In Silico Biology 4, 0037 (2004); ©2004, Bioinformation Systems e.V.

Recognition of remotely related structural homologues using sequence profiles of aligned homologous protein structures

Seema Namboori, Narayanaswamy Srinivasan* and Shashi B. Pandit

Molecular Biophysics Unit, Indian Institute of Science
Bangalore 560 012, India

* Corresponding author
Email: ns@mbu.iisc.ernet.in
Phone: +91-80-2293 2837
Fax: +91-80-2360 0535

Edited by E. Wingender; received April 25, 2004; revised June 17 and July 22, 2004; accepted July 24; published August 06, 2004

Abstract

In order to bridge the gap between proteins with three-dimensional (3-D) structural information and those without 3-D structures, extensive experimental and computational efforts for structure recognition are being invested. One of the rapid and simple computational approaches for structure recognition makes use of sequence profiles with sensitive profile matching procedures to identify remotely related homologous families. While adopting this approach we used profiles that are generated from structure-based sequence alignment of homologous protein domains of known structures integrated with sequence homologues. We present an assessment of this fast and simple approach. About one year ago, using this approach, we had identified structural homologues for 315 sequence families, which were not known to have any 3-D structural information. The subsequent experimental structure determination for at least one of the members in 110 of 315 sequence families allowed a retrospective assessment of the correctness of structure recognition. We demonstrate that

correct folds are detected with an accuracy of 96.4% (106/110). Most (81/106) of the associations are made correctly to the specific structural family. For 23/106, the structure associations are valid at the superfamily level. Thus, profiles of protein families of known structure when used with sensitive profile-based search procedure result in structure association of high confidence. Further assignment at the level of superfamily or family would provide clues to probable functions of new proteins. Importantly, the public availability of these profiles from us could enable one to perform genome wide structure assignment in a local machine in a fast and accurate manner.

Key words: protein fold, superfamily, profiles, protein family, fold recognition, structural genomics, remote homologues

---

Introduction

The known three-dimensional (3-D) structure of proteins are classified in a hierarchical way in Structural Classification of Proteins (SCOP) based on their structural and functional features [Murzin *et al.*, 1995]. CATH [Orengo *et al.*, 1997] and DALI [Holm and Sander, 1993] also provide hierarchical classification of protein structures. In general, proteins with sequence identities of 30% or greater and having similar functions are classified into a family. Families with low sequence (<30%) identities, with similar structural and functional characteristics are indicative of possible common evolutionary origin and such families are clustered into a superfamily. Protein members of families and superfamilies are said to belong to the same fold if they exhibit similar spatial orientation of their major secondary structural elements in addition to similar topological connections. The importance of 3-D structural information lies in the fact that it provides useful insights into the molecular function and evolution of proteins. The number of proteins with known amino acid sequences is significantly higher as compared to those with known three-dimensional structures [Holm and Sander, 1996]. The problem is compounded by rapid accumulation of a very large number of putative protein sequences from genome sequencing project. This gap between proteins with known 3-D structural information and those with known sequences has significantly driven the efforts towards experimental determination of structures for as many proteins as possible. However, experimental methods could be time consuming and have their own limitations. Computational methods leading to reliable protein fold recognition and modeling could result in a useful structural framework until a high-resolution experimental structure becomes available. Computational recognition of protein folds could also aid in efforts of large-scale determination of structures in setting priority targets [Teichmann *et al.*, 1999; Brenner, 2000].

Various computational methods such as GenThreader [Jones, 1999], UCLA/DOE Fold assignment [Mallick *et al.*, 2002], 3D-PSSM [Kelley *et al.*, 2000] and FUGUE [Shi *et al.*, 2001] are frequently used for protein fold recognition. Such methods

undergo periodic evaluation through Critical Assessment of protein Structure Prediction (CASP) [Moult *et al.*, 2001]. The automated methods of assessment for structure prediction include CAFASP [Fischer *et al.*, 1999], EVA [Eyrich *et al.*, 2001] and LiveBench [Bujnicki *et al.*, 2001]. The traditional fold recognition approaches are, generally, based on sequence to structure compatibility. Two of the popular fold recognition methods, GenThreader and 3D-PSSM, involve alignment of protein sequence into a library of structural templates. This is performed by taking into consideration a combination of factors like secondary structure, solvent accessibility or residue-residue contact preferences [Sippl, 1990; Russell *et al.*, 1996; Rost *et al.*, 1997; Jones, 1999]. On the other hand, ab initio fold prediction methods involve use of various energy potentials to predict the fold starting from the primary structure, without using any prior structural information [Saunders *et al.*, 2002; Bonneau *et al.*, 2001].

Furthermore, one of the recent advancements in the area of structure predictions is to use multiple prediction servers (meta servers), such as 3D Jury [Ginalski *et al.*, 2003] and 3D-SHOTGUN [Fischer, 2003]. These take into consideration the common structural motifs from a compilation of 3D models generated by various structure prediction servers. This results in a more confident and reliable prediction. One of the meta prediction method 3D Jury has been shown to be comparable to other meta servers, however, it has the highest combined specificity and sensitivity [Ginalski and Rychlewski, 2003; von Grotthuss *et al.*, 2003]. Such an approach has been used to associate structures to Pfam families without a member of known structure in a database "Pfam-no-3D" (http://bioinfo.pl/Pfam-no-3D/). The current release of this database has 76 Pfam families without a member of known 3-D structure, but for which structures could be predicted.

