Biochemistry & Pharmacology: Open Access

Bidkar et al., Biochem Pharmacol 2014, 3:3
DOI: 10.4172/2167-0501.1000136

Research Article Open Access

In-silico Structural and Functional Analysis of Hypothetical Proteins of *Leptospira Interrogans*

Anil P Bidkar[1]*, Krishan K Thakur[1], Nityanand B Bolshette[1], Jyotibon Dutta[1] and Ranadeep Gogoi[2]

[1]*Laboratory of Biotechnology, Department of Biotechnology, National Institute of Pharmaceutical Education and Research (NIPER), Guwahati Medical College, Guwahati-781032, Assam, India*

[2]*Department of Biotechnology & Bioengineering, Institute of Science and Technology, Guwahati University, Guwahati-781014, Assam, India*

Abstract

Though after a start of genome sequencing most of the protein sequences are deposited in databases, some proteins remain to be annotated and functionally characterized. Analyzing and annotating the function of Hypothetical proteins of Leptospira interrogans is very important which is a pathogenic spirochete causes various complications in human and animals. Randomly we have selected 12 sequences of hypothetical proteins and were analyzed through web tools as Pfam, CD Blast, Expasy's to determine physicochemical properties and protein family information based on conserved domains. Proteins from families AdoMetDC, LRR, and PilZ are most conserved in many microorganisms can be targeted to develop efficacious drug molecules. Ligand binding site and protein structure prediction was done by using Qsite finder and PS2 server which will help to develops therapeutic molecules in docking studies. Present study has shown intracellular interactions of proteins involved in drug resistance, transcription factors, and transport channels.

Keywords: Hypothetical protein; *Leptospira interrogans*; Bioinformatics tools; Functional analysis

Introduction

Genome sequencing projects and genetic engineering has revealed many aspects of complex cellular environment containing large number of proteins. Despite sequences of most of organisms are available and proteins coded are studied experimentally, there are some proteins whose functions are unknown, need to be characterised [1]. Such proteins are known as Hypothetical Proteins (HP) sequences [2] of which are known but there is no evidence of experimental study [3]. There is extensive need to study and classify these hypothetical proteins which can open new way to design drug molecules against infectious organisms. Functional annotation of HPs involved in infection, drug resistance, and essential biosynthetic pathways is important for development of the potent antibacterial against infectious agents. Improved understanding of these proteins may make them potential targets of antimicrobial drugs [4]. *Leptospira interrogans* is gram negative spirochete, having an internal flagella is pathogenic which causes Leptospirosis [5-7], other serovars (strains) are distinguished on the basis of cell surface antigens. These are infectious to animals, but through animal urine can be spread to human [8]. Leptospira enters in body via broken skin, mucosa and spreads in body, if immune system fails to stop the growth of bacteria it cause severe hepatic and renal dysfunctions [9,10]. This present study highlights the *in silico* studies to characterize HPs from *Leptospira interrogans*.

Methods

Sequence retrieval

KEGG (Kyoto Encyclopedia of Gene and Genomes) is a large collection of databases having entries of genes, proteins, pathways in metabolism and diseases, drug and ligands of organism [11]. We have selected the Sequences of 12 hypothetical proteins of *Leptospira interrogans* randomly from KEGG database (www.genome.jpg/kegg).Gene IDs for selected proteins are gi|24214908, gi|24215649, gi|24215664, gi|24215909, gi|24216444, gi|24217373, gi|24213620, gi|24214752, gi|24214753, gi|294827583, gi|294827687, gi|24213945.

Pfam

Pfam is curated Protein families database, it uses jackhmmer programme (HMMR3). To give profile HMM (Hiden Markov Model) with PSI-BLAST, which were searched against UniProt [12]. However, to include protein in a family its domain and sequence bit scores must be equal or above the Gathering Thresholds (GA). Pfam gives Pfam A families which are manually curated and Pfam B families generated automatically [13].

Batch CD search

Hypothetical Protein sequences were searched for conserved domains at batch CD search, which gives results by using MSA and 3D structures for homologous domains available on Pfam and SMART [12,14].

ExPASy-ProtParam tool

ProtParam tool (www.expasy.org/tools/protparam.html) was used to estimate physicochemical parameters of hypothetical proteins [15]. Query protein can be submitted in form of SWISS/TrEMBL ID or protein sequence. Server provides directly calculated values of pI/MW (Isoelectric point, Molecular Weight), Percentage of each amino acid, Extinction Coefficient (EC), Instability Index (II) [16], Aliphatic Index (AI) and GRAVY (Grand Average of Hydrophobicity).

