Regions of minimal structural variation among members of protein domain superfamilies: application to remote homology detection and modelling using distant relationships ☆

Saikat Chakrabarti, R. Sowdhamini*

National Centre for Biological Sciences, UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India

Received 7 April 2004; accepted 13 May 2004

Available online 28 May 2004

Edited by Robert B. Russell

Abstract Structurally conserved regions or structural templates have been identified and examined for features such as amino acid content, solvent accessibility, secondary structures, non-polar interaction, residue packing and extent of structural deviations in 179 aligned members of superfamilies involving 1208 pairs of protein domains. An analysis of these structural features shows that the retention of secondary structural conservation and similar hydrogen bonding pattern within the templates is 2.5 and 1.8 times higher, respectively, than full-length alignments suggesting that they form the minimum structural requirement of a superfamily. The identification and availability of structural templates find value in different areas of protein structure prediction and modelling such as in sensitive sequence searches, accurate sequence alignment and three-dimensional modelling on the basis of distant relationships.

Keywords: Structural motifs; Structural invariants; Evolutionary relationship; Sequence searches; Structure prediction

1. Introduction

Three-dimensional folding patterns are shared among different proteins and several of them can be classified into groups characterized by a common evolutionary origin [1–3]. Non-homologous proteins might share folds due to the limitation in the number of protein folds in nature [4–6]. Knowledge of these relationships provides important contribution to the prediction of fold and function of new proteins. Several databases [7–11] exist that classify protein entries [12,13] into structural classes, folds, superfamilies and families of domains at different levels of structural hierarchy. Protein domains, that share high structural similarity and biological function but share poor sequence identity due to evolutionary divergence, are grouped together at the superfamily level.

Concentrated efforts have been made in the past in studying the structural invariants of specific families or superfamilies by analysing the structures of members: these include folds like neurotoxin-agglutinin folds [14], Ig domains [15], β-barrels [16], α/β barrels [17], TIM barrels [18], β-trefoils [19], jellyrolls [20] and α/β hydrolases [21]. Results derived from such analyses have pointed to particular residues important for function and a small number of structural elements that are required for retention of fold and function. However, the intriguing question remains in understanding structural determinants that are common to most superfamilies. Studies [22,23] have shown the prevalence of short segmental conservation among similar pairs of proteins.

In this analysis, we have examined structurally aligned representative members from the superfamily alignment databases [8,11]. Several conserved regions of protein domains have been identified for 179 multi-member superfamilies using criteria such as amino acid preference and solvent accessibility. Such conserved segments are termed as ‘structural templates’ and are characteristic of the superfamily with high retention of similarities in other structural features (such as secondary structural content, hydrogen bonding pattern, non-polar interaction and residue packing). Structural templates maintain a particular spatial pattern when compared across different proteins belonging to the same superfamily; together with their interaction pattern and spatial orientation, they can be projected as the bare minimum requirement for maintaining the common core structure of the particular superfamily. We further discuss areas where information on structural templates can be immediately valuable. These patterns can be utilized to identify distant homologues and can be used as additional restraints in improving the quality of sequence alignments and models derived by homology modelling that involves distant relationships.

2. Materials and methods

2.1. Identification and characterization of structural templates

Solvent accessibility was measured using the program PSA [24] and residues that have accessible surface area less than 7% were treated as being buried as used in earlier analysis in proteins [25]. At every alignment position, all possible pairs of superfamily members and their observed amino acids were scored using standard 20 × 20 substitution matrix [26]. 179 multi-member structure-based sequence alignments [27] from CAMPASS [8] and PASS2 databases [11] have been used as initial inputs for the identification of the templates for each superfamily. Structural templates were identified by the presence of at least three consecutive solvent-buried residues that have higher amino acid exchange scores. A structural template was, however, allowed to propagate on both directions until the two primary features (solvent

☆ Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2004.05.028.

* Corresponding author. Fax: +91-80-3636421.
E-mail address: mini@ncbs.res.in (R. Sowdhamini).

0014-5793/$22.00 © 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2004.05.028

accessibility and amino acid exchange) were conserved in at least 60% of residues within the template.

