Research Article

Exploring the function in the hypothetical junction of *Yersinia enterocolitica* subsp. *palearctica Y11* - A bioinformatics approach

Mayur Dange[1], Archana Umate[1], Dilip Gore[2]*, Maithili Hedaoo[1], Shreya Fadnavis[1], M A Soni[1], Anup D Chahande[1]

ABSTRACT

Aim: *Yersinia enterocolitica* subsp. *palearctica Y11* is a bacterium known for spreading human gastroenteritis and in relation study investigated functional features in the hypothetical proteins encoded by *Y. enterocolitica* subsp. *palearctica Y11* by conserved domain search and protein structure prediction. **Materials and Methods:** Five conserved domain analysis web tools used to perform the search as conserved domain database BLAST, INTERPROSCAN, Pfam, Uniprot, and CATH along with PS2 protein structure prediction server for structure modeling. **Results and Conclusion:** These web tools successfully predicted the function in 13 hypothetical proteins. Overall, results of *Y. enterocolitica* subsp. *palearctica Y11* hypothetical proteins were documented for coding enzymes and other functions. Further, investigation is suggested to understand the role of these hypothetical proteins in metabolic pathway which will be useful for organism bio-controlling.

KEY WORDS: Hypothetical proteins, bioinformatics, conserved domains, protein function

INTRODUCTION

Yersinia enterocolitica identified as enteropathogenic bacterium bringing human gastroenteritis with featured symptoms such as diarrhea, abdominal pain, fever, and vomiting.[1] Even though the symptoms are self-limiting, in few cases, sever clinical features are reported that lead to pseudoappendicular syndromes or septicemia in old age patients with immune compromised situation.[2-4]

The major transmission for enteric yersiniosis is reported through fecal–oral route[5,6] and species *Y. enterocolitica* has been recognized with six subtypes[7] except biotype 2A all five (1B, 2-5) reported to be highly pathogenic to human or animals. Moreover, among them, biotype is more common to infect human and relates with the serotype O:3 (4/O:3) and 2/O:9 which is very common in pigs as a reservoir, especially the 4/O:3 strains.[8] Pig acts as a carrier for the pathogen in their tonsils and intestinal tract and spread in the environment through stools or by contaminated pork meat.[9] Hence, it is indicated that pork meat and pig are the main source of the transmission of the pathogen to the human and defined surveillance has been suggested for the given pathogen.

Access this article online

Website: jprsolutions.info	ISSN: 0974-6943

In today's era with the ever-increasing genome sequencing rate, numerous genome sequencing is going on, but the quality of their gene annotation and function linking still remains the real problem. As a result, a number of open reading frames still remain clueless about their function and named as conserved hypothetical proteins and most regions (20–50%) of genome represents the same.[10] In a view, now, assigning functions to the hypothetical proteins have been possible using the biological databases and orthologous searches using comparative genomics approach. Many of these "hypothetical proteins" occur in fact in more than one bacterial species, and can thus be combined into orthologous groups; this subset of proteins contains the majority of biologically relevant sequences (less likely to be artifactual), and it is amenable to analysis by comparative genomics techniques.[11-14]

[1]Department of Biotechnology, Priyadarshini Institute of Engineering and Technology, Mouza Shivangaon, Behind CRPF Campus, Hingna Road, Nagpur – 440 019, Maharashtra, India, [2]Sai Biosystems Private Limited, Plot No. 271 Raghuji Nagar, Nagpur – 440 009, Maharashtra, India

***Corresponding Author:** Dilip Gore, Sai Biosystems Private Limited, Plot No. 271 Raghuji Nagar, Nagpur – 440 019, Maharashtra, India. E-mail: saibiosystems@gmail.com

Received on: 23-01-2017; Revised on: 27-02-2018; Accepted on: 28-03-2018

In the present study, an attempt has been made to understand the role of hypothetical proteins encoded by the *Yersinia enterocolitica* subsp. *palearctica Y11* by involving the bioinformatics approach in which the sequence homology and structural closeness with the related family protein members were ascertained, and probable role of these proteins has been predicted as per homology concept used earlier.

