G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(3):125-135
http://dx.doi.org/10.5808/GI.2016.14.3.125

ORIGINAL ARTICLE

Sequence Analysis of Hypothetical Proteins from *Helicobacter pylori* 26695 to Identify Potential Virulence Factors

Ahmad Abu Turab Naqvi[1§], Farah Anjum[2§], Faez Iqbal Khan[3],
Asimul Islam[1], Faizan Ahmad[1], Md. Imtaiyaz Hassan[1]*

[1]Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India,
[2]Female College of Applied Medical Science, Taif University, Al-Taif 21974, Kingdom of Saudi Arabia,
[3]School of Chemistry and Chemical Engineering, Henan University of Technology, Henan 450001, China

Helicobacter pylori is a Gram-negative bacteria that is responsible for gastritis in human. Its spiral flagellated body helps in locomotion and colonization in the host environment. It is capable of living in the highly acidic environment of the stomach with the help of acid adaptive genes. The genome of *H. pylori* 26695 strain contains 1,555 coding genes that encode 1,445 proteins. Out of these, 340 proteins are characterized as hypothetical proteins (HP). This study involves extensive analysis of the HPs using an established pipeline which comprises various bioinformatics tools and databases to find out probable functions of the HPs and identification of virulence factors. After extensive analysis of all the 340 HPs, we found that 104 HPs are showing characteristic similarities with the proteins with known functions. Thus, on the basis of such similarities, we assigned probable functions to 104 HPs with high confidence and precision. All the predicted HPs contain representative members of diverse functional classes of proteins such as enzymes, transporters, binding proteins, regulatory proteins, proteins involved in cellular processes and other proteins with miscellaneous functions. Therefore, we classified 104 HPs into aforementioned functional groups. During the virulence factors analysis of the HPs, we found 11 HPs are showing significant virulence. The identification of virulence proteins with the help their predicted functions may pave the way for drug target estimation and development of effective drug to counter the activity of that protein.

Keywords: drug discovery, drug target, *Helicobacter pylori*, hypothetical proteins, pathogenesis, virulence

Introduction

Helicobacter pylori is a Gram-negative bacteria that is associated with several gastric problems in human. It is a slow growing microaerophilic bacteria [1]. Its spiral shape flagellated body helps in locomotion and invasion on the host cells. It belongs to the class of bacteria that are responsible for most common bacterial infections in human [2]. It is adapted to the acidic gastric environment for survival. It is also indigenous to the worldwide human population. It was first isolated by Marshall and Warren in 1984 [3-5]. Prolonged infection of the organism can be transformed into a chronic infection that causes severe gastric diseases such as duodenal ulcer, gastric ulcer, gastric lymphonema and cancer [6, 7]. Nonchronic infection of the bacteria is usually asymptomatic. There is usually no development of clinical disease observed in the infected person. The prevalence of infection is also guided by the variations in geographical conditions, age, race, and socioeconomic status of the infected persons [8-10]. A person having bacterial infection at an early age is more prone to develop a chronic infection [11-13]. *H. pylori* infection in developing countries is higher in comparison to the developed countries. The reason behind this may be poor hygiene practices in the developing countries [14].

The *H. pylori* genome was first sequenced in 1997 [5]. The genome of *H. pylori* 26695 strain (NC_000915.1) contains

Received July 20, 2016; Revised August 5, 2016; Accepted August 29, 2016

*Corresponding author: Tel: +91-11-2698-3409, Fax: +91-11-2698-3409, E-mail: mihassan@jmi.ac.in

[§]These two authors contributed equally to this work.

1,555 coding genes and 65 pseudogenes. The GC content of the genome is 38.9%. The coding genes in the genome encode 1,445 proteins, seven rRNAs, and 36 tRNAs. The genome contains 340 predicted gene products characterized as hypothetical proteins (HPs).

In this study, we have analyzed the sequences of all the HPs from *H. pylori* to assign probable functions. The objective is to identify putative virulence proteins in the proteome that help in pathogenesis. We have used an established protocol [15, 16] for the function prediction of the HPs that comprises leading bioinformatics tools and databases [17-19]. The analysis goes in a systematic way of predicting physicochemical properties of the proteins using ProtParam. Then, subcellular localization using different programs is carried out to assist the function prediction. Identification of transmembrane helices (TMHs) in the HPs to find out membrane protein is carried out using TMHMM and HMMTOP. We have analyzed the HPs for similarity searching using Basic Local Alignment Search Tool (BLAST). Protein-protein interaction is helpful in assessing the function of novel proteins. We have used Search Tool for the Retrieval of Interacting Genes (STRING) database for predicting protein-protein interaction networks for the HPs. The classification of the HPs is done using CATH, Structural Classification of Proteins (SCOP), Pfam, SVMProt, and Protein Analysis through Evolutionary Relationships (PANTHER) database. Conserved domain discovery and motif search in the HPs are carried out using Conserved Domain Architecture Retrieval Tool (CDART), Simple Modular Architecture Research Tool (SMART), InterProScan, and Motif, respectively. We have made final predictions on the basis of a consensus approach [20-22]. The putative function predicted by four or more programs for an HP is considered the probable function of that HP with high precision and high confidence [17, 23]. Finally, we have successfully assigned putative functions to 104 HPs out of 340 HPs with high precision. Furthermore, we have classified proteins on the basis of their involvement in the various biological process and predicted molecular functions into diverse functional groups such as enzymes, binding proteins, transporters, and proteins involved in cellular processes and into the proteins exhibiting miscellaneous functions.

Methods

Data abstraction

In this study, the primary source of genome data is National Center for Biotechnology Information (NCBI) genome database. We extracted preliminary information using "*Helicobacter pylori*" string that redirects to the genome-wide project report of *H. pylori* genome. We selected *H. pylori* 26695 strain from the database with the Accession Code RefSeq NC_000915.1. The genome contains 1,555 genes coding for 1,445 proteins. We then extracted the hypothetical proteins from the pool of 1,445 proteins. We used Uniprot for retrieving Uniprot IDs and fasta sequences of the HPs using their Protein Product IDs (e.g., NP_206816.1). Fasta sequences of the HPs retrieved from Uniprot were used for further analysis.

