Nucleic Acids Research Advance Access published September 12, 2007

Nucleic Acids Research, 2007, 1–7
doi:10.1093/nar/gkm698

MALISAM: a database of structurally analogous motifs in proteins

Hua Cheng[1,2], Bong-Hyun Kim[2] and Nick V. Grishin[1,2,*]

[1]Howard Hughes Medical Institute and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390-9050, USA

Received July 2, 2007; Revised August 16, 2007; Accepted August 22, 2007

ABSTRACT

MALISAM (manual alignments for structurally analogous motifs) represents the first database containing pairs of structural analogs and their alignments. To find reliable analogs, we developed an approach based on three ideas. First, an insertion together with a part of the evolutionary core of one domain family (a hybrid motif) is analogous to a similar motif contained within the core of another domain family. Second, a motif at an interface, formed by secondary structural elements (SSEs) contributed by two or more domains or subunits contacting along that interface, is analogous to a similar motif present in the core of a single domain. Third, an artificial protein obtained through selection from random peptides or in sequence design experiments not biased by sequences of a particular homologous family, is analogous to a structurally similar natural protein. Each analogous pair is superimposed and aligned manually, as well as by several commonly used programs. Applications of this database may range from protein evolution studies, e.g. development of remote homology inference tools and discriminators between homologs and analogs, to protein-folding research, since in the absence of evolutionary reasons, similarity between proteins is caused by structural and folding constraints. The database is publicly available at http://prodata.swmed.edu/malisam.

INTRODUCTION

Homology and analogy are two alternative scenarios to explain structural similarities among proteins. Homologs inherit similarities from their common ancestor, while structural analogs converge to similar structures due to a limited number of energetically favorable ways to pack secondary structural elements (SSEs) (1,2). Homology is inferred from sequence, structure and functional similarities, and several databases exist to catalog homologous proteins, e.g. Pfam (3), SCOP (4) and CATH (5). In contrast, analogy is more difficult to assess, and currently, no database exists that is devoted to structural analogs. Traditionally, studies aimed at discriminating homologs and analogs from their structures use domains classified in the same SCOP superfamily as homologs and domains classified in the same SCOP fold but different superfamilies as analogs (6–8). However, SCOP authors do not explicitly state that members of different superfamilies are necessarily analogous. Domains are grouped in SCOP superfamilies when convincing evidence for their homology exists. When additional evidence comes to light, for instance, through newly determined structures, superfamilies may be merged (9). Moreover, many studies argue for homology between certain SCOP superfamilies or even folds (10–13). Hence, using domains in the same SCOP fold but different superfamilies as structural analogs is problematic because some domains in this category may be homologous.

In an attempt to assemble a reliable set of structural analogs, we construct the MALISAM (manual alignments for structurally analogous motifs) database. By definition, analogy refers to those structural similarities that are not caused by evolutionary relatedness. Thus, to identify structural analogs convincingly, we need to minimize the possibility of homology. To this end, we explore three situations where we are more confident that the structural similarities have arisen in the absence of evolutionary relationships. In the first situation, a hybrid motif, formed by an insertion together with part of the evolutionary core of one domain family, may fortuitously match the core motif in another evolutionarily unrelated domain family. Second, an interface motif, formed by part of each domain or subunit contacting along a domain–domain or subunit–subunit interface, may happen to be structurally similar to the core motif of another unrelated domain family. And third, an artificial protein, obtained through selection from random peptides or sequence design experiments not biased by sequences of a particular homologous family, is a *bona fide* analog to a structurally similar,

*To whom correspondence should be addressed. Tel: +214 645 5952; Fax: +214 645 5948; Email: grishin@chop.swmed.edu

natural protein. To determine if a structural motif is a hybrid motif, an interface motif, or a core motif, we use the descriptions of the involved fold, superfamily or family in the SCOP database, in particular, SCOP definitions of fold cores (4).