MATERIALS AND METHODS

Data Collection

The hypothetical protein sequences for *Y. enterocolitica* subsp. *palearctica Y11* from KEGG using limit search as "hypothetical protein" were recovered. In response, a total of 698 hypothetical proteins were retrieved from the genome for functional annotation by conserved domain search.

Conserved Domain Search

To search for function in the hypothetical proteins, five conserved domain analysis software was used:

a. Conserved domain database (CDD)-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/)[15-17]
b. INTERPROSCAN (http://www.abi.ac.uk/interpro)[18]
c. PFAM (http://www.pfam.sanger.ac.uk/)[19]
d. CATH (http://www.cathdb.info/[20]
e. UNIPROT: http://www.uniprot.org/.[21]

These servers are linked with number of CDD and linked matrix been able to find the pattern-based match for query sequence with known family protein and ultimately any conserved domain if present could be reported.

Confidence Level

The parameter of confidence limit set as 100%, 80%, 60%, 40%, 20%, and 0% considering the following rules which does indicate the match of each server with another server for a given query sequence for its function. If the given five tools indicate the same functions, then the confidence level was to be 100%; four tools 80%; three tools 60%; two tools 40%; and one tool as 20%, and the scoring of each protein was made accordingly for better function prediction.

Protein Structure Prediction of Hypothetical Proteins

The protein structure prediction of hypothetical proteins was predicted by PS square-protein structure prediction server. $(PS)^2$ (Pronounced PS square): $(PS)^2$ is an automated homology modeling server. The method uses an effective consensus strategy by combining PSI-BLAST, IMPALA, and T-Coffee in both template selection and target-template alignment. The final three-dimensional structure is built using the modeling package MODELLER. The web address is http://www.ps2.life.nctu.edu.tw/.[22]

OBSERVATION AND RESULTS

Result

Functional annotation to the hypothetical proteins

As per conserved domain search carried out for the 698 hypothetical proteins of *Y. enterocolitica* subsp. *palearctica Y11* with the five servers described earlier, results showcased that 19 hypothetical proteins possess enzymatic conserved domain along with some other functions with variable confidence limits as high as 100% to lowest 60%. The confidence limits of proteins those were grouped are as follows: 100% for 13 proteins, 80% for 04 proteins, 60% for 01 protein, and exact function predicted with these proteins is shown in Table 1.

Protein Structure Prediction

All 100% confidence gaining hypothetical proteins ($n = 13$) were used for the protein structure prediction possessing the respective conserved domains. The prediction server has generated the predicted structures for 12 proteins successfully. The protein structure prediction of hypothetical proteins was detailed by the template information by which the hypothetical protein sequence has shown the best homology along with related information of percent, identity, score, and E-value as shown in Table 2 and Figure 1.

DISCUSSION

In the present study, speed of the genome sequencing and parallel use of the bioinformatics has been used effectively to determine the function in the hypothetical proteins marked on the genome of *Y. enterocolitica* subsp. *palearctica Y11*. Its genome represented to code more than 15% of the loci with the hypothetical proteins and in the current study when investigated by the five conserved domain searching servers, it has been put forward that only 13 proteins in the total hypothetical proteins ($n = 698$) able to actually possess the conserved domain as per homology search again with 100% confidence level. Of these 13 proteins, 12 proteins were further able to modeled themselves for the three-dimensional structures of tertiary type as per modeling program; hence, it does indicate that many other organisms possess the homolog protein and must be functioning in metabolism as number of workers remain interested to surely crystallized these proteins and determined the structures by X-ray crystallography and also justifies our finding for function prediction in few hypothetical proteins. In similar studies, workers reported the same approach and linked many hypothetical proteins of number of bacterial pathogens such as *Bacillus anthracis, Shigella flexneri, Helicobacter pylori,* and *Haemophilus influenza* with their exact function.[23-26] Hence, by involving further study such as gene cloning