Physicochemical parameterization

Physicochemical properties of the proteins help in deducing the biochemical characteristics of the proteins and functional characterization. We used Expassy's ProtParam [24] server for estimation of physicochemical parameters of the HPs. ProtParam server is equipped with modules that are capable of predicting an array of physicochemical properties using predefined formulas and experimental inferences. We predicted relative molecular weight, theoretical pI, extinction coefficient, instability index, aliphatic index, and grand average of the hydropathicity of the HPs using ProtParam. All these properties help in identifying the probable function of the proteins. Data for physicochemical parameterization are listed in Supplementary Table 1.

Sub-cellular localization

The function of a protein is very well influenced by its location in the cellular space. For instance, proteins of the exoproteomic pool and secretory proteins often play an essential role in virulence related activities such as adherence to host cells. We used an array of tools to carry out the subcellular localization of the HPs. We used PSORTb [25], PSLpred [26], and Cello [27] to predict the location of HPs in the cell. These predictors use experimental data from known proteins to make predictions for query proteins using their fasta sequences. They predict the possible occurrence of protein in diverse cellular or extracellular localities such as cytoplasm, periplasm, inner membrane, outer membrane, or extracellular space. To predict signal peptides in the HPs, we used SignalP [28] prediction platform for the existence of signal peptides in the HPs, which is a characteristic feature of membrane-bound proteins. SecretomeP [29] server was used to find out nonclassical secretory proteins among the HPs. Prediction of TMHs in the proteins helps in the identification of membranous proteins. We used HMMTOP [30] and TMMHMM [31] for this purpose. Both these programs use Hidden Markov Model (HMM) profiles of training data set to predict TMHs in query sequences. The supplementary data are given in Supplementary Table 2.

Identification of virulence proteins

The present work put stress upon the identification of potential virulence proteins in the pool of HPs. Pathogenic bacteria contain a range of virulence proteins in their pathogenesis machinery. There are adhesins, exotoxins, endotoxins, and secretion systems, etc., that comprise the virulence moiety of pathogenic bacteria. We used VirulentPred [32] and VICMpred [33] for the identification of virulence factors among the HPs. Both these tools are Support Vector Machine (SVM) based using 5-fold cross-validation processes to validate the results. VirulentPred uses the strategy of two-way predictions, i.e., non-Virulent or Virulent whereas VICMpred categorizes proteins into four classes namely proteins involved in cellular processes, metabolism protein, information molecule, and virulence factors. It has a training set of 670 proteins from Gram-negative bacteria including 70 known virulence factors. Information for virulence factors analysis is provided in Supplementary Table 3.

Homology and function prediction

The assertion of homology between proteins derived on the basis of sequence similarity provides insights into the functional properties of an unknown protein showing similarity with a protein of known function. BLAST [34-37] is a commonly used and most reliable tool for the purpose. Structure and function prediction help to identify novel drug targets which can be further utilized for therapeutic intervention [38-46]. We used blastp module to search for homologous proteins to the HPs against a database of nonredundant protein sequences. To decrease the redundancy in the results, a threshold was set for the e-value less than 0.0005 and sequence identity more than 30%. SMART [47] was used for the function prediction. It uses information about domain architecture from known proteins and provides functional annotation of query sequences. Function prediction based on motif discovery was performed using InterProScan [48] and Motif. InterProScan searches the query sequence against Interpro consortium to bring about the function of the proteins using motif information. Motif operates as an interface between user and motif library of known databases. It searches the query sequence against Pfam, TIGRFAM, COG, SMART, PROSITE Patterns, and PROSITE profiles. The user has the facility to choose any of these databases. We also used STRING [49] to predict protein-protein interaction networks for the HPs. It gives functional insights for the HPs based on protein-protein interaction. Information for homology and function prediction is listed in Supplementary Table 4.

Classification and domain assignment

Protein classification and domain assignment using sequence similarity search may give ample evidence for function prediction of the HPs. We have used an array of databases and retrieval tools such as CATH [50], SUPERFAMILY [51], PANTHER [52], Pfam [53], CDART [54], SVMProt [55], and ProtoNet [56] for the classification of the HPs. CATH provides the classification of Protein Data Bank (PDB) protein structure repository. CATH v4.0 release contains 235,858 domains, 2,738 superfamily and 69,058 annotated PDBs. SUPERFAMILY database provides structure and functional annotation of proteins based on HMM using SCOP classification system. PANTHER is another efficient protein classification database based on HMM profiles. PANTHER provides a multi-way classification of proteins on the basis of family and subfamily, molecular function, involvement in a biological process, and association with a pathway in any cellular process. It reduces the risk of redundancy by applying strict HMM scoring strategy. We also used Pfam for the classification of HPs. Pfam is a database of protein families with representative multiple sequence alignments and HMMs for each family. SVMProt was also used for functional classification of the HPs. It is a SVM based classification software trained with the dataset of about 54 functional families of protein. We performed cluster-based classification of the HPs using ProtoNet. It gives a hierarchical classification of proteins using clusters of proteins showing functional similarity. The information about the classification of the HPs is given in Supplementary Table 5.

Results

Sequences of 340 HPs from *H. pylori* 26695 strain tested with exclusive pipeline developed by our group [23, 57]. We used several tools for the sequence analysis such as, BLAST, CATH, SCOP, CDART, InterProScan, Motif, protein family databases, conserved domain databases, protein cluster database, protein-protein interaction database, and other such analysis tools such as virulence predictors, subcellular localization prediction programs, etc. Data produced by all these methods and prediction programs help us deducing results. We successfully assigned probably functions to 104 HPs with high confidence (Table 1). As mentioned earlier, the basis of the confidence level was consensus based, i.e., the similar function for an HP predicted by four or more programs was considered function for the HP with high confidence and precision. To reduce redundancy and to maintain the reliability of the results, we deliberately omitted the HPs for which functions were predicted with low level and less precision.

