Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution

S Balaji[†] and N Srinivasan*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India
[†]Present address: National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20894, USA

**Corresponding author (Fax, 91-80-2360 0535; Email, ns@mbu.iisc.ernet.in)*

Several studies based on the known three-dimensional (3-D) structures of proteins show that two homologous proteins with insignificant sequence similarity could adopt a common fold and may perform same or similar biochemical functions. Hence, it is appropriate to use similarities in 3-D structure of proteins rather than the amino acid sequence similarities in modelling evolution of distantly related proteins. Here we present an assessment of using 3-D structures in modelling evolution of homologous proteins. Using a dataset of 108 protein domain families of known structures with at least 10 members per family we present a comparison of extent of structural and sequence dissimilarities among pairs of proteins which are inputs into the construction of phylogenetic trees. We find that correlation between the structure-based dissimilarity measures and the sequence-based dissimilarity measures is usually good if the sequence similarity among the homologues is about 30% or more. For protein families with low sequence similarity among the members, the correlation coefficient between the sequence-based and the structure-based dissimilarities are poor. In these cases the structure-based dendrogram clusters proteins with most similar biochemical functional properties better than the sequence-similarity based dendrogram. In multi-domain protein families and disulphide-rich protein families the correlation coefficient for the match of sequence-based and structure-based dissimilarity (SDM) measures can be poor though the sequence identity could be higher than 30%. Hence it is suggested that protein evolution is best modelled using 3-D structures if the sequence similarities (SSM) of the homologues are very low.

[Balaji S and Srinivasan N 2007 Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution; *J. Biosci.* **32** 83–96]

1. Introduction

The extent of divergence in the three-dimensional (3-D) structures of homologous proteins in a family is related to the extent of similarity in their amino acid sequences (Lesk and Chothia 1980). It is well known that the 3-D structures and structural features of homologous proteins are conserved better than their amino acid sequences (Chothia and Lesk 1986; Hubbard and Blundell 1987; Flores *et al* 1993; Russell and Barton 1994; Balaji and Srinivasan 2001). It has been demonstrated several times (Murzin 1993a b, 1998; Murzin *et al* 1995; Holm and Sander 1997; Russell *et al* 1997; Todd *et al* 2001) that the homologous proteins could diverge beyond recognition at the level of their amino acid sequences but maintain similar structure and function. While the relationship between sequence divergence and structural divergence in homologous proteins suggest a gross exponential decrease of structural divergence as sequence similarity increases, this relationship does not exist at low sequence similarities (below 30%) (Balaji and Srinivasan 2001). In several cases of low sequence similarity proteins retain the fold as well as retain the broad biochemical features and/or functional properties, suggesting an evolutionary connection (Murzin *et al* 1995; Russell and Sternberg 1996, 1997; Sowdamini *et al* 1998; Bray *et al* 2000).

To model evolution of proteins within a set of divergently-evolved proteins it is useful to construct the phylogenetic trees based on the similarities in the amino acid sequences and the base sequences of the genes. However, given the fact that at sequence identities less than about 30% direct relationship

Keywords. Homologous proteins; phylogenetic relationships; protein evolution; protein structures; structural comparison

between sequence divergence and structural divergence does not exist, there is a problem in using base sequence of coding genes and/or amino acid sequences of proteins in modelling evolution. Level of similarity between two divergently-evolved proteins increases as the proteins are viewed from “Sequences of bases in the genes” to “Amino acid sequences of the proteins” to “3-D structures of proteins”. Hence it seems worthwhile to use the 3-D structures in modelling evolution. Construction of phylogenetic trees using 3-D structures was first applied for a variety of protein families by Johnson and coworkers (Johnson *et al* 1992a, b) and later by others (Balaji and Srinivasan 2001; Holm and Sander 1993; Sowdamini *et al* 1996; Grishin 1997; Bujnicky 2000; Goh *et al* 2000). Comparison of distance matrices, used in phylogenetic tree construction methods, has been considered as an equivalent of comparison of phylogenetic trees (Balaji and Srinivasan 2001; Goh *et al* 2000; Pazos and Valencia 2001) to describe co-evolution of interacting partners and to study protein evolution. In this paper we investigate the correspondence between phylogenetic relationships within homologous protein families derived purely based on amino acid sequences with that obtained solely on 3-D structures. We ask the question, how often the correspondence is poor? In these cases therefore to understand the underlying reasons for the poor correlation is sought.

2. Methods

The data set analysed comprises of structure-based and sequence-based dissimilarity matrices derived from the structure-based alignments of homologous proteins. A total of 108 families having 10 or more members in the PALI database (Balaji *et al* 2001; Sujatha *et al* 2001; Gowri *et al* 2003) (release 2.1) were analysed. PALI database (*http://pauling.mbu.iisc.ernet.in/~pali*) comprises of protein structural families in which every member is structurally aligned pairwise with every other member in that family and multiple structural alignment of all members in the family is also available. The alignments were made using STAMP (Russell and Barton 1992) which encodes a rigid body superposition procedure. The structure-based and sequence-based phylogenetic trees for every family with three or more members are also available in the PALI database. The structure based phylogenetic trees were constructed using a structural distance metric (SDM) (Johnson *et al* 1992 a, b) defined as:

$$\mathrm{SDM} = -100*\log\,(\mathrm{w1*PFTE+w2*SRMS}),$$

where, w1 = (1-PFTE+1-SRMS)/2 and w2 = (PFTE+ SRMS)/2;

$$\mathrm{PFTE} = \frac{\text{Number of topologically equivalent residues}}{\text{Length of the smallest protein}},$$

SRMS = 1-(RMSD/3.0), where RMSD is root mean square deviation in Å.

SDM is the measure of structural distance which has been used in constructing structure-based distance matrices for all 108 families in the data set. In the current analysis, the sequence dissimilarity metric for all the aligned positions (calculated for every pair in the families), was taken as a measure of sequence-based distance metric.

Sequence dissimilarity metric; (SEDM) is defined as:

$$\mathrm{SEDM} = -100*\log\,(\mathrm{SSM});$$

SSM = sequence similarity score, defined as

$$\mathrm{SSM} = \Sigma(2*S_{ij}/(S_{ii}*S_{jj}))/N_{al}.$$

Where N_{al} is number of residue-residue alignment positions, S_{ij} is substitution probability score calculated for substitution of *i*th amino acid by *j*th amino acid derived out from structural alignments (Johnson *et al* 1993). Summation is for all aligned positions in the pairwise alignment. It should be noted that the sequence dissimilarity is calculated soley based on the 3-D structure-based alignment of amino acid sequences of the homologous proteins. Linear statistical correlation coefficient was computed between SDM and SEDEM for all 108 families with 10 or more members in the data set. Linear correlation coefficient is defined as:

$$\rho = \Sigma(x_i\text{-}x)(y_i\text{-}y)/\mathrm{sqrt}(\Sigma(x_i\text{-}x)^2\ \Sigma(y_i\text{-}y)^2),$$

where $x = \Sigma x_i/N$ and $y = \Sigma y_i/N$. The the standard deviation in the correlation coefficient was calculated using bootstrap method as proposed by Efron (1979 a, b). In the bootstrap method, involving comparison of *N* structure-based distances (*p*) with corresponding *N* sequence-based distances (*q*), a random number generator is used to draw *N* times (p_i, q_i) independently with replacement from the data set, so that each new pair drawn is an independent random selection of one of the pairs in the data set, to make new data set. This has been repeated 1000 times; each time calculating for linear correlation coefficient. Central interval of distribution of correlation coefficient of the random experiment (performed 1000 times) is evaluated as [*a*,*b*] where *a* and *b* are such that; (Number of outcomes less than *a*)/1000 = 0.16, and (Number of outcomes less than *b*)/1000 = 0.84.