Currently, MALISAM does not include examples of structural analogy between stand-alone, compact domains because, for stand-alone domains, we cannot argue for analogy convincingly by ruling out homology, especially when the domains share stronger structural similarity. It is our hope that, by studying the reliable analogs in MALISAM, researchers can gain a better understanding of structural analogy and develop better approaches to discriminating between remote homologs and analogs.

CONTENTS AND METHODS

In this work, 'a structural motif' refers to a set of SSEs with a specific spatial arrangement (architecture) and sequential connectivity (topology), allowing for circular permutations (14). Two structural motifs are regarded as analogs if they come from evolutionarily distinct domain families, but share the same architecture and topology, allowing for circular permutations. 'A domain' refers to a possible evolutionary unit with a well-defined hydrophobic core. We do not always follow the SCOP definition of a 'domain'. For instance, SCOP has a whole 'multi-domain protein' class, in which, by definition, a SCOP 'domain' is composed of more than one structural domain. In addition, our 'domains' frequently correspond to duplicated 'structural repeats' in SCOP, when such repeats constitute evolutionary units and form a well-defined hydrophobic core within each repeat with fewer interactions between repeats.

The 130 pairs of analogous motifs in MALISAM are organized in the following three categories corresponding to the three situations discussed in Introduction section.

A hybrid motif and a core motif

In evolution, a domain family usually preserves a common core structure while accumulating insertions and deletions in the periphery (15). A core motif is composed entirely of SSEs belonging to the evolutionary core of a domain family, while a hybrid motif is composed of both core elements and peripheral insertions. A hybrid motif in one family may happen to be structurally similar to the core motif in another evolutionarily unrelated family. For instance, a hybrid β-grasp motif found in formate dehydrogenase is formed by part of the ferredoxin core and a large C-terminal extension. This hybrid β-grasp motif is analogous to the core β-grasp motif in protein L (16). This structural phenomenon is a special case of a general property called gregariousness (17), i.e. proteins displaying partial structural similarity to many, frequently unrelated proteins.

Consider two ancient domain families with evolutionary cores denoted as AB and *BC*, respectively. As represented by B and *B*, these two families show partial similarity between their cores. Although most members in the first family have structure AB, some members carry an additional part C and have structure ABC. The BC motif in these proteins is structurally similar to the core motif *BC* in the second family. To justify analogy between BC and *BC*, it is necessary to demonstrate that C is an insertion rather than a deletion, i.e. the first family evolved from AB to ABC instead of from ABC to AB. If C is not present in the majority of the family members and is restricted to a particular phylogenetic or functional group within the family, it is most likely that proteins with structure AB are ancient and proteins with structure ABC are comparatively new. In other words, AB is the ancient core of the family, and C is a recent insertion. Thus, though AB and *BC* show partial structural similarity, they are not homologous, both being ancient and having different cores. Since AB and ABC are homologous, ABC and *BC* cannot be homologous. Therefore, the hybrid motif BC and the core motif *BC* are structural analogs.

Figure 1 illustrates an analogous pair comprised of a hybrid motif and a core motif. The core of the SCOP superfamily 'glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase' consists of four β-strands and one α-helix with the sequential connection of βαβββ. One representative of this superfamily, the first domain in the antibiotic resistance protein FosA from *Pseudomonas aeruginosa* (18), is diagramed in Figure 1a. In this domain (termed 'structural repeat' having glyoxalase fold in SCOP), the βαβββ motif is a core motif, and no other SSEs are present. Figure 1b depicts another domain that contains a βαβββ motif, namely, the first domain in the oligopeptide-binding protein OPPA from *Salmonella typhimurium* (19), which belongs to the SCOP superfamily 'periplasmic-binding protein-like II'. The core of this superfamily includes five β-strands and their connecting helices, as illustrated in Figure 1c by a more typical member in this superfamily, the first domain in lysine/arginine/ornithine-binding protein (LAO) from *S. typhimurium* (20). Although FosA and OPPA are classified in different SCOP superfamilies, folds and even classes, the two βαβββ motifs boxed in Figure 1a and b are similar in both architecture and topology. A comparison of Figure 1b and c suggests that the β-hairpin highlighted in red, which is present in OPPA but absent in LAO, is an insertion. Indeed, a manual inspection of the structures in this superfamily shows that the β-hairpin is present in only 4 out of the superfamily's 32 proteins. Moreover, these four proteins form a more closely related subgroup as indicated by their higher sequence identities to one another than to other proteins in this superfamily. All these observations suggest that the β-hairpin does not belong to the ancient core of this superfamily and is instead an insertion that appeared later in evolution. Thus, the βαβββ motif in OPPA (Figure 1b) has a hybrid origin: the red β-hairpin is an insertion while the remaining two strands and one helix are core elements. Therefore, the hybrid βαβββ motif in OPPA and the core βαβββ motif in FosA (Figure 1a) define a pair of analogs.