Table 1. List of 104 HPs with predicted functions from *Helicobacter pylori*

No.	Uniprot ID	Function	Molecular weight (Da)	Theoretical PI	Subcellular localization
1	O24860	TrbC/VIRB2 family protein	10,525.7	8.98	Inner membrane
2	P56066	ATP-dependent Clp protease (ClpS)	10,344.0	5.61	Cytoplasmic
3	O24894[a]	His-Me finger endonucleases-like superfamily	49,556.6	8.64	Cytoplasmic
4	O24904	Cell wall assembly and cell proliferation coordinating protein, KNR4-like	16,102.3	5.06	Cytoplasmic
5	O24914	NLP/P60 family protein-like domain	52,340.1	9.24	Inner membrane
6	O24934	Class II Aldolase and adducin N-terminal domain	27,143.3	8.62	Cytoplasmic
7	P56080	Radical SAM superfamily 4Fe-4S single cluster domain	34,405.4	8.81	Cytoplasmic
8	O24951	Cysteine-rich domain	27,491.8	6.94	Cytoplasmic
9	O24963	VitK2_biosynth family	32,948.5	8.56	Cytoplasmic
10	O24965	AMIN domain protein	23,248.0	9.03	Cytoplasmic
11	O24976	Chemotaxis phosphatase CheZ	28,621.8	4.63	Cytoplasmic
12	P56117	Phospholipase D/nuclease superfamily	58,287.1	9.39	Cytoplasmic
13	O25010	Ribbon-helix-helix protein, copG family	8,598.8	9.52	Cytoplasmic
14	O25022	Cytochrome c-like domain	14,438.9	8.44	Cytoplasmic
15	O25038	MgtE intracellular N domain	24,814.9	8.40	Cytoplasmic
16	P56132	Putative zinc- or iron-chelating domain	15,175.6	8.33	Cytoplasmic
17	O25053	Indole-3-glycerol phosphate synthase	21,035.3	5.34	Cytoplasmic
18	O25058	TrkA-C domain	56,163.3	8.93	Cytoplasmic
19	O25075	Alginate lyase-like domain	37,588.9	8.64	Extracellular
20	O25076	YceI-like domain protein	20,384.8	9.32	Periplasmic
21	O25146	Sporulation/cell division region	29,031.0	9.33	Periplasmic
22	O25155	Calcineurin-like phosphoesterase	29,506.3	8.72	Cytoplasmic
23	O25156	Alanine racemase, N-terminal domain	25,011.0	7.68	Cytoplasmic
24	O25174	Thioesterase/thiol ester dehydrase-isomerase	16,129.6	5.24	Cytoplasmic
25	O25177	DHH phosphoesterase	47,547.4	6.23	Cytoplasmic
26	O25178	Von Willebrand factor type A (vWA) domain	21,136.3	8.44	Cytoplasmic
27	O25192	Toprim-like	58,853.9	8.98	Cytoplasmic
28	O25195	ATPase AAA	41,437.3	5.36	Cytoplasmic
29	O25201	AAA domain	27,447.4	8.31	Cytoplasmic
30	O25213	Tellurite resistance protein TerB	29,846.0	4.87	Cytoplasmic
31	O25255	L,D-transpeptidase catalytic domain	38,618.3	9.22	Outer membrane
32	O25292	Iron-sulfur cluster-binding domain	29,345.4	9.21	Cytoplasmic
33	O25301	Sulfatase	77,616.6	9.16	Inner membrane
34	O25309[a]	Aminodeoxychorismate lyase	37,615.9	9.32	Cytoplasmic
35	O25317	Disulfide bond formation protein DsbB	55,481.2	8.46	Inner membrane
36	O25373	SurA N-terminal domain	47,633.3	8.93	Cytoplasmic
37	O25408	Transcriptional regulatory protein tyrr	43,414.4	6.31	Cytoplasmic
38	O25431	GTP-binding protein, HSR1-related	66,056.1	5.65	Cytoplasmic
39	O25442	Fibronectin type-III domain	48,082.6	9.18	Outer membrane
40	O25450	Molybdopterin biosynthesis protein (MoeB)	23,847.4	8.97	Cytoplasmic
41	O25456	5-Formyltetrahydrofolate cyclo-ligase family	23,647.9	9.98	Cytoplasmic
42	O25468	VitK2_biosynth/menaquinone biosynthesis	26,738.5	9.90	Cytoplasmic
43	O25510	Outer membrane protein transport protein	63,653.5	9.50	Outer membrane
44	O25520	Type I restriction endonuclease subunit S	11,289.9	6.73	Extracellular
45	O25562	Cupin, RmlC-type	11,030.0	6.19	Cytoplasmic
46	O25564	Flagellar hook-length control protein FliK	58,160.8	9.14	Extracellular
47	O25576	DHBP synthase RibB-like alpha/beta domain	16,136.6	10.13	Outer membrane
48	O25579[a]	Toxin-like outer membrane protein	274,562.7	5.78	Extracellular
49	O25589[a]	Acetyltransferase family protein	18,418.3	5.84	Cytoplasmic
50	O25616	50S ribosome-binding GTPase	51,795.4	5.01	Cytoplasmic
51	O25618	Dynamin family protein/GTPase	50,173.9	6.61	Cytoplasmic
52	O25619	Dynamin family protein/GTPase	62,576.6	5.43	Cytoplasmic

