3 Biotech (2015) 5:317–336
DOI 10.1007/s13205-014-0231-z

ORIGINAL ARTICLE

Structure-based functional annotation of putative conserved proteins having lyase activity from *Haemophilus influenzae*

Mohd. Shahbaaz · Faizan Ahmad · Md. Imtaiyaz Hassan

Received: 13 April 2014 / Accepted: 28 May 2014 / Published online: 17 June 2014

Abstract *Haemophilus influenzae* is a small pleomorphic Gram-negative bacteria which causes several chronic diseases, including bacteremia, meningitis, cellulitis, epiglottitis, septic arthritis, pneumonia, and empyema. Here we extensively analyzed the sequenced genome of *H. influenzae* strain Rd KW20 using protein family databases, protein structure prediction, pathways and genome context methods to assign a precise function to proteins whose functions are unknown. These proteins are termed as hypothetical proteins (HPs), for which no experimental information is available. Function prediction of these proteins would surely be supportive to precisely understand the biochemical pathways and mechanism of pathogenesis of *Haemophilus influenzae*. During the extensive analysis of *H. influenzae* genome, we found the presence of eight HPs showing lyase activity. Subsequently, we modeled and analyzed three-dimensional structure of all these HPs to determine their functions more precisely. We found these HPs possess cystathionine-β-synthase, cyclase, carboxymuconolactone decarboxylase, pseudouridine synthase A and C, D-tagatose-1,6-bisphosphate aldolase and aminodeoxychorismate lyase-like features, indicating their corresponding functions in the *H. influenzae*. Lyases are actively involved in the regulation of biosynthesis of various hormones, metabolic pathways, signal transduction, and DNA repair. Lyases are also considered as a key player for various biological processes. These enzymes are critically essential for the survival and pathogenesis of *H. influenzae* and, therefore, these enzymes may be considered as a potential target for structure-based rational drug design. Our structure–function relationship analysis will be useful to search and design potential lead molecules based on the structure of these lyases, for drug design and discovery.

Keywords *Haemophilus influenzae* · Hypothetical protein · Sequence analysis · Structure analysis · Function prediction · Gene annotation · Structure–function relationship

Introduction

Haemophilus influenzae strain Rd KW20 was the first organism whose genome was successfully sequenced in 1995 by Craig Venter's group (Fleischmann et al. 1995). The genome of *H. influenzae* contains 1,740 protein-coding genes, 2 transfer RNA genes, and 18 other RNA genes in a single circular chromosome of 1,830,140 base pairs (Fleischmann et al. 1995). This organism belongs to the family *Pasteurellaceae* (Christensen 2008; Kuhnert 2008), a nontypeable (NTHi) causing bacteremia and acute bacterial meningitis in infants, children and adults (Murphy and Sethi 1992; Sethi and Murphy 2001). *H. influenzae* is part of the normal nasopharyngeal flora in humans and is often coupled with otitis media, chronic bronchitis and community-acquired pneumonia (Apisarnthanarak and Mundy 2005; Eldika and Sethi 2006). *H. influenzae* is obligatory to human and requires β-nicotinamide adenine dinucleotide

Electronic supplementary material The online version of this article (doi:10.1007/s13205-014-0231-z) contains supplementary material, which is available to authorized users.

Mohd. Shahbaaz
Department of Computer Science, Jamia Millia Islamia, New Delhi 110025, India

F. Ahmad · Md. Imtaiyaz Hassan (✉)
Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India
e-mail: mihassan@jmi.ac.in

and heme for growth (Markel et al. 2007; Morton et al. 2004a). Due to its inability to produce iron-containing heme for the cytoplasmic enzymes, it uses multiple mechanisms to obtain heme (Stojiljkovic and Perkins-Balding 2002) that include various heme acquisition proteins like the heme utilization protein, Hup (Morton et al. 2004b) and heme-binding lipoprotein Hbp A (Morton et al. 2005). Mechanism for heme acquisition indicates a strict regulation of iron homeostasis in *H. influenzae,* and it is important for its survival and virulence (Morton et al. 2004a).

The sequenced genome of various organisms on comparative genomics analysis shows that a significant portion of the genes encodes "hypothetical proteins (HPs)", i.e. functionally uncharacterized proteins but found in organisms (Galperin and Koonin 2004). HPs are predicted to be expressed from an open reading frame, but no experimental evidences are available for their function. The majority of HPs is believed to be a product of pseudogenes in human as well as of other organisms and constitute an extensive fraction of their proteomes (Desler et al. 2012; Galperin 2001). HPs have unique sequences and are essential determinants of species-specific phenotypic properties, such as pathogenicity in a given organism (Desler et al. 2012; Galperin 2001). These determinants are considered as a potent drug targets in pathogenic organisms (Kumar et al. 2014; Tsoka and Ouzounis 2000). Experimental characterization of these HPs in a model organism, such as *Escherichia coli*, *Saccharomyces cerevisiae, H. influenzae* etc., will be helpful in complete understanding of their biological systems at the molecular level (Nimrod et al. 2008; Park et al. 2012; Pidugu et al. 2009). The recent functional annotation of previously uncharacterized tRNA modification enzymes (Alexandrov et al. 2002; Jackman et al. 2003; Soma et al. 2003) of the deoxyxylulose pathway (Eisenreich et al. 2001, 2004), and of the central role of cyclic diguanylate in bacterial signaling (Galperin 2004; Jenal 2004) emphasizes the importance HP functional characterization.

A precise prediction of protein functions depends on accurate protein folding, which is revealed by their three-dimensional structure (Dobson 2003). Similarity in protein structure leads to similarity in biological function that cannot be predicted by sequence analysis alone, for the protein sequence is less conserved than the tertiary structure of a protein (Illergard et al. 2009). Available information on protein structure is useful for characterization of binding motifs and catalytic cores present in proteins (Shapiro and Harris 2000). Knowledge of protein structure is important for rational drug design that mainly relies on the structure of the target protein to narrow the searching of lead compound and for refinement of the compound (Capdeville et al. 2002; Klebe 2000). Hence, structure prediction of HPs may deliver some evidences about their biological function and help in the drug discovery of better therapeutics for treating diseases caused by *H. influenzae*.

Table 1 List of hypothetical protein with lyase activity in *H. influenzae* strain Rd KW20

S. no.	Accession no	Gene ID	Protein product	Uniprot ID	Protein name
1	NC_000907.1	949660	NP_438613.1	P44717	HP HI0452
2	NC_000907.1	950684	NP_438775.1	P44782	HP HI0617
3	NC_000907.1	950339	NP_439586.2	P44197	HP HI1435
4	NC_000907.1	950454	NP_439740.1	P45267	HP HI1598
5	NC_000907.1	950030	NP_439212.1	Q57498	HP HI1053
6	NC_000907.1	949991	NP_439177.1	P44095	HP HI1016
7	NC_000907.1	950004	NP_439172.1	P44093	HP HI1011
8	NC_000907.1	950653	NP_438618.1	P44720	HP HI0457

Recently, we have annotated the function of HPs from *H. influenzae* (Shahbaaz et al. 2013). During extensive sequence analysis of HPs from *H. influenzae,* we predicted that eight HPs possess lyase activity as listed in Table 1. Since working on the structure-based drug design therefore we are looking for the novel therapeutic targets (Hassan et al. 2007a, b; Thakur et al. 2013; Thakur and Hassan 2011). Lyase enzymes are important for virulence and survival of pathogens in the host. These enzymes provide essential nutrients and are involved in modifying the local environment for the favorable growth (Bjornson 1984). Two examples of such enzymes are cystathionine β-lyase and isocitrate lyase. The formation and degradation of cystathionine cause the mobilization of sulfur from Cys, which is required for the biosynthesis of methionine in bacteria and is important for the virulence (Ejim et al. 2004). Similarly, the isocitrate lyase is an enzyme of the glyoxylate pathway (an anaplerotic pathway of the TCA cycle), which cleaves isocitrate to glyoxylate and succinate, which are potential drug targets for control of many human diseases caused by various pathogens (Britton et al. 2001). Thus, lyase enzymes are used for sequence and structure analysis for further study in this paper.

Materials and methods

Sequence retrieval

The genome analysis of *H. influenzae* shows that 1,657 proteins are present in its proteome (http://www.ncbi.nlm.nih.gov/genome/). We have selected all 429 proteins (Table S1 and S2) with a name HP separately and perform rigorous sequence analyses using fasta sequences retrieved from UniProt (http://www.uniprot.org/) (Shahbaaz et al.

2013). Sequences of proteins having lyase-like activity were used for further structure prediction and analysis.

Sub-cellular localization

To characterize a protein as a drug and/or vaccine target, sub-cellular localization is essentially important. We used sub-cellular localization tools, PSORTb (Yu et al. 2010b), PSLpred (Bhasin et al. 2005) and CELLO (Yu et al. 2006). SignalP 4.1 (Emanuelsson et al. 2007) was used for signal peptide prediction and SecretomeP (Bendtsen et al. 2005) for identifying protein involvement in non-classical secretory pathway. TMHMM (Krogh et al. 2001) and HMMTOP (Tusnady and Simon 2001) were used to predict the propensity of a protein to be a membrane protein.

Sequence comparisons

We predicted the function of HPs using conserved sequence patterns in protein families (Chen and Jeong 2000). BLASTp (Altschul et al. 1990) and HHpred (Soding et al. 2005) based on hidden Markov models were used for remote protein homology detection against various protein databases such as PDB (Bernstein et al. 1978), SCOP (Hubbard et al. 1999), CATH, etc.

Function prediction

The domain analysis allows more precise function prediction (Reid et al. 2007). We used various publicly available databases for functional annotation such as Pfam (Punta et al. 2012), PANTHER (Mi et al. 2005), SMART (Letunic et al. Letunic et al. 2012), SUPERFAMILY (Gough et al. 2001), CATH (Sillitoe et al. 2013), CDART (Geer et al. 2002), SYSTERS (Meinel et al. 2005), ProtoNet (Rappoport et al. 2012) and SVMProt (Cai et al. 2003). CDART and SMART were used for similarity search based on domain architecture and profiles rather than by direct sequence similarity. MOTIF (Kanehisa 1997) and InterProScan (Quevillon et al. 2005) help to identify signature sequence in a particular protein. The MEME suite (Bailey et al. 2009) has been used to perform motif-sequence database searching.

Virulence factor analysis

Virulence factors are considered as potential drug/vaccine targets (Baron and Coombes 2007). We used VICMpred (Saha and Raghava 2006) and Virulentpred (Garg and Gupta 2008) for identifying potential virulence factors. Both are SVM-based methods used to predict bacterial virulence factors from protein sequences with a significant accuracy.

Functional protein association networks

To understand the role of a protein in certain biological pathways, protein–protein interaction analysis is very useful. Here, we used STRING (version-9.05) (Szklarczyk et al. 2011) to predict protein interaction partners of HPs. The interactions include direct (physical) and indirect (functional) associations, experimental or co-expression.

Structure prediction

Three-dimensional structure of proteins P44197 and P44720 was predicted using the homology modeling (Marti-Renom et al. 2000) on MODELLER module present in the Discovery Studio 3.5 (Accelrys 2013). We have identified templates using sequence similarity search methods PSI-BLAST (Altschul et al. 1997) and found pseudouridine synthases RluC (Mizutani et al. 2004) and aminodeoxychorismate lyase from *Escherichia coli* (PDB id: 2RIF) as a suitable template for P44197 and P44720, respectively. The template and query sequences were aligned and finally used for modeling protein structure in MODELLER (Eswar et al. 2007).

