Protein Engineering vol.14 no.4 pp.219–226, 2001

Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins

S.Balaji and N.Srinivasan[1]

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

[1]To whom correspondence should be addressed.
E-mail: ns@mbu.iisc.ernet.in

The database PALI (Phylogeny and ALIgnment of homologous protein structures) consists of families of protein domains of known three-dimensional (3D) structure. In a PALI family, every member has been structurally aligned with every other member (pairwise) and also simultaneous superposition (multiple) of all the members has been performed. The database also contains 3D structure-based and structure-dependent sequence similarity-based phylogenetic dendrograms for all the families. The PALI release used in the present analysis comprises 225 families derived largely from the HOMSTRAD and SCOP databases. The quality of the multiple rigid-body structural alignments in PALI was compared with that obtained from COMPARER, which encodes a procedure based on properties and relationships. The alignments from the two procedures agreed very well and variations are seen only in the low sequence similarity cases often in the loop regions. A validation of Direct Pairwise Alignment (DPA) between two proteins is provided by comparing it with Pairwise alignment extracted from Multiple Alignment of all the members in the family (PMA). In general, DPA and PMA are found to vary rarely. The ready availability of pairwise alignments allows the analysis of variations in structural distances as a function of sequence similarities and number of topologically equivalent Cα atoms. The structural distance metric used in the analysis combines root mean square deviation (r.m.s.d.) and number of equivalences, and is shown to vary similarly to r.m.s.d. The correlation between sequence similarity and structural similarity is poor in pairs with low sequence similarities. A comparison of sequence and 3D structure-based phylogenies for all the families suggests that only a few families have a radical difference in the two kinds of dendrograms. The difference could occur when the sequence similarity among the homologues is low or when the structures are subjected to evolutionary pressure for the retention of function. The PALI database is expected to be useful in furthering our understanding of the relationship between sequences and structures of homologous proteins and their evolution.

Keywords: comparative modelling/homologous proteins/phylogeny/structural comparison/structure-based alignments

Introduction

Homologous proteins are characterized by significant sequence similarity, similar three-dimensional (3D) structures and, often, common function (Rossmann and Argos, 1976; Argos and Rossmann, 1979; Lesk and Chothia, 1980, 1982; Chothia and Lesk, 1986; Overington *et al.*, 1990; Sander and Schneider, 1991; Orengo, *et al.*, 1992; Flores *et al.*, 1993; Hilbert *et al.*, 1993; Mirny and Shakhnovich, 1999; Wood and Pearson, 1999). The most useful first step in the comparison of homologous protein structures is a database of classified protein structures and their structural alignments. Value-added protein structural databases include SCOP (Murzin *et al.*, 1995), CATH (Orengo *et al.*, 1997), FSSP (Holm and Sander, 1994), CAMPASS (Sowdhamini *et al.*, 1996, 1998), structural comparison of SCOP domains (Gerstein and Levitt, 1998; Levitt and Gerstein, 1998), Entrez3D (Hogue *et al.*, 1996) and ASTRAL (Brenner *et al.*, 2000). Whereas SCOP and CATH classify protein structures at various levels of hierarchy, FSSP organizes similar 3D structures together. Databases such as 3D_ALI (Pascarella and Argos, 1992; Pascarella *et al.*, 1996), HSSP (Sander and Schneider, 1991), HOMSTRAD (Overington *et al.*, 1990; Mizuguchi *et al.*, 1998a), ALBASE (Sali and Overington, 1994) and LPFC (Schmidt *et al.*, 1997) focus largely on the homologous proteins and these databases provide structure-based alignments.

Analysis of these databases could have an implication for the comparative modelling. One of the approaches to improve the accuracy of the models generated using comparative modelling techniques is to equip the modelling procedure with the information on sequence-dependent structural variations within homologous proteins (Hilbert *et al.*, 1993; Srinivasan and Blundell, 1993). For example, several groups (Flores *et al.*, 1993; Yee and Dill, 1993; Chelvanayagam *et al.*, 1994; Russell and Barton, 1994; Rost, 1997) have analysed variations in a variety of structural features in pairs of homologous proteins. The features studied included solvent accessibility, secondary structure and side-chain conformation as a function of sequence variation.

While the databases such as those mentioned above are certainly very useful, the simultaneous availability of pairwise and multiple alignments of protein structures and the ready availability of structure-based phylogeny can form basic steps to aid further understanding of relationship between sequence and structural variability. One of the principal objectives behind setting-up the database PALI (Phylogeny and ALIgnment of homologous protein structures) is the ready availability of derived data to study variations of various structural features of homologous proteins as a function of sequence similarity. Such a study can be significantly aided by the availability of structure-based sequence alignments performed by considering two proteins at a time (pairwise). PALI contains a large number of pairwise alignments characterized by a wide range of sequence identity between topologically equivalent residues.