Table 1. Continued

No.	Uniprot ID	Function	Molecular weight (Da)	Theoretical PI	Subcellular localization
53	O25624	Outer membrane efflux protein	47,710.0	8.24	Cytoplasmic
54	O25630	Peptidase M50 family protein	11,409.5	8.98	Inner membrane
55	O25642	Nucleotidyl transferase	30,853.8	5.64	Outer membrane
56	O25704	Prokaryotic metallothionein family protein	10,871.9	9.42	Cytoplasmic
57	O25708	Zn-dependent exopeptidases	50,233.2	9.00	Cytoplasmic
58	O25713[a]	Neuraminyllactose-binding hemagglutinin precursor (NLBH)	23,736.8	9.52	Inner membrane
59	O25721	PD-(D/E)XK nuclease superfamily	90,792.8	5.54	Cytoplasmic
60	O25747	Anti-sigma-28 factor, FlgM	8,598.9	8.96	Cytoplasmic
61	O25749	Tetratricopeptide repeat containing protein	38,377.7	9.37	Outer membrane
62	O25761	AAA domain protein	88,752.5	5.56	Cytoplasmic
63	O25768	KH domain RNA binding protein	12,710.6	7.80	Periplasmic
64	O25803	Flagellar motility protein	10,282.0	9.93	Periplasmic
65	O25808	Haloacid dehalogenase-like hydrolase	25,560.1	5.91	Cytoplasmic
66	O25816	RDD family proetin	18,699.6	8.20	Inner membrane
67	O25843	SH3 domain protein	21,396.7	9.06	Cytoplasmic
68	O25848	RDD family proetin	17,843.4	10.07	Inner membrane
69	O25864[a]	Tetratricopeptide repeat containing protein	92,524.5	8.90	Cytoplasmic
70	O25866	Telomere-length maintenance and DNA damage repair	10,736.1	6.09	Cytoplasmic
71	O25870[a]	Glycosyltransferase family 9 (heptosyltransferase)	56,944.3	9.34	Cytoplasmic
72	O25872	HAD superfamily, subfamily IIIB (acid phosphatase)	26,297.4	9.30	Outer membrane
73	O25873	YceI-like domain protein	20,614.7	9.20	Periplasmic
74	O25884	YtkA-like family protein	13,829.3	9.62	Periplasmic
75	O25886[a]	HlyD-like secretion protein	38,611.7	8.94	Outer membrane
76	O25888	Branched-chain amino acid transport protein (AzlD)	13,303.2	9.51	Inner membrane
77	O25892	NYN domain	26,487.8	9.28	Cytoplasmic
78	O25894	DnaJ molecular chaperone homology domain	29,728.8	9.14	Cytoplasmic
79	O25906	Restriction endonuclease-like	34,798.8	7.03	Cytoplasmic
80	O25930	Outer membrane protein assembly factor BamD	26,256.4	9.32	Cytoplasmic
81	O25933	DNA/RNA non-specific endonuclease	15,475.4	8.76	Cytoplasmic
82	O25934[a]	Type-1 restriction enzyme ecoki specificity protein	18,193.6	10.04	Cytoplasmic
83	O25942	Fibronectin-binding protein A N-terminus (FbpA)	51,095.6	9.34	Cytoplasmic
84	O25960	Iojap superfamily ortholog	13,011.9	4.84	Cytoplasmic
85	O25966	S4 domain protein	9,385.0	9.55	Cytoplasmic
86	O25990	Jag N-terminus	23,434.2	9.83	Cytoplasmic
87	O25993[a]	LPP20 lipoprotein	33,870.7	9.31	Outer membrane
88	O25998	Heat shock protein HSLJ	20,480.2	6.54	Cytoplasmic
89	O26000	Mce related family protein	30,489.2	8.34	Outer membrane
90	O26006	Type I restriction modification DNA specificity domain	45,084.6	8.14	Outer membrane
91	O26007	Type I restriction modification DNA specificity domain	77,049.7	8.18	Outer membrane
92	O26014	TPR repeat family protein	96,653.8	6.18	Cytoplasmic
93	O26015	Carbon-nitrogen hydrolase	30,759.2	5.20	Cytoplasmic
94	O26020	ABC-2 family transporter protein	42,545.2	7.16	Inner membrane
95	O26021	ABC-2 family transporter protein	41,089.8	8.80	Inner membrane
96	O26022	Outer membrane efflux proteins (OEP)	57,011.9	9.16	Outer membrane
97	O26025	NIF system FeS cluster assembly, NifU, C-terminal	10,120.9	6.57	Cytoplasmic
98	O26035	Riboflavin biosynthesis protein (ribG)	39,025.2	8.51	Cytoplasmic
99	O26042[a]	Ferrichrome iron receptor-related	97,384.6	9.05	Outer membrane
100	O26046	Type IIS restriction enzyme R and M protein (ECO57IR)	149,716.0	7.14	Cytoplasmic
101	O26058	Purine nucleoside phosphorylase (PunB)	20,200.3	5.42	Cytoplasmic
102	O26095	Flagellar biosynthetic protein flhb	9,981.6	5.26	Cytoplasmic
103	O26100	PAP2 family protein	24,548.7	9.62	Inner membrane
104	O26107	Ubiquinol-cytochrome C chaperone	28,417.6	5.45	Cytoplasmic

[a] Hypothetical proteins (HPs) predicted virulent in virulence factors analysis.

Discussion

Classification of the HPs

For the ease of the approach for understanding the probable involvement of these HPs in pathogenesis, we categorized all 104 HPs into various functional groups on the basis of their individual molecular function and their involvement in various biological processes (Fig. 1). We found 27 HPs showing similarities with various enzyme classes like oxidoreductases, hydrolases, transferases, etc. Ten HPs are categorized as transporters, 26 showing features of binding proteins, 23 HPs have predicted to be involved in various cellular and regulatory processes and 18 HPs are listed in the category of proteins showing miscellaneous functions. These HPs are further studied and extensively analyzed using previously available literature and experimental studies.

Enzymes, having catalytic properties, play a substantial role in the life of a living organism to provide biochemical machinery for various cellular and regulatory processes. We found 27 HPs showing similarities experimentally characterized enzymes representatives of enzyme classes. HP O25317 showed similarity with disulfide bond formation protein DsbB. Disulfide bonds provide stability and maturation strength to the protein thus, DsbB has a critical role in the development of substantial protein machinery that may be involved in the metabolic or regulatory pathways [58] of that pathogen, therefore, helping in the pathogenesis. Out of 27 enzymes, five HPs are categorized as transferases. HPs O25589 and O25870 are showing similarity with acetyltransferase family protein and glycosyltransferase family 9 (heptosyltransferase), respectively. Both these HPs are predicted virulent in virulent factors analysis. Glycosyltranferases facilitate the "biosynthesis of disaccharides, oligosaccharides, and polysaccharides" by catalyzing the transfer of sugar moieties [59]. HP O25870 is predicted heptosyltransferase may be a potential drug target. Heptosyltranferase help in the formation of the core region of lipopolysaccharides which constitute the major component outer membrane structure in Gram-negative bacteria [60]. About 60% of all predicted enzymes belong to hydrolases enzyme class and most of them are involved in metabolic pathways. In the predicted hydrolases, there are ATPases, restriction endonucleases, phosphoesterases, etc., that facilitate the processes of transcription, translation, functional group localization, and other such essential activities that help in the development and propagation of the pathogen inside the host. There are four HPs showing similarities with member proteins of lyase enzyme class. HP O25309 is showing similarity with aminodeoxychorismate lyase and is predicted as virulent factor. Aminodeoxychorismate lyase is a class member of pyridoxal-phosphate-binding protein class IV which helps in the biosynthesis of tetrahydrofolate by aminodeoxychorismate to para-aminobenzoate. Tetrahydrofolate is an essential precursor in purine biosynthesis [61].