Since we do not find any suitable template for other HPs with lyase activity, MODELLER has no role in structure prediction of such proteins. We used threading or fold recognition and ab initio modeling protocol for predicting the structure of such proteins. The threading or fold recognition (Xu et al. 2008) methods were applied for aligning the primary sequence with those proteins present in the Protein Data Bank (PDB) with similar folds. It uses sequence–structure alignment approaches for modeling. We used Sparksx server (Yang et al. 2011) for predicting the structure of P44782, P45267 and Q57498, which uses the SPINE X server for predicting secondary structure, torsion angles and solvent accessibility with higher accuracy. For protein P44093, we used RaptorX (Peng and Xu 2011), as sparksx models are not found suitable for further studies which use NEFF to measure the amount of information in the sequence profile of a protein. However, sparksx along with PSI-BLAST and HHpred is considered to be the best server for predicting the hard targets in Critical Assessment of protein Structure Prediction (CASP) proceedings (Kryshtafovych et al. 2011).

ROBETTA server (Kim et al. 2004) was used for structure prediction of HP P44717, which uses ab initio or de novo methods to predict the structure of proteins whose structural analogs do not exist in the PDB or could not be successfully identified by threading. ROBETTA server generates structure from scratch, since it uses a new alignment method, called K*Sync, to align the query sequence onto the parent structure. Then it models variable regions by allowing them to explore conformational space

with fragments in a fashion similar to the de novo protocol in context of the template. When no structural homolog is available, domains are modeled with the Rosetta de novo protocol (Misura et al. 2006), which allows the full length of the domain to explore conformational space via fragment insertion, producing a large decoy ensemble from which the final models are selected.

We used I-TASSER (Roy et al. 2010) for predicting the structure of HP P44095 that first generates three-dimensional (3D) atomic models from multiple threading alignments and iterative structural assembly simulations. Using the structural matching of the 3D models with other known proteins it inferred function of the protein and the output contains full-length secondary and tertiary structure predictions, and functional annotations on ligand-binding sites, Enzyme Commission numbers and Gene Ontology terms in the results (Roy et al. 2010). Since the unavailability of proper prediction, we used the COACH (Yang et al. 2013) for protein–ligand-binding site prediction and COFACTOR (Roy et al. 2012) for structure-based biological function annotation of HP, for proper function analysis of P44095.

Predicted models were further refined using a side-chain refinement protocol of Discovery Studio 3.5, which uses force fields such as CHARMM (Brooks et al. 2009) and SCWRL4 (Krivov et al. 2009), and predicts positions of the side chains which are used for refinement of predicted protein structures. The final models were further validated by PROCHECK suite (Laskowski et al. 1996), a module of SAVES (Structural Analysis and Verification Server).

Structure analysis

Three-dimensional structure of homologous proteins often remains more conserved than their sequence (Chothia and Lesk 1986). Structural similarities are more reliable than that of sequences for grouping together distant homologues (Taylor and Orengo 1989). Three-dimensional structure of all models was analyzed by various servers for function prediction. Identifying the functional region of protein is an important step towards characterizing its molecular function. We used POCASA (Yu et al. 2010a), Pocket-Finder (Laurie and Jackson 2005) and firestar server (Lopez et al. 2007) for prediction of functionally important residues in HP and the PPM server for calculating rotational as well as translational positions of transmembrane and peripheral proteins in cell membranes (Lomize et al. 2012). It can be applied to newly determine experimental protein structures or theoretical models. Since structural motifs are associated with catalytic functions, motif finding is also essential for precise function prediction (Singh and Saha 2003). We used ProFunc (Laskowski et al. 2005) web server for predicting the function of proteins from its atomic coordinates based on identification of functional motifs. DALI server (Holm and Rosenstrom 2010) was used for structure comparison and fold similarity search. PyMOL (DeLano 2002), a molecular graphics system, is used for visualization of protein structure.

Results and discussion

After the extensive sequence and structure analysis, we found eight HPs with characteristic function, namely, cystathionine-β-synthase, cyclase family protein, carboxymuconolactone decarboxylase, pseudouridine synthase A and C, tagatose-1,6-bisphosphate aldolase and aminodeoxychorismate lyase proteins (Table 2). We successfully predicted and analyzed the three-dimensional structure of HPs with lyase activity, using various computational tools. Final models were further validated to check their stereo-chemical parameters such as bond angle and bond length. All eight models P44782, P44717, P45267, Q57498, P44197, P44095, P44720 and P44093 show significant validation score on SAVES server (http://nihServer.mbi.ucla.edu/SAVES/). Structure analysis helped us to predict the function of each HP precisely (Table 3). Detailed structural analysis outcomes for each protein are described here, separately.

HP P44717

Sub-cellular localization prediction of HP P44717 indicates that this HP is localized in the cytoplasmic membrane. Signal peptide prediction suggests the presence of a signal peptide, and indicates that its translocation occurs through non-classical secretory pathway. HMMTOP (Tusnady and Simon 2001) and TMHMM (Krogh et al. 2001) analyses suggest that HP P44717 consists of four transmembrane helix, and may work as a transporter protein (Table S3). Similarity search revealed its high closeness to the cystathionine-β-synthase (CBS) domain protein. HHpred (Soding et al. 2005) also suggests high similarity with magnesium and cobalt efflux protein CorC (PDB ID: 4HG0) and CBS domain protein (PDB ID: 3LHH) (Table S4). Sequence analyses of P44717 suggest that P44717 is comprised of two domains, CorC/HlyC and CBS domain. By applying the clustering algorithms, we analyzed that HP P44717 is a member of cluster 143846 (CBS domain protein) of SYSTERS (Meinel et al. 2005) and cluster 4144744 (Transporter-associated region) of ProtoNet (Rappoport et al. 2012), and a member of transmembrane family protein. Sequence-based motifs are discovered using InterProScan (Quevillon et al. 2005) and MOTIF tools (Kanehisa 1997). We further observed that its sequence possesses cystathionine-β-synthase motif and

Table 2 List of sequence-based predicted function of HPs with lyase activity and motif discovered using MEME of *H. influenzae* strain Rd KW20

S. no.	Cluster[a]	Uniprot ID	MEME results						MAST function prediction	Consensus[b] function
			Motif 1		Motif 2		Motif 3			
			Start	Site	Start	Site	Start	Site		
1	Cluster 118	P44717	262	GYIESH	199	TQEHYL	405	YGKYKF	UPF0053 protein	Cystathionine-beta-synthase
2	Cluster 187	P44782	152	VKLTPVTGRSHQLRL HMLALGHPILGDKFY	56	FCEPAHRLDMATSGIIVFALSKAADRELK RQFREREPKKHYQAIVWGH	18	YQDNHLCVVNKPSG	Ribosomal large subunit pseudouridine synthase A	RNA pseudouridylate synthase rluA (lyase)
3	Cluster 187	P44197	152	VKLIPHTGRKHXLRXHMK HVFHPIXGDTQY	46	HVFPIHRLDRPTSGVLLFALSSEIANL MCEQFEQKYVQKSYLAVVRGY	6	YQDGFLVAVNKPAG	tRNA pseudouridine synthase C	Ribosomal large subunit pseudouridine synthase C rluC
4	Cluster 114	P45267	335	RVYFERM	14	AISPQI	253	FSQDFM	No result	CYTH-like adenylate cyclase
5	Cluster 196	Q57498	1	MFTDWK	59	TRCESC	21	KQYPKM	No result	Carboxymuconolactone decarboxylase
6	Cluster 131	P44095	128	GIWFPCTW	21	PNHCGTHM	11	TPFSSF	No result	Cyclase family protein
7	Cluster 45	P44093	68	WLKENGCTQFYFKYCST	281	IHNENYIE	151	NLMRLM	No result	D-tagatose-1,6-bisphosphate aldolase
8	Cluster 89	P44720	315	DGSGGH	123	WRKDLENAPH	338	RWYRSQ	UPF0755 protein	Aminodeoxychorismate lyase

[a] CLUSS predicted clusters

[b] Consensus result form on the basis of values present in supplementary Table S2 & S3

Table 3 List of structure-based predicted function and validation of HP with lyase activity in *H. influenzae* strain Rd KW20

S. no.	Uniprot ID	Template	Identity (%)	RMSD	Ramachandran	Proposed function
1	P44717 (robetta model)	Magnesium and cobalt efflux protein CorC, 4HG0 CorC/HlyC transporter-associated domain of a CBS domain protein, 2PLS	23 60	3.400 0.598	99 % (91.7 % core 6.5 % allow 0.8 % gener 1.0 % disall)	CBS domain associated CorC/HlyC transporter
2	P44782 (sparksx model)	Template-pseudouridine synthase RluA, 2I82	60	0.174	99.5 % (91.1 % core 8.4 % allow 0.0 % gener 0.5 % disall)	Pseudouridine synthase RluA
3	P44197 (Modeller structure)	Pseudouridine synthase RluC, 1XPI Pseudouridine synthase RluC, 1V9K	34 33	0.379 0.334	99.5 % (88.2 % core 8.8 % allow 2.5 % gener 0.5 % disall)	Pseudouridine synthase RluC
4	P45267 (Sparksx model)	Putative adenylate cyclase, 2GFG	39	0.918	98.8 % (88.9 % core 9.6 % allow 0.3 % gener 1.2 % disall)	Adenylate cyclase
5	Q57498 (Sparksx model)	Template—Carboxymuconolactone decarboxylase family protein, 1VKE	31	0.300	99 % (94.2 % core 2.9 % allow 1.9 % gener 1.0 % disall)	Carboxymuconolactone decarboxylase
6	P44095 (I-TASSERmodel)	Manganese dependent isatin hydrolase, 4J0N (A) Metal-dependent hydrolase with cyclase activity, 1R61	54 27	2.752 1.139	98.3 % (88.4 % core 9.1 % allow 0.8 % gener 1.7 % disall)	Metal-dependent hydrolase with cyclase activity
7	P44093 (Raptorx model)	Putative tRNA synthase, 3DQQ	56	0.215	99.7 % (93.1 % core 6.6 % allow 0.0 % gener 0.3 % disall)	Putative tRNA synthase
8	P44720 (Modeller)	Predicted aminodeoxychorismate lyase protein, 2R1F	49	0.378	99.7 % (89.8 % core 8.3 % allow 1.6 % gener 0.3 % disall)	Aminodeoxychorismate lyase protein

CBS domain motif. Another motif discovery search tool, MEME suite (Bailey et al. 2009) has detected three motifs, namely, 262′-GYIESH, 199′-TQEHYL and 405′-YGKYKF (Table 2). Based on all these observations, we suggest that HP P44717 may function as cystathionine-β-synthase and present in the inner plasma membrane to work as a transporter for various inorganic salts.

The virulence factor analysis shows that HP P44717 is a non-virulent protein. We further analyzed the functional protein association networks of HP P44717 using an online server STRING. This protein is showing close interaction with thymidylate kinase, 16S ribosomal RNA methyltransferase, virulence-associated protein D, hemolysin, DNA repair protein and GTP-binding protein (Figure S1).

