The SHS2 Module Is a Common Structural Theme in Functionally Diverse Protein Groups, Like Rpb7p, FtsA, GyrI, and MTH1598/Tm1083 Superfamilies

V. Anantharaman and L. Aravind*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

ABSTRACT **Using structural comparisons, we identified a novel domain with a simple fold in the bacterial cell division ATPase FtsA, the archaeo-eukaryotic RNA polymerase subunit Rpb7p, the GyrI superfamily, and the uncharacterized MTH1598/Tm1083-like proteins. The fold contains a core of 3 strands, forming a curved sheet, and a single helix in a strand–helix–strand–strand (SHS2) configuration. The SHS2 domain may exist either in single or duplicate copies within the same polypeptide. The single-copy versions of the domain in FtsA and Rbp7p are most closely related, and appear to mediate protein–protein interactions by means of strand 1, and the loop between strand 2 and strand 3 of the domain. We predict that the interactions between FtsA and its functional partners in bacterial cell division are likely to be similar to the interactions of Rbp7p in the archaeo-eukaryotic RNA polymerase complex. The dimeric versions typified by the GyrI superfamily appear to have been adapted for small-molecule binding. Sequence profiles searches helped us to identify several new versions of the GyrI superfamily, including a family of secreted forms that is found only in animals and the bacterial pathogen *Leptospira*. Through sequence–structure comparisons, we predict the positions that are likely to be important for ligand specificity in the GyrI superfamily. In the MTH1598/Tm1083-like proteins, a SHS2 domain is inserted into the loop between strand 1 and helix 1 of another SHS2 domain. This has resulted in a structure that has convergent similarities with the Hsp33 and green fluorescent protein folds. The sequence conservation pattern and its phyletic profile suggest that it might function as an enzyme in some conserved aspect of nucleic acid metabolism. Thus, the SHS2 domain is an example of a simple module that has been adapted to perform an entire spectrum of functions ranging from protein–protein interactions to small-molecule recognition and catalysis. Proteins 2004;56:795–807.** © **2004 Wiley-Liss, Inc.**

Key words: cell division; dodecin; RNA polymerase; duplication; transcription; BmrR; Zsig11

INTRODUCTION

The rapid, concomitant growth of the sequence and structure databases, due to the various ongoing efforts in genomics, has enabled us to address several key problems in the early evolution of proteins. In particular, the combination of sequence and structure comparisons has made it possible to identify and trace the evolutionary history of certain simple monophyletic protein modules that have been conserved since the earliest epochs of the protein universe. Examples of these include the simple α-helical folds, such as the helix–turn–helix (HTH),1[–3] the helix–extended–helix (HEH),[4,5] and the helix–hairpin–helix (HhH)[6] domains, which are present in a wide range of proteins from the three principal superkingdoms of life. Despite their small size, they may either exist as stand-alone proteins or as a domain in multidomain proteins, or they may be incorporated into larger globular folding units.[2,7] Other simple modules, such as various types of 3-stranded units or the monomeric cystathionine beta synthase (CBS) domain unit do not appear to exist in stand-alone form. Instead, these appear to have given rise to larger globular folds such as the Double Psi beta barrel, the elongation factor-isomerase (EI) barrel, the sandwich barrel hybrid motif, the SET domain, and the dimeric CBS domain through dimerization and duplication.[8–12] Often, sequence and structure comparisons show that these simple units have clearly identifiable unique features that appear to have been preserved over long periods of evolution, both in stand-alone versions and in the forms that are parts of larger folding units. The retention of these features, despite the functional diversification of the individual protein families containing the units, strongly favors their divergent evolution from single ancestral units.

The identification of these ancient simple protein domains helps in understanding the provenance and evolutionary trajectories of more complex structures in proteins. The detection of these modules also enables us to glean common underlying structural themes that unify a range

Correspondence to: L. Aravind, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894. E-mail: aravind@ncbi.nlm.nih.gov

Received 9 October 2003; Accepted 30 January 2004

Published online 7 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20140

of disparate biological functions. As a result, they may aid in framing new hypotheses regarding certain poorly understood biochemical or biological functions. Extreme sequence divergence that accompanies functional diversification and incorporation into larger units, through duplication or domain insertion, can easily obscure the affinities of these simple ancient domains. On account of their insertion into or association with larger globular units, or accretion to form larger folding units, these simple domains are occasionally overlooked in both manual and semiautomatic structural classification schemes such as the SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) and the CATH (www.biochem.ucl.ac.uk/bsm/cath/) databases. As a result, investigating these proteins require detailed case-by-case investigation using sensitive sequence and structure similarity search methods.

In this work, we identify and characterize a novel simple protein domain in a wide range of proteins with distinct biological functions. The bacterial cell-division protein, FtsA, is an ATPase that is most closely related to Hsp70, the eukaryotic cytoskeletal protein actin, and MreB. All these proteins and several other more distantly related proteins, such as the sugar kinases, contain two copies of the RNAse H-fold domain. However FtsA differs from all these proteins in possessing a unique insert in first of the RNAse H-fold repeats.[13] The crystal structure of FtsA revealed that this insert folds into a distinct globular module, termed the 1C, and is believed to be mediate critical interactions of the FtsA protein.[13,14] However, the provenance of this domain in FtsA has been thus far unclear. Using sequence profile searches and transitive structure similarity searches, we show that homologous domains are also present in the archaeo-eukaryotic subunit of the DNA-dependent RNA polymerase, namely, Rpb7p/Rpc25p/MJ0397 (henceforth called Rpb7p), Rob, GyrI, and related regulatory domains of bacterial transcription factors and certain conserved uncharacterized proteins, which are highly conserved in archaea and eukaryotes. This module contains a conserved core of three strands and helix, and defines a distinct, evolutionarily mobile domain with a simple fold. We trace the evolutionary history of this module and present evidence that it has been utilized as a common theme in a variety of biological functions. We show that the monomeric forms of the domain have been principally utilized in functional contexts related to protein–protein interactions. Additionally, different duplicated versions of this simple structure have given rise to more complex structures that contain clefts. These clefts form binding sites for diverse ligands or may even act as scaffold for uncharacterized catalytic activities. We also provide evidence that some of the more complex structures formed through the duplication of this simple module could convergently resemble other complex folds, which have arisen through the duplication of entirely unrelated founding units.

MATERIAL AND METHODS

The nonredundant (NR) database of protein sequences (National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD) was searched using the BLASTP program.[15] Profile searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with a default profile inclusion expectation (E) value threshold of 0.01 (unless specified otherwise), and was iterated until convergence.[15,16] In PSI-BLAST searches that were initiated with distinct globular domains without any compositional bias (as indicated by the SEG program), we did not use the compositional–bias–correction option. This increased sensitivity of the searches, without bringing false positives into the profiles. Multiple alignments were constructed using the T-Coffee[17] or PCMA[18] programs, followed by manual correction based on the PSI-BLAST results. All large-scale sequence analysis procedures were carried out using the SEALS package: http://www.ncbi.nlm.nih.gov/cbbresearch/walker/seals/index.html.

For structural comparisons, the DALI/FSSP program was used.[19,20] It has been shown that DALI Z scores > 10 are characteristic of obvious relationships, such as those between two members of the same family. Z scores between 6 and 10, typically, correspond to more distant relationships that are detectable through sequence profile analysis. Z scores < 3 fall in the realm of remote structural relationships and require additional analysis, such as comparisons of topologies, to make inferences of homology.[19,21] Protein secondary structure prediction was performed using the PHD program through the PredictProtein server.[22] The Swiss-PDB viewer and PyMol (http://pymol.sourceforge.net/) programs were used to carry out structural superpositions and other manipulations of Protein Data Bank (PDB) files.[23] Figures were rendered using PyMol[24] (http://www.pymol.sourceforge.net/) or POV-Ray (http://www.povray.org/).

Similarity-based clustering of proteins was carried out using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt). Phylogenetic analysis was carried out using the maximum-likelihood, neighbor-joining, and least-squares methods.[25,26] Briefly, this process involved the construction of a least-squares tree using the FITCH program,[27] or a neighbor-joining tree using the NEIGHBOR[27] or the MEGA program,[28] followed by local rearrangement using the Protml program of the Molphy package[26] to arrive at the maximum likelihood (ML) tree. The statistical significance of various nodes of this ML tree was assessed using the relative estimate of logarithmic likelihood bootstrap (Protml RELL-BP), with 10,000 replicates.

RESULTS AND DISCUSSION

Identification of the SHS2 Module

The FtsA protein, which is an ATPase involved in bacterial cell division,[13] contains two repeats of the RNase H-fold domain[29] (also see the SCOP database: http://scop.mrc-lmb.cam.ac.uk/scop/). Each RNAse H-fold module has a conserved core with a $\beta_3\alpha\beta\alpha\beta\alpha$ topology. The first repeat of the RNAse H-fold module in FtsA contains a unique insert between the fourth strand and the second helix. This insert is a distinctive structure and has been termed the 1C domain in the FtsA crystal structure.[13] Experimental analyses have suggested that this domain is important for the specific biological function of FtsA in bacterial cell division (see below).[14] PSI-BLAST searches seeded with the sequence of this module from the *Thermotoga* FtsA showed that this module was present in practically all FtsA proteins, but it was not detectable in any of its closely related homologs such as Actin, Hsp-70, or MreB. A direct comparison of the FtsA structure with all other RNAse H-fold proteins with duplicated RNAse H-fold modules, such as actin, Hsp70, and the sugar kinases,[29] confirmed the observation that this module was only present in FtsA. This, observation also strongly suggested that the 1C domain was inserted into the N-terminal RNAse H-fold domain after the divergence of FtsA from the other proteins with two RNAse H-fold modules. Given that FtsA can be traced back to the common ancestor of all extant bacteria, the above events appear to have occurred prior to the divergence of the bacteria from their common ancestor.

The 1C domain of FtsA adopts a simple fold with 3 strands and single helix arranged in a $\beta\alpha\beta_2$ pattern and a 132-strand order in the β-sheet (Fig. 1). To understand the origins of this domain of FtsA, we sought to detect more distantly related versions of this module in other proteins using structure similarity searches of the PDB database. Given that this is a simple module, and units with a similar topology are found within other well-characterized folds such as the RRM[30,31] and IF3-C[32,33] (Fig. 2), the results of the structure similarity searches needed to be further critically evaluated. A search of the PDB database, using the coordinates of the 1C domain from FtsA, with the DALI program recovered the N-terminal domain of the RNA polymerase subunit Rpb7p[34] as the best hit ($Z = 5.2$). A visual examination of this domain of the Rpb7p showed that it had an identical topology as the FtsA 1C domain and also shared specific structural peculiarities, such as the elongated strands and a twisted β-sheet (Fig. 1). A reciprocal search with the Rbp7p-N terminal domain recovered the FtsA 1C domain as the best hit and the two domains shared a root-mean-square deviation (RMSD) of ~3 Å over the aligned segment of approximately 65 residues. These observations suggested that FtsA and Rpb7p share a novel simple domain that appears to have descended from a common ancestor. We refer to this module as the strand–helix–strand–strand, or SHS2, fold based on the pattern of secondary structure elements present in it.

The structure similarity searches with the above SHS2 domains also consistently recovered representatives of two other families of proteins, namely, the gyrase inhibitor,[35,36] GyrI (and its homologs, such as the transcription factors Rob[37] and BmrR[38]), and the uncharacterized prokaryotic proteins MTH1598 (also termed archease in the GenBank database) from *Methanobacterium*[39] and Tm1083 from *Thermotoga* ($Z \sim 4.3$–4.8). Both these proteins contained two domains with a $\beta\alpha\beta_2$ pattern with the same order of strands as the above SHS2 modules. In the case of the GyrI-like proteins, the two $\beta\alpha\beta_2$ domains were adjacent to each other in the form of a tandem duplication[36] (Fig. 1). Furthermore, these two individual repeats of the GyrI-like proteins detected each other in sequence similarity searches (see below). In the case of MTH1598-like proteins, one of the $\beta\alpha\beta_2$ modules was inserted into the second topological equivalent $\beta\alpha\beta_2$ module (Fig. 1). Reciprocal structural similarity searches with the individual $\beta\alpha\beta_2$ units from MTH1598 recovered the SHS2 modules of FtsA and Rpb7p as their neighbors. Visual examination of all these structures showed that, in addition to an identical topology, they shared other commonalities in terms of the strand length, sheet curvature, and the presence of a long connective between the helix and strand 2, (Figs. 1 and 2). In all these structures, strands 2 and 3 crossed each other in the same fashion, and after crossover, the paths of the backbones are oppositely directed (Figs. 1 and 2). Furthermore, when the stand-alone $\beta\alpha\beta_2$ modules of FtsA 1C and Rpb7p, the duplicate $\beta\alpha\beta_2$ modules of the GyrI-like and the MTH1598-like proteins, and topologically similar parts of more complex domains, such as the RRM-like fold, the IF3-C fold, and the YjgF-like fold,[40] were clustered based on pairwise Z scores, the former set grouped together to the exclusion of other structures (Fig. 2). This suggested that, like the FtsA 1C and Rpb7p N-terminal domains, the two $\beta\alpha\beta_2$ modules of the GyrI-like proteins and MTH1598-like proteins are also members of the SHS2 fold.