Following the work of Eventoff and Rossmann (Eventoff and Rossmann, 1975), it was established by Johnson *et al.* (Johnson *et al*., 1990a,b) and later by Grishin (Grishin, 1997) that structure-based phylogenetic tree diagrams can also be useful in understanding the evolution of proteins. Structural similarity-based and as structure-dependent, sequence

similarity-based phylogenetic tree diagrams of various families are readily available in PALI and these give an immediate picture of the most closely related homologues to a protein structure. Incorporation of the sequence of a new protein, belonging to a family, in such a phylogenetic diagram in PALI could provide clues to choosing basis structures in the comparative model building of the new protein.

We also report a validation of the multiple rigid-body structural alignments in PALI by comparing them with those obtained from a more sophisticated procedure (COMPARER). The direct pairwise alignments in PALI are also assessed by comparing them with the pairwise alignments obtained from multiple alignment of all the members in the family. Using the data in PALI we report the relationship between variations in sequence and structural similarities among homologous protein structures. Although for most of the families structure-based dendrograms are similar to the corresponding structure-dependent, sequence-based dendrograms, we discuss the case of a representative protein family where differences in the two kinds of dendrograms exist.

Materials and methods

Data set

The homologous protein structural families and proteins in each family used in PALI release 1.1 (available at http://pauling.mbu.iisc.ernet.in/~oldpali) are derived based on rigorous consultation of HOMSTRAD (Mizuguchi *et al.*, 1998a) and SCOP (Murzin *et al.*, 1995). The release of PALI 1.1 used in this work comprises 225 families involving 990 protein domains, 3850 structural alignments, about 520 000 residue–residue alignments and 450 dendrograms. A subsequent update of PALI (release 1.2; http://pauling.mbu.iisc.ernet.in/~pali) contains over 500 families (Balaji *et al.*, 2001).

Structural alignments

Every protein in a family is structurally aligned, pairwise, with every other member in the family. All the proteins within a family are also simultaneously superimposed to obtain the alignment of multiple structures. Obviously, in families with only two members the pairwise and multiple alignments are identical. The latest version (4.2) of the STAMP suite of programs (Russell and Barton, 1992), which provides rigid-body treatment to structures, has been used for the superposition of structures. Although the procedure is automated to suit the large-scale application as in setting-up PALI, the result files of the superposition program have been manually inspected to ensure that there is no erroneous result.

One of the common measures of structural divergence between two homologous protein structures is the root mean square deviation (r.m.s.d.) of topologically equivalent Cα atoms. It has been shown that the r.m.s.d. value for a given pair of proteins could depend on the number of topological equivalences (e.g. Swindells, 1996). Further, identical r.m.s.ds in two superpositions do not guarantee the same extent of structural divergence since the number of topologically equivalent Cα atoms in the two pairs could be very different. Hence we calculated the Structural Distance Metric (SDM) (Johnson *et al.*, 1990a,b) for every pairwise alignment in PALI. SDM combines the r.m.s.d. and the number of equivalences and it was defined by Johnson *et al.* as

$$\mathrm{SDM} = -100\log[(w_1 \times \mathrm{SRMS}) + (w_2 \times \mathrm{PFTE})]$$

where

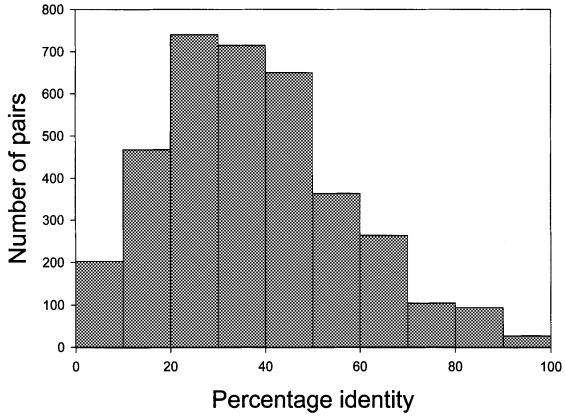

Fig. 1. Histogram showing the number of pairs of homologous proteins, used in the present analysis, at various ranges of pairwise sequence identity of topologically equivalent residues.

$$\mathrm{SRMS} = 1 - \text{r.m.s.d. (in Å)}/3.5$$

$$\mathrm{PFTE} = \text{No. of equivalent C}\alpha\text{ atoms/No. of residues in the smallest protein}$$

$$w_1 = [(1 - \mathrm{SRMS}) + (1 - \mathrm{PFTE})]/2$$

and

$$w_2 = (\mathrm{SRMS} + \mathrm{PFTE})/2$$

The definitions of the weights w_1 and w_2 are such that SDM is a more effective representation than r.m.s.d., especially in the case of distantly related proteins.