We use the following procedure to find analogous pairs consisting of a hybrid motif and a core motif. First, we represent the most commonly found structure motifs (21,22) or their circularly permuted forms by matrices and

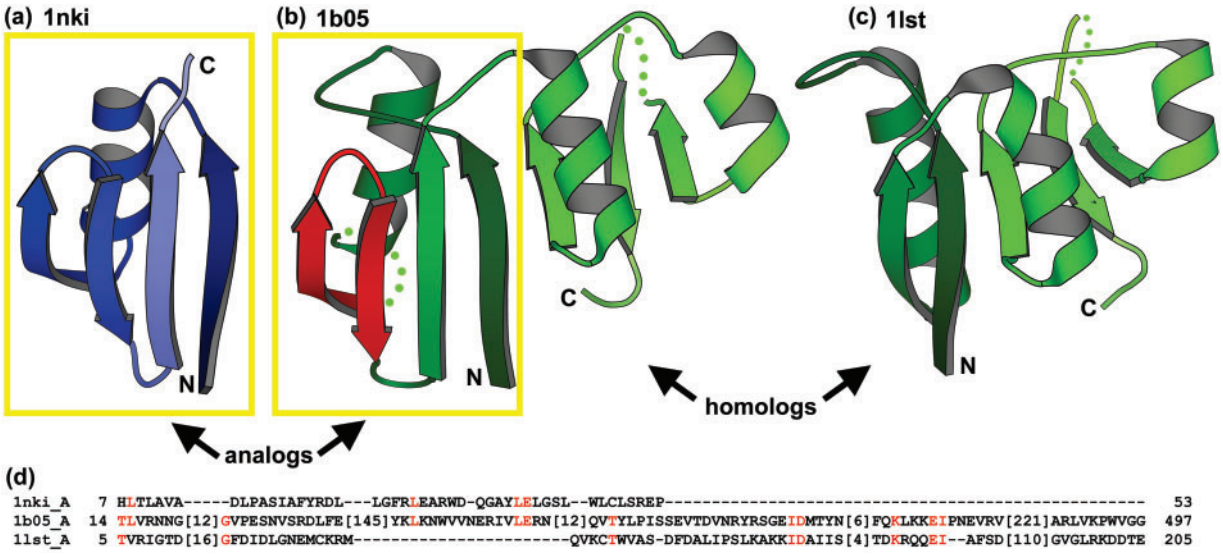

Figure 1. An analogous pair composed of a hybrid motif and a core motif. (**a**) The first domain in fosfomycin resistance protein A (FosA) from *P. aeruginosa* (PDB code 1nki). The core region is colored from deep blue (N-terminus) to pale blue (C-terminus). The core βαββ motif is framed in yellow. (**b**) The first domain in oligopeptide-binding protein (OPPA) from *S. typhimurium* (PDB code 1b05). The core region is colored from deep green (N-terminus) to pale green (C-terminus), and the inserted β-hairpin is colored in red. The hybrid βαββ motif is framed in yellow. (**c**) The first domain of lysine/arginine/ornithine-binding protein (LAO) from *S. typhimurium* (PDB code 1lst). The core region is colored from deep green (N-terminus) to pale green (C-terminus). Discontinuous regions are represented by dotted curves. Diagrams are generated by MOLSCRIPT (37). (**d**) Structure-based sequence alignment of the three domains. The PDB code, chain identifier and starting and ending residue numbers are given for each sequence. Number of omitted residues is indicated in brackets. Identical residue pairs are highlighted in red.