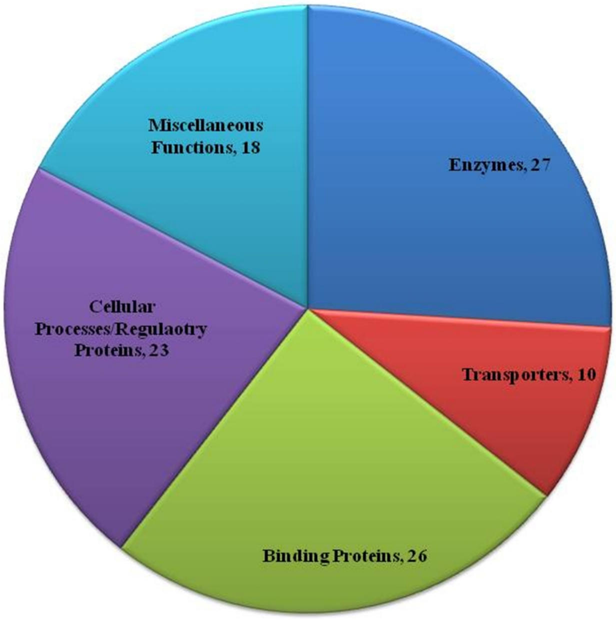

Fig. 1. Classification of hypothetical proteins into enzymes (n=27), transporters (n=10), binding proteins (n=26), cellular processes/ regulatory proteins (n=23) and miscellaneous functions (n=18).

Transporters have always remained a subject of interest during the process of novel drug discovery against the pathogenic diseases. Transporters, due to their specific evolution making them capable of transporting essential molecules, are involved in a wide range of metabolic pathways and other important cellular processes. *H. pylori* genome has an ample amount of genes that encode a large number of transporter proteins, mainly ATP-binding cassette (ABC) transporters. In the predicted HPs, we found 10 HPs showing characteristic similarity with transporters. HPs O26020, and O26021 are showing similarity with ABC-2 family transporter proteins. ABC transporters, specific to prokaryotes, are the leading molecules that fulfill the energy requirement of the organism for diverse biological processes [62]. The required amount energy that they provide comes from the hydrolysis of ATP molecules performed by ABC transporters [63] having specifically evolved domains for ATP hydrolysis. We found HP O26042 is showing similarity with ferrichrome iron receptor (fhuA). Iron uptake is believed to be preferential activity in *H. pylori* for the survival in the host system [64]. fhuA is an outer membrane transport protein which catalyzes the transport of ferrichrome and also acts as a receptor for T5 phages in *Escherichia coli* and other toxic substances [65]. HP O26042 is also predicted virulent in virulence factors analysis. Thus, it can be considered potential drug target.

Twenty-six HPs are characterized as binding proteins. These proteins are further specified according to their functions as adhensin, DNA-, RNA-, protein-, nucleotide-, metal- and lipid-binding proteins. Some of the representative members of this group are may be known involved in leading cell activities, transcription, translation, and other regulatory processes. In this group, we have identified four HPs showing characteristics of restriction modification proteins, three of which belong to type I and one belong to type II. All these proteins may have an essential role in DNA modification. HP O25934 is showing similarity with type-1 restriction enzyme ecoKI specificity protein (hsdS) and predicted virulent by both VICMpred and VirulentPred. Type-1 restriction enzyme ecoKI specificity protein belongs to the class of S-adenosyl-L-methionine dependent endonucleases that are constituents of bacterial DNA restriction-modification mechanisms, which guards the organism from foreign DNA invasion [66]. We identified HP O25749 showing positive virulence and exhibiting similarity with tetratricopeptide repeat (TPR) protein. TPR is a signature motif of proteins regulating protein-protein interaction and the formation of multiprotein complexes [67]. Proteins with TPR motifs are involved in important biological processes such as cell cycle, protein folding, transcriptional regulation, etc. [68]. Involvement in leading processes makes them liable to be treated as potential drug targets. We found two HPs O25618 and O25619 are showing significant similarity to dynamin like GTPases. The function of dynamin GTPases is well studied in eukaryotes. They are involved in membrane fusion and fission mediated by the hydrolysis of GTP molecules but the exact function of their prokaryotic counterparts, despite the existence of structural data, is not well understood and needs a further probe to straighten out their role in prokaryotes [69].

We have identified 23 HPs may be involved in diverse cellular processes and regulatory mechanisms. Proteins mediating the formation of cell envelope such flagellar biosynthesis proteins, flagellar motility proteins are signature members of this group. Flagella is responsible for bacterial motility in a host environment which helps in the colonization of the pathogen [70]. *H. pylori* is equipped with “five to seven unipolar” flagella that are protected against gastric acidity due to the presence of a covering sheath formed of phospholipids [71]. There are a relatively higher number of genes in *H. pylori* that encodes flagellar proteins supporting the fact that motility facilitates the colonization of the pathogen in the host body; thus, their association with bacterial virulence is also subjected to consideration in the course of drug discovery. HP O26095 is showing similarity with flagellar biosynthetic protein flhb that mediates the formation of flagella. It may be a potential drug target. We found HP O25564 similar to flagellar hook-length control protein FliK that controls the length of the flagellar hook during flagellar biosynthesis [72]. In the *H. pylori* genome, there are seven known genes encoding molecular chaperons. We have identified HP O25894 is showing homology with molecular chaperon. DnaJ, is signature member of the family of molecular chaperons that exhibit a diverse number of molecular functions such ATP binding, metal ion binding, unfolded protein binding and is involved in a number of leading biological processes like protein folding, protein unfolding, DNA replication, and response to heat shock, etc. [73]. The involvement of chaperons in essential cellular processes required for survival and propagation of pathogen make them potential drug targets for the development of effective drugs against pathogenicity.