The Structural Contexts of the SHS2 Fold Domains

The SHS2 fold domains could be clearly differentiated into two classes that occurred in distinct structural contexts. One class is composed of the forms in which the SHS2 fold occurs in a single copy in the polypeptide. This class includes the FtsA insert (the 1C domain) and the Rpb7 N-terminal domain, which is fused to a C-terminal S1 domain that adopts the oligomer binding (OB) fold. In both these versions, the SHS2 fold does not form many close contacts with the other domains in the same polypeptide. These forms are also unified by the presence of a longer flap-like insert between the second and third strands (Figs. 1 and 2). Experimental studies on FtsA 1C[14] and Rpb7p[34,41] suggest that they are mainly involved in homo- and heteromeric contacts with other proteins in the complexes in which they occur.

The second class of SHS2 domains includes those that typically occur as duplicate copies within the same polypeptide, namely, the GyrI-like and the MTH1598-like proteins. In these proteins, the SHS2 fold domains may possess an N-terminal region, which adopts an extended conformation (Figs. 1 and 2). In the GyrI-like proteins, which have emerged through the simple tandem duplication of the SHS2 fold domain, the individual modules are arranged in the opposite orientation with respect to each other. The dimer interface between the two modules is formed via two distinct sets of interactions: (1) N-terminal regions in extended conformation from each repeat form hydrogen bonds with each other, as well as with the strand 1 of the other repeat; and (2) strand 2 from each of the SHS2 modules, which are oriented in an antiparallel

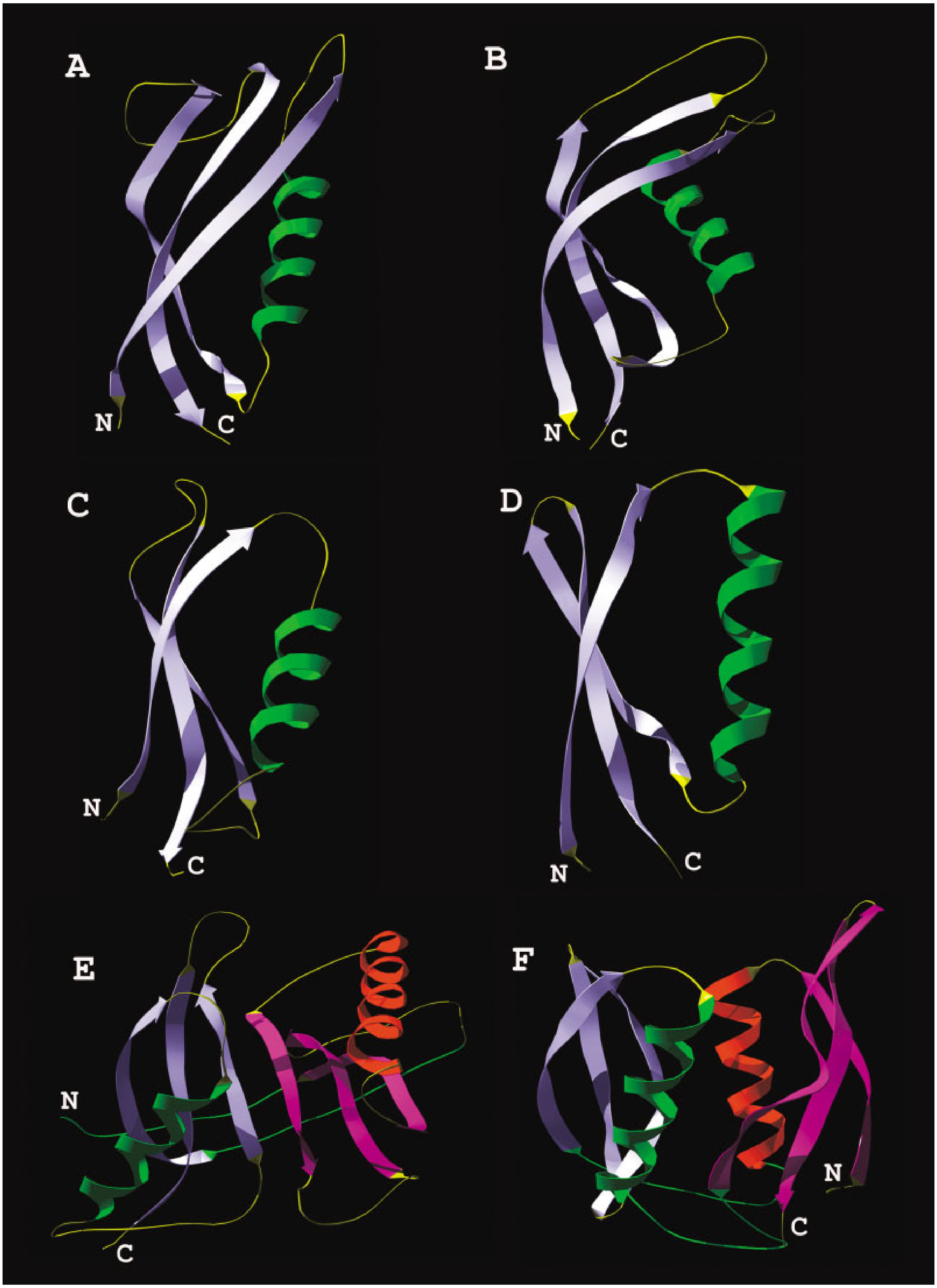

Figure 1.

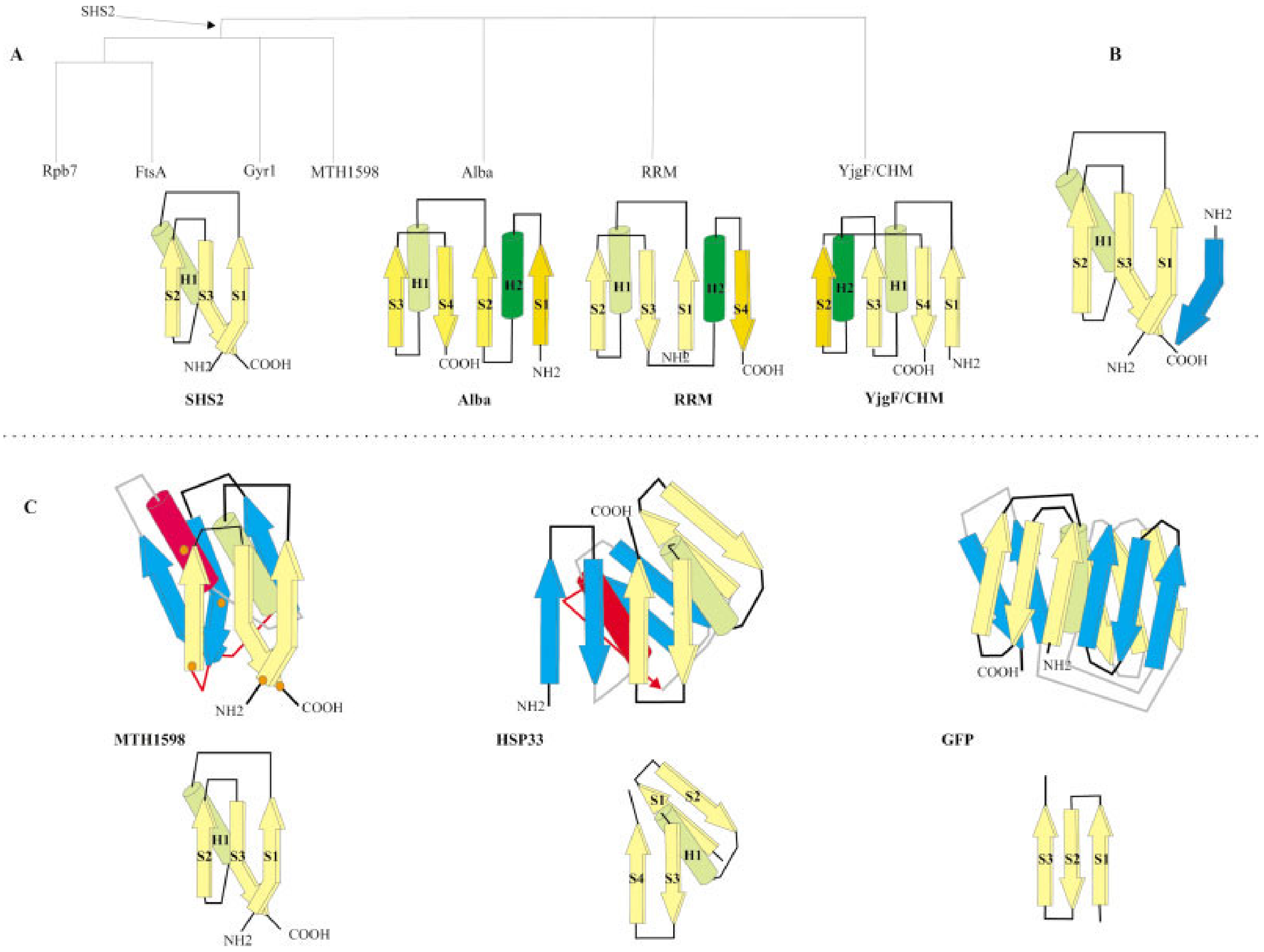

Fig. 2. Topology of SHS2 fold: similarities and differences with other folds. (**A**) Relationship of the SHS2 domains and their monophyly with respect to other folds with topologically or structurally equivalent units [IF3-C (PDB code: 1tig), RRM (PDB code: 1ris), and YjgF/CHM (CHM, chorismate mutase) (PDB code: 1qu9)] are shown. The clustering is based on pairwise *Z* scores between the different structures. The equivalent core is shown in light yellow and light green, while the extra strands and helices are shown in orange and dark green, respectively. (**B**) The interaction of the N-terminal strand of SHS2 with other antiparallel extended regions in dimerization. The antiparallel element, which may come from the same protein (as in the case of the GyrI-like proteins), or from another protein in the complex (as in Rpb7-Rpb4 complex) is shown. (**C**) The topological representations of MTH1538/Tm1083-like proteins (PDB code: 1jw3; 1j5u), HSP33 (PDB code: 1hw7), and GFP (PDB code: 1emb) and the individual precursor units from which they are likely to have arisen through duplication. The conserved residues in the MTH1538/Tm1083-like proteins are shown as orange dots.

direction in the dimer, form hydrogen bonds with each other. As result, the 2 modules form a broad, contiguous 6-stranded sheet, bounded on either side by the single helix from each SHS2 module, and reinforced from behind by the two interacting N-terminal extensions (Fig. 1). Thus, the face of the dimer provides a large depression that could potentially accommodate other molecules (Fig. 1).

In MTH1598-like proteins, the arrangement of the two SHS2 modules is dramatically different from that of the GyrI-like dimers. Here, the second SHS2 module is inserted into the loop between the helix and the second strand of the first SHS2 module (Fig. 1). As a result, the two modules are arranged similar to joined palms, with the helices from the respective modules packing against each, and lying within a sheath formed by the strands (Figs. 1 and 2). This arrangement is strikingly reminiscent of certain other protein folds, such as the heat shock protein 33 (HSP33),[42] nidogen-1 perlecan binding domain (NPBD),[43] and the green fluorescent protein fold (GFP),[44,45] which contain helical segments within an outer sheath formed by β-strands (Fig. 2). Consistent with this, only the search with entire MTH1598/Tm1083 structure, but not the individual SHS2 units, using the DALI program, recovers Hsp33 (Z score ~ 4.1). A careful analysis of the

Fig. 1. Single copy and the dimeric units of SHS2 modules. The single SHS2 units of (**A**) FtsA 1C domain (PDB code: 1e4gT); (**B**) Rpb7p (PDB code: 1go3E); (**C**) BmrR/Rob family protein C terminal domain (PDB code: 1jyh); and (**D**) MTH1598/Tm1083 inserted SHS2 domain (PDB code: 1j5u); and the dimeric SHS2 units of (**E**) BmrR/Rob family (PDB code: 1jyh); and (**F**) MTH1598 (PDB code: 1jw3; 1j5u) are shown in ribbon representation. In the case of the dimeric forms, the two repeats have been colored differently. The figures were generated using SWISS-PDB viewer and POV-ray.

structures of HSP33, NPBD, and GFP reveals that their structural similarity with MTH1598 is superficial and is most likely to be a product of convergent evolution (Fig. 2). The core of the Hsp33 fold can be decomposed into two $\beta_2\alpha\beta_2$ units that interlock together to give rise to the overall structure, where the two helical segments are sandwiched between two β-sheets (Fig. 2). Despite extensive sequence divergence, the structures of the two units superimpose very well (RMSD ~ 2.6 Å) and detect each other as the best hits in structure similarity searches of the PDB database ($Z \sim 6.8$). Thus, the HSP33 domain appears to have emerged from a duplication of an ancestral $\beta_2\alpha\beta_2$ module. An examination of the spatial arrangement of the two repeats in Hsp33 shows that each repeat contributes 2 strands to both the sheets that surround the 2 central helices. This arrangement suggests that the stand-alone ancestral $\beta_2\alpha\beta_2$ units of the HSP33 fold are likely to have interlocked to assemble as a dimer, which was stabilized by the formation of the 4-stranded β-sheets on either side of the helices (Fig. 2). GFP and NPBD share a common fold that contains 11 core strands, which form a barrel surrounding a single helix. The strand order and arrangement in the GFP/NPBD fold resembles neither MTH1598/Tm1083 nor HSP33 (Fig. 2). In the case of the GFP/NPBD barrel, there are at least 3 topologically identical antiparallel β_3 units (strand order 123) that constitute its core and could potentially represent the ancient ancestral units from which this fold was assembled. Thus, the origins of these structures with certain global similarity from unrelated ancestral precursor units provide a clear illustration of convergent evolution in proteins.