We predicted three-dimensional structure of P44717 using online server Robetta (Kim et al. 2004). The predicted structure was energy minimized, and validated using online server SAVES showing 99 % of residues in the allowed region of the Ramachandran plot (Hooft et al. 1997; Ramachandran et al. 1963). The root-mean-square deviation (rmsd) of the model with respect to the templates 4HG0 and 2PLS was 3.400 and 0.598 $Å^2$, respectively (Table 3), indicating close functionality. Structure comparison and analysis further revealed that HP P44717 shows the presence of three functional domains, namely, CBS domain, cytochrome *c* oxidase polypeptide and CorC_HlyC domain (Fig. 1a). The overall structure of P44717 is comprised of 18 α-helices and nine β-strands including a CBSX subunit that contains two conserved CBS domains which form a pair facing each other, and are connected by two fundamental β-strands (β1–β2 and β3–β5) and bordered by six α-helices. The presence of helix α14 and helix α16 is the main feature of identifying dimeric CBS domain proteins, and both helices are critical for the dimer assembly of CBSX subunit (Fig. 1b). The CorC_HlyC domain contains four β-strands (β6–β9) arranged as antiparallel β-strand and two α-helices (α17 and α18) that contain binding sites for metal ion binding to amino acids Ile413, Asp414, Thr415 (β8) and Asp421 (β9) (Figure S2 A and B). We also detected an AMP-binding site in P44717 structure at Val227 using Firestar server analysis (Figure S2 C). The active site cavity was predicted by PocketFinder (Laurie and Jackson 2005) showing residues Val227, Tyr228, Leu229, Asp230, Phe233, Glu237, Val238, Thr241, Leu242 and Ile250 which are presumably responsible for the lyase activity (Fig. 1c).

In order to find structurally similar proteins in the PDB related to the P44717, we used DALI server (Holm and

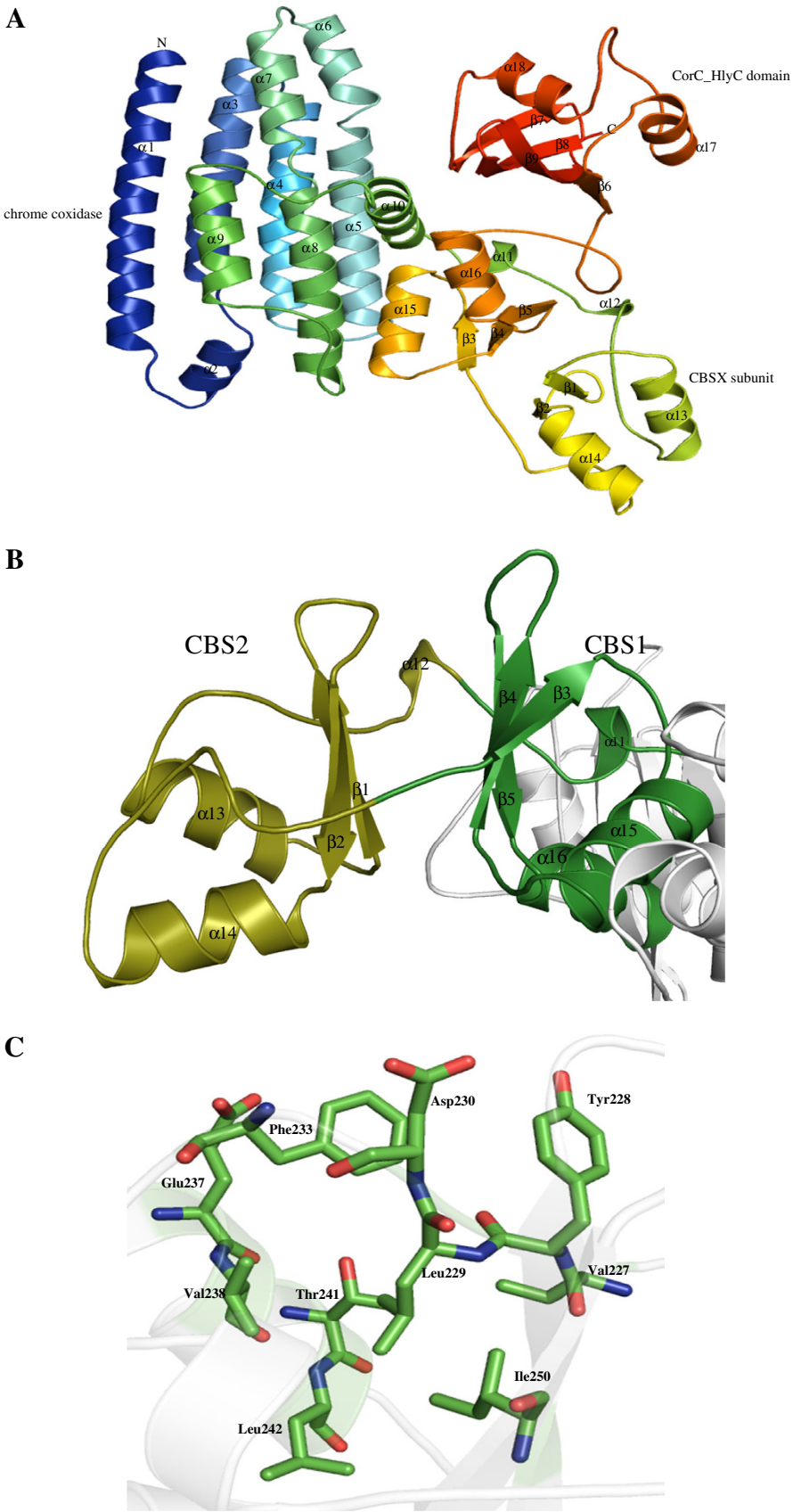
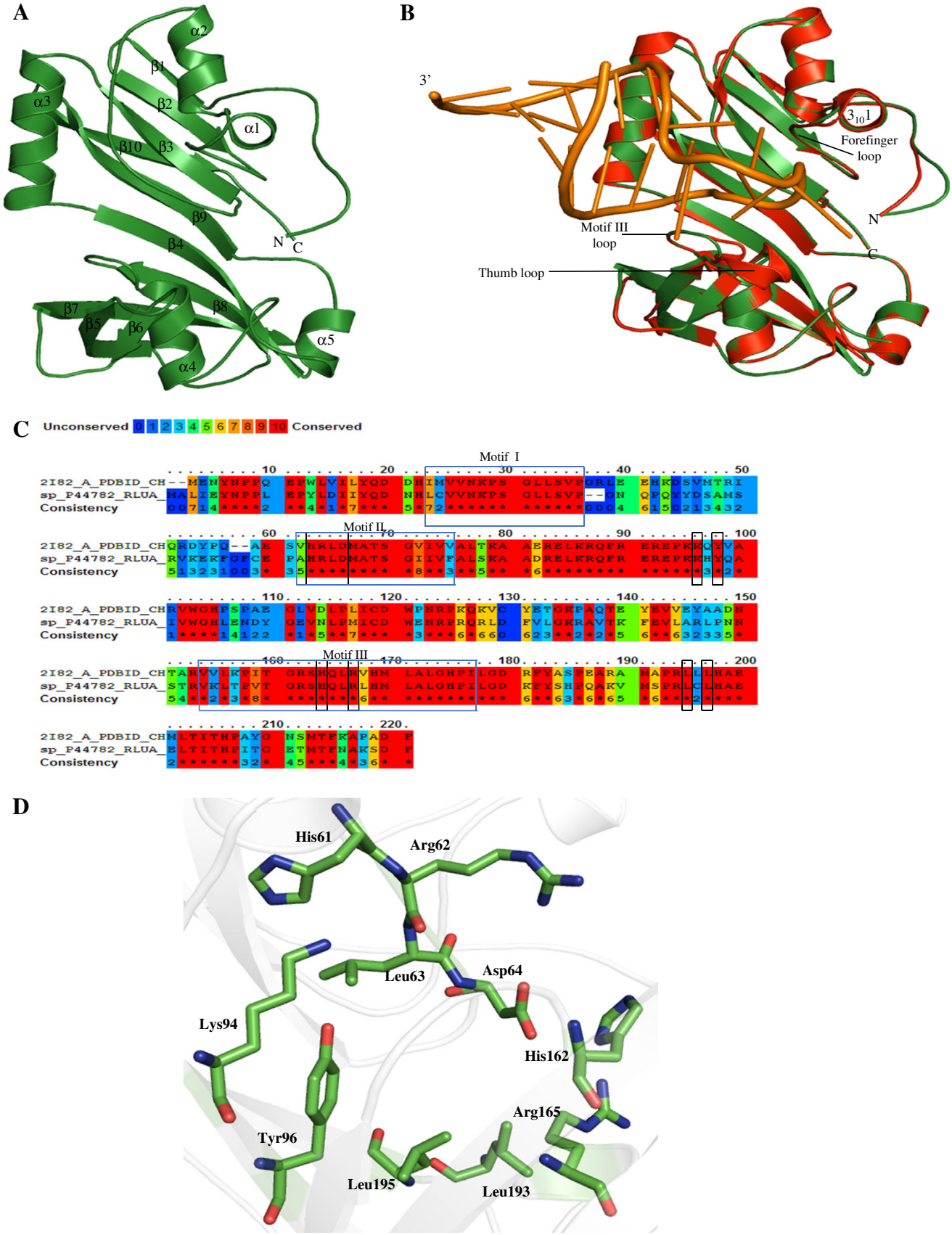


Fig. 1 Representation of model structure of HP P44717. **a** Showing CBS, CorC_HlyC and cytochrome c oxidase domains. **b** Diagram showing the monomer of CBSX with CBS1 and CBS2 domain. **c** Residues present in the active site pocket are illustrated in stick


A
α2
β1
α3
β2
α1
β10
β3
β9
β4
N
C
β7
β5
β6
β8
α5
α4
B
3'
$3_{10}1$
Forefinger loop
N
C
Motif III loop
Thumb loop
C
Unconserved 0 1 2 3 4 5 6 7 8 9 10 Conserved
Motif I
Motif II
Motif III
2I82_A_PDBID_CH
sp_P44782_RLUA_
Consistency
D
His61
Arg62
Leu63
Asp64
Lys94
His162
Arg165
Tyr96
Leu195
Leu193

◀ **Fig. 2** Representation of model structure of HP P44782. **a** Overall structure is represented in cartoon. **b** Superposition of the RluA–RNA complex (PDB ID: 2I82; *red*) with free RluA (P44782; *dark green*). The forefinger and thumb loops that "grasp" the RNA are indicated with motif III loop, and contrast the absence of the thumb loop in P44782 structure. **c** Multiple sequence alignment of P44782 and Rlu A (PDB ID: 2I82) with *red* color indicates residues that are conserved in two proteins. Motifs I, II, and III are as defined by Koonin (1996) highlighted in *blue boxes* and conserved active site residues in *black boxes*. **d** Predicted active site residues shown in stick

Rosenstrom 2010). The results clearly indicate a significant similarity of the CBS domain containing proteins such as magnesium and cobalt efflux protein (Z score = 13.8), hemolysin-like protein (Z score = 13.4), etc., which are among the top hits. In each case, residues at positions 80–150 in the sequence show a close resemblance with rmsd of 0.8–3.2 $Å^2$, despite a sequence similarity of 34 % in the aligned region. We also found a close structural similarity to the magnesium and cobalt efflux protein. Furthermore, ProFunc server (Laskowski et al. 2005) was used to predict the function of the HP on the basis of sequence and structure comparisons. We found 11 motifs in the InterPro database (Hunter et al. 2011) with CorC_HlyC, CBS and hemolysin related sequence motifs. There are two significant ligand-binding templates also present in the P44717 structure. All these analyses strongly suggest that HP P44717 contains CBS domain and it has CorC/HlyC transporter function (Table 3).