***Corresponding author:** Anil P Bidkar, Department of Biotechnology, National Institute of Pharmaceutical Education and Research (NIPER), Gauhati Medical College, Guwahati-781032, Assam, India, Tel: +91-970-678-9537; E-mail: anilbidkar1@gmail.com, nbolshette@gmail.com

Received March 29, 2014; **Accepted** April 25, 2014; **Published** April 30, 2014

Citation: Bidkar AP, Thakur KK, Bolshette NB, Dutta J, Gogoi R (2014) In-silico Structural and Functional Analysis of Hypothetical Proteins of *Leptospira Interrogans*. Biochem Pharmacol 3: 136. doi:10.4172/2167-0501.1000136

SOSUI server

Amphiphilicity index and Hydropathy index of query protein sequences were calculated by SOSUI server which categorises protein into cytoplasmic or transmembrane nature [17].

Protein-Protein Interaction network

Protein in the cell environment interacts with other proteins, *in silico* these interactions were studied by STRING v9.1 (Search Tool for Retrieval of Interacting Genes). STRING is a large repository of protein-protein interactions involving functional interactions, stable complexes, and regulatory interactions among proteins [18,19]. Figure 1 Shows resulting protein-protein interaction network of selected hypothetical proteins, for better understanding interaction networks should be seen on server site.

Disulfide-Bonding in protein

Disulfide bonds among cysteine residues in protein plays an important role in folding it into functional and stable conformation. DISULFIND server utilizes SVM binary server to predict bonding state of cysteins, these cysteins are paired by Recursive Neural Network to show disulfide bridges [18]. Information about disulphide bonding helps in experimental structure determination and defining stability of the proteins.

Protein structure prediction

Protein structure prediction server $(PS)^2$ [20] requires query sequence in fasta format to generate 3D structure by comparative modelling. Server utilizes consensus strategy to find template using PSI-BLAST and IMPALA. Query sequence and template aligned by T-coffee, PSI-BLAST, and IMPALA [17]. 3D structures are predicted from template using MODELLER and visualised by CHIME, Raster3D. Resulting 3D structural model of selected hypothetical proteins are shown in Figure 2.

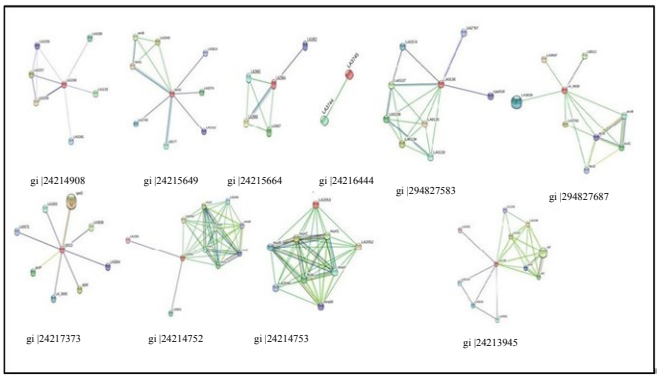

Figure 1: Showing Protein-protein Interaction of hypothetical proteins generated by STRING tool

Ligand binding site prediction

Q-site finder server was used for binding site prediction in selected proteins [21]. Server uses energy based methods to find clefts on protein surface for ligands [22]. These hot spots for ligand binding have predicted after ranking their physicochemical properties as hydrophobicity, desolvation, electrostatic & van der waal potentials.

Discussion

ProtParam tool computes different physicochemical parameters depending on the queries submitted to the databases. Isoelectric focusing separates proteins according to pI where pH gradients are developed [23]. Predicted pI via server may not be adequate because

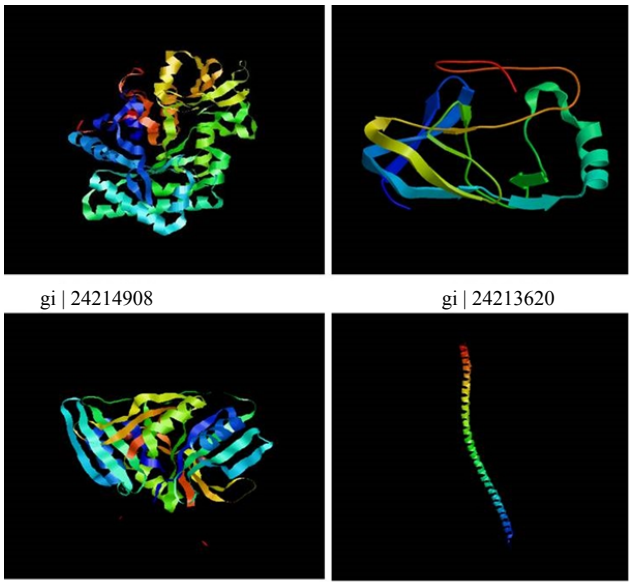

Figure 2: Predicted structures of hypothetical proteins of Leptospira interrogans by PS2 server