The above-discussed structural contexts of the SHS2 domain suggest that, like other simple folds, such as the HTH, the HEH, and the chromodomain-like folds, it may exist either in stand-alone forms or as the building blocks of larger units.

Functional Diversity of the SHS2 Module

The small sheets of the SHS2 tend to form more stable aggregates by hydrogen-bonding with other extended segments. As discussed above, one instance of this is observed in the GyrI-like proteins, in the form of the hydrogen bonds between the first strands of the SHS2 modules and the antiparallel N-terminal extensions of the adjacent repeats. The structure of the complex of the RNA polymerase subunits, Rpb7 and Rpb4, shows a very similar hydrogen-bonding interaction between strand 1 of the Rpb7 SHS2 domain and an extended N-terminal segment of Rpb4[34,41] (Fig. 2). The presence of this interaction in two distinct forms of the SHS2 domain suggests that it may be also occur in other SHS2 domains. In particular, the 1C domain of FtsA has been proposed to play a role in interaction with other bacterial cell-division proteins, such as FtsZ.[13,14] Hence, it is possible that a similar bonding between an extended segment of FtsZ and strand 1 of the SHS2 domain of FtsA could serve as the basis of this interaction. The crystal structure of the eukaryotic RNA polymerase complex suggests that Rpb7 also interacts with the catalytic subunit Rpb1 by means of the loop between strand 2 and strand 3 of its SHS2 domain.[41] This loop has a characteristic sequence motif of the form Gxxs (where s is typically a small residue) (Fig. 3) and appears to be buried in a prominent depression on the surface of the Rpb1 structure (close to residues 1443–1445 in Rbp1).[41] A loop with a similar sequence signature is also seen in the identical position in the SHS2 domain of FtsA (Fig. 3). Recent mutational analysis of FtsA has suggested that disruption of this loop interferes with the dimerization of FtsA.[14] Based on this, it was proposed that the loop might be inserted into a cleft formed between the two RNAse H-fold modules, thereby mediating the dimerization of FtsA.[14] Thus, both the solo SHS2 domains appear to mediate similar interactions through the region between strand 2 and strand 3, suggesting that it was ancestral feature of this class of SHS2 domains.

Sequence searches suggest that the GyrI-like proteins define a vast superfamily (hereinafter referred to as the GyrI superfamily) of proteins,[36] most of which contain two copies of the SHS2 domain (Fig. 4). However, sequence profile searches seeded with different GyrI-like proteins also recovered certain members of this superfamily from *Desulfitobacterium* and *Magnetospirillum* that apparently possessed only a single SHS2 domain. It is possible that these members could homodimerize to adopt an overall structure similar to those forms, which contain the internal duplication. The functional diversity of the GyrI-like proteins with duplicate SHS2 domains is not entirely known. The region corresponding to the first strand of the second SHS2 in the GyrI dimer has been shown to interact with the DNA gyrase holoenzyme, thus inhibiting the gyrase's supercoiling activity.[46] Thus, at least in the case of GyrI, the SHS2 domain could mediate its functions through protein–protein interactions, similar to the other SHS2 domains. In the case of Rob and BmrR, the GyrI-like unit is fused to different types of N-terminal HTH domains that bind DNA.[37,38,47] In numerous bacterial single-component signaling systems,[48,49] the HTH domain is fused to C-terminal domains that bind effector small molecules, suggesting that this may be the general function of the duplicate SHS2 units that are associated with HTH domains. Consistent with this, in BmrR, the C-terminal GyrI-like module has been shown to be the principal determinant of drug binding.[38,47,50]

To obtain a better understanding of the regions that are potentially involved in ligand binding, we mapped the conservation pattern specific to individual families (Fig. 4) onto a representative three-dimensional (3D) structure of the GyrI-like module (Fig. 5). Such techniques have previously been used with considerable success in mapping potential interaction sites of different domains.[51–53] For this purpose, we constructed a midpoint-rooted tree of the GyrI superfamily using the entire alignment encompassing the 2 SHS2 domains, with the neighbor-joining algorithm (Fig. 4). The tree was then sectioned at different points, such that at each section the alignment was divided into n monophyletic lineages (in this case we choose sections with 3 and 8 monophyletic lineages) (Fig. 4). The consensus conservation pattern was then obtained

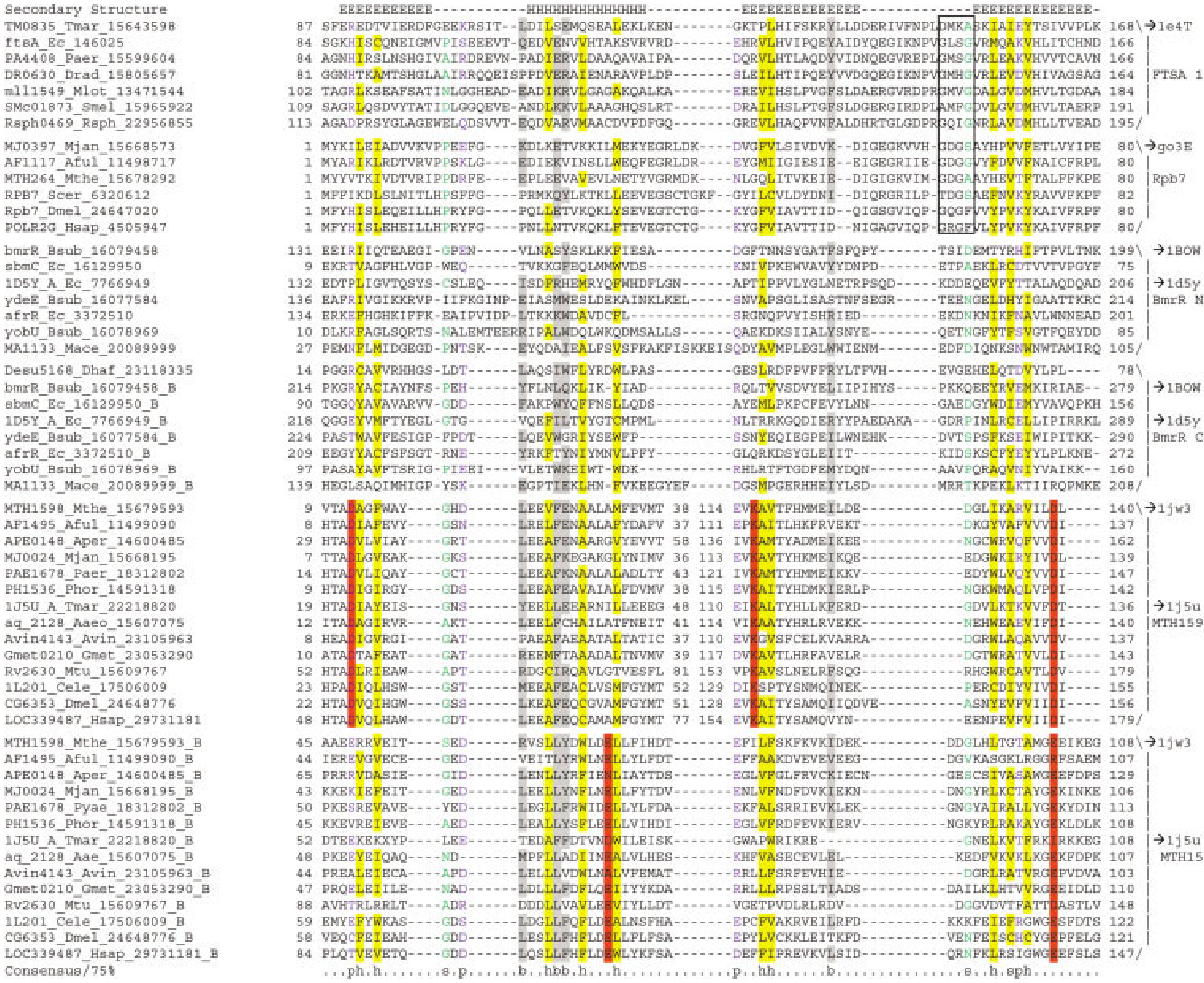

Fig. 3. Multiple sequence alignment of the SHS2 fold domains. Multiple alignments of the individual groups of the SHS2 fold were first generated (see legend to Fig. 4). These were then further aligned based on the structural alignment obtained through the superpositioning of the SHS2 domains structures. The 75% consensus was used for coloring. Residues specific to a particular group have been shaded red, while the loop shared by the FtsA and Rpb7p has been boxed. The species abbreviations are as shown in the legend to Fig. 4. Additional abbreviations: Aae, *Aquifex aeolicus*; Drad, *Deinococcus radiodurans*; Gmet, *Geobacter metallireducens*; Mtu, *Mycobacterium tuberculosis*; Tmar, *Thermotoga maritima*; Aful, *Archaeoglobus fulgidus*; Aper, *Aeropyrum pernix*; Mjan, *Methanococcus jannaschii*; Pyae, *Pyrobaculum aerophilum*; Phor, *Pyrococcus horikoshii*; Dmel, Drosophila *melanogaster*; Scer, *Saccharomyces cerevisiae*.

