

## Research Article

# Cross-Genome Comparisons of Newly Identified Domains in *Mycoplasma gallisepticum* and Domain Architectures with Other *Mycoplasma* species

Chandra Sekhar Reddy Chilamakuri,<sup>1,2</sup> Adwait Joshi,<sup>2</sup> Sane Sudha Rani,<sup>2</sup>  
Bernard Offmann,<sup>1</sup> and R. Sowdhamini<sup>2</sup>

<sup>1</sup>Equipe de Bioinformatique, Laboratoire de Biochimie et Génétique Moléculaire, Université de La Réunion, 15 avenue René Cassin, La Réunion, 97715 Saint Denis Messag Cedex 09, France

<sup>2</sup>National Centre for Biological Sciences, GKVK Campus, Bellary Road, Bangalore 560065, India

Correspondence should be addressed to R. Sowdhamini, mini@ncbs.res.in

Received 30 July 2010; Revised 21 February 2011; Accepted 23 May 2011

Academic Editor: G. Pesole

Copyright © 2011 Chandra Sekhar Reddy Chilamakuri et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Accurate functional annotation of protein sequences is hampered by important factors such as the failure of sequence search methods to identify relationships and the inherent diversity in function of proteins related at low sequence similarities. Earlier, we had employed intermediate sequence search approach to establish new domain relationships in the unassigned regions of gene products at the whole genome level by taking *Mycoplasma gallisepticum* as a specific example and established new domain relationships. In this paper, we report a detailed comparison of the conservation status of the domain and domain architectures of the gene products that bear our newly predicted domains amongst 14 other *Mycoplasma* genomes and reported the probable implications for the organisms. Some of the domain associations, observed in *Mycoplasma* that afflict humans and other non-human primates, are involved in regulation of solute transport and DNA binding suggesting specific modes of host-pathogen interactions.

## 1. Introduction

Progress in DNA sequencing technology has produced the whole genomes of many important organisms including humans. The proper utilization of such sequence information requires understanding of the function of each protein in the database. The ever-increasing gap between the number of sequences deposited in databases and the numbers with accurate functional annotation is a big concern to the scientific community. The goal of functional genomics is to determine the function of proteins predicted from the sequencing projects [1, 2]. To reach this goal, computational approaches can assist in the classification of functional genomics targets.

Functional and evolutionary relationships can be inferred from sequence comparisons, especially at high sequence identities. The established computational methods to function detection primarily depend on homology matching to genes with known functions by employing programs such as FASTA [3] and BLAST [4].

Nevertheless, establishing homology is not straightforward and provides limited coverage. Over the past few years, many new methods have emerged to organize the proteins; some of them are highly automated, and others are curated. Position-specific iterative BLAST (PSI-BLAST) can be used to extend the search to distantly related homologues [5]. Some of the other methods rely on the hierarchical classification of proteins into families such as the superfamilies/families in the PIR-PSD [6] protein groups in ProtoMap [7]. Few other methods organize proteins to families of domains such as Pfam [8] and SMART [9]. Others rely on sequence motifs or conserved regions, such as in PROSITE [10] and PRINTS [11]. Databases like CATH [12], SCOP [13], and FSSP [14] employ structural data to organize proteins in to domains. Others are integrations of various family classifications, such as InterPro [15]. However, each of these databases is useful for particular needs, and most of them rely on high sequence similarity for accurate function annotation

transfer, and no classification scheme is by itself adequate for addressing all genomic annotation needs [16]. The Gene Ontology (GO) consortium provides a controlled vocabulary to describe the function of a protein [17].

Identification of domains at the sequence level most often relies on the detection of global and local sequence alignments between a given target sequence and domain sequences found in databases such as Pfam [8] and SMART [9]. However, sequence-based methods often fail under low sequence identity conditions. Intermediate sequence approach has been shown to be more effective in enhancing the coverage in homology search and in connecting remotely related proteins of common function [18]. It was shown that about 70% improvement over direct search [18] is possible using this method. Using similar approach in the domain assignment to sequences, earlier, we showed that the domain assignment could be substantially enhanced in the family of genes containing adenyl cyclases [19]. PURE, this computation-intensive search protocol, was further developed as a web tool [20]. Next, we had implemented our method at the whole genome level by taking smaller genome organism *Mycoplasma gallisepticum* as a specific example [21]. This paper reports the cross-genome comparisons of 14 *Mycoplasma* genomes to study the conservation of domains and domain architectures involving new domain associations identified by us in *Mycoplasma gallisepticum*.

As shown in the earlier paper, PURE approach is effective in establishing remote domain relationships [20, 21] and can be useful when the user fails to assign domains to the sequence by using direct search methods like Pfam [8]. We also showed, by comparing different versions of Pfam databases, that the PURE approach can give a good hint at the domains, which are going to be assigned in the updated Pfam database [19].

*Mycoplasma* constitutes a unique group of bacteria best characterized as lacking peptidoglycan and having one of the smallest genomes of all free-living prokaryotes. Members of this group also represent important pathogens of humans, animals, and plants. Over the last few years, the genomes of many *Mycoplasma* species were sequenced, reinforcing comparative genome studies which permit a better understanding of their metabolism and the relations with their hosts. Phylogenetic analyses indicate that Mycoplasmas have undergone a degenerative evolution from related, low G+C content, Gram-positive eubacteria [22, 23]. Mycoplasmas possess no complete routes for amino acids synthesis and degradation, implying that these monomers must be acquired either from their hosts or from a culture medium, depending upon membrane transporters [24]. Exogenous peptides are an important source of amino acids. Indeed, bacteria have evolved peptide transport systems that also assist in responses to environmental changes, mediating functions such as quorum sensing, sporulation, pheromone transport, and chemotaxis [25].

## 2. Materials and Methods

Complete protein sequences of 14 different *Mycoplasma* genomes were obtained from National Center for Biotechnology

Information website [26]. The species we considered for our study were *Mycoplasma gallisepticum* strain R (total number of proteins in the genome 726), *Mycoplasma genitalium* strain G37 (477), *Mycoplasma agalactiae* strain PG2 (742), *Mycoplasma arthritidis* strain 158L3-1 (631), *Mycoplasma capricolum* subsp. *capricolum* (812), *Mycoplasma hyopneumoniae* strain 232 (691), *Mycoplasma hyopneumoniae* strain 7448 (657), *Mycoplasma hyopneumoniae* strain J (657), *Mycoplasma mobile* strain 163K(633), *Mycoplasma mycoides* subsp. *mycoides* SC str. PG1 (1016), *Mycoplasma penetrans* strain HF-2 (1037), *Mycoplasma pneumoniae* strain M129 (689), *Mycoplasma pulmonis* strain UAB CTIP (782), and *Mycoplasma synoviae* (659) (Table 1).