Though we have categorized HPs in the definite functional classes on the basis of their molecular functions and their involvement in diverse biological processes, but there HPs which exhibit some unique functions or functions are not clearly classified in the available literature. We put those HPs in the group of proteins exhibiting miscellaneous functions. HP O25579 is identified as toxin like outer membrane protein and showing significant virulence in virulence factors analysis. We found HP O25993 similar to lipoprotein with positive virulence. Despite the fact that *H. pylori* infects the host in the free environment, evidence for adherence to epithelial cells of the gastric tissues of the host are also found [64]. Outer membrane proteins and lipoproteins have an effective role in cell adhesion in *H. pylori* [5]; thus, they may be taken as strong candidates for drug targets. We identified HP O25713 similar to neuraminyllactose-binding hemagglutinin (NLBH) with substantial virulence. In *H. pylori*, NLBH, which is also a lipoprotein, has an effective role in adhesion to the gastric epithelium of the host [74]. We identified three characterized genes in the *H. pylori* genome that encodes NLBH proteins at distant locations. The specific function of NLBH signifies its virulence making it a potential therapeutic target.

Virulence factors

As discussed in the last section, we have performed virulence factors analysis for all the 340 HPs to bring about virulent proteins that play an effective role in the propagation of disease. We preferable selected consensus-based approach for the purpose of taking the results of both predictors VICMpred and VirulentPred as positive. Thus, we found 22 HPs predicted virulent by both these programs (Table 2). While looking for the virulent proteins in the array

Table 2. List of predicted virulent proteins from *Helicobacter pylori*

No.	Uniprot ID	VirulentPred	VICMPred
1	O24863	Yes	Yes
2	O24894	Yes	Yes
3	O24909	Yes	Yes
4	O25085	Yes	Yes
5	O34410	Yes	Yes
6	O25309	Yes	Yes
7	O25457	Yes	Yes
8	O25579	Yes	Yes
9	O25589	Yes	Yes
10	O25601	Yes	Yes
11	O25713	Yes	Yes
12	O34410	Yes	Yes
13	K4NT00	Yes	Yes
14	O25864	Yes	Yes
15	O25870	Yes	Yes
16	O25886	Yes	Yes
17	O25934	Yes	Yes
18	O25979	Yes	Yes
19	O25993	Yes	Yes
20	O26042	Yes	Yes
21	K4NEW8	Yes	Yes

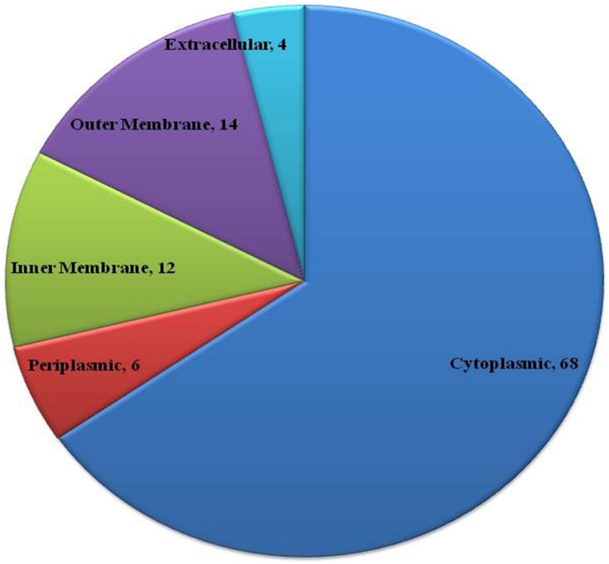

Fig. 2. Classification of HPs on the basis of subcellular localization.

of 104 predicted HPs, we found 11 HPs showing positive virulence that are mentioned in Table 1. Concrete specification of virulence proteins amongst the predicted functional candidates paves the way for further studies on drug discovery and development in a more focused way. Therefore, results of virulence factors analysis hold significant in the lookups for further study and experimental characterization of predicted HPs.

Subcellular localization

Identification of subcellular location of the protein in a computer based functional analysis is significant because there is a strong relation between the function and location of the protein in cellular space [75, 76]. It also gives insight into the determination of probable drug target or vaccine target among the identified virulent proteins. For the newly assigned 104 HPs, we deduced their relative subcellular locations from the results of subcellular localization prediction discussed earlier on the basis of consensus-based approach. Relative subcellular locations of predicted HPs are given in Table 1. We also classified the predicted HPs based on their subcellular locations (Fig. 2). Associating the results of subcellular localization with those of virulence factors analysis may help in the identification of probable drug or vaccine targets.

In conclusion, computational sequence analysis of HPs in order to find out possible functional clues is an extensive work and need much patience for each gene is individually analyzed with an array of tools and databases. The inferences are drawn with a sensitive approach to discard the possibilities of false-positives. Due to the occurrence of similar looking patterns, prediction software may predict different function for similar HP than that predicted by another tool. Therefore, we have selected a more sensitive consensus-based approach, cross-checking the results of all used programs and then deducing inferences on the basis of majority rule. Majority rule is the criteria taking the function predicted by four or more tools as the probable function of the HP. This way, we have successfully predicted probable functions of 104 HPs with high level confidence. A wide range of HPs showing functional similarities with the proteins those play an essential role in bacterial pathogenesis. The study may pave the way for experimentalists to look forward to the possibilities of *in vitro* functional characterization of virulent proteins that may be considered potential therapeutic targets in the process of drug discovery.

Supplementary materials

Supplementary data including five tables can be found with this article online at http://www.genominfo.org/src/sm/gni-14-125-s001.pdf.

Acknowledgments

This work is supported by the Indian Council of Medical Research, Government of India (BIC/12(04)/2012) to MIH. The central instrumentation facility of Jamia Millia Islamia is highly acknowledged for providing high-speed server.