It is well known that homologous proteins are characterized by high degree of structural resemblance [Balaji and Srinivasan, 2001]. Proteins with no obvious sequence similarities could adopt similar structures [Murzin *et al.*, 1995; Orengo *et al.*, 1997; Chothia and Gerstein, 1997; Holm and Sander, 1994]. Thus, methods for identifying fold using homologous sequence relationships are of practical importance. When the sequence similarities are high the relationships are detected easily using sequence based search methods like BLAST [Altschul *et al.*, 1990] and FASTA [Pearson and Lipman, 1988]. But, if the sequence similarity between the query and homologues in the sequence database is low the similarities are difficult to detect. However, the incorporation of multiple sequence alignment in the form of profiles or Position Specific Scoring Matrices (PSSMs) [Gribskov *et al.*, 1987], has enabled detection of remotely related proteins effectively. The profiles encapsulate the conserved patterns and variations therein, hence enabling the detection of distantly related proteins. There are numerous profile-based search tools such as PSI-BLAST [Altschul *et al.*, 1997], IMPALA [Schäffer *et al.*, 1999], HMMER [Eddy, 1998] and SAM-T98 [Karplus *et al.*, 1998] that identify remote homologues in a fast and effective manner. Moreover, sequence profiles have been shown to be useful in fold recognition [Koonin *et al.*, 2000].

The in-house PALI [Balaji *et al.*, 2001; Gowri *et al.*, 2003] resource provides 3-D structure-based sequence alignments of homologous proteins of known structure

that are integrated with homologous sequences identified from sequence databases. Thus, structure-based sequence alignments of large number of protein domains in various families could be used to generate sensitive profiles. The homologues in a family in PALI database have been derived from SCOP. The PALI family profiles have been used earlier to establish relationships between two homologous protein families, one with known structure from PALI and the other with unknown structure from Pfam as described in SUPFAM [Pandit *et al.*, 2002]. During such an exercise about one year ago we had proposed distant relationship between 315 Pfam families, with apparently no structural information, and families of known structure (http://pauling.mbu.iisc.ernet.in/~osupfam). Hence, for these Pfam families we could propose a possible framework structure. For 110 of these 315 families, the structure was subsequently elucidated using nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography methods. In retrospect, we have analyzed the correctness of remote similarity associations for these 110 Pfam families using SCOP hierarchical classification as a reference. We could associate correct folds for an overwhelming majority of Pfam families using PALI profiles. Thus, we propose that the profiles generated using multiple structure-dependent sequence alignments in PALI are valuable in associating structures for distantly related proteins. Tailoring these PSSMs, which are freely available from the authors, with programs such as RPS-BLAST that is available freely (at NCBI, USA), could serve as an important tool for genome-wide structure assignments. The advantage of this approach is that it could be set-up and run at the user's site and is fast enough to assign structures of proteins at the entire genome level in a reasonably short time. This approach is complementary to other effective and publicly available methods as SUPERFAMILY [Gough *et al.*, 2001] and Gene3D [Buchan *et al.*, 2002].

Materials and methods

Databases

The sequence-based domain families were obtained from Pfam (Version 7.2) [Sonnhammmer *et al.*, 1997; Bateman *et al.*, 2002] (http://www.sanger.ac.uk/Software/Pfam). The information as to whether or not a Pfam family has at least one of the members with known structure was retrieved from the flat files provided at the Pfam site. The integrated structure-sequence alignment, which corresponds to structural families, was obtained from the in-house PALI database (Release 2.1) (http://pauling.mbu.iisc.ernet.in/~pali). The non-redundant database (NRDB) was obtained from National Center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov/blast/db).

Profile generation for PALI families

RPS-BLAST searchable profiles or PSSMs were generated using PSI-BLAST [Altschul *et al.*, 1997] for every PALI family. As a first step in profile generation, PALI family sequences were integrated with homologous sequences from the corresponding Pfam families or Non-Redundant sequence Database (NRDB) using procedures described previously [Gowri *et al.*, 2003]. For this purpose, collection of seed alignments available for Pfam A families were used. These seed alignments comprise of representative sequences of various families and hence, in general, these homologues have low sequence similarities among themselves. When these sequences are integrated with corresponding families in PALI, the sequences are re-aligned to be consistent with the 3-D structural superposition based alignment available in PALI. 3-D structure-based alignments are expected to be more accurate than sequence similarity-based alignments especially for homologues characterized by low sequence similarities. In general, lengths of families in Pfam are longer than the corresponding PALI families. Hence, integration of PALI and Pfam families is also accompanied by pruning of domain boundaries of Pfam entries so that it is same as the domain boundaries defined in corresponding structural families.

In the next step, profile is generated for every PALI family using PSI-BLAST, by querying a reference sequence from the PALI family against its own integrated structure sequence database. A high quality of profile is ensured by providing structure-based multiple sequence alignment and structure-based domain boundaries as an input to PSI-BLAST.

Identification of structural homologues

The relationships between Pfam and PALI families were derived by searching each member of Pfam family against PALI profile database, using sensitive profile-based search method of RPS-BLAST. All these relationships derived about one year ago, corresponding to SUPFAM 1.2, have been opened publicly since then (http://pauling.mbu.iisc.ernet.in/~osupfam). An e-value cut off of 3×10^{-5} was used in order to extract valid hits.

Retrospective Assessment of remote structural homologue recognition

The approach used in the retrospective analysis is represented in Figure 1. The identification of a possible structure for a Pfam family relies on its association to a family of known structure from PALI or another intermediately related Pfam family that is related to a structural family. For the retrospective assessment two datasets of parsable SCOP codes/identifiers that remain unchanged across SCOP updates, were generated. The first dataset consists of SCOP identifiers for associated structures of the 110 Pfam families (Version 7.2). The SCOP identifiers of subsequently solved structures (observed structures), of the same 110 Pfam families, constitute the second dataset. These SCOP codes present in the two datasets were compared for verifying the validity of the structure associations.