Phylogenetic relationships

Structure-based and structure-dependent, sequence-based phylogenetic tree diagrams were generated for every family in PALI. The PHYLIP package of programs (Felsenstein, 1989) involving KITSCH was used to generate dendrograms. The input to structure-based phylogeny of a family is a matrix of SDM between various protein domains in the family. The percentage sequence non-identity matrices were used to generate structure-dependent, sequence-based phylogenetic dendrograms. Using the Web interface to PALI it is possible to generate a dendrogram which can incorporate a query sequence on to the phylogenetic relationship of an existing homologous protein family (Sujatha *et al.*, 2001).

Results and discussion

Variation in sequence identity within pairs of homologous proteins

Figure 1 shows the distribution of the number of pairwise alignments at various levels of percentage sequence identity for topologically equivalent residues. Over 600 pairs of proteins occur in each of the ranges 20–30, 30–40 and 40–50%. The distribution falls markedly under 20% and over 50%. Thus, much of the data used in the present analysis are characterized by pairs of proteins with sequence identity lying in the range 20–50%. The availability of pairwise alignments at various levels of sequence similarity should provide a convenient means of studying variations in the structural properties of two homologues such as solvent accessibility, lengths and orientation of equivalent secondary structures and conformation of equivalent loops and side chains.

Assessment of the quality of the multiple structural alignments

We compared the quality of the multiple structural alignments in PALI, which were obtained by rigid-body superposition using STAMP, with those obtained using COMPARER (Sali and Blundell, 1990; Zhu *et al.*, 1992). COMPARER uses structural properties, at every residue position, such as solvent accessibility class and secondary structure and relationships such as hydrogen bonding pattern. To facilitate detailed comparison of the multiple structural alignments we chose families in all α class in PALI at random to represent distinct average pairwise sequence identities. The extent of sequence identities ranged from 20% (family of calponin homology domains) to ~61% (family of acyl carrier proteins) and there are three members in each of these families. Figure 2a and b show the alignment in PALI and from COMPARER, respectively, for the family of calponin homology domains which corresponds to a low average sequence identity (20%).

Detailed comparison of multiple structural alignments, for the four families, obtained from COMPARER and STAMP shows that number of alignment positions where difference in alignment exists varies from 8.2% (acyl carrier proteins) to 23.3% (TMV-like viral coat proteins). The percentage difference in alignments in the conserved secondary structural regions is zero for three of the families where the average pairwise sequence identity is above the 'twilight' zone defined by Doolittle (1981). In 2.9% of the aligned positions of calponin homology domains a difference in alignment exists (Figure 2a and b). All of these differences occur in the alignment positions involving termini of the helices or in loops. These differences could occur since the lengths of the equivalent helices and conformations at the termini of the helices are known to vary markedly in distant homologues. Thus very few differences in the alignments are seen even for the family with low (20%) average pairwise identity. The reason for high correspondence of rigid body-based and COMPARER-based multiple structural alignments may be the close similarity of tertiary structures within the homologous proteins

Comparison of direct pairwise alignments with the pairwise alignments extracted from multiple alignments

It is conceivable that multiple structural alignments may be more accurate than pairwise alignments. A further assessment of the quality of the alignments in PALI was made by comparing Pairwise alignment extracted from Multiple Alignment of all the members in the family (PMA) and the alignment obtained by directly superposing the two proteins (DPA; Direct Pairwise Alignment). We considered 154 families in PALI with three or more members in each family for the comparison of DPA and PMA. We asked following questions:

1. How often the differences occur in the alignment positions in DPA and PMA?
2. How many of these differences correspond to equivalent residues?
3. How many of these differences involve helices and β-strands?