Boot strap estimate of standard deviation is defined as half length of the interval [*a*,*b*]

$$\sigma = (b\text{-}a)/2.$$

As already mentioned correlation coefficient is considered as a measure of similarity between sequence-based and structure-based phylogenetic trees. As the sequence and structure-based dissimilarity measures are the basic inputs into the construction of phylogenetic trees that are based solely on sequences and solely on structures, this method of comparing the sequence dispersion in a family and structure dispersion is independent of phylogenetic tree construction algorithms.

3. Results

The extent of structural divergence between two homologous proteins is measured by a (SDM) and an analogous measure, SEDM, has been used to quantify the dissimilarity at the level of amino acid sequences. The sequence alignment used in calculating SEDM is the outcome of the optimal superposition of C^{α} atoms in the two proteins; and hence the alignments used to calculate SDM and SEDM are essentially the same.

Figure 1a is the plot of average SDM as a function of average SEDM for all pairs in the 108 families considered in the analysis. For simplicity and clarity in the trend of points in figure 1a, observed SEDM values at every 5 units are averaged and the corresponding SDM values are also averaged. It is evident from the figure that there is a large slope change at about 60 units for SEDM. A value of 60 for SEDM corresponds to about 30% sequence identity which is the approximate upper threshold of the 'twilight' zone (Doolittle 1981). When the percentage sequence identity (%I) becomes less than 'twilight' zone, the direct correspondence between sequence and structural variations falls drastically as seen in figure 1b. Figure 1b is the plot of %I vs. SDM for all 32863 pairwise alignments in the PALI structure-alignment database. From figure 1b, it is evident that there is a large spread in the points below 30% sequence identity. Figure 1c shows the distribution of points as in figure 1b except that %I values at every 5% interval and the corresponding SDM values are averaged. It may be noted that fall of average SDM with increasing %I

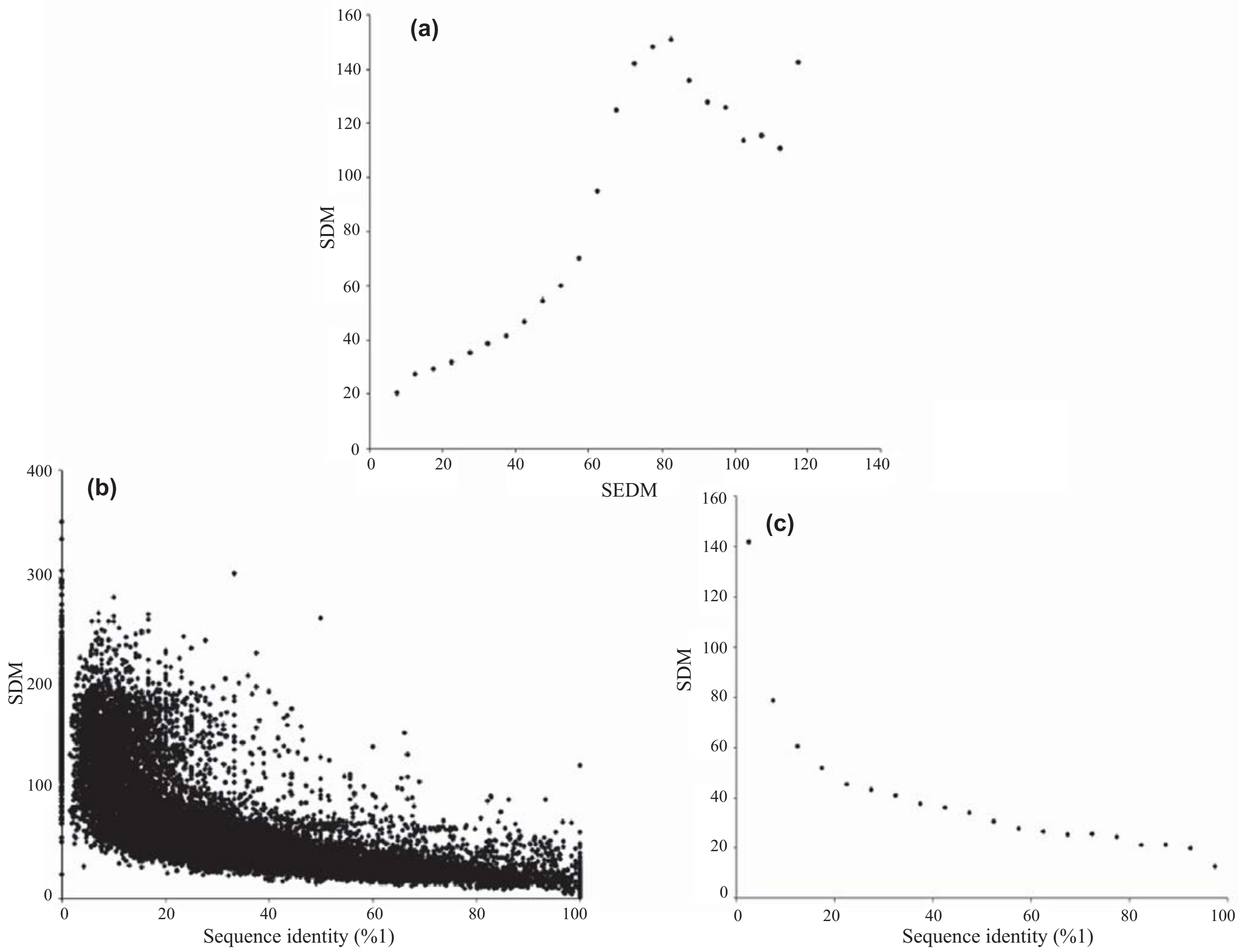

Figure 1. **(a).** Plot of SEDM versus average SDM, which is averaged for every 5 units of SEDM. (**b**) Plot of percentage sequence identity (%I) versus SDM for all 32863 pairwise alignments of homologous protein structures in the PALI database. **(c)** Plot of percentage sequence identity (%I) versus average SDM averaged for every 5% interval of %I.

