Interaction Interfaces of Protein Domains Are Not Topologically Equivalent Across Families Within Superfamilies: Implications for Metabolic and Signaling Pathways

N. Rekha,[1] S.M. Machado,[2] C. Narayanan,[3] A. Krupa,[1] and N. Srinivasan[1]*
[1]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India*
[2]*Institute of Bioinformatics and Applied Biotechnology, International Technology Park, Bangalore, India*
[3]*Centre for Biotechnology, Anna University, Chennai, India*

ABSTRACT **Using a data set of aligned protein domain superfamilies of known three-dimensional structure, we compared the location of interdomain interfaces on the tertiary folds between members of distantly related protein domain superfamilies. The data set analyzed is comprised of interdomain interfaces, with domains occurring within a polypeptide chain and those between two polypeptide chains. We observe that, in general, the interfaces between protein domains are formed entirely in different locations on the tertiary folds in such pairs. This variation in the location of interface happens in protein domains involved in a wide range of functions, such as enzymes, adapters, and domains that bind protein ligands, or cofactors. While basic biochemical functionality is preserved at the domain superfamily level, the effect of biochemical function on protein assemblies is different in these protein domains related by superfamily. The divergence between proteins, in most cases, is coupled with domain recruitment, with different modes of interaction with the recruited domain. This is in complete contrast to the observation that in closely related homologous protein domains, almost always the interaction interfaces are topologically equivalent. In a small subset of interacting domains within proteins related by remote homology, we observe that the relative positioning of domains with respect to one another is preserved. Based on the analysis of multidomain proteins of known or unknown structure, we suggest that variation in protein–protein interactions in members within a superfamily could serve as diverging points in otherwise parallel metabolic or signaling pathways. We discuss a few representative cases of diverging pathways involving domains in a superfamily. Proteins 2005;58:339–353.** © 2004 Wiley-Liss, Inc.

Key words: homologous proteins; molecular recognition; protein domain superfamilies; protein evolution; protein–protein interactions

INTRODUCTION

Interactions between protein domains play essential roles in almost all metabolic and signaling pathways. Multimeric proteins are found in every cellular location, including cell organelles, the cytosol, and cell membranes. Interactions between different proteins are critical in all known cellular processes, such as signaling and transcriptional regulation, metabolism of small and large molecules, protein folding assisted by chaperones, enzyme regulation and allostery, immune response, and control of cell regeneration.[1,2] The delicate balance of the cellular metabolites and chemical entities or environmental stimuli is known to govern a number of interactions between proteins. Thus, the study of the quaternary structure of a protein yields valuable information on its biological active state and its overall role in the functionality of the cell.

Attempts have been made to study the structural and chemical aspects of protein–protein interfaces based on the available three-dimensional (3D) structures. Some of the methods aim at predicting protein–protein interaction interfaces from the tertiary structures based on the surface features such as geometry[3] and energetics of association.[4] Analysis of the composition of residue types at the tertiary surfaces of closely related family of proteins[5–8] gives insight into the chemical nature of the interfaces.

During the course of evolution, the sequences diverge, retaining only those residues that are necessary for the protein to function.[9] This might hold true for residues involved in protein interfaces too, as observed in the cases of SH2 and SH3 domains, where the surface residues

---

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: Wellcome Trust, UK, in the form of International Senior Fellowship in Biomedical Sciences (to N. Srinivasan). Grant sponsor: Computational Genomics Project funded by the Department of Biotechnology, India. Grant sponsor: CSIR, India, in the form of fellowships to N. Rekha and A. Krupa.

*Correspondence to: N. Srinivasan, Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India. E-mail: ns@mbu.iisc.ernet.in

Received 13 April 2004; Accepted 2 August 2004

Published online 23 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20319

involved in binding to other proteins are better conserved than the other residues.[9,10] Evolutionary information derived from the large number of protein sequences and structures can powerfully guide both the analysis and prediction of protein–protein interfaces. The residue conservation at the interface among close homologues is statistically significant.[11] Fraser et al.[12] demonstrated that the interacting proteins in *Saccharomyces cerevisiae*, when compared with their orthologues from *Caenorhabditis elegans*, evolve at similar rates. Recent studies, however, have shown that the interacting proteins in closely related genomes have a slight influence on the rate at which these proteins diverge.[13] Additionally, Jordan et al.[14] have shown that the proteins involved in a large number of interactions, thus acting as hubs of an interaction network, tend to evolve more slowly than the bulk of the other proteins in the genome.

Since the tertiary fold and gross functionality are preserved in several remote homologues [referred to as "superfamily" according to the Structural Classification of Proteins (SCOP) definition[15]], it would be important to study the extent of topological equivalence of protein–protein interaction in these remotely related proteins. Thornton and coworkers[16] have shown that in 23 out of 167 superfamilies of enzymes, the subunit assembly varies between the constituting members. Aloy et al.[17] have shown that the root-mean-square deviation for the structural superposition of protein–protein interfacial regions is high if the sequence identity between the similarly folded proteins is lower than about 30%. However, it is not clear if the locations of protein–protein interactions among superfamily members are in the equivalent surfaces in the tertiary structures. Thus, in this article, we address the next level of the question: Are the protein–protein interaction interfaces between the domains related by the superfamily topologically equivalent? Further implications of the variation in the protein–protein interfaces to the functions of the superfamily related proteins are also investigated.

The current analysis is carried out at the protein-domain level, as defined in SCOP. One would be able to define interfaces between two domains in a multidomain protein (interdomain interface) and between two different polypeptide chains in the same structure (interchain interface). During the course of the present analysis, and as noted earlier,[5] it was observed that the interfacial properties of interdomain interfaces and interchain interfaces show similarities. Hence, the data sets of two types of interfaces are pooled together and are here commonly referred as "interdomain interface." In this work we demonstrate that, in general, the residues involved in the interaction between protein domains in members across the families within a superfamily are not topologically equivalent, and the interaction interfaces are in distinct locations in the tertiary structures. We then present an analysis of remotely related sequence domains, without the knowledge of 3D structure, and explore the implications of remotely related sequence domains in rendering variations in their biochemical roles and interaction properties. We argue that distinct protein–protein interaction interfaces across protein families in a superfamily, as observed in the domains of known structure, imply variations in the metabolic and signaling pathways, and in gene interaction networks in general.

MATERIALS AND METHODS

The Data Set

The data set consists of alignments between proteins of known structures belonging to different families within a superfamily and is obtained from the database of Protein Alignments organized as Structural Superfamilies (PASS2)[18] (Supplementary Data 1). PASS2 is a nearly automated version of Cambridge database of Protein Alignments organised as Structural Superfamilies (CAMPASS) and contains sequence alignments of proteins belonging to different families within a superfamily.[19] This database has been designed to be in correspondence with the database of SCOP.[15]

The release of the PASS2 database used in the current analysis consists of 110 superfamilies that have more than one member (multimember superfamilies, MMS) and 613 superfamilies consisting of only one family in each superfamily ("orphans" or single-member superfamilies, SMS). Of the Protein Data Bank (PDB) entries, 374 entries are distributed across 110 MMS. In MMS, protein chains with no more than 25% sequence identity have been considered for the alignment. The alignments are generated by using COMPARER[20] and are represented in the JOY format.[21]

Dimeric protein families with multiple members in each family described in SCOP[15] are considered as controls. Gross results of conservation at domain–domain interfaces across families within superfamilies are compared with the corresponding data from domain–domain interfaces within families. This list of families of dimeric proteins used in the analysis is given in Supplementary Data 2. The structural alignments of the homologous members within each family were arrived at using STAMP.[22]

The quaternary structure of the various members of the 110 MMS from PASS2 was derived by consulting the Protein Quaternary Structure file server (PQS).[23] Quaternary structure is defined as that level of form in which the units of tertiary structure aggregate to form homo- or heteromultimers. The PQS server makes available coordinate sets for probable quaternary states for structures contained in the PDB that were determined by X-ray crystallography.

PASS2 is constructed at the level of domains in the known structures. However PQS offers the quaternary structure for a PDB entry that might have more than one domain. Hence, we query PQS on the basis of the PDB code and identify the oligomeric interfaces and explore if a domain/chain of interest from PASS2 participates in the oligomer formation.

Identification of Interfacial Residues

A simple method of identification of protein–protein interfaces is defined based on the change in the solvent-accessible surface area of the residues when the accessibili-

ties are calculated at the monomeric state and the oligomeric state separately. To define this change in solvent accessibility, the program Protein Surface Accessibility (PSA), which has an implementation of the algorithm by Lee and Richards,[24] has been used.

The criterion used to define an interfacial residue is as follows: The accessibility of the residue in the complex (oligomer) form should be $\leq 7\%$, and the same residue in the isolated domain should have an accessibility of $\geq 10\%$. The accessibility of $< 7\%$ indicates the buried state, and the accessibility of $> 10\%$ represents the exposed state. This is a stringent criterion, the use of which allows us to identify the interacting residues at the middle of the interaction patches that are highly buried when bound to the other subunit and well exposed in the unbound state.