2.2. Conservation of structural features

SSTRUC program, that is part of JOY4.0 suite of programs [28], was used to identify secondary structural positions. The HBOND program, part of JOY4.0 suite, has been used to identify hydrogen bonds. Each non-polar residue (Ala, Val, Leu, Ile, Met, Pro, Tyr, Phe and Trp) in the structural template was examined for the presence of a neighbouring non-polar residue within a sphere of radius 4 Å by examining C^{β}–C^{β} distances. In the case of glycine, a virtual C^{β} atom was considered. Residue packing has been measured in terms of Ooi number [29] that provides the number of residues surrounding each C^{α} atom of residues in a protein. Higher Ooi numbers correspond to high residue packing and suggest that the residue is in a well-packed environment. Both primary features (like solvent accessibility and amino acid preference) and secondary features (like secondary structure, hydrogen bonding, non-polar contacts and residue packing) were considered for characterization of the structural motifs. They have also been utilized in scoring the structural conservation of templates (see Supplementary Materials for details).

2.3. Spatial deviations of structural templates

Structural templates are converted into vector representation using SCHELAX [30,31] and the distances between all possible pairs of templates and virtual torsion angles were calculated using standard vector algebra.

3. Results

3.1. Structural templates

1045 structural templates were identified from 179 multi-member superfamilies consisting of 620 proteins. On an average, structural templates could be assigned to less than a quarter (18%) of the total alignment length positions (see Table in Supplementary Materials for details).

3.2. Conservation of structural features in structural templates

The mapping of structural templates on the superfamily alignments allows further investigation of the conservation of structural features not directly used in the identification of templates. These include features like secondary structure, hydrogen bonding, residue packing and non-polar-residue contacts. Conservation of a structural feature at an alignment position was confirmed when the same structural type was assigned to several members of a superfamily at an equivalent position in the alignment.

Fig. 1 summarizes the conservation of individual structural features across the superfamilies, both for full-length alignments and at the template regions alone (also see Section 2 and Supplementary Materials) for a representative set of superfamilies.

Data obtained from all the 179 superfamilies show that the conservation of secondary structure at the structural templates was nearly twice (2.6 times) compared to full-length proteins. The conservation of residue packing (Ooi number) in equivalent regions within structural templates was about threefold higher (3.46) and the ratio of percentage conservation of hydrogen bonding patterns in structural motifs vs. full-length alignments was 1.78. In contrast, the conservation of other secondary features such as non-polar interactions was nearly indifferent to the position of structural templates (ratio of percentage conservation of interactions of non-polar residues: 1.12).