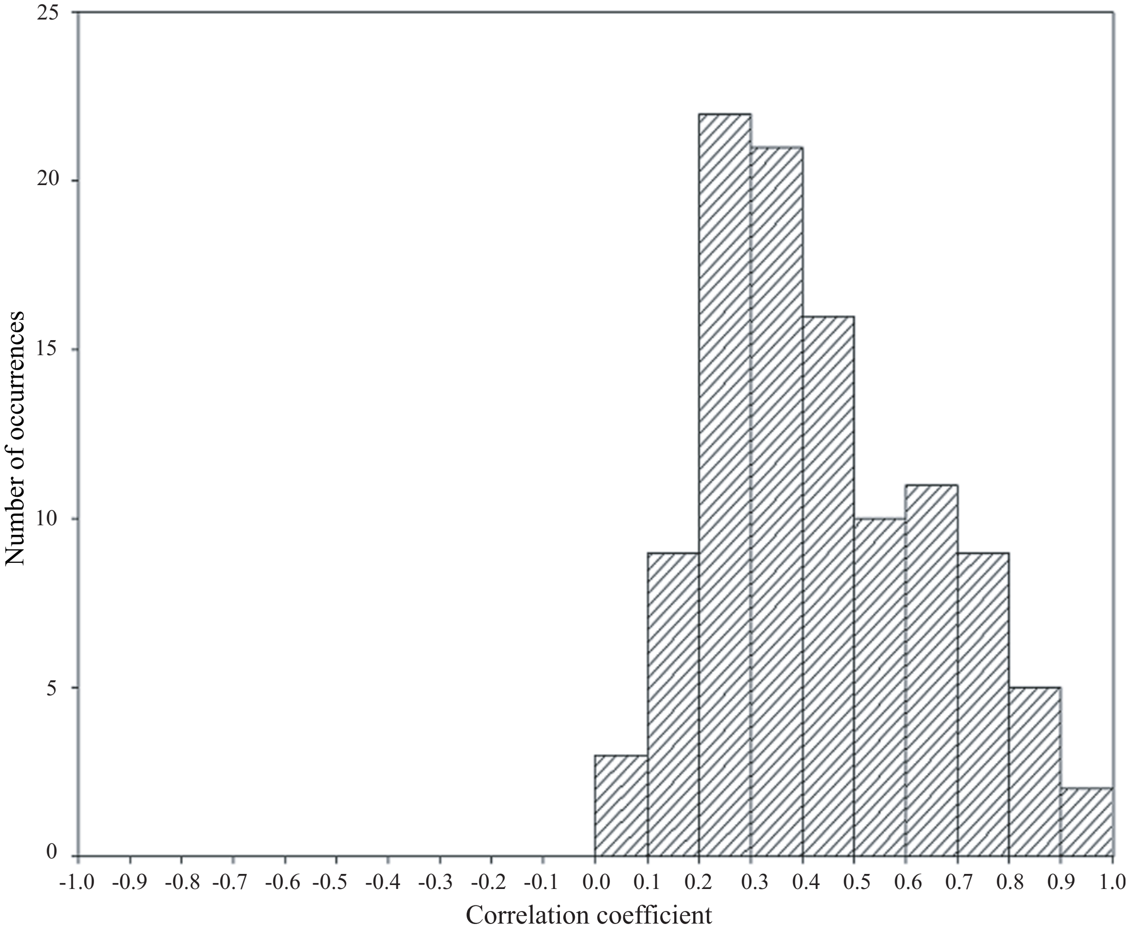

Figure 2. Histogram depicting the distribution of the number of occurrences for every 0.1 interval of correlation coefficient (between sequence-based and structure-based phylogenetic trees) for 108 families.

(figure 1c) is consistent with the earlier studies (Chothia and Lesk 1986; Hubbard and Blundell 1987; Flores *et al* 1993; Russell and Barton 1994; Balaji and Srinivasan 2001). From figure 1b, c it is clear that under about 30% of sequence identity the SDM values are widely varying and direct relationship between extent of sequence identity and SDM is almost non-existent.

For every family two dissimilarity matrices were constructed, one based on all-against-all SEDM (sequence-based) values and the other for corresponding SDM (structure-based) values. Matrices such as these are fed as inputs for phylogenetic tree generation and hence the correlation coefficient for the agreement between these two matrices can be viewed as a measure of extent of agreement between the two tree structures. Figure 2 shows the distribution of the correlation coefficient, between SEDM and SDM, of all 108 families. It is quite clear from the figure that none of the families have negative correlation between the structure-based and sequence-based distance matrices. The distribution could be categorized in to 3 groups.

(i) Families with good correlation (correlation coefficient > 0.6) between sequence-based and structure-based phylogenetic trees.
(ii) Families with poor correlation (correlation coefficient < 0.2).
(iii) Families with intermediate correlation (correlation co-efficient greater than or equal 0.2 but less than 0.6).

Further detailed analysis described is restricted to the first two groups only.

3.1 *Families with high correlation between sequence-based and structure-based phylogenetic trees*

There are 27 families with considerably high correlation co-efficient of greater than 0.6 and the bootstrap standard deviations in correlation coefficient are within 20% of the original value. Figure 3 shows, for these 27 families, the deviation of average SDM for the family from the value suggested by the plot of %I vs. average SDM (figure 1c). In the figure the closed circles represent the average SDM and triangles represent the SDM suggested by the figure 1c. For 18 families, average family SDM is higher than the one suggested by figure 1c. The length of the line joining closed circles and the triangles in the figure is a measure of extent of deviation of the average SDM from the one suggested by figure 1c. It is deduced, from figure 3, that 17 families out of 27 families do not deviate by more than 10 units of SDM. The list of these 27 families is given in table 1. It could be seen from the figure that the deviation of average SDM has a maximum of 28.21. It is important to note that all the families, except three, have average %I above the 'twilight' zone (30%), and the remaining three families too have the average sequence identity close to 30%. Hence the direct relationship between sequence divergence to structural divergence by and large hold for these families which is reflected by the high correlation coefficients. Most of these families are characterized by average sequence identities more than the "twilight zone" values and low average SDM values. Sixty-nine families with intermediate correlation co-efficient values between 0.2 to 0.6 also include some of the protein families that have high-sequence identity among the pairs but not as high correlation between sequence-based and structure-based phylogenetic trees. Such families include members that bind metal ions or other small molecules like 2Fe-2S ferredoxins, Nitrogenase iron protein-like, cold shock DNA-binding domain-like, fatty acid binding protein-like, retinol binding protein-like, RecA protein-like (ATPase-domain) and DnaQ-like 3′-5′ exonuclease. In most of these examples the structural similarities seem to be influenced by presence of non-protein atoms (such as metal ions)