*Mycoplasma* species can be categorized into different groups based on motility and host specificity [27]. *Mycoplasma gallisepticum*, *Mycoplasma genitalium*, *Mycoplasma mobile*, *Mycoplasma pneumoniae*, and *Mycoplasma pulmonis* were grouped as motile and the remaining species *Mycoplasma agalactiae* PG2, *Mycoplasma arthritidis* 158L31, *Mycoplasma capricolum* ATCC 27343, *Mycoplasma hyopneumoniae* 232, *Mycoplasma hyopneumoniae* 7448, *Mycoplasma hyopneumoniae*, *Mycoplasma mycoides*, *Mycoplasma penetrans*, and *Mycoplasma synoviae* 53 were grouped as non-motile. Mycoplasmas were also classified based on the host specificity. *Mycoplasma genitalium*, *Mycoplasma penetrans*, *Mycoplasma pneumoniae*, and *Mycoplasma pulmonis* were primate specific, *Mycoplasma synoviae*.53 and *Mycoplasma gallisepticum* grouped as avian specific, *Mycoplasma hyopneumoniae* 232, *Mycoplasma hyopneumoniae* 7448, and *Mycoplasma hyopneumoniae* -J were grouped as swine-specific Mycoplasmas. *Mycoplasma arthritidis* 158L3 1 and *Mycoplasma pulmonis* are grouped as rodent specific, *Mycoplasma agalactiae* PG2, *Mycoplasma capricolum* ATCC 27343, and *Mycoplasma mycoides* are grouped as ovine specific, and lastly *Mycoplasma mobile* is fish-specific *Mycoplasma* in targeting its host for survival.

We assigned domain region to the *Mycoplasma gallisepticum* protein sequences by scanning the sequences against HMM profiles in the PfamA database (version 21.0) [8] which consists of 8957 families by using standalone version of Hmmpfam of the HMMER suite [28] with *E*-value cutoff 0.1.

HMMTOP [29] server was used for transmembrane helix prediction, and a standalone version of COILS [30] program was used for coiled-coil region prediction. We used PSI-BLAST [5] (with three iterations and expectation cutoff value of 0.001) for search for similar sequences. During the blast searches, low complexity filter was turned on. Non-redundant database [31] was used for sequence similarity searches. Standalone version of PSIPRED [32] was used for secondary structure prediction. Multiple sequence alignments were performed using CLUSTALW program [33].

## 3. Results and Discussion

Earlier analysis revealed 71 new domain relationships in the *Mycoplasma gallisepticum* genome which corresponds to 62 unassigned regions [21]. 22 domains, which are in the border

TABLE 1: 14 *Mycoplasma* species considered in this study. Host-group specificity and motility information is provided with genome size and total number of proteins present in the individual species.

Organism	Genome size (nt)	No. of proteins	Host-group specificity	Motility
<i>M. agalactiae</i> _PG2	877,438	742	Ovine/caprine	Nonmotile
<i>M. arthritidis</i> _158L3_1	820,453	631	Rodents	Nonmotile
<i>M. capricolum</i> _ATCC_27343	1,010,023	812	Ovine/caprine	Nonmotile
<i>M. gallisepticum</i>	1,012,800	726	Avian	Motile
<i>M. genitalium</i>	580,076	477	Human/primates	Motile
<i>M. hyopneumoniae</i> _232	892,758	691	Swine	Nonmotile
<i>M. hyopneumoniae</i> _7448	920,079	657	Swine	Nonmotile
<i>M. hyopneumoniae</i> -8	897,405	657	Swine	Nonmotile
<i>M. mobile</i>	777,079	633	Fish	Motile
<i>M. mycoides</i>	1,211,703	1016	Ovine/caprine	Nonmotile
<i>M. penetrans</i>	1,358,633	1037	Human/primates	Nonmotile
<i>M. pneumoniae</i>	816,394	689	Human/primates	Motile
<i>M. pulmonis</i>	963,879	782	Human/primates and Rodents	Motile
<i>M. synoviae</i> _53	799,476	659	Avian	Nonmotile

regions of cut-off expectation value, were excluded from the cross-genome analysis, and 49 domains which belong to 42 unassigned regions are used in the analysis. Detailed domain architectures, along with newly predicted domains, are shown in Table 2. 24 sequences out of 42 sequences picked up one or more domains, which were initially full-length unassigned sequences. Interestingly, some of the newly predicted domains such as Chase 3, DUF 1393, DUF 30, DUF 31, LMP, and HTH 12 are not present in the other *Mycoplasma* genomes. These domains could only be identified in *Mycoplasma gallisepticum* genome in the indirect searches. This could be because these domains may have species-specific functions or Mollicutes may have evolved by degenerative or reductive evolution, accompanied by significant losses of genomic sequences [34], wherein some of these domains might have lost their function and diverged beyond recognition by direct search methods. The intermediate sequences through which these domain relationships are established are predominantly of prokaryotic in origin and have relatively fewer hits in the PSI-BLAST search.

Our analysis also revealed the presence of extra copy of domains such as RMMBL, Lactamase\_B, ABC\_membrane, ABC\_tran, Lipoprotein\_X, SBP\_bac\_5, ATP\_synt\_ab\_N, Helicase\_C, tRNA\_anti, and GTP\_EFTU in the *Mycoplasma gallisepticum* genome. Because of the limited coding capacity of their genome, *Mycoplasmas* lack many enzymatic pathways characteristic of most bacteria; consequentially, *Mycoplasma* genes encode many proteins with functions related to catabolism and metabolite transport while encoding few anabolic proteins [35]. Most of these newly predicted domains related to transportation function. Despite low sequence identities, these domains could have critical function in the nutrient transportation. Some of the interesting examples are explained below.

Protein NP\_853190.1 was a completely unassigned protein. Our method predicted peptidase\_M23 (Peptidase family M23) domain relationship in the protein. Members of this family are zinc metallopeptidases and have a characteristic

HxH motif [36], and the current gene product also preserved this functional motif in the unassigned region. We found this domain in *Mycoplasma gallisepticum* only through indirect searches, and the unassigned sequence has less than 20% sequence identity with the typical peptidase\_M23 members, albeit with few indels in the alignment (Figure 1). Perhaps, the low sequence identity could explain why this is not associated with domain in the direct searches. Peptidase\_M23 domain is present in only two other *Mycoplasma* members (*Mycoplasma mobile* and *Mycoplasma pulmonis*). Interestingly, chaperonin (cpn60 or GroEL) domain is absent from these species but is present in *Mycoplasma gallisepticum* genome. Peptidases and chaperonins are components of protein homeostatic mechanisms. Molecular chaperones promote protein folding and prevent protein misfolding and aggregation, while certain proteases function primarily to degrade improperly folded proteins [37, 38]. It has been hypothesized that the protein homeostatic process in Mollicute organisms has shifted through evolution towards favoring protein degradation rather than protein folding [39]. Since peptidase\_M23 is present only in *M. mobile* and *M. pulmonis* (Figure 2) along with other peptidases where GroEL is completely absent from the genomes, this may explain the need for higher peptidases to degrade improperly folded proteins. Whereas, in *M. gallisepticum*, the presence of GroEL reduces the pressure on peptidases like peptidase\_M23 and sequences could have diverged substantially.