	**Figure 1:** Overview of methodology used in the retrospective assessment of structure associations using PALI profiles.

The PALI family has information about SCOP superfamily and fold as well. This provides us an opportunity to assess the correctness of association at all of these levels. If all members of a Pfam family show association to only one PALI family profile, the structural relation was correct at the level of family. In case, the members of a Pfam family were related to more than one PALI family profile classified under same superfamily or same fold the association was deemed to be correct to that superfamily or fold, respectively.

Comparison of performance of pure sequence profiles and sequence profiles derived from structure-based alignments

We have also addressed the question: "Are there any genuine improvements in the identification of distant relationship when structure-based sequence profiles (PALI profiles) are used compared to sequence profiles obtained from sequence-based alignments?" For this purpose RPS-BLAST searchable profiles or PSSMs were generated using PSI-BLAST [Altschul *et al.*, 1997] for every Pfam family. As mentioned already there are 110 Pfam families for which families of known structure (PALI families) have been associated. Every sequence from each of 110 Pfam families was searched against Pfam family sequence profiles, using RPS-BLAST, and an e-value criteria of 3×10^{-5} was used to extract valid hits. We have specifically looked for valid hits with Pfam families corresponding to 110 PALI families, which were associated with 110 Pfam families of apparently no structural information. This exercise unearthed the relative performance of profiles obtained from sequence-based and structure-based alignments.

Results and discussion

The approach to assign homologous families of known structure uses Reverse PSI-BLAST (RPS-BLAST) searches on a database of profiles of known 3-D structures. Sensitive profiles have been generated using structure-based amino acid sequence alignments integrated with homologous sequences. The domain boundary definitions used are same as defined in SCOP. Since these domain boundaries are derived based on 3-D structure they are precise and robust. In the instance of sequence identities below about 30%, the structure-based sequence alignments are more accurate compared to sequence-based alignments. Hence, we have incorporated structure-based sequence alignments in profile generation, since the average sequence identity of about 60% of PALI families is below 30% [Balaji and Srinivasan, 2001]. These profiles pertaining to structural families have been searched using RPS-BLAST with an e-value

threshold of 0.00003 for extracting reliable associations. This e-value threshold is based on a similar profile-based procedure of IMPALA and also on the benchmarking with RPS-BLAST (B. Anand and N. Srinivasan unpublished). The use of such a stringent e-value might fail to recognize a few true positives. However, the number of false positives would also be minimized considerably. This is apparent from the assessment discussed subsequently.

Overall accuracy of remote homology recognition

In the present assessment, we identified a total of 110 Pfam families for which structures could be associated using PALI profiles one year ago and experimental structures were solved subsequently. Using PALI profiles, we were able to assign correct folds for 106 (96.4%) of 110 Pfam families. Table 1 presents an overview of the structure association for these 106 Pfam families, along with the hierarchy at which the structure was correctly associated.

Table 1: List of Pfam families for which correct fold assignment. The structural hierarchical level (fold/ superfamily/ family) at which structure was correctly recognized for every Pfam family, when compared to experimentally solved structures, is also shown. The PALI family is represented with reference PDB [Berman *et al.*, 2002] structure, wherein the first four letters corresponds to PDB code. The query coverage in the profile alignment is also listed for each relationship. All the associations have been made at an e-value of 3x10-5 or better.

	Pfam Family	PALI Family	Pfam coverage (In%)	PALI coverage (In %)	Prediction upto Fold/ Superfamily/ Family
1	DHBP_synthase	1g57a_o542	100	93.66	Family
2	DNA_primase_S	1g71a_o740	100	87.21	Family
3	GNT-I	1fo8a_o486	78.75	99.39	Superfamily
4	Metalloenzyme	1eqja_o501	100	43.43	Family
5	GidB	1dl5a_o474	95.58	54.26	Superfamily
6	AICARFT_IMPCHas	1g8ma_o314	80.86	66.84	Family
7	TrkA-N	1id1a_o356	100	78.43	Superfamily
8	Ribosomal_L3	1jj2b_o249	100	79.53	Family

9	Ribosomal_L5	1iq4a_737	100	31.84	Family
10	Peptidase_M1	1hs6a_o223	51.03	94.71	Family
11	Peptidase_M3	1i1ip_o733	100	67.82	Family
12	DPPIV_N_term	1crza_o269	31.35	53.99	Family
13	NtA	1jb3a_o239	94.07	100	Family
14	Fe-ADH#	1dqsa_o747	99.5	92.13	Superfamily
15	TonB	1ihra_o722	85.71	41.1	Family
16	ParA	1ihua1_453	100	38.28	Family
17	NNMT_PNMT_TEMT	1vid_o468	72.2	71.36	Superfamily
18	UDPGP	1h7ea_o480	57.71	74.29	Superfamily
19	Lipase_2	1c4xa_o487	94.8	80.07	Superfamily
20	Peptidase_U3	1c8ba_o453	100	95	Family
21	Ribosomal_L18p	1jj2m_491	100	65.05	Family
22	Ribosomal_L19e	1jj2o_o179	96.62	100	Family
23	Peptidase_S51	1fyea_o373	100	87.27	Family
24	Methyltransf_2	1fp2a_512	95.4	89.34	Superfamily
25	Bgal_small_C	1dp0a_o220	100	31.4	Family
26	CoaE	1e6ca_o398	99.44	91.18	Superfamily
27	Bgal_small_N	1dp0a_o220	100	63.14	Family
28	APH	1j7la_o574	100	97.72	Family
29	TRNA-synt_1e#	1qu2a1_426	70.88	90.67	Family
30	TRNA-synt_1f	1qu2a1_426	61.78	91.78	Family