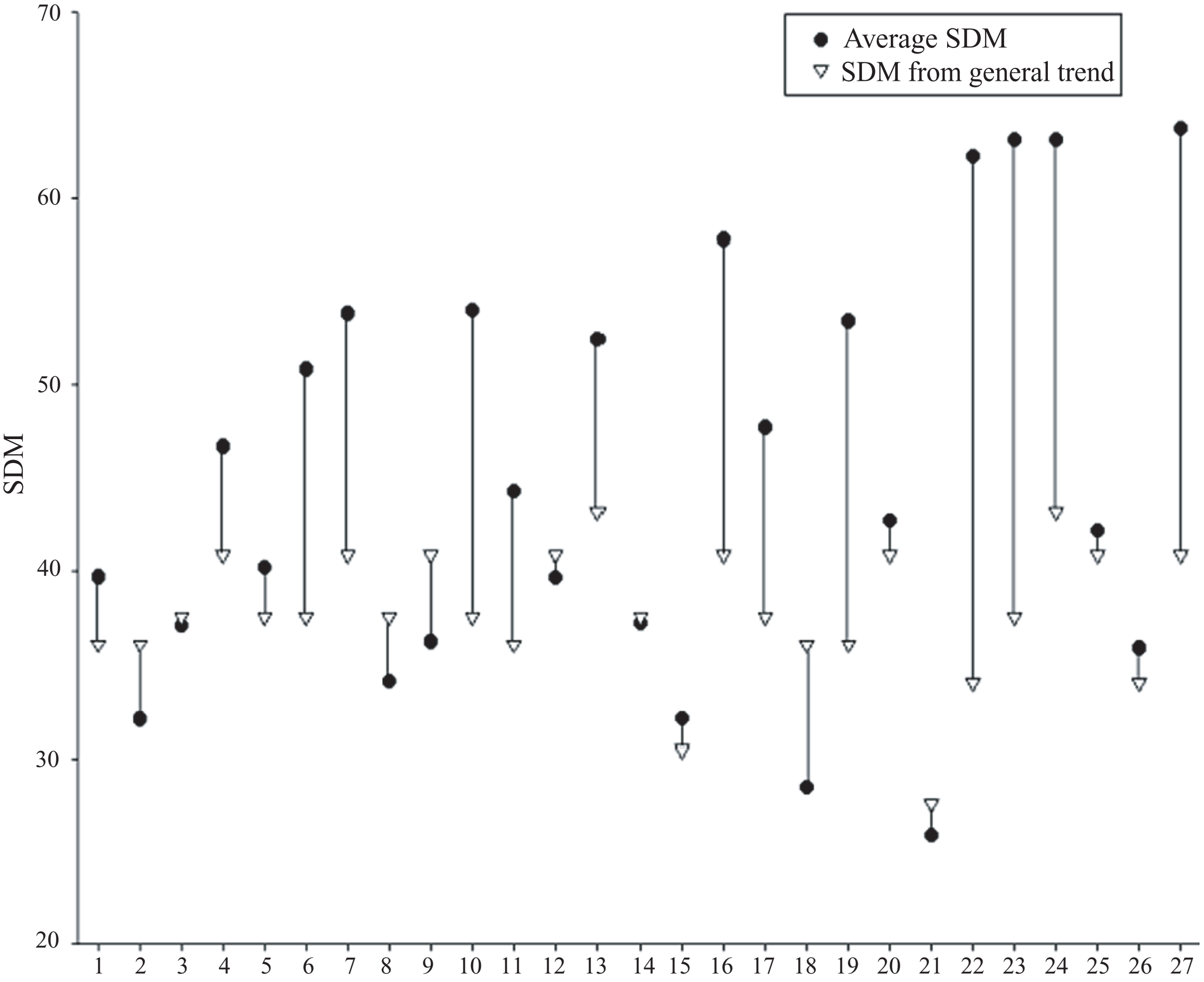

Figure 3. Scatter plot depicting the deviation of average SDMs (marked as filled circles) for 27 families from the one (marked as triangles) suggested by the general trend figure 1c. These 27 families show good correlation (greater than 0.6) between sequence-based and structure-based phylogenetic trees.

Table 1. List of families with correlation coefficient greater than 0.6 in same order as in figure 3.

Families with correlation coefficient greater than 0.6
Phycocyanin-like
Fe, Mn superoxide dismutase (SOD), N-terminal domain
Homeodomain
Cytochrome c peroxidase-like
Cu, Zn superoxide dismutase-like
Crystallins/Ca-binding development proteins
SH3-domain
Eukaryotic proteases
Triosephosphate isomerase (TIM)
Reductases
Glutathione S-transferases, N-terminal domain
Superantigen toxins, C-terminal domain
Lactate
Plant cytotoxins
C-type lysozyme
MHC antigen-recognition domain
Papain-like
Fe, Mn superoxide dismutase (SOD), C-terminal domain
GAPDH-like (glyceraldehyde-3-phosphate dehydrogenase-like)
FAD/NAD-linked reductases, dimerization (C-terminal) domain
Matrix metalloproteases, catalytic domain
Insulin-like
Animal Kazal-type inhibitors
Spider toxins
Classic zinc finger, C2H2
Zn2/Cys6 DNA-binding domain
Defensin

and, therefore, slight discrepancy exists in the relationship between sequence divergence and structural divergence which is reflected in the lower correlation coefficient. It is important to note that nearly one-third, 22 out of 69 families have correlation coefficient, close to 0.2 (less than 0.28).

3.2 *Families with poor correlation between sequence-based and structure-based phylogenetic trees*

Twelve families have correlation coefficient less than 0.2, which are considered as extreme cases of poor correspondence between sequence-based and structure-based phylogenetic tree structures. Table 2 lists the families, correlation coefficient, average SDM, average %I, average SDM as suggested by the general trend (figure 1b) for these 12 families. Expected SDM values have been adopted from the general trend of SDM variations as a function of percentage sequence identity (%I) (figure 1b). The value shown within the bracket is the standard deviation of correlation coefficient evaluated by bootstrap method.

3.2a *Poor correlation with high sequence identities:* It could be noted from table 2 that, while the majority of the families have low average sequence identities, three families have high average %I but have poor correlation coefficient. These families are either from multi-domain or small protein classes. Further investigations on these three families revealed possible reasons for the poor correlation between structure-based and sequence-based phylogenetic trees. In the case of multi-domain system, serpins, the poor correlation could be due to large variations in the relative orientation of the two domains even among its closely related members. An example shown in figure 4 which depicts the superposed structures of two serpins with highly similar amino acid sequences (88%) but with an observed SDM (42.45) nearly twice (21.29) of the one suggested by the general trend (figure 1c). The SDM values when the two domains are superimposed separately are 23.6 and 21.4. Further, it is interesting to note that the serpins also have 85% common substructure. Despite that difference in the spatial orientation of the two structural domains can be seen. It is interesting to note that none of the three multi-domain protein families, *β*-lactamase/D-Ala carboxypeptidase, sugar phosphatases and serpins, considered in this analysis have correlation better than 0.6. These suggests that the poor correspondence indeed could stem from inter-domain dynamics in multi-domain protein families.

For small protein families, conotoxins and small kunitz-type inhibitors, the average sequence length is about 25 residues and about 60 residues respectively. Due to the small size of proteins in these two families there could be a large structural dispersions due to inherent flexibility in the structures. Moreover for such disulfide-rich systems it could be expected that the formation of disulfide bonds is a major dominant factor that determines the maintenance of the stable folded state in addition to the hydrophobic effect. Hence, conotoxins and small kunitz–type inhibitors may lack a hydrophobic core similar to protein domains of bigger size. Hence in these circumstances it could be concluded that it is appropriate to model the evolution in these families using structure-based phylogenetic trees.

3.2b *Poor correlation with low sequence identities:* Nine families with poor correlation coefficient given in table 2 fall in various classes of protein structures namely *α*, *β*, *α*/*β* and *α*+*β*. These families have low average %I and highly deviating average SDM from the average value (see table 2). Further analysis of the origins of poor correlation in each of these 9 families has been made. Table 3 lists

Table 2. Families with poor correlation between sequence-based and structure-based phylogenetic trees shown below with the correlation co-efficient, mean %I and mean SDM.

Family	Average %I	Average SDM	Average SDM estimated from the general trend	Correlation coefficient
Short chain cytokines	13.3	106.2	60.3	0.09 (0.03)
Calmodulin-like	19.5	113.0	51.6	0.19 (0.01)
Immunoglobulin E-set domains	12.2	101.0	60.3	0.16 (0.02)
Plant-virus proteins	12.7	126.7	60.3	0.14 (0.03)
PH domain	15.5	95.3	51.6	0.19 (0.03)
L-arabinose-binding domain	15.8	109.8	51.6	0.17 (0.03)
Phosphate-binding domain-like	14.0	116.0	60.3	0.12 (0.01)
Extended AAA ATPase domain	14.6	131.8	60.3	0.09 (0.02)
Aminoacyl t-RNA synthethase Class II	16.4	106.4	51.6	0.15 (0.02)
Serpins	35.4	48.6	37.5	0.10 (0.05)
Conotoxin	48.9	43.4	34.0	0.18 (0.12)
Small Kuntiz type (BPTI) inhibitors	40.3	36.1	36.1	0.0 (0.1)

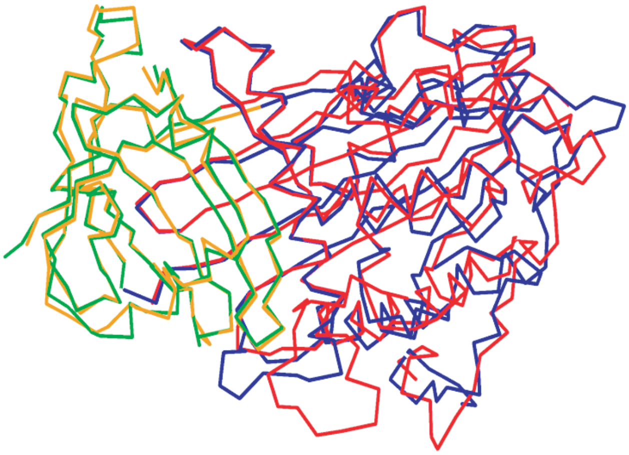

Figure 4. Illustration of variations in the relative spatial disposition of domains of two homologous proteins with high sequence similarity. Two homologous structures of Serpins are shown with optimal match of C$^{\alpha}$ positions. The two structural domains in each structure are highlighted in different colours. This figure was prepared using Setor (Evans 1993).

these structural families and their further classification in the Pfam sequence family database (Bateman *et al* 2000). The potential reasons for the poor correlation between sequence-based and structure-based trees are discussed for the individual families below.

