphate Aldolase
BZ19_1178	2r6aA	420	93	27.92	510.1	2.80E-22	BH1
BZ19_1294	1wy2A	351	96.79	23.66	474.9	2.50E-20	Prolidase
BZ19_1436	1nktA	836	98.91	11.76	126.4	0.66	Translocation atpase
BZ19_1618	1jpdX	318	99.07	66.36	516.3	1.30E-22	L-Ala-D/L-Glu Epimerase
BZ19_1645	3cqvA	370	99.46	47.7	1070	0	Unknown protein (SO_1313)
BZ19_1682	1wvfA	515	53.34	14.97	269.5	7.00E-09	p-Cresol Methylhydroxylase
BZ19_1877	1vkdA	327	87.18	31.87	312.9	2.70E-11	Glycosidase
BZ19_2392	1bgxT	828	98.01	29.66	287.6	6.90E-10	Taq polymerase
BZ19_2531	2v3zA	439	99.08	81.76	1023.6	0	Aminopeptidase P
BZ19_2555	1nu0A	131	97.86	72.26	389.8	1.40E-15	Hypothetical protein

Protein structure prediction of Hypothetical proteins

The best score enzyme coding 34 hypothetical protein linked with the structure-function relationship by using PS^2 protein structure prediction server. Here only 15 proteins reported the homology with predetermined three-dimensional structures of template proteins. The details of homology and predicted tertiary structure of hypothetical proteins derived from it showcased in Table 3.

DISCUSSION

The role of bioinformatics certainly increasing in proteomic research especially to explore the hidden potential in many uncharacterized islands of pathogens. The bioinformatics algorithm designed to find out the pattern of a conserved domain in such uncharacterized protein-making protein realistic to link with particular metabolic function. In the present study, *Y. pseudotuberculosis* a human pathogen expressed an enzymatic feature in many hypothetical proteins once detected by sequence and structure-based homology. Here *Y. pseudotuberculosis* encoded 759 hypothetical proteins and 358 predicted to showcase enzymatic coding ability and with the enzyme coding ability, 190 proteins recorded with at least 50% confidence as per server search record. Further, those 34 proteins predicted with 100% enzyme coding ability also confirmed for protein structure-based homology, but only 15 of them showcased existing template matching with their sequences. In a similar manner success stories of many bacterial hypothetical proteins put forward by using the similar approach of bioinformatics once evidenced with *Shigella flexneri* (Gore, 2009); *Bacillus anthracis* (Gore and Raut, 2009), *Haemophilus influenza* (Dogra and Gore, 2010) and *Helicobacter pylori* (Gore *et al.*, 2010). Overall study highlighted that only 15 hypothetical proteins are worth for further investigation in *Y. pseudotuberculosi*

CONCLUSION

The organism *Y. pseudotuberculosis* encodes 759 hypothetical proteins and among them 358 proteins worth to investigate for enzyme function. Further 15 hypothetical proteins are most prominently able to encode enzyme function once bioinformatics evidences confirmed about their conserved domain presence and defined homology with previously known enzymes for structure and sequence. Study showcased the need of investigation on these enzyme coding hypothetical proteins in coming time.

REFERENCES

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic Acids Res.* 25:3389-3402.

An D., Apidianakis Y., Boechat A.L., Baldini R.L., Goumnerov B.C. and Rahme L.G. (2009). The pathogenic properties of a novel and conserved gene product, KerV, in proteobacteria.. *PLoS One.* 4(9):e7167.

Chih-Chieh C., Jenn-Kang H., and Jinn-Moon Y. (2006). (PS)2: protein structure prediction server Nucleic Acids Res.

Dogra P. and Gore D. (2010). Prediction of Enzymatic Function and Structure of H. influenzae Hypothetical Proteins - An In silico Approach. *Int. J. Soft Computing Bioinformatics* .1:67-77.

Garbom S., Forsberg A., Wolf-Watz H. and Kihlberg B.M. (2004). Identification of novel virulence-associated genes via genome analysis of hypothetical genes. *Infect Immun.* 72:1333-1340.

Gore D. (2009). In silico Prediction of Structure and Enzymatic Activity for Hypothetical Proteins of *Shigella flexneri. Biofrontiers.* 1: Pg 1-10.

Gore D. and Raut A. (2009). Computational Function and Structural Annotations for Hypothetical proteins of Bacillus anthracis. *Biofrontiers.* 1:27-36.

Gore D., Denge P. and Amrute M. (2010). Homology Modeling and Enzyme Function Prediction in the Hypothetical Proteins of Helicobacter pylori - an Insilico Approach. *Biomirror.*1-5/ bm-1111251610.

Heine W., Beckstette M., Heroven A.K., Thiemann S., Heise U., Nuss A.M., Pisano F., Strowig T. and Dersch P. (2018). Loss of CNFY toxin-induced inflammation drives Yersinia pseudotuberculosis into persistency. *PLoS Pathog.* 14(2):e1006858.

Hoffman J.M., Sullivan S., Wu E., Wilson E. and Erickson D.L. (2017). Differential impact of lipopolysaccharide defects caused by loss of RfaH in Yersinia pseudotuberculosis and Yersinia pestis. *Sci. Rep.* 7:10915.

Kim J., Fukuto H.S., Brown D.A., Bliska J.B. and London E. (2018). Effects of host cell sterol composition upon internalization of Yersinia pseudotuberculosis and clustered β1 integrin. *J. Biol. Chem.* 293:1466-1479.

Schaffer A.A., Aravind L., Madden T.L., Shavirin S., Spouge J.L., Wolf Y.I., Koonin E.V. and Altschul S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches With composition-based statistics and other refinements. *Nucleic Acids Res.* 29(14):2994-3005.

Schrimpe-Rutledge A.C., Jones M.B., Chauhan S., Purvine S.O., Sanford J.A., Monroe M.E., Brewer H.M., Payne S.H., Ansong C., Frank B.C., Smith R.D., Peterson S.N., Motin V.L., and Adkins J.N. (2012). Comparative omics-driven genome annotation refinement: application across Yersiniae. *PLoS One.* 7: e33903.

Seecharran T., Kalin-Manttari L., Koskela K., Nikkari S., Dickins B., Corander J., Skurnik M., and McNally A. (2017). Phylogeographic separation and formation of sexually discrete lineages in a global population of Yersinia pseudotuberculosis. *Microb. Genom.* 3(10):e000133.

Willcocks S.J., Stabler R.A., Atkins H.S., Oyston P.F., and Wren B.W. (2018). High-throughput analysis of Yersinia pseudotuberculosis gene essentiality in optimised in vitro conditions, and implications for the speciation of Yersinia pestis. *BMC Microbiol.* 18:46.