The results are summarized in Figure 3. Out of >510 000 residue–residue alignments in 377 534 (73.8%) positions there is no difference in the alignment between DPA and PMA. Hence in most of the positions the alignments from DPA and PMA match. Out of 146 304 (26.2%) mismatch positions only 14 965 positions (10.2%) involved topologically equivalent residues. Hence about 90% of the positions with disagreement in the alignment come from structurally variable regions which are often loops. Out of 14 965 equivalent positions with disagreement between DPA and PMA, 9695 positions (64.8%) involve at least one residue in the loop. Many of these are likely to correspond to termini of helices and β-strands where structural variability is more pronounced than in the middle of the helix or β-strand. There are only 5270 positions where the alignments between DPA and PMA disagree and the residues involved come from helices or β-strands. This is a very small proportion (0.9%) of the total number (564 095) of residue–residue or residue–gap alignments in the database. As many as 4720 of these 5270 positions correspond to residues from identical secondary structures. Preliminary examination of some of these disagreeing DPA and PMA suggests that shifts in alignments in helical regions by three or four residues (corresponding roughly to the number of residues per turn of the helix) and shifts by two residues in β-strand regions are common. Thus a slide in the alignment by one turn is the most common kind of disagreement which occurs in only 0.9% of all the residue–residue alignments in the database.

Disagreement between DPA and PMA for a pair of proteins can be pronounced for at least two reasons: (1) sequence identity between the proteins is low and hence more divergent in 3D structures and the mean pairwise sequence similarity in the family is low and (2) the number of proteins within the family is large. The family of globins has both of these features with 35 members and sequence identities between certain members falling to below 20%. As a result, many pairs in the globins family are expected to show differences between DPA and PMA. Hence we investigated the alignments in the globin family in more detail.

Out of 595 pairs of globins, 431 pairs show at least one difference between DPA and PMA. A minority of 174 pairwise alignments shows differences between DPA and PMA involving residues present in helices in the two structures. In 63 pairs of globins a three or four residue shift (about one turn) is seen in the alignment of equivalent helices. We performed further analysis on the cases with differences between DPA and PMA involving residues in the helices. The main objective of this analysis was to find out if, in general, DPA or PMA is better. For this purpose we compared the following local environments around various residues present in helices and involved in differences between DPA and PMA:

1. Percentage solvent accessibility at the residue.
2. Ooi number (number of Cα atoms around a residue within a sphere of 9 Å radius).
3. Number of interacting side chains within Ooi sphere (number of non-polar residues within Ooi sphere with Cβ–Cβ distance less than Cα–Cα distance).
4. Packing density (ratio of the sum of the volumes of interacting residues within the Ooi sphere to volume of the Ooi sphere).

Correlation of these structural features for the aligned residues (in helices) in DPA and PMA were evaluated by means of the statistical correlation coefficient. Table I shows that the correlation coefficients between DPA and PMA for various structural environments are low. This suggests a pronounced structural difference in the pairs of globins showing differences between DPA and PMA involving residues in helices. Differences in correlation coefficients between DPA and PMA are

Fig. 2. Legend on facing page

so low as to favour clearly one of the two alignments. This result may be viewed in the light of the fact that there is a significant difference in packing between helices among pairs of globins with low sequence similarity although the geometry of the packing of helices involved in positioning the haem group is well conserved (Lesk and Chothia, 1980). The nature of the differences in the structures is such that many of the structural environments considered around 'equivalent' residues, as suggested by DPA and PMA, do not correlate very well.

Gross relationships between sequence and structural variability

The relationship between r.m.s.d. and sequence identity among homologous protein structures was first studied by Chothia and Lesk (Chothia and Lesk, 1986) using a small dataset and subsequently studied by others using larger datasets (Hubbard and Blundell, 1987; Flores *et al.*, 1993; Chelvanayagam *et al.*, 1994; Russell and Barton, 1994).

We analysed the SDM for 3625 pairwise alignments as a function of percentage sequence identity calculated for the topologically equivalent Cα atoms. A small number of pairs corresponding to less than ~10% sequence identity show a widespread distribution of SDM (data not shown). Figure 4 shows the distribution of average SDM calculated at every 5% interval of sequence identity. This distribution is very similar to that reported by Chothia and Lesk (1986) and others. This suggests that the use of SDM has the advantage of combining r.m.s.d. and number of equivalences and it behaves similarly to r.m.s.d. The points in Figure 4 could be fitted to the equation

$$\mathrm{SDM} = C_1 + C_2\exp[-(ID - C_3)/C_4]$$

where *ID* is the sequence identity and C_1–C_4 are constants with values 28.6, 185.6, 0 and 11.5, respectively. The similarity of the overall nature of the fitted curve suggests that SDM is analogous to r.m.s.d. which was used in previous studies. As SDM combines r.m.s.d. and number of equivalences, SDM appears to be a more effective representation than r.m.s.d.