Fig. 4. Multiple sequence alignment of the GyrI family, which was constructed using T-Coffee after parsing high-scoring pairs from PSI-BLAST search results. The secondary structure from the crystal structures is shown above the alignment with E representing a strand and H a helix. The 85% consensus shown below the alignment was derived using the following amino acid classes: hydrophobic (h: ALICVMYFW, yellow shading); small (s: ACDGNPSTV, green); polar (p: CDEHKNQRST, blue) and its charged subset (c: DEHKR, pink); and big (b: FILMQRWYEK; grey shading). The limits of the domains are indicated by the residue positions on each end of the sequence. The numbers within the alignment are nonconserved inserts that have not been shown. A neighbor-joining phylogenetic tree is shown on the right. The terminal red dots on the tree denote the 8 monophyletic lineages obtained in the shallower cross section (separated by spaces), while the blue dots represent the 3 monophyletic lineages of the deeper cross section. The b3023 group in which the conserved E has been replaced by a hydrophobic residue is indicated by brackets to the right. The sequences are denoted by their gene name followed by the species abbreviation and GenBank Identifier (gi). Species abbreviations: Ana, *Nostoc* sp.; Atum, *Agrobacterium tumefaciens*; Avin, *Azotobacter vinelandii*; Bant, *Bacillus antharacis*; Bfun, *Burkholderia fungorum*; Bhal, *Bacillus halodurans*; Bjap, *Bradyrhizobium japonicum*; Blon, *Bifidobacterium longum*; Bsub, *Bacillus subtilis*; Bthe, *Bacteroides thetaiotaomicron*; Cace, *Clostridium acetobutylicum*; Ccre, *Caulobacter crescentus*; Ctep, *Chlorobium tepidum*; Cthe, *Clostridium thermocellum*; Ddes, *Desulfovibrio desulfuricans*; Dhaf, *Desulfitobacterium hafniense*; Ec, *Escherichia coli*; Efae, *Enterococcus faecium*; Fnuc, *Fusobacterium nucleatum*; Lgas, *Lactobacillus gasseri*; Linn, *Listeria innocua*; Lint, *Leptospira interrogans*; Llac, *Lactococcus lactis*; Lpla, *Lactobacillus plantarum*; Mdeg, *Microbulbifer degradans*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Oihe, *Oceanobacillus iheyensis*; Paer, *Pseudomonas aeruginosa*; Pput, *Pseudomonas putida*; Rpal, *Rhodopseudomonas palustris*; Rrub, *Rhodospirillum rubrum*; Rsol, *Ralstonia solanacearum*; Rsph, *Rhodobacter sphaeroides*; Saur, *Staphylococcus aureus*; Scoe, *Streptomyces coelicolor*; Sent, *Salmonella enterica*; Smel, *Sinorhizobium meliloti*; Smut, *Streptococcus mutans*; Sone, *Shewanella oneidensis*; Styp, *Salmonella typhimurium*; Tfus, *Thermobifida fusca*; Vvul, *Vibrio vulnificus*; Xcam, *Xanthomonas campestris*; Ypes, *Yersinia pestis*; Mace, *Methanosarcina acetivorans*; Mthe, *Methanothermobacter thermautotrophicus*; Cele, *Caenorhabditis elegans*; Hsap, *Homo sapiens*.

```
Secondary Structure            -EEEEEEE—EEEEEEEEEE-----------------HHHHHHHHHHHHHHH-----------EEEEEEEE----------EEEEEEEEEEEE----
BH0594_Bhal_15613157       128 MNYRMEQKEAFRVVGMKKRVKLVHRGTNTD------ITDMLGTISDETYMQIENL  3 EPGGILNVCTNFSEG----LEDEGELDYYIAAATIK--
ydeE_Bsub_16077584         128 MNYRIEEKEAFRIVGIKKRVPIIFKGINPE------IASMWESLDEKAINKLKEL  3 APSGLISASTNFSEGR---TEENGELDHYIGAATTK--
Desu2056_Dhaf_23113568     128 MNYRLVEKDEFYIVGFKKRITMQFKGINPE------MDSLVQKLTPQIIAELKSL  3 EPKGMLSVSANFDER----TAEGSQLDQYIGVATSQ--
Efae2873_Efae_22993358     135 MEFRIEKKEQFYVAGVAKKVPVKFEGVNQE------IQKLAQTITETQREEMHKL  3 YPNQVINVSYGFDGDR---LEEKGSLMHMIGFATTKE-
Desu2520_Dhaf_23114056     124 MEFRIEDKPAFNLVGVSKRVPMQFEGVNKE------IVKLAQSITAEQREEMHAL  3 EFYEIINASYDADADF---LKEEGDLTHLIGILTTEN-
OB0722_Oihe_23098177       124 MDVRIVEKPAFNLVGVTKRVPMQFEGVNNE------IVKLAQSITDAQQEEMHSL  3 EPYEVVNASYDADANF---LKEEGDLTHLIGVLTTLD-
SCO4223_Scoe_21222619      132 MATRFTDRPAFRLIGHATRVPLIHEGVNPH------IQQHIAALPREAHARLKSL  3 EPAGLLQVSDDVAQD----SPEGTELTYLHGVAVSAGT
Tfus0882_Tfus_23017809     179 MRYRIVEKDAFQVVGKKARVPLVYEGPNPA------IVEFVRSIDPGTLQRLAEL  3 EPRGVVAAVANIAGERVS-APEGTELDYYHGVVTSA--
BH3506_Bhal_15616068       130 MEPKFVTKPAFHIIGYELKTKNADGQNNKD------IPEFWQHYLKNKLGCTIPN  2 HKHVELGICTEF-------NPETGEFVYVIGMEVEKGT
CAC0426_Cace_15893717      128 MNYKIETKKSFKIAGVSKRISTKEGNNFKI------IPEFWDEVKKSGQCEIIEK  2 GKLGVMGVCTNF-------DCEYQEFDYLIAVEGDKI-
Desu2573_Dhaf_23114109     149 MNYQIEHWPAFKVMGISHKVKTAAAFEV--------IPGLWEKAWQDGTMGRFME  5 RPAGFLGIAAGGQ------WGDSEGMNYIIAVTNHVDV  6
ORF17_Llac_2467228         128 MEYRIENLDFELRIVGKSKPVKTSRAFKT-------IPTLWNTAKKDGFMQELVD  8 TLESLLGVCGKEA------AITDEQFSYFMGVRYDG--
VV21338_Vvul_27367708      123 VEVKVDTKEAFLLKGVRGEMNGLFSLKPNFAQV---VPLLWKNLEEAATNLPP--    LMNRRLGVVDVTQA-----SFDGSHIKYWAGIELDSNI  7
FN0315_Fnuc_19703660       127 MEISIKKKEKFIVGGIKAENIETFQ-----------CPKVWEELFKKVSFNDLEK  1 GNGNSYGVCYE--------TTSSKSINYIAAFDVKNVS  1
ORF00052_Llac_10957153     105 MNYLITNKPEMMITGIKESYPNITVGQAS-------IPKFWDRFNESDLFNQVIN  5 SPNTILGICLP---------REGEAYDYFIGVYTDEK-

LA0598_Lint_24213298         7 KIMKECLMPKKYLIGIGTRTKNENEMGSNGN-----IPKLWEKFFGEVLPKLKT-    TDKFIYAVYKDYE------SDENGEYSYFIGVPS----
SO1757_Sone_24373324         1 MTMELVFMAAQPMLGLCTRTNNRTEMASDGGK----IAGLWRAFFESSQLTSM--    LDSPMYGVYYDYE------SDMTGDYSVLVGKCVDSA-
yobU_Bsub_16078969           2 GFSHITHLDLKRFAGLSQRTSNALEMTEERR-----IPALWDQLWKQDMSALLSQ  2 KDKSIIALYSNYE------QETNGFYTFSVGTFQEYDD
CAC3493_Cace_15896730        1 MEYEIVKLEEKIVLGVSAVTSNDDPNMGKV------IGGLWEKLYQGGINETIKN  1 VNEYAIGLYSDY---------EDNKYVVTVGNEVCK--
Desu1867_Dhaf_23113368       1 MNYEIVTLPEKIVVGVTARTSHTDPQCQQV------IGGLWQKFMGEGIWVSIQN  1 ANPYCLGLYSGY---------DETSYDVTVGAEVTK--
PA4878_Paer_15600071       120 MHARIVERPAFSVVGMEYFGSAPGDT----------IGQLWERFIPREHEIAGKH    DPEVSYGICAQQPN-----GEFHYVAGFEVQEGW----
BH3633_Bhal_15616195         7 LPFTIVEQNERKMIGLKLEGPYTRMNE---------IGMLWETFNQRVSEIDHLV    QDDLSFGIVQDRE------RDFTYYAAVEVSSFT----
Desu2067_Dhaf_23113580     132 VHPKIVELPEIKVAGLRGETTLRDNR----------LRELWDRANSLYIQIPNRI    PNGRAFGICEACAENTLYTMNDDILFTEVAGIEVSSFA
BH3634_Bhal_15616196       130 HKPYIVTRQPLQVVGLQLEGLSLPGGSDREQTTQVAIPSLWHQLEARIHEIERRV    DPFVSYGISSP--------SRDGSHFTYMACVEIEAGD  1
SMc03170_Smel_15966654     126 EPPRFEESPGLLLAGLAETYDYNRTEG---------IPSLWQRFNAYFGSIPGQ-    HGNIAYGACTQS-------DGEAGRFRYMAAAEIRDAE
AGR_L_698_Atum_15890465    126 QPPEMRRADAPMVVGLSTPCSLENNST---------IPALWQRFNLRESEVEEV-    VSGAAYGVCSA--------ADEAGNCTYLAGVKAL---
bll6698_Bjap_27381809      137 APPRPETATAPLVAGISERISCDNGAI---------IPGMWQRFHQEVADIPAR-    VGNVAYGVCCNG-------DDAGNFDYIAGVEVSDYS

SAV1506_Saur_15924496      128 YPYRLEETDDISLVGYARFIDTKYLSHPFN------VPDFLEDLLIDGKIKELRR  4 SPFELFVISCP----------LENGLEIFVGVPSE---
lin0911_Linn_16799983        4 RQGRLEEWEGFTGIGLVHEGLKTEAFHTG-------IKTAFKEMLKLAQELDDFS    ELKEVYGISVHNIE-----DGITHYAVIPVEQKYAHL-
BH2119_Bhal_15614682         1 MSYDILTLAAYRAIGLKWEGTFSEIVPN--------LKNVIQQMEDRADELEHKI    NSNIQLGLSYHTIE-----NGFVHYAVYEVSEEQEI--
YPO2243_Ypes_16122471      128 PEIKQVTLPGKELVGFTRRLDFEEYNG---------CAVQRSSCMAMKDEILLDF  8 QRIYSLFSVKDVDG-----QQGQKSVYYSTAIDKERKH  2
1D5Y_A_Ec_7766949          124 PEHKFVTLEDTPLIGVTQSYSCSLEQ----------ISDFRHEMRYQFWHDFLGN  3 IPPVLYGLNETRPSQD---KDDEQEVFYTTALAQDQAD
STM4586_Styp_16767827      126 PEHQFVTLEDTPLLGVTQSYSCSLEQ----------ISDFRHEMRVQFWHDFLGH  3 IPPVLYGLNETRPSME---KDDEQEVFYTTALPQEQAD

YPO0456_Ypes_16120785      126 PQPEFITLPEQHLVGITQSYSCTLEQ----------ISTHRAELRLHFWQQYLGD  3 LPPVLYGLHHSRPNPE---KDDEQEIFYTTAIEPQHIP
STM1671_Styp_16765014      127 YQFEICQLTSKEIPGFQTSHQIATND----------LPKKASPIKWKIIHETLRT    WGENVYCLSSFKPD-----NTKDQVIAVSSFFGMEHNS  1
