Profile–profile comparisons by COMPASS predict intricate homologies between protein families

RUSLAN I. SADREYEV,[1] DAVID BAKER,[2] AND NICK V. GRISHIN[1]

[1]Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA
[2]Howard Hughes Medical Institute and Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA

(RECEIVED May 13, 2003; FINAL REVISION July 8, 2003; ACCEPTED July 9, 2003)

Abstract

Recently we proposed a novel method of alignment–alignment comparison, COMPASS (the tool for COmparison of Multiple Protein Alignments with Assessment of Statistical Significance). Here we present several examples of the relations between PFAM protein families that were detected by COMPASS and that lead to the predictions of presently unresolved protein structures. We discuss relatively straightforward COMPASS predictions that are new and interesting to us, and that would require a substantial time and effort to justify even for a skilled PSI-BLAST user. All of the presented COMPASS hits are independently confirmed by other methods, including the ab initio structure-prediction method ROSETTA. The tertiary structure predictions made by ROSETTA proved to be useful for improving sequence-derived alignments, because they are based on a reasonable folding of the polypeptide chain rather than on the information from sequence databases. The ability of COMPASS to predict new relations within the PFAM database indicates the high sensitivity of COMPASS searches and substantiates its potential value for the discovery of previously unknown similarities between protein families.

Keywords: Protein structure prediction; COMPASS; ROSETTA; domains of unknown function; helix–turn–helix; rRNA methylase; PPR; viral coat proteins

Sequence comparison has proven to be a valuable tool in the study of protein structure, function, and evolution. In a series of successful efforts to improve the detection of remote sequence similarities, the most powerful methods involve the comparison of multiple protein alignments to single sequences or to other multiple alignments (Gotoh 1993, 1994; Pietrokovski 1996; Altschul et al. 1997; Eddy 1998; Karplus et al. 1999; Schaffer et al. 1999, 2001; Jaroszewski et al. 2000; Rychlewski et al. 2000; Kunin et al. 2001; Yona and Levitt 2002; Sadreyev and Grishin 2003). The underlying assumption of these approaches is that the information extracted from aligned related sequences may represent general features of the family and allow prediction of similarity to a remote sequence (or family), even if its similarity to each of the individual aligned sequences is insignificant. Well-known and widely used methods involving sequence-alignment comparison include PSI-BLAST (Altschul et al. 1997; Schaffer et al. 2001), IMPALA (Schaffer et al. 1999), SAM-T99 (Karplus et al. 1999), and HMMER (Eddy 1998). As a further step in this direction, several methods have been developed for the comparison of multiple alignments to multiple alignments. They include iterative protocols for multiple alignment construction based on the sum-of-pairs scoring system (Gotoh 1993, 1994); the LAMA protocol for the comparisons of block alignments with no gaps permitted (Pietrokovski 1996), which is further used in the CYRCA method (Kunin et al. 2001) for the search of multiple consistently aligned blocks within two compared alignments; and FFAS (Jaroszewski et al. 2000; Rychlewski et al. 2000) and prof_sim (Yona and Levitt 2002) methods, which involve the construction of local gapped alignments of the two families.

Reprint requests to: Nick V. Grishin, Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA; e-mail: grishin@chop.swmed.edu; fax: (214) 648–9099.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03197403.

Recently we proposed a novel method of alignment–alignment comparison, COMPASS (the tool for COmparison of Multiple Protein Alignments with Assessment of Statistical Significance). It derives numerical profiles from alignments, constructs optimal local profile–profile alignments, and analytically estimates *E*-values for the detected similarities. The scoring system and *E*-value calculation are a generalization of the PSI-BLAST approach to profile-sequence comparison, which is adapted for the profile–profile case. Tested along with previously reported methods, COMPASS shows increased abilities for sensitive and selective detection of remote sequence similarities, as well as improved quality of local alignments (Sadreyev and Grishin 2003).

In this work, we present several examples of relations between protein families that were detected by COMPASS and that lead to predictions of presently unresolved protein structures. In the search for such relationships, we used the PFAM database (Bateman et al. 2002) families with unknown structure as queries to run COMPASS against the data set of PFAM families containing at least one protein with a solved structure. The statistically significant hits were examined, and meaningful predictions were chosen for further analysis. In particular, among the profile–profile alignments produced by COMPASS, we chose the ones that (1) covered long enough regions sufficient for the fold prediction, (2) were not discussed previously in the literature, and (3) were impossible to find by automatic iterative PSI-BLAST searches with default parameters. If the family of interest contained bacterial sequences, we demanded that it correspond to an uncharacterized Cluster of Orthologous Groups of proteins (COG) in the COG database (Tatusov et al. 1997, 2001), a powerful resource for bacterial sequence classification. In brief, we were looking for the straightforward COMPASS predictions that are new and interesting to us, and that would require a substantial time and effort even for a skilled PSI-BLAST user.

On the other hand, we sought an independent confirmation of the validity of our predictions using extensive PSI-BLAST searches with manual inspection of the hits, secondary structure predictions, predictions by fold-recognition packages, and other information. As an approach that is most independent of the existing sequence databases, we used ab initio prediction of protein structure by the ROSETTA method (Bonneau et al. 2001; Simons et al. 2001). Given a protein sequence, ROSETTA generates putative protein structures based on two assumptions: (1) Local interactions play the major role in protein folding at the scale of short segments of polypeptide chain; and (2) nonlocal interactions between distant parts of the chain are capable of stabilizing native-like arrangements of local structural segments. Using the libraries of the short structural fragments adopted by the related sequences with known structure, ROSETTA applies a Monte Carlo procedure to optimize an energy function that favors native-like compound structures (Bonneau et al. 2001; Simons et al. 2001).

This combination of independent methods provided additional information to confirm COMPASS predictions and to make further refinements of COMPASS alignments.

Materials and methods

The search with COMPASS was performed using the PFAM database of multiple sequence alignments (Bateman et al. 2002; version 6.6). Specifically, we used 1717 PFAM alignments that do not contain sequences with known protein structures for the COMPASS search with default parameters (filtering out alignment columns with more than 50% effective counts of gaps, BLOSUM62 residue substitution matrix, gap penalties 11 + k) against the set of 1354 PFAM families with solved structure, which had members included in the PDB database. The "full" PFAM alignments were used (as opposed to the "seed" alignments).

In parallel, we performed PSI-BLAST searches in the same database. For this task, we extracted all of the individual sequences from the PFAM alignments of the families with known structure, resulting in a database of 311,753 sequences. In this sequence database, we ran PSI-BLAST searches using each of the families with unresolved structures as a query (one round of the PSI-BLAST 2.2.1 search with a PSI-BLAST numerical profile derived from the alignment; the template sequence was set to the first sequence of the query alignment; the maximal number of displayed hits and the maximal *E*-value were both set to 10,000). After producing a list of sequence hits for the query alignment, we considered the families with known structures, finding a sequence with the best *E*-value in each of them. This best *E*-value was assigned to the PSI-BLAST comparison of the query alignment and the given family. This setup for the COMPASS and PSI-BLAST searches has been previously used for the comparisons within the group of PFAM alignments with known structures (Sadreyev and Grishin 2003).