The CBS domain, also known as 'Bateman domain', is a conserved domain present in prokaryotes and eukaryotes (Ignoul and Eggermont 2005). Structure analyses of bacterial CBS domains show that two CBS domains form an intramolecular dimeric structure (CBS pair). Human hereditary diseases, such as hypertrophic cardiomyopathy, homocystinuria, myotonia congenital, retinitis pigmentosa, etc., can be caused by mutations in CBS domains of cystathionine-β-synthase, inosine 5′-monophosphate dehydrogenase, AMP kinase, and chloride channels, respectively, which also affect multimerization and sorting of proteins, channel gating, and ligand binding (Bateman 1997). Recent experiments show that CBS domains can bind adenosine-containing ligands such as ATP, AMP, or *S*-adenosylmethionine, which indicate that CBS domains may function as sensors of intracellular metabolites (Bateman 1997).

HP P44782

HP P44782 is localized in the cytoplasm and lacks signal peptide, therefore, it is not secreted (Table S3). The sequence analysis showed a significant similarity of HP P44782 with 23S rRNA/tRNA pseudouridine synthase A (Table S1 and S2), which was further confirmed by domain annotation and cluster analysis using online tool SYSTERS and ProtoNet. Furthermore, a pseudouridine synthase, RsuA/RluB/C/D/E/F motif, and pseudouridine synthase, the RluA motif, are present in the sequence of HP P44782. MEME suite has further confirmed three distinct motifs, namely, 152′-VKLTPVTGRSHQLRLHMLALGHPILGD KFY, 56′-FCEPAHRLDMATSGIIVFALSKAADRELKR QFREREPKKHYQAIVWGH, 18′-YQDNHLCVVNKPSG (Table 2), an indication of ribosomal large subunit pseudouridine synthase A. Interacting partners of HP P44782 are ATP-dependent helicase, tRNA pseudouridine synthase B, tRNA pseudouridine synthase A, tRNA-dihydrouridine synthase A, glp protein, lipoprotein signal peptidase, glycerol-3-phosphate regulon repressor, 16S pseudouridylate 516 synthase, *S*-adenosyl-L-methionine-dependent methyltransferase (Figure S1) indicating its role in the pseudouridine synthesis.

Sparksx predicted the structure of P44782 using ribosomal large subunit pseudouridine synthase A (PDBID: 2I82), a lyase, as a template. The final model was validated on PROCHECK (Laskowski et al. 1996) which shows that 99.5 % residues are in the allowed region of Ramachandran plot. Model structure showed rmsd of 0.174 $Å^2$ from the template. The overall structure of P44782 is elongated and is comprised of mixed α/β fold (Fig. 2a, b). The structure is a conserved Ψ synthase fold, with an eight-stranded β sheet core, and additional strand extends the core β sheet which is also encircled by helices and loops on one face. ASL-RNA binds with its helical axis nearly parallel to the plane of the β sheet core of the predicted P44782 model (Fig. 2b). We found three motifs in sequence of HP P44782 (Fig. 2c) (Koonin 1996). Motif II is the part of the active site while motif I plays an architectural role by supporting motif II (Hoang and Ferre-D'Amare 2001, 2004). The superimposition of structures of P44782 and pseudouridine synthase RluA (PDB code: 2I82) indicates that C-terminal side of the motif I forms a projection that packs against minor groove face of the anticodon loop. N-terminal half of motif III (conserved among Ψ synthases of the RluA and RsuA families) interacts with the RNA backbone on the 5′- side of the ASL (Fig. 2b). The "thumb loop" and "forefinger loop" are found in pseudouridine synthase RluA (Hoang et al. 2006). Structure alignment showed the presence of forefinger loop but not the thumb loop in the HP P44782. The forefinger loop makes RluA enzyme bind with the minor groove of the substrate RNA. Previously determined crystal structure shows that families of pseudouridine synthases have universal conservation in the active sites as an aspartate, a basic residue, and a tyrosine (Del Campo et al. 2004; Koonin 1996; Ramamurthy et al. 1999). Amino acids corresponding to Asp64, Tyr96, and Arg165 are conserved in all ψ synthases, and the active site of RluA (Hoang et al. 2006) is comprised of residues His 61, Arg 62, Leu 63, Asp64, Lys94, Tyr96, His162, Arg165, Leu193 and Leu195 (Fig. 2c, d).

Structure similarity search using DALI server provides many significant hits for P44782 having ribosomal large subunit pseudouridine synthase activity, with ribosomal large subunit pseudouridine synthase C (Z score = 27.3), ribosomal large subunit pseudouridine synthase F (Z score = 12.8), ribosomal large subunit pseudouridine synthase E (Z score = 12.7), etc., among the top hits. We observe an rmsd of 0.4–3.0 $Å^2$ for residues 180–250 in each match with 60 % sequence identity. Structure of P44782 was further analyzed using ProFunc server showing five sequence pseudouridine synthase motif and six significant ligand-binding templates. The ProFunc and Dali server searches clearly indicate the presence of pseudouridylate synthase activity in the HP P44782 (Table S4).

The isomerization of uridine to pseudouridine (Ψ) within RNA, discovered often in highly conserved locations in ribosomal and transfer RNA, is a ubiquitous process reported to be present in early stages in the evolution of life (Ofengand et al. 2001). Pseudouridine (Ψ) synthases are enzymes that catalyze site-specific isomerization of uridine residues in cellular RNAs. Ψ synthases can be classified into five families, namely, RluA, RsuA, TruA, TruB, and TruD on the basis of amino acid sequence (Kaya and Ofengand 2003), and structure analyses revealed that in spite of minimal sequence similarity, the cores of all Ψ synthases adopt the same fold (Ferre-D'Amare 2003; Foster et al. 2000; Sivaraman et al. 2002) and enzymatic core includes an aspartate residue that is conserved among all five families of Ψ synthases. On site-directed mutagenesis this Asp was found to be essential for catalytic activity of such enzyme (Conrad et al. 1999; Ramamurthy et al. 1999).

HP P44197

HP P44197 is localized in the cytoplasm, and lacks signal peptide and transmembrane helix (Table S3). This protein showed high sequence similarity to tRNA pseudouridine synthase C. The domain annotation shows that P44197 is a member of pseudouridine synthase RsuA/RluD family, showing pseudouridylate synthase activity (Table S2). Inter ProScan and MOTIF tools further confirmed a pseudouridine synthase motif in the HP P44197 (Table 2). Predicted functional partners of HP P44197 are exodeoxyribonuclease VII small subunit, exodeoxyribonuclease VII small subunit, 1-deoxy-D-xylulose-5-phosphate synthase, aspartate-semialdehyde dehydrogenase, tRNA pseudouridine synthase B, and tRNA pseudouridine synthase A which provide an insight of pseudouridine synthase activity for this HP (Figure S1).

Ribosomal large subunit pseudouridine synthase C (PDB ID: 1V9K) is a lyase used as template for predicting the structure of HP. Ramachandran plots show 99.5 % of residues in allowed region. An rmsd of 0.379 $Å^2$ with template was observed (Table 3). The predicted structure of P44197 is also an elongated, mixed α/β fold and possesses six α-helices and 10 β-strands (Fig. 3a), with an eight-stranded β sheet core. This structure is also a conserved Ψ synthase fold. Two sequence motifs (motif I and II) are present in the sequence of P44197 (Fig. 3b) (Koonin 1996), where motif II is the part of the active site, and architectural role was played by motif I (Koonin 1996). The active site of P44197 includes amino acid residues His51, Arg52, Leu53, Asp54, Lys84, Tyr86, His162, Arg164, Leu199, and Leu201 (Fig. 2c), in which Asp54 and Arg164 are conserved in the active site and mainly responsible for the catalytic activity (Del Campo et al. 2004).

The structure of HP P44197 was further analyzed using ProFunc and Dali servers. This analysis revealed a close similarity to the ribosomal large subunit pseudouridine synthase C (Z score = 34.5), ribosomal large subunit pseudouridine synthase D, (Z score = 26.7), etc. The best match showed 35 % sequence identity over aligned 217 residues. ProFunc server finds five pseudouridine synthase signature sequence, six ligand-binding templates and one DNA-binding template in the HP P44197. All these observations strongly suggest pseudouridylate synthase C activity in this HP (Table 3).

Similarly, RluC and RluD are homologous enzymes which convert three specific uridine bases to pseudouridine in *E. coli* ribosomal 23S RNA to pseudouridine, namely, bases 955, 2504, and 2580 in the case of RluC and 1911, 1915, and 1917 in the case of RluD and both contain N-terminal S4 RNA-binding domain (Mizutani et al. 2004). A loss of RluD which acts as an RNA chaperone is important for ribosome assembly. The loss of RluC results in reduced growth rate, and has no significant effect of impairing growth (Mizutani et al. 2004).

HP P45267

The sequence analysis of HP P45267 showed that it is localized in the cytoplasm and is a non-secretary protein (Table S3). This protein showed a significant sequence similarity to the adenylate cyclase, which was further confirmed by domain analysis which suggested the presence of adenylate cyclase-like domain. The cluster analysis also suggests that HP P45267 belongs to the cluster of protein involved in the cAMP biosynthetic process. Analyzing SVMProt suggests that the HP P45267 is a member of zinc-binding protein and belongs to the DNA-binding protein family. Motif search showed the presence of CYTH-like phosphatases and adenylate cyclase motifs (Table 2). This is a virulent protein, which may be involved in the cellular process. The interacting partners of HP P45267 are Hsf-like protein, phosphatase, glutamate-ammonia-ligase, adenylyl transferase, DNA repair protein,