YPMT1.81c_Ypes_16082873    132 LQPRICYLKERNIIGQCFNFRD--------------LVFYSGIDSKCRLGKLYDS  1 KKNTAITVSNRIPF-----HDKTNDIIARTVVWDR---
afrR_Ec_3372510            126 PSPDIRYMERKEFHGHKIFFKEAIP-----------VLIDPTKKKWDAVDCFLSR  1 NQPVYISHRIEDE-------KDNKNIKFNAVLWN----
ydeR_Ec_13096081           126 PIPELYYLPQRKFTGISLKYKEKIP-----------YTPASSKIKWDVVQSLLLK    QTSLFISNNTMQ--------GSRRKNEFIINSIIWE--

MA0989_Mace_20089866       128 TEPVIKEIPELRVLGKREKGTFVVT-----------IGKLINEICACVSSPENQR  1 RVKTTGPIMFLCHD------EEYKETGADIEITLPVSG  3
MTH628_Mthe_15678656        21 MEISEKMVEGIRVAYIECRGSYER------------IPEYISEVAGWVLKNGLQ-    MTGRVYGTYYNTPE-----EVDEEELLYEIGVSIAGEA
MA0499_Mace_20089388         3 SEVTIVELSPQFVLGIRKTGAYRE------------IPVMLNRICEFAFSKNI--    -QITGYPVFLWHET-----TVEEAQKAEVDENADIEV 10
BT1189_Bthe_29346599         5 SEIMLLQQPEQPALAIEVQTYMKGMSQA--------IGENFVRIDSLFKKQGE--    VTTDIPFVEYPDFE-----SLTEDRIKMIIGLKSSKPL
all0345_Ana_17227841       139 YEVVIKKVAPIQVASIRQILPDNPS-----------IGQLYGEISEYLAQNGVK-    AGDYYAGIWHDPGY-----KDTDIDAEAVISIEGSI--

bltR_Bsub_16079711         124 SSISFEYLNEETFMLSRKTLNLPERK----------YVAAISELIHEVQQYELD-    EGYPIGGIFAREQI-----LEKDFYNYSYFYIKVK---
CAC3443_Cace_15896684      122 FEKVRFTYEDKKAYMIEPCYGKDEKS----------YMQSFISICNKSKELQID-    FQNPICSIITKEAL-----KSEDYKDVSYFGIRIPKD-
bmrR_Bsub_16079458         123 GEVFVLDEEEIRIIQTEAEGIGPENV----------LNASYSKLKKFIESADGF-    TNNSYGATFSFQPY-----TSIDEMTYRHIFTPVLTNK  3
Ddes0467_Ddes_23473562     128 QEVSIKYIEPSRYLFLNQTYDTNIKA----------AIINIDFTNYVESLNNE--    ITGPVILNFSSHAA-----RMKDKEQPVRILQKTL---
FN1743_Fnuc_19705064       121 DHSYILHFEKRYVVAVKILENEPKED----------FHIRLNELRNNEKYKNLK-    YMRQFLYIADYDA------LIEGNLKPYYLGMPIKES-
ydfL_Bsub_16077613         121 HQISKIKLPAMRVAYLQHEYVLGHD-----------IEHSLAELRTHLNVNED--    IFIGKIGLSISAAN-----VKAKQFDKYSSIFMILED-
Desu0349_Dhaf_23111654     121 DKIEERTLKERKIAVLKKELAISDD-----------LEQPIRELAKRNSLHAV--    MFLGKVGVSISSAN-----LQRGNFDKFSAVFVVIEP-
CT0179_Ctep_21673020         7 FVCELKELAPVPALLIRTQTTMSELGSL--------FEAGYHDILQLLAGQG---    KSPSGPPFARYF-------GMSAGTFEVEFGFPVEGGV
Magn3564_Mmag_23010076       1 --MDPVTLPAKPVAILSGQTKWEAARAN--------LRASFKTIGETLAKLGLK-    PAGRPIALYTK---------TEDDGFQYEAMIPIESAP  3
SCO0140_Scoe_21218699      120 RAVTLEELPARRVLAVTLDVPEGAG-----------LDWYDEAMCDVDSAAGER-    SVLPPGGRYEHTLF-----TEGHGRATVYVPFEAPLP  1
SMU.1470c_Smut_24379865     17 AKPIFLEVEEQRFITIKGKGNPNDQDFSNR------VSALYALAYGIKMAYKQAM 12 AVYPLEGLWQQAKD-----AKEDTLEKDKLSYTIMIRQ 22
lp_3071_Lpla_28379488       19 KQPQLLTIPAQTFMSIHGTGNPNGPEFQTH------LQTLYPAAYGLKHAYKQYA 10 VVFPLEGVWSLTIKG----QQLDHLDKDEFSYDIMIRV 21
MA1133_Mace_20089999        19 KEVSIIDVPEMNFLMIDGEGDPNTSKE---------YQDAIEALFSVSFKAKFIS  8 AVMPLEGLWWIENM-----EDFDIQNKSNWNWTAMIRQ 22
Magn3916_Mmag_23010599      19 REFCEIHVPTLTYLKVDGAGDPNSAAA---------YREAITWLYGVSYAVKFAA  7 VVPPLEALWWADDP-----GSFVRREKETWRWTVMIPA 21
BL0980_Blon_23465549        17 RMPAIVTVPAMRFVAVDGVGDPNEEGGD--------YAKAMQLLYGISFTIKMNK 15 TVPPLEGLWSMEK----GVPGVDYTRKTDFYWTSMIRL 22
Chte2144_Cthe_23022091      17 TEPEIIDVPQMNFIAVRGKGDPNEEDGA--------YKQAVNILYAIAYTIKMSN 12 VVPPLEGFWWQE-----GVEGVDYSQKDKFNWISVIRL 21
Lgas0746_Lgas_23002903      17 KQPEIIRVPKMNYIAVSGSGDPNQEDGT--------YQKALGLLYGLAYTIKMSK 12 VVPPLEGLWWSKDQ-----QKIDYAHKENFAWISMIRL 21

LA0433_Lint_24213133        28 VLVQEEMKGPPYVLSHERIGDYRN------------VGLTFEALQKELPEKGI--    RNFKLFSIYLDNPN-----EVPKEKLRCEVGALFSEPL
ZSIG11_Hsap_7706708         36 VSAGSPPIRNVTVAYKFHMGLYGE------------TGRLFTESCSISPK-----    --LRSIAVYYDNPH-----MVPPDKCRCAVGSILSEG-  2
2G526C_Cele_17536487        35 TTSPKNLDKPLTVYYKYHLGPYQN------------VMNVIGEAKQLLASSP---    TPATFFGIYYDNPE-----VTDSHFLQSAVGVVFGSDG  1
sbmC_Ec_16129950             1 MNYEIKQEEKRTVAGFHLVGPWEQT-----------VKKGFEQLMMWVDSKNI--    VPKEWVAVYYDNPD-----ETPAEKLRCDTVVTVPGYF  3
STY2266_Sent_16760993        1 MDYEIRQEQKRKIAGFHMVGPWEHT-----------VKQGFEQLMTWVDRQRI--    VPVEWIAVYYDNPD-----VVPAEKLRCDTVVSVAENF  3
VV20787_Vvul_27367197      150 STMDIQTFPASKIAYIRVTGEYGKN-----------YESATQKLYQWAGPRGL--    AGNTCIFIYHDNPE-----ITPADKCRTDICLLNCENA
CAC3490_Cace_15896727        1 MDTNIEMIPAYKIAYIRRTGAYGLDN----------VQI-MEQLKSWAREENLFN    ESSIILGIAQDNPK-----VTEPKDCRYDACLVVSDEF  1
AAN07145_Saur_22773947       1 MNYKIEILDDCNVIYVRNKGKYGSNKN---------YEM-MKNFKEWIKENCYWR  2 ETNGILGVALDDPQ-----IVEEESCRYDLVLKIDEDV
BH0401_Bhal_15612964       143 MNISIKELPDYEVAFVRHVGSYLET-----------YKA-WATLGAWASENRLDP    PQSYFIGISLDDPK-----EVEEHLCRYDACVTLPEGF  2
BA2733_Bant_30262706         1 MKVIIKNLPSLEVAFIRRTGSYFE------------PQDHWGKLLNWSIENKLYP    LEQSFIGISLDNPE-----LVASHMCRHDACVTIPKNF  2
VV21661_Vvul_27368002      142 PEPKITEVPERMAAYVRHVGYNRS------------IKNAWLILKAWANSEGR--    SFEVQFGLHHSNPA-----WVELDKCRYVACIAIDKPL
BT2372_Bthe_29347782       161 TKIEVKEMPDMKAVYCRHMGAFKE------------IVKAYEKLIKWAEPRGLYI  1 NVTKSATVTHDDPS-----VTELSKVRQSACIIVEGDV
lin1814_Linn_16800881      141 GKVTITTLQNIPVIYKRIVGSYKELE----------LTNPLSELFQYGMEHDLLD  1 DSSFPLTIYHSHPD-----ITTADNQRASSCIIIKQNV  2
BT1904_Bthe_29347314       131 LKSEIKSIPARNVIYIRLSGDYKLND----------YGGTWGRLWQFIKEQKLPM    GDFSPLCIYHDDPK-----VTPAEKLRTDVCMVMPVQV
BT4142_Bthe_29349550       216 IEPSIERVPHTRIAYLKLERTHHVSHS---------FSVLWKQVLQFSESYGLLS    KGCKYVSLTLDYPF-----ITLEEQSRFMVGVTLPQSF

Ddes0156_Ddes_23473255     159 LEPEVKVLAPVRVAFVRHTGPYAQ------------CEAAWKTLCDWAFPQGLVM    AQTQFMGICYDDPE-----VTAPDKIRYDACISVPDNV
Ddes0836_Ddes_23473930       1 MDIRVVMLSAFTVAGIRAYGPYSVS-----------APEAWNTLAPWLARMEKSG    AELSYWGIMHDDIH-----ITQPEKIRYDAMVAVPADL
Rrub0570_Rrub_22965975       2 PALDVRPLPAMIFASLRHIGPYGE------------AGETFRKITDWAVSEGVMT    PETQILGLSYDDPK-----TTPAEILRYDACVTLKTPV
Avin0350_Avin_23102175     170 AQVRVESFPSQPVLAMRFYGSYAL------------VEDNWRRFAECLDRAGFPL    ADAQAVGIVQDNPE-----ITPNDLVRYDCAIVDAGFD
Rpal2152_Rpal_22962859     142 WSIEIQRFDAIPAFAIRHDGPYIE------------IGKAFGMLFGQLAARGALP    ERIEMIAVYLDDPT-----AVPLERLRSFAALAAPGGK
Bcep3966_Bfun_22986028      38 REVVIRHVEPIELLSVDHVGPYPQ------------IGKAFDALFGWLAKHNLLA    AQMRVIGVYYDDPS-----VVEESALRSKAGVLLPHPV  4
Mdeg0819_Mdeg_23026984     147 HTVDIEKFNKVELGGLAHQGDYLD------------IGAVFEKVFVSAGSKQFLN    EHSRSFGIYYDDPT-----SKDKSALRSHACVTLNPQQ
Rsph1750_Rsph_22958167       4 FPVHIRETSPRRLAALRHTGSFEE------------IGATFGQICQILEERSLLS    QAGCMIGVYWDNPL-----LAPPGALHSHAGIDLPEAM
RSp1247_Rsol_17549468      157 PHVELIELPPIEVVCLRHDGPVAT------------IGQTFRTLMRMLHTGQALP    GTPERIGICCGDPE-------VRDTFRYYAAAVPPAR  1
XCC3056_Xcam_21232486      164 LQVTVQWLEPLEVVVLRQRGAFDD------------LDRGFGRIAAWAERAGVIE    HLHALIGVPLSDHR-----DVPAQQHLFECGIAFATAV
b3023_Ec_16130919            6 LDVNIIDFPSIPVAMLPHRCSPEL------------LNYSVAKFIMWRKETGLSP  1 NQSQTFGVAWDDPA-----TTAPEAFRFDICGSVSEPI  1
STM3175_Styp_16766475      134 MDVKIVEFPPTRVAMLTHLGHPDK------------VNASAAKFIAWRRETGQSP  1 ASSQTFGIAWHDPQ-----TTPPAQFRFDICGSVRQPI  1
PP2173_Pput_26988897       135 MQIRIVNFAETHVAALEHQGPPGL------------VSESVARFRQWRMHSGQSP  1 ASSRTFGIPYDNPD-----TTPAHAFRFAVCGEIDEAV  1
blr4000_Bjap_27379111      140 DDVTIRDVPPTRVAIMEHRGDPAT------------LPATIQRFIAWRRAANLHP    RTSPTFNVWRSERR-----PASPADYSVDLCAGTDQPI  1
mll4902_Mlot_13474099      129 DDVTIRDVPPTPVAIMEHRGDRAT------------LQDTIQRFIAWRKAAGLSP    ETSPTFNVFRSERE-----PAIAADYSMDICVGTDQPI  1
CC2527_Ccre_16126766       187 DDVRIVDFPDTPVAVMRHEGDPAL------------LGDTIRRFIAWRRAAGLPP    RVSATFNVFHDDPD-----DTPAEQYRLDLCAATARVA
RSc0132_Rsol_17544851      112 HAVEIIEREDVPVAAIEHRGDPAR------------LGETLRAFIAWRRANRLPP    AVSATYNIVYDNPD-----DTPPEAFRMDICAATPAPV  1
Consensus/85%                  ....h.......hh.hp..................h...h..h...................hsh....................hs.......
```