Sequence ID	No of AA	MW	pI	EC	II	AI	GRAVY
gi| 24214908	535	61891	8.27	100980	42.45	69.66	-0.761
gi| 24215649	445	50918	9.20	43445	39.23	79.26	-0.396
gi| 24215664	426	49102	8.7	13535	60.67	124.98	-0.460
gi| 24215909	251	26175.9	5.24	17795	28.55	65.58	-0.161
gi| 24216444	241	28062	8.74	11920	47.89	93.44	-0.297
gi| 24217373	288	32486.7	5.49	26275	36.72	83.30	-0.267
gi| 24213620	428	50351.8	5.13	31290	45.52	106.73	-0.065
gi| 24214752	399	46207.3	6.71	112550	34.23	121.93	0.631
gi| 24214753	182	20886.2	8.6	19285	37.95	95.33	-0.276
gi| 294827583	309	33712.5	8.68	11710	36.34	88.61	-0.358
gi| 294827687	199	22228.3	9.62	13075	33.11	116.03	0.323
gi|24213945	212	23671.6	3.90	1490	51.06	90.71	-0.743

Table 1: Physicochemical properties of proteins using Protparam tool.

in case of high number of basic amino acids and lower buffer capacity. By using pH gradients and calculated pI, proteins can be separated experimentally. MW of proteins along with pI is used for the 2D gel electrophoresis. EC shows a light absorbed by a protein relative to their composition at a specific wavelength. EC given (Table 1) are calculated with reference to Tryptophan, Cysteine, Tyrosine [17]. Instability Index (II) refers to the stability of the protein in test tube [24]. Among studied proteins gi|24214908, gi|24215664, gi|24216444, gi|24213620, gi|24213945 were found to be unstable, and rest are stable (proteins with II above 40 are unstable). Aliphatic amino acid constitutes the aliphatic index (a relative volume of aliphatic side chains). Increased AI results into a hydrophobic interactions and thus gives thermostatic stability to protein, predicted AI and II shows inverse relation for stability except these two proteins gi|24215664 andgi|24215909. GRAVY [26] values are a ratio of all hydropathy values of amino acids to the number of residues in sequence. Smaller the GRAVY [25] more hydrophilic is protein, gi|24214908 and gi|24213945 proteins found the most hydrophilic. In case of 3D structure hydrophilic domains tends to be on exterior surface, while hydrophobic domains avoids external environment and forms internal core of the protein. Search of family for hypothetical proteins based on conserved domains having consensus sequence in their structure is given in Table 2. Hypothetical protein gi|24214908 found to be a member of GH18_CFLE_spore_hydrolase, Cortical fragment Lytic Enzyme bearing a catalytic domain from glycosylhydrolase, an enzyme used in breaking spore peptidoglycans so as to activate it for germination when favourable conditions are available. Hypothetical protein gi|24215649 from PDZ_serine_ protease involved in protein reassembly and work as a heat shock protein. Protein gi|24215664 belongs to Leucine-Rich Repeats (LRR), ribonuclease inhibitor like family. LRR are motifs having role in protein interactions in complex networks (Table 3). S-adenosylmethionine decarboxylase (AdoMetDC) enzyme for biosynthesis of spermine and spermidine by decarboxylation of SAM belongs to Ado_Met_dc family (gi|24217373). Pilz domain in gi|24213620 is found in bacterial cellulose synthase and other proteins that forms biofilm around a bacterium and involve in effluxing drug [26]. Hypothetical protein gi|294827583 (FecR superfamily) is involved in Iron transport system in bacterial membranes, Fe^{3+} (insoluble) loaded on citrate carrier is sensed by FecR protein found in periplasmic space in bacterial membrane [27]. Protein sites are predicted as cytoplasmic, host associated, extracellular, cytoplasmic membrane proteins (Table 4). SOSOI server predictions (Table 5) shows that positively charged amino acids are more at the end of transmembrane region. Analysis by DISULFIND server (Table 6) shows the Cysteine residues involved in disulphide

Sequence ID	Family
gi| 24214908	GH18_CFLE_spore_hydrolase,GH18_chitinase-like superfamily
gi| 24215649	PDZ_serine_protease
gi| 24215664	LRR_RI superfamily
gi| 24216444	Macro superfamily
gi| 24217373	AdoMet_dc superfamily, SET superfamily
gi| 24213620	DUF1577 superfamily, PilZ superfamily
gi| 24214752	DUF2157 superfamily
gi| 24214753	GDYXXLXY superfamily
gi| 294827583	FecR superfamily
gi| 294827687	DUF2179 superfamily
gi|24213945	DUF904 superfamily

Table 2: Family distributions of HP's by Batch CD-BLAST analysis.