Generation of Interdomain Interfaces

For each of the proteins belonging to 110 MMS from PASS2, the corresponding chain entry from the quaternary structure file from PQS was extracted. In the case of interdomain interfaces, the domains were extracted from the quaternary structure file from the PDB using the domain definitions in the PASS2 alignment. The PSA program was run on both the extracted domain (chain/domain) and the intact complex. Using the criteria described above, the interfacial residues were identified.

The identified interfacial residues were then mapped back onto the PASS2 superfamily alignment. The MMS were analyzed further in pairs, as discussed later, and scoring parameters were defined to quantify the extent of conservation of interfacial nature.

There are three scoring schemes used in the further analysis of the selected pairs: the Conservation of Interface Location score (CIL), the Conservation of Interface Region score (CIR), and the Equivalent Secondary Structure score (ESS).

Conservation of Interface Location (CIL) score

This score indicates the presence of interface in topologically equivalent residue positions across the domains in the pair considered and is not designed to measure the extent of similarity of aligned residues in the interface. Each position in the pairwise alignment is assessed for the property of the aligned residues to be found at the interface. This score does not consider the type of the aligned residues in the topologically equivalent positions at the interface. The final score is obtained by normalizing the total number of interfacial residue positions in the topologically equivalent locations in the alignment by the total number of residues at the interface in the longer member. This score is represented as a percentage:

$$\text{CIL} = (\text{Number of residues with interfacial nature conserved})/(\text{Total number of residues in the interface of the longer protein}) * 100$$

Conservation of Interface Region (CIR) score

The superfamily-related proteins are highly divergent, and many insertions and deletions are expected in the alignments. In order to give a slight allowance for the high divergence, we have used the CIR score. This crude measure is devised to react to the fact that there exists no unique solution to the structural alignment of two proteins, especially if the the proteins are highly divergent.[25] Using this rough measure, we are now able to assess whether the interfacial locations are topologically equivalent, irrespective of potential slight misalignments. CIR defines interface location as conserved if an interfacial residue is observed in both the proteins at either the topologically equivalent position or within 5 residues up or down in the alignment. The final score is normalized for the total number of residues falling in the interface of the longer sequence of the alignment and represented as a percentage:

$$\text{CIR} = (\text{Number of interfacial residues aligned or present within plus or minus 5 positions in the alignment})/(\text{Total number of interfacial residues in the longer protein}) * 100$$

Equivalent Secondary Structure score

In the remote homologues, the relative orientation of the secondary structures may be varying markedly. The ESS score reflects the number of ESSs falling at the interfaces of both the members in the alignment. While scanning along the longer member in the alignment, when a residue at the interface is found, the ESS, if present in the other member, is scanned for one or more interfacial residues. If interfacial residues are found in the ESS, then this contributes to the score. The final score is normalized for the total number of secondary structures with at least one interfacial residue falling at the interface of the longer member and is represented as a percentage:

$$\text{ESS} = (\text{Number of equivalent aligned secondary structures participating in interface})/(\text{Total number of secondary structures involved in interface formation in the longer protein}) * 100$$

The above scores are useful in assessing the conservation of interfacial property (whether or not an interfacial position is present at an aligned position). It is important to note that the residue type is not considered in any of the above three scores.

Residue Conservation at Topologically Equivalent Interface Locations in Proteins Belonging to Family and Superfamily

The residue substitution matrices for the locations within the alignment where the interface locations are conserved were constructed. The positions where interfaces are conserved in two proteins in an alignment were identified, and the frequency of residues present at these aligned positions were calculated, thereby generating a substitution matrix for the interfacial residues. Out of 4822 total interfacial residues in the data set for superfamilies, we observe only 1025 interface conserved locations.

The interface residue substitution matrix thus obtained is normalized by considering the substitution score for the residue with another in the PASS2 data set, in general, considering all aligned positions:

$$\text{Normalized interfacial } G_{i,j} = \text{Raw interfacial } G_{i,j}/\text{Raw } G_{i,j} \text{ in full data set}$$

where $G_{i,j}$ represents a matrix element. The normalized interfacial matrix is given as Supplementary Data 3.

A similar normalized matrix was constructed also for the interfaces derived from the members of families, and this matrix is also provided in the Supplementary Data 3. The matrices for family and superfamily are then compared with a generalized residue substitution matrix, BLOSUM 62. The comparison was done by calculating a simple linear correlation coefficient between the corresponding elements of the matrices and assessing the significance of these correlations when compared to the correlation between 1000 pairs of random numbers.

Identification of Superfamily-Related Protein Domains That Have Distinct Biological Roles

SWISS-PROT[26] database comprises amino acid sequences of proteins that are experimentally studied, and their molecular functions and overall functional roles are well understood. TrEMBL[26] is a similar database, but it includes proteins that are less well studied. Pfam[27] has a collection of amino acid sequence alignments, where each alignment corresponds to a functional domain family. Sequence-based domain families in Pfam can be grouped into superfamilies[28] by relating each of the constituent members with a structural superfamily as presented in the SCOP.[15] Similarly, Pfam families with no structural information can also be grouped together into sequence superfamilies by virtue of remote sequence similarities and profile matches.[28] A compendium of superfamily-related clusters of Pfam families is made available in the SUPFAM database (http://pauling.mbu.iisc.ernet.in/~supfam[28]).

A data set of SWISS-PROT and TrEMBL proteins with Pfam domains assigned (SwissPfam) is available at the Pfam data download site (//ftp.genetics.wustl.edu/pub/pfam/). Proteins that have at least one domain related by superfamily to a domain in the other proteins from the SwissPfam data set are grouped together. We then extract small sets of two or more proteins, where at least one domain family in a protein bears a superfamily relationship with a domain family in another protein in the set. Most of the other domain families in the protein set are identical. We then assess both the molecular functions and gross implications of these proteins in pathways.

RESULTS AND DISCUSSION

Pairwise alignments were extracted from the multiple structural alignment of the members of various protein domain superfamilies that are available as an alignment database.[18,19] Out of 110 multimember superfamilies, 606 pairs of distantly related protein entries and their structure-based alignments were obtained.

The interfacial residues in these alignments were separately identified, wherever applicable, for both types of interfaces—one within the same polypeptide chain (interdomain interface) and the other, the interface formed across the chains (interchain interface). The situation of comparison of interchain interfacial residues does not arise if at least one of the members in the pair is a monomer. In other words, the domain under consideration does not participate in interchain interaction interface defined by the PQS server[23] and by our criteria (see Materials and Methods section).

Out of 606 alignment pairs that are extracted from the data set, only 304 pairs were found to be appropriate for interchain interface comparison. Of the remaining 302 pairs, 288 pairs comprised of alignment involving monomer member(s), and domains represented in the remaining 14 pairs do not participate in interchain interface formation. This suggests a high degree of variation in the protein–protein interaction properties in distantly related protein pairs.

The other type of interface considered is between two domains in the same polypeptide chain. This type of interfaces can be defined or compared only for those pairs where both the members of the pair are multidomain proteins. Additionally, any pair that does not have an interfacial residue identified by our criteria (see Materials and Methods section) is eliminated from further analysis. Out of 606 pairs extracted, only 112 pairs were selected for studying interdomain interfaces. There are 374 structural domains (members) spanned over 110 superfamilies, and only 146 of these members are found to be multidomain entries in the PDB. Pairwise alignments within every superfamily were extracted by considering these 146 members. Only those pairs with both the members are part of multidomain proteins, so that interface can be defined for both the members, are considered. Further, 23 pairs are eliminated from the set, as either one or both members of the pair do not have an interfacial residue defined by our criteria.

Finally, pairs extracted from 80 out of 110 multimember superfamilies were found suitable for analysis. A plot of number of members of the superfamily considered for the analysis versus number of members in the same superfamily is shown in Figure 1. The scatter of the data points in the plot shows that there exists no major bias in the data set chosen, due to highly populated superfamilies.

We have pooled interdomain interface and interchain interface data sets, as both correspond to interaction between modular protein domains. Henceforth we refer to them collectively as interdomain interfaces, which refers to interfaces within a polypeptide chain and across polypeptide chains.