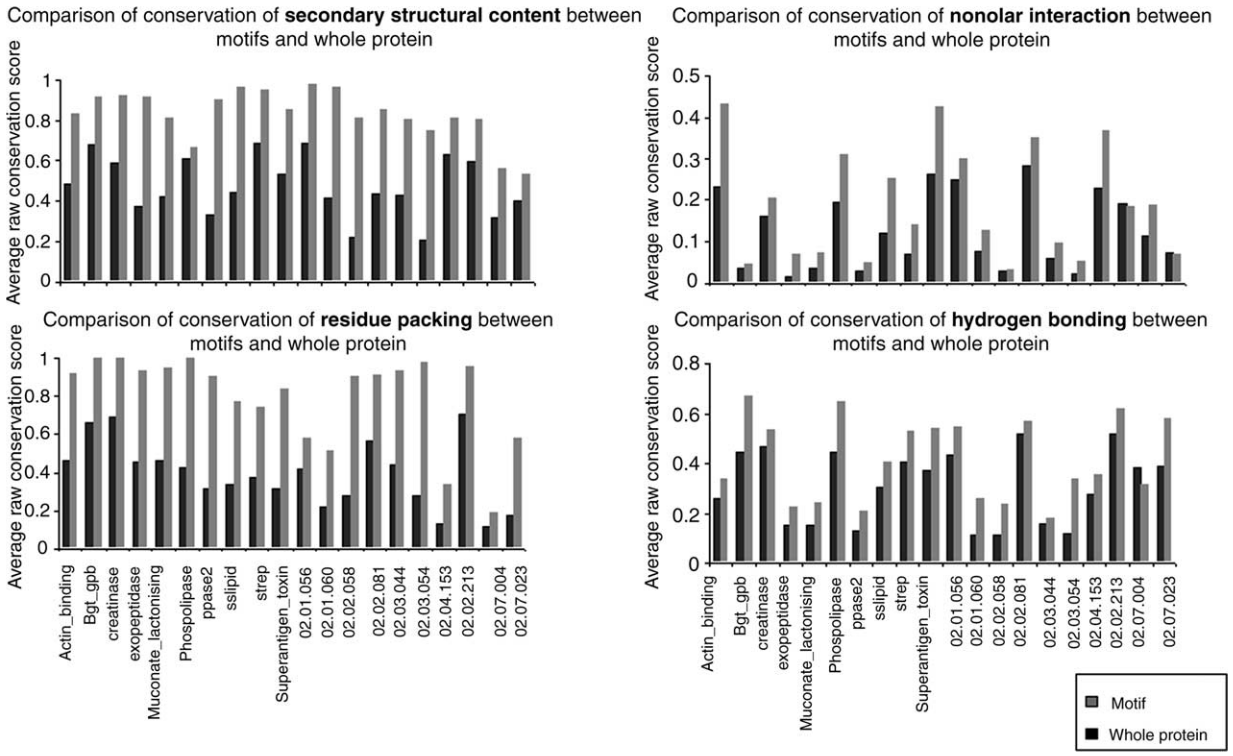

Fig. 1. Comparison of structural criteria between structural templates and whole protein for a representative dataset of 20 protein superfamilies belonging to different structural classes. Content of important structural features (secondary structure, residue packing, non-polar interaction and hydrogen bonding) within the templates is compared with the whole-length protein. Scores are calculated as described in Section 2. Higher content of structural features signifies the importance of the structural templates.

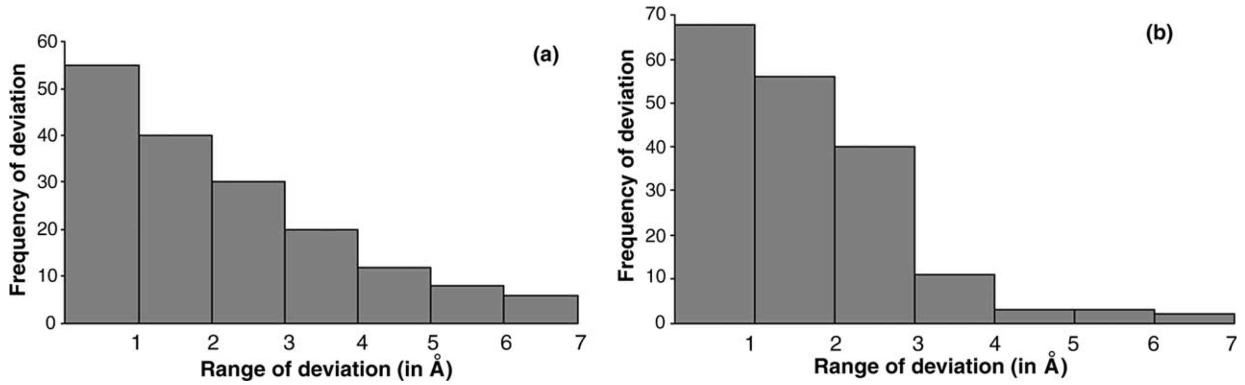

Fig. 2. Mean deviation in spatial distances. The mean spatial distances between the equivalent templates are calculated for all the proteins within 179 superfamilies. Distances are calculated after a vectorial representation of the templates followed by normal vector algebra. Deviation for intertemplate distances (a) are shown to be quite low for superfamily members. Deviation of the average distances of the templates with respect to the centre of mass of the protein (b) also shows similar distribution.

3.3. Analysis of spatial deviations in structural templates

Average deviations in the distances for the equivalent template segments were calculated between all possible pairs of proteins within each superfamily. Distances of each template with respect to the centre of mass of the protein were also calculated and the deviation of these distances between all possible pairs of members of the same superfamily was observed.

Superposed coordinates could not be obtained for two superfamilies due to difficulties in rigid-body superposition. The mean distances between structural templates can vary, but significant numbers were less than 3 Å (Fig. 2(a), 125 out of 177 superfamilies; 69%). As shown in Fig. 2(b), in a vast majority of the cases, the average deviations in distances between a structural template and the centroid of the protein were less than 3 Å (164 out of 177 superfamilies; 91%). Structural parameters analysed in this work do not include positional coordinates. Therefore, structural conservation is not always accompanied by spatial rigidity. Short structural templates and superfamilies with poor sequence identities are much more likely to undergo a spectrum of spatial variations (please see Supplementary Material).