use these matrices as queries in ProSMoS (23) searches to find SCOP domains that contain the query motif. The found domains are then grouped by SCOP superfamily. We inspect one randomly selected representative domain from each superfamily to determine whether the query motif occurring in this superfamily is a hybrid or not, according to SCOP description of the core of this superfamily. SSEs that do not belong to the core are considered to be insertions, and a hybrid motif should contain both core elements and inserted elements. The coordinates of a hybrid motif are extracted from the original PDB file and used as a query in DALI (24) searches against a culled PDB database, in which any two proteins share <50% sequence identity. The hits are sorted by DALI Z score and inspected manually to select the most similar core motif(s) to the query. The hybrid motif and the selected core motif are then added to the database as a pair of analogs. Currently, there are 91 pairs of such analogs in MALISAM.

An interface motif and a core motif

An interface motif is formed by part of each domain or subunit contacting along a domain–domain or subunit–subunit interface. In compact multi-domain proteins or obligate multi-chain complexes, where domains or subunits are closely associated, an interface motif may be well defined enough to mimic the structure and topology of a core motif, which is composed entirely of SSEs belonging to the evolutionary core of a domain family. For the interface motif and the core motif to be a pair of analogs, they should come from evolutionarily unrelated domain families. Particularly, the interface motif should not cover the complete core of any of the contributing domains or subunits because in that case, the core motif might be a duplicated or elaborated version of that completely covered domain or subunit.

Figure 2 shows an analogous pair comprised of an interface motif and a core motif. The actin filament capping protein CapZ (Figure 2a) is a heterodimer of two homologous subunits α and β (25). SCOP divides each subunit into three domains, with the N-terminal domain being a 3-helical bundle (diagramed in ribbons in Figure 2a). The two 3-helical bundle domains from the two subunits make many hydrophobic contacts at the dimer interface, forming a 4-helical bundle (framed in red). This 4-helical bundle defines an interface motif, consisting of part of the 3-helical bundle from the α subunit followed by part of the 3-helical bundle from the β subunit. Figure 2b depicts one monomer in the conserved hypothetical protein Xcc0516 homopentamer from *Xanthomonas campestris* (26). This 4-helical bundle in Xcc0516 comprises the whole polypeptide chain and is a core motif. The interface 4-helical bundle in CapZ (boxed in red in Figure 2a) shares the same overall structure and topology with the upper portion of the core 4-helical bundle in Xcc0516 (boxed in red in Figure 2b). Yet the interface motif comes from a 3-helical bundle domain family, while the core motif comes from an evolutionarily unrelated 4-helical bundle domain family. Therefore, the interface motif in CapZ and the core motif in Xcc0516 can be regarded as a pair of analogs.

The same procedure described above to find hybrid motifs is used to identify interface motifs. In addition, we manually inspected the SCOP entries in the 'multi-domain proteins' class or with 'duplication' in their annotations

in a search for interface motifs. The coordinates of an interface motif are extracted from the original PDB file and used as a query in DALI (24) searches against the aforementioned, culled PDB database to find the most similar core motif(s) for it. Note that, before running DALI, the coordinates file of a subunit–subunit interface motif is preprocessed so that it contains only one chain. Currently, MALISAM stores 33 pairs of analogs that are composed of an interface motif and a core motif.

An artificial protein and a natural protein

An artificial protein and a natural protein are a pair of analogs if they are structurally similar. An example of this situation is provided by Krishna and Grishin (27). We analyze the entries in the SCOP class 'designed proteins' and select the ones with *de novo* sequences and reasonably complex structures to serve as queries in DALI searches against the aforementioned representative PDB database. We pay attention to whether an artificial protein is designed without a sequence bias towards a particular homologous family, e.g. the artificially designed WW domain (PDB 1e0m) will not work as an analog to natural WW domains, because it is based largely on the consensus sequence of WW domain family. By inspecting top DALI hits, we select the most similar natural proteins for an artificial protein query. Currently, MALISAM contains six pairs of analogs that consist of an artificial protein and a natural protein, but we expect this number to increase due to active research in the field of protein sequence design.