(i) *L-arabinose-binding domain-like*: This structural family has been further classified into two sequence/functional families in the Pfam database (Bateman *et al* 2000) namely periplasmic-binding proteins and anf-receptor-like. Periplasmic-binding protein family constitutes of mainly oligosaccharide-binding domains, while the anf-receptor-like family has proteins that bind to peptides/amino acids. This structural family could be considered as a “sequence superfamily” of two sequence/functional families. Fourty-eight SDMs out of 78 possible SDMs corresponding to pairwise alignments are greater than 100 (SDM of 100 are usually encountered in the cases of sequence identity less than 5%) and almost all of them correspond to comparisons across the two sequence families. The correlation coefficient computed between the sequence-based and structure-based distance matrices individually for these sequence families yielded a much improved correlation coefficient of 0.56 for periplasmic binding domains and 0.38 for anf-receptor-like domain families. This contrasts with a poorer correlation coefficient of 0.17 if all the proteins within the structural family are taken together. Thus the structure-based phylogenetic tree clusters the two sequence/function-based families separately and hence the structure-based phylogenetic tree models the evolution better than the sequence-based phylogenetic tree.

(ii) *Class II amino acyl t-RNA synthetase (catalytic domain)*: This family of 16 members are involved in a grossly similar function of being t-RNA synthatase. However according to Pfam this structural family could be split further into tRNA synthetase class II core domain (G, H, P, S and T), tRNA synthetases class II core domain (F), tRNA synthetases class II (D, K and N) and aspartate-ammonia ligase sequence families (table 3). 34 out of 120 pairs have SDM greater than 100, which correspond to across sequence/functional families comparisons. For the two subfamilies, tRNA synthetases class II (D, K and N) and tRNA synthetase class

Table 3. Sequence families within structural families with poor correlation between sequence-based and structure-based phylogenetic trees.

Structural family	Sequence families (functional families) in Pfam classification
Short chain cytokines	Interleukin 2
	Interleukin 4
	Stem cell factor
	FLT3 ligand
	Granulocyte macrophage colony stimulating factor
	Interleukin 5
	Interleukin 13
	Interleukin 3
	Erthropoietin
Calmodulin-like	EF hand
	Sacroplasmic calcium binding protein-like*: this includes the calcium binding proteins in obelin
Immunoglobulin E-set domains	TIG/IPT family of cell surface receptors and DNA binding domains
	Peptidase family C25 C terminal ig-like domain
	Alpha amylase C-terminal all-beta domain
	ML domain-MD-2-related lipid recognition domain
	Filamin/ABP280 repeat
	BNR/Asp-box repeat
L-arabinose binding domain-like	Periplasmic binding proteins
	Anf-receptor-like
Phosphate binding domain-like	Porphobilinogen deaminase, dipyromethane cofactor binding domain
	Phosphate-binding protein
	Bacterial extracellular solute binding proteins; Family 1 and family 3
	Sulphate binding proteins/thiosulphate binding proteins
Extended AAA ATPase domain	ADP binding domain-like #: this family includes the proteins that have their structures determined in complex with ADP or adenine
	DNA polymerase III-like #: this family includes the proteins that are Zn and SO_4 co-ordinated ATP binding domains
Class II aminoacyl t-RNA synthetases	tRNA synthetase class II core domain (G, H, P, S and T): Gly, His, Pro, Ser and Thr tRNA synthetases core catalytic domains
	tRNA synthetases class II core domain (F): Phenylalanine tRNA synthetase core catalytic domain
	tRNA synthetases class II (D, K and N): Asp, Lys, and Asn tRNA synthetases core catalytic domains
	Aspartate-ammonia ligase
Plant virus proteins	Proteins in this family represent various virus families (subfamilies) ##: Bromoviridae, Satellites, Sobemoviridae, Necroviridae, Reoviridae, Tymoviridae and Comoviridae. This family has proteins from different viral classes like (+) sense RNA viruses, dsRNA viruses and satellites
PH domain	Highly diverged family within the beta barrel framework. Some of the PH domains are known for binding to different phospholipids with variation in the region of binding to phospholipids. A few PH domains are known to be translocated to the membrane by binding to *β*,*γ*-subunits of heterotrimeric G-proteins.

In table 3, * classification is not according to Pfam database (based on SCOP). # Classification is made on the basis of the ligands binding to the domains (it is not according to Pfam classification). ## Based on classification of the virus in which the proteins exist.