31	DUF108	1b7go1_407	44.87	57.87	Family
32	HSP33	1hw7a_o564	83.39	99.56	Family
33	GSPII_E#	1g6oa_456	98.24	86.44	Family
34	UPP_synthetase	1f75a_447	97.38	94.01	Family
35	Neur_chan_LBD	1i9ba_o199	99.52	98.05	Family
36	E1#	1f08a_o725	32.18	93.24	Family
37	Binary_toxA	1g24a_645	100	59.72	Family
38	Lum_binding	1i8da_o247	100	91.4	Family
39	TauD#	1ds1a_o293	96.01	85.45	Superfamily
40	Fz	1ijxa_055	93.55	92.8	Family
41	Ribosomal_L13	1jj2i_o357	100	79.58	Family
42	BAG	1i6za_180	100	60	Family
43	Ribosomal_L23	1jj2r_o547	96.39	98.77	Family
44	Ribosomal_L29	1jj2u_o072	95.31	93.85	Family
45	PUF	1ib2a_o033	100	10.87	Family
46	Ribosomal_L39	1jj21_o059	100	86.96	Family
47	Ribosomal_L44	1jj22_o804	100	83.7	Family
48	IMS	1im4a_o754	66.5	95.61	Family
49	Ribosomal_L24e	1jj2t_o793	80.3	100	Family
50	SIR2	1j8fa_442	100	59.29	Family
51	LuxS	1j6wa_729	99.38	99.35	Family
52	Hpr_kinase	1jb1a_o310	55.19	98.73	Family
53	Reovirus_cap	1fn9a_o620	100	100	Family
54	YjeF_N	1jzta_o496	100	75	Family

55	Adaptin_N	1b3ua_030	77.02	64.8	Superfamily
56	Thioesterase	1c4xa_o487	98.68	88.97	Superfamily
57	IP_trans	1fvza_o556	100	94.42	Family
58	Ribosomal_L5_C	1iq4a_737	100	53.07	Family
59	Glyco_hydro_25	1jfxa_o352	100	52.53	Family
60	MerR	1exja_o166	100	31.36	Family
61	RNA_pol_L	1i50k_o709	100	81.58	Family
62	CMAS	1dl5a_o474	98.84	55.52	Superfamily
63	DeoC#	1rpxa_o338	96.55	94.35	Fold
64	DNK#	1e2ka_451	100	45.78	Superfamily
65	Tropomodulin	1a4ya_371	44.13	30	Superfamily
66	Ribosomal_L31e	1jj2w_o632	90.11	100	Family
67	Ribosomal_L32e	1jj2x_o511	100	76.06	Family
68	MoaE	1fm0e_o648	100	79.58	Family
69	Pyridoxal_deC	1c4ka_o476	88.8	70.13	Superfamily
70	IspD	1i52a_o479	100	98.67	Superfamily
71	Ribosomal_L37e	1jj2z_o803	100	98.21	Family
72	SurE	1j9la_o505	100	75.71	Family
73	Rota_Capsid_VP6	1qhda_o024	100	99.53	Family
74	NAD_Gly3P_dh#	1evya_o010	47.53	90.96	Family
75	Flagellin_C	1io1a_o751	69.44	12.66	Family
76	Flagellin_N	1io1a_o75	80.14	28.61	Family

		1			
77	F420_oxidored	1qmga_410	47.39	45.58	Family
78	Transmembrane4	1g8qa_o057	40.45	98.89	Family
79	DHHA1	1i74a_o308	87.3	16.78	Superfamily
80	DHHA2	1i74a_o308	100	41.12	Family
81	Glyco_transf_20#	1f6da_o517	83.94	95.36	Superfamily
82	Pantoate_ligase	1ihoa_o384	100	99.29	Family
83	V-ATPase_H	1ho8a_o034	100	90.38	Family
84	Sua5_yciO_yrdC	1hrua_o541	98.86	93.48	Family
85	MMR_HSR1#	1f5na_460	57.2	55.93	Family
86	Spermine_synth	1dusa_o472	75.95	97.4	Superfamily
87	ATP-sulfurylase	1g8fa_428	68.11	100	Family
88	DNA_RNApol_7kD	1i50l_o805	100	69.57	Family
89	PABP	1g9la_130	100	50	Family
90	DEP	1fsha_o129	100	79.79	Family
91	DHH	1i74a_o308	100	48.03	Family
92	PdxJ	1ho1a_o343	100	97.87	Family
93	Dockerin_1	1dava_o091	100	30.98	Family
94	Methyltransf_5	1dl5a_o474	31.61	29.65	Family
95	Peptidase_S15	1jkma_528	57.06	74.42	Family
96	Ku	1jeya_o741	100	42.19	Family
97	Ku_C	1jeyb_o742	72.45	13.4	Family
98	Ku_N	1jeya_o74	100	43	Family

		1			
99	Shikimate_DH#	1dxy-_408	52.61	60.3	Superfamily
100	Glyco_transf_8	1ga8a_o483	99.61	91.14	Superfamily
101	RNA_POL_M_15KD	1i50i2_831	*	*	Fold
102	GTP1_OBG#	1f5na_460	*	*	Family
103	SecA_protein	1heia_o397	*	*	Superfamily
104	IL3	1eera_078	*	*	Family
105	IL12	1f6fa_077	*	*	Family
106	Pectate_lyase	1bn8a_336	*	*	Family

* There is no direct alignment present for these Pfam and PALI families since the structure association is made through one or more intermediate Pfam families. Hence, the query and profile coverage cannot be determined.