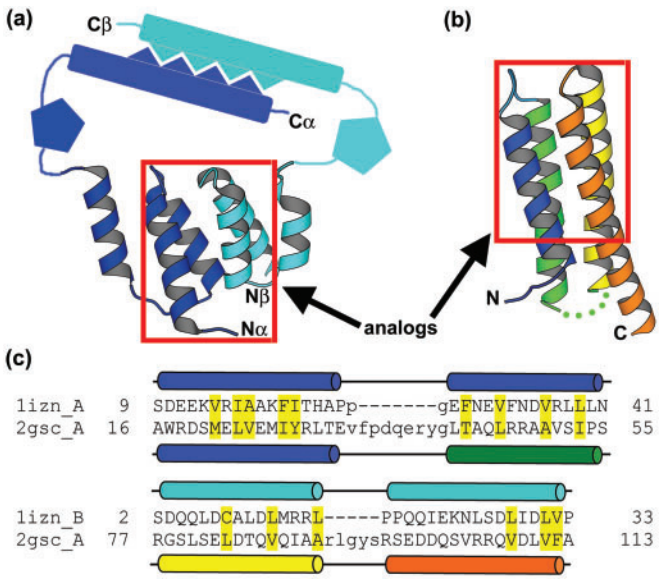

Figure 2. An analogous pair composed of an interface motif and a core motif. (**a**) The actin filament capping protein CapZ from *Gallus gallus* (PDB code 1izn). The α subunit is in blue, and the β subunit is in cyan. The three domains in each subunit are represented as a ribbon diagram, a pentagon and a rectangle, respectively. The interface 4-helical bundle motif is in red frame. (**b**) The conserved hypothetical protein Xcc0516 from *X. campestris* (PDB code 2gsc). The polypeptide chain is colored in rainbow from N-terminus (blue) to C-terminus (orange). The part that is aligned to the hybrid 4-helical bundle motif in CapZ is in red frame. The disordered region between the 2nd and the 3rd helices is represented by a dotted curve. Diagrams are generated by MOLSCRIPT (37). (**c**) Structure-based sequence alignment of the two 4-helical bundle motifs. The helices in the red frames in (**a**) and (**b**) are diagramed as cylinders in the same colors above or below their corresponding sequences. The PDB code, chain identifier and starting and ending residue numbers are given for each sequence. Columns with two hydrophobic residues are highlighted in yellow.

Alignments and scores

Each analogous pair is aligned manually by visual inspection and automatically by three programs [DALI (28), TM-align (29) and FAST (30)]. In constructing the manual alignment for a pair, we first superimpose the two motifs in the Insight II software according to their corresponding SSEs, then transform the structural superposition into sequence alignment following these general rules: (i) corresponding SSEs are aligned and loops are frequently ignored; (ii) H-bonding networks in β-sheets are followed, i.e. if two residues are aligned, their respective H-bond partners are also aligned and (iii) gap openings are avoided as much as possible.

Different aligners, manual or automatic, may give different structural alignments for the same pair of structures. For example, since β-strands are repeats of two-residue units, one can superimpose two β-strands in several ways or registers by shifting one strand relative to the other by two residues at a time. Such kind of ambiguity in structural alignments has been discussed in the literature (31,32). To quantify the similarity between a manual alignment and an automatic alignment, we compute an agreement by dividing the number of positions aligned the same way in the two alignments by the total number of aligned positions in the manual alignment. The distribution of these agreements is shown in Figure 3. Out of the three programs used, manual alignments agree the most with DALI. Frequently, the disagreements between different aligners result from the aforementioned register or shifting problem.