Protein	Site Volume (cubic A^0)	Amino acid involved
gi| 24214908	548	TYR 6, TRP 10, PHE 37, ALA 38, GLY 39,THR 48, GLY 76, GLY 77, TRP 78, LYS 79, PHE 80, ASP 117, GLN 119, TYR 120, THR 162, ALA 164, GLY 165, ILE 166, MET 189, TYR 191, ASP192, LEU 193.
gi| 24213620	465	PRO 34, ALA 35, LYS 134, LYS 135, GLY 136, TRP 137, GLY 138, ASP 175, SER 176, ASN 177, TYR 178, ARG 206, PHE 207, LEU 208, ASN 209, HIS 210, SER 211.
gi| 24214753	201	ARG 9, VAL 10, GLY 11, ARG 15, ASP 64, LYS 65, TYR 66, GLU 67, LEU 68, PRO 116, ASP 117, GLY 118, TYR 119.
gi|24213945	73	ASN 42, ASP 43, GLN 44, LYS 46, LEU 47, GLU 50

Table 3: Prediction of Amino acids involved in Ligand binding (Q-site finder).

Sequence ID	Protein Localization
gi| 24214908	Unknown
gi| 24215649	Extracellular
gi| 24215664	Cytoplasmic
gi| 24216444	Cytoplasmic
gi| 24217373	Cytoplasmic
gi| 24213620	Cytoplasmic Membrane
gi| 24214752	Cytoplasmic
gi| 24214753	Unknown
gi| 294827583	Cytoplasmic Membrane
gi| 294827687	Cytoplasmic
gi|24213945	Unknown

Table 4: Protein localisation by PSORTB.

Gene ID	N- Terminal	Trans membrane region	C-Terminal	Type	Length
gi| 24215649	60	FYFFKIFVLVLCSLTISVPLKAE	82	PRIMARY	23
	130	AIVADQDFLLALLLPGEFPLFS	109	SECONDARY	22
gi| 24215664	71	LMILLILICFSCKLQAQ	87	PRIMARY	17
gi| 24214752	94	YYSFLILGVVIIGIGVLAIIAAN	116	PRIMARY	23
	121	DDLVKLGFSFIVLTFVAGLSFW	142	PRIMARY	22
	149	LFTVFIVLYSILILGMIGLISQV	171	PRIMARY	23
	188	LSCLFLITTDSKTFLHLWILGFQ	210	PRIMARY	23
	303	FSFPWVIMICRLLILTPIFYLLI	325	PRIMARY	23
gi| 24214753	62	KLILPIALVFPILFFVSEIITLE	84	PRIMARY	23
gi| 294827583	61	YLSIVILCTFAMLLLVC	77	PRIMARY	17
gi| 294827687	73	VLLPCFIFLSRVTDVSIGTIRVI	95	PRIMARY	23
gi| 294827687	103	GIAASLGFLEVLLWVVVITQVIK	125	PRIMARY	23
gi| 294827687	138	GGFATGTFIGMILEEKLAIGFSL	160	SECONDARY	23

Table 5: SOSOI results for proteins.

Sequence ID	Bonded Cysteines
gi| 24215909	55-210, 192-217, 200-236
gi| 24214753	69-107, 87-118, 89-126

Table 6: Cysteine residues involved in disulphide bonding.

bonding of hypothetical proteins. Protein-protein interaction study has shown some hypothetical proteins are involved in essential cellular process such as transport across membrane, biosynthesis of molecules, translational regulation. Hypothetical protein gi| 24214908 (Figure 1) interacts with SUA5 protein which is known as one of translational regulator from YrdC/SUA5 family. Search for gi| 24215909 shown to be involved in chloride transport with chloride channel protein (EriC gene). Protein gi| 24217373 found to be interacted with S-layer like protein (slpM) which forms layer around bacteria to attach other surfaces and protect it from environment. Additionally it involves in cell dividing processes and transport across membrane. Protein gi| 294827687 had shown interaction with proteins for bleomycin resistance, chorismatesynthase (Trp biosynthesis) and Mammalian Cell Entry (MCE) like proteins. Figure 2 shows 3D structures of proteins gi| 24214908, gi| 24213620, gi| 24214753, gi|24213945 predicted from amino acid sequence on PS^2 server by using templates 1vf8A, 3bo5A, 1f9zA, and c2efsA respectively.