Figure 5 shows the distribution of SDM plotted against number of equivalences which is averaged at every five equivalences. There is a steep fall in SDM until the number of equivalences increases to ~40. The fall in SDM is much gentler after about 40 equivalences, suggesting that SDM is a sensitive descriptor of structural distance between two proteins when there is only a small number of overlapping Cα atoms. The nature of the curve in Figure 6b can be modelled as a double exponential function:

Fig. 2. Structure-based sequence alignment for the family of calponin homology domains (**a**) as in PALI and (**b**) as deduced using COMPARER (Sali and Blundell, 1990; Zhu *et al.*, 1992). The first four letters of each code represent the code used in protein databank and the fifth character is the chain identifier. The structural features at various residue positions are represented using the program JOY (Mizuguchi *et al.*, 1998b). Key to JOY notation: solvent inaccessible, UPPER CASE (O); solvent accessible, lower case (o); positive ϕ, *italic* (*o*); *cis* peptide, breve (ŏ); hydrogen bond to other side chain, tilde (õ); hydrgen bond to main-chain amide, **bold (o)**; hydrogen bond to main-chain carbonyl, underline (o); disulphide bond, cedila (ç).

$$\mathrm{SDM} = D_1 + D_2\exp[-(n_{eq} - D_3)/D_4] + D_5\exp[-(n_{eq} - D_3)/D_6]$$

where n_{eq} is the number of equivalences and D_1–D_6 are constants with values 0, 147.8, 0, 24.4, 33.3 and 1.55×10^{10}, respectively.

Comparison of dendrograms generated from structural similarities with those derived from structure-dependent sequence-based similarities

A structure-based dendrogram was derived for every family in PALI using SDM obtained from all the pairwise alignments within a family. Equivalent residues within pairwise alignments were used to obtain the measure of sequence dissimilarity between two proteins and another dendrogram was generated for every family. The structure-based and structure-dependent, sequence identity-based relationships were compared for all the 154 families with three or more members in the family. For every family, the correlation coefficient was calculated between the matrix of SDMs and the matrix of sequence dissimilarity.

Figure 6 shows the distribution of correlation coefficient values in 154 families; 44 out of 154 structures (29%) have a high correlation coefficient of 0.9 and are also identified to have similar SDM-based and sequence-based dendrograms. Nine families have a negative correlation coefficient and most of these have differences in the relative order of homologous proteins in the two dendrograms. However, in general, the correlation coefficients are found to have no connection with the congruency or otherwise of the two types of dendrograms (S.Balaji and N.Srinivasan, unpublished results).

A radical difference in the relative ordering of proteins in these two types of tree diagram could occur owing to, among various reasons, a low sequence similarity between homologous proteins and the nature of the functional states of the homologous protein structures (S.Balaji and N.Srinivasan, unpublished results). The interleukin 8 family is discussed below to demonstrate a typical case of variability in the two kinds of dendrograms.

Figure 7a and b show dendrograms generated on the basis of a matrix of amino acid dissimilarity of topologically equivalent residues and 3D structural dissimilarity matrix, respectively, for the family of interleukin 8. All the proteins

