BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 15 2007, pages 1909–1918
doi:10.1093/bioinformatics/btm274

*Structural bioinformatics*

# Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions

Mainak Guharoy and Pinak Chakrabarti*
Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Kolkata 700054, India

Received on February 13, 2007; revised on May 8, 2007; accepted on May 14, 2007
Advance Access publication May 17, 2007
Associate Editor: Martin Bishop

## ABSTRACT

**Motivation:** The increasing amount of data on protein–protein interaction needs to be rationalized for deriving guidelines for the alteration or design of an interface between two proteins.

**Results:** We present a detailed structural analysis and comparison of homo- versus heterodimeric protein–protein interfaces. Regular secondary structures (helices and strands) are the main components of the former, whereas non-regular structures (turns, loops, etc.) frequently mediate interactions in the latter. Interface helices get longer with increasing interface area, but only in heterocomplexes. On average, the homodimers have longer helical segments and prominent helix–helix pairs. There is a surprising distinction in the relative orientation of interface helices, with a tendency for aligned packing in homodimers and a clear preference for packing at 90° in heterodimers. Arg and the aromatic residues have a higher preference to occur in all secondary structural elements (SSEs) in the interface. Based on the dominant SSE, the interfaces have been grouped into four classes: $\alpha$, $\beta$, $\alpha\beta$ and non-regular. Identity between protein and interface classes is the maximum for $\alpha$ proteins, but rather mediocre for the other protein classes. The interface classes of the two chains forming a heterodimer are often dissimilar. Eleven binding motifs can capture the prominent architectural features of most of the interfaces.

**Contact:** pinak@boseinst.ernet.in

**Supplementary information:** A separate file is provided with 3 tables and 2 figures, which are referred to with a prefix 'S' in text.

## 1 INTRODUCTION

The association and dissociation of protein molecules regulate most biological processes and considerable efforts have gone into understanding protein interactions. X-ray crystallography provides the direct snapshot of the interface formed when two protein molecules associate and the Protein Data Bank (PDB) (Berman *et al.*, 2000) is the repertoire of the wealth of such information.

Association between two (or more) protein chains can be classified as strong and permanent ('obligate') or as weak and transient ('non-obligate') (Jones and Thornton, 1996). In the former, the protein subunits occur only in the complexed state, as exemplified by protein quaternary structures such as the homodimers (Bahadur *et al.*, 2003). Protein molecules that usually exist independently but form complexes depending on factors such as physiological conditions, chemical modifications, binding of ligands, etc., form non-obligate interactions (Chakrabarti and Janin, 2002; Janin and Chothia, 1990; Lo Conte *et al.*, 1999). These two types of interfaces can differ in physicochemical characteristics, most notably interface area (Bahadur *et al.*, 2004). Biological interfaces have been characterized in terms of the secondary structure elements at their interaction sites (Argos, 1988; Dou *et al.*, 2004; Hoskins *et al.*, 2006; Miller, 1989; Neuvirth *et al.*, 2004). However, there has been no attempt to compare the properties of the secondary structure elements in these two interface categories, and more importantly, if increasing interface size affects these properties. Although various interaction databases and prediction servers exist – 3DID (Stein *et al.*, 2005), PIBASE (Davis and Sali, 2005), InterPreTS (Aloy and Russell, 2003), DOCKGROUND (Douguet *et al.*, 2006), PROTCOM (Kundrotas and Alexov, 2006), SCOPPI (Winter *et al.*, 2006), etc.—these do not usually distinguish intrachain (domain–domain) interactions from interchain interactions, and in the latter category, between obligatory and non-obligatory interactions. This obfuscates the visualization of any pattern involving geometrical and structural aspects of protein–protein interactions.

The basic forces (close packing, hydrophobic effects, shape complementarity between associating parts, electrostatic considerations, etc.) that determine the tertiary structure of proteins appear to be similar to the ones that regulate the processes of protein–protein recognition and binding (Tsai and Nussinov, 1997; Tsai *et al.*, 1996, 1997; Saha *et al.*, 2007). Investigating the structural properties of the recognition sites in the two distinct types of interfaces and their comparison to what is seen in protein tertiary structures should provide insights into the inter-related processes of protein folding and protein binding. This article focuses on the secondary structures of interface residues; the characterization of peptide segments at the interface in terms of secondary structure and their association across the recognition surface. These are in turn organized to form certain recognizable motifs that recur in the interfaces between unrelated proteins.

*To whom correspondence should be addressed.

The universe of protein folds has been divided into classes. It has been suggested that the total number of interaction types (proteins sharing similar sequences tend to interact similarly) is limited. According to estimates, most interactions in nature will conform to one of about 10 000 types (Aloy and Russell, 2004), like the 1000 protein folds suggested by Chothia (1992). Although there are databases dealing with protein interfaces, there have been no attempts to classify them along the terms used for fold classification, something that we have attempted here. An offshoot would be to study the correlation between protein class and interface class, and also if the binding sites of the two partner molecules have identical interface classes.

## 2 METHODS

### 2.1 Datasets used and initial calculations

This study uses two sets of non-redundant protein–protein interfaces—the first being a group of 122 homodimers (Bahadur *et al*., 2003) and the second of 204 protein–protein heterocomplexes (Pal *et al*., 2007). Atomic coordinates were obtained from the Protein Data Bank (PDB) (Berman *et al*., 2000). Identification of interface residues was carried out using the program ProFace (Saha *et al*., 2006). DSSP (Kabsch and Sander, 1983) was used for secondary structure assignments. The secondary structure types considered were: $\alpha$- and $3_{10}$-helix, $\beta$-strand, turn (involving or not involving hydrogen bond) and the unclassified residues (assigned ‘ ’ by the program). Turn and unclassified residues were together assumed to constitute the non-regular (NR) regions.

### 2.2 Calculation of propensities

The propensity $(P_i)^{\mathrm{SSE}}$ of a residue *i* to occur in a given secondary structural element (SSE) is calculated as follows:

$$(P_i)^{\mathrm{SSE}} = \frac{(n_{\mathrm{i,sse,int}}/N_{\mathrm{sse,int}})}{(n_{\mathrm{i,sse,total}}/N_{\mathrm{sse,total}})},$$

where $\boldsymbol{n}_{\mathbf{i,sse,int}}$ and $\boldsymbol{N}_{\mathbf{sse,int}}$ are the counts of residue *i* and of