Virtual torsion angles between the templates were calculated and an average value for each template segment for that superfamily was stored. Deviations in the torsion angles for all the superfamilies were calculated (similar to distances). The absolute angle of each template vector with respect to the centre of mass of the protein was calculated and compared with the other superfamily members. The variation in the values of virtual angles between the equivalent structural templates can be high (only 85 out of 177 superfamilies have a mean angle variation within 20° from an average structure; Fig. 3). In some superfamilies, the deviation in virtual angles between three structural templates are as high as 100°. Fig. 4 shows two extreme examples – one superfamily with a rigid core and another superfamily that accommodates vivid and dramatic spatial deviations even at the structural templates.

3.4. Applications of structural templates

Scanning sequence databases using structural templates. Structural templates can be employed to scan in sequence databases and similar sequences potentially belonging to the superfamily can be identified. SCANMOT is a procedure that searches for similar sequences in entire sequence database using conserved regions and inter-motif spacing as sole restraints and attributes significance to the scores. This program is available via http://caps.ncbs.res.in/scanmot/scanmot.html. Structural templates from all multi-member superfamilies,

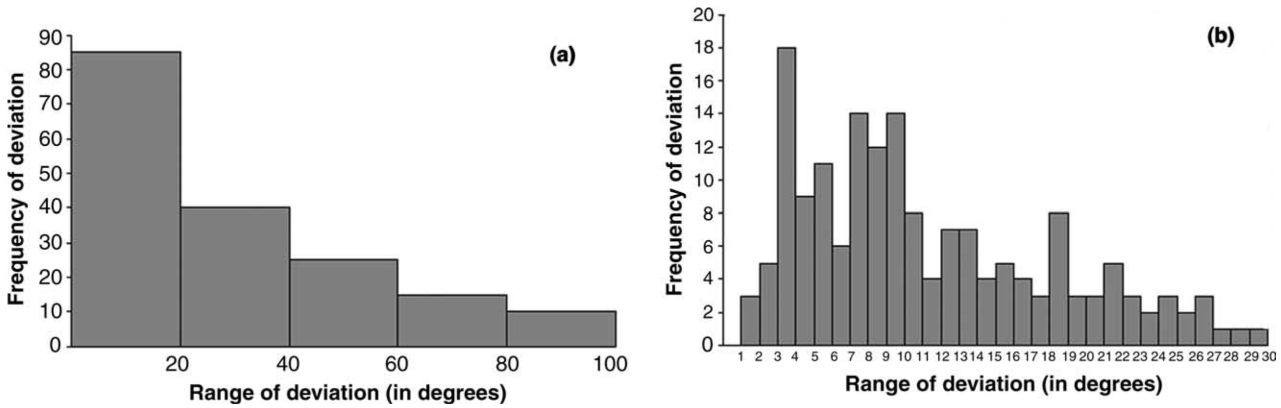

Fig. 3. Mean deviation in torsion and absolute angle. The mean spatial angular patterns between the equivalent templates are calculated for all the proteins within 179 superfamilies. Angles are calculated after vectorial representation of the templates followed by normal vector algebra. Deviation for intertemplate torsion angle (a) are shown to be quite low for superfamily members. Deviation of the absolute angles of the templates with respect to the centre of mass of the protein (b) shows much wider distribution.

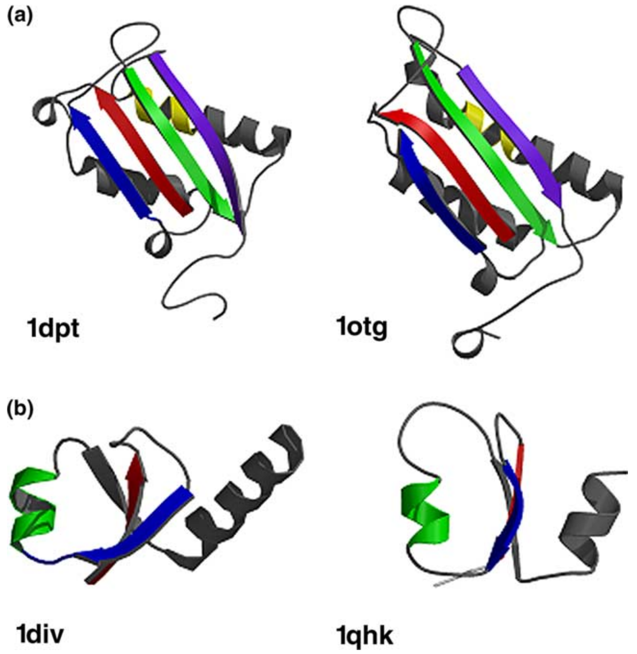

Fig. 4. Deviation in spatial orientation patterns within the structural templates. Structural templates identified for the proteins d-dopachrome tautomerase (PDB code: 1dpt) and 5-carboxymethyl-2-hydroxymuconate isomerase (PDB code: 1otg) belonging to Tautomerase/MIF superfamily show very little deviation whereas templates for ribosomal protein l9 (PDB code: 1div) and ribonuclease hi (PDB code: 1qhk) show marked difference in spatial orientation.