Extent of Conservation of Topologically Equivalent Positions in Interdomain Interfaces

We have asked the question: How often are the residues of a protein located at the interdomain interface topologically equivalent to the interdomain interfacial residues in a distantly (superfamily) related protein? It should be noted that this question does not address the similarity of the amino acid residue types in the interfaces of the two superfamily-related proteins, and instead addresses the issue of conservation of interfacial property, which is purely structural. As the families compared are in the same superfamily, the gross biochemical functions of the

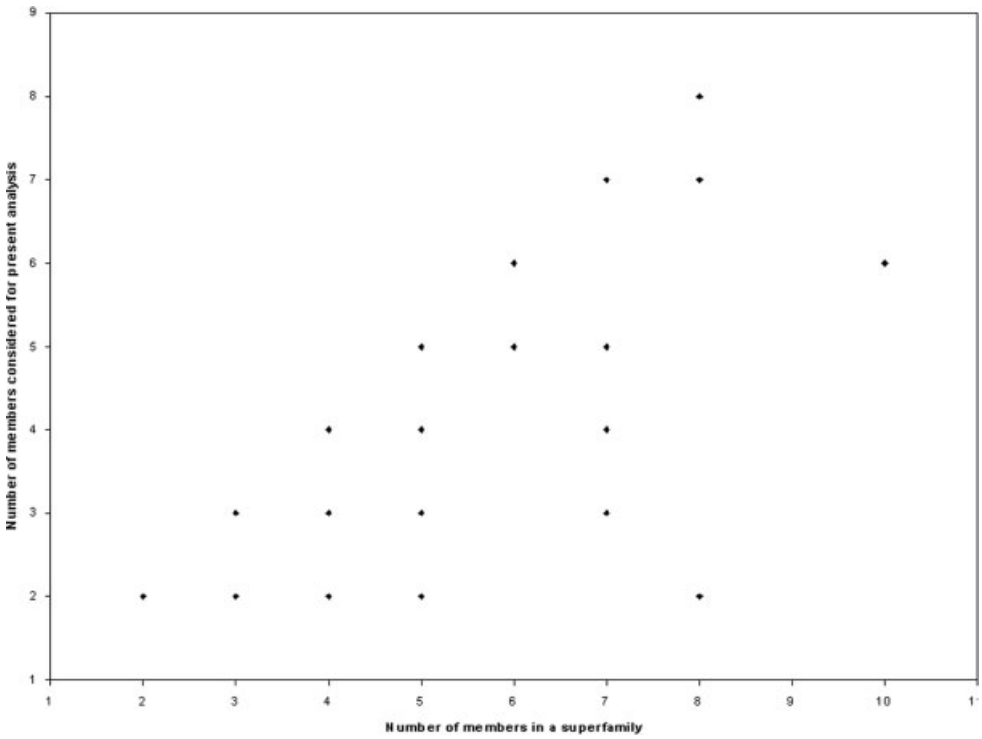

Fig. 1. Scatterplot showing the number of members in a superfamily and number of members considered in the present analysis. The scatter in the data points depicts no inherent bias due to presence of overpopulated superfamilies.

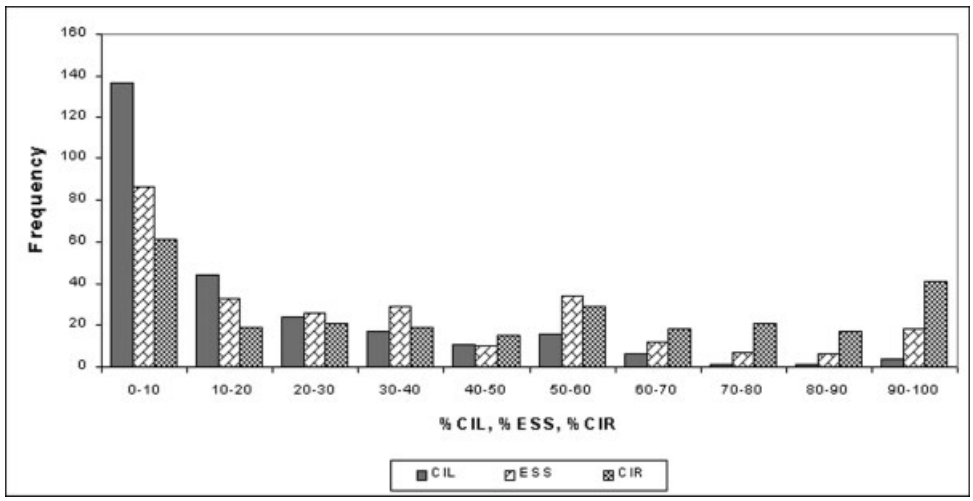

Fig. 2. Extent of occurrence of protein–protein interaction interfaces in topologically equivalent positions in protein domains related by superfamily. The *x* axis represents the percentage range of score; the *y* axis represents the number of cases that falls in the given frequency.

proteins in these families are either same or similar. The graph shown in Figure 2 corresponds to consolidated data based on the three scores (for detailed discussion on scores refer to the Material and Methods section) that are used to quantify the extent of conservation of topologically equivalent positions in interdomain interfaces in superfamily-related proteins. The *x* axis represents the percentage range of each score. Within a range in the *x* axis, the first bar represents the score corresponding to the conservation of interface location; the second bar represents the equivalent secondary structure score; and the third bar indicates the conservation of interface region score. The *y* axis represents the number of cases that fall in these ranges (frequency).

Figure 2 shows that the interfacial nature conservation score (CIL score) drops along the *x* axis and has the highest number of observations in the 0—10% range of the interfacial nature conservation. The property of the secondary

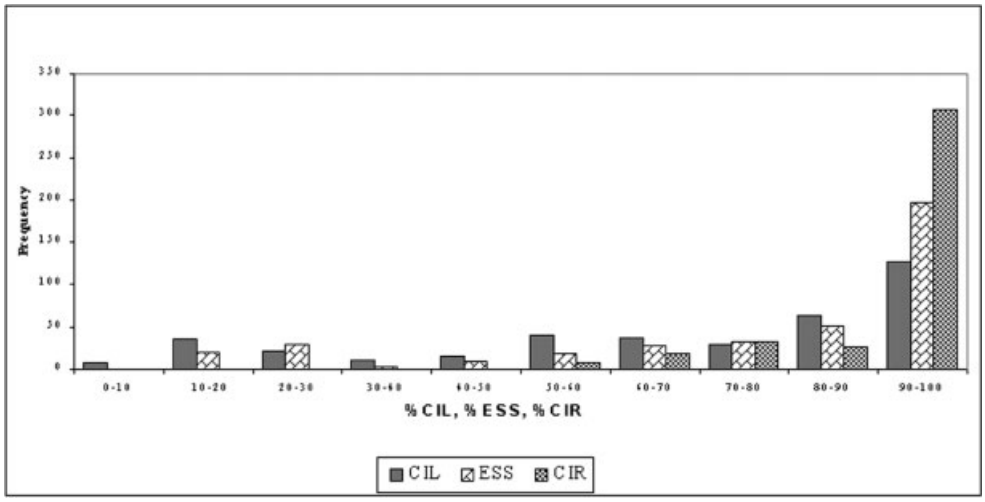

Fig. 3. Extent of occurrence of protein–protein interaction interfaces in topologically equivalent positions in protein domains related by family. The *x* axis represents the percentage range of score; the *y* axis represents the number of cases that falls in the given frequency.

structure involvement in the interface formation is also best observed in the 0–10% cases, and it drops further along the *x* axis. Thus, in general, the location of the interface is not well conserved when proteins are from different families of the same superfamily. Additionally, in many of these cases of protein domains related by superfamily, the oligomerization states are also different.

The CIR is a relaxed score as compared to the other two scores. In this score, we relax our criteria and identify interfacial residues in close proximity in the alignment, but they need not fall into topologically equivalent positions or secondary structures. This is done so as to account for the structural variation that is observed in remote homologues as a result of extensive evolution. The largest peak for CIR was observed at the 0–10% range, indicating that the property of interface location at the same site is not well preserved. The proximal scores are generally higher than other scores beyond 60%. This means that formation of interface is influenced by the high structural divergence during extensive evolution of proteins, and the manual inspection of the alignments suggests that it is not uncommon for insertions or deletions to occur at the interfaces. A smaller peak for CIR at 90–100% indicates that structural variation does contribute to variation at the interface location, and in these cases, the interface locations are mostly proximal to one another in the alignment but are not in topologically equivalent positions. This tiny set of superfamilies with members falling in this region shows reasonable conservation of the interfacial location, which is ascertained for these cases by visual inspection.

In the case of families (Fig. 3) consisting of closely related homologues, it can be seen from the graph that the location of the interface is well conserved, with all the three scores peaking at $> 90\%$. Hence, it can be seen that in the homologues within the family, the location of interface between the two members is conserved, which was also shown by Valdar and Thornton,[11] and in almost all cases the oligomerization state is also preserved.

Topological Equivalence of Interdomain Interfaces in Specialized Cases of Extensive Evolution of Both the Interacting Domains

In earlier sections, we observed that if the amino acid sequence changes are extensive between two homologues, the protein–protein interaction sites are not in the topologically equivalent positions. Most commonly, we observe that when the divergence is extensive, the pair of domains related by superfamily associate with unrelated domains/subunits. However, we identified a small minority of cases from the PDB that are remotely related and have two superfamily-related structural domains (as defined by SCOP). This means that in the case of a protein containing domains A and B, we have identified related proteins with similar domain composition, say A′ and B′, which are distantly (superfamily) related to A and B, respectively. However, no relationship constraint was placed between domains A and B within the same protein. The details of the examples identified are listed in Table I.

Apic et al.[29] reported that the relative positioning of the two superfamily-related domains in unrelated proteins are dissimilar. Bashton and Chothia[30] have shown that different superfamilies have different types of connections with the Rossmann domain, but those with a particular superfamily are of same type. In the present report, we have analyzed the extent of interface occurrence in topologically equivalent positions and the nature of residue substitutions observed where the interfacial residues are in topologically equivalent positions in a set of 6 domain–domain interaction pairs of the kind (Table I) discussed above.