To choose COMPASS hits of potential interest, we discarded all similarities between protein families detected by COMPASS that had PSI-BLAST *E*-values <0.1. COMPASS hits that corresponded to higher PSI-BLAST *E*-values and thus passed the initial filtering were subjected to a more thorough analysis. From 518 such hits that did not include apparent transmembrane proteins, we chose ~100 COMPASS alignments of interest, which covered substantial protein regions and were not restricted to trivial similarities (e.g., P-loops). After this step, we further excluded the results that could be easily reproduced by running multiple iterations of PSI-BLAST against the NCBI nr sequence database, with various family members as queries, and the families that contained prokaryotic sequences assigned to a characterized COG (Tatusov et al. 1997, 2001). Sequences from the remaining PFAM alignments were submitted to the PHD server (Rost 1996; Przybylski and Rost 2002) for secondary structure predictions (with the option of iterated PSI-BLAST searches against SWISS-PROT, TrEMBL, and PDB databases). Using the available option of the PHD server, we also submitted PFAM alignments for the secondary structure prediction. Both PSI-BLAST and PFAM profiles produced similar PHD predictions. The sequences from the PFAM alignments were also submitted to the fold-recognition servers: 3D-PSSM (Kelley et al. 2000; version 2.6.0, using fold library 1.53.7660 and sequence database 2002.9.4) and bioinbgu (Fischer 2000; version as of September 2002). Iterated PSI-BLAST searches were performed in the NCBI nr database using the sequences from PFAM alignments as queries. The PSI-BLAST hits with *E*-values higher than the default

cutoff were manually inspected. In cases in which the sequence length was <150 residues, it was submitted to the ROSETTA package for the ab initio prediction of protein structure (standalone version as of March 2002). To generate the input profile for ROSETTA, a PSI-BLAST search with the *E*-value cutoff of 0.001 was used; minimum sequence identity for inclusion in the profile was set to 25%; the PHD secondary structure prediction was used along with the profile to generate two sets of fragments of sizes 9 and 3; and the minimum allowed confidence for fragment prediction was set to 0.25. For each sequence, 2000 decoy structures (output in the PDB format) were generated and clustered by structure similarity. The decoys serving as the centers for the largest 10 clusters were examined.

Based on this additional information, the COMPASS hits were validated, and the initial profile–profile alignments were manually refined. For any method involving automated sequence alignment, such a refinement increases the quality of the initial prediction by using data other than the sequence information, and by detailed manual assessment of these data. The refinement was based on the analysis of secondary structure elements, residue properties (hydrophobicity, charge, size, etc.), and conservation at the alignment positions, potential functionally and structurally important sites in the protein families. As shown previously (Sadreyev and Grishin 2003), the accuracy of the COMPASS alignments is higher than the accuracy of alignments produced by other tested automated methods for sequence–profile or profile–profile comparison. As shown in Figures 1–4, the manually refined alignments were in general similar to those initially produced by COMPASS. The main differences included slight local shifts of aligned positions and inclusion of longer profile regions after the manual analysis. Based on the resulting alignments, the tertiary structure of the families of interest was proposed.

We also compared the results of COMPASS to those of two other available methods of profile–profile comparison, LAMA (Pietrokovski 1996) and prof_sim (Yona and Levitt 2002). To perform the LAMA search, we first submitted the PFAM alignment to the Blocks Multiple Alignment Processor server (http://blocks.fhcrc.org/blocks/process_blocks.html) and generated the alignment blocks, which were used as queries to search with LAMA (version as of 04/28/2000) in the Blocks+ database with the default parameters (minimal length of reported alignments set to 4, *Z*-score cutoff set to 5.6). The version of prof_sim, which was generously provided by G. Yona (Cornell University), performs pairwise comparisons between submitted profiles. We submitted the pairs of interest and compared the resulting alignments and *E*-values to those by produced COMPASS. For the generated profile–profile alignments, prof_sim estimates *P*-values, which should be multiplied by the number of profiles in the database (1354 in our case) to obtain *E*-values (G. Yona, pers. comm.).

Results

DUF185 is homologous to methyltransferases

DUF185 is a family of protein domains found in bacteria, plants, and animals. In PFAM 7.6, it is described as "uncharacterized domain in proteins of unknown function." In the COG database (Tatusov et al. 1997, 2001), this domain corresponds to uncharacterized ACR (COG1565).

Using this PFAM alignment as the query, COMPASS search in the set of PFAM families with known structure detected the similarity of DUF185 to the RrnaAD family (Table 1). RrnaAD contains ribosomal RNA adenine dimethylases found in archaea, bacteria, plants, fungi, and animals. According to SCOP (Murzin et al. 1995) classification, these proteins belong to the family of RNA methylases, with an *S*-adenosyl-L-methionine-dependent methyltransferase fold.

Figure 1A shows the alignment of two sequences that represents the profile–profile alignment constructed by COMPASS, and the multiple alignment of representative sequences from the two families that was produced by the manual refinement of the COMPASS result. This multiple alignment includes a portion of DUF185 and the region of RrnaAD that contains the specific signature of this family (PROSITE entry PS01131): [LIVM]–[LIVMFY]–[DE]–x–G–[STAPV]–G–x–[GA]–x–[LIVMF]–[ST]–x(2)–[LIVM]–x(6)–[LIVMY]–x–[STAGV]–[LIVMFYHC]–E–x–D (Fig. 2A). In the rRNA methylase structure, this region corresponds to the two adjacent β/α units that are involved in *S*-adenosyl methionine binding. The produced alignment reveals a remarkable conservation of this signature in the DUF185 family, including the invariant ligand-binding glutamate residues. In addition to the conserved signature pattern, the aligned regions have similar profiles of hydrophobicity and location of small residues. The secondary structure prediction for this part of DUF185 is consistent with the location of the known secondary structure elements in rRNA methylase (PDB ID 1yub; Fig. 2A,B). The predicted secondary structure of the full-length DUF185 proteins (data not shown) comprises a pattern of consequent β–α

Table 1. *The results produced by various methods on the PFAM families of interest*

PFAM 6.6 name (PFAM Acc)	COMPASS hit (PFAM Acc)	CMPSS *E*-value	3D-PSSM top hit (*E*-value)	Bioinbgu top hit (consens. score)	LAMA top hit blocks Acc (*Z*-score)	Prof_sim *P*-value/*E*-value
DUF185 (PF02636)	RrnAD (PF00398)	5.3×10^{-6}	1kp9A (1.88×10^{-1})	1kpiA (27.5)	IPB001737A (9.4)	2.46×10^{-3}/3.33
DUF128 (PF01995)	HTH_5 (PF01022)	1.48×10^{-6}	1jmrA (1.04×10^{-1})	1dprA (16.7)	IPB001845B (7.2)	0.049/66.3
PPR (PF01535)	Clathrin_repeat (PF00637)	6.13×10^{-3}	1paa (94.5)	—	IPB000132A (6.3)	9.73×10^{-3}/13.2
Astro_capsid (PF03115)	Viral_coat (PF00729)	7.35×10^{-9}	1bmv (3.00×10^{-2})	1bmv (30.7)	IPB001218C (8.8)	0.321/435

PFAM name and accession number are indicated for the families of unknown structure and the related families with solved structure that were found by COMPASS, as well as the COMPASS *E*-value for the found similarity. The top hits for the families of unknown structure produced by 3D-PSSM (PDB ID and *E*-value), bioinbgu (PDB ID and consensus score), and LAMA (Blocks accession number and Z-score) are shown, along with the *P*-values and *E*-values produced by prof_sim on the PFAM alignment pairs.

units similar to the secondary structure arrangement of methylases. These similarities indicate that the DUF185 family possesses a Rossmann-like fold and may function as methylases.