recorded in PASS2 [11] database, have been scanned in a database of protein entries from the structural databank [12,13] using SCANMOT algorithm. This multi-motif procedure does not employ the query sequence except to note the inter-motif spacing. Further, this method is not iterative in nature. But, the average specificity for all the 110 superfamilies in PASS2 database [11] is 86% and the average sensitivity is around 70% (Fig. 5(a)). This method is especially likely to perform well for sparsely represented superfamilies (for results, see Fig. 5(b)). Further, the results of largely populated superfamilies (Fig. 5(c)) indicate that the results are influenced by additional factors such as sequence dispersion. The structural templates encode position-specific permitted amino acid exchanges within representative members at the superfamily level. We find that these exchanges, though not weighted for frequency of occurrence, are more effective in eliminating false positives than internally consulting an amino acid exchange library (data not shown).

Fifteen superfamilies of proteins together with their identified structural motifs were utilized to scan into a curated dataset of hypothetical proteins present in the non-redundant sequence database. Hypothetical proteins that show significant pattern matching to the query sequence of known function were selected as probable candidates for distant relationship with the query sequence. To reconfirm the distant relationship between the identified hypothetical proteins and the query protein of known function, separate PSI-BLAST [32] runs were initiated, using such hypothetical hits as query against a database of SCOP [7] protein domain together with their close homologues as well as a non-redundant sequence database. Around 80% of the selected hypothetical proteins could retrieve proteins belonging to the initial query (Table 1).

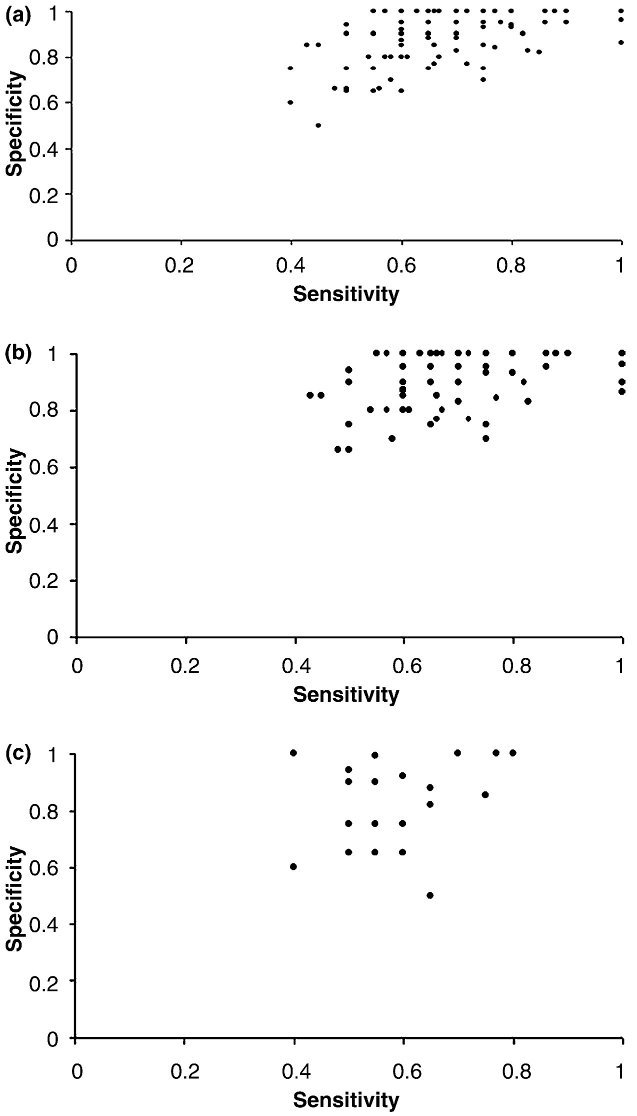

Fig. 5. Mining the PDB sequence database for similar sequences using structural template information. All the templates are scanned in the sequence database with superfamily member proteins as query sequence by a multimotif-scanning program. True positives have been characterized by same SCOP [7] structure classification status. (a) Results obtained for all (110) multi-member superfamilies in PASS2 [11] database. (b) Specificity–sensitivity plots of poorly represented superfamilies in PASS2 (with less than 25 structural entries in 95%-non-redundant PDB [12,13] dataset). (c) Same as (b) but for largely populated PASS2 superfamilies (with more than 25 entries in 95%-non-redundant PDB dataset). High sensitivity of large superfamilies is dependent on the sequence diversity between individual entries.