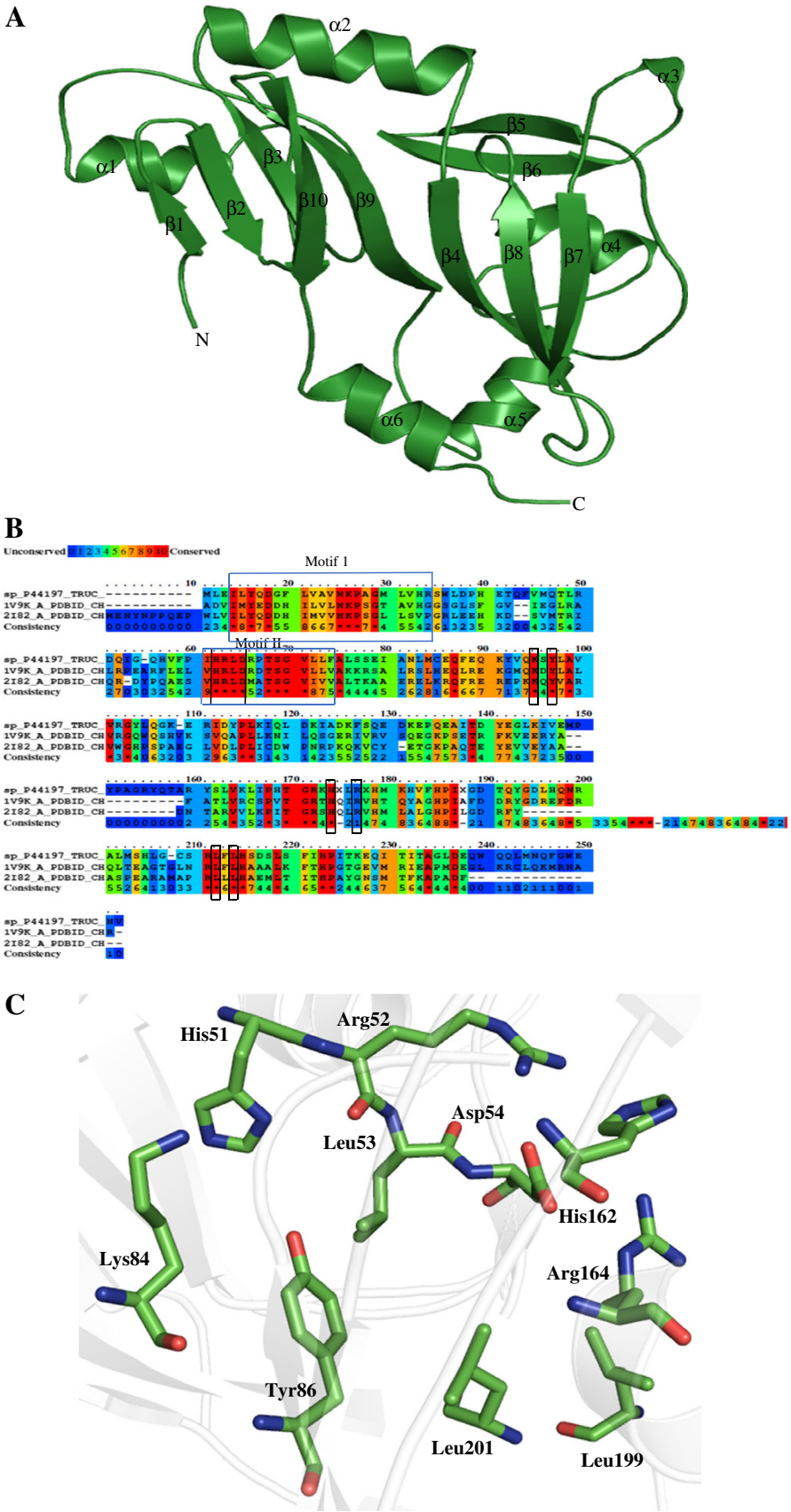


Fig. 3 Representation of model structure of HP P44197. **a** Cartoon model showing overall structure. **b** Multiple sequence alignment showing conserved residues in *red color*, with a catalytic domain of P44197, Rlu C and *E. coli* Rlu A (PDB ID: 2I82) distinguished residues in *black box* and motifs I and II in *blue box*. **c** A detailed description of P44197 active site in which Asp54 is surrounded by arginine residues from both the sides namely Arg52 and Arg164

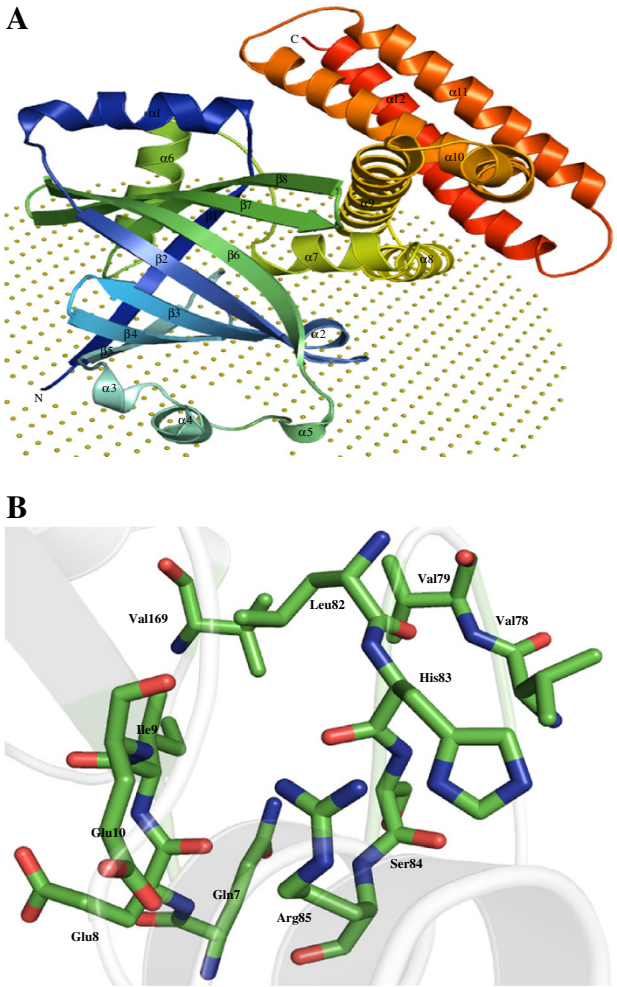


Fig. 4 Representation of model structure of HP P45267. **a** Cartoon diagram showing N-terminal part is shown in *blue* while C-terminal in *red*. Collection of non-bonded spheres represents membrane. **b** Stick representation of P45267 active site with Glu10 residue is proposed to be involve in cAMP binding

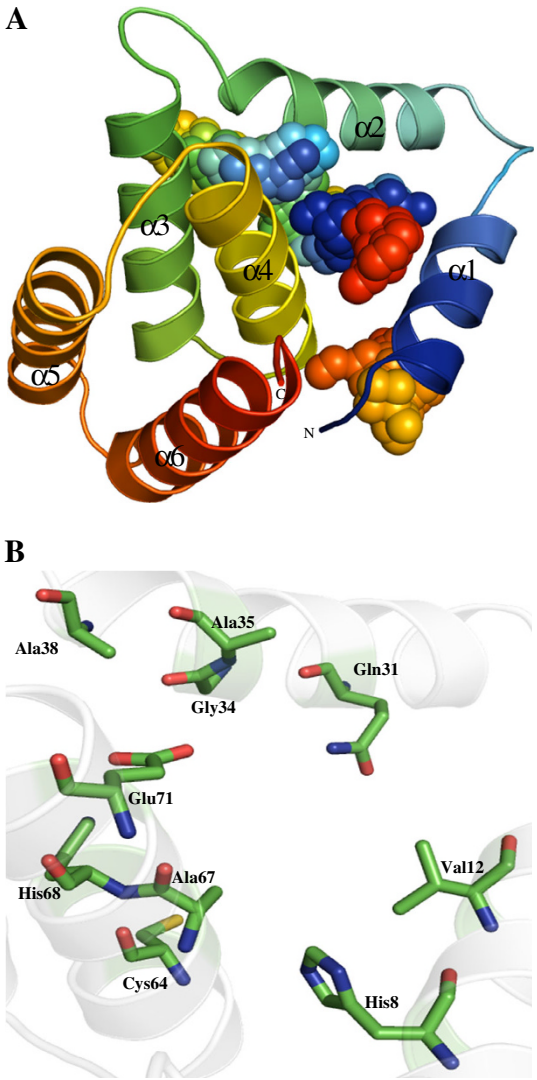


Fig. 5 Representation of model structure of HP Q57498. **a** Cartoon model showing overall structure with active site residues. **b** A description of active site with His68 residue represents the citrate-binding site

outer membrane protein P2 (Figure S1). All these findings strongly suggest that the HP P45267 possesses CYTH-like adenylate cyclase activity and play significant roles in the survival of *H. influenzae,* and may be considered as a potential drug target (Carbonetti 2010).

Structure of P45267 was predicted by fold recognition algorithm on the sparksx server using adenylate cyclase (PDB ID: 2GFG) as template. PROCHECK reports that 98.8 % of residues are present in the allowed region of the Ramachandran plot. Overall structure contains 12 α-helices and 8 β-strands. Interestingly, eight stranded anti-parallel barrels (β1–β8) surrounded by nine long α-helices and three small helical regions fold were adopted by this protein (Fig. 4a) which looks like a cup or bottle with helical handles on both sides. The structure analysis shows that residues Met1, Val79, Pro108 and Phe109 are attached to the membrane. We found three AMP-binding residues Glu10, Arg62 and Arg127 in the P45267 (Figure S3). The active site may contain residues Gln7, Glu8, Ile9, Glu10, Val78, Val79, Leu82, His83, Ser84, Arg85, and Val169 (Fig. 4b).

The structure analysis results showed the presence of adenylate activity in this HP due to the high structural resemblance with the CYTH-like phosphatase (Z score = 34.5), adenylate cyclase 2 (Z score = 13.5), etc. We also found a similar search in ProFunc analysis, which shows the presence of 3 CYTH-like phosphatase motifs in this HP. These observations suggest that the HP P45267 may be a CYTH-like phosphatase and adenylate cyclase enzyme, and it is essential for pathogenesis (Table 3).

Adenylyl cyclase (AC, EC 4.6.1.1) catalyzes the cyclization of ATP to form cyclic AMP (cAMP), and is an important signaling molecule present in most of cells (Taussig and Gilman 1995). The liver cells produce cAMP, which is responsible for phosphorylase activation in response to extracellular adrenaline. It is found that adenylyl cyclase is involved in a wide variety of signal transduction mechanisms in different cell types (Sutherland 1972). In prokaryotes, expression of metabolic pathways such as the lactose operon, which are otherwise inactive when glucose is abundant, are triggered by cAMP (Black et al. 1980), and in eukaryotes, cAMP is involved in hormonal signal transduction or transmits stimulus from extracellular receptors to activate cell-specific mechanisms (Sutherland 1972). Danchin (1993) identified six distinct non-homologous classes of AC.

HP Q57498

We obtained variable results for sub-cellular localization of HP Q57498. PSLpred (Bhasin et al. 2005) analysis revealed that this protein is localized in the periplasmic region. However, CELLO (Yu et al. 2006) tool suggests that it is a cytoplasmic protein. This is a non-secreted protein and shows the absence of transmembrane helix (Table S3). Similarity search shows a significant similarity to the alkyl hydroperoxidase (AhpD) core protein and carboxymuconolactone decarboxylase (CMD). Domain annotation and cluster analysis further confirmed that this protein belongs to a CMD family protein. Motif discovery shows that this HP contains a typical CMD motif, which was further confirmed by a motif search on MEME suite (Table 2). This HP is found to be a virulent protein on the basis VirulentPred. This structural information of this protein may be utilized for drug design and discovery. This HP interacts with transcriptional regulator, autonomous glycyl radical cofactor GrcA, dihydrolipoamide dehydrogenase, and type III restriction–modification system endonuclease-like protein. This indicates the functional significance of HP Q57498 (Figure S1).

Sparksx server (Yang et al. 2011) modeled the sequence of HP Q57498 where CMD proteins (PDB ID: 1VKE and 2QEU) were used as a template and 99 % residues in allowed region of the Ramachandran plot. The Q57498 predicted structure is showing all α-helix topology (six α-helices), with active cavities present between α1, α2 and α4 (Fig. 5a). We found citrate-binding sites in Q57498 which include residues corresponding Tyr30 and His68 (Figure S4). His68 is also present in the predicted active cavity along with His8, Val12, Gln31, Gly34, Ala35, Ala38, Cys64, Ala67, His68, and Glu71 (Fig. 5b). The ProFunc and Dali server searches, further used to validate the predicted function for Q57498, indicate the CMD activity in the HP. The Dali search mainly includes gamma-CMD (Z score = 11.6), CMD family protein (Z score = 10.7), etc., which are in top hits, showing the overall rmsd range of 0.7–3.1 $Å^2$ over 97–133 residues. ProFunc validated the Dali results and showed the presence of five CMD and AhpD-like sequence motifs. ProFunc and Dali servers have also confirmed the structural similarity of HP Q57498 with that of CMD (Table 3).

The 3-oxoadipate pathway branches, namely catechol and protocatechuate, cause degradation of aromatic compounds. This pathway is important for the bacteria, unites at the common intermediate 3-oxoadipate enol-lactone, and the enzyme, CMD is involved in protocatechuate catabolism and gene fusion event leads to expression of CMD in bacteria (Eulberg et al. 1998).