The example of the protein pair DNA polymerase from *Archea* (1d5a) and its homologue reverse transcriptase (1mu2) shows circular permutation of the constituent domains. In this case, we observe that the interdomain

TABLE I. Homologous Pairs of Domain–Domain Interactions Where the Homologous Domains Are Superfamily-Related

PDB id	Description	Interacting domains		PDB id	Description	Interacting domains	
1a8p	Ferredoxin reductase from *Azotobacter vinelandii*	**Superfamily** Riboflavin synthase domain–like	**Superfamily** Ferredoxin reductase–like, C-terminal NADP-linked domain	1ddg Chain A	Sulfite reductase NADPH flavoprotein	**Superfamily** Riboflavin synthase domain–like	**Superfamily** Ferredoxin reductase–like, C-terminal NADP-linked domain
		Family Ferredoxin reductase FAD-binding domain–like	**Family** Reductase			**Family** NADPH-cytochrome p450 reductase FAD-binding domain–like	**Family** NADPH-cytochrome p450 reductase–like
Domain Boundaries		2–100	101–258	**Domain Boundaries**		226–446	447–559
1jak Chain A	β-*N*-acetyl hexosaminidase	**Superfamily** Transglycosidases	**Superfamily** β-*N*-acetylhexosaminidase–like domain	1gqi Chain A	α glucuronidase	**Superfamily** Transglycosidases	**Superfamily** β-*N*-acetylhexosaminidase–like domain
		Family β-*N*-acetylhexosaminidase catalytic domain	**Family** β-*N*-acetylhexosaminidase domain			**Family** α-D-glucuronidase catalytic domain	**Family** α-D-glucuronidase, N-terminal domain
Domain Boundaries		8–150	151–506	**Domain Boundaries**		5–151	152–712
1gpu Chain A	Transketolase (TK)	**Superfamily** Thiamin diphosphate–binding fold (THDP–binding)	**Superfamily** TK C-terminal domain–like	1 dtw Chain B	Branched chain α ketoacid dehydrogenase	**Superfamily** Thiamin diphosphate–binding fold (THDP–binding)	**Superfamily** TK C-terminal domain-like
		Family TK-like THDP-binding domains	**Family** TK C-terminal domain–like			**Family** Branched-chain α-keto acid dehydrogenase THDP-binding domains	**Family** Branched-chain α-keto acid dehydrogenase β-subunit, C-terminal-domain
Domain Boundaries		338–554	535–680	**Domain Boundaries**		17–204	205–342
1d5a Chain A	Archeal DNA polymerase	**Superfamily** Ribonuclease H–like	**Superfamily** DNA/RNA polymerases	1mu2 Chain A	Reverse transcriptase	**Superfamily** DNA/RNA polymerases	**Superfamily** Ribonuclease H–like
		Family DnaQ-like 3′-5′-exonuclease	**Family** DNA polymerase I			**Family** Reverse transcriptase	**Family** Ribonuclease H
Domain Boundaries		1–347	348–756	**Domain Boundaries**		3–429	430–555
1e4e chain A	D-ala-D-lactate ligase	**Superfamily** Pre-ATP-grasp domain	**Superfamily** Glutathione synthetase ATP-binding domain–like	1m0w chain A	Glutathione synthase	**Superfamily** Pre-ATP-grasp domain	**Superfamily** Glutathione synthetase ATP-binding domain–like
		Family D-Alanine ligase N-terminal domain	**Family** ATP-binding domain of peptide synthetases				
Domain Boundaries		2–131	132–342				
1a9x Chain A	Carbamoyl phosphate synthatase, large chain	**Superfamily** Pre-ATP-grasp domain	**Superfamily** Glutathione synthetase ATP-binding domain–like			**Family** Eukaryotic glutathione synthetase, substrate-binding domain	**Family** Eukaryotic glutathione synthetase ATP-binding domain
		Family BC N-terminal domain–like	**Family** BC ATP-binding domain–like	**Domain Boundaries**		216–323	6–210, 324–491
Domain Boundaries		1–127	128–402				

Abbreviations: ATP, adenosine triphosphate; FAD, flavin adenine dinucleotide; NADP, nicotinamide adenine dinucleotide phosphate; NADPH, dihydro-nicotinamide-adenine-dinucleotide phosphate.

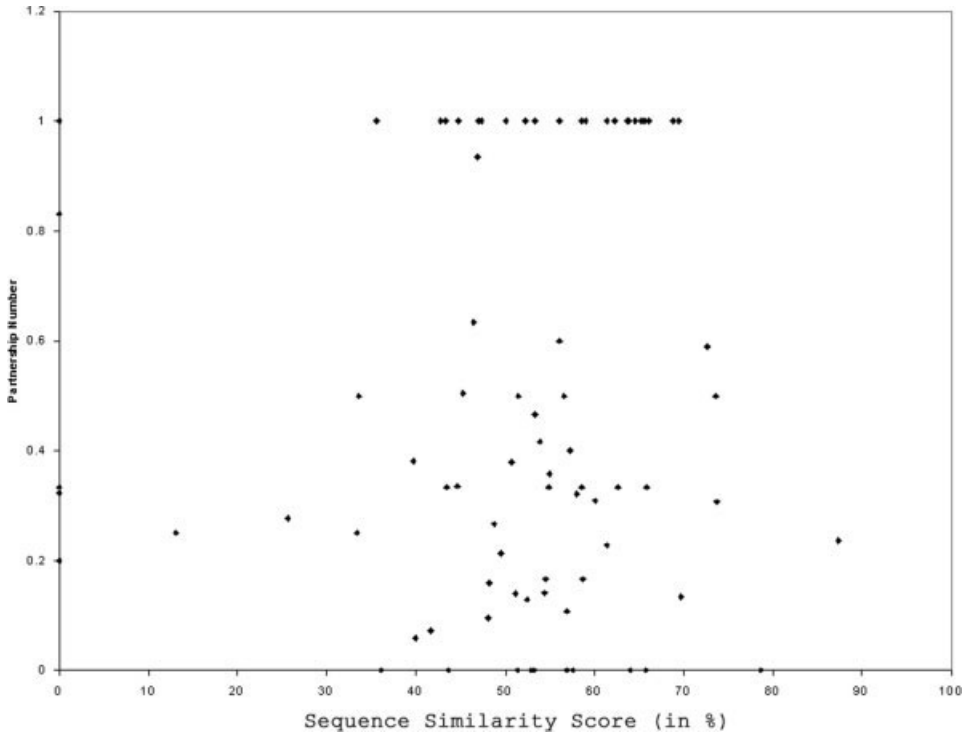

Fig. 4. Plot showing the relationship between sequence similarity score at the interface and the partnership number. The partnership number is a normalized measure of number of superfamilies that interact with a given superfamily.

interface is not topologically equivalent. This is consistent with the observation of Apic et al.,[29] where they showed that the relative positioning of circularly permuted protein domains is not conserved. In all the remaining cases, we observe that the interface locations are in topologically equivalent surfaces of the tertiary structure. We have also calculated the frequency of occurrence of the superposed interfacial residues on the tertiary surface. We observe that 28% of the interfacial residues identified (out of 139 pairs of interfacial residues) are in topologically equivalent positions. However, when we consider the rough topological equivalence, giving allowance of up to 5-residue mismatch, we observe that 72% of the interfacial residues can now be deemed to be in the same location on tertiary structure.

In the cases where the interfacial residues are in topologically equivalent positions, we have studied the residue substitution patterns in these positions. The most commonly observed event is that the short, nonpolar amino acids are conservatively substituted in most of the alignment positions (47% of 40). A few positions also show conservation of the residues (22% of 40), and drastic substitutions of residues are also not infrequent (30% of 40). Thus, we observe here that the interfacial residues, even if present at topologically equivalent positions, do not show strong residue conservation pattern.

Residue Conservation at the Interface

Similarity of interfacial residues in the topologically equivalent alignment positions has been analyzed by considering all the interfaces in the superfamily data set. Only the small proportion of positions of the pairwise alignments, where interfacial nature is conserved, are taken into consideration. In 277 interdomain interfaces, we find there are 1025 interface conserved positions (out of 4822 total interfacial positions). In these pairs of remote relatives (where CIL score = 100%), we observe that the sequence identity at the interface is in the range of 0–50%. However, we observe from the data set that the nature of the partner domain associated with the given superfamily is not influenced by the extent of divergence (indicated by percentage identity) at the interface. Additionally, the number and types of superfamilies that associate are not influenced by the sequence similarity at the interface. This feature can be visualized in Figure 4, which shows the plot of partnership number versus the sequence similarity scores for all the superfamilies of proteins used in present analysis. The partnership number is a population-independent measure of the number of different superfamilies of proteins attached with the members of the superfamily in consideration. The sequence similarity score, which considers conserved and substituted positions in the alignment, is calculated for the interface regions irrespective of conservation of interface location across the members of superfamily. From the plot, it is clear that higher sequence similarity does not translate into similar types of domains being associated with the superfamily.