For the majority of individual DUF185 proteins, PSI-BLAST 2.2.4 searches with default parameters in the NCBI nr database did not produce methylase hits up to convergence, although methylases could be found among the hits with *E*-values higher than the inclusion cutoff of 0.005. For several DUF185 proteins, PSI-BLAST searches found methylases with *E*-values lower than the default cutoff, but the hits required as many as 8–12 iterations of PSI-BLAST. When the DUF185 alignment was submitted to the Blocks Multiple Alignment Processor server and the resulting blocks were used for the LAMA search in the Blocks+ database, a high *Z*-score was assigned to a block of the rRNA adenine dimethylase family (Table 1). One of the other three weaker hits (block IPB001566A, LAMA *Z*-score 6.3) also represented a methyltransferase family, RNA methyltransferase trmA. The prof_sim method produced the alignment of the DUF185 and RrnaAD profiles that reflected the similarity between the main conserved sites found by COMPASS, although with a marginal *P*-value (Table 1).

This prediction based on a result of COMPASS search and confirmed by PSI-BLAST searches is further supported by the predictions produced by the fold-recognition servers. The bioingbu server assigned a high consensus score (Table 1) to the top consensus prediction, mycolic acid cyclopropane synthase CmaA1 (PDB ID 1kpiA), which is classified within the same SCOP superfamily (*S*-adenosyl-L-methionine-dependent methyltransferases) as rRNA methylases. rRNA methylase Ermc′ (PDB ID 1qaoA) was assigned a consensus score of 3.9. Similarly, the top two 3D-PSSM predictions were mycolic acid cyclopropane synthase CmaA1 (PDB ID 1kp9A; Table 1) and rRNA methyltransferase Ermc′ (PDB ID 1qamA), with an *E*-value of 3.32×10^{-1}.

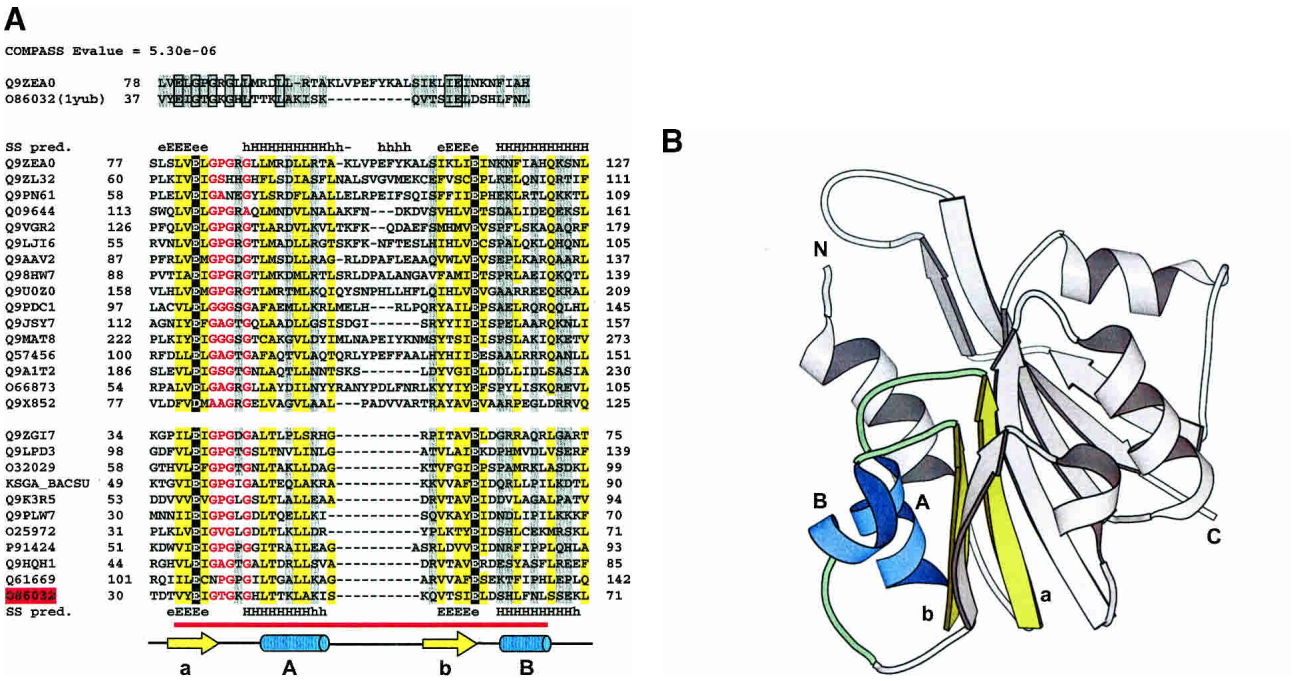

Figure 1. Sequence similarity between the DUF185 and RrnaAD families of the PFAM database indicates that DUF185 is a putative methylase domain. (*A*) Profile–profile alignment, as constructed by COMPASS (shown as the alignment of two representative sequences), and manually refined multiple alignment including representatives from DUF185 (*top*) and RrnaAD (*bottom*) are illustrated. The TrEMBL identifier and first residue number are shown for each sequence. In the initial COMPASS alignment, matches with positive scores are highlighted in gray, and invariant residues are boxed. In the refined multiple alignment, the uncharged residues (all amino acids except D, E, K, and R) in mostly hydrophobic sites are highlighted in yellow, the nonhydrophobic residues (all amino acids except W, F, Y, M, L, I, and V) at mostly hydrophilic sites are highlighted in light gray, and the small residues (G, P, A, S, C, T, V) at positions occupied by mostly small residues are shown in red letters. The invariant glutamate residues are boxed in black. The identifier of the sequence with known spatial structure in RrnaAD (PDB Id 1yub) is highlighted in red. The PHD secondary structure predictions (SS pred.) are shown for this sequence and for the *top* sequence of the DUF185 alignment. The actual secondary structure of the 1yub fragment is shown *below* the alignment, with secondary structure elements labeled and colored according to the scheme shown in *B*. α-Helices and β-strands are displayed as arrows and cylinders, respectively. The region covered by the initial COMPASS alignment is shown with a red line *below* the multiple alignment. (*B*) A ribbon diagram of the fragment of rRNA methylase (PDB ID 1yub) that was drawn by MOLSCRIPT (Kraulis 1991). N and C termini are labeled. The highlighted region corresponds to the alignment in *A*; α-helices are colored in blue; β-strands are colored in yellow.