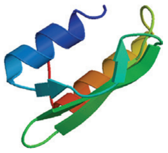
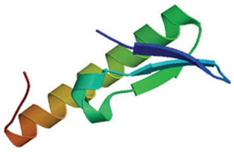
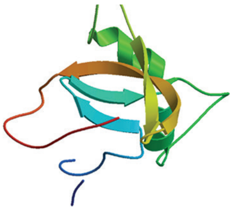
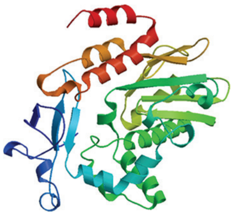
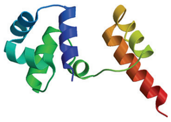
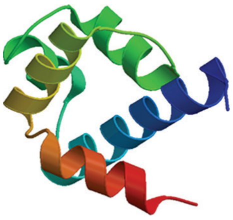
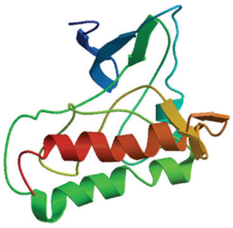
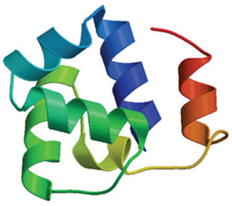
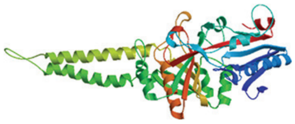
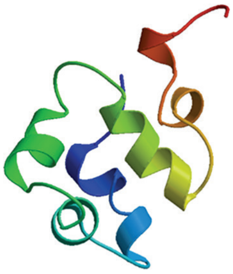
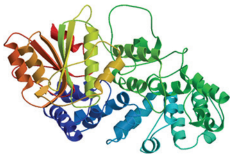
	
Y11_22881	Y11_22891
	
Y11_23541	Y11_24431 Contd...
	
Y11_23541	Y11_24431 Contd...
	
Y11_27051	Y11_27681 contd....
	
Y11_28531	Y11_28651 contd....
	
Y11_29281	Y11_29601 contd...

Figure 1: Tertiary structures predicted for the hypothetical proteins of *Yersinia enterocolitica* subsp. *palearctica Y11- A*

Table 1: Conserved domain recorded in the several hypothetical proteins of *Y. enterocolitica* subsp. *palearctica Y11* with variable confidence percent