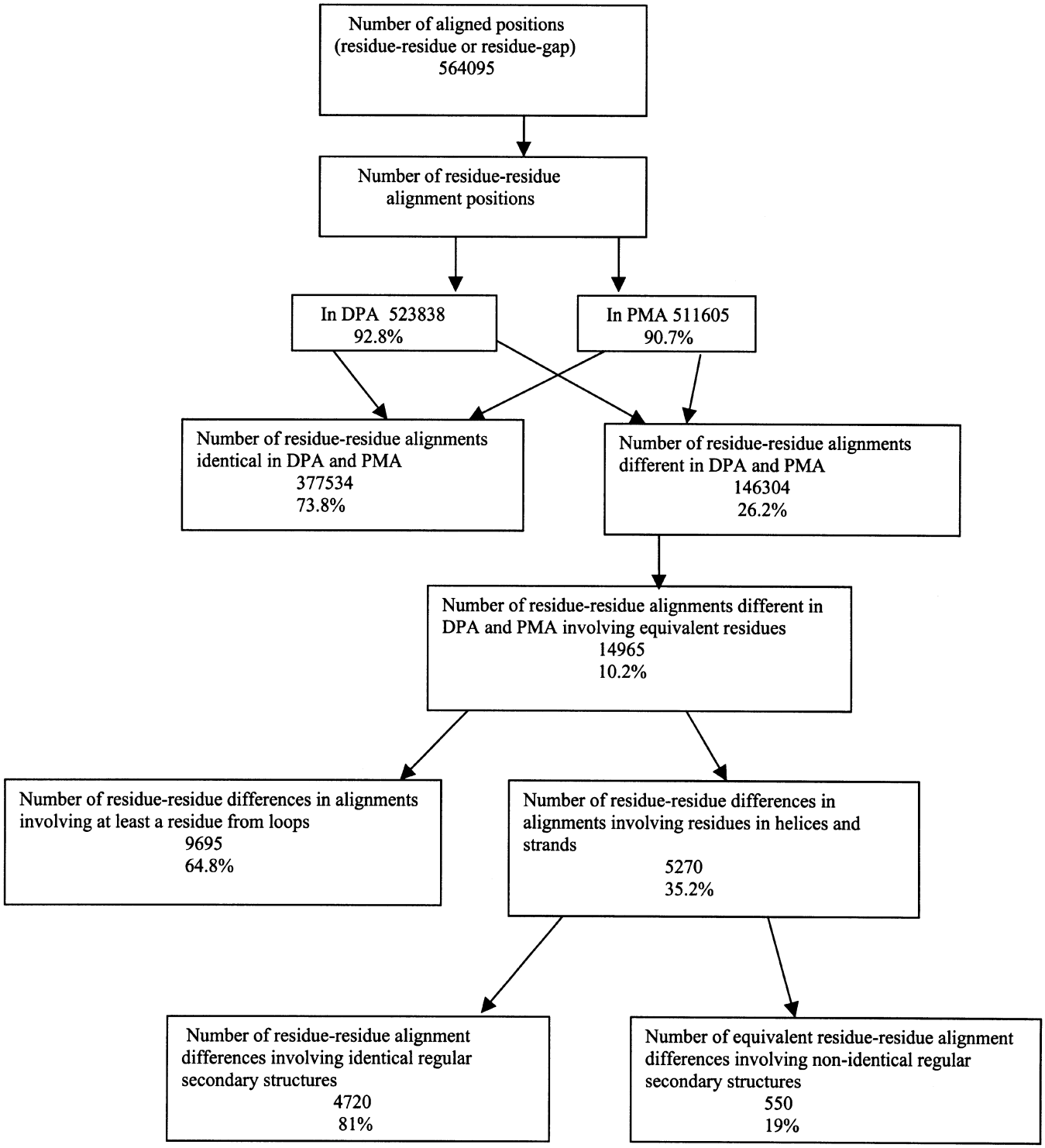

Fig. 3. Comparative analysis of Direct Pairwise Alignment (DPA) between two homologous proteins and Pairwise alignment derived from Multiple Alignment (PMA) of all the structures in the family.

except 1plf (bovine platelet factor 4) are from humans. Platelet factor 4 from human (1rhp) has about 76% of the topologically equivalent residues identical with the homologue from bovine. The sequence similarity-based dendrogram (Figure 7a) shows two major clusters, one containing ranties (1tro) and macrophage inflammatory protein (1hum) and the other containing the rest, including the two homologues of platelet factor 4. One of the clear differences between the two dendrograms is that the cluster of platelet factor 4 is separated from the rest of the proteins in the structure-based dendrogram (Figure 7b). The sequence identity for the topologically equivalent residues between human/bovine platelet factor 4 and other members in the family ranges from 0 to 19%. It appears that distantly related homologues characterized by such low sequence identity [below the 'twilight zone' defined by Doolittle (Doolittle, 1981)] need not conform to the inverse relationship between sequence similarity and SDM.

Table I. Correlation coefficients for four structural parameters for the pairs of globins aligned as in DPA and PMA

Alignment type	Percentage solvent accessibility	Ooi number	Number of side chain to side chain interactions	Packing density
DPA	0.514	0.352	0.403	0.428
PMA	0.639	0.577	0.192	0.163

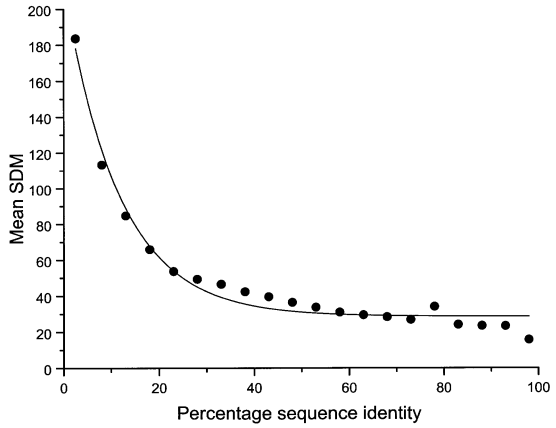

Fig. 4. Plot of structure-based distance metric for pairs of homologous proteins averaged over every 5% range of sequence identity.