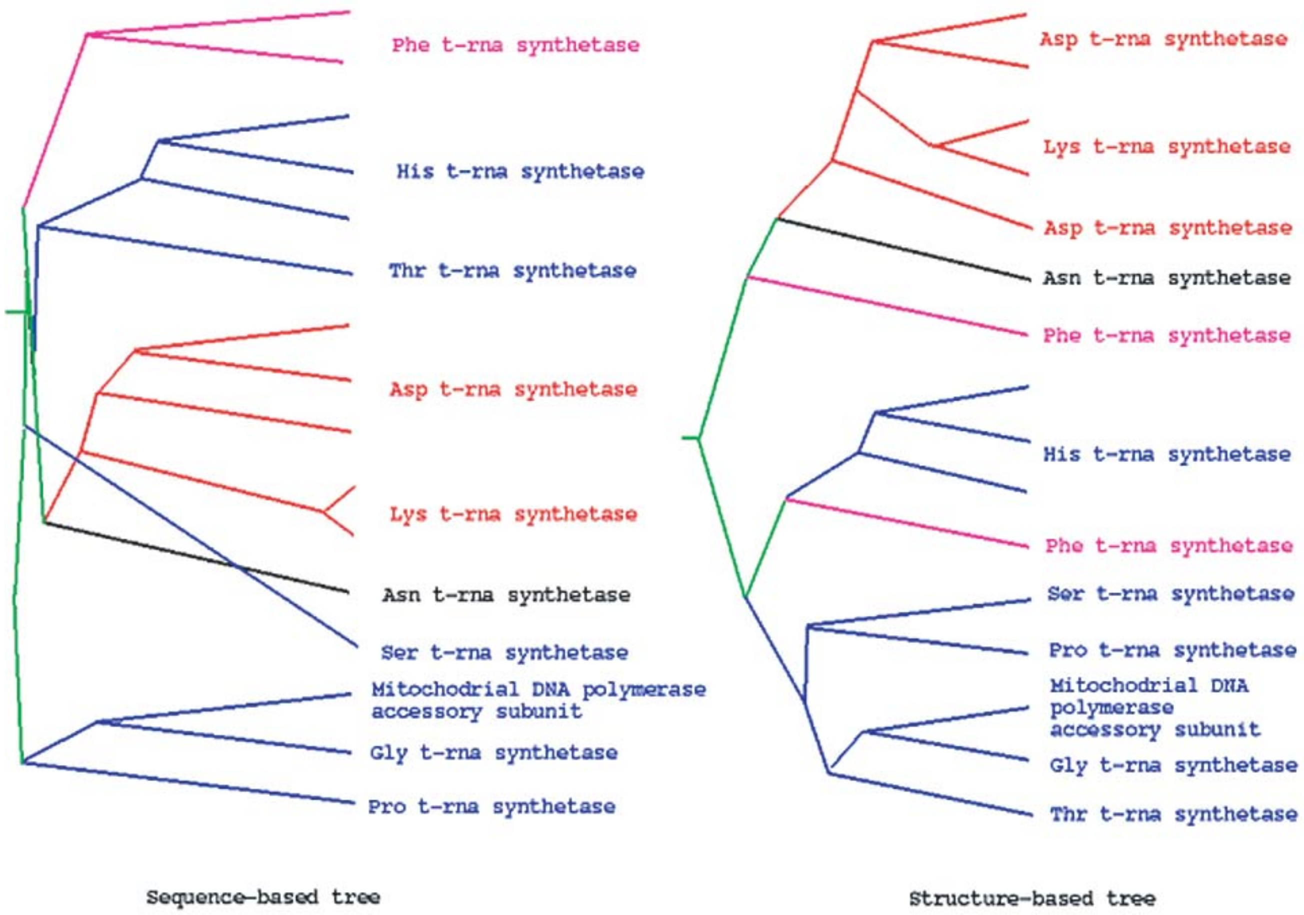

Figure 5. Dendrograms depicting the sequence-based and structure-based phylogenetic relationships for class II amino acyl t-RNA synthetase (catalytic domain) family. The members belonging to same Pfam family are in same colour.

II core domain (G, H, P, S and T), the correlation coefficients are 0.21 and 0.15 respectively. It is interesting to note that tRNA synthetase class II core domain (G, H, P, S and T) did not show better correlation between the sequence and structure-based trees. However further sub-divided families tRNA synthetase class II core domian (G, P, S and T) and tRNA synthetase class II core domain (H) gave better correlations of 0.69 and 0.74 between the corresponding sequence and structure-based trees. Hence the correlation in sequence/functional families are higher than the correlation for the entire structural family considered. Figure 5 shows the sequence-based and structure-based phylogenetic trees of the family constructed using PHYLIP suite of software (Felsenstein 1995). It could be deduced from the figure that the clustering of protein in to subfamilies in the structure-based phylogenetic tree is closer to detailed function-based classification than in the sequence-based tree.

(iii) *Phosphate binding protein-like*: This family has 17 members could be subdivided in to 6 sequence/functional families of Pfam classification as listed in table 3. This family includes a spectrum of grossly similar functional sequence/functional families ranging from solute binding proteins to sulphate and phosphate binding domains. Intricate functional/sequence diversity is evident from the fact that 84 pairs out of possible 136 pairs in the family have SDM greater than 100 which corresponds to comparisons across the subfamilies (sequence families). The correlation coefficient for the subfamily of bacterial extracellular solute binding proteins is improved to 0.39 as opposed to 0.12 for the entire structural family. Thus the structure-based clustering (phylogenetic tree) groups proteins which make subfamilies in terms of higher similarity in terms of function.

(iv) *Calmodulin-like*: This family is second largest of the 9 families that has the poor correlation between the structure-based and sequence-based phylogenetic trees. The family could be further divided to two sequence families in Pfam as EF-hand and sacroplasmic calcium binding protein-like as there is low sequence similarity and large SDM of greater than 100 across the two sequence families. The correlation coefficients computed for these sequence families individually are 0.21 and 0.54, which are higher than the correlation coefficient of the entire structural family (0.15). This validates the further classification of EF-hand-like family in to two sub families. This forms another example which emphasises that functional diversity could be better modelled using structure-based phylogeny than the sequence-based phylogeny.

(v) *Immunoglobulin E-set domains*: This forms the largest family among the families that have poor correlation between sequence and structure-based phylogenetic trees. Twenty-six domains in the family could be further classified in to 8 sequence families according to Pfam classification. The eight families include TIG/IPT family of cell surface receptors and DNA binding domains, Peptidase family C25 C terminal Ig-like domain and α-amylase C-terminal all-β domain sequence families of the Pfam classification (see table 3). The correlation coefficient evaluated for the largest of the sequence families (TIG/IPT family, 9-member subfamily) between the sequence-based and structure-based phylogeny is 0.92. Such a high correlation at the level of structural subfamilies or sequence families emphasises that general poor correlation between sequence-based and structure-based trees can occur when homologues of low sequence similarity with a subtle variation in the functional properties between sub-clusters are grouped together as a family.

(vi) *Short chain cytokines*: Table 3 shows the protein names and functions of all the 10 members in the family. As it could be seen that functions of the members are grossly similar as all of them are involved in cellular signalling events to effect growth and activation by binding to their receptors. Nearly 90% of SDMs are greater than 80 units, which corresponds to very low sequence identity pairs (see figure 2). Almost every member in this family is classified in to different sequence families in the Pfam database (see table 3), based on sequence/functional similarity. Hence it could be concluded that this family is tending to a 'superfamily' context of functionally related proteins with similar fold with low sequence similarity. Thus it appears that, for this family, structure-based phylogenetic tree would be an appropriate descriptor of protein evolution