KEGG No.	Pfam	CATH	CDD BLAST	InterPro	UniProt	% confidence
22701	FaeA-like protein	FaeA-like protein	FaeA-like protein	FaeA-like protein	Uncharacterized protein	80
22881	HicA toxin of bacterial toxin-antitoxin	Orotate phosphoribosyltransferase	HicA toxin of bacterial toxin-antitoxin	Orotate phosphoribosyltransferase	Orotate phosphoribosyltransferase	100
22891	HicB-like antitoxin of bacterial toxin-antitoxin	HicB-like antitoxin of bacterial toxin-antitoxin	HicB-like antitoxin of bacterial toxin-antitoxin	HicB-like antitoxin of bacterial toxin-antitoxin	Uncharacterized protein	100
23541	Putative helicase	Putative helicase	Integrating conjugative element relaxase,	Putative helicase	Helicase	100
23671	ParD-like antitoxin	None	ParD-like antitoxin of type II	ParD-like antitoxin	None	60
24431	Outer membrane efflux protein	Outer membrane efflux protein	Outer membrane protein	Outer membrane efflux protein	Outer membrane efflux protein	100
25281	Reverse transcriptase	Reverse transcriptase	Reverse transcriptase	Reverse transcriptase	Reverse transcriptase	100
25291	Helix-turn-helix	Helix-turn-helix	Helix-turn-helix	Helix-turn-helix	Helix-turn-helix	100
26791	Glycine-rich SFCGS	None	Glycine-rich SFCGS	Sugar-phosphate isomerase	None	60
26801	PRD domain protein EF_0829/AHA_3910	PRD domain protein EF_0829/AHA_3910	PRD domain protein EF_0829/AHA_3910	PRD domain protein EF_0829/AHA_3910	PRD domain protein EF_0829/AHA_3910	100
27051	Helix-turn-helix	Putative transcriptional regulator	Helix-turn-helix	Helix-turn-helix	Putative transcriptional regulator	100
27681	Serine/threonine protein kinase	Serine/threonine protein kinase	Serine/threonine protein kinase	Serine/threonine protein kinase	Putative type III secreted effector...	100
28371	Papain fold toxin 1, glutamine deamidase	None	Papain fold toxin 1, glutamine deamidase	Tox-PL domain	None	80
28451	Phage P2 GpE	Phage P2 GpE	Phage P2 GpE	Phage P2 GpE	None	80
28531	Helix-turn-helix domain	Helix-turn-helix domain	Helix-turn-helix domain	Helix-turn-helix domain	Predicted transcriptional regulator	100
28651	SEC-C motif	Preprotein translocase subunit SecA	SEC-C motif	SEC-C motif	Preprotein translocase subunit SecA	100
29031	Transposase domain (DUF772)	Transposase domain (DUF772)	Transposase domain (DUF772)	Transposase domain (DUF772)	None	80

(Contd...)

Table 1: ***(Continued...)***

KEGG No.	Pfam	CATH	CDD BLAST	InterPro	UniProt	% confidence
29281	Winged helix-turn-helix DNA-binding	DNA-binding transcriptional regulator Nlp	Winged helix-turn-helix DNA-binding	DNA-binding transcriptional regulator Nlp	Putative sugar fermentation stimula...	100
29601	VWA domain containing CoxE-like protein	Protein ViaA	hypothetical protein	VWA domain containing CoxE-like protein	Protein ViaA	100

CDD: Conserved domain database, *Y. enterocolitica: Yersinia enterocolitica*

Table 2: Protein structure prediction of selected hypothetical proteins of *Y. enterocolitica* subsp. *palearctica Y11* using best-scored templates of RCSB PDB

KEGG NO.	Template	Seq-len	Aligned (%)	Identity (%)	Bit-score	E-value
Y11_22881	1whzA	70	96.97	25.76	180.3	0.00065
Y11_22891	2dsyD	81	42.22	21.05	127.3	0.58
Y11_23541	2ipqX	121	28.29	31.4	240.8	0.00000028
Y11_24431	1ek9A	428	55.49	16.26	186.9	0.00028
Y11_25281	1xr6A	460	94.52	16.67	141	0.1
Y11_25291	1b0nA	103	69.93	25.24	153.1	0.021
Y11_26801	No significant templates could be found					
Y11_27051	2b5aA	77	81.52	16.88	123.4	0.96
Y11_27681	1zarA	267	18.01	15.73	145.8	0.054
Y11_28531	1b0nA	103	89.01	18.82	162.4	0.0065
Y11_28651	1lrzA	400	80.98	10.63	119.3	1.6
Y11_29281	1neqA	74	83.95	55.88	285.1	9.4E-10
Y11_29601	1yvrA	520	98.98	15.04	133.1	0.28

Y. enterocolitica: Yersinia enterocolitica and expression in laboratory conditions, we could decipher functions in promising hypothetical proteins of *Y. enterocolitica* subsp. *palearctica Y11* and may link them with metabolic pathways.

CONCLUSION

The present study highlighted the 13 hypothetical proteins containing the conserved domains as per homology search carried out by the bioinformatics approach. Study concluded that available web tools linked with rich protein databases information proved to be beneficial and resultant filtered out some of the important enzyme and other function coding hypothetical proteins out of the whole pool of *Y. enterocolitica* subsp. *palearctica Y11* proteomics. In future, obtained data may assist in relational linking of today's hypothetical proteins as tomorrow's functional counterpart probably by involving gene cloning, microarray, and marker-based experiments.