The full-length region of the sequence ID NP\_852865.1 was unassigned; that is, no sequence domains were observed and recorded. Our method indirectly assigned amino terminal Lactamase\_B and carboxy terminal RMMBL (RNA-metabolizing metallo-beta-lactamase) domains in the sequence. In the initial PSI-BLAST search against nonredundant database, it has picked up which belongs to more than 100 different species, including *Homo sapiens*, at very low expectation values. In the Hmmpfam search, all the hits showed identical domain architectures in all the sequences with amino terminal Lactamase\_B and carboxy terminal

TABLE 2: New domain architectures of 42 *Mycoplasma gallisepticum* proteins. For each protein reference ID is given in column two along with protein length. Fully associated domains are indicated with blue color, partially associated domains with brown color, and yellow color indicated the domains already associated with protein. Each domain is indicated by its name, starting and ending residues. In column four D represents Domain, Da indicates domain architectures. If particular domain or domain architecture present in the proteome it is indicated by P indicates present and domain/domain architecture not present in the existing proteome is indicated by NP meaning not present. Symbol \* indicates that; this protein sequence is completely unassigned before out method, @ indicated unique domain/domain architecture and & indicated average sequence identities.


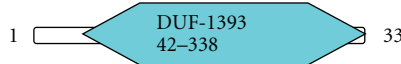




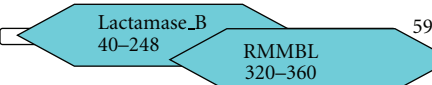
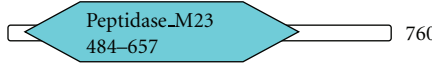
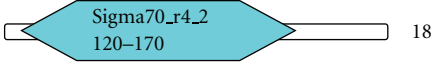

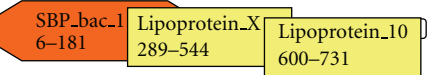
S. no.	Protein ID & unassigned regions	Domain architectures	<i>m. gallisepticum</i>
1* <sup>@</sup>	NP_853387.1 1-582 &18%	1  582	D: NP DA: NP
2* <sup>@</sup>	NP_853341.1 1-338 &68%	1  338	D: NP DA: NP
3*	NP_853479.1 1-810 &19%	1  810	D: P DA: P
4* <sup>@</sup>	NP_853440.1 1-613 &16%	1  613	D: P DA: NP
5* <sup>@</sup>	NP_853441.1 1-681 &17%	1  681	D: P DA: NP
6* <sup>@</sup>	NP_853488.1 1-809 &14%	1  809	D: P DA: NP
7* <sup>@</sup>	NP_852865.1 1-598 &14%	1  598	D: P DA: P
8* <sup>@</sup>	NP_853190.1 1-760 &22%	1  760	D: NP DA: NP
9* <sup>@</sup>	NP_852863.1 1-182 &25%	1  182	D: P DA: NP
10* <sup>@</sup>	NP_853458.1 1-124 &31%	1  272	D: NP DA: NP
11 <sup>@</sup>	NP_852814.1 <b>Q7NC49.MYCGA</b> 1-288 &18%	1  751	D: NP DA: NP

TABLE 2: Continued.







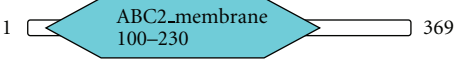
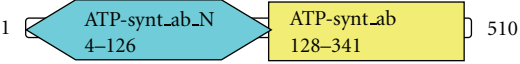
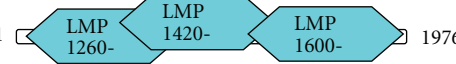
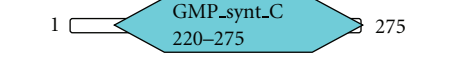
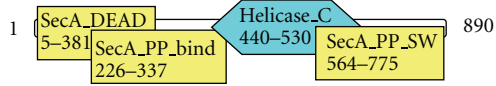
S. no.	Protein ID & unassigned regions	Domain architectures	<i>m. gallisepticum</i>
12* <sup>@</sup>	NP_852786.1 1-129 &30%	1  129	D: <i>P</i> DA: <i>NP</i>
13*	NP_853051.1 1-523 &22%	1  523	D: <i>P</i> DA: <i>P</i>
14* <sup>@</sup>	NP_852899.1 1-481 &18%	1  481	D: <i>P</i> DA: <i>NP</i>
15* <sup>@</sup>	NP_853298.1 1-904 &21%	1  904	D: <i>NP</i> DA: <i>NP</i>
16*	NP_852891.1 1-130 &27%	1  130	D: <i>P</i> DA: <i>P</i>
17*	NP_853257.1 1-83 &98%	1  83	D: <i>P</i> DA: <i>P</i>
18*	NP_853068.1 1-369 &14%	1  369	D: <i>P</i> DA: <i>P</i>
19 <sup>@</sup>	NP_853438.1 <b>Q7NAJ4_MYCGA</b> 1-127 &17%	1  510	D: <i>P</i> DA: <i>NP</i>
20* <sup>@</sup>	NP_853333.1 1-1976 &17%	1  1976	D: <i>NP</i> DA: <i>NP</i>
21* <sup>@</sup>	NP_852801.1 1-275 &22%	1  275	D: <i>NP</i> DA: <i>NP</i>
22 <sup>@</sup>	NP_852813.1 <b>Q7NC50_MYCGA</b> 338-563 &53%	1  890	D: <i>P</i> DA: <i>NP</i>

TABLE 2: Continued.