Pfam families for which structure prediction could be made also by using purely sequence profiles.

We analyzed the correctness of the association at the level of fold, superfamily and family for the above Pfam families. We could assign the structure correctly to the level of family for 81 out of 106 Pfam families (Table 1). For 23 Pfam families, the structural relation was shown to be correct only at superfamily level. There are two Pfam families for which we could not associate a definite superfamily but these were assigned correct folds.

The associations, which are correct to the level of families, indicate these proteins might share good sequence identities in addition to significant structural similarities. We further explored the reasons for being unable to identify the exact family. Interestingly, 14 out of 23 Pfam families for which structure associations could be made correctly at the superfamily level belonged to new structural family in the same superfamily. In other words, the structural family concerned was not existent at the time of recognizing the remote relationship. Hence, in these cases structure association was possible only by matching it to the closest possible PALI family profile in the same superfamily. For remaining 9 cases, although the Pfam family shares gross sequence similarity to the associated SCOP family, it belonged to the observed SCOP family by virtue of a greater resemblance in the functional site residues between Pfam family and observed SCOP family.

It is important to note that some of the structure associations of the Pfam families were established indirectly through intermediately related Pfam families which, in turn, were related to a PALI family. For example, GTP1/OBG could be related to Ferrous iron transport protein B Pfam family that, in turn, is related to the family of G-proteins in PALI. So we could not predict, directly, the

relationship between GTP1/OBG family in Pfam and the G-protein family in PALI. However, its relationship to G-protein PALI family was established only through Ferrous iron transport protein B Pfam family.

The query and profile coverage for Pfam and PALI, in the light of structure association, were assessed. The percentage coverage for best pairwise sequence alignment between query and profile is shown in Table 1. Since the structural domains are known to have well-defined domain boundaries in comparison to sequence-based domains, the coverage could aid in improved domain delineation for sequence-based families. In a number of instances the Pfam query sequences were covered completely (~100% coverage) in the alignments. In such cases, there is a possibility that the Pfam family domain definition can be extended so as to encompass the structural domain definitions. On the other hand, there are a few instances where the entire length of a PALI family is aligned within a much longer Pfam family. For example, the best query coverage for Pfam family E1 is ~31% whereas the PALI profile is covered about 93% suggesting there could be more than one folding unit being encompassed by the Pfam domain. Interestingly, in later Pfam updates this Pfam family E1 is split into 2 Pfam families PPV_E1_N and PPV_E1_C domains. Since structural domains are independent folding units, such precise domain boundaries would contribute to a more robust Pfam family definition.

The present result that the use of structure-based alignment improves remote homologue detection apparently differs from one of the previous work [Griffiths-Jones and Bateman, 2002], wherein the use of structure-based alignment did not show improvement in homologue detection, using HMM profiles. The disparity in the two conclusions is likely due to a few important differences in both the analyses: (i) Our dataset predominantly consists of homologous structures of low sequence identity [Balaji and Srinivasan, 2001]. Most of the pairwise sequence identities are less than 30%. In such cases, when sequence identity is low structure-based alignments are better compared to multiple sequence alignment. (ii) We have generated integrated structure-sequence alignment by incorporating homologous sequences either from Pfam-A seed alignments, which are characterized by a diverse set of protein sequences, or NRDB. Hence, profiles generated considering structure-based sequence alignment enriched with homologous sequence would be more effective in remote homologue detection.

False-positives

We were unable to identify correct folds for 4/110 (3.6%) Pfam families. All these four Pfam families were indirectly related to a PALI family, through an intermediate Pfam family. The four cases were further explored in-depth. The region of alignment for each of the Pfam families Herpes_glycop_D and Reo_sigma1, with their respective profiles, was very short (<50 amino acids). For the remaining two Pfam families, viz. Clat_adaptor_s and Sigma54_activat, the erroneous association was made due to short similar regions of alignment, sharing common secondary structures, that do not associate appropriately in 3-D space. However, applying a query or profile coverage criterion in addition to the consideration of an e-value of 3×10^{-5}, could help in eliminating the incorrect

associations.

Examples of Family association

The structure of Clostridial binary toxin A (Binary_toxA) Pfam family was remotely related to the ADP-ribosylating toxins structural family. The relationship between these two families was established with an e-value of 7 x 10^{-25}. The ADP-ribosylating toxin family is involved in transfer of ADP-ribose group of NAD. The ADP-ribosylation of regulatory proteins is an important pathological mechanism by which various bacterial toxins affect the eukaryotic cell function. The Binary_toxA Pfam family consists of a group of bacterial exotoxins from the Clostridium species of Gram-positive, spore-forming, anaerobic, bacilli present in soil. This family includes neurotoxin from *Clostridium botulinum*, which affects the nervous system of the host. *Clostridium perfringens* Iota toxin also a member of Binary_toxA is an ADP-ribosylating toxin. It is a binary toxin composed of enzymatic component (Ia), which is concerned with ADP-ribosylation and a binding component (Ib) playing a role in membrane-transport [Tsuge *et al.*, 2003]. Iota-toxin catalyses the transfers of ADP-ribose group of NAD to a target protein with nicotinamide release. The prediction that the structure of Binary_toxA is similar to ADP-ribosylating toxin family is consolidated by their similar functions.