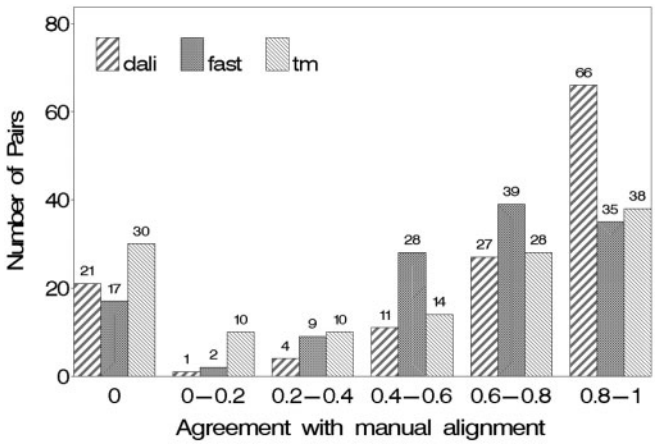

Figure 3. Agreement between manual and automatic alignments. The three programs are represented by three different patterns. The horizontal axis corresponds to the ranges of the agreement bins, and the vertical axis corresponds to the number of pairs that fall into each bin.

To characterize a manual or automatic alignment between two analogous motifs, we compute six alignment-based scores, namely, aligned length, sequence identity, Cα RMSD, GDT_TS (33), COMPASS-like score and Consensus. RMSD and GDT_TS quantify how close the two motifs are in the structure superposition. Sequence identity and COMPASS-like score measure how similar the two motifs are in sequence by comparing two single sequences or two multiple sequence alignments, respectively. In calculating the COMPASS-like score, two sequence profiles, one for each motif, are generated by running PSI-BLAST (34), aligned according to the manual or automatic structure alignment, and then compared and scored by the scoring function used in the COMPASS program (35). In calculating the consensus score, we compare the four alignments generated by the four aligners (Manual, DALI, TM-align and FAST), assuming that similarities captured by different aligners are more likely to be correct. For each aligner, the consensus score equals the percent of its alignment that is agreed by at least one of the other three aligners. A more detailed description of score calculation can be found in Ref. (36) or by clicking a score name on a pair-specific page at the MALISAM website.

Table 1 shows the mean and the standard error of each score and each aligner over the 130 pairs in MALISAM. Manual alignments are typically shorter than DALI and TM alignments, but longer than FAST alignments. Also, manual alignments have the best average RMSD, GDT_TS and consensus score, implying their higher overall quality than the automatic alignments.

WEB INTERFACE

The main webpage of MALISAM lists all the analogous pairs in the database. Clicking a pair name redirects the browser to that pair's specific page (Figure 4), which contains a summary table, a short justification for

Table 1. Mean and standard error of various scores for each aligner

	DALI	TM-align	FAST	Manual
Aligned length (a.a.)	60.96 ± 1.04	**61.12** ± 1.23	50.83 ± 1.02	56.69 ± 1.00
Sequence identity (%)	8.21 ± 0.32	8.06 ± 0.31	**8.61** ± 0.35	8.54 ± 0.36
RMSD (Å)	3.11 ± 0.05	3.06 ± 0.05	3.23 ± 0.22	**2.92** ± 0.06
GDT_TS (%)	59.05 ± 0.59	60.24 ± 0.63	57.18 ± 0.94	**60.53** ± 0.73
COMPASS- like	−11.46 ± 0.74	−12.08 ± 0.76	**−10.45** ± 0.76	−10.75 ± 0.72
Consensus (%)	67.02 ± 2.48	47.41 ± 2.90	66.84 ± 2.76	**73.93** ± 2.37

The mean and the standard error for each score and each aligner. For RMSD, a smaller value indicates higher similarity; for all other scores, a larger value means higher similarity. The best mean in each row is bolded.