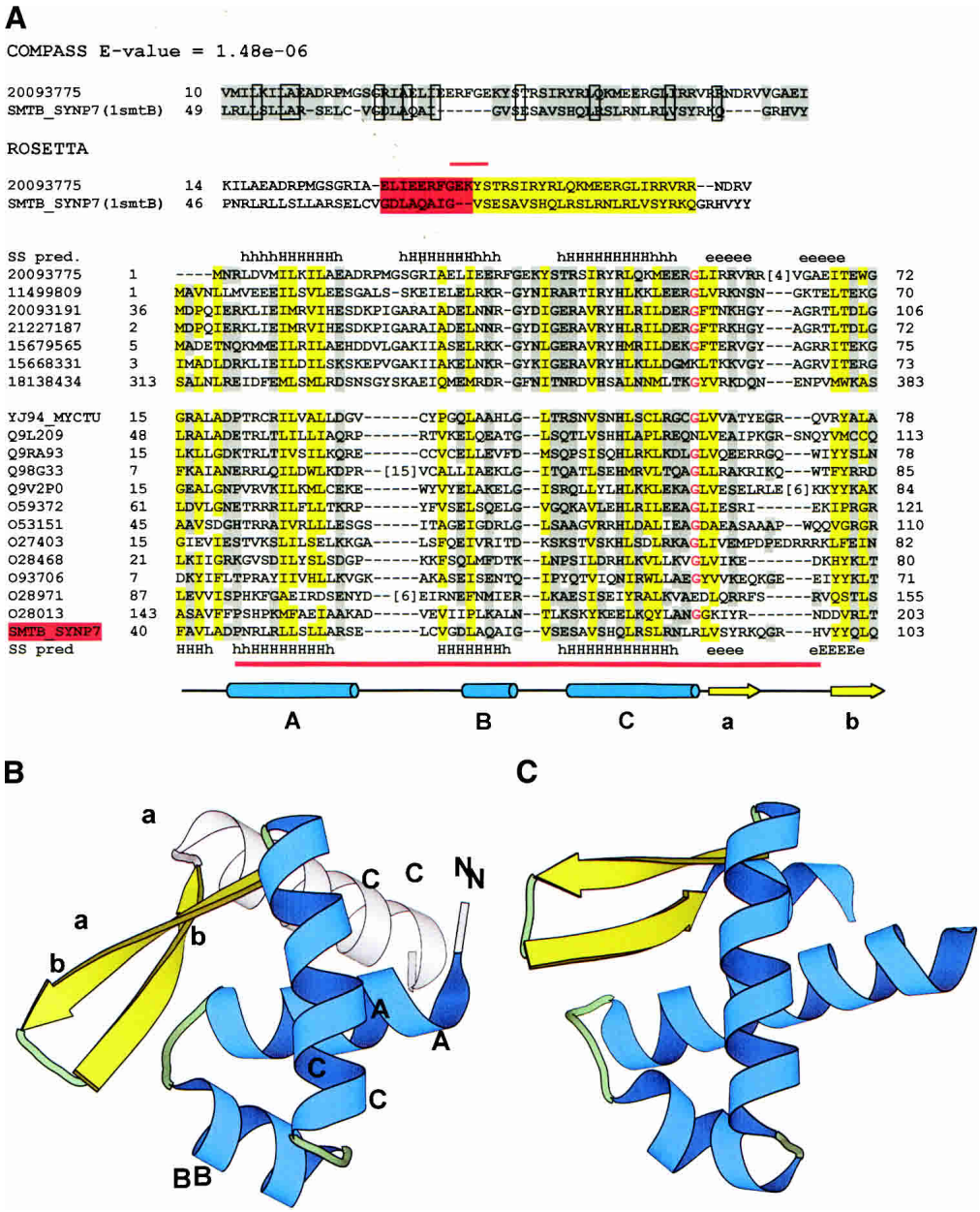

Figure 2. Sequence similarity between the DUF128 and HTH_5 families of PFAM implies that DUF128 contains a "winged helix" domain. (*A*) Profile–profile alignment, as constructed by COMPASS (shown as the alignment of two representative sequences), structure-based alignment of ROSETTA prediction, and manually refined multiple alignment including representatives from DUF128 (*top*) and HTH_5 (*bottom*) are illustrated. GenBank identifiers (GI) are shown for the sequences of the extended DUF128 family (*top*); TrEMBL identifiers are shown for the sequences of the HTH_5 family (*bottom*). The first residue number is shown for each sequence. In the initial COMPASS alignment, matches with positive scores are highlighted in gray, and invariant residues are boxed. In the structure-based alignment of ROSETTA prediction and a winged helix domain (PDB ID 1smtB), the regions in the vicinity of the functional site that were used in the manual refinement of COMPASS alignment are highlighted. (Yellow) The region consistent with COMPASS alignment; (red) the region that includes the more reasonable ROSETTA-based alignment than that produced by COMPASS. The functionally important turn between the helices B and C, with a two-residue insertion, is marked by the red line *above* the alignment. In the refined multiple alignment, the uncharged residues (all amino acids except D, E, K, and R) in mostly hydrophobic sites are highlighted in yellow, the nonhydrophobic residues (all amino acids except W, F, Y, M, L, I, and V) at mostly hydrophilic sites are highlighted in light gray, and the small residues (G, P, A, S, C, T, V) at positions occupied by mostly small residues are shown in red letters. Long insertions are not displayed: The numbers of omitted residues are specified in brackets. The identifier of the sequence with known spatial structure in HTH_5 (PDB Id 1smtB) is highlighted in red. The PHD secondary structure predictions (SS pred.) are shown for this sequence and for the *top* sequence of the DUF128 alignment. The actual secondary structure of the 1smtB fragment is shown *below* the alignment, with secondary structure elements labeled and colored according to the scheme shown in *B*. α-Helices and β-strands are displayed as arrows and cylinders, respectively. The region covered by the initial COMPASS alignment is shown with a red line *below* the multiple alignment. (*B*) A ribbon diagram of the fragment of a "winged helix" domain (PDB ID 1smtB) that was drawn by MOLSCRIPT (Kraulis 1991). N and C termini are labeled. The highlighted region corresponds to the alignment in *A*; α-helices are colored in blue, β-strands are colored in yellow. (*C*) The ribbon diagram of ab initio prediction of the tertiary structure for the *top* sequence in the DUF128 alignment (GenBank GI 20093775) made by ROSETTA. N and C termini are labeled. Secondary structure elements are labeled and colored according to the scheme shown in *B*; α-helices are colored in blue; β-strands are colored in yellow.

DUF128 contains a "winged helix" domain

DUF128 is a family of proteins found in *Archaea*; the function of these proteins is unknown, according to PFAM 7.6. In the COG database, they correspond to COG1693, which is labeled as an uncharacterized ArCR. Several additional close homologs of DUF128 were detected by PSI-BLAST searches.