S. no.	Protein ID & unassigned regions	Domain architectures	<i>m. gallisepticum</i>
23 <sup>@</sup>	NP_853467.1 <b>Q7NAH2.MYCGA</b> 455-1051 &14%	1 [ HSDR_N 6-217 ] [ ResIII 267-454 ] [ Helicase_C 660-730 ] 1051	D: <i>P</i> DA: <i>NP</i>
24 <sup>@</sup>	NP_853482.1 <b>Q7NAF8.MYCGA</b> 435-641 &17%	1 [ DNA_ligase_aden 9-319 ] [ DNA_ligase_ZBD 407-434 ] [ HHH 589 ] [ DNA_ligase_OB 321-402 ] [ BRCT 642- ] 717	D: <i>NP</i> DA: <i>NP</i>
25 <sup>*@</sup>	NP_853456.1 1-1274 &24%	1 [ HNH 650-708 ] 1274	D: <i>NP</i> DA: <i>NP</i>
26 <sup>@</sup>	NP_853136.1 1-118 &15%	1 [ HTH_11 28-73 ] [ HrcA 119-335 ] 362	D: <i>NP</i> DA: <i>NP</i>
27 <sup>@</sup>	NP_853240.1 <b>Q7NB22.MYCGA</b> 1-257 &14%	1 [ HTH_12 38-89 ] [ RNB 257-586 ] [ S1 644-698 ] 707	D: <i>NP</i> DA: <i>NP</i>
28 <sup>@</sup>	NP_853425.1 <b>GIDA.MYCGA</b> 406-622 &30%	1 [ GIDA 14-405 ] [ HTH_5 580-622 ] 622	D: <i>NP</i> DA: <i>NP</i>
29 <sup>@</sup>	NP_852895.1 <b>Q7NBZ6.MYCGA</b> 134-625 &21%	1 [ NusA_N 8-133 ] [ S1 140-210 ] [ KH_1 333-393 ] 625	D: <i>P</i> DA: <i>NP</i>
30 <sup>*@</sup>	NP_852906.1 1-190 &23%	1 [ Methyltransf_3 7-185 ] 190	D: <i>NP</i> DA: <i>NP</i>
31 <sup>@</sup>	NP_853364.1 <b>Q7NAQ3.MYCGA</b> 453-559 &35%	1 [ PGM_PMM_I 54-197 ] [ PGM_PMM_III 330-452 ] [ PGM_PMM_IV 481-550 ] [ PGM_PMM_II 222-325 ] 559	D: <i>NP</i> DA: <i>NP</i>
32 <sup>*@</sup>	NP_853326.1 1-125 &17%	1 [ PTS_EIPB 47-85 ] 125	D: <i>P</i> DA: <i>NP</i>
33 <sup>@</sup>	NP_853386.1 63-127 &18%	1 [ RuvA_N 1-62 ] [ HHH 63-92 ] [ HHH 98-127 ] [ RavA_C 158-203 ] 226	D: <i>NP</i> DA: <i>NP</i>



TABLE 2: Continued.

S. no.	Protein ID & unassigned regions	Domain architectures	<i>m. gallisepticum</i>
34 <sup>@</sup>	NP_853171.1 <b>Q7NB90.MYCGA</b> 1-398 &18%	1  644	D: <i>NP</i> DA: <i>NP</i>
35 <sup>@</sup>	NP_852968.1 <b>Q7NBS7.MYCGA</b> 360-723 &13%	1  723	D: <i>NP</i> DA: <i>NP</i>
36 <sup>@</sup>	NP_853282.1 <b>Q7NAY5.MYCGA</b> 1-179 &22%	1  403	D: <i>NP</i> DA: <i>NP</i>
37 <sup>@</sup>	NP_852812.1 <b>Q7NC51.MYCGA</b> 521-666 &20%	1  666	D: <i>P</i> DA: <i>NP</i>
38 <sup>@</sup>	NP_852876.1 <b>Q7NC15.MYCGA</b> 1-319 &18%	1  1501	D: <i>P</i> DA: <i>NP</i>
39 <sup>@</sup>	NP_853404.1 <b>Q7NAL4.MYCGA</b> 1-118 &20%	1  278	D: <i>P</i> DA: <i>NP</i>
40 <sup>@</sup>	NP_853200.1 1-158 &14%	1  366	D: <i>P</i> DA: <i>NP</i>
41 <sup>*@</sup>	NP_853174.1 1-241 &12%	1  241	D: <i>NP</i> DA: <i>NP</i>
42 <sup>*@</sup>	NP_852793.1 1-261 &19%	1  261	D: <i>NP</i> DA: <i>NP</i>

RMMBL domains and with very good *E* values. The metallo-beta-lactamase fold contains five sequence motifs. The first four motifs are found in Lactamase\_B (PF00753) and are common to all metallo-beta-lactamases. The fifth motif appears to be specific to function. RMMBL represents the fifth

motif from metallo-beta-lactamases involved in RNA metabolism.

Multiple sequence alignment of predicted regions with typical Lactamase\_B and RMMBL (Figures 3 and 4) revealed that the most residues that are typical to the family are not

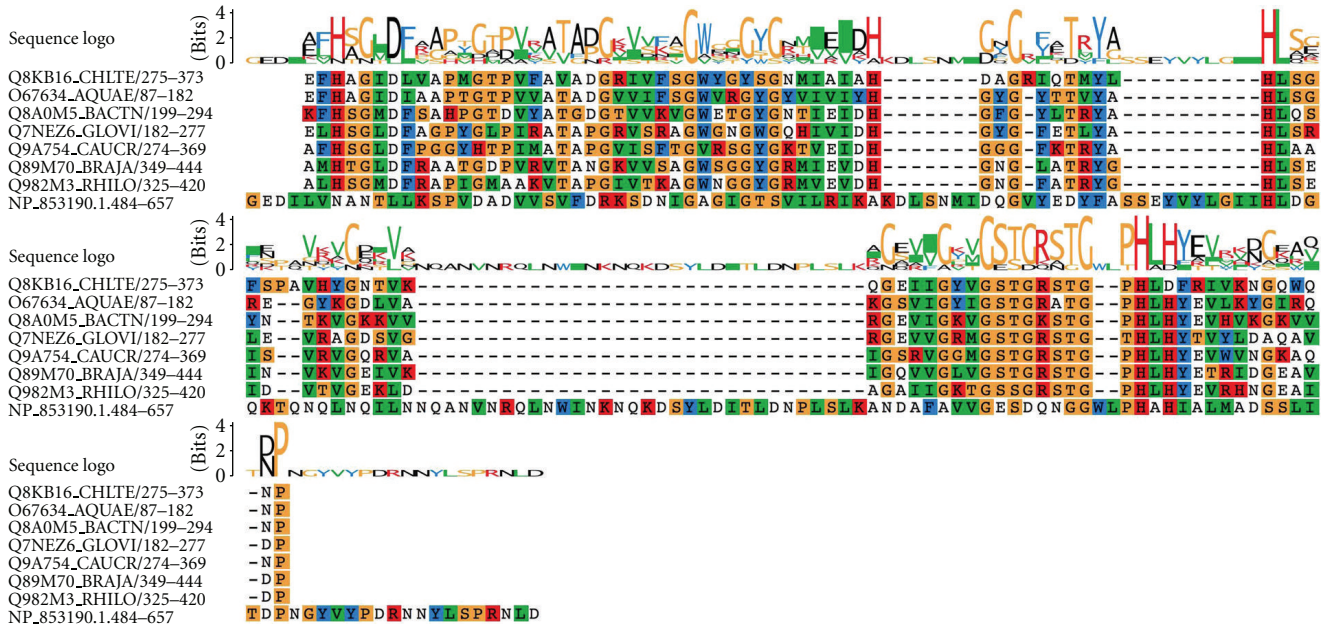


FIGURE 1: Multiple sequence alignment between peptidase\_M23 family representative sequences and unassigned protein sequence (NP\_853190.1) from *M. gallisepticum* genome. Peptidase\_M23 sequences obtained from Pfam database. Characteristic HxH motif is conserved and has few insertion regions in the unassigned sequence.