The structure for Pfam family GTPase of unknown function (MMR_HSR1) was not elucidated when we associated the same with a structural family. We were able to associate the structure for this family by virtue of its homology with human guanylate binding protein, which belongs to the same structural family of G-proteins, with an e-value of 3 x 10^{-09}. GTPases are group proteins that can switch conformation with GTP or GDP bound to it and bring about downstream effect. These play key regulatory roles in signaling, translation and protein targeting [Wittinghofer and Pai, 1991; Sprang, 1997; Vetter and Wittinghofer, 2001]. The alignment between a member of GTPase of unknown function family and human guanylate binding protein shows the required GTP binding motifs G-1 (GXXXXGK[T/S]), G-3 (DXXG) as conserved (Figure 2).

	Figure 2: Alignment, showing conserved GTP binding motifs (G-1, G-3), between a GTPase of unknown function Pfam family (MMR_HSR1) member and a GTP binding protein of known structure. (ENGA_CHLTR-Probable GTP-binding protein engA from *Chlamydia trachomatis*, 1f5na_460-Guanylate binding protein, a member G protein structural family).

Examples of Superfamily association

Potential common evolutionary relationship is suggested by similarities in structural and functional features among various families. This suggests a common evolutionary origin for them and hence they are classified under a particular superfamily.

The members of Dephospho-Coenzyme A kinase (CoaE) Pfam family catalyze the last step in Coenzyme A biosynthesis from pantothenate, involving phosphorylation of 3'-hydoxyl group of dephosphocoenzyme A to form Coenzyme A, utilizing ATP as the phosphate donor [Obmolova *et al.*, 2001]. Coenzyme A plays an essential role in cellular metabolism. Structural association linked CoaE family with that of Shikimate kinase with an e-value at $2x10^{-08}$. Both, CoaE and Shikimate kinase belong to the superfamily of P-loop containing nucleotide triphosphate hydrolases. They both probably share similar ATP binding sites as suggested by their functionality though they do not have the same biochemical function.

Iron-containing alcohol dehydrogenase (Fe-ADH) family proteins catalyze the reversible oxidation of ethanol to acetaldehyde while concomitantly reducing NAD to NADH. Fe-ADH family consists of zinc-containing long-chain alcohol dehydrogenases, short-chain alcohol dehydrogenases and iron-containing alcohol dehydrogenases. The prediction that Fe-ADH family is a member of the Dehydroquinate synthase-like superfamily was made with e-value of 8×10^{-64}. When the crystal structure of a member, glycerol dehydrogenase belonging to Fe-ADH family, was determined it revealed a similarity to dehydroquinate synthase supplying a striking example of divergent evolution [Ruzheinikov *et al.*, 2001]. The alignment between one of Fe-ADH members and a member of dehyrdoquinate synthase family exhibits conservation of residues involved in NAD and metal binding (Figure 3).

Figure 3: Alignment between Fe-ADH Pfam family member and dehydroquinate synthase protein of known structure. The conserved NAD binding residues are shown in red where metal binding residues are in blue. (ADH1_CLOSA- NADPH-dependent butanol dehydrogenase from *Clostridium saccharobutylicum*, 1dqsa_o747- Dehydroquinate synthase protein of Dehydroquinate synthase structural family).

Example of fold association

The family of deoxyribose-phosphate aldolase (DeoC) is involved in nucleotide metabolism. One of the members in this family from *E. coli*, the 2-Deoxyribose-5-

phosphate aldolase, catalyses the reversible aldol reaction between acetaldehyde and D-glyceraldehyde-3-phosphate to generate D-2-deoxyribose-5-phosphate. It is unique among the aldolases as it catalyses the reversible condensation of two aldehydes [DeSantis *et al.*, 2003]. This particular member has applications in synthesis of the antitumor agent Epothilone A [Liu and Wong, 2002]. DeoC family adopts the TIM beta/alpha-barrel fold. DeoC is classified in Class I aldolase family in Aldolase superfamily. The structure of this family was assigned correctly to the fold with an e-value of 2×10^{-13}. The structure, with which it aligned so as to allow fold recognition, is D-ribulose-5-phosphate 3-epimerase, which belongs to TIM beta/alpha-barrel fold.

Specificity and sensitivity of the approach

We have evaluated this approach using standard measures of sensitivity (or coverage) and selectivity (or reliability). The evaluated set consists of 284 protein families for which the structures that were solved during the year. Using our approach we predicted correct structures for 106/284 (sensitivity) families. The predictions were correct in 106/110 cases (selectivity). This suggests that this approach might not detect all the relationships. However, the positive aspect is that whenever homologue detection is made, it is with a high accuracy.

Further on, we assessed the instances of 174 Pfam families for which structures were unknown one year ago and we could not associate any structure at that time. Of these, 51 Pfam families are membranous, short-length or coiled-coil proteins. Some of the other false-negatives can also be accounted by the fact that PALI database does not include the profiles for the following SCOP classes: membrane and cell surface proteins and peptides, peptides, coiled coil proteins, low resolution protein structures and designed proteins. Furthermore, PALI profiles are generated using a reference sequence as the input for PSI-BLAST and so this reference sequence plays a significant role in determining the characteristic of the profile [Aravind and Koonin, 1999; Koonin *et al.*, 2000]. Thus, whether homologues are detected or not is determined, at least partly, by the choice of initial reference sequence that is used as an input to generate profiles. Hence, the reason for a few false negatives could be the feature of reference sequence that is used to generate PALI profiles. The inability to establish structural associations, for some Pfam families, at reliable e-values (3×10^{-5}) is an essential reason for discarding them as invalid relationships.

Comparison of performances of sequence profiles and structure-based sequence profiles

In order to assess the enhancement in remote homology detection using structural information we have compared the remote homology detection using pure sequence profiles by considering Pfam family sequences with structure-based sequences profiles as in PALI. Using sequences profiles we could associate 15 of 110 Pfam families to another Pfam family, which has a known 3-D structure or could be related to a family of known structure. Further, when the association of correct structure is assessed, we could correctly predict structure for 12 of out of 15 Pfam families. These correct associations are indicated by "#" in Table

1.