To assess the extent of residue substitutions in these positions, we derive two substitution matrices—one from all the aligned positions in the pairwise superfamily

alignment data set, and the other from the 1025 aligned interfacial positions. The raw matrices obtained for interfacial positions were normalized with respect to the background substitution values obtained for all aligned positions in the data set. The linear correlation coefficient between the raw substitution values for interface and for the general data set was calculated. Similar comparisons were made to BLOSUM matrices also using the normalized interface matrices. Details of matrix construction and analysis are given in the Materials and Methods section. Similar operations were performed also on the domains within families that have closely related homologues. The normalized interfacial matrices are given in Supplementary Data 3.

The correlation coefficient value of 0.717 is obtained when the substitution matrices derived from all aligned positions in the superfamily data set are compared with matrices derived from only interfacial positions for the same data set. Similarly, a correlation coefficient value of 0.830 is obtained for family related proteins, on comparison between general residue conservation and interfacial residue conservation. The interfacial matrices derived for both family and superfamily were compared to a general substitution matrix, BLOSUM. This analysis allowed us to assess the residue substitution rates at the interface locations, in the light of background substitutions observed in proteins in general. We observe that the residue substitutions at the interfaces derived from superfamily-related proteins correlate poorly with BLOSUM, with a correlation coefficient of 0.451, and the residue substitution at the interfaces of family related proteins are correlated better, with a coefficient of 0.733. All the correlation coefficients were found to be significant when they were calculated between pairs of 1000 randomly generated numbers.

Thus, it appears that being at the topologically equivalent interface region does not translate into markedly better conservation of residues at these aligned positions. This is especially true for protein interfaces from superfamilies. The interface formation is thus found to be tolerant to amino acid substitutions. It is observed that the small, nonpolar amino acids, like alanine, valine, isoleucine and leucine, are conservatively substituted in the location considered in all types of proteins (i.e., proteins either related by family or superfamily).

Wherever the sequences show large divergence, the residues at topologically equivalent interfacial regions are often substituted. However, as also shown by Valdar and Thornton,[11] in the case of closely related proteins, we note that not only are the interface locations conserved very well, but the residues involved in interface formation are also conserved, or are conservatively substituted. Additionally, we observe that the identity of residues in the family-related proteins is preserved very well, such that the diagonal elements of the interfacial matrix generated for family-related proteins correspond to high scores. This feature of diagonal elements showing high scores is completely absent in matrices generated for superfamily-related proteins.

Variation in Interaction Between Domains in Superfamilies: A Strategy for Bringing About Functional Diversity

Nature seems to have utilized variation in oligomerization and domain–domain interactions as a strategy to bring about interesting variation in overall functionality of the proteins in different families in a superfamily. During evolution, as the sequence divergence becomes extensive, as will be seen in the following sections, new domains are recruited with the existing domain. We also show that, under such circumstances, spatial positioning of the composite domains is often variant, such that a desired biological effect is achieved.

Although the general trend is that the locations of interfaces are not topologically equivalent in structurally aligned superfamily-related proteins, we observe exceptionally good conservation of interface location in the cases of a few superfamilies, where oligomerization is a prerequisite for function. A classic example of this category is that of enzymes belonging to the triosephosphate isomerase (TIM) fold, where dimerization is essential for catalysis. Some other examples of preservation of protein–protein interaction sites in topologically equivalent positions of remote homologues include cyclins (interdomain example with CIL score = 100) bound to cyclin dependent–kinases and that of bacterial adhesins (CIL = 100). Bacterial adhesins are surface proteins in the bacterial cell wall that bind to receptor molecules on the surface of a susceptible host cell, enabling the bacterium to make intimate contact with the host cell, adhere, colonize, and resist flushing.[31,32] Table II shows a few examples of superfamily-related proteins that maintain their basic activity but achieve functional specialization by variations in their interaction with other domains. We discuss the example of chorismate mutase in some detail here.

Chorismate mutase–like proteins

The homodimeric enzyme chorismate mutase (CM) from *Escherichia coli* (ECM) and yeast (YCM) catalyzes the first committed step of aromatic amino acid biosynthesis, the rearrangement of chorismate to prephenate via similar catalytic mechanism.[33] The subunits of both the enzymes adopt a 4-helical bundle fold, and each subunit harbors an active site.[34,35] YCM is allosterically regulated by means of a regulatory domain that is also in the oligomerization region and is sensitive to varying concentrations of Trp.[36] ECM, on the other hand, shows no sensitivity to varying concentrations of Trp and no evidence of regulatory mechanisms, and is constitutively active.[36] The crystal structures[37] reveal that the mode of dimerization of these two enzyme is different.

Structure-based superposition of ECM and YCM dimers reveals that one subunit of the ECM dimer is superposed well with one of the subunits of YCM dimer, while the other ECM subunit occupies the regulatory region of the same YCM subunit (Fig. 5). Thus, we see that the interchain interfaces presented by both these enzymes have very little in common, and this feature is

TABLE II. Consequences for Biological and Biochemical Properties When Domain–Domain Interaction Interfaces Are Not Topologically Equivalent in Representative Pairs Within a Superfamily

Superfamily		Overall CIR score (%)	Oligomerization state	Difference in biological role–activity	References
Chorismate mutase (CM)	*S. cerevisiae* CM	15.38	Homodimer	Has an additional regulatory domain, sensitive to Trp concentrations	32–36
	E. coli CM		Homodimer	No evidence of regulation of this enzyme is reported.	
Four Helical cytokines	Interleukin 4, leptin	39.44	Monomer	Each of these ligands binds to its respective receptors in the oligomeric state as discussed. The receptor seems to recognize different shape of the oligomeric ligand that is produced by different modes of oligomerization.	40–46
	Interleukin 6		Parallel face-to-face homodimer		
	Interleukin 5		S–S linked homodimer		
	Ftl3 ligand		Homodimer, head-to-head orientation		
	Ciliary neutropic factor		Antiparallel face-to-face homodimer		
Robosome-inactivating proteins (RIPs)	Mistletoe lectin (type II RIP)	50.00	Heterodimer	The toxic chain is oligomerized with a lectin moiety, by an S–S linkage. This facilitates the cell entry for the toxic chain by binding to the galactose-containing cell surface receptors via the lectin chain.	47–51
	Shiga toxin (type II RIP)		(Hetero) hexamer	The toxic chain is associated with a pentameric cell-binding subunit. The pentamer binds to globotriaosyl ceramide on the cell surface, thus facilitating the cell entry for the toxic chain.	
	Saporina SO6 (type I RIP)		Monomer	Cell entry of the toxic chain by simple internalization.	
MurD-like peptide ligases	Folyl poly glutamate synthase (FPGS)	14.29	Two-domain protein	The presence of additional domains assists the enzyme in recognizing different types of substrates, while catalyzing the same biochemical reaction.	52–54
	MurD enzyme		Three-domain protein		
Class II aaRS biotin synthase	Glycyl tRNA synthases	43.33	Two-domain protein	The additional domains–chains in these cases assist the presentation of substrate to the catalytic domain–chain.	55–60
	Seryl tRNA synthases		Two-domain protein		
	Phenylalanyl tRNA synthases		Heterodimeric protein		
	Thereonyl-tRNA synthases		Three-domain protein		
	Asparagine synthase		Monomer, single-domain protein	Conversion of Asp to Asn	
	Biotin holoenzyme synthase (BirA)		Two-domain protein	Has an additional DNA-binding domain to bring about transcriptional regulation of biotin biosynthesis by BirA.	

translated as low CIL, CIR, and ESS scores in the present analysis.

Thus, from Table II and from the previous example, we observe that the overall catalytic mechanism or cofactor required for function is identical in superfamily-related protein domains. But the ultimate biological role of such proteins is governed by the variation in their association with other domains. The catalytic domain is associated with a number of other domains in a variety of different ways to accommodate new substrates or to generate different products, or to impose novel regulatory mechanisms on the pathways.