COMPASS detected profile similarity between DUF128 and the HTH_5 family (Table 1), which includes bacterial transcription regulatory proteins from the arsR family. In the SCOP database, members of HTH_5 belong to the superfamily of "winged helix" DNA-binding domains, the DNA/RNA-binding 3-helical bundle fold (Fig. 2B). The alignment produced by COMPASS includes the N-terminal region of DUF128 that corresponds to the bundle of the three helices and the following β-hairpin (the "wing") in HTH_5. This alignment is shown in Figure 2A along with the manually refined multiple alignment of the two families. The multiple alignment reveals a significant similarity in the patterns of hydrophobicity over the major portion of both domains. The secondary structure prediction for the DUF128 representatives is consistent with the structure of the aligned regions in the HTH_5 family. These similarities indicate the putative secondary structure elements in the DUF128 domain (Fig. 2A).

Our prediction that the DUF128 and the HTH_5 families are homologs and possess similar structures was supported by the fold-recognition predictions produced by bioinbgu and 3D-PSSM servers. The top consensus prediction of the bioinbgu server was 1dprA, a structure that contains a winged helix DNA-binding domain, according to SCOP. The produced alignment and predicted secondary structure were consistent with the alignment generated by COMPASS. The top predictions by 3D-PSSM (with *E*-values ~ 0.1–0.3) included 1jmrA (newer version 1mkmA; Table 1), 1fk7A, and 1bib, all containing a winged helix DNA-binding domain.

When the N-terminal part of the DUF128 alignment was submitted to the Blocks Multiple Alignment Processor server and the resulting blocks were used for the LAMA search in the Blocks+ database, similarities to three blocks of HTH motifs were detected, the highest similarity being to the ArsR family (Table 1). Prof_sim generated the alignment that included approximately the same regions as were included in the COMPASS alignment, although the *P*-value corresponded to the high *E*-value of 66.3 (Table 1). As an additional support of our hypothesis, one of the DUF128 homologs from an archaeal genome (Slesarev et al. 2002) detected by PSI-BLAST in the NCBI nr database (GenBank GI 20093775) was recently annotated as a predicted transcriptional regulator containing a wHTH DNA-binding domain. However, to our knowledge, the prediction has not been discussed in the literature and was not reflected in public databases of protein families.

A relatively small size of the complete domain predicted within DUF_128 allowed us to use the ROSETTA method for ab initio structure prediction. The N-terminal sequence region of 71 residues from a DUF128 protein (GenBank GI 20093775), which was included in the DUF128/HTH_5 alignment, was submitted to ROSETTA. One of the top-ranking structure predictions produced by ROSETTA, the center of the second largest decoy cluster (29 decoys), was strikingly similar to the known structure of an HTH_5 protein, 1smtB (Fig. 2C). This structure was consistent with the prediction made by COMPASS. The higher-ranking center of the biggest decoy cluster (50 decoys) did not represent a reasonably folded structure. The situation when the correct prediction corresponds not to the top decoy cluster but to one of the lower-ranking clusters frequently occurs when using ROSETTA (Bonneau et al. 2001; Simons et al. 2001), and at least several largest decoy clusters should be examined for each set of ROSETTA results.

Furthermore, the structural identities of residues that were indicated by ROSETTA appeared to be a useful additional source of information for the manual refinement of the DUF128/HTH_5 alignment. The alignment of the structure predicted by ROSETTA and the structure of HTH_5 protein 1smtB (Fig. 2A) was consistent with the COMPASS alignment in the region of helix C (Fig. 2A, highlighted with yellow), but it was significantly different in the region of helix B and the turn between helices B and C (Fig. 2A, highlighted with red). The alignment of this region proposed by ROSETTA was more biologically reasonable, because (1) it lined up the DUF128 regions with strong helical propensities (B and C) to the helices of the same length in the HTH protein; and (2) it located a structurally reasonable position of the functionally important turn between the two helices. Therefore, the ROSETTA alignment was used in the refinement of the initial COMPASS result. The alignment in the turn region, which comprises the functional DNA-binding site, was especially challenging, because in DUF128 this region presumably contains a two-residue insertion and is different from the typical turn in HTH proteins (Fig. 2A, the site marked with the red line in the ROSETTA alignment).

Clathrin repeats are closest homologs of PPR motif

PPR (pentatricopeptide repeats) is a family of ~35-residue repeats of unknown function, which are found in eukaryotic proteins, with especially wide distribution in plants. No structures of PPR have been solved. This family was hypothesized to be related to a large and functionally diverse superfamily of TPRs (tetratricopeptide repeats; Small and Peeters 2000) involved in protein–protein interactions (Blatch and Lassle 1999; Groves and Barford 1999; Andrade et al. 2001).

COMPASS detected a significant similarity of PPR to one particular family of the TPR-related motifs, clathrin repeats (Clathrin in PFAM 7.6 and later versions; Table 1). This family contains repeats from the arm region of the clathrin heavy chain and from vacuolar protein-sorting (VPS) proteins. The heavy chains of clathrin include seven such tandem repeats of two antiparallel α-helices, which form a groove that may be used for protein–protein interactions (Fig. 3).

COMPASS alignment of PPR to clathrin repeats included major portions of these motifs and was manually extended to the full-length alignment (Fig. 3A). This alignment reveals similar patterns of hydrophobicity between the two families. The predicted secondary structure elements in the PPR motif (two helices connected by a loop) are consistent with the known secondary structure of clathrin repeats.

The sequences of the PPR family were too short to produce any hits when used as PSI-BLAST queries. Apparently the length of the query presented a problem for the fold-recognition servers as well: bioingbu did not accept the submission of short sequences, whereas 3D-PSSM produced no predictions with *E*-value less than 90, the top prediction being a C2H2 zinc-finger motif (Table 1). On the contrary, the scoring system implemented in COMPASS allows statistically significant alignments of short regions because it takes into consideration the effective number of residues in the alignment columns. Thus, the method may assign low *E*-values to the similarities between short but "thick" alignment segments, which reflects the higher reliability of statistical sampling derived from thick alignments.

When the PFAM alignment of the PPR family was submitted to the Blocks Multiple Alignment Processor and the resulting blocks were used as queries for the LAMA search in the Blocks+ database, no hits were found that would correspond to the similarity between PPR and clathrin repeats. The alignment of PPR and Clathrin_repeat profiles generated by prof_sim was assigned the *P*-value corresponding to the high *E*-value of 13.2 (Table 1).

When a sequence of the PPR family (GenBank GI 1705915) was submitted to ROSETTA for ab initio structure prediction, the top-ranking structure produced by ROSETTA was highly similar to the clathrin repeat structure (Fig. 3B,C). It included two antiparallel helices whose location in the sequence was consistent with COMPASS alignment. This ROSETTA prediction further confirms our hypothesis that the clathrin repeat motif is the closest homolog of the PPR motif.

A product of the protein precursors from Astro_capsid family is homologous to Viral_coat family

Astro_capsid is the family of viral capsid protein precursors encoded by astrovirus ORF2, one of the three astrovirus ORFs. The members of this family are found in *Astroviridae* and in avian nephritis virus. The 87-kD precursor undergoes an intracellular cleavage to form a 79-kD protein, and the subsequent extracellular trypsin cleavage produces the three proteins that form the infectious virion (Bass and Qiu 2000).