HP P44095

The HP P44095 is localized in the cytoplasm with no signal peptide and transmembrane helix (Table S3). Similarity search and sequence analysis show that HP P44095 is highly similar to the protein of cyclase family and metal-dependent hydrolase with cyclase activity (Table S2). Furthermore, cluster analysis shows that this protein belongs to the cyclase activity proteins containing clusters. Motif analysis tools have further predicted a putative cyclase motif and glucose transporter type 3 (GLUT3) signatures, respectively, in the HP P44095. All these analyses led to a conclusion that HP P44095 presumably works as a protein with cyclase activity. String analysis suggests various proteins such as gluconate permease, 3-hydroxyisobutyrate dehydrogenase, putative aldolase, and glycerol-3-phosphate regulon as a functional networking partner of the HP P44095 (Figure S1).

We used I-TASSER structure prediction server to compute the structure of P44095 using metal-dependent hydrolase with cyclase activity (PDB ID: 1R61) and manganese-dependent isatin hydrolase (PDB ID: 4J0N) as templates. A final model was validated with PROCHECK which shows that 98.3 % residues are in the allowed region of the Ramachandran plot. P44095 adopts mixed α/β fold with five β-strands and six α-helices (Fig. 6a). Residues Glu57, Pro60, Pro74, Tyr75, Ala76, Ala77, and Ile78 are responsible for membrane attachment in P44095. The structure and firestar server, COACH and COFACTOR analyses show the presence of three zinc-binding sites present at His23, His27, and Asp29 (Figure S5). PocketFinder predicts the active site residues as Met1, His23, Cys24 and Gly25 (Fig. 6b). We further analyzed this HP using ProFunc and Dali servers. Results showed a close structural similarity to the hydrolase (Z score = 13.9) and alpha amylase (Z score = 3.2). Furthermore, ProFunc found three cyclase motifs. ProFunc and Dali servers

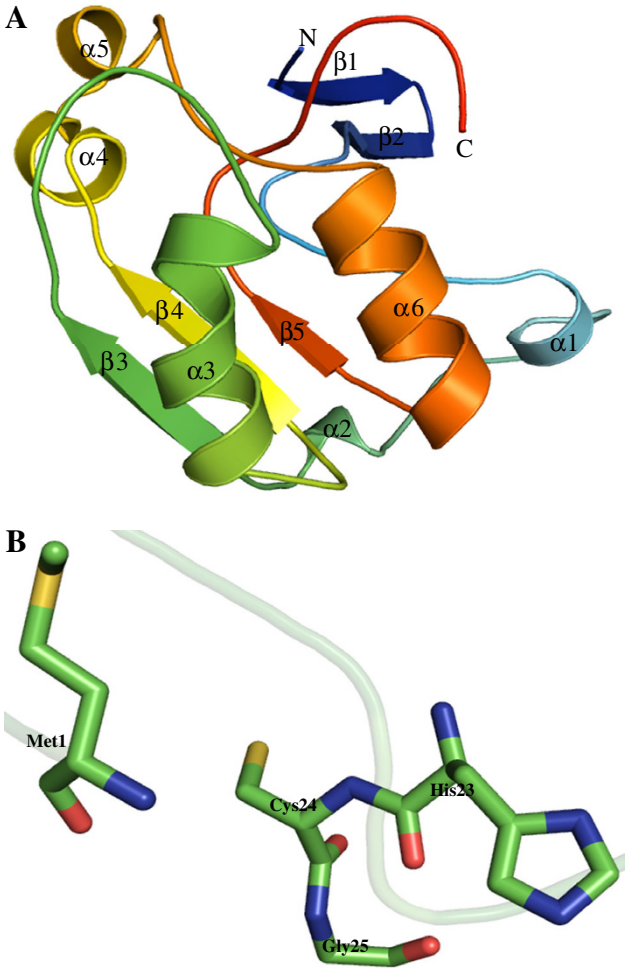


Fig. 6 Representation of model structure of HP P44095. **a** Overall structure of P44095 shown in cartoon model with N-terminal is shown in *blue* and C-terminal in *red*. **b** Representation of the active site residues of P44095 in stick model

suggest that HP P44095 belongs to the cyclase family and possesses metal-dependent hydrolase activity (Table S4).

HP P44093

HP P44093 shows cytoplasmic localization devoid of any transmembrane helix as well as the signal peptide, hence it may not be secreted (Table S3). This protein shows high sequence similarity to the 4-hydroxy-3-methyl but-2-enyl diphosphatereductase. We are unable to find any significant hit in most of the searches performed in various databases. However, SUPERFAMILY (Gough et al. 2001) shows that HP P44093 is a member of YgbK-like family, and PANTHER (Mi et al. 2005) predicts that this HP belongs to D-tagatose-1,6-bisphosphate aldolase family. This protein belongs to the cluster of proteins with ygbK-like activity. The classification system SVMProt (Cai et al. 2003) search shows that this HP belongs to the carbon–carbon lyases family. Motif finding tools suggest the presence of D-tagatose-1,6-bisphosphate aldolase motif and glycosyl hydrolases family 1 active site motifs in the HP P44093. Interaction partners of HP P44093 are aldolase, 3-hydroxyisobutyrate dehydrogenase, glycerol-3-phosphate regulon repressor, gluconate permease, L-fuculose phosphate aldolase, and D-xylose transporter subunit XylF indicating this protein plays a significant role in the carbohydrate metabolism (Figure S1).

We used RaptorX sever for predicting the structure of P44093 using putative tRNA synthase (PDB ID: 3DQQ) as a template (Table 3) and found that 99.7 % residues in allowed region of the Ramachandran plot. The structure of P44093 contains 17 β-strands and 18 α-helices, and attains α/β fold with N-terminus covering one end and the C-terminus the other in a mixed parallel and anti-parallel fold (Fig. 7a). The residues Pro54, Asn56, Lys90, Ile281, Leu315, Ile318, Gln319, His320, Gln321, and Phe322 are membrane-attaching residues in P44093 (Fig. 7b). The firestar server was unable to detect any ligand-binding sites in P44093, while COACH and COFACTOR consensus show Ile7, Asn8 and Ile10 may be magnesium- and beta-D-glucose-binding sites. The P44093 also shows a conserved dihydroxyacetone phosphate-binding site (Figure S6), and the amino acid residues Pro243, Thr244, Thr248, Asn349, Trp382, Ala408, Gln409 and Phe412 may form the active cavity of the protein as suggested by the Pocket-Finder. Results of ProFunc and Dali servers show a close resemblance to the formimidoylglutamase (Z score = 7.2) and arginase-1 (Z score = 7.2) with overall rmsd of 1.5–4.0 $Å^2$. Furthermore, three motifs with D-tagatose-1,6-bisphosphate aldolase were predicted in the sequence of HP P44093. Function assignment by ProFunc and Dali servers confirms the previously annotated results that the given HP is a D-tagatose-1,6-bisphosphate aldolase and tRNA synthase, respectively (Table S4).

It is known that tagatose-1,6-bisphosphate aldolase (TBPA), a tetrameric class II aldolase, catalyzes the reversible condensation of dihydroxyacetone phosphate with glyceraldehyde 3-phosphate to produce tagatose 1,6-bisphosphate. A comparison of TBPA with related fructose-1,6-bisphosphate aldolase (FBPA) shows common features associated with the mechanism. Major products of these enzymes catalyzed reactions differ in the chirality at a single position (Hall et al. 2002).

HP P44720

Sub-cellular localization prediction with Psortb and PSLpred shows that HP P44720 is localized in cytoplasm, while CELLO suggests a periplasmic localization.

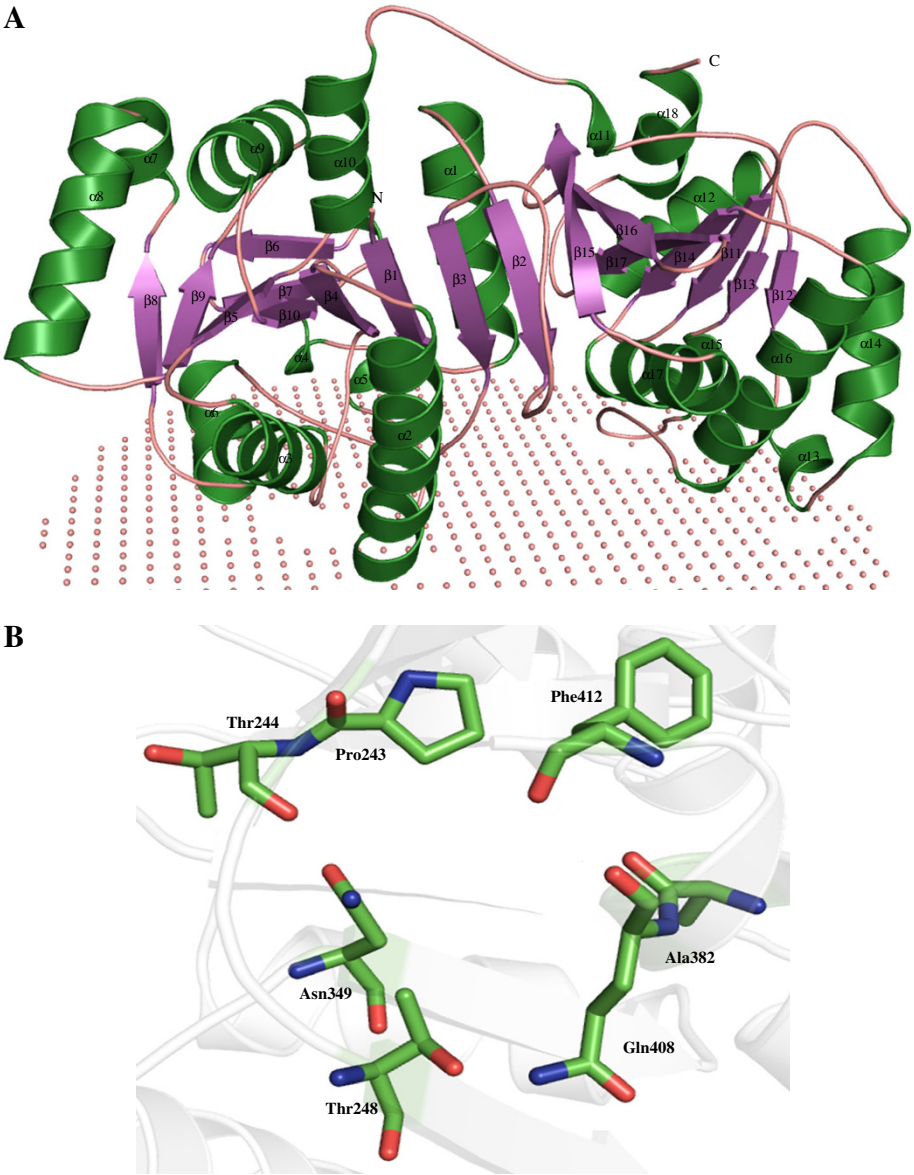


Fig. 7 Representation of model structure of HP P44093. **a** Three-dimensional structure represented in cartoon model with membrane represented as non-bonded spheres. **b** Representation of the active site residues of P44093 in stick model

Similarity search for HP P44720 shows high sequence similarity to aminodeoxychorismate lyase (Table S1 and S2). SUPERFAMILY shows that this HP is a member of zinc finger protein 425-like domain family. The cluster analysis using SYSTERS and ProtoNet shows that the protein belongs to the cluster of proteins with aminodeoxychorismate lyase activity. The motif discovery suggested the presence of zinc finger C2H2-type/integrase DNA-binding domain motif and aminodeoxychorismate lyase active site motif, respectively. The interaction partners of HP P44720 are DNA polymerase III subunit delta, thymidylate kinase, β-hexosaminidase, protease, Holliday junction resolvase-like protein, 3-oxoacyl-(acyl carrier protein) synthase I, acyl carrier protein indicating this protein plays a significant role in the pathogen metabolic system (Figure S1).