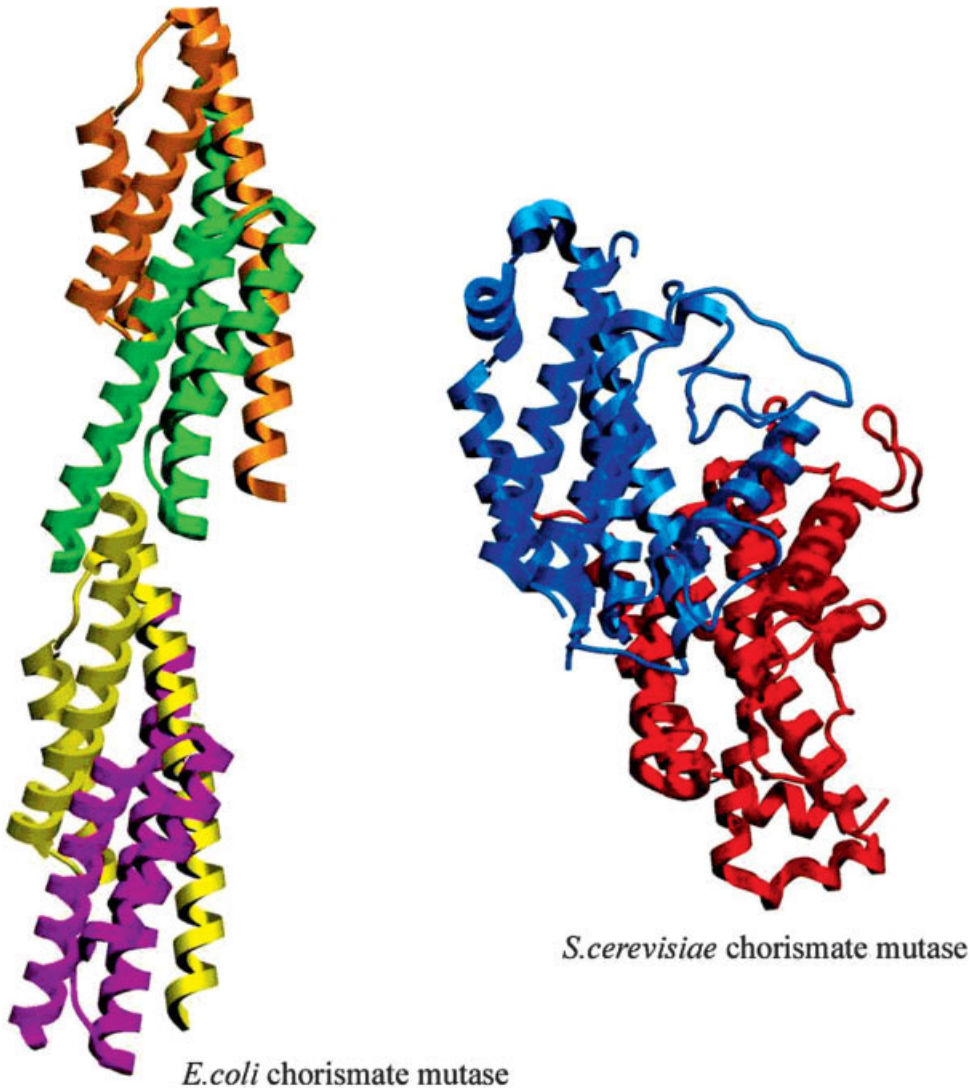

Fig. 5. Difference in oligomerization between homodimeric chorismate mutase from *E. coli* and *S. cerevisiae* (green- and blue-colored ribbons indicate the superposed domains containing the active site from ECM and YCM, respectively). This figure has been produced using SETOR software.[39]

Superfamily-related protein domains bring changes in signaling and metabolic pathways and in gene networks

The present study shows that, most often, the remotely related protein homologues have no similarity in the way they associate with other protein domains. In the previous section we discussed the examples from the data set, where the superfamily-related proteins show differential interactions with other domains, while retaining the basic biochemical function. Here we show that the domains that are remotely related in two multidomain proteins, while all other composite domains are identical, have influence on the ultimate biological function of the gene products under consideration and, in many cases, the pathways to which they belong. Here, the "domain" indicates a set of amino acid sequences that are similar, and is most often associated with a specific molecular function. The procedure for identifying such cases is discussed in the Materials and Methods section. In more than 80% of the cases identified from the Swiss-Prot database, the distantly related domains are involved in domain recruitment, gene duplication, and divergence.

This suggests that the coupled effect of sequence divergence that results in superfamily-related domain families and recruitment of new domains is the most common step toward incorporating new gross functionality to the gene product and introducing radical changes in the metabolic and signaling pathways. However, as clearly indicated in the examples in Figure 6, that extensive sequence divergence can bring about drastic changes in the protein–protein interaction that is influential in altering the course of a signaling or a metabolic pathway. One of the representative examples is discussed in details here.

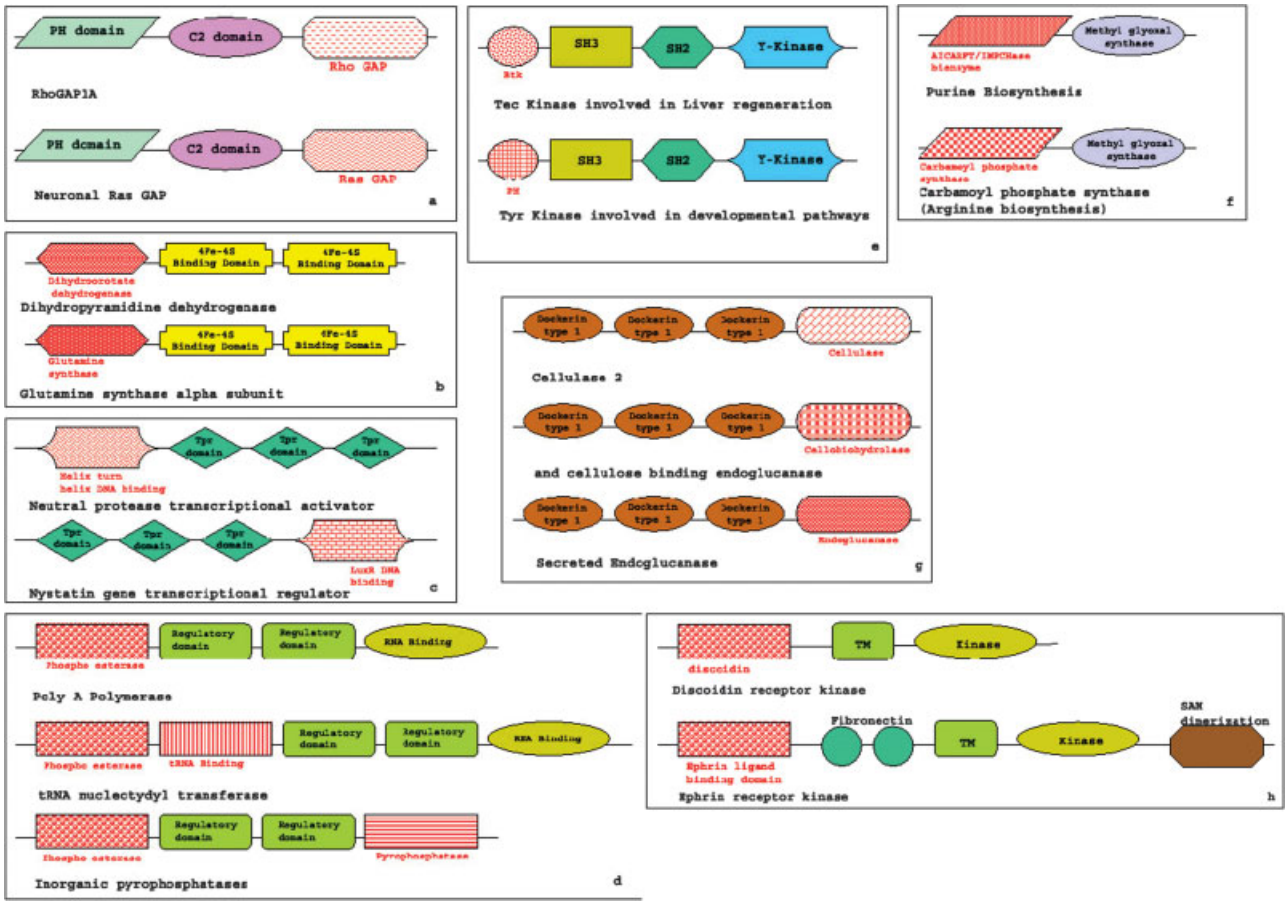

Figure 6.