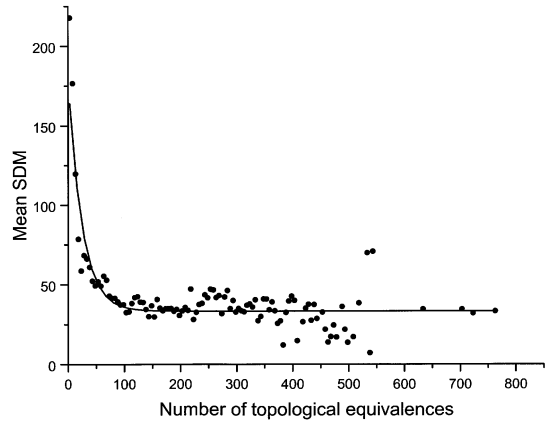

Fig. 5. Plot of structure-based distance metric for pairs of homologous proteins averaged over every five topological equivalences.

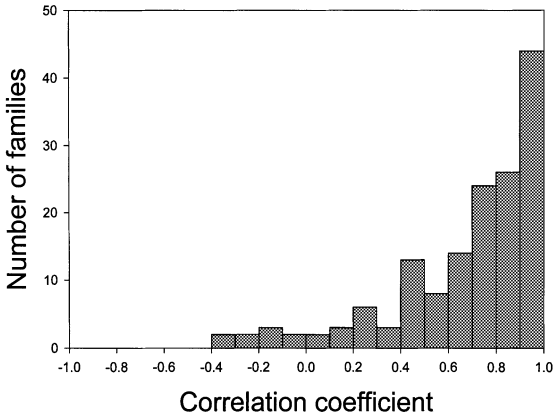

Fig. 6. Distribution of correlation coefficient between sequence-based and structure-based distance matrices for families with at least three members.

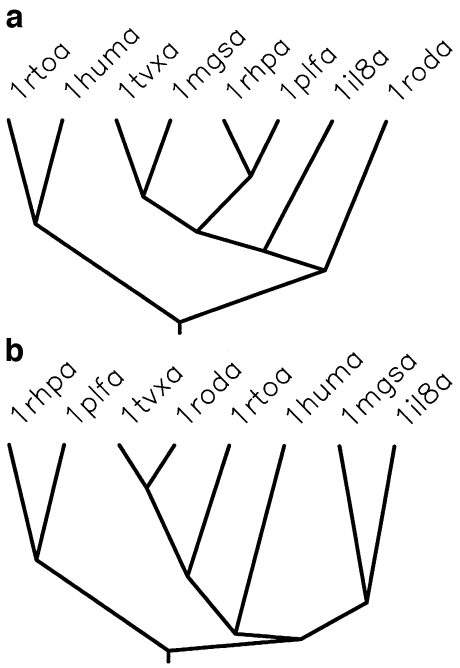

Fig. 7. Dendrograms for the interleukin 8 family of proteins based on (**a**) sequence similarity of topologically equivalent residues and (**b**) structural distance metric. The first four letters of each code represent the code used in the protein databank and the fifth character is the chain identifier.

Conclusions

The use of databases of protein structural alignments forms an important step in the understanding of structure, sequence and functional constraints in the evolution of proteins. They are also helpful in learning about relationships between sequences and structures. Such studies can help in improving the comparative modelling procedures.

Alignment of multiple structures within a family is likely to be more accurate than the pairwise alignments. However, multiple structural alignment could depend on the number of structures within the family that is increasing with the increase in the number of known structures. On the other hand, assessed pairwise alignment establishes the direct relationship between two homologous proteins. It has been shown that pairwise alignments are not, in general, significantly different from multiple structural alignments, perhaps owing to a high similarity of structures within the homologous proteins.

The ready availability of structure-based and structure-dependent, sequence-based dendrograms permits studies on mutual relationships among sequences and structures of homologous proteins. Especially for the families involving low sequence similarities, sequence alignment could be unreliable and a dendrogram using alignment of structures is more appropriate.

Acknowledgements

We thank Mr S. Sai Chetan Kumar for his help in assessing the quality of multiple structural alignments in PALI and Ms S.Sujatha for producing the Web interface to PALI. We also thank Dr Kenji Mizuguchi for sending us the HOMSTRAD database. S.B. is supported by a Fellowship from the Council of Scientific and Industrial Research, India. This research is supported by the award of an International Senior Fellowship in Biomedical Sciences to N.S. from the Wellcome Trust, London.