COMPASS assigned a low *E*-value to an extended alignment of Astro_capsid to the Viral_coat family (Table 1). Viral_coat contains the S domain of the capsid proteins from plant icosahedral positive-strand RNA viruses (Dolja and Koonin 1991), the domain that forms the virion shell. COMPASS detected the similarity between this domain and the first of the three Astro_capsid proteins (Fig. 4). After the manual refinement, it was possible to extend COMPASS alignment to the full length of Viral_coat sequences (Fig. 4A). The resulting multiple alignment reveals the significant similarity in the patterns of hydrophobicity and in the location of small residues over the whole length of the first Astro_capsid product and the Viral_coat domains. This similarity infers their homology and indicates that the first Astro_capsid protein possesses an all-β structure similar to that of the S domain of the capsid proteins from plant icosahedral positive-strand RNA viruses.

The top fold-recognition predictions for the N-terminal part of the Astro_capsid precursor were consistent with COMPASS results: both 3D-PSSM and bioinbgu servers predicted the highest similarity to a plant icosahedral virus protein (PDB ID 1bmv, Table 1).

This homology could not be detected using Astro_capsid members as queries for extensive PSI-BLAST searches with default parameters. However, for some of the Astro_capsid sequences, PSI-BLAST predicted similarity to the viral coat proteins with *E*-values higher than the default cutoff. For the blocks prepared from the Astro_capsid alignment with the Blocks Multiple Alignment Processor, several similar blocks of viral capsid proteins were detected in the Blocks+ database by LAMA: coronavirus nucleocapsid proteins (Table 1), as well as herpesvirus UL25 proteins involved in virus penetration and capsid assembly (IPB002493A), tymovirus coat proteins (IPB000574A), and Gag retroviral nucleocapsid proteins (IPB000721B). Prof_sim produced a shorter alignment, with the residue equivalences different from that of COMPASS and a high *P*-value (Table 1).

Discussion

Here we present several examples of novel similarities between PFAM protein families detected by COMPASS. We consider these family relationships interesting because they allow structure and function predictions for proteins with unresolved structure and reveal the evolutionary connections that were not previously reported in the literature. These similarities were assigned low *E*-values after a single COMPASS search in the set of 1354 PFAM alignments. However, they could not be automatically found by iterative PSI-BLAST searches with default parameters in the NCBI

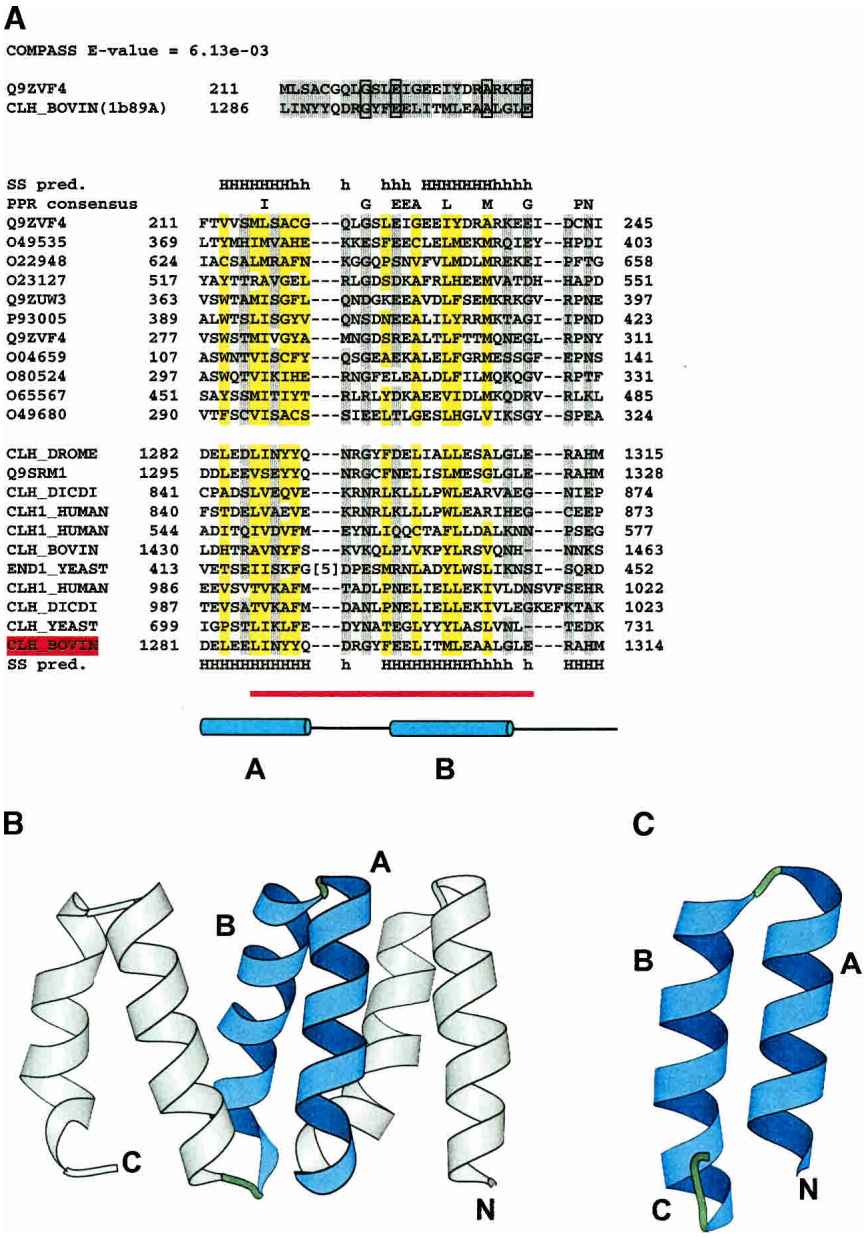

Figure 3. Sequence similarity between the PPR and clathrin families of PFAM indicates that clathrin repeats are the closest homologs of the PPR motif. (*A*) Profile–profile alignment, as constructed by COMPASS (shown as the alignment of two representative sequences), and manually refined multiple alignment including representatives from PPR (*top*) and clathrin (*bottom*). The TrEMBL identifier and first residue number are shown for each sequence. In the initial COMPASS alignment, matches with positive scores are highlighted in gray, and invariant residues are boxed. In the refined multiple alignment, the uncharged residues (all amino acids except D, E, K, and R) in mostly hydrophobic sites are highlighted in yellow, the nonhydrophobic residues (all amino acids except W, F, Y, M, L, I, and V) at mostly hydrophilic sites are highlighted in light gray, and the small residues (G, P, A, S, C, T, V) at positions occupied by mostly small residues are shown in red letters. Long insertions are not displayed: The numbers of omitted residues are specified in brackets. The identifier of the sequence with known spatial structure in the clathrin family (PDB Id 1b89A) is highlighted in red. The PHD secondary structure predictions (SS pred.) are shown for this sequence and for the *top* sequence of the PPR alignment. The actual secondary structure of the 1b89A fragment is shown *below* the alignment, with secondary structure elements labeled and colored according to the scheme shown in *B*. α-Helices are displayed as cylinders. The region covered by the initial COMPASS alignment is shown with a red line *below* the multiple alignment. (*B*) A ribbon diagram of the fragment of the clathrin heavy chain (PDB ID 1b89A) that was drawn by MOLSCRIPT (Kraulis 1991). N and C termini are labeled. The highlighted region corresponds to the alignment in *A*; α-helices are colored in blue. (*C*) The ribbon diagram of ab initio prediction of the tertiary structure for the *top* sequence in the PPR alignment (GenBank GI 1705915) made by ROSETTA. N and C termini are labeled. α-Helices are labeled and colored according to the scheme shown in *B*.