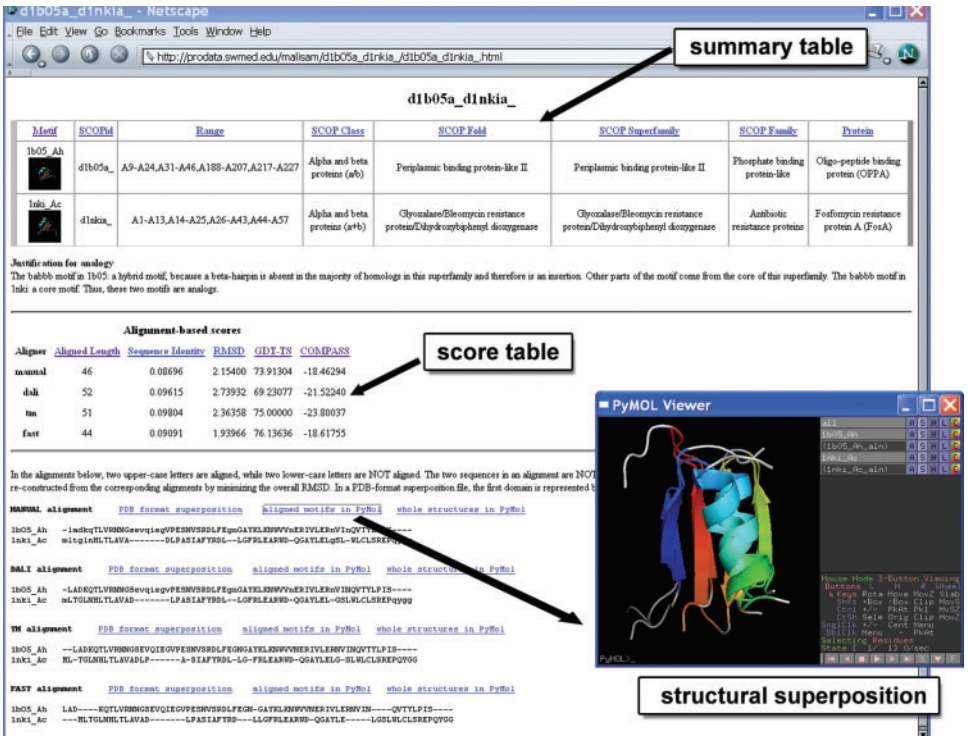

Figure 4. Web layout of an analogous pair in MALISAM.

analogy, a table of alignment-based scores and structure-based sequence alignments. The summary table provides basic information about the two motifs in that pair: PDB code, chain identifier, residue ranges and SCOP code and classification. The structural superposition reconstructed from a manual or automatic alignment can be viewed in PyMOL (http://pymol.sourceforge.net/) by selecting the 'aligned motifs in PyMol' link or can be downloaded by selecting the 'PDB format superposition' link. A compressed file of the entire database with alignments of all pairs can be downloaded from ftp://iole.swmed.edu/pub/cheng/analogs/analogs.tar.

CONCLUSION

The MALISAM database is a compilation of manual as well as automatic alignments between structurally analogous motifs. To the best of our knowledge, this is the first database devoted entirely to structural analogs. MALISAM may be used in protein-folding studies, for the similarities between analogs result from folding and packing constraints rather than evolutionary relatedness. A reliable set of structural analogs provided in MALISAM could be used in the training of remote homology detection methods, since these methods frequently involve discrimination between homologs and structural analogs. Additionally, MALISAM manual alignments should be useful for the development of better automatic structure alignment techniques.