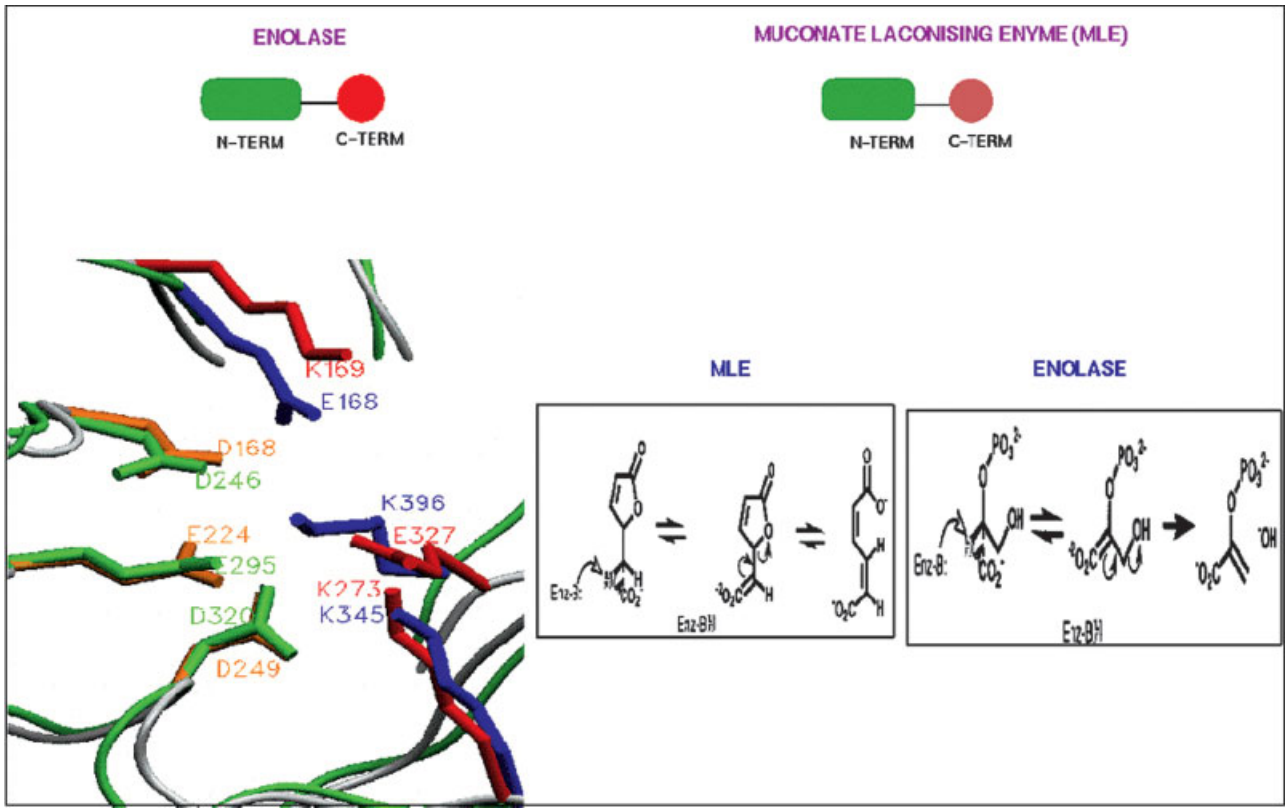

Figure 7.

Enolase/muconate-lactonizing enzyme (MLE)

Enolase is the key enzyme of the glycolysis that forms the central component of metabolic pathways in both eukaryotes and prokaryotes. It catalyzes the conversion of 2-phosphoglycerate to phospho enol pyruvate. The enzyme is composed of two distinct domains (Fig. 7). The N-terminal domain is classified in the "Enolase_N-term–like" family and included in the "Enolase_N-term–like" superfamily. The C-terminal catalytic domain is grouped under the "Enolase_C-term like–domain" and "Enolase family."

Another enzyme with a similar domain combination at the superfamily level is the muconate-lactonizing enzyme (MLE; Fig. 7). This enzyme is a part of the biochemical machinery involved in the degradation of aromatic acids in only bacterial species through the β-ketoadipate pathway. The N-terminal domain of the enzyme is classified under the same superfamily and family as the N-terminal domain of the enolase. The C-terminal domain of the enzyme also belongs to the same superfamily as the enolase C-terminal domain, but is grouped under a different family, β-D-glucarate deydratase, as compared to the family of the C-terminal domain of enolase.[38]

The classification of the C-terminal domain under the same superfamily is further consistent with their sharing of the core chemical reaction of the extraction of proton from the carbon adjacent to the carboxyl group (Fig. 7). However the comparison of active sites reveals the variation in the specific features of the two enzymes. In addition to the distinct nature of substrates (muconate for MLE and 2-phosphoglycerate for enolase), the following differences are observed at the catalytic site.

The relative positions of the atoms in the catalytic region (Fig. 7) are similar in the two proteins, while the residues that contain them are placed at different points in the protein fold. Despite the conservation of the position of the catalytic residues in the protein scaffold, their identities and roles vary. For instance, E168 of enolase is important for both the processes of extraction of proton and stabilization of the intermediate. The corresponding residue K169 of MLE acts only as a catalytic base and has no role in stabilization of the intermediate. In addition, K396 of enolase helps in the intermediate stabilization, while the functionally analogous residue in MLE is E327. Such unusual conservation of catalytic sites has enabled the enzymes to retain the core reaction, while they have diverged in terms of substrate specificity and reactions beyond proton extraction.

In the example described above, the MLE is an intermediate protein converting muconate to muconolactone, which is finally converted to β-keto adipate in a series of steps in the pathway. The β-keto adipate also serves as a precursor for the generation of acetyl coenzyme A (coA). The enolase enzyme, on the other hand, is involved in the intermediate step of glycolysis generating acetyl coA as the end product. It therefore appears that during evolution, the enzymes sharing similar catalytic mechanisms have originated from a common ancestor and have been recruited in independent pathways based on their subsequent functional specialization. The conservation of the relative positions of the catalytic residues, despite their distinct identity and roles, suggests a higher preference to conserve the location of catalytic sites during evolution of enzymes as against recruiting residues at new sites in the 3D fold for catalysis. The members of a superfamily hence appear to be the ideal templates for generating alternate functions by the modification of functional sites, including the interfaces of protein–protein interactions, as is evident from the previous sections.

This example and those summarized in Figure 6 throw light on the fact that as sequences diverge, their domain–domain interaction modes, substrate specificities, and mode of catalysis vary, and these changes could result in the final outcome of the respective metabolic or signaling pathways. Thus, the divergence of the proteins of the superfamily to distinct families leads not only to functional specialization but is also probably one of the first steps toward the formation of newer and specialized pathways.

CONCLUSIONS

It has been shown by Valdar and Thornton[11] and also as a part of the current analysis that the interfacial location is conserved for the closely related homologues (members of a family). However, the nature of protein–protein interaction and structure appear to be varying under extreme divergent evolution, as reported by Teichmann[13] and Jordan et al.[14] Here, we have addressed the conservation of the oligomeric state in both interdomain and interchain cases when the homology is barely detectable, at the sequence level, between the proteins. The results we obtain here suggest that the gross location of interface in the superfamily-related proteins need not be preserved.

The accumulation of protein–protein interaction data contributed by both high-throughput experimental and computational methods, coupled with application of the powerful algorithms for detecting remote homologues and structure prediction, could provide meaningful insights into the quaternary structure and oligomerization of proteins. With significant progress in the identification of remote homologues using powerful profile matching meth-

Fig. 6. Domain organization of functionally distinct proteins that show same or similar domain organization, but for some domains related by superfamily: (**a**) Rho GAP and RasGAP superfamily; (**b**) Dihydro orotate dehydrogenase and glutamine synthase superfamily; (**c**) LuxR and helix–turn–helix DNA-binding domain superfamily; (**d**) Phosphoesterase, tRNA-binding domain, and pyro phosphatase superfamily; (**e**) Btk and PH domain superfamily; (**f**) AICARFT/IMPCHAse bienzyme and carbamoyl phosphate synthase superfamily; (**g**) cellulase, cellobiohydrolase, and endoglucanase superfamily; (**h**) discoidin and ephrin receptor kinases.

Fig. 7. Enolase and MLE. Domain arrangement of the two enzymes, enolase and MLE, is shown (**top panel**). Structural superposition of the catalytic sites of enolase and MLE (**lower left panel**) highlights the similar and distinct catalytic residues. Metal binding residues conserved in both the enzymes are colored green. The identities of the residues involved in the abstraction of protons vary between the two enzymes and are shown in blue (enolase) and red (MLE). The schematic represents the core biochemical reaction catalyzed by enolase and MLE. Extraction of proton from the carbon adjacent to the carboxylic acid is shown (**lower right panel**).

ods like the PSI-BLAST, Hidden Markov model (HMM), and fold recognition, the present observation suggests that one should be cautious while extrapolating the protein–protein interaction features based on remote homology.

It is known that in the fold, the biochemical and functional properties are retained in extensive divergent evolution. However, we also observe that wherever quaternary structure has a bearing on the intricate function of a protein, the type of oligomer formed and the superfamilies involved are generally preserved. It is observed in the current work that homologous proteins with high divergence in sequences could be implicated in the divergence of the pathways of their involvement, and this change is brought about by differential interactions of member proteins. Such variations in protein interactions could be the key in bringing about interesting divergence in pathways, regulation, and cross talks.

With the emergence of structures on large machinery such as the ribosome, the DNA and the RNA polymerase, and chaperones and virus assembly, prospects of extrapolating the basis of this machinery in other organisms given sequence similarity between the component proteins has increased. However, the nature of protein associations found in those complexes with extensive divergent evolution may vary significantly even though the tertiary structures may be preserved. In this light, the current analysis suggests that protein–protein interaction sites should be viewed to be different as compared to functional sites such as active sites and small molecular ligand–binding sites. Protein–protein interaction sites may be considered fundamentally different in that variation in interaction is used in setting the direction for the given biochemical function.

Thus, it appears that high sequence divergence and differential protein–protein interaction is the minimal step toward bringing about variations in pathways, which is further compounded with domain recruitment, domain duplication, and divergence of the duplicated domain.

ACKNOWLEDGMENTS

Our thanks to Anirban Bhaduri for kindly providing us with the superfamily alignments and superpositions in the PASS2 database, to Shashi Pandit for providing flat files of the SUPFAM database, to Swanand Gore for providing us with the domain comparison algorithm, and to V. S. Gowri for the parsed SCOP files. Our thanks also to Cyrus Chothia, MRC, Cambridge, for taking keen interest in our work and giving valuable suggestions, and to R. Sowdhamini and S. Balaji for their comments and suggestions.