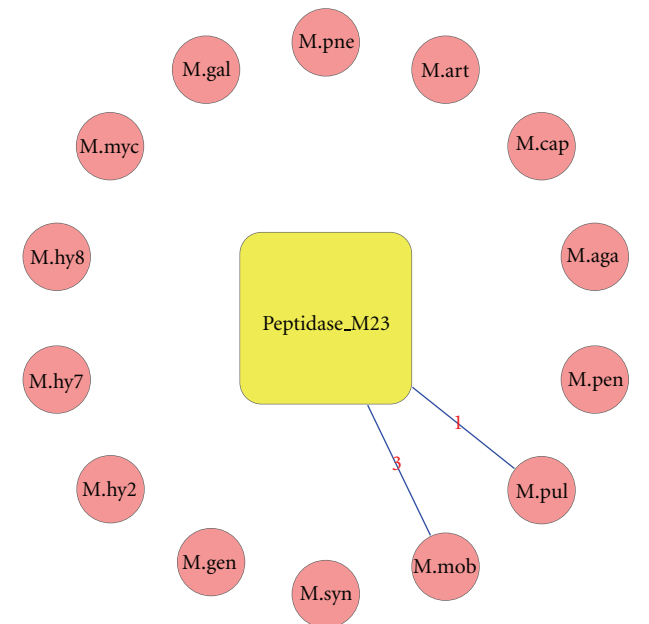


FIGURE 2: Peptidase domain presence in different *Mycoplasma* genomes. Domain represented by square box and species by circle. Lines connecting domain and species indicate the presence of domain in that species. Edge numbers indicate number of domains copies in genome.

conserved. The domains and domain architecture is conserved across Mycoplasmataceae members (Figure 5). It has been documented that presence of paralogs in *Mycoplasma genitalium* (MG139 and MG423) and *Mycoplasma pneumo-*

*niae* (MPN280 and MPN261) along with other bacteria [40, 41] could be as inactive forms. These inactive forms could be confined to modularity function helping in regulating enzymatic activity as already suggested by Aravind [41]. Acquisition of new functions beyond the ancestral enzymatic one is also possible [41]. Due to low sequence identities (<20%) with typical Lactamase\_B members, in *Mycoplasma gallisepticum* initially there was only one copy of Lactamase\_B domain in the genome (NP\_852802.1). Our analysis revealed that there is a putative paralog of this domain in this genome, like other *Mycoplasma* genomes.

SBP\_bac.5 (bacterial extracellular solute-binding proteins, family 5) domain relationship is established in NP\_853298.1 (see Table 2), which was initially full-length unassigned sequence. Cross-genome comparisons revealed that this domain is present in all the *Mycoplasma* species, except *Mycoplasma mobile*, *Mycoplasma pneumoniae*, and *Mycoplasma synoviae* (Figure 6). This domain is involved in peptide and nickel transportation. Mycoplasmas have reduced genome size and are highly dependent on the environment for nutrient abortion [35]. The presence of extra SBP\_bac.5 domain could help in the peptide uptake by the organisms.

*Mycoplasma* species were classified into six different groups according to host specificities (as mentioned earlier), and the newly predicted domains were classified based on the host specificities (Table 3; see Supplementary Table S1 in Supplementary Material available online at doi: 10.1155/2011/878973). There were few domains, which are group specific, while the majority are found in all the groups. The group specific domains perhaps imply their selectivity in the hosts owing to function which may be directly or indirectly



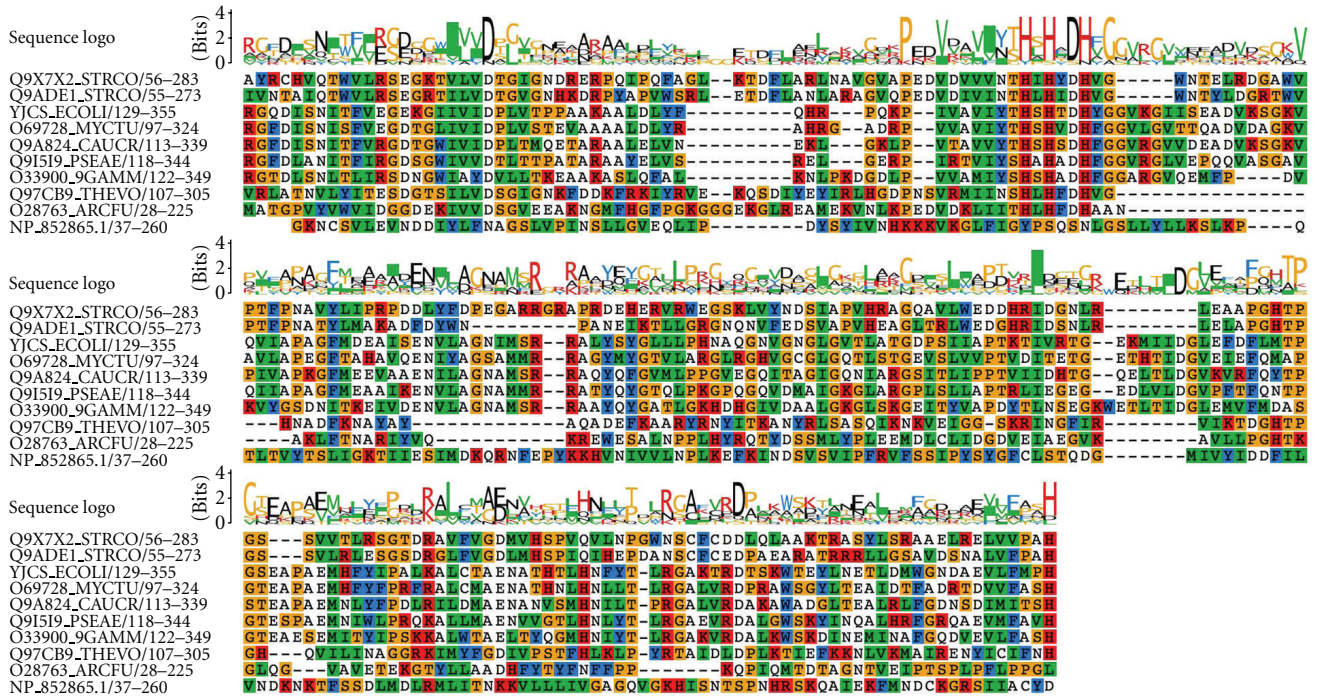


FIGURE 3: Multiple sequence alignment between Lactamase\_b family representative sequences and unassigned protein sequence (NP\_852865.1) from *M. gallisepticum* genome. Lactamase\_b sequences obtained from Pfam database.