Figure 4.

```
Secondary Structure          ------EEEEE-EEEEEEEEEE------------HHHHHHHHHHHHHHH------------EEEEEEE-----------EEEEEEEEEEE
BH0594_Bhal_15613157         KCPEHLVELEIPTPTWAVFTVEGSW------EDVQEMWGRIYSDWFPTSDY------EHAEGPEIL-----------SSANEKSEIWIPVV 279
ydeE_Bsub_16077584           RCPDNFSRLEVPASTWAVFESIGPFP-----DTLQEVWGRIYSEWFPSSNY------EQIEGPEILWNEHKDV----TSPSFKSEIWIPIT 288
Desu2056_Dhaf_23113568       AMSNHYDLLHVPAATWAVFKAVGTFP-----EALQDTWAKIYAEWFPASGY------EMTGGPELLWNETPDT----SKPDYKSEIWIPVR 287
Efae2873_Efae_22993358       NPYADLEMLAIEEHTWAIFPNKGPFP-----QTLQDTWGKIYAEWLPSSDY----GLVEAPEISFTKWGE-------DFSNVYSEIWIAVK 295
Desu2520_Dhaf_23114056       QVSDLLEKVPVEACTWAIFPNEGPFP-----SMLQDTMAKIYSEWLPSSNY----EVIKAPAFSFTKMNQH------KKDCAYSEVWIPVR 285
OB0722_Oihe_23098177         QVSDRLEKVSIPACTWAVFPNEGPFP-----DTLQQTMARTYSEWFPTSDY----EVIEAPSFSFTKMNEY------KENYAYSEIWIPVR 285
SCO4223_Scoe_21222619        PAPDDLDAIEVPAGTWAVFRSSGPYP-----DALQTTWASTASEWFPSNPW----RLRPGPSIVAVLDRAD------DFTTATTELWLPVE 293
Tfus0882_Tfus_23017809       PAPEDMAVLPVAAGLWAVFEVEGEAP-----YAIQYLWRDVFTQWFPANPY------QSVPGPELLRTTMSQ-----DGAYVEAELWVPVE 340
BH3506_Bhal_15616068         KAPEGMVYKSFPELEYAVFTTPKANEESFT-SSIQSTWNYIFTEWFPQSDY----EHNGVVEFELYDERCH------GTENIEMDIYIPVK 291
CAC0426_Cace_15893717        EGLDNYVVIEVPELSFAVFESIGPMP-----DALQDVTRRIFSEWFPATKY----EEAEGPEIEVYLPGNP------QDKDYRAEIWVPVV 284
Desu2573_Dhaf_23114109       PVLEGMEEFAFPAAAWAIFEANGELP-----DATQKVYKQFYTEWLPNSGY----ALADLPVIECY------------MQENRQEVWIALV 308
ORF17_Llac_2467228           ESPENMETLIIPASTWAVFPN------------IVDAWKRLYSEWVPTSEY----ELANLPCIECYYG---------PKHKPRHELWVPVI 279
VV21338_Vvul_27367708        LISDQLETLAVPQQTYAIVKHKGPIE------RLPKTLEWFIIHWLPNSGY----RGIDGYELEVYPTDYHP-----NSLNAEMEYWVPIQ 288
FN0315_Fnuc_19703660         AKKLGLDTMEIPEAEYAVVKLKGKIP-----NCIHEGWKYVMEVFFPEHGY----KHAGTPDFELYSEGDM------GSDNYEMELWVPII 278
ORF00052_Llac_10957153       TTADKLETITLPASDWAVFKAVGKVP-----EAIHQTYQDIYESFFPSTTY----TQKKAPDFESYPLDLNP-----MSENHITEIWIPIH 262

LA0598_Lint_24213298         DEINIFETVQLPEGKYLELTSLKGKSP----NIVIELWQKVWTNSDIKR------RRAFEVDYEIYPID--------FSETPETQVHLLLS 156
SO1757_Sone_24373324         SETGPFIPLQLCEGNYLKFSAQGEMP-----HCVINLWGEVWGYFSASDCPH---RRCYQTDFEVYRSA------------DKVEIYIGVL 151
yobU_Bsub_16078969           ILPGPYENIDLPASAYAVFTSRIGPIE----EIVLETWKEIWT-WDKRH------LRTFTGDFEMYDQNAA------VPQRAQVNIYVAIK 159
CAC3493_Cace_15896730        AENEALTIKKIPAGKYAKFSIEGHME-----KAVAEAWSKIWQMNLDR-SY---------EADFEEYLNS--------DFNNAKVDIYISLK 146
Desu1867_Dhaf_23113368       NGNPELTEKIIPAGSYALFRIKGDVV-----KDVAEAWDKIWTLPLER-SF---------TGDFEEYLSN--------ENGVAEIKIYIALK 146
PA4878_Paer_15600071         PVPEGMVRFQVPAQKYAVFTHKGTA------PQIAESFQAIYSHLLAERGL----EPKAGVDFEYYDQRFRGP----LDPNSQVDLYIPIY 270
BH3633_Bhal_15616195         KVPEGMATIILPACRYAVFTHKGPQDKFSETVVAALKSLKESGYEKDSDNY------WLEGYDDHRFSP--------ECLDNSEDIYFPLK 157
Desu2067_Dhaf_23113580       GLTEPFVQKIIPGGRYAVFTHRGTL------GMLPQTFDYIWGTWFLTTKE----EMDWREDFELYDERFLGY----DHPDSEVDLYIPIR 291
BH3634_Bhal_15616196         NIPQGFIQKEIKGGTYASFHYKGAA------NQLSALTNYIYGSWLPSSQY----QSHFGVEIEVYDEREPK-----EDGEVAVEIWTPIK 291
SMc03170_Smel_15966654       ALPSGFSTLKLPRQRYAIFVHRGHI------SGIANTAHHIFTTWFPQSGY----RHGELPDLMERYDERFDP----HSGMGAVEIWVPLK 278
AGR_L_698_Atum_15890465      TKTPGMDYIELPAQSYAVFTHNGHI------SDLPRTVYTIWNKALPEAGL----KTADSPDFERYDQRFNP-----ETGRGSVEIWIPVI 273
bll6698_Bjap_27381809        DLPRRFGRIRIPEQRYAVFAHTDHV------ASIRRTVNTIWNQWLPASGL----KAADAPSFERYDEKF-DP----ATGNGGFEIWVPVR 287

SAV1506_Saur_15924496        RYPAHLESRFLPGKHCAKFNLQGEID-----YATNEAWYYIESSLQLTLPY-----ERNDLYVEVYPLDISF-----NDPFTKIQLWIPVK 281
lin0911_Linn_16799983        KEPLEWIEVPAHTYFVAEHIAETDILESYEEIARAIQQKEYKPYITANNPV----FDPLPFKLEIYAKN--------GIEESADEIRIPVV 162
BH2119_Bhal_15614682         PDGMMEIKVPEWTYVKTTHNKGEDIQKTYQDLHQWLFDSDYTVIREDGVDYY---DPYMPIKHEHYPVDR-------FDPNDPHDIYIPVV 159

YPO2243_Ypes_16122471        QGSREIDHIAIPEGEFLSISHQGSA------KECVKFATYLFNNVLPKLRD----EVGRGIEMEVIEIDHCHAESKLRDISVVYHYLMAVN 297
1D5Y_A_Ec_7766949            GYVLTGHPVMLQGGEYVMFTYEGLG------TGVQEFILTVYGTCMPMLNL----TRRKGQDIERYYPAEDAKAGD-RPINLRCELLIPIR 286
STM4586_Styp_16767827        GYVQSAHPVLLQGGEYVMFTYEGLG------TGVQDFILTVYGTCMPMLNL----TRRKGQDIERYYPSEDTKTGD-RPINLRCEFLIPIR 288
YPO0456_Ypes_16120785        CNVPEGQPVILQGGEYVQFSYDGPL------DGLQNFILTLYGTILPQLAL----IRRRGYDIERFYPQGRPKDG--PPATLKCDYFIPIR 287
STM1671_Styp_16765014        DGGKIPMSRVIKCGKYAKFHFVGHK------EQYQQFSNTIYMCILPKLNL----IRREGEDIEYFHLASVQKQQN-NESIVDLYYYIPVL 285
YPMT1.81c_Ypes_16082873      NKHFSDSEIKVDKGLYAYFFFNDTY------DQYVHHMYNIYYNSLPIYNL----NKRDGYDVEVIKRR--------NDNTIDCHYFLPIY 276
afrR_Ec_3372510              NEADNSISEELEEGYYACFSFSGTR------NEYRKFTYNIYMNVLPFYGL----QRKDSYGLEIITKI--------DSKSCFYEYYLPLK 270
ydeR_Ec_13096081             KSETSDCEYSIEGGVYALFRYTGKP------AGYSDHINNIYLNALPFYGL----QKKNSFDLEIICID--------AHGGRYFEYYLPVK 270

MA0989_Mace_20089866         VEDPKMEIKTLPAMKAISVIYRGPY------HEVEVAYNRIFS-FAEENSL-----ETILPSRELYFNDPAEV----PEEELMTEVQVQFR 282
MTH628_Mthe_15678656         EADERVKIKNLPSHRVIAALHEGPY------TEVGPVIHALIE-YAMENGY-----EITGPVTEVYLNSPLEV----DESELLTEVQIPVK 170
MA0499_Mace_20089388         TGDEEIKVYELPGGKMAKVVHEGPY------EESILTYEKLFF-WITENGK-----RIAGPVREVYLNDPQEV----KPEEILTEIYAPLE 159
BT1189_Bthe_29346599         QGDEKIQSVILPARRIVVCLHRGNY------NELAQLYNEMTE-WIKTNGY-----KASGTSIEYYYSNP-DV----PEEEHVTRVEMPLL 156
all0345_Ana_17227841         KGNERIKIYELPGSETTACLIHHGSY-----ETLAQAYATLVS-WIEANGY-----NITAPNREVYIIGGNEQ----NNDSYVTELQFPVA 288

bltR_Bsub_16079711           DGAENINYHVRPKGLYAVGYEIGG--------NTEEAYRRIIE-FIERNGM-----QIGENAYEEYMLDEMVVD---GYENTYAKILLQVK 271
CAC3443_Cace_15896684        FIGHIVSRFEKPDGIYASTYHKGSY------DTMYLSYKRLIE-NIKEQGY-----EVCGNAYEVDLLSTLTSV---SSDEYLKLISIQVC 273
bmrR_Bsub_16079458           SITPDMEITTIPKGRYACIAYNFSP------EHYFLNLQKLIK-YIADRQL-----TVVSDVYELIIPIHYSPK---KQEEYRVEMKIRIA 278
Ddes0467_Ddes_23473562       MPCHDEYTTDMGGHMMAACYHIGPH------EDIAETYKKIRR-WARQHGY-----TLGDESYERYVTDYWTTR---NSAKFVTEIMIKAS 276
FN1743_Fnuc_19705064         PNILSENIIEIPAGNYLCFKARILS------ETWNPYFAKLFF-HGKEKPT-----IVLANEYENNLHEYL---------SSVYELQILIL 265
ydfL_Bsub_16077613           ENEKTSSEIIFPSREYLQIRFKGSH------PEAEPYYKKLLA-YMKEHHY-----EVAGDSIEITLIDYGITN---NLDNYVTEILLPIK 270
Desu0349_Dhaf_23111654       EDNCQENLSCLPGGNYLTLRFQGTH------PDAEASYIKMLQ-YMKENGY-----ELAGDSLEFALIDYGLTN---DPSKFITELQIPYK 270
CT0179_Ctep_21673020         EGSGRVVTGLTPSGKAASSLYIGPY------GEIEAVYDALMK-WVDDNGF-----DLSGEAYEIYLDNPAET----APDQLRTRVSLMLH 156
Magn3564_Mmag_23010076       TETNGVTFGSSPSGKALRFKHSGTY------EEIDGTYETLTA-YLDANEI-----DVQDRFLDEYVTALGEG----ADDKLDIDIYALPP 151
SCO0140_Scoe_21218699        EMPGRVRELHLPRRTAAVATHQGRH------DDLDLTYGAVGS-FAARNDL-----RSQDLVEEVYLVGPRDTD---RPELWRTLVAWLIA 271
SMU.1470c_Smut_24379865      SFYESISFESIKDGKCIQVLHIGSY------DDEPQSFAKMDA-FAKEHGL----KRSSDIHREIYLSNAQ------RTEKSKLKTILRYQ 206
lp_3071_Lpla_28379488        LPTEQINVQQFDAMQVAQILHLGAY------DDEPASFAKIDE-LVANQGM----HRTSKIHREIYLSDAR------RVAPDKLKTILRYR 206
MA1133_Mace_20089999         PALARLRFESPHEGLSAQIMHIGPY------SKEGPTIEKLHN-FVKEEGYEFDGSMPGERHHEIYLSDMR------RTKPEKLKTIIRQP 205
Magn3916_Mmag_23010599       PRPVSLRFEPYAEGRSLQILHIGRY------DDEGPTLARLHHEVMPGRGV-----TFNGPHHEIYLSDPR------KTEAAKLRTILRQP 199
BL0980_Blon_23465549         VDINRAYLFDFDEGVVAQVMHKGPY------DDEPATVTILDG-YARSQGYGLD-FSDARRHHEIYLSDPR------RAKPENLKTIIRHP 211
Chte2144_Cthe_23022091       IDCSLAEFLTIKEGLCVQALHLGPF------DDEPKTVAKMDE-YLIKEGYEKD-FSNERLHHEIYLSDPR------KVEPAKYKTVIRHP 206
Lgas0746_Lgas_23002903       QDFSQAKFFTYDEGLCVQIMHTGSY------DNEPATIEEMHQ-FVKKENYQID-IQNPRYHHEIYLSDPR------RTKTERLKTVIRLP 206

LA0433_Lint_24213133         EKIPNGLSLELKIRTIPSKKYL---------TAEFPLRNFLSIPLGIYKVY----PKLFRACEERGCDLKGRASEIYEPLTEHKTTYLLPL 180\Secreted
ZSIG11_Hsap_7706708          SPSPELIDLYQKFGFKVFSFPAPSHVV----TATFPYTTILSIWLATRRVH----PALDTYIKERKLCAYPRLE---IYQEDQIHFMCPLA 185|forms
2G526C_Cele_17536487         DFHSEKYSKDLTDNGFEKLVLPNVERAVQ--VTQPSTGGFASFLALVWFTY----STIRKYISENKLETTYAVE---FYTNTEINVIFPLD 190/
sbmC_Ec_16129950             ENSEGVILTEITGGQYAVAVARVVG------DDFAKPWYQFFNSLLQDSAY----EMLPKPCFEVYLNNGA------EDGYWDIEMYVAVQ 153
STY2266_Sent_16760993        DNSEGVIVTAIEGGEYATAVARVED------RDFAKPWERFFDVLEQDSAY----QIASAPCFETYLNNGM------EDGYWDIEMYIPVQ 153
VV20787_Vvul_27367197        EGNGELEIKDFPGGEYAFIRKTITD------NAQYATAWDELMARLVERGL----ESDERPCFELYHSYDP------QTQHADVSFCTAVT 299
CAC3490_Cace_15896727        VNNNYINFGETPGGKYCVFKISHTA-------DKVQKAWMEIFSELSKRNYK---LDDRRPILERYAVQ--------MINNHYCEICVPIL 151
AAN07145_Saur_22773947       KINDFINSRSFTGGKYVVFALPHTT------KDVKDFYSNLPN-IIDNNNL----RVKNEPILERYKEE--------EGKDKYCEVLIPIL 152
BH0401_Bhal_15612964         KPHEDIQFKGLSGGTYALYSFYDRV------DKLGIVYQSIFTQWLPASGF----VADDKPSLEFCMNDPALD----PEGKAKVDLYIPIR 297
BA2733_Bant_30262706         EQHEDVQFKSVDGGLYALYQFYDEP------HKLSEVYRYMYAVWLPNSEYS---ADYDRDNLEFCMNNVAED----LEGKLKVDLFVPIK 156
VV21661_Vvul_27368002        KYRGVVNQMVIPGGLHAVFRLHGVY------GELLPQISMVLEKWLPASGF----KLRSTPAYVHYHSNHFVN----ESEAFELDFYLPIS 292
BT2372_Bthe_29347782         KGEGEIGNLVIPGGQYAVGHFELGT------EDFEKAWNTMCK-WFTESGY----QQGDGCTYELYHNSHRTH----PENKHIVDICIPIK 313
lin1814_Linn_16800881        LDSEEIGLMEIPSGLYGVIHFEIRQ------KEYPDAWDFVYGNWLPSSGY----LPANSFPFEVYLNNPLED----PLQKHIVDIYIPIE 298
BT1904_Bthe_29347314         APKGDVGFKTLPAGRYAIFLYKGTY------DNLQAVYDTIYGKCLPEMEC----TLRDEPSAERYLNNPCET----APEELLTEIYIPVE 285
BT4142_Bthe_29349550         KIPKGFGVYEVPAGEYAIFRFKGLY------HELNRVYRYIYLDWLPANDY----SLREPFTFETYINTPEKT----PVSELITDIYIPVK 371

Ddes0156_Ddes_23473255       QAEHPVAIREVEGGEYVCAVWKGPY------TGLTNAYAALCGQWGGRSGR----EFRAAPSVEVYLNDCTST----PEQELLTEIRMPLM 311
Ddes0836_Ddes_23473930       RLDGSVFARRMDACEYAMAEHKGPY------HEVQAIWLSLFWDWLPFSGR----CPATLPCLEQYMGNPVRT----PSALLTTRLYLPLA 154
Rrub0570_Rrub_22965975       ALPPGMQLMALPACSWVMMTHKGPY------ETMGETFTKLYAALAEHPTL----VPIATGAIEIYVNDPATT----APADLITELGVAVL 154
Avin0350_Avin_23102175       REHPALSRLILPAARFACLEHRAPY------SHIFPVYRTIGSVWGPRSGECF--GAGGTGTLERYRQPPWLN----GGGEQWLDVTIALK 324
Rpal2152_Rpal_22962859       DLDPPVERIEIAGGDYAVLRHKGPY------ADMSAAYNWLFGTWLPQSDR----EADDRPVVEKYLNSPQDT----KPSDLLTDLCLPLR 294
Bcep3966_Bfun_22986028       TLSPPVSVAHLKGGDYAVLEHRGPY------SDLRAAYDWLYGTWLVQSGR----KAADAPSFEEYLNSPKDT----APADLITEICVPLV 194
Mdeg0819_Mdeg_23026984       VKEAGLDYLTLEAGNHAVLTFTGPY------SELEQAYEWFYGEWLPQSGY----QLADRPPFEEYLNDPKTV----PPSELLTKIYLPLA 299
Rsph1750_Rsph_22958167       EIAPPLEELRLSGGRHAVLQYTGPY------SGLARAYRYFFSQWLPEARE----IQAEGPLIEIYRNAAMET----PSDQLVTEICVPLL 156
RSp1247_Rsol_17549468        LPSGSLEPARIEGGLYARHRLVGPY------ALIAPTFEALFGGWLPHSSY----EPDDRPALERYRSPPSPP----QQRDCVTDLLIPIR 308
XCC3056_Xcam_21232486        APPAPLHVRVLGEGAHAVLRHTGSY------AGLEDALDRLLADWLPGSGH----ALRDAPLHYLYLDDPETV----AEADLRADIRVPVV 316
b3023_Ec_16130919            DNRYGVSNGELTGGRYAVARQLGEL------DDISHTVWGIIRHWLPASGE----KMRKAPILFHYTNLAEGV----TEQRLETDVYVPLA 160\b3023
STM3175_Styp_16766475        ENDVGVVNSEIPGGRCAVVRHQGSL------DSLPESVWYLFREWLPASGE----TPRDFPVFFQYLNFVHEV----AEHELLTDIYLPLR 288|
PP2173_Pput_26988897         PNDFGVHERVIPAGRCAVIRHTGSP------DYIGETIYPLYRDWLPSSNE----ELRDHPLFFHYLSVYPET----PLEQWQTDIYVPLK 289|
blr4000_Bjap_27379111        ANGEAIKAGEIPGGRCAVLRVVGHT------DNLEPAALYLYRDWLPASGE----EARDFPIYCQRLSLPPEV----PEHEAVAELFLPLK 293|
mll4902_Mlot_13474099        EKDGQMKAGVIPGGRCAVLRYPGNT------NNLEPAALYLYREWLPASGE----EVRDFPVYCQRRLSRLAEG---PLHEVLVELFLPLK 283|
CC2527_Ccre_16126766         SNPDGVTSGVIPGGPCAVLRQIGSS------DDLRAASRFLYGEWLPRSGR----ELRDAPFFAQRVSFFPDV----PEHEAVTDLFLPLS 339|
RSc0132_Rsol_17544851        PNEPGVVGKTLAGGRYAVLRHAGSD------DTLDQTVAYLYAEWLPASGE----EPRDAPLLFQRVRFYPDV----PEHEAVTDVLLPLR 265/
Consensus/85%                .....h....h....hs.h..................h..h...hh...sb............p.b.................ch.h.h.
```