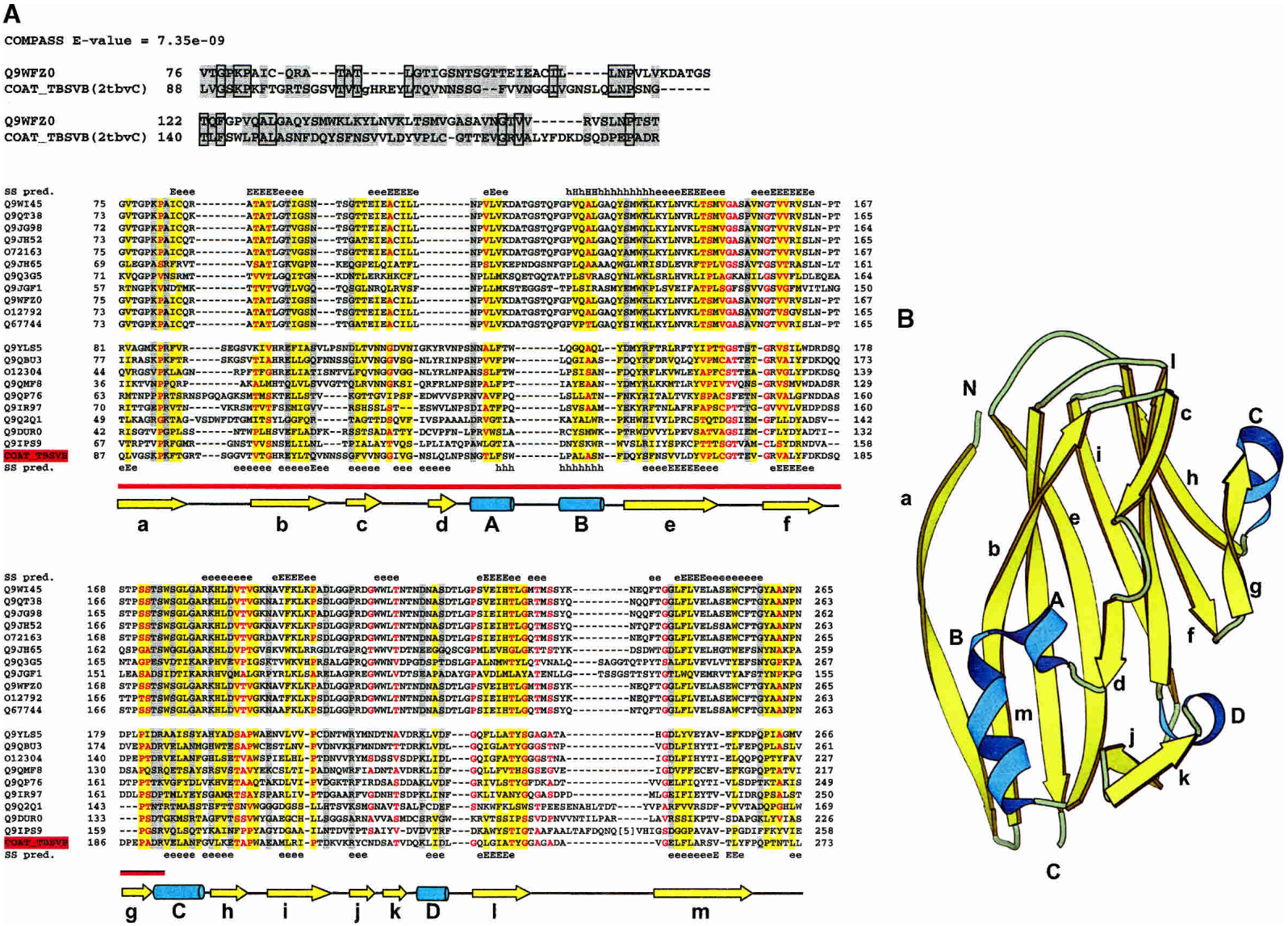

Figure 4. (Legend on next page)

nr database. In several cases (DUF128, DUF185, and Astro_capsid families), the similarities found by COMPASS were also detected by another method for profile–profile comparison, LAMA, which produced high *Z*-scores for ungapped alignments between short alignment blocks. The third available profile–profile comparison method, prof_sim, assigned too conservative estimates of *E*-values for the produced alignments, although in several cases, the alignments were similar to those constructed by COMPASS.

All of the presented COMPASS hits were independently confirmed by other methods for protein structure predictions. Most of the predicted similarities could be found by a skilled PSI-BLAST user among hits with high *E*-values, after a number of PSI-BLAST iterations. The secondary structure predictions for the families with unknown structure were consistent with the alignments produced by COMPASS. The predictions made by the fold-recognition (threading) servers (3D-PSSM and bioinbgu), which use versatile additional information (matching of secondary structure elements, propensities of the residues in the respect of solvent accesibility, etc.), were also generally in a good accord with COMPASS results. Finally, in the two cases in which the predicted protein regions were short enough to use ROSETTA, it produced the ab initio structural predictions that were strikingly similar to the proteins with known structure found by COMPASS. Although the presented similarities between the PFAM families could be predicted by other methods, the search by COMPASS was particularly well-fitted for this task, because this method uses only the minimal sequence information from the two compared alignments, and it makes the statistically significant predictions after a single search of the profile database.

The difference between the scoring systems implemented in COMPASS and PSI-BLAST becomes especially pronounced in the case of short regions of similarity between two alignments, which is presented by the found similarity between 34–35-residue motifs, PPR, and clathrin repeats (Fig. 3). The scoring system implemented in COMPASS allows statistically significant alignments of short regions because it explicitly takes into account the effective counts of amino acids in the alignment columns rather than the residue frequencies alone. Thus, two alignment columns with similar amino acid frequency distributions will receive a higher matching score if they contain higher effective numbers of residues, because such a positive match is statistically more significant. In many cases (not shown), COMPASS reveals local structural similarities between the families, which are insufficient for homology prediction but represent important localized structural motifs, such as P-loops (Walker A), FAD/NAD binding motifs, Zn fingers, and so on. The evolutionary implications of such local structural similarities between different protein folds have been discussed, arguments being made for both convergent and divergent evolution (Fetrow and Godzik 1998; Copley et al. 2001; Lupas et al. 2001).

As for any method of sequence similarity detection, the produced alignments of interest should be further manually refined using the additional predictions provided by other methods. In this respect, ab initio structure predictions made by ROSETTA appeared to be a valuable independent source of information about the reasonable structural identities between the two families. When the tertiary structure suggested by ROSETTA possesses the correct protein fold, it may be superimposed with the known structure of the homolog predicted by another method and produce an alignment of structurally identical residues. These residue identities may be especially useful for the improvement of a sequence-derived alignment, because they are based on a reasonable folding of the polypeptide chain rather than on the information derived from sequence databases.