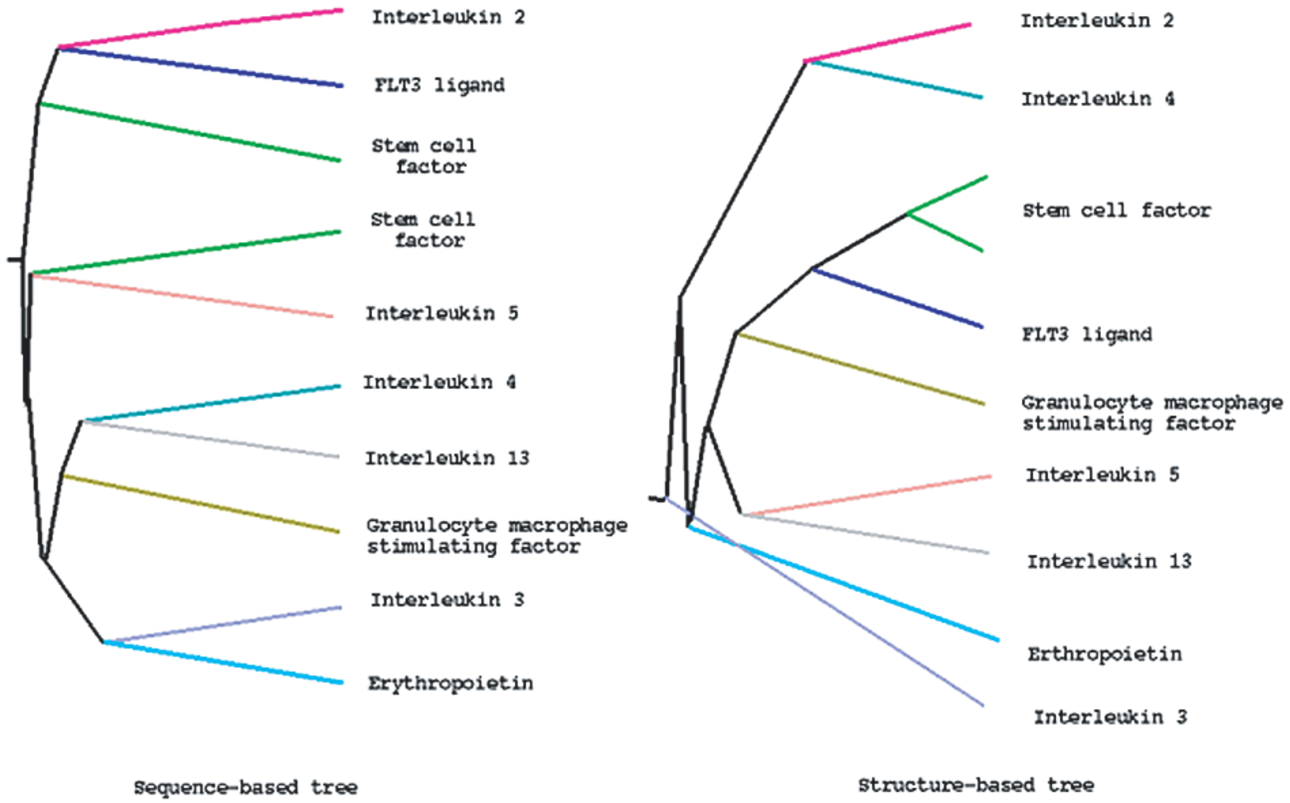

Figure 6. Dendrograms depicting the sequence-based and structure-based phylogenetic relationships for short chain cytokines. The members belonging to same Pfam family are in same colour.

rather than the sequence-based phylogenetic tree. Figure 6 shows the sequence-based and structure-based phylogenetic trees of this family constructed using PHYLIP suite of software it could be evident from the trees the clustering of proteins are quite different in sequence-based and structure-based trees. It is interesting to note that the interleukin 2 and interleukin 4 are distantly related but the evolutionary relationship is not detectable in the sequence-based phylogeny but it is evident in the structure-based phylogeny.

(vii) *Extended AAA ATPase domain*: By classifying the domains in this structural family in to subfamilies based on the similarity of ligands they bind (have their structures experimentally determined complexed with them) these domains could be classified in to ADP binding domain-like and DNA polymerase III-like. But such a classification lead to practically negligible increase in correlation for ADP binding domain-like subfamily but no increase in correlation between DNA polymerase III-like family between corresponding sequence and structure-based phylogenetic trees. It is interesting to note that significant majority of the members in the family have less than 20% sequence identity between themselves so it is an example of a family which has high divergence. Moreover the influence of the non-proteins atoms on the protein structures also could play a major in the poor correlation between sequence and structure-based phylogenetic trees (Balaji and Srinivasan 2001) as most of the members in the family have their structures determined in the complexed form with either adenosine-5′-diphosphate or adenine or 5′-adenyly-imido-triphosphate or Zn ions or sulphate ions

(viii) *Plant virus proteins*: Members of this family come from different viral families like Bromoviridae, Satellites and Reoviridae (see table 3). It is interesting to note that 43 out of 45 possible pairs have sequence identity of less than 30%. Moreover 35 pairs have less than 20% sequence identity. Hence extent of divergence in the sequences is higher than in the corresponding structures. The evolutionary relationship in the family of plant virus proteins is practically undetectable in their sequences. But in order to model the evolution of plant viruses it is inevitable to consider the similarities in the nucleic acid enclosed in the viral capsid and the nature of the assembly of the capsid itself. Hence the structure-based phylogenetic relationship coupled with other phylogenetic relationships dependent on parameters like those mentioned above would serve as a better model for the evolution of plant virus.

(ix) *PH-domain*: This family is made of distantly related members. Nearly half the number (36 out of 78 pairs) of pairs have SDM of greater than 100 units and about 60 pairs have sequence identity of less than 20%. It is also important to note that significant number of structures have their structures determined by NMR technique and in the current analysis single representative from the ensemble of structures has been chosen at random. It is noted that within the ensemble of structures of PH domains there could be quite large deviations leading to about 45 units of SDM between the structures within the ensemble. This in turn could lead to poor correlation between sequence and structure-based phylogenetic trees.

3.2c *Structure-based phylogeny of superfamilies:* Comparison of sequence-based and structure-based phylogenetic tree structures in the previously described 9 families suggest that there could be inherent sequence families that are forming the “sequence superfamily” and in these situations structure-based phylogenetic trees model the evolution better than sequence-based phylogenetic trees. As a sequel to the above mentioned conclusion it would be appropriate to extend and model the evolution of structural superfamily by the structure-based phylogeny. In the case of superfamilies the evolutionary relationship for the members across the families within a superfamily is practically undetectable in their sequences. Such a tree is presented for the case of 4Fe-4S ferredoxins superfamily. The structure-based phylogenetic tree is shown in figure 7 for this superfamily with three families in it. It could be noted from the figure that the members within the same family cluster together at a node. Hence it is suggested that the structure-based phylogeny model the evolution better than the sequence-based phylogeny within structural/sequence superfamilies that is in the case of distantly related proteins.

4. Discussion

Comparison of sequence and structure-based phylogenetic trees leading to high correlation coefficient are usually encountered in high sequence similarity families (above about 30% sequence identity) and the analysis on the families with poor correlation between structure-based and sequence-based trees suggest that there are:

(i) Inherent sequence superfamily relationship within structural families for low sequence similarity families.
(ii) Size of the proteins has an effect on the correlation due to inherent flexibility in the structures due to small size of the proteins in the case of small proteins and potential domain movements in the proteins in the case of multi-domain systems.

Hence structure-based phylogenetic trees model the evolution better than sequence-based phylogenetic trees for these families. The structure-based phylogenetic tree for a 4Fe-4S ferredoxins structural superfamily lead to

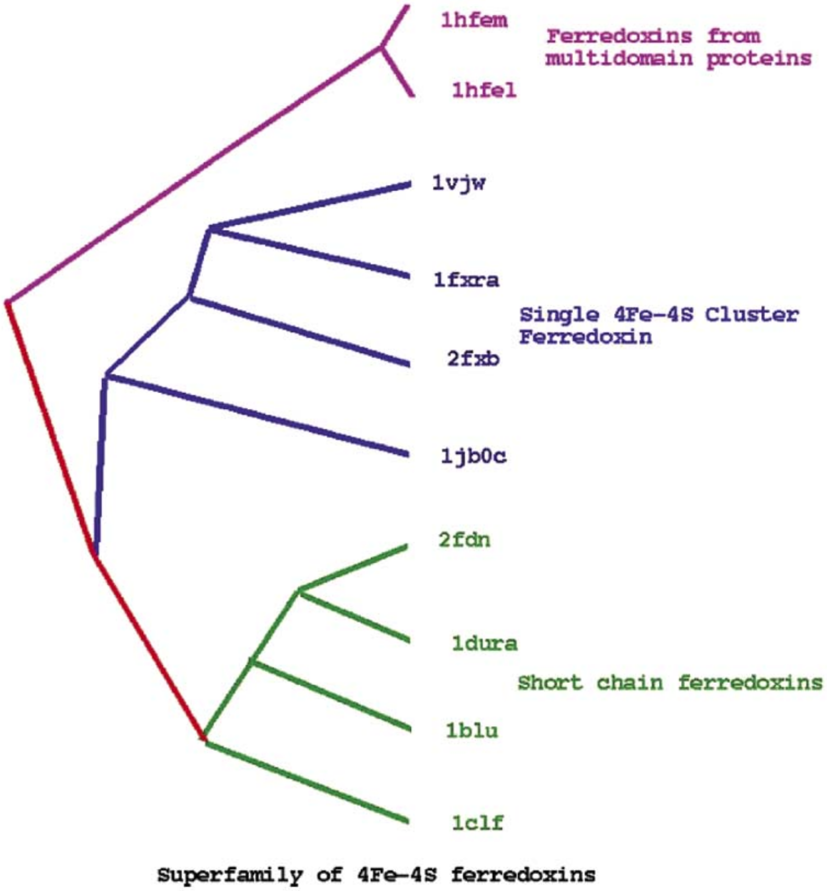
1hfem
1hfel
Ferredoxins from multidomain proteins
1vjw
1fxra
2fxb
Single 4Fe-4S Cluster Ferredoxin
1jb0c
2fdn
1dura
Short chain ferredoxins
1blu
1clf
Superfamily of 4Fe-4S ferredoxins