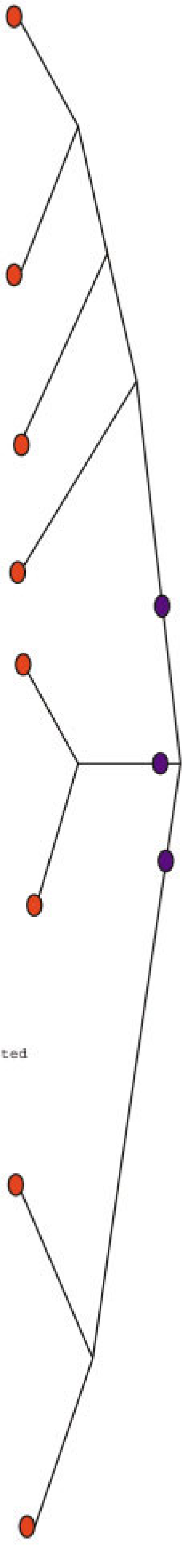

Figure 4. (Continued)

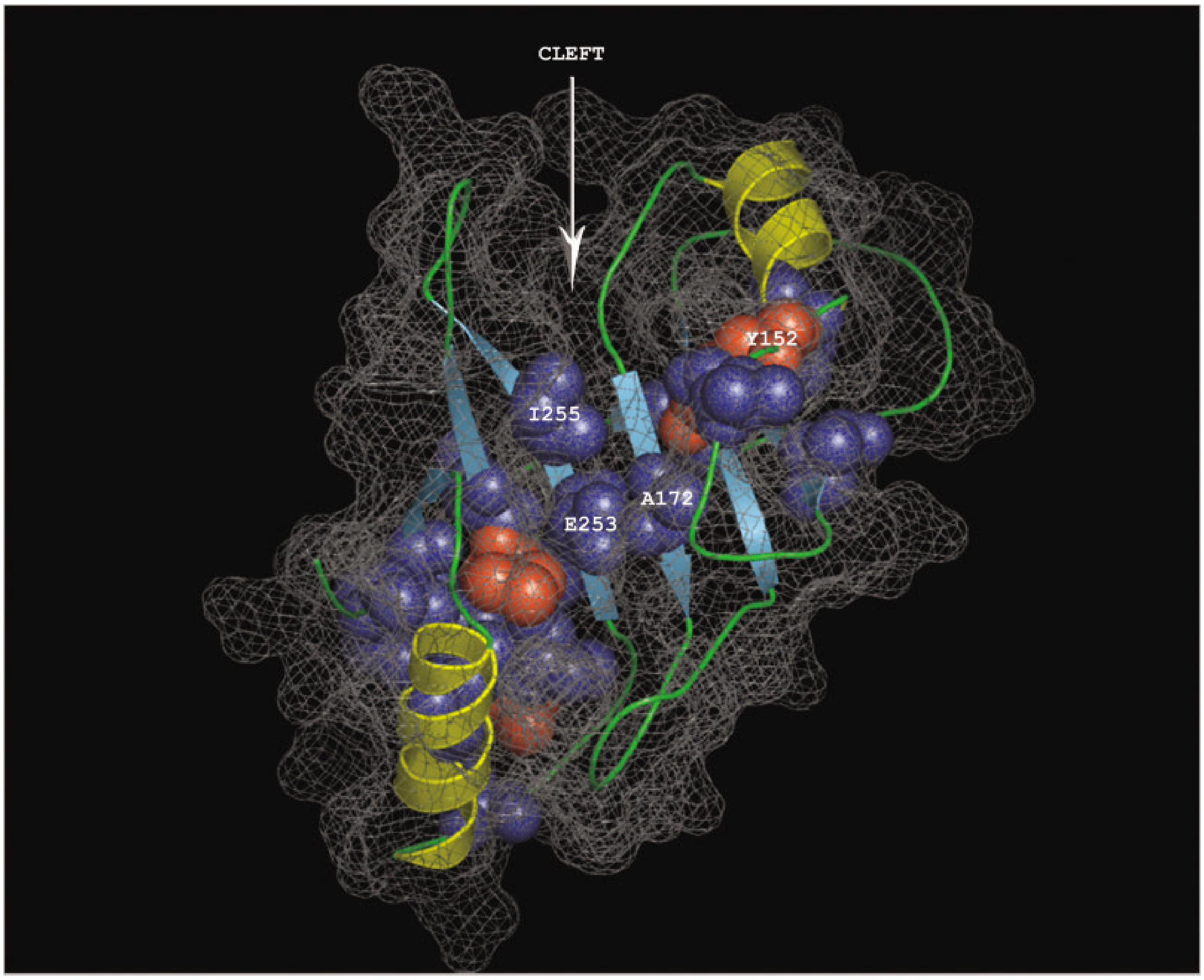

Fig. 5. Conserved positions of the GyrI superfamily mapped on to the crystal structure of BmrR. The conservation pattern specific to individual monophyletic lineages at the two cross sections of the BmrR tree are shown onto BmrR structure (PDB code: 1exj). The conserved residues from the cross section with 3 monophyletic lineages are V126, I134, I135, T137, L148, L155, A172, Y217, A218, I220, L236, L245, E253, I255, E271, and I274, shown in blue, while the residues from the cross section with 8 lineages are Y152, M272, and I276, shown in red. The positions that show specific effects on ligand binding in the mutational analysis of BmrR have been labeled. The figure was generated using PyMol.

for each of these monophyletic lineages. The consensus patterns for the individual lineages at each section were then compared to determine those positions in the alignment at which conserved residues were present in all the monophyletic groups at given section. Thus, at the same position, one monophyletic group may contain a conserved polar residue, while the same position in another monophyletic groups may conserve a hydrophobic residue. This would suggest that the said position might be critically placed in the 3D structure for a particular interaction, though its actual character may vary between groups depending on the interaction specificity of each group. When plotted on the 3D view of the representative structure [in this case, BmrR (PDB code: 1exj)],[47] the majority of conserved positions mapped to the exposed face of the 6-stranded beta sheet bounded on either side by the 2 helices (Fig. 5). The center of this face contained a cleft, and the conserved positions tended to line it, suggesting that the ligands of the GyrI-like superfamily may be lodged in this region (Fig. 5).

Studies on the *Bacillus subtilis* BmrR protein have shown that it is a promiscuous binder of diverse ligands, and the binding of one of its ligands, tetraphenylphosphonium (TPP), proceeds via the unwinding of the helix of the first SHS2 module to access the cleft.[38,47] Furthermore, mutational analysis of BmrR has suggested that different ligands could make different specific contacts within binding cleft.[38,50] In light of these observations, it is likely that the conserved positions detected above are likely to determine only the principal constraints of the ligand-binding pocket of this superfamily. In particular, one of the positions, corresponding to E253 in BmrR, emerged as a conserved position in our analysis. The mutational analysis suggests that it is one residue necessary for binding all the ligands of BmrR.[38,50] In most members of the GyrI superfamily, the corresponding position is occupied by an acidic residue (typically glutamate), suggesting that these proteins are likely to bind positively charged ligands, just like BmrR (Fig. 4). However, there is a single bacterial family, typified by the b3023 protein from *Escherichia coli* K12, in which this position is conserved but occupied by a hydrophobic residue (Fig. 4). Hence, this family may depart from the usual affinity for positively charged ligands and bind hydrophobic ligands instead. Mutations of three other conserved positions, in the set we identified, had noticeable effects on ligand interactions.[38,50] These correspond to the positions of A172, Y152, and I255, and in *B. subtilis* BmrR protein (Fig. 5). The I255 position was typically occupied by either an aliphatic hydrophobic residue or by tyrosine in most lineages of the superfamily. As tyrosine could potentially form a hydrogen bond, as against an aliphatic residue, this position could affect the general ligand preferences of different proteins, depending on the residue occupying this conserved position. The Y152 position is typically occupied by either aromatic or aliphatic hydrophobic residues. Its location at the base of the helix of the first SHS2 module, which partially unwinds upon TPP binding, suggests that it might play a role in regulating access of ligands to the binding cleft. Additionally, 4 positions (corresponding to V126, Y217, L245, and I276 of BmrR) form a patch of conservation on the face opposite to the ligand binding cleft (Figs. 4 and 5). The exact role of these positions is unclear, but their location suggests that they might mediate interactions with other proteins of the transcription machinery.

The iterative sequence profile searches with the PSI-BLAST program helped us to identify several novel members of the GyrI superfamily (Fig. 4). These include several versions that were not fused to HTH domains, but retained the conserved positions of the predicted ligand-binding cleft (e.g., the family typified by MA1133). These GyrI-like modules could act as intracellular sensors of small molecules that may signal via other signaling systems, or may even possess as yet uncharacterized catalytic functions with respect to their ligands. Interestingly, these searches also identified eukaryotic members of the GyrI superfamily, in the form of an orthologous group of proteins typified by ZSIG11 from humans. These proteins possess a strongly predicted signal peptide, suggesting that they are secreted proteins. These proteins also contain the conserved glutamate (corresponding to E253; Fig. 4), suggesting that they may either act as novel extracellular carriers or sensors of positively charged small molecules in animals. Outside of the animals, this family is encountered only in the animal pathogen *Leptospira interrogans* (Fig. 4). It is possible that this pathogen-encoded version may help it to interact with the same ligands as the endogenous host encoded forms as a part of the infection process.