The PFAM database has been developing for a long time (Sonnhammer et al. 1997, 1998; Bateman et al. 1999, 2000, 2002), and has been extensively investigated for possible similarities between families using different approaches, which include manual inspection of alignment seeds by experts (Sonnhammer et al. 1997, 1998; Schultz et al. 1998; Bateman et al. 1999, 2000, 2002; Yona et al. 2000; de Bakker et al. 2001; Aloy et al. 2002; Pandit et al. 2002). The fact that COMPASS predicted new relations within this database of alignments indicates the high sensitivity of this method and its potential value for the discovery of previously unknown similarities. In particular, COMPASS

Figure 4. Sequence similarity between the Astro_capsid and Viral_coat families of PFAM indicates structural similarity between two viral capsid proteins. (*A*) Profile–profile alignment, as constructed by COMPASS (shown as the alignment of two representative sequences), and manually refined multiple alignment including representative sequences from Astro_capsid (*top*) and Viral_coat (*bottom*). The TrEMBL identifier and first residue number are shown for each sequence. In the initial COMPASS alignment, matches with positive scores are highlighted in gray, and invariant residues are boxed. In the refined multiple alignment, the uncharged residues (all amino acids except D, E, K, and R) in mostly hydrophobic sites are highlighted in yellow, the nonhydrophobic residues (all amino acids except W, F, Y, M, L, I, and V) at mostly hydrophilic sites are highlighted in light gray, and the small residues (G, P, A, S, C, T, V) at positions occupied by mostly small residues are shown in red letters. Long insertions are not displayed: The numbers of omitted residues are specified in brackets. The identifier of the sequence with known spatial structure in Viral_coat (PDB Id 2tbvC) is highlighted in red. The PHD secondary structure predictions (SS pred.) are shown for this sequence and for the *top* sequence of the Astro_capsid alignment. The actual secondary structure of the 2tbvC fragment is shown *below* the alignment, with secondary structure elements labeled and colored according to the scheme shown in *B*. α-Helices and β-strands are displayed as arrows and cylinders, respectively. The region covered by the initial COMPASS alignment is shown with a red line *below* the multiple alignment. (*B*) A ribbon diagram of the fragment of the S domain of the capsid proteins from plant icosahedral positive-strand RNA viruses (PDB ID 1smtB) that was drawn by MOLSCRIPT (Kraulis 1991). N and C termini are labeled. α-Helices are colored in blue; β-strands are colored in yellow.

search may be a useful initial step in the characterization of a novel alignment, because it provides an opportunity for fast and relatively easy detection of similarities to existing protein families.

Acknowledgments

We thank Lisa Kinch and James Wrabl for discussion and critical reading of the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Aloy, P., Oliva, B., Querol, E., Aviles, F.X., and Russell, R.B. 2002. Structural similarity to link sequence space: New potential superfamilies and implications for structural genomics. *Protein Sci.* **11:** 1101–1116.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. 2001. Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.* **134:** 117–131.

Bass, D.M. and Qiu, S. 2000. Proteolytic processing of the astrovirus capsid. *J. Virol.* **74:** 1810–1814.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27:** 260–262.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. 2000. The Pfam protein families database. *Nucleic Acids Res.* **28:** 263–266.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.

Blatch, G.L. and Lassle, M. 1999. The tetratricopeptide repeat: A structural motif mediating protein–protein interactions. *Bioessays* **21:** 932–939.

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C.E., and Baker, D. 2001. Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins* **Suppl 5:** 119–126.

Copley, R.R., Russell, R.B., and Ponting, C.P. 2001. Sialidase-like Asp-boxes: Sequence-similar structures within different protein folds. *Protein Sci.* **10:** 285–292.

de Bakker, P.I., Bateman, A., Burke, D.F., Miguel, R.N., Mizuguchi, K., Shi, J., Shirai, H., and Blundell, T.L. 2001. HOMSTRAD: Adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics* **17:** 748–749.

Dolja, V.V. and Koonin, E.V. 1991. Phylogeny of capsid proteins of small icosahedral RNA plant viruses. *J. Gen. Virol.* **72 (Pt 7):** 1481–1486.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.

Fetrow, J.S. and Godzik, A. 1998. Function driven protein evolution. A possible proto-protein for the RNA-binding proteins. *Pac. Symp. Biocomput.* **3:** 485–496.

Fischer, D. 2000. Hybrid fold recognition: Combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.* **5:** 119–130.

Gotoh, O. 1993. Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.* **9:** 361–370.

———. 1994. Further improvement in methods of group-to-group sequence alignment with generalized profile operations. *Comput. Appl. Biosci.* **10:** 379–387.

Groves, M.R. and Barford, D. 1999. Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9:** 383–389.

Jaroszewski, L., Rychlewski, L., and Godzik, A. 2000. Improving the quality of twilight-zone alignments. *Protein Sci.* **9:** 1487–1496.

Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L., and Hughey, R. 1999. Predicting protein structure using only sequence information. *Proteins* **37:** 121–125.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299:** 499–520.

Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24:** 946–950.

Kunin, V., Chan, B., Sitbon, E., Lithwick, G., and Pietrokovski, S. 2001. Consistency analysis of similarity between multiple alignments: Prediction of protein function and fold structure from analysis of local sequence motifs. *J. Mol. Biol.* **307:** 939–949.

Lupas, A.N., Ponting, C.P., and Russell, R.B. 2001. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134:** 191–203.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Pandit, S.B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S.S., Mhatre, N.S., Sowdhamini, R., and Srinivasan, N. 2002. SUPFAM—A database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: Implications for structural genomics and function annotation in genomes. *Nucleic Acids Res.* **30:** 289–293.

Pietrokovski, S. 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* **24:** 3836–3845.

Przybylski, D. and Rost, B. 2002. Alignments grow, secondary structure prediction improves. *Proteins* **46:** 197–205.

Rost, B. 1996. PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266:** 525–539.

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9:** 232–241.

Sadreyev, R.I. and Grishin, N.V. 2003. COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326:** 317–336.

Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L., and Altschul, S.F. 1999. IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15:** 1000–1011.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., and Altschul, S.F. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29:** 2994–3005.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci.* **95:** 5857–5864.

Simons, K.T., Strauss, C., and Baker, D. 2001. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306:** 1191–1199.

Slesarev, A.I., Mezhevaya, K.V., Makarova, K.S., Polushin, N.N., Shcherbinina, O.V., Shakhova, V.V., Belova, G.I., Aravind, L., Natale, D.A., Rogozin, I.B., et al. 2002. The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci.* **99:** 4644–4649.

Small, I.D. and Peeters, N. 2000. The PPR motif—A TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci.* **25:** 46–47.

Sonnhammer, E.L., Eddy, S.R., and Durbin, R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **28:** 405–420.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26:** 320–322.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:** 22–28.

Yona, G. and Levitt, M. 2002. Within the twilight zone: A sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.* **315:** 1257–1275.

Yona, G., Linial, N., and Linial, M. 2000. ProtoMap: Automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.* **28:** 49–55.