required for its survival. We found that the two domains namely, HNH (endonuclease) and HTH\_5 (helix turn helix motif containing transcription factor), are specific to *M. mobile* (found in fresh water Tench fish—*Tinca tinca*). Five domains namely, GMP\_synt\_C (GMP synthase CTD), HHH (helix-hairpin-helix motif involved in DNA binding), Methyltransf\_3 (O-methyltransferases), SBP\_bac\_1 (Bacterial extracellular solute-binding protein), and Transposase\_mut (Transposase, Mutator family with DNA-based transposition activity), were found to be primarily in human-specific and primate group-specific pathogens. Most of these species-specific domains are involved in DNA binding and have transcription factor functions. One of them, GMP\_synt\_C (GMP synthase CTD), is associated with GATase (Glutamine amidotransferase class-I) and Peptidase.C26 domains to form a gene product in *M. penetrans* involved in GMP biosynthesis. Amongst the human- and primate-specific pathogens, *M. penetrans* has the largest genome (1,358,633 nt) and maximum number of proteins (1037) among all 14 *Mycoplasma* species analyzed in this study (Table 1), suggesting that this organism may possess additional genetic information involved in its unique infection process. This organism lacks pyrimidine biosynthetic machinery but using orotate-related metabolism (again unique to *M. penetrans*) circumvents this problem [33]. On the other hand, presence of purine biosynthesis (GMP synthase) related protein assists on the purine part of nucleotide biosynthesis. Also, the larger size of genome and number of proteins present underlines presence of GMP\_synt\_C domain specific to *M. penetrans*. Such an inspection of domain architectures in proteins containing these newly predicted domains was carried out for all

host-group specific domains. It revealed that, except for GMP\_synt\_C, all other domains are present as single domains in complete protein sequences. Most of the newly predicted domains are transcription factors not only involved in nucleotide biosynthesis but also specifically involved in the regulation of solute transport. This fact emphasizes the importance of solute transfer across the membrane in conditions of minimal genomes. Host-group-wise comparative analysis revealed that the TGS domain is present in two groups, rodents and ovine/caprine. Even within rodent-specific pathogens, it is present in only *M. arthritidis\_158L3\_1*, whereas; it is present in all three species of the ovine/caprine host group. TGS domain is named after threonyl-tRNA synthetase (ThrRS), GTPase, and guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase (SpoT). Its presence in proteins like GTPases suggests its role in ligand (nucleotide) binding or some regulatory function, but it has no direct information about function [35]. However, in *M. mycoides*, it is present in association with other domains in two different proteins. One of them is GTP diphosphokinase involved in guanosine tetraphosphate metabolic process explaining the possible involvement of TGS domain in nucleotide biosynthetic machinery. Here, *M. mycoides*, which also infects cattle (causing contagious bovine pleuropneumonia (CBPP)), has the second largest genome (1,211,703 nt) and number of proteins (1016) in the 14 *Mycoplasma* species under consideration (Table 1), explaining the presence of additional genetic information [34].

Some of the domains are specific to motile group, for example, HHH, HNH, HTH\_5, Peptidase\_M23, and SBP\_bac\_1 are specific to motile group, whereas GMP\_synt\_C,



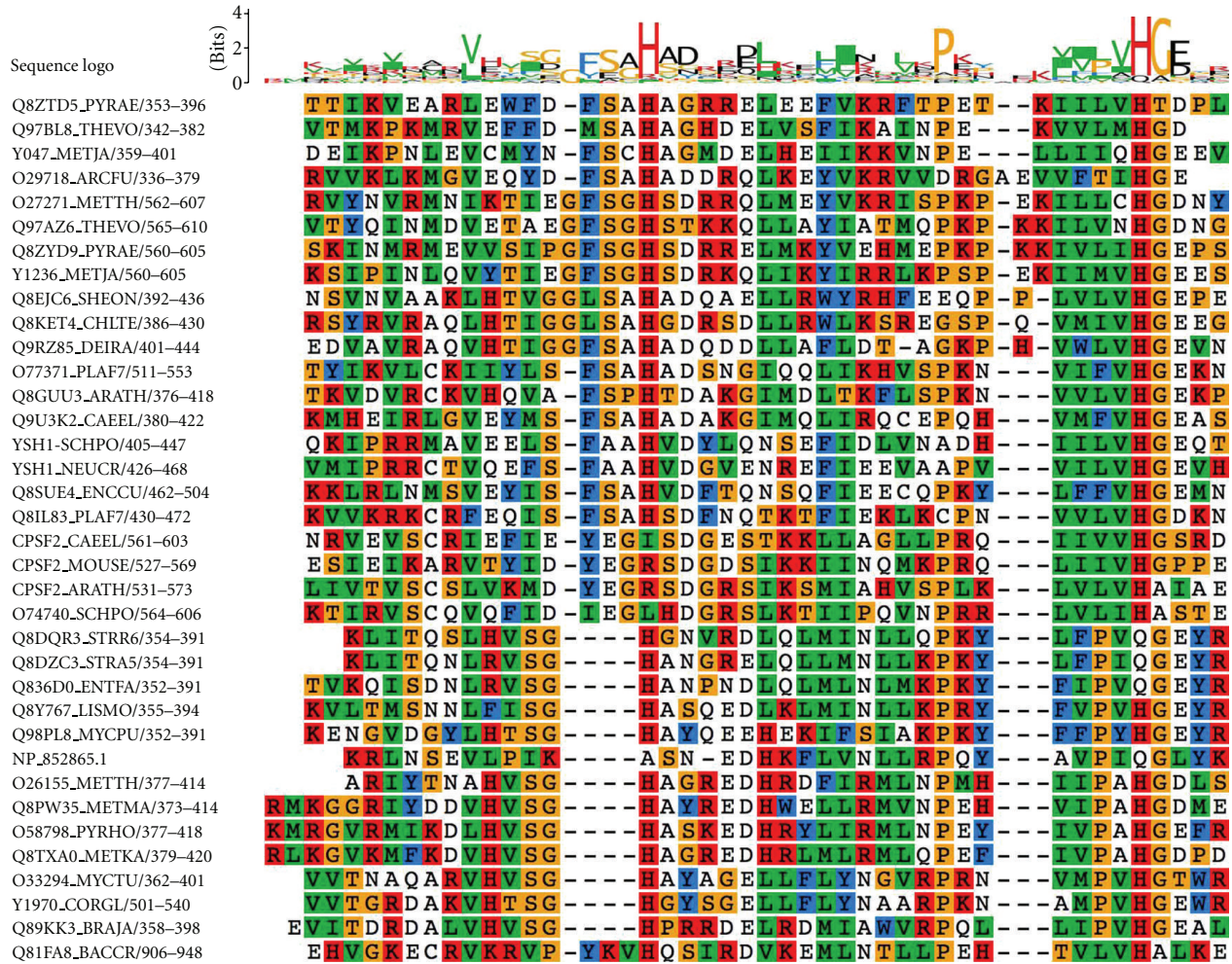


FIGURE 4: Multiple sequence alignment between RMMBL family representative sequences and unassigned protein sequence (NP\_852865.1) from *M. gallisepticum* genome. RMMBL sequences obtained from Pfam database.