ACKNOWLEDGEMENTS

We are grateful to Sara Cheek for useful scripts. We thank Lisa Kinch for critical reading of the manuscript and S. Sri Krishna for helpful discussions. This work was supported by NIH grant GM67165 to N.V.G. Funding to pay the Open Access publication charges for this article was provided by Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Orengo,C.A., Sillitoe,I., Reeves,G. and Pearl,F.M. (2001) Review: what can structural classifications reveal about protein evolution? *J. Struct. Biol.*, **134**, 145–165.
2. Finkelstein,A.V. and Ptitsyn,O.B. (1987) Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys Mol. Biol.*, **50**, 171–190.
3. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
4. 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.
5. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH – a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
6. Russell,R.B., Saqi,M.A., Sayle,R.A., Bates,P.A. and Sternberg,M.J. (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.*, **269**, 423–439.
7. Matsuo,Y. and Bryant,S.H. (1999) Identification of homologous core structures. *Proteins*, **35**, 70–79.
8. Dietmann,S. and Holm,L. (2001) Identification of homology in protein structure classification. *Nat. Struct. Biol.*, **8**, 953–957.
9. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
10. Ponting,C.P. and Russell,R.B. (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J. Mol. Biol.*, **302**, 1041–1047.
11. Copley,R.R. and Bork,P. (2000) Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.*, **303**, 627–641.
12. Aravind,L., Anantharaman,V. and Koonin,E.V. (2002) Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA world. *Proteins*, **48**, 1–14.
13. Kinch,L.N., Cheek,S. and Grishin,N.V. (2005) EDD, a novel phosphotransferase domain common to mannose transporter EIIA, dihydroxyacetone kinase, and DegV. *Protein Sci.*, **14**, 360–367.
14. Lindqvist,Y. and Schneider,G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.*, **7**, 422–427.
15. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
16. Krishna,S.S. and Grishin,N.V. (2005) Structural drift: a possible path to protein fold change. *Bioinformatics*, **21**, 1308–1310.
17. Harrison,A., Pearl,F., Mott,R., Thornton,J. and Orengo,C. (2002) Quantifying the similarities within fold space. *J. Mol. Biol.*, **323**, 909–926.
18. Rigsby,R.E., Rife,C.L., Fillgrove,K.L., Newcomer,M.E. and Armstrong,R.N. (2004) Phosphonoformate: a minimal transition state analogue inhibitor of the fosfomycin resistance protein, FosA. *Biochemistry*, **43**, 13666–13673.
19. Sleigh,S.H., Seavers,P.R., Wilkinson,A.J., Ladbury,J.E. and Tame,J.R. (1999) Crystallographic and calorimetric analysis of peptide binding to OppA protein. *J. Mol. Biol.*, **291**, 393–415.
20. Oh,B.H., Pandit,J., Kang,C.H., Nikaido,K., Gokcen,S., Ames,G.F. and Kim,S.H. (1993) Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand. *J. Biol. Chem.*, **268**, 11348–11355.
21. Ruczinski,I., Kooperberg,C., Bonneau,R. and Baker,D. (2002) Distributions of beta sheets in proteins with application to structure prediction. *Proteins*, **48**, 85–97.
22. Brenner,S.E., Chothia,C. and Hubbard,T.J. (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.*, **7**, 369–376.
23. Shi,S., Zhong,Y., Majumdar,I., Sri Krishna,S. and Grishin,N.V. (2007) Searching for three-dimensional secondary structural patterns in proteins with ProSMoS. *Bioinformatics*, **23**, 1331–1338.
24. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
25. Yamashita,A., Maeda,K. and Maeda,Y. (2003) Crystal structure of CapZ: structural basis for actin filament barbed end capping. *EMBO J.*, **22**, 1529–1538.
26. Lin,L.Y., Ching,C.L., Chin,K.H., Chou,S.H. and Chan,N.L. (2006) Crystal structure of the conserved hypothetical cytosolic protein Xcc0516 from *Xanthomonas campestris* reveals a novel quaternary structure assembled by five four-helix bundles. *Proteins*, **65**, 783–786.
27. Krishna,S.S. and Grishin,N.V. (2004) Structurally analogous proteins do exist! *Structure (Camb)*, **12**, 1125–1127.
28. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
29. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
30. Zhu,J. and Weng,Z. (2005) FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–627.
31. Godzik,A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.

32. Cheng,H. and Grishin,N.V. (2005) DOM-fold: a structure with crossing loops found in DmpA, ornithine acetyltransferase, and molybdenum cofactor-binding domain. *Protein Sci.*, **14**, 1902–1910.
33. Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.
34. 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.
35. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
36. Cheng,H., Kim,B.-H. and Grishin,N. (2007) MALIDUP: a database of manually constructed structure alignments for duplicated domain pairs. *Proteins: Structure, Function, and Bioinformatics*.
37. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946–950.