References

1. Shiota S, Suzuki R, Yamaoka Y. The significance of virulence factors in *Helicobacter pylori*. *J Dig Dis* 2013;14:341-349.
2. Testerman TL, Morris J. Beyond the stomach: an updated view of *Helicobacter pylori* pathogenesis, diagnosis, and treatment. *World J Gastroenterol* 2014;20:12781-12808.
3. Marshall BJ, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* 1984;1:1311-1315.
4. Cover TL, Blaser MJ. *Helicobacter pylori* infection, a paradigm for chronic mucosal inflammation: pathogenesis and implications for eradication and prevention. *Adv Intern Med* 1996;41:85-117.
5. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, *et al*. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 1997;388:539-547.
6. Yamaoka Y, Graham DY. *Helicobacter pylori* virulence and cancer pathogenesis. *Future Oncol* 2014;10:1487-1500.
7. Cid TP, Fernández MC, Benito Martínez S, Jones NL. Pathogenesis of *Helicobacter pylori* infection. *Helicobacter* 2013;18 Suppl 1:12-17.
8. Nakamura S, Matsumoto T. *Helicobacter pylori* and gastric mucosa-associated lymphoid tissue lymphoma: recent progress in pathogenesis and management. *World J Gastroenterol* 2013; 19:8181-8187.
9. Tomoda A, Kamiya S, Suzuki H. *Helicobacter pylori* and pathogenesis. *Biomed Res Int* 2015;2015:304768.
10. Watari J, Chen N, Amenta PS, Fukui H, Oshima T, Tomita T, *et al*. *Helicobacter pylori* associated chronic gastritis, clinical syndromes, precancerous lesions, and pathogenesis of gastric cancer development. *World J Gastroenterol* 2014;20:5461-5473.
11. de Bernard M, Josenhans C. Pathogenesis of *Helicobacter pylori* infection. *Helicobacter* 2014;19 Suppl 1:11-18.
12. De Falco M, Lucariello A, Iaquinto S, Esposito V, Guerra G, De Luca A. Molecular mechanisms of *Helicobacter pylori* pathogenesis. *J Cell Physiol* 2015;230:1702-1707.
13. Hagiwara T, Mukaisho K, Nakayama T, Hattori T, Sugihara H. Proton pump inhibitors and *Helicobacter pylori*-associated pathogenesis. *Asian Pac J Cancer Prev* 2015;16:1315-1319.
14. Brown LM. *Helicobacter pylori*: epidemiology and routes of transmission. *Epidemiol Rev* 2000;22:283-297.
15. Naqvi AA, Shahbaaz M, Ahmad F, Hassan MI. Identification of functional candidates amongst hypothetical proteins of *Treponema pallidum* ssp. *pallidum*. *PLoS One* 2015;10:e0124177.
16. Naqvi AA, Ahmad F, Hassan MI. Identification of functional candidates amongst hypothetical proteins of *Mycobacterium leprae* Br4923, a causative agent of leprosy. *Genome* 2015;58: 25-42.
17. Shahbaaz M, Ahmad F, Hassan MI. Structure-based function analysis of putative conserved proteins with isomerase activity from *Haemophilus influenzae*. *3 Biotech* 2015;5:741-763.
18. Shahbaaz M, Bisetty K, Ahmad F, Hassan MI. Towards new drug targets? Function prediction of putative proteins of *Neisseria meningitidis* MC58 and their virulence characterization. *OMICS* 2015;19:416-434.
19. Shahbaaz M, Bisetty K, Ahmad F, Hassan I. Current advances in the identification and characterization of putative drug and vaccine targets in the bacterial genomes. *Curr Top Med Chem* 2016;16:1040-1069.
20. Kumar K, Prakash A, Islam A, Ahmad F, Hassan MI. Identification of functional candidates amongst hypothetical proteins of *Neisseria gonorrhoeae*. *Lett Drug Des Discov* 2016;13:451-464.
21. Kumar K, Prakash A, Anjum F, Islam A, Ahmad F, Hassan MI. Structure-based functional annotation of hypothetical proteins from *Candida dubliniensis*: a quest for potential drug targets. *3 Biotech* 2015;5:561-576.
22. Kumar K, Prakash A, Tasleem M, Islam A, Ahmad F, Hassan MI. Functional annotation of putative hypothetical proteins from *Candida dubliniensis*. *Gene* 2014;543:93-100.
23. Shahbaaz M, Hassan MI, Ahmad F. Functional annotation of conserved hypothetical proteins from *Haemophilus influenzae* Rd KW20. *PLoS One* 2013;8:e84263.
24. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 2003;31:3784-3788.
25. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, *et al*. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010;26:1608-1615.
26. Bhasin M, Garg A, Raghava GP. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* 2005;21:2522-2524.
27. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci* 2004;13: 1402-1406.
28. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 2011;8:785-786.
29. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. *BMC Microbiol* 2005;5:58.
30. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 2001;17:849-850.
31. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001; 305:567-580.
32. Garg A, Gupta D. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics* 2008;9:62.
33. Saha S, Raghava GP. VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics* 2006;4:42-47.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403-410.
35. Khan FI, Shahbaaz M, Bisetty K, Waheed A, Sly WS, Ahmad F, *et al*. Large scale analysis of the mutational landscape in β-glucuronidase: a major player of mucopolysaccharidosis type VII. *Gene* 2016;576(1 Pt 1):36-44.
36. Shahbaaz M, Bisetty K, Ahmad F, Hassan MI. *In silico* ap-

proaches for the identification of virulence candidates amongst hypothetical proteins of *Mycoplasma pneumoniae* 309. *Comput Biol Chem* 2015;59 Pt A:67-80.