REFERENCES

1. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
2. Pawson T, Nash P. Protein–protein interactions define specificity in signal transduction. Genes Dev 2000;14:1027–1047.
3. Connolly ML. Shape complementarity at hemoglobin alpha 1 beta 1 subunit interface. Biopolymers 1986;25:1229–1247.
4. Smith GR, Sternberg MJ. Prediction of protein–protein interaction by docking methods. Curr Opin Struct Biol 2002;12:28–35.
5. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng 2000;13:77–82.
6. Glaser F, Steinberg DM, Vasker IA, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 2001;43:89–102.
7. Brinda KV, Kannan N, Vishveshwara S. Analysis of homodimeric protein interfaces by graph spectral methods. Protein Eng 2002;15: 265–277.
8. Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Biol 2003;325:377–387.
9. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
10. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
11. Valdar WSJ, Thornton JM. Protein–protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
12. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science 2002;296:750–752.
13. Teichmann SA. The constraints protein–protein interaction place on sequence divergence. J Mol Biol 2002;324:399–407.
14. Jordan IK, Wolf YI, Koonin EV. No simple dependence between protein evolution rate and the number protein–protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 2003;3:1–21.
15. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
16. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.
17. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–998.
18. Mallika V, Bhaduri A, Sowdhamini R. PASS 2: a semi-automated database of protein alignments organised as structural superfamilies. Nucleic Acids Res 2002;30:284–288.
19. Sowdhamini R, Burke DF, Huang JF, Mizuguchi K, Nagarajaram HA, Srinivasan N, Steward RE, Blundell TL. CAMPASS: a database of structurally aligned protein superfamilies. Structure 1998;6:1087–1094.
20. Sali A, Blundell TL. Definition of general topological equivalence in protein structures. J Mol Biol 1990;212:403–428.
21. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP. JOY: protein sequence–structure representation and analysis. Bioinformatics 1998;14:617–623.
22. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–323.
23. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
24. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
25. Godzik A. The structural alignment between two proteins: Is there a unique answer? Protein Sci 1996;5:1325–1338.
26. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31: 365–370.
27. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
28. Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS, Sowdhamini R, Srinivasan, N. SUPFAM—a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. Nucleic Acids Res 2002;30:289–293.
29. Apic G, Gough J, Teichmann SA. Domain combinations in archeal, eubacterial and eukaryotic proteomes. J Mol Biol 2001;310:311–325.
30. Bashton M, Chothia C. The geometry of domain combination in proteins. J Mol Biol 2002;315:927–939.

31. Christensen AM, Massiah MA, Turner BG, Sundquist WI, Summers MF. Three-dimensional structure of the HTLV-II matrix protein and comparative analysis of matrix proteins from the different classes of pathogenic human retroviruses. J Mol Biol 1996;264:1117–1131.
32. Hill CP, Worthylake D, Bancroft DP, Christensen AM, Sundquist WI. Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. Proc Natl Acad Sci USA 1996;93:3099–3104.
33. Braus GH. Aromatic amino acid biosynthesis in the yeast *Saccharomyces cerevisiae*: a model system for the regulation of a eukaryotic biosynthetic pathway. Microbiol Rev 1991;55:349–370.
34. Strater N, Hakansson K, Schnappauf G, Braus G, Lipscomb WN. Crystal structure of the T state of allosteric yeast chorismate mutase and comparison with the R state. Proc Natl Acad Sci USA 1996;93:3330–3334.
35. Xue Y, Lipscomb WN. Location of active site of allosteric chorismate mutase from *Saccharomyces cerevisiae*, and comments on catalytic and regulatory mechanisms. Proc Natl Acad Sci USA 1995;92:10595–10598.
36. Strater N, Schnappauf G, Braus G, Lipscomb WN. Mechnisms of catalysis and allosteric regulation of yeast chorismate mutase from crystal structures. Structure 1997;5:1437–1452.
37. Lee AY, Karplus PA, Ganem B, Clardy J. Atomic-structure of the buried catalytic pocket of *Escherichia coli* chorismate mutase. J Amer Chem Soc 1995;117:3627–3628.
38. Neidhart DJ, Kenyon GL, Gerlt JA, Petsko GA. Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. Nature 1990;347:692–694.
39. Evans SV. SETOR: Hardware lighted three-demensional solid model representations of macromolecules. J Mol Graphics 1993;11: 134–138.
40. DeVos R, Guisez Y, Van der Heyden J, White DW, Kalai M, Fountoulakis M, Plaetinck G. Ligand-independent dimerisation of the extracellular domain of the leptin receptor determination of the stoichiometry of leptin binding. J Biol Chem 1997;272:18304–18310.
41. Somers W, Stahl M, Sheera JS. 1.9 Å crystal structure of interleukin 6: implications for a novel mode of receptor dimerization and signaling. EMBO J 1997;16:989–997.
42. Milburn MV, Hassell AM, Lambert MH, Jordan SR, Proudfoot AE, Graber P, Wells TN. A novel dimer configuration revealed by the crystal structure at 2.4 Å resolution of interleukin-5. Nature 1993;363:172–176.
43. Savvides SN, Boone T, Karplus AP. Flt3 ligand structure and unexpected commonalities of helical bundles and cystine knots. Nat Struct Biol 2000;7:486–491.
44. McDonald NQ, Panayotatos N, Hendrickson WA. Crystal structure of dimeric human ciliary neurotropic determined by MAD phasing. EMBO J 1995;14:2689–2699.
45. Wlodaver A, Pavlovsky A, Gustchina A. Crystal structure of human recombinant interleukin-4 at 2.25Å resolution. FEBS Lett 1992;309:59–64.
46. Hage T, Sebald W, Reinemer P. Crystal structure of the interleukin-4/receptor alpha chain complex reveals a mosaic binding interface. Cell 1999;97:271–281.
47. Krauspenhaar R, Eschenburg S, Perbandt M, Kornilov V, Konareva N, Mikailova I, Stoeva S, Wacker R, Maier T, Singh T, Mikhailov A, Voelter W, Betzel C. Crystal structure of mistletoe lectin I from *Viscum album*. Biochem Biophys Res Commun 1999;257:418–424.
48. Fraser ME, Chernaia MM, Kozlov YV, James MN. Crystal structure of the holotoxin from *Shigella dysenteriae* at 2.5 Å resolution. Nat Struct Biol 1994;1:59–64.
49. Savino C, Federici L, Ippoliti R, Lendaro E, Tsernoglou D. The crystal structure of saporin SO6 from *Saponaria officinalis* and its interaction with ribosome. FEBS Lett 2000;470:239–243.
50. Stirpe F, Barbieri L, Battelli MG, Soria M, Lappi DA. Ribosome-inactivating proteins from plants: present status and future prospects. Biotechnology (NY) 1992;10:405–412.
51. Peumans WJ, Hao Q, Van Damme EJ. Ribosome-inactivating proteins from plants: more than RNA N-glycosidases? FASEB J 2001;15:1493–1506.
52. Bertrand JA, Auger G, Fanchon E, Martin L, Blanot D, van Heijenoort J, Dideberg O. Crystal structure of UDP-*N*-acetylmuramoyl-L-alanine:D-glutamate ligase from *Escherichia coli*. EMBO J 1997;16:3416–3425.
53. Bertrand JA, Auger G, Martin L, Fanchon E, Blanot D, Le Beller D, van Heijenoort J, Dideberg O. Determination of MurD mechanism through crystallographic analysis of enzyme complexes. J Mol Biol 1999;289:579–590.
54. Sun X, Bognar AL, Baker EN, Smith CA. Structural homologies with ATP- and folate-binding enzymes in the crystal structure of folylpolyglutamate synthetase. Proc Natl Acad Sci USA 1998;95: 6647–6652.
55. Nakatsu T, Kato H, Oda J. Crystal structure of asparagine synathase reveals a close evolutionary relationship to class II amino-acyl tRNA synthetase. Nat Struct Biol 1998;5:15–19.
56. Wilson KP, Shewchuck LM, Brennan R, Otsuka A, Matthews BW. *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains. Proc Natl Acad Sci USA 1992;89:9257–9261.
57. Reshetnikova L, Moor N, Lavrik O, Vassylyev DG. Crystal structures of phenylalanyl-tRNA synthetase complexed with phenylalanine and phenylalanyl-adenylate analogue. J Mol Biol 1999;287: 555–568.
58. Sankaranarayanan R, Dock-Bregeon AC, Romby P, Caillet J, Springer M, Rees B, Ehresmann C, Ehresmann B, Moras D. The structure of thereonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. Cell 1999;97:371–381.
59. Logan DT, Mazauric MH, Kren D, Moras D. Crystal structure of glycyl-tRNA synthetase from *Thermus thermophilus*. EMBO J 1995;14:4156–4167.
60. Belrhali H, Yaremchuk A, Tukalo M, Larsen K, Berthet-Colominas C, Leberman R, Beijer B, Sproat B, Als-Nielsen J, Grubel G, Legrand JF, Lehmann M, Cusack S. Crystal structures at 2.5 angstrom resolution of seryl-tRNA synthetase complexed with two analogs of seryl adenylate. Science 1994;263:1432–1436.