Thus, compared to association of 110 Pfam families with remotely related PALI families, only 15 of these Pfam families could be associated with a distantly related Pfam family. Hence, there is a substantial difference in the performance of sequence-based and structure-based profiles in terms of number of distant relationships detected. Further, 12 out of 15 distant relationships among Pfam families turned-out to be correct as compared to 106 correctly detected distant relationships out of 110 relationships made between Pfam and PALI families. This appreciable enhancement in the performance of structure association could be because of a) integration of distant sequence homologues in the profile generation b) accurate sequence alignment generated using 3-D structural information and c) more robust and shorter domain definition as indicated by structures compared to sequence alignments. Due to shorter domain length of PALI profiles, the similarity measure for the sequence versus profile match involving both query and profile from PALI is better than the similarity measure for the match between Pfam query and Pfam profile. The basic reason for this feature is the challenge for the alignment program to obtain good sub-optimal alignment if the lengths of the query and profile are substantially different.

In a related work Tang *et al.*, 2003, have shown that use of structural information is effective in improving remote homology detection. They have suggested that the use of secondary structure information is more valuable than the use of entire 3-D information and also that sequence-based profiles, used transitively based on structural relationships, could perform better than profiles directly incorporating multiple structure alignments.

Recommendations for using sequence profiles of structural families

The carefully generated PALI profiles when used along with sensitive profile-based search method, RPS-BLAST, can effectively detect distantly related structural homologues. The statistical significance of association, represented in form of e-value, should be considered as the initial criteria for structural associations. We recommend use of an e-value cut-off of 3×10^{-5} for reliable homologue detection if the database comprises of PALI PSSMs or like. In the absence of a significant e-value, one might assess the remote relationships based on individual discretion.

It is also recommended that query sequences correspond to domains, instead of full length multi-domain gene products. Another important suggestion is that the lengths of the domain predicted/identified from a multi-domain protein should be minimal and as short as possible. It should not be too short (under about 75 residues) as it is likely to result in high occurrence of false positives.

The profile-based search procedure being sequence dependent, one could use many homologous sequences as queries in order to establish relationship with a structural family. Furthermore, one can explore intermediate relationship with another sequence for possible structure association. Application of the above criteria in structural assignments for gene products encoded in the genome of *Mycoplasma genitalium* resulted in structure association for ~60 % of the gene

products.

Public availability

PALI profile-based homology detection can be carried out at a publicly accessible (http://pauling.mbu.iisc.ernet.in/~pali) site. A query sequence is searched in the PALI family profiles database using sensitive profile-based search procedure of RPS-BLAST (http://pauling.mbu.iisc.ernet.in/~pali/rpsblast_pali.html). Alternatively, it is also possible to download the PALI profiles (http://pauling.mbu.iisc.ernet.in/~osupfam/download.html), for various structural families. Searching these PALI profiles by using RPS-BLAST, available freely from NCBI, should enable genome-wide fold recognition that is feasible at the user's local site.

Conclusions

The use of sequence profiles arrived at on the basis of 3-D structural similarity and sensitive profile-based search methods has enabled us to detect correctly many of the non-trivial relationships between sequence (Pfam) and structural (PALI) families and has aided in structural association. This resulted in correct fold recognition for 96.4% of the cases. The present work confirms the correct structure association for 106 out of 110 Pfam families that were analyzed. Descending further into the hierarchy of structure classification, 104/106 predictions were correct to the level of the superfamily and 81/106 were correct right up to the specific family. Thus, it is established from the retrospective analysis that using sequence profiles of structural families would be an important source of structure association.

The higher optimal performance of PALI profiles is achieved due to (a) Accurate alignment of distantly related homologues and (b) Longer domain size in Pfam compared to corresponding PALI domains. The well-defined, usually shorter, domain boundaries as suggested by 3-D structures along with reliable alignment of homologues impart distinct advantages for use of these profiles in remote similarity detection. PALI profiles and the profile-based search programs are freely available at publicly accessible sites. In addition, the PALI profiles can be obtained from http://pauling.mbu.iisc.ernet.in/~osupfam/download.html site and could be used locally in conjunction with the freely available versions of RPS-BLAST or IMPALA available for download from NCBI. This makes the present approach as one of the very few methods, which could facilitate genome-wide structure assignment in a fast, efficient and reliable manner at user's local site.

Acknowledgements

S. N. is supported by computational genomics project sponsored by Department of Biotechnology, New Delhi. S. B. P. is supported by the Council of Scientific and Industrial Research (CSIR), New Delhi. This research is supported by the award of International Senior Fellowship in Biomedical Sciences to N. S from the Wellcome Trust, London, computational genomics project sponsored by the Department of Biotechnology, New Delhi and the NMITLI project supported by CSIR, New Delhi. We thank the anonymous referee for valuable comments.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

- Altschul, S. F., Madden, T. L., Schäffer, 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.

- Aravind, L. and Koonin, E. V. (1999). Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J. Mol. Biol. 287, 1023-1040.

- Balaji, S. and Srinivasan, N. (2001). Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. Protein Eng. 14, 219-226.

- Balaji, S., Sujatha, S., Kumar, S. S. C. and Srinivasan, N. (2001). PALI-a database of Phylogeny and ALIgnment of homologous protein structures. Nucleic Acids Res. 29, 61-65.

- 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. L. (2002). The Pfam Protein Families Database. Nucleic Acids Res. 30, 276-280.

- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. and Zardecki, C. (2002). The Protein Data Bank. Acta Crystallogr. D Biol. Crystallogr. 58, 899-907.

- 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.

- Brenner, S. E. (2000). Target selection for structural genomics. Nat. Struct. Biol. 7 Suppl., 967-969.

- Buchan, D. W., Shepherd, A. J., Lee, D., Pearl, F. M., Rison, S. C., Thornton, J. M. and Orengo, C. A. (2002). Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. Genome Res. 12, 503-514.

- Bujnicki, J. M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001). LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci. 10, 352-361.

- Chothia, C. and Gerstein, M. (1997). Protein evolution. How far can sequences diverge? Nature 385, 579-581.

- DeSantis, G., Liu, J., Clark, D. P., Heine, A., Wilson, I. A. and Wong, C. H. (2003). Structure-Based Mutagenesis Approaches Toward Expanding the Substrate Specificity of D-2-Deoxyribose-5-phosphate Aldolase. Bioorg. Med. Chem. 11, 43-52.

- Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics 14, 755-763.

- Eyrich, V. A., MartÍ-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2001). EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17, 1242-1243.

- Fischer, D. (2003). 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 51, 434-441.

- Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D.,

Karplus, K. J., Kelley, L. A., MacCallum, R. M., Pawowski, K., Rost, B., Rychlewski, L. and Sternberg, M. (1999). CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins Suppl. 3, 209-217.

- Ginalski, K. and Rychlewski, L. (2003). Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res. 31, 3291-3292.

- Ginalski, K., Elofsson, A., Fischer, D. and Rychlewski L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-1018.

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

- Gowri, V. S., Pandit, S. B., Karthik, P. S., Srinivasan, N. and Balaji, S. (2003). Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. Nucleic Acids Res. 31, 486-488.

- Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. USA 84, 4355-4358.

- Griffiths-Jones, S. and Bateman, A. (2002). The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. Bioinformatics. 18, 1243-1249.

- Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233, 123-138.

- Holm, L. and Sander, C. (1994). The FSSP database of structurally aligned protein fold families. Nucleic Acids Res. 22, 3600-3609.

- Holm, L. and Sander, C. (1996). Mapping the protein universe. Science 273, 595-603.

- Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797-815.

- Karplus, K., Barrett, C. and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846-856.

- 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.

- Koonin, E. V., Wolf, Y. I. and Aravind, L. (2000). Protein fold recognition using sequence profiles and its application in structural genomics. Adv. Protein Chem. 54, 245-275.

- Liu, J. and Wong, C. H. (2002). Aldolase-catalyzed asymmetric synthesis of novel pyranose synthons as a new entry to heterocycles and epothilones. Angew. Chem. Int. Ed. 41, 1404-1407.

- Mallick, P., Weiss, R. and Eisenberg, D. (2002). The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds. Proc. Natl. Acad. Sci. USA 99, 16041-16046.

- Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins Suppl 5, 2-7.

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

- Obmolova, G., Teplyakov, A., Bonander, N., Eisenstein, E., Howard, A. J. and Gilliland, G. L. (2001). Crystal Structure of dephospho-Coenzyme A kinase from Haemophilus influenzae. J. Struct. Biol. 136, 119-125.

- 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.

- 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.

- Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444-2448.

- Rost, B., Schneider, R. and Sander, C. (1997). Protein fold recognition by prediction-based threading. J. Mol. Biol. 270, 471-480.

- Russell, R. B., Copley, R. R. and Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. J. Mol. Biol. 259, 349-365.

- Ruzheinikov, S. N., Burke, J., Sedelnikova, S., Baker, P. J., Taylor, R., Bullough, P. A., Muir, N. M., Gore, M. G. and Rice, D. W. (2001). Glycerol dehydrogenase: structure specificity and mechanism of a family III polyol dehydrogenase. Structure 9, 789-802.

- Saunders, J. A., Gibson, K. D. and Scheraga, H. A. (2002). *Ab initio* folding of multiple-chain proteins. Pac. Symp. Biocomput. 7, 601-612.

- Schäffer, 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 12, 1000-1011.

- Shi, J., Blundell, T. L. and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. 310, 243-257.

- Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213, 859-883.

- Sonnhammer, E. L. L., Eddy, S. R. and Durbin, R. (1997). Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments. Proteins 28, 405-420.

- Sprang, S. R. (1997). G protein mechanisms insights from structural analysis. Annu. Rev. Biochem. 66, 639-678.

- Tang, C. L., Xie, L., Koh, I. Y., Posy, S., Alexov, E. and Honig, B. (2003). On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J. Mol. Biol. 334, 1043-1062.

- Teichmann, S. A., Chothia, C. and Gerstein, M. (1999). Advances in structural genomics. Curr. Opin. Struct. Biol. 9, 390-399.

- Tsuge, H., Nagahama, M., Nishimura, H., Hisatsune, J., Sakaguchi, Y., Itogawa, Y., Katunuma, N. and Sakurai, J. (2003). Crystal Structure and Site-directed Mutagenesis of Enzymatic Components from Clostridium perfringens Iota-toxin. J. Mol. Biol. 325, 471-483.

- Vetter, I. R. and Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. Science 294, 1299-1304.

- von Grotthuss, M., Pas, J., Wyrwicz, L., Ginalski, K. and Rychlewski, L. (2003). Application of 3D-Jury, GRDB, and Verify3D in fold recognition. Proteins Suppl. 6, 418-423.

- Wittinghofer, A. and Pai, E. F. (1991). The structure of Ras protein a model for a universal molecular switch. Trends Biochem. Sci. 16, 382-387.