Improvement of alignment accuracy using structural templates. A rate-limiting step in large-scale analysis of evolutionary trends in superfamilies has been in obtaining good quality alignments. More often, when homologous sequences are to be appended to a carefully curated pre-existing alignment of a superfamily, this is a tedious and slow process. The occurrence of structural templates can be useful to multiply align distantly related sequences with an existing alignment by seeding or fixing conserved regions as initial equivalences. An alignment algorithm (FMALIGN) has been employed to utilize the structural template regions by combining progressive

Table 1
Characterization of unknown proteins

Code and name of the superfamily	Number of motifs identified	Number of true positives identified	Number of true positives confirmed by PSI-BLAST[a]
02.01.001	4	6	4
02.01.056	6	2	2
02.02.058	4	8	5
02.02.152	8	4	4
02.03.054	6	4	3
02.03.071	6	10	8
02.03.100	2	2	2
02.04.153	3	1	1
02.04.156	5	9	7
02.07.006	3	15	12
Actin_bin	5	2	2
Repressor_like	3	3	2
Ppase2	4	3	2
Sslipid	4	2	2
Strep	6	1	1

[a] The *E*-value threshold for each PSI-BLAST run was 0.001 with five iterations.

dynamic algorithm, local substructure alignment and iterative refinement to achieve an improved alignment (available at: http://caps.ncbs.res.in/FMALIGN/home.html). This server considers the local similarity of the sequences in the conserved motif regions and allows local conserved regions of the sequences to be fixed and aligns the rest based on normal progressive alignment. The chances of global misalignment are thereby reduced and the possibility of obtaining better overall alignment is increased.

Previously identified hypothetical proteins, which could be distantly related to superfamilies and are putative members, have been aligned with the existing structural alignments. A careful structure-based sequence alignment, derived using COMPARER [27], followed by a manual curation, formed the reference alignment for comparing the quality of alignments derived by this approach. We also compare the template-fixed alignment with other automatic multiple sequence alignment programs such as CLUSTALX [33] and T-Coffee [34] with default parameters. Comparison of the Sum-of-Pair Score (SPS) values [35] for the template-seeded alignment and the normal multiple sequence alignment shows better quality for fixed-motif alignments in comparison to other popular alignment procedures (Fig. 6).

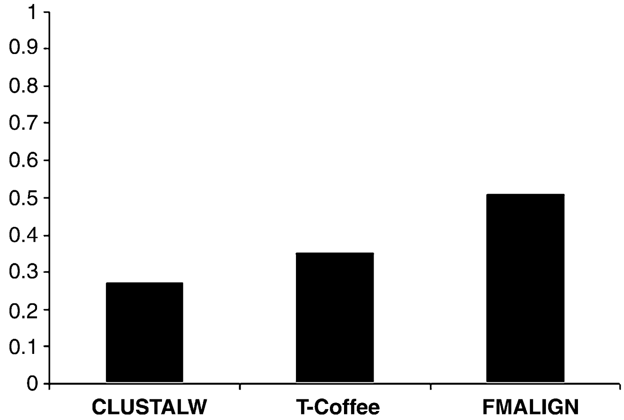

Fig. 6. Utilization of structural templates in improvement of alignment accuracy. Template regions have been utilized to multiply align distantly related hypothetical sequences with an existing alignment by seeding or fixing conserved regions as initial equivalences (FMALIGN). A careful structure-based sequence alignment, derived using COMPARER [25], followed by a manual curation, formed the reference alignment for comparing the quality of alignments. Standard alignment comparing scheme using Sum-of-Pair Score (SPS), described in BAliBASE benchmark alignment methods [33], has been used to compare the alignment accuracy with respect to the reference alignment. The vertical bars denote average SPS obtained by considering 26 representative superfamilies in PASS2 (please see Supplementary Material for *-marked superfamilies) and compared with other popular alignment procedures like CLUSTALW [31] and T-Coffee [32]. Individual SPS values for each superfamilies are provided in Supplementary Materials.