Structure of P44720 was predicted using amino deoxychorismate lyase protein (PDB ID: 2R1F) as a template (Table 3). The energy minimized model shows 99.7 % residues in the allowed region of the Ramachandran plot. The structure of P44720 shows six β-strands and 12 α-helices, and it assumes α/β fold with N-terminus covering one end and the C-terminus the other in a parallel fold (Fig. 8a). The membrane-attached residues in P44720 is Lys2, Phe4, Leu5, Ile6, Ala7 and Ile8 (Fig. 8b). Since Lys and Glu residues are important for catalytic activity of aminodeoxychorismate lyase (Nakai et al. 2000), we assume that residues corresponding to Tyr168, Pro169, Met191, Lys197, Ala198, Pro293, Ser294, Glu295, Leu298, Gln299, and Ala302 may form the active site cavity as suggested by the Pocket-Finder. Both ProFunc and Dali server searches suggested the presence of amino

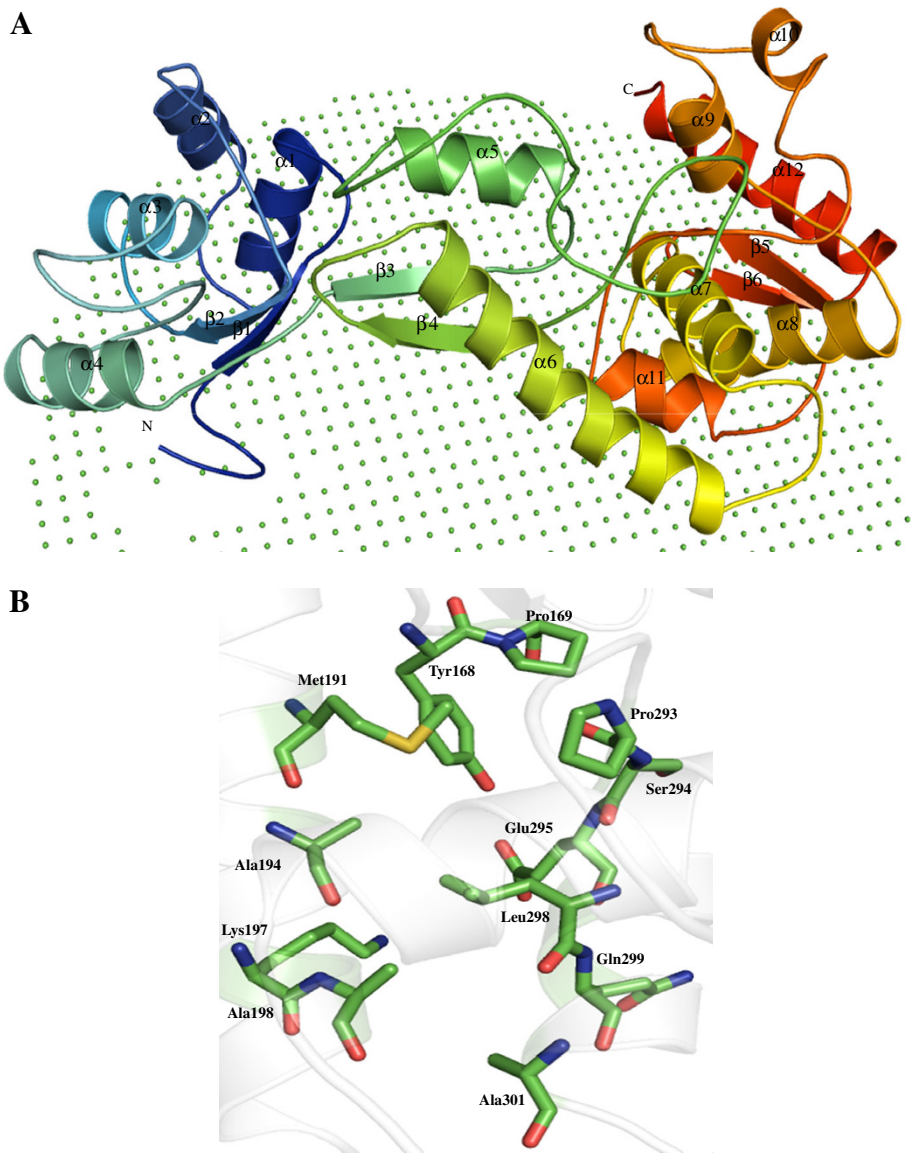


Fig. 8 Representation of model structure of HP P44720. **a** Cartoon model representation of overall structure in which membrane is represented by non-bonded atoms. **b** Representation of the active site residues of P44720 in stick model

deoxychorismate lyase in this HP. Moreover, ProFunc shows seven motifs in the HP P44720. Function allocation by ProFunc and Dali servers further confirms that given HP is an amino deoxychorismate lyase (Table S4).

Chorismate is important in the biosynthesis of many important aromatic products such as anthranilate, prephenate, *p*-aminobenzoate and *p*-hydroxybenzoate, since it acts as a branch point precursor for the intermediary metabolites in bacteria (Green and Nichols 1991). 4-amino-4-deoxychorismate (ADC), formed from chorismate by the action of the enzyme p-aminobenzoate synthase, encoded by pabA and pabB, the product of pabC, is amino deoxychorismate lyase (ADCL) with pyridoxal 5′-phosphate (PLP) as a cofactor, converts ADC to *p*-aminobenzoate and pyruvate (Nichols et al. 1989; Ye et al. 1990), and along with *p*-aminobenzoate synthase is a key enzyme in the biosynthesis of *p*-aminobenzoate.

Conclusions

Our study combines a number of bioinformatics tools for function predictions of previously not assigned proteins in the *H. influenzae* genome. We stressed that besides sequence analysis, structure can be used as a framework to explain known functional properties. Here we have combined the latest versions of several protein databases, protein motifs, features from the amino acid sequence, structure prediction, structure analysis, structure–function relationship, as well as a pathway and genome context methods to assign a precise function to hypothetical proteins for which no any experimental information is available. We used the sequence and structure-based methods for functional annotation of HPs from *H. influenzae* with lyase activity. The structure of these enzymes was predicted by most of the available methods, but we chose

those models with best quality. HPs P44717, P44782, P44197, P45267, Q57498 and P44720 are categorized as lyase enzyme which was further confirmed by structure prediction and analysis. We also observed the reliability of different methods for predicting the three-dimensional structure of HPs and found that homology and fold recognition methods were only applicable to highly evolutionary conserved proteins and ab initio servers, especially I-TASSER gave good result for P44095 which is not predicted by other methods very precisely. Our in silico approach may be combined with experimental methods and make a leading contribution in functional elucidation of the protein structures.

Acknowledgments Authors sincerely thank Indian Council of Medical Research for financial assistance (Project No. BIC/12(04)/2012).

Conflict of interest Authors declare no conflict of interest.

References

Accelrys (2013) Discovery studio modeling environment, Release 3.5. Accelrys Software Inc., San Diego

Alexandrov A, Martzen MR, Phizicky EM (2002) Two proteins that form a complex are required for 7-methylguanosine modification of yeast tRNA. RNA 8:1253–1266

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Apisarnthanarak A, Mundy LM (2005) Etiology of community-acquired pneumonia. Clin Chest Med 26:47–55

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208

Baron C, Coombes B (2007) Targeting bacterial secretion systems: benefits of disarmament in the microcosm. Infect Disord Drug Targets 7:19–27

Bateman A (1997) The structure of a domain common to archaebacteria and the homocystinuria disease protein. Trends Biochem Sci 22:12–13

Bendtsen JD, Kiemer L, Fausboll A, Brunak S (2005) Non-classical protein secretion in bacteria. BMC Microbiol 5:58

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1978) The Protein Data Bank: a computer-based archival file for macromolecular structures. Arch Biochem Biophys 185:584–591

Bhasin M, Garg A, Raghava GP (2005) PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 21:2522–2524

Bjornson HS (1984) Enzymes associated with the survival and virulence of gram-negative anaerobes. Rev Infect Dis 6(Suppl 1):S21–S24

Black RA, Hobson AC, Adler J (1980) Involvement of cyclic GMP in intracellular signaling in the chemotactic response of *Escherichia coli*. Proc Natl Acad Sci USA 77:3879–3883

Britton KL, Abeysinghe IS, Baker PJ, Barynin V, Diehl P, Langridge SJ, McFadden BA, Sedelnikova SE, Stillman TJ, Weeradechapon K et al (2001) The structure and domain organization of *Escherichia coli* isocitrate lyase. Acta Crystallogr D Biol Crystallogr 57:1209–1218

Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S et al (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614

Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31:3692–3697

Capdeville R, Buchdunger E, Zimmermann J, Matter A (2002) Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. Nat Rev Drug Discov 1:493–502

Carbonetti NH (2010) Pertussis toxin and adenylate cyclase toxin: key virulence factors of *Bordetella pertussis* and cell biology tools. Future Microbiol 5:455–469

Chen R, Jeong SS (2000) Functional prediction: identification of protein orthologs and paralogs. Protein Sci 9:2344–2353

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Christensen H (2008) Taxonomy and biodiversity of members of Pasteurellaceae. Academic Press, Caister

Conrad J, Niu L, Rudd K, Lane BG, Ofengand J (1999) 16S ribosomal RNA pseudouridine synthase RsuA of *Escherichia coli*: deletion, mutation of the conserved Asp102 residue, and sequence comparison among all other pseudouridine synthases. RNA 5:751–763

Danchin A (1993) Phylogeny of adenylyl cyclases. Adv Second Messenger Phosphoprotein Res 27:109–162

Del Campo M, Ofengand J, Malhotra A (2004) Crystal structure of the catalytic domain of RluD, the only rRNA pseudouridine synthase required for normal growth of *Escherichia coli*. RNA 10:231–239

DeLano WL (2002) The PyMOL molecular graphics system. In: Schrödinger L (ed) DeLano Scientific, San Carlos, CA, USA.