Figure 7. Dendrogram showing the clustering of families within the superfamily of 4Fe-4S Ferredoxins. The members belonging to same superfamily are in same colour.

clustering of families together hence could be also be used to model the evolution of the structural superfamilies. This idea could be extrapolated to study the evolution of protein with no significant sequence similarities within a common fold. We already made an attempt along this direction on some of the highly populated folds listed in the SCOP database. The outcome of the analysis suggests that the members within the same superfamily usually cluster together in the structure-based phylogenetic trees (S Balaji, S Chakrabarti, R Sowdhamini and N Srinivasan, manuscript in preparation).

The proposed approach of using 3-D structures to model the evolution of proteins is particularly sensitive if the sequence divergence between the homologues is extensive and under such a circumstance it is inappropriate to use the traditional sequence similarity based approach to model the evolution. Given the distant relatedness between homologous proteins the structure-based approach appears to be very appropriate to identify specific features of protein evolution. For example, it is common to find two distinct sets of homologous proteins, of high sequence divergence, from same set of organisms. These sets of proteins with suggested putative functional link between the members of the two families may be studied for parallel evolution. The approach can be very much like the traditional sequence similarity based method, but, one might use 3-D structures and SDM values to compare the phylogentic profiles of the two distinct sets of homologues each characterized by poor average sequence similarity. Under a situation of parallel evolution, it is expected to find similar phylogentic profiles of two sets of highly divergent homologues if the phylogentic profiles are calculated using SDM values.

Characterization of orthologous and paralogous proteins demands understanding of functions and biological roles of proteins concerned in great detail. Identification of paralogous proteins is eased by the occurrence of the homologues in the same organism. If the paralogues are characterized by poor sequence similarity their evolutionary relatedness can not be easily discerned from the measures of sequence similarity. However clustering of these proteins on the basis of 3-D structural similarities is expected to indicate functional differences as suggested from the structure-based phylogeny at the superfamily level (figure 7) in which members within families, in the superfamily, characterized by high similarity in functions are clustered together although the gross functions of members of different families in the superfamily can be same. Identification of putative orthologues of common 3-D fold, if they are diverged extensively, is expected to be possible by the SDM-based clustering of proteins of closest functional similarity. Our future applications of structure-based method to model phylogeny would include characterization of orthologous and paralogous proteins.

Acknowledgements

This research has been supported by the award of International Senior Fellowship in biomedical sciences to NS by the Wellcome Trust, London and by the computational genomics project from the Department of Biotechnology, New Delhi. SB was supported by a fellowship from the Council of Scientific and Industrial Research, New Delhi and by the Wellcome Trust during the course of this work.

References

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

Bray J E, Todd A E, Pearl F M, Thornton J M and Orengo C A 2000 The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues; *Protein Eng.* **13** 153–165

Bujnicki J M 2000 Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures; *J. Mol. Evol.* **50** 39–44

Chothia C and Lesk A M 1986 The relation between the divergence of sequence and structure in protein; *EMBO J.* **5** 823–826

Doolittle R F 1981 Similar amino acid sequences: chance or common ancestry?; *Science* **214** 149–159

Efron B 1979a Bootstrap methods: Another look at the jackknife; *Ann. Stat.* **7** 1–26

Efron B 1979b Computers and Theory of Statistics:Thinking the Unthinkable; *SIAM Rev.* **21** 460–480

Evans S V 1993 SETOR: hardware-lighted three-dimensional solid model representations of macromolecules; *J. Mol. Graph.* **11** 127-128, 134–138

Felsenstein J 1995 *PHYLIP (Phylogeny Inference Package) version 3.57*c (Department of Genetics, University of Washington, Seattle, USA)

Flores T P, Orengo C A, Moss D S and Thornton J M 1993 Comparison of conformational characteristics in structurally similar protein pairs; *Protein Sci.* **2** 1811–1826

Goh C S, Bogan A A, Joachimiak M, Walter D and Cohen F E 2000 Co-evolution of proteins with their interaction partners; *J. Mol. Biol.* **299** 283–293

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

Grishin N V 1997 Estimation of evolutionary distances from protein spatial structures; *J. Mol. Evol.* **45** 359–369

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 1997 An evolutionary treasure: unification of a broad set of amidohydrolases related to urease; *Proteins: Struct. Funct. Genet*. **28** 72–82

Hubbard T J and Blundell T L 1987 Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modeling**;** *Protein Eng.* **1** 59–71

Johnson M S, Overington J P and Blundell T L 1993 Alignment and searching for common protein folds using a data bank of structural templates; *J. Mol. Biol*. **231** 735–752

Johnson M S, Sali A and Blundell T L 1992a Phylogenetic relationships from three-dimensional protein structures; *Methods Enzymol*. **183** 670–690

Johnson M S, Sutcliffe M J and Blundell T L 1992b Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins; *J. Mol. Evol.* **1** 43–59

Lesk A M and Chothia C 1980 How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins; *J Mol. Biol.* **136** 225–270

Murzin A G 1993a Can homologous proteins evolve different enzymatic activities?; *Trends Biochem. Sci.* **18** 403–405

Murzin A G 1993b Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors; *J. Mol. Biol.* **230** 689–694

Murzin A G 1998 How far divergent evolution goes in proteins?; *Curr. Opin. Struct. Biol.* **8** 380–387

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

Pazos F and Valencia A 2001 Similarity of Phylogenetic trees as indicator of protein-protein interaction; *Prot. Eng.* **14** 609–614

Russell R B and Barton G B 1992 Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels; *Proteins: Struct. Funct. Genet.* **14** 309–323

Russell R B and Barton G J 1994 Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility; *J. Mol. Biol.* **244** 332–350

Russell R B and Sternberg M J 1996 A novel binding site in catalase is suggested by structural similarity to the calycin superfamily; *Protein Eng.* **9** 107–111

Russell R B and Sternberg M J 1997 Two new examples of protein structural similarities within the structure-function twilight zone; *Protein Eng.* **10** 333–338

Russell R B, Saqi M A, Sayle R A and Sternberg M J 1997 Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation; *J. Mol. Biol.* **269** 423–439

Sowdhamini R, Burke D F, Huang J F, Mizuguchi K, Nagarajaram H A, Srinivasan N, Steward R E and Blundell T L 1998 CAMPASS: a database of structurally aligned protein superfamilies; *Structure* **6** 1087–1094

Sowdhamini R, Rufino S D and Blundell T L 1996 A database of globular protein structural domains: clustering of representative family members into similar folds; *Fold. Des.* **1** 209–220

Sujatha S, Balaji S and Srinivasan N 2001 PALI: a database of alignments and phylogeny of homologous protein structures; *Bioinformatics* **17** 375–376

Todd A E, Orengo C A and Thornton J M 2001 Evolution of function in protein superfamilies, from a structural perspective; *J. Mol. Biol*. **307** 1113–1143

*e*Publication: 12 December 2006