REFERENCES

1. Rosner BM, Werber D, Hoehle M, Stark K. Clinical aspects and self-reported symptoms of sequelae of *Yersinia enterocolitica* infections in a population-based study, Germany 2009-2010. BMC Infect Dis 2013;13:236.
2. Perdikogianni C, Galanakis E, Michalakis M, Giannoussi E, Maraki S, Tselentis Y, *et al*. *Yersinia enterocolitica* infection mimicking surgical conditions. Pediatr Surg Int 2006;22:589-92.
3. Savin C, Carniel E. Diarrhea of bacterial origin: The case of *Yersinia enterocolitica*. Rev Francoph Lab 2008;400:49-58.
4. Bottone EJ. *Yersinia enterocolitica*: The charisma continues. Clin Microbiol Rev 1997;10:257-76.
5. Fredriksson-Ahomaa M, Korkeala H. Molecular epidemiology of *Yersinia enterocolitica* 4/O:3. Adv Exp Med Biol 2003;529:295-302.
6. Nuorti JP, Niskanen T, Hallanvuo S, Mikkola J, Kela E, Hatakka M, *et al*. A widespread outbreak of *Yersinia pseudotuberculosis* O:3 infection from iceberg lettuce. J Infect Dis 2004;189:766-74.
7. Wauters G, Kandolo K, Janssens M. Revised biogrouping scheme of *Yersinia enterocolitica*. Contrib Microbiol Immunol 1987;9:14-21.
8. Drummond N, Murphy BP, Ringwood T, Prentice MB, Buckley JF, Fanning S. *Yersinia enterocolitica*: A brief review of the issues relating to the zoonotic pathogen, public health challenges, and the pork production chain. Foodborne Pathog Dis 2012;9:179-89.
9. Kapperud G. *Yersinia enterocolitica* in food hygiene. Int J Food Microbiol 1991;12:53-65.
10. Roberts RJ. Identifying protein function-a call for community action. PLoS Biol 2004;2:E42.
11. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci 1998;23:324-8.
12. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86-90.
13. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions. Science 1999;285:751-3.
14. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999;96:2896-01.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-402.
16. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, *et al*. Improving the accuracy of PSI-BLAST protein database searches with composition based statistics and other refinements. Nucleic Acids Res 2001;29:2994-3005.
17. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, *et al.* CDD: A conserved domain database for interactive domain family analysis. Nucleic Acids Res 2007;35:D237-40.
18. Zdobnov EM, Rolf A. Interproscan-an integration platform for the signatures recognition methods in Interpro. Bioinformatics 2001;17:847-8.
19. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, *et al.* The pfam protein families database. Nucleic Acids Res 2002;30:276-80.
20. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, *et al.* New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 2013;41:D490-8.
21. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, *et al.* UniProt: The universal protein knowledgebase. Nucleic Acids Res 2004;32:D115-9.
22. Chen CC, Hwang JK, Yang JM. (PS)2-v2: Template-based protein structure prediction server. BMC Bioinf 2009;10:366.
23. Dogra P, Gore D. Prediction of enzymatic function and structure of H. Influenzae hypothetical proteins-an *in silico* approach. Int J Soft Comput Bioinf 2010;1:77-87.
24. Gore DG, Denge AP, Amrute NM. Homology modeling and enzyme function prediction in the hypothetical proteins of *Helicobacter pylori* - an *in silico* approach. BIOMIRROR 2010;1:1-5.
25. Gore D. *In silico* prediction of structure and enzymatic activity for hypothetical proteins of *Shigella flexneri*. Biofrontiers 2009;1:1-10.
26. Gore D, Raut A. Computational function and structural annotations for hypothetical proteins of *Bacillus anthracis*. Biofrontiers 2009;1:27-36.