Conclusion

Development of potential bioinformatics tools and databases has opened new platform for in-silico study. Currently it is very needful to annotate and characterize hypothetical proteins in *Leptospira interrogans serovar*. These hypothetical proteins may have an imperative role in producing many virulence factors and cause serious infection or disease.We have analyzed 12 hypothetical proteins from KEGG database and categorized its physicochemical properties and recognized domains and families using various bioinformatics tools and databases. The structures were modeled and their ligand binding sites were identified.Physicochemical predictions made for hypothetical proteins, which can be used to find therapeutic agents against infections caused by *Leptospira interrogans*. Some of hypothetical proteins serves as channel proteins, ribosomal proteins or are involved in cell cycle process. Families which were identified for these hypothetical proteins are involved in normal cellular processes and the resistance against drugs. Ligand binding hotspots were found with Q-sitefinder which shown amino acids involved in interaction with ligands. It will help in study of molecular docking for development of potent and effective target against Leptospira infection.

Acknowledgement

This study was supported by NIPER Guwahati academic staff. We are very grateful for their excellent support in every manner.

References

1. Adinarayana KP, Sravani TS, Hareesh C (2011) A database of six eukaryotic hypothetical genes and proteins. Bioinformation 6: 128-130.
2. Shahbaaz M, Hassan MI, Ahmad F (2013) Functional annotation of conserved hypothetical proteins from Haemophilus influenzae Rd KW20. PLoS One 8: e84263.
3. Hsieh WJ, Pan MJ (2004) Identification Leptospira santarosai serovar shermani specific sequences by suppression subtractive hybridization. FEMS Microbiol Lett 235: 117-124.
4. Galperin MY, Koonin EV (1999) Searching for drug targets in microbial genomes. Curr Opin Biotechnol 10: 571-578.
5. Chou LF, Chen YT, Lu CW, Ko YC, Tang CY, et al. (2012) Sequence of Leptospira santarosai serovar Shermani genome and prediction of virulence-associated genes. Gene 511: 364-370.
6. Langston CE, Heuter KJ (2003) Leptospirosis. A re-emerging zoonotic disease. Vet Clin North Am Small Anim Pract 33: 791-807.
7. Tubiana S, Mikulski M, Becam J, Lacassin F, Lefèvre P, et al. (2013) Risk factors and predictors of severe leptospirosis in New Caledonia. PLoS Negl Trop Dis 7: e1991.
8. Kohn B, Steinicke K, Arndt G, Gruber AD, Guerra B, et al. (2010) Pulmonary abnormalities in dogs with leptospirosis. J Vet Intern Med 24: 1277-1282.
9. Picardeau M, Brenot A, Saint Girons I (2001) First evidence for gene replacement in Leptospira spp. Inactivation of L. biflexa flaB results in non-motile mutants deficient in endoflagella. Mol Microbiol 40: 189-199.
10. Ko AI, Goarant C, Picardeau M (2009) Leptospira: the dawn of the molecular genetics era for an emerging zoonotic pathogen. Nat Rev Microbiol 7: 736-747.
11. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277-280.
12. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290-301.
13. http://pfam.sanger.ac.uk/search/
14. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40: D302-305.
15. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, et al. (1999) Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 112: 531-552.
16. Mohan R, Venugopal S (2012) Computational structural and functional analysis of hypothetical proteins of Staphylococcus aureus. Bioinformation 8: 722-728.
17. Mitaku S, Hirokawa T, Tsuji T (2002) Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. Bioinformatics 18: 608-616.
18. Lewis AC, Saeed R, Deane CM (2010) Predicting protein-protein interactions in the context of protein evolution. Mol Biosyst 6: 55-64.
19. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808-815.
20. Chen CC, Hwang JK, Yang JM (2006) (PS)2: protein structure prediction server. Nucleic Acids Res 34: W152-157.
21. Laurie AT, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21: 1908-1916.
22. Burgoyne NJ, Jackson RM (2006) Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. Bioinformatics 22: 1335-1342.
23. Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, et al. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis 14: 1023-1031.
24. Guruprasad K, Reddy BV, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng 4: 155-161.
25. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105-132.
26. Amikam D, Galperin MY (2006) PilZ domain is part of the bacterial c-di-GMP binding protein. Bioinformatics 22: 3-6.
27. Van Hove B, Staudenmaier H, Braun V (1990) Novel two-component transmembrane transcription control: regulation of iron dicitrate transport in Escherichia coli K-12. J Bacteriol 172: 6749-6758.