Utilization of structural templates as additional distance restraints in improvement of comparative modelling. A serious limitation in homology modelling, where there is a distant relationship between query and template, has been that the model is a simple replica of the template and does not allow structural deviations. Average template distance patterns for a superfamily can be utilized to attain a more accurate model of a new member of a superfamily using homology modelling procedures [36].

4. Discussion

Since the structural environment of individual residues is a strong regulatory factor to determine the fold of the overall protein [37,38], we have examined physical properties like solvent accessibility of residues and amino acid preferences to identify the conserved core region of proteins. Other independent structural features such as the retention of secondary structure and hydrogen bonding patterns have been studied and compared with full-length sequences to show that they are primarily concentrated in the structural templates. The preservation of secondary structural positions and residue packing is higher within structural templates of superfamily members. During the comparison of similarities in structural features between proteins related by low sequence identities, the spread in data is so high that it is often impossible to distinguish homologous and analogous situations. In this paper, we suggest a reductionist approach where the structural templates can be considered for comparisons as they represent the conserved core structure of each superfamily and also provide the minimal requirement of sequence and structural information to retain each superfamily fold.

The structural templates have been ranked on the basis of conservation of structural features like solvent accessibility, amino acid conservation, secondary structure, hydrogen bonding, residue packing and non-polar contacts [39]. This study provides a graded set of templates for each superfamily depending upon the extent of structural conservation. Access to the consolidated results of the analysis of all these structural features for superfamily alignments is provided through the World Wide Web [39].

We further demonstrate the application of structural templates in three different areas: sequence search, multiple sequence alignment and homology modelling. In each case, the inclusion of the information of structural templates like position in sequence, permitted amino acid exchanges and spatial information give rise to sensitive and accurate results. (a) Sensitive search for homologues is possible using profile-based techniques against a database of sequences. In addition, patterns or templates can be provided as additional restraints prior to iterative profile searches such as PHI-BLAST [40] that lead to higher sensitivity and lower risks of false positives; however, such searches are usually limited to the use of one template in one run. We have used multiple structural templates as simultaneous constraints and show that they do not lead to false positives despite poor coverage. (b) Inclusion of sequence homologues to a pre-existing structure-based sequence alignment requires careful manual curation. Usually, automatic methods such as MALIGN [41] and CLUSTAL [42] provide reliable alignments for closely related sequences. However, utilization of structural templates through a template-fixed alignment algorithm provides better results even at lower sequence identity range, such as superfamily level. (c) It is well known that homology modelling fails to provide reliable models where the sequence identity between the query and the template structure is low. Such low-resolution models still find value either in supporting fold recognition exercises or to understand the gross distribution of residues and charges. However, there are clear limitations in the applications of such models, since the positions and orientations of the structurally conserved regions of the model remain highly similar to that in the template structure. It is also not very useful to consider large number of templates in homology modelling during distant relationships [43]. The orientation of structural templates, as described by distances and angles, when supplied as additional constraints in homology modelling, gives rise to models that are closer to the experimental structure. This has been tested on known examples of distantly related proteins whose structural information is available [36]. In most instances, the model obtained by using one of the structural homologues as a query and providing the spatial orientations at templates observed in the other structural homologue as restraints give rise to models that are closer to the experimental structure. The availability of structural information of conserved regions can also be applied to other areas in modelling, molecular dynamics and docking.

Acknowledgements: We thank Dr. N. Srinivasan for critical reading of the manuscript and for useful discussions. R.S. is a Senior Research Fellow funded by the Wellcome Trust Grant. We also acknowledge National Centre for Biological Sciences (TIFR) for financial support.