Desler C, Durhuus JA, Rasmussen LJ (2012) Genome-wide screens for expressed hypothetical proteins. Methods Mol Biol 815:25–38

Dobson CM (2003) Protein folding and misfolding. Nature 426:884–890

Eisenreich W, Rohdich F, Bacher A (2001) Deoxyxylulose phosphate pathway to terpenoids. Trends Plant Sci 6:78–84

Eisenreich W, Bacher A, Arigoni D, Rohdich F (2004) Biosynthesis of isoprenoids via the non-mevalonate pathway. Cell Mol Life Sci 61:1401–1426

Ejim LJ, D'Costa VM, Elowe NH, Loredo-Osti JC, Malo D, Wright GD (2004) Cystathionine beta-lyase is important for virulence of *Salmonella enterica serovar typhimurium*. Infect Immun 72:3310–3314

Eldika N, Sethi S (2006) Role of nontypeable *Haemophilus influenzae* in exacerbations and progression of chronic obstructive pulmonary disease. Curr Opin Pulm Med 12:118–124

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2007) Comparative protein structure

modeling using MODELLER. Curr Protoc Protein Sci Chapter 2: Unit 2.9

Eulberg D, Lakner S, Golovleva LA, Schlomann M (1998) Characterization of a protocatechuate catabolic gene cluster from *Rhodococcus opacus* 1CP: evidence for a merged enzyme with 4-carboxymuconolactone-decarboxylating and 3-oxoadipate enol-lactone-hydrolyzing activity. J Bacteriol 180:1072–1081

Ferre-D'Amare AR (2003) RNA-modifying enzymes. Curr Opin Struct Biol 13:49–55

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Foster PG, Huang L, Santi DV, Stroud RM (2000) The structural basis for tRNA recognition and pseudouridine formation by pseudouridine synthase I. Nat Struct Biol 7:23–27

Galperin MY (2001) Conserved 'hypothetical' proteins: new hints and new puzzles. Comp Funct Genomics 2:14–18

Galperin MY (2004) Bacterial signal transduction network in a genomic perspective. Environ Microbiol 6:552–567

Galperin MY, Koonin EV (2004) 'Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res 32:5452–5463

Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinform 9:62

Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. Genome Res 12:1619–1623

Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313:903–919

Green JM, Nichols BP (1991) p-Aminobenzoate biosynthesis in *Escherichia coli*. Purification of aminodeoxychorismate lyase and cloning of pabC. J Biol Chem 266:12971–12975

Hall DR, Bond CS, Leonard GA, Watt CI, Berry A, Hunter WN (2002) Structure of tagatose-1,6-bisphosphate aldolase. Insight into chiral discrimination, mechanism, and specificity of class II aldolases. J Biol Chem 277:22018–22024

Hassan M, Kumar V, Somvanshi RK, Dey S, Singh TP, Yadav S (2007a) Structure-guided design of peptidic ligand for human prostate specific antigen. J Pept Sci 13:849–855

Hassan MI, Kumar V, Singh TP, Yadav S (2007b) Structural model of human PSA: a target for prostate cancer therapy. Chem Biol Drug Des 70:261–267

Hoang C, Ferre-D'Amare AR (2001) Cocrystal structure of a tRNA Psi55 pseudouridine synthase: nucleotide flipping by an RNA-modifying enzyme. Cell 107:929–939

Hoang C, Ferre-D'Amare AR (2004) Crystal structure of the highly divergent pseudouridine synthase TruD reveals a circular permutation of a conserved fold. RNA 10:1026–1033

Hoang C, Chen J, Vizthum CA, Kandel JM, Hamilton CS, Mueller EG, Ferre-D'Amare AR (2006) Crystal structure of pseudouridine synthase RluA: indirect sequence readout through protein-induced RNA structure. Mol Cell 24:535–545

Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38:W545–W549

Hooft RW, Sander C, Vriend G (1997) Objectively judging the quality of a protein structure from a Ramachandran plot. Comput Appl Biosci 13:425–430

Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C (1999) SCOP: a structural classification of proteins database. Nucleic Acids Res 27:254–256

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S et al (2011) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40:D306–D312

Ignoul S, Eggermont J (2005) CBS domains: structure, function, and pathology in human proteins. Am J Physiol Cell Physiol 289:C1369–C1378

Illergard K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins 77:499–508

Jackman JE, Montange RK, Malik HS, Phizicky EM (2003) Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. RNA 9:574–585

Jenal U (2004) Cyclic di-guanosine-monophosphate comes of age: a novel secondary messenger involved in modulating cell surface structures in bacteria? Curr Opin Microbiol 7:185–191

Kanehisa M (1997) Linking databases and organisms: GenomeNet resources in Japan. Trends Biochem Sci 22:442–444

Kaya Y, Ofengand J (2003) A novel unanticipated type of pseudouridine synthase with homologs in bacteria, archaea, and eukarya. RNA 9:711–721

Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32:W526–W531

Klebe G (2000) Recent developments in structure-based drug design. J Mol Med (Berl) 78:269–281

Koonin EV (1996) Pseudouridine synthases: four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases. Nucleic Acids Res 24:2411–2415

Krivov GG, Shapovalov MV, Dunbrack RL Jr (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77:778–795

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Kryshtafovych A, Moult J, Bartual SG, Bazan JF, Berman H, Casteel DE, Christodoulou E, Everett JK, Hausmann J, Heidebrecht T et al (2011) Target highlights in CASP9: experimental target structures for the critical assessment of techniques for protein structure prediction. Proteins 79(Suppl 10):6–20

Kuhnert P (2008) Pasteurellaceae: biology, genomics and molecular aspects. Caister Academic Press, Norwich

Kumar K, Prakash A, Tasleem M, Islam A, Ahmad F, Hassan MI (2014) Functional annotation of putative hypothetical proteins from *Candida dubliniensis*. Gene 543:93–100

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33:W89–W93

Laurie AT, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. Bioinformatics 21:1908–1916

Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40:D302–D305

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res 40:D370–D376

Lopez G, Valencia A, Tress ML (2007) Firestar—prediction of functionally important residues using structural templates and alignment reliability. Nucleic Acids Res 35:W573–W577

Markel TA, Crisostomo PR, Wang M, Herring CM, Meldrum KK, Lillemoe KD, Meldrum DR (2007) The struggle for iron:

gastrointestinal microbes modulate the host immune response during infection. J Leukoc Biol 81:393–400
Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325
Meinel T, Krause A, Luz H, Vingron M, Staub E (2005) The SYSTERS protein family database in 2005. Nucleic Acids Res 33:D226–D229
Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ et al (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33:D284–D288
Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 103:5361–5366
Mizutani K, Machida Y, Unzai S, Park SY, Tame JR (2004) Crystal structures of the catalytic domains of pseudouridine synthases RluC and RluD from *Escherichia coli*. Biochemistry 43:4454–4463
Morton DJ, Bakaletz LO, Jurcisek JA, VanWagoner TM, Seale TW, Whitby PW, Stull TL (2004a) Reduced severity of middle ear infection caused by nontypeable *Haemophilus influenzae* lacking the hemoglobin/hemoglobin–haptoglobin binding proteins (Hgp) in a chinchilla model of otitis media. Microb Pathog 36:25–33
Morton DJ, Smith A, Ren Z, Madore LL, VanWagoner TM, Seale TW, Whitby PW, Stull TL (2004b) Identification of a haem-utilization protein (Hup) in *Haemophilus influenzae*. Microbiology 150:3923–3933
Morton DJ, Madore LL, Smith A, Vanwagoner TM, Seale TW, Whitby PW, Stull TL (2005) The heme-binding lipoprotein (HbpA) of *Haemophilus influenzae*: role in heme utilization. FEMS Microbiol Lett 253:193–199
Murphy TF, Sethi S (1992) Bacterial infection in chronic obstructive pulmonary disease. Am Rev Respir Dis 146:1067–1083
Nakai T, Mizutani H, Miyahara I, Hirotsu K, Takeda S, Jhee KH, Yoshimura T, Esaki N (2000) Three-dimensional structure of 4-amino-4-deoxychorismate lyase from *Escherichia coli*. J Biochem 128:29–38
Nichols BP, Seibold AM, Doktor SZ (1989) Para-aminobenzoate synthesis from chorismate occurs in two steps. J Biol Chem 264:8597–8601
Nimrod G, Schushan M, Steinberg DM, Ben-Tal N (2008) Detection of functionally important regions in "hypothetical proteins" of known structure. Structure 16:1755–1763
Ofengand J, Malhotra A, Remme J, Gutgsell NS, Del Campo M, Jean-Charles S, Peil L, Kaya Y (2001) Pseudouridines and pseudouridine synthases of the ribosome. Cold Spring Harb Symp Quant Biol 66:147–159
Park SJ, Son WS, Lee BJ (2012) Structural analysis of hypothetical proteins from *Helicobacter pylori*: an approach to estimate functions of unknown or hypothetical proteins. Int J Mol Sci 13:7109–7137
Peng J, Xu J (2011) RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins 79(Suppl 10):161–171
Pidugu LS, Maity K, Ramaswamy K, Surolia N, Suguna K (2009) Analysis of proteins with the 'hot dog' fold: prediction of function and identification of catalytic residues of hypothetical proteins. BMC Struct Biol 9:37
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J et al (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301
Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120
Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99
Ramamurthy V, Swann SL, Paulson JL, Spedaliere CJ, Mueller EG (1999) Critical aspartic acid residues in pseudouridine synthases. J Biol Chem 274:22225–22230
Rappoport N, Karsenty S, Stern A, Linial N, Linial M (2012) ProtoNet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. Nucleic Acids Res 40:D313–D320
Reid AJ, Yeats C, Orengo CA (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. Bioinformatics 23:2353–2360
Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738
Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40:W471–W477
Saha S, Raghava GP (2006) VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. Genomics Proteomics Bioinform 4:42–47
Sethi S, Murphy TF (2001) Bacterial infection in chronic obstructive pulmonary disease in 2000: a state-of-the-art review. Clin Microbiol Rev 14:336–363
Shahbaaz M, Hassan MI, Ahmad F (2013) Functional Annotation of conserved hypothetical proteins from *Haemophilus influenzae* Rd KW20. PLoS One 8:e84263
Shapiro L, Harris T (2000) Finding function through structural genomics. Curr Opin Biotechnol 11:31–35
Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R et al (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 41:D490–D498
Singh R, Saha M (2003) Identifying structural motifs in proteins. Pac Symp Biocomput 8:228–239
Sivaraman J, Sauve V, Larocque R, Stura EA, Schrag JD, Cygler M, Matte A (2002) Structure of the 16S rRNA pseudouridine synthase RsuA bound to uracil and UMP. Nat Struct Biol 9:353–358
Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–W248
Soma A, Ikeuchi Y, Kanemasa S, Kobayashi K, Ogasawara N, Ote T, Kato J, Watanabe K, Sekine Y, Suzuki T (2003) An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. Mol Cell 12:689–698
Stojiljkovic I, Perkins-Balding D (2002) Processing of heme and heme-containing proteins by bacteria. DNA Cell Biol 21:281–295
Sutherland EW (1972) Studies on the mechanism of hormone action. Science 177:401–408
Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39:D561–D568
Taussig R, Gilman AG (1995) Mammalian membrane-bound adenylyl cyclases. J Biol Chem 270:1–4
Taylor WR, Orengo CA (1989) Protein structure alignment. J Mol Biol 208:1–22
Thakur PK, Hassan I (2011) Discovering a potent small molecule inhibitor for gankyrin using de novo drug design approach. Int J Comput Biol Drug Des 4:373–386

Thakur P, Kumar J, Ray D, Anjum F, Hassan MI (2013) Search of potential inhibitor against New Delhi metallo-beta-lactamase 1 from a series of antibacterial natural compounds using docking approach. J Nat Sci Biol Med 4:51–56

Tsoka S, Ouzounis CA (2000) Recent developments and future directions in computational genomics. FEBS Lett 480:42–48

Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850

Xu J, Jiao F, Yu L (2008) Protein structure prediction using threading. Methods Mol Biol 413:91–121

Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27:2076–2082

Yang J, Roy A, Zhang Y (2013) Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29:2588–2595

Ye QZ, Liu J, Walsh CT (1990) p-Aminobenzoate synthesis in *Escherichia coli*: purification and characterization of PabB as aminodeoxychorismate synthase and enzyme X as aminodeoxychorismate lyase. Proc Natl Acad Sci USA 87:9391–9395

Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Proteins 64:643–651

Yu J, Zhou Y, Tanaka I, Yao M (2010a) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics 26:46–52

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ et al (2010b) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26:1608–1615