The MTH1598/Tm1083 superfamily, which contains a different kind of arrangement of the two SHS2 modules, remains functionally uncharacterized. This protein is highly conserved in both archaea and eukaryotes, and present sporadically in bacteria. This phyletic pattern is similar to several proteins involved in core aspects of nucleic acid metabolism, such as DNA replication, repair and recombination, and various aspects of transcription, translation, and RNA metabolism. These proteins contain 4 absolutely conserved charged residues that are distributed along the outer rim of the surface formed by the 2 β-sheets of the SHS2 unit (Figs. 1 and 2). This would suggest that the MTH1358/Tm1083-like proteins are likely to be enzymes, and based on their phyletic pattern, they are likely to be involved in some core aspect of nucleic acid metabolism.

The above observations suggest that the SHS2 domains have been utilized in functionally distinct contexts and this reflects in the sequence conservation patterns in the individual families (Fig. 3). The SHS2 domains of RBP7p and FtsA are most closely related in structural terms, and they accordingly appear to share certain conserved residues that are likely to mediate similar interactions (Fig. 3). In contrast, the dimeric forms have been subject to very different selective forces. The interaction with small molecules in the GyrI superfamily has resulted in the peculiar conservation pattern focused on the potential interaction surface formed by the strands from the two SHS2 units. In the MTH1598 family, the conservation is similarly associated with the exposed faces of the strands (Figs. 2 and 3). However, it is restricted to an outer rim of one end of the sheet and appears to mainly comprise charged residues, which may be involved in a catalytic process (Fig. 2). Thus, the different SHS2 domains have considerably diverged from each other in sequence, and could be detected only by means of structure similarity searches.

Reconstruction of the Overall Evolutionary History of the SHS2 Fold

The SHS2 domain of FtsA can be clearly traced back to last common ancestor of all bacteria. Likewise, the SHS2 domain of Rpb7 can be traced back to the last common ancestor of the archaeo-eukaryotic branch of life. Hence, it is likely that the two were most probably derived from a precursor that was present in the last universal common ancestor of all life forms (LUCA). Though these two versions of the domain share several functional and structural features, the actual biological contexts in which they function are unrelated. Hence, while their precursor in LUCA is likely to have participated in protein–protein interactions, its exact biological context cannot be determined. Of the dimeric forms, the GyrI superfamily is widespread in bacteria and likely to have been present from early in bacterial evolution (Fig. 4). The sporadic presence of this domain in archaea and eukaryotes suggests that it might have entered these lineages due to lateral gene transfer from bacteria. The MTH1358/Tm1083 superfamily is well-conserved in the archaeo-eukaryotic lineage, but is only sporadically found in bacteria, suggesting that this version emerged specifically during the early evolution of the former lineage. In principle, the two dimeric versions could have been derived at the base of the archaeo-eukaryotic and bacterial lineages from the monomeric version that is traceable to the LUCA. However, given the extreme sequence divergence, it is quite possible that other versions of the SHS2 domain have eluded detection. If such forms are recovered, they might provide a clearer view of the early evolution of this fold.

Interestingly, the SHS2 fold appears to be more limited in its distribution than more complex folds such as the RRM-like and the IF3-C folds, which contain topologically equivalent secondary structural elements (Fig. 2). One possibility is that the SHS2 fold has been derived relatively recently through the drastic degeneration of one of these other folds. The other possibility is that the SHS2 fold is an ancient fold that emerged along with the other more widespread folds, but the small 3-stranded sheet allowed it lesser functional versatility than the RRM-like or IF3-C folds. As a result, it could be utilized only in a few contexts and remained restricted in its distribution.

GENERAL CONCLUSIONS

Using structure similarity searches, we have delineated a novel domain with a simple fold in the bacterial cell-division ATPase FtsA, the RNA polymerase subunit Rpb7p, the GyrI superfamily, and the uncharacterized MTH1598/Tm1083-like proteins. The fold contains a core of 3 strands forming a curved sheet and a single helix. The SHS2 domain may either exist in a single copy or, alternatively, in two copies in the same polypeptide, which may fold into more complex dimeric units. The single-copy forms of the domain in FtsA and Rbp7 appear to mediate different protein–protein interactions, by means of strand 1, and the loop between strand 2 and strand 3 of the domain. The dimeric versions typified by the GyrI superfamily appear to have been adapted for small-molecule binding. Sequence profiles searches helped us to identify several new versions of the GyrI superfamily, including a family of secreted forms that is found only in animals and the bacterial pathogen *Leptospira*. In the MTH1598/Tm1083 family, a SHS2 domain is inserted into the loop between strand 1 and helix 1 of another SHS2 domain. The sequence conservation pattern and its phyletic profile suggest that it might be an enzyme involved in some conserved aspect of nucleic acid metabolism.

NOTE ADDED IN PROOF

After the acceptance of this article for publication, a structure of the flavin-binding protein, dodecin, became available in the PDB database (PDB: 1MOG; Bieger, B., Essen, L.-O., Oesterhelt, D.: Crystal Structure of Halophilic Dodecin a Novel, Dodecameric Flavin Binding Protein from Halobacterium Salinarum Structure 11 pp. 375, 2003). The dodecin family is found in several bacteria and few archaea and represents a new stand-alone version of the SHS2 domain. It most closely resembles the SHS2 domains of FtsA and Rpb7p, and appears to represent a single domain small-molecule binding form, in contrast to the duplicated versions in the GyrI superfamily.

REFERENCES

1. Aravind L, Koonin EV. DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res 1999;27: 4658–4670.
2. Grishin NV. Two tricks in one bundle: helix–turn–helix gains enzymatic activity. Nucleic Acids Res 2000;28:2229–2233.
3. Rosinski JA, Atchley WR. Molecular evolution of helix–turn–helix proteins. J Mol Evol 1999;49:301–309.
4. Aravind L, Koonin EV. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. Genome Res 2001;11:1365–1374.
5. Aravind L, Mazumder R, Vasudevan S, Koonin EV. Trends in protein evolution inferred from sequence and structure analysis. Curr Opin Struct Biol 2002;12:392–399.
6. Doherty AJ, Serpell LC, Ponting CP. The helix–hairpin–helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. Nucleic Acids Res 1996;24:2488–2497.
7. Shao X, Grishin NV. Common fold in helix–hairpin–helix proteins. Nucleic Acids Res 2000;28:2643–2650.
8. Aravind L, Iyer LM. Provenance of SET-domain histone methyltransferases through duplication of a simple structural unit. Cell Cycle 2003;2:369–376.
9. Iyer LM, Koonin EV, Aravind L. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. BMC Struct Biol 2003;3:1.
10. Castillo RM, Mizuguchi K, Dhanaraj V, Albert A, Blundell TL, Murzin AG. A six-stranded double-psi beta barrel is shared by several protein superfamilies. Struct Fold Des 1999;7:227–236.
11. Coles M, Diercks T, Liermann J, Groger A, Rockel B, Baumeister W, Koretke KK, Lupas A, Peters J, Kessler H. The solution structure of VAT-N reveals a "missing link" in the evolution of complex enzymes from a simple βαββ element. Curr Biol 1999;9: 1158–1168.
12. Bateman A. The structure of a domain common to archaebacteria and the homocystinuria disease protein. Trends Biochem Sci 1997;22:12–13.
13. van den Ent F, Lowe J. Crystal structure of the cell division protein FtsA from *Thermotoga maritima*. EMBO J 2000;19:5300–5307.
14. Carettoni D, Gomez-Puertas P, Yim L, Mingorance J, Massidda O, Vicente M, Valencia A, Domenici E, Anderluzzi D. Phage-display and correlated mutations identify an essential region of subdomain 1C involved in homodimerization of *Escherichia coli* FtsA. Proteins 2003;50:192–206.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
16. Aravind L, Koonin EV. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J Mol Biol 1999;287:1023–1040.
17. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
18. Pei J, Sadreyev R, Grishin NV. PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 2003;19:427–428.
19. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
20. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
21. Holm L, Sander C. The FSSP database: fold classification based on structure–structure alignment of proteins. Nucleic Acids Res 1996;24:206–209.
22. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
23. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
24. DeLano WL. The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific; 2002.
25. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 1996;266:418–427.
26. Hasegawa M, Kishino H, Saitou N. On the maximum likelihood method in molecular phylogenetics. J Mol Evol 1991;32:443–445.
27. Felsenstein J. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 1989;5:164–166.
28. Kumar S, Tamura K, Jakobsen IB, Nei M. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 2001;17: 1244–1245.

29. Bork P, Sander C, Valencia A. An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. Proc Natl Acad Sci USA 1992;89:7290–7294.
30. Sedelnikova SE, Agalarov S, Eliseikina IA, Fomenkova NP, Nikonov SV, Garber MB, Svensson LA, Liljas A. Crystals of protein S6 from the 30 S ribosomal subunit of *Thermus thermophilus*. J Mol Biol 1991;220:549–550.
31. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 2002;30:1427–1464.
32. Wardleworth BN, Russell RJ, Bell SD, Taylor GL, White MF. Structure of Alba: an archaeal chromatin protein modulated by acetylation. EMBO J 2002;21:4654–4662.
33. Aravind L, Iyer LM, Anantharaman V. The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism. Genome Biol 2003;4:R64.
34. Todone F, Brick P, Werner F, Weinzierl RO, Onesti S. Structure of an archaeal homolog of the eukaryotic RNA polymerase II RPB4/RPB7 complex. Mol Cell 2001;8:1137–1143.
35. Nakanishi A, Oshida T, Matsushita T, Imajoh-Ohmi S, Ohnuki T. Identification of DNA gyrase inhibitor (GyrI) in *Escherichia coli*. J Biol Chem 1998;273:1933–1938.
36. Romanowski MJ, Gibney SA, Burley SK. Crystal structure of the *Escherichia coli* SbmC protein that protects cells from the DNA replication inhibitor microcin B17. Proteins 2002;47:403–407.
37. Kwon HJ, Bennik MH, Demple B, Ellenberger T. Crystal structure of the *Escherichia coli* Rob transcription factor in complex with DNA. Nat Struct Biol 2000;7:424–430.
38. Zheleznova EE, Markham PN, Neyfakh AA, Brennan RG. Structural basis of multidrug recognition by BmrR, a transcription activator of a multidrug transporter. Cell 1999;96:353–362.
39. Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez, Denisov A, Lee CH, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM, Arrowsmith CH. An NMR approach to structural proteomics. Proc Natl Acad Sci USA 2002;99:1825–1830.
40. Volz K. A test case for structure-based functional assignment: the 1.2 A crystal structure of the yjgF gene product from *Escherichia coli*. Protein Sci 1999;8:2428–2437.
41. Bushnell DA, Kornberg RD. Complete, 12-subunit RNA polymerase II at 4.1-Å resolution: implications for the initiation of transcription. Proc Natl Acad Sci USA 2003;100:6969–6973.
42. Vijayalakshmi J, Mukhergee MK, Graumann J, Jakob U, Saper MA. The 2.2 Å crystal structure of Hsp33: a heat shock protein with redox-regulated chaperone activity. Structure (Camb) 2001;9:367–375.
43. Hopf M, Gohring W, Ries A, Timpl R, Hohenester E. Crystal structure and mutational analysis of a perlecan-binding fragment of nidogen-1. Nat Struct Biol 2001;8:634–640.
44. Ormo M, Cubitt AB, Kallio K, Gross LA, Tsien RY, Remington SJ. Crystal structure of the *Aequorea victoria* green fluorescent protein. Science 1996;273:1392–1395.
45. Wall MA, Socolich M, Ranganathan R. The structural basis for red fluorescence in the tetrameric GFP homolog DsRed. Nat Struct Biol 2000;7:1133–1138.
46. Nakanishi A, Imajoh-Ohmi S, Hanaoka F. Characterization of the interaction between DNA gyrase inhibitor and DNA gyrase of *Escherichia coli*. J Biol Chem 2002;277:8949–8954.
47. Heldwein EE, Brennan RG. Crystal structure of the transcription activator BmrR bound to DNA and a drug. Nature 2001;409:378–382.
48. Aravind L, Anantharaman V. HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. FEMS Microbiol Lett 2003;222:17–23.
49. Anantharaman V, Koonin EV, Aravind L. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. J Mol Biol 2001;307:1271–1292.
50. Vazquez-Laslop N, Markham PN, Neyfakh AA. Mechanism of ligand recognition by BmrR, the multidrug-responding transcriptional regulator: mutational analysis of the ligand-binding site. Biochemistry 1999;38:16925–16931.
51. Xie T, Chen J, Ding DF. An evolutionary trace method for functional prediction of genomes. Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai) 1999;31:433–439.
52. Sowa ME, He W, Slep KC, Kercher MA, Lichtarge O, Wensel TG. Prediction and confirmation of a site critical for effector regulation of RGS domain activity. Nat Struct Biol 2001;8:234–237.
53. Lichtarge O, Sowa ME, Philippi A. Evolutionary traces of functional surfaces along G protein signaling pathway. Methods Enzymol 2002;344:536–556.