References

[1] Lesk, A.M. and Chothia, C. (1980) J. Mol. Biol. 136, 225–270.
[2] Barton, G.J. (1990) J. Mol. Biol. 212, 389–402.
[3] Swindells, M.B. and Thornton, J.M. (1993) Protein Eng. 6, 711–715.
[4] Chothia, C. (1992) Nature 357, 543–544.
[5] Chothia, C., Hubbard, T., Brenner, S., Barns, H. and Murzin, A. (1997) Annu. Rev. Biophys. Biomol. Struct. 26, 597–627.
[6] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) Structure 5, 1093–1108.
[7] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.
[8] Sowdhamini, R., Burke, D.F., Huang, J.F., Mizuguchi, K., Nagarajaram, H.A., Srinivasan, N., Steward, R.E. and Blundell, T.L. (1998) Structure 6, 1087–1094.
[9] Pandit, S.B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S.S., Mhatre, N.S., Sowdhamini, R. and Srinivasan, N. (2001) Nucleic Acids Res. 30, 289–293.
[10] Bray, J.E., Todd, A.E., Pearl, F.M., Thornton, J.M. and Orengo, C.A. (2000) Protein Eng. 13, 153–165.
[11] Mallika, V., Bhaduri, A. and Sowdhamini, R. (2002) Nucleic Acids Res. 30, 284–288.
[12] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) Eur. J. Biochem. 80, 319–324.
[13] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) Nucleic Acids Res. 28, 235–242.
[14] Drenth, J., Low, B.W., Richardson, J.S. and Wright, C.S. (1980) J. Biol. Chem. 255, 2652–2655.
[15] Lesk, A.M. and Chothia, C. (1982) J. Mol. Biol. 160, 325–342.
[16] Farber, G.K. and Petsko, G.A. (1990) Trends Biochem. Sci. 15, 228–234.
[17] Kannan, N., Selvaraj, S., Gromiha, M.M. and Vishveshwara, S. (2001) Proteins 43, 103–112.
[18] Nagano, N., Orengo, C.A. and Thornton, J.M. (2002) J. Mol. Biol. 321, 741–765.
[19] Murzin, A.G., Lesk, A.M. and Chothia, C. (1992) J. Mol. Biol. 223, 531–543.
[20] Chelvanayagam, G., Heringa, J. and Argos, P. (1992) J. Mol. Biol. 228, 220–242.
[21] Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I. and Schrag, J. (1992) Protein Eng. 5, 197–211.
[22] Flores, T.P., Orengo, C.A., Moss, D.S. and Thornton, J.M. (1993) Protein Sci. 2, 1811–1826.
[23] Russell, R.B. and Barton, G.J. (1994) J. Mol. Biol. 244, 332–350.
[24] Lee, B. and Richards, F.M. (1971) J. Mol. Biol. 55, 379–400.
[25] Hubbard, T.J. and Blundell, T.L. (1987) Protein Eng. 1, 159–171.
[26] Johnson, M.S. and Overington, J.P. (1993) J. Mol. Biol. 233, 716–738.
[27] Sali, A. and Blundell, T.L. (1990) J. Mol. Biol. 212, 403–428.
[28] Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. and Overington, J.P. (1998) Bioinformatics 14, 617–623.
[29] Nishikawa, K. and Ooi, T.J. (1986) Biochemistry 100, 1043–1047.
[30] Chou, K.C., Nemethy, G. and Scheraga, H.A. (1984) J. Am. Chem. Soc. 106, 3161–3170.
[31] Srinivasan, N., Sowdhamini, R., Ramakrishnan, C. and Balaram, P. (1991) in: Molecular Conformation and Biological Interactions (Balaram, P. and Ramaseshan, S., Eds.), pp. 59–63, Indian Academy of Sciences, Bangalore.
[32] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[33] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) Nucleic Acids Res. 24, 4876–4882.
[34] Notredame, C., Higgins, D.G. and Heringa, J. (2000) J. Mol. Biol. 302, 205–217.
[35] Thompson, J.D., Plewniak, F. and Poch, O. (1999) Nucleic Acids Res. 27, 2682–2690.
[36] Chakrabarti, S., John, J. and Sowdhamini, R. (2003) J. Mol. Model 10, 69–75.
[37] Wako, H. and Blundell, T.L. (1994) J. Mol. Biol. 238, 682–692.
[38] Overington, J.P., Johnson, M.S., Sali, A. and Blundell, T.L. (1990) Proc. R. Soc. Lond. B. Biol. Sci. 241, 132–145.
[39] Chakrabarti, S., Venkatramanan, K. and Sowdhamini, R. (2003) Protein Eng. 16, 791–793.
[40] Zhang, Z., Schaffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V. and Altschul, S.F. (1998) Nucleic Acids Res. 26, 3986–3990.
[41] Johnson, M.S., Overington, J.P. and Blundell, T.L. (1993) J. Mol. Biol. 231, 735–752.
[42] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[43] Srinivasan, N. and Blundell, T.L. (1993) Protein Eng. 6, 501–512.