References

Argos,P. and Rossmann,M.G. (1979) *Biochemistry*, **18**, 4951–4960.
Balaji,S., Sujatha,S., Sai Chetan Kumar,S. and Srinivasan N. (2001*) Nucleic Acids Res.*, **29**, 61–65.
Brenner,S.E., Koehl,P. and Levitt,R. (2000) *Nucleic Acids Res.*, **28**, 254–256.
Chelvanayagam,G., Roy,G. and Argos,P. (1994) *Protein Eng*, **7**, 173–184.
Chothia, C and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.
Doolittle, R.F. (1981) *Science*, **214**, 149–159.
Eventoff,W. and Rossmann,M.G. (1975) *CRC Crit. Rev. Biochem*., **3**, 111–140.
Felsenstein,J. (1989) *Cladistics*, **5**, 164–166.
Flores,T.P., Orengo,C.A., Moss,D.S and Thornton,J.M. (1993) *Protein Sci*., **2**, 1811–1826.
Gerstein,M. and Levitt,M. (1998) *Protein Sci.*, **7**, 445–456.
Grishin,N.V. (1997) *J. Mol. Evol.*, **45**, 359–369.
Hilbert,M., Bohm,G. and Jaenicke,R. (1993) *Proteins*, **17**, 138–151.
Hogue,C., Ohkawa,E. and Bryant,S.H. (1996) *Trends Biochem. Sci.*, **21**, 226–229.
Holm,L. and Sander,C. (1994) *Nucleic Acids Res*., **22**, 3600–3609.
Hubbard,T.J.P. and Blundell,T.L. (1987) *Protein Eng*., **1**, 159–171.
Johnson,M.S, Sutcliffe,M.J. and Blundell,T.L (1990a) *J. Mol. Evol*., **1**, 43–59.
Johnson,M.S., Sali,A. and Blundell,T.L. (1990b) *Methods Enzymol*., **183**, 670–690.
Lesk,A.M. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.
Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol*, **160**, 325–342.
Levitt,M. and Gerstein,M. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.
Mirny,L.A. and Shakhnovich,E.I. (1999) *J. Mol. Biol.*, **291**, 177–196.
Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998a) *Protein Sci.*, **7**, 2469–2471.
Mizuguchi,K., Deane,C.M., Johnson,M.S., Blundell,T.L. and Overington,J.P. (1998b) *Bioinformatics*, **14**, 617–623.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol*., **247**, 536–540.
Orengo,C.A., Brown,N.P. and Taylor,W.R. (1992) *Proteins*, **14**, 139–167.
Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M (1997) *Structure*, **5**, 1093–1108.
Overington J., Johnson,M.S., Sali,A. and Blundell,T.L (1990) *Proc. R. Soc. London*, **241**, 132–45.
Pascarella,S. and Argos,P. (1992) *Protein Eng*., **5**, 121–137.
Pascarella,S., Milpetz,F. and Argos,P. (1996) *Protein Eng.*, **9**, 249–251.
Rossmann,M.G. and Argos,P. (1976) *J. Mol. Biol.*, **105**, 75–95.
Rost,B. (1997) *Fold. Des.*, **2**, S19–S24.
Russell,R.B. and Barton,G.J. (1992) *Proteins*, **2**, 309–323.
Russell,R.B. and Barton,G.J. (1994) *J. Mol. Biol*., **244**, 332–350.
Sali,A. and Blundell,T.L. (1990) *J. Mol. Biol*., **212**, 403–28.
Sali,A. and Overington,J.P. (1994) *Protein Sci.*, **3**, 1582–1596.
Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.
Schmidt,R., Gerstein,M. and Altman,R. (1997). *Protein Sci*., **6**, 246–248.
Sowdhamini,R., Rufino,S.D. and Blundell,T.L (1996) *Fold. Des*., **1**, 209–220.
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.
Srinivasan,N. and Blundell,T.L. (1993) *Protein Eng*., **6**, 501–512.
Sujatha,S., Balaji,S. and Srinivasan,N. (2001) *Bioinformatics*, **17**, 375–376.
Swindells,M.B. (1996) *Methods Enzymol.*, **266**, 643–653.
Wood,T.C. and Pearson,W.R. (1999) *J. Mol. Biol.*, **291**, 977–995.
Yee,D.P. and Dill,D.A. (1993) *Protein Sci.*, **2**, 884–899.
Zhu,Z.-Y., Sali,A. and Blundell,T.L. (1992) *Protein Eng*., **5**, 43–51.

Received August 8, 2000; revised December 12, 2000; accepted January 23, 2001