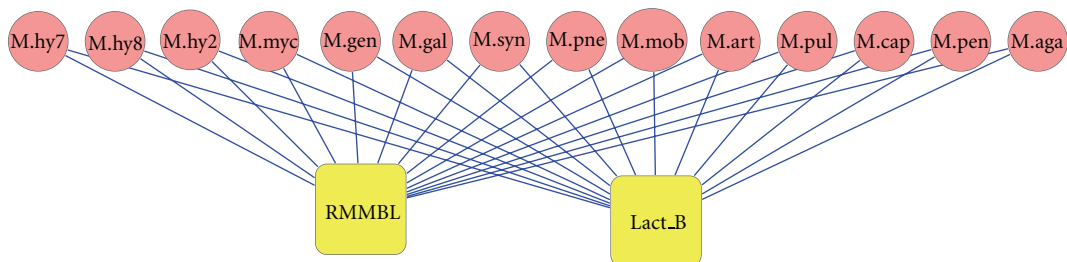


FIGURE 5: Presence of RMMBL and Lactamase.B domains and domain architecture in different *Mycoplasma* genomes. Domains are in square box and species in the circles. Edges represent presence of that domain in that species.

Methyltransf\_3, NusB, TGS, and Transposase\_mut domains are specific to nonmotile group (Supplementary Table S2). Inspecting the domain architectures for all the domains specific to the motility group, we found that they were not associated with any other domain in the complete protein sequence, except for the HHH domain in *M. pneumoniae*. Even in *M. pneumoniae*, HHH (helix-hairpin-helix motif—small DNA-binding motif) was associated with three different ligase domains involved in replication, repair, and recombination events. Therefore, although there is no

obvious link between the presence and absence of these domains and motility function, these distant relationships perhaps acquired new function, which may be required for motility of the pathogens.

#### 4. Conclusions

The investigation in the sequence information among closely related genomes helps in tracing of appearance, disappearance, and reappearance of genes or their close homologues

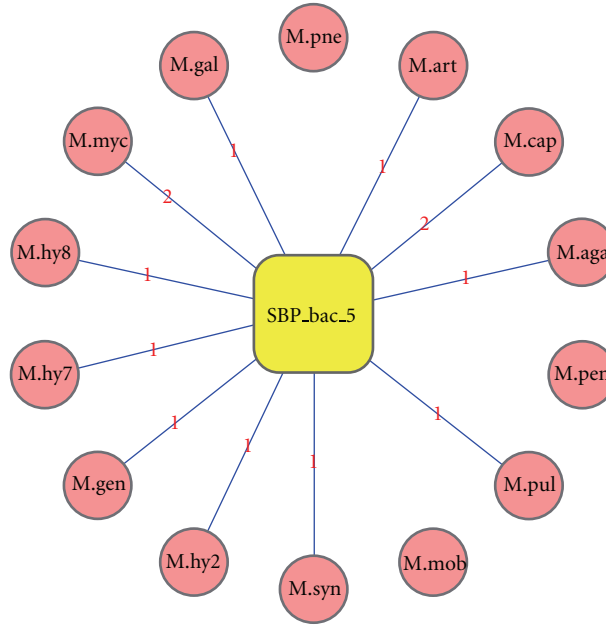


FIGURE 6: Comparison of SBP\_bac\_5 domain across different *Mycoplasma* genomes. SBP\_bac\_5 domain present in all the *Mycoplasma* genomes but in *Mycoplasma mobile*, *Mycoplasma pneumoniae*, and *Mycoplasma synoviae*. All the genomes have single copy of SBP\_bac\_5 domain except *Mycoplasma mycoides* and *Mycoplasma capricolum* where two copies were found in the genome.

TABLE 3: Comparison of PURE-predicted domains in *M. gallisepticum* to the Pfam-reported domains in the 14 *Mycoplasma* species. Species grouped based on host specificities of the individual *Mycoplasma* species.

Host group	Human (+primates)	Avian	Swine	Rodents	Ovine/caprine	Fish
No. of <i>Mycoplasma</i> spp.	4	2	3	2	3	1
Species	<i>M. genitalium</i> <i>M. penetrans</i> <i>M. pneumoniae</i> <i>M. pulmonis</i>	<i>M. synoviae</i> <sub>53</sub> <i>M. gallisepticum</i>	<i>M. hyopneumoniae</i> <sub>232</sub> <i>M. hyopneumoniae</i> <sub>7448</sub> <i>M. hyopneumoniae</i> <sub>8</sub>	<i>M. arthritis</i> <sub>158L3.1</sub> <i>M. pulmonis</i>	<i>M. agalactiae</i> <sub>PG2</sub> <i>M. capricolum</i> <sub>ATCC_27343</sub> <i>M. mycoides</i>	<i>M. mobile</i>
No. of domains	(25+26+24+27) 36	(23+24) 27	(24+24+24) 24	(24+27) 29	(25+24+27) 29	25
No. of intragroup common domains	19	20	24	22	21	25
No. of group-specific domains	5	0	0	0	0	2

in closely related bacterial genomes. Generally, functional annotation transfer is accomplished by phylogenomics-based methods that exploit strong phylogenetic relationship and based on the closest orthologue identified [42]. Apart from different sequence homology-based methods, microarray expression data along with machine learning techniques like Support Vector Machines (SVM) are integrated together for functional annotations [43]. Although use of such meth-

ods will be useful, GO annotations could be more comprehensive with regards to the biological process part or the cellular component part than for the exact molecular function [44]. Protein classification methods along with gene ontology terms are very useful tools in protein functional annotation. However, the best hit with respect to sequence identity may not be the correct protein to be used for annotation transfer since paralogous protein sequences from

the same organism do share high identity but function may vary.

In this study, newly and indirectly identified domains in *Mycoplasma gallisepticum* have been compared across 14 *Mycoplasma* species. This study showed that some of the newly identified domains are specific to *Mycoplasma gallisepticum* genome. Such genome-specific domains will perhaps provide important clues to the physiological and pathogenic specificities of the genome.

## Acknowledgments

C. S. R. Chilamakuri is supported by a Ph.D. grant from Conseil Régional de La Réunion. A. Joshi is supported by a fellowship by Council of Scientific and Industrial Research, India. This work is in part supported by PPF FRROI (Bioinflam) from the University of La Réunion and the French Ministry of Research. B. Offmann is thankful to Conseil Régional de La Réunion for financial support. R. Sowdhamini thanks the University of La Réunion for the visiting professorship position and NCBS for financial and infrastructural support. R. Sowdhamini was a Senior Research Fellow of the Wellcome Trust (UK).