37. Zaidi S, Hassan MI, Islam A, Ahmad F. The role of key residues in structure, function, and stability of cytochrome-c. *Cell Mol Life Sci* 2014;71:229-255.
38. Devika NT, Amresh P, Hassan MI, Ali BM. Molecular modeling and simulation of the human eNOS reductase domain, an enzyme involved in the release of vascular nitric oxide. *J Mol Model* 2014;20:2470.
39. Hassan MI. Editorial. Recent advances in the structure-based drug design and discovery. *Curr Top Med Chem* 2016;16:899-900.
40. Hassan MI, Bilgrami S, Kumar V, Singh N, Yadav S, Kaur P, *et al*. Crystal structure of the novel complex formed between zinc alpha2-glycoprotein (ZAG) and prolactin-inducible protein (PIP) from human seminal plasma. *J Mol Biol* 2008;384:663-672.
41. Hassan MI, Kumar V, Singh TP, Yadav S. Structural model of human PSA: a target for prostate cancer therapy. *Chem Biol Drug Des* 2007;70:261-267.
42. Hassan MI, Kumar V, Somvanshi RK, Dey S, Singh TP, Yadav S. Structure-guided design of peptidic ligand for human prostate specific antigen. *J Pept Sci* 2007;13:849-855.
43. Hassan MI, Waheed A, Grubb JH, Klei HE, Korolev S, Sly WS. High resolution crystal structure of human β-glucuronidase reveals structural basis of lysosome targeting. *PLoS One* 2013;8:e79687.
44. Hoda N, Naz H, Jameel E, Shandilya A, Dey S, Hassan MI, *et al*. Curcumin specifically binds to the human calcium-calmodulin-dependent protein kinase IV: fluorescence and molecular dynamics simulation studies. *J Biomol Struct Dyn* 2016;34:572-584.
45. Khan FI, Aamir M, Wei DQ, Ahmad F, Hassan MI. Molecular mechanism of Ras-related protein Rab-5A and effect of mutations in the catalytically active phosphate-binding loop. *J Biomol Struct Dyn* 2016:1-14.
46. Naz F, Shahbaaz M, Khan S, Bisetty K, Islam A, Ahmad F, *et al*. PKR-inhibitor binds efficiently with human microtubule affinity-regulating kinase 4. *J Mol Graph Model* 2015;62:245-252.
47. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 2012;40:D302-D305.
48. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;30:1236-1240.
49. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, *et al*. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013;41:D808-D815.
50. 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-D498.
51. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. *Nucleic Acids Res* 2000;28:257-259.
52. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* 2013;41:D377-D386.
53. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2004;32:D138-D141.
54. Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. *Genome Res* 2002;12:1619-1623.
55. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* 2003;31:3692-3697.
56. Rappoport N, Karsenty S, Stern A, Linial N, Linial M. ProtoNet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. *Nucleic Acids Res* 2012;40:D313-D320.
57. Khan S, Shahbaaz M, Bisetty K, Ahmad F, Hassan MI. Classification and functional analyses of putative conserved proteins from *Chlamydophila pneumoniae* CWL029. *Interdiscip Sci* 2015 Dec 9 [Epub]. http://dx.doi.org/10.1007/s12539-015-0134-7.
58. Kadokura H, Katzen F, Beckwith J. Protein disulfide bond formation in prokaryotes. *Annu Rev Biochem* 2003;72:111-135.
59. Campbell JA, Davies GJ, Bulone V, Henrissat B. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J* 1997;326(Pt 3):929-939.
60. Gronow S, Brabetz W, Brade H. Comparative functional characterization *in vitro* of heptosyltransferase I (WaaC) and II (WaaF) from *Escherichia coli*. *Eur J Biochem* 2000;267:6602-6611.
61. Green JM, Merkel WK, Nichols BP. Characterization and sequence of *Escherichia coli* pabC, the gene encoding aminodeoxychorismate lyase, a pyridoxal phosphate-containing enzyme. *J Bacteriol* 1992;174:5317-5323.
62. Higgins CF. ABC transporters: physiology, structure and mechanism: an overview. *Res Microbiol* 2001;152:205-210.
63. Schneider E, Hunke S. ATP-binding-cassette (ABC) transport systems: functional and structural aspects of the ATP-hydrolyzing subunits/domains. *FEMS Microbiol Rev* 1998;22:1-20.
64. Labigne A, de Reuse H. Determinants of *Helicobacter pylori* pathogenicity. *Infect Agents Dis* 1996;5:191-202.
65. Bonhivers M, Ghazi A, Boulanger P, Letellier L. FhuA, a transporter of the *Escherichia coli* outer membrane, is converted into a channel upon binding of bacteriophage T5. *EMBO J* 1996;15:1850-1856.
66. Kisiela M, Skarka A, Ebert B, Maser E. Hydroxysteroid dehydrogenases (HSDs) in bacteria: a bioinformatic perspective. *J Steroid Biochem Mol Biol* 2012;129:31-46.
67. Cerveny L, Straskova A, Dankova V, Hartlova A, Ceckova M, Staud F, *et al*. Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms. *Infect Immun* 2013;81:629-635.
68. Goebl M, Yanagida M. The TPR snap helix: a novel protein repeat motif from mitosis to transcription. *Trends Biochem Sci* 1991;16:173-177.

69. Praefcke GJ, McMahon HT. The dynamin superfamily: universal membrane tubulation and fission molecules? *Nat Rev Mol Cell Biol* 2004;5:133-147.
70. Marais A, Mendz GL, Hazell SL, Mégraud F. Metabolism and genetics of *Helicobacter pylori*: the genome era. *Microbiol Mol Biol Rev* 1999;63:642-674.
71. Geis G, Suerbaum S, Forsthoff B, Leying H, Opferkuch W. Ultrastructure and biochemical studies of the flagellar sheath of *Helicobacter pylori*. *J Med Microbiol* 1993;38:371-377.
72. Kawagishi I, Homma M, Williams AW, Macnab RM. Characterization of the flagellar hook length control protein fliK of *Salmonella typhimurium* and *Escherichia coli*. *J Bacteriol* 1996;178:2954-2959.
73. Bukau B. Regulation of the *Escherichia coli* heat-shock response. *Mol Microbiol* 1993;9:671-680.
74. Chaturvedi G, Tewari R, Mrigank, Agnihotri N, Vishwakarma RA, Ganguly NK. Inhibition of *Helicobacter pylori* adherence by a peptide derived from neuraminyl lactose binding adhesin. *Mol Cell Biochem* 2001;228:83-89.
75. Scott MS, Calafell SJ, Thomas DY, Hallett MT. Refining protein subcellular localization. *PLoS Comput Biol* 2005;1:e66.
76. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. *J Mol Biol* 1998;283:707-725.