## References

- [1] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan, "Predicting function: from genes to genomes and back," *Journal of Molecular Biology*, vol. 283, no. 4, pp. 707–725, 1998.
- [2] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, "Protein function the post-genomic era," *Nature*, vol. 405, no. 6788, pp. 823–826, 2000.
- [3] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 8, pp. 2444–2448, 1988.
- [4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
- [5] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.
- [6] W. C. Barker, F. Pfeiffer, and D. G. George, "Superfamily classification in PIR-international protein sequence database," *Methods in Enzymology*, vol. 266, pp. 59–70, 1996.
- [7] G. Yona, N. Linial, and M. Linial, "ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space," *Proteins*, vol. 37, no. 3, pp. 360–378, 1999.
- [8] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin, "Pfam: multiple sequence alignments and HMM-profiles of protein domains," *Nucleic Acids Research*, vol. 26, no. 1, pp. 320–322, 1998.
- [9] J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting, "SMART, a simple modular architecture research tool: identification of signaling domains," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 11, pp. 5857–5864, 1998.
- [10] N. Hulo, A. Bairoch, V. Bulliard et al., "The PROSITE database," *Nucleic Acids Research*, vol. 34, pp. D227–D230, 2006.
- [11] T. K. Attwood, M. J. Blythe, D. R. Flower et al., "PRINTS and PRINTS-S shed light on protein ancestry," *Nucleic Acids Research*, vol. 30, no. 1, pp. 239–241, 2002.
- [12] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH—a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093–1108, 1997.
- [13] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.
- [14] L. Holm and C. Sander, "The FSSP database: fold classification based on structure-structure alignment of proteins," *Nucleic Acids Research*, vol. 24, no. 1, pp. 206–209, 1996.
- [15] R. Apweiler, T. K. Attwood, A. Bairoch et al., "InterPro—an integrated documentation resource for protein families, domains and functional sites," *Bioinformatics*, vol. 16, no. 12, pp. 1145–1150, 2000.
- [16] C. H. Wu, H. Huang, L. S. L. Yeh, and W. C. Barker, "Protein family classification and functional annotation," *Computational Biology and Chemistry*, vol. 27, no. 1, pp. 37–47, 2003.
- [17] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The gene ontology consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [18] J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia, "Intermediate sequences increase the detection of homology between sequences," *Journal of Molecular Biology*, vol. 273, no. 1, pp. 349–354, 1997.
- [19] C. S. Reddy, A. Manonmani, M. Babu, and R. Sowdhamini, "Enhanced structure prediction of gene products containing class III adenyl cyclase domains," *In Silico Biology*, vol. 6, no. 5, pp. 351–362, 2006.
- [20] C. C. S. Reddy, K. Shameer, B. O. Offmann, and R. Sowdhamini, "PURE: a webserver for the prediction of domains in unassigned regions in proteins," *BMC Bioinformatics*, vol. 9, article 281, 2008.
- [21] C. C. Reddy, S. S. Rani, B. Offmann, and R. Sowdhamini, "Systematic search for putative new domain families in *Mycoplasma gallisepticum* genome," *BMC Research Notes*, vol. 3, article 98, 2010.
- [22] M. J. Rogers, J. Simmons, and R. T. Walker, "Construction of the mycoplasma evolutionary tree from 5S rRNA sequence data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 4, pp. 1160–1164, 1985.
- [23] C. R. Woese, J. Maniloff, and L. B. Zablen, "Phylogenetic analysis of the mycoplasmas," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 1, pp. 494–498, 1980.
- [24] A. T. R. Vasconcelos, H. B. Ferreira, C. V. Bizarro et al., "Swine and poultry pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*," *Journal of Bacteriology*, vol. 187, no. 16, pp. 5568–5577, 2005.
- [25] J. L. Wang, M. Y. Ho, and E. Y. Shen, "Mycoplasma pneumoniae infection associated with hemolytic anemia—report of one case," *Acta Paediatrica Taiwanica*, vol. 45, no. 5, pp. 293–295, 2004.
- [26] National center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/>.



- [27] K. Dybvig and L. L. Voelker, "Molecular biology of mycoplasmas," *Annual Review of Microbiology*, vol. 50, pp. 25–57, 1996.
- [28] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.
- [29] G. E. Tusnády and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, no. 9, pp. 849–850, 2001.
- [30] A. Lupas, M. Van Dyke, and J. Stock, "Predicting coiled coils from protein sequences," *Science*, vol. 252, no. 5010, pp. 1162–1164, 1991.
- [31] D. L. Wheeler, T. Barrett, D. A. Benson et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 34, pp. D173–D180, 2006.
- [32] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.
- [33] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.
- [34] C. R. Woese, "Bacterial evolution," *Microbiological Reviews*, vol. 51, no. 2, pp. 221–271, 1987.
- [35] C. M. Fraser, J. D. Gocayne, O. White et al., "The minimal gene complement of *Mycoplasma genitalium*," *Science*, vol. 270, no. 5235, pp. 397–403, 1995.
- [36] N. M. Hooper, "Families of zinc metalloproteases," *Federation of European Biochemical Societies Letters*, vol. 354, no. 1, pp. 1–6, 1994.
- [37] D. A. Dougan, A. Mogk, and B. Bukau, "Protein folding and degradation in bacteria: to degrade or not to degrade? That is the question," *Cellular and Molecular Life Sciences*, vol. 59, no. 10, pp. 1607–1616, 2002.
- [38] F. U. Hartl and M. Hayer-Hartl, "Protein folding. Molecular chaperones in the cytosol: from nascent chain to folded protein," *Science*, vol. 295, no. 5561, pp. 1852–1858, 2002.
- [39] P. Wong and W. A. Houry, "Chaperone networks in bacteria: analysis of protein homeostasis in minimal cells," *Journal of Structural Biology*, vol. 146, no. 1-2, pp. 79–89, 2004.
- [40] I. Callebaut, D. Moshous, J. P. Mornon, and J. P. De Villartay, "Metallo- $\beta$ -lactamase fold within nucleic acids processing enzymes: the  $\beta$ -CASP family," *Nucleic Acids Research*, vol. 30, no. 16, pp. 3592–3601, 2002.
- [41] L. Aravind, "An evolutionary classification of the metallo-beta-lactamase fold proteins," *In Silico Biology*, vol. 1, no. 2, pp. 69–91, 1999.
- [42] K. Sjölander, "Phylogenomic inference of protein molecular function: advances and challenges," *Bioinformatics*, vol. 20, no. 2, pp. 170–179, 2004.
- [43] M. P. S. Brown, W. N. Grundy, D. Lin et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.
- [44] I. Friedberg, "Automated protein function prediction—the genomic challenge," *Briefings in Bioinformatics*, vol. 7, no. 3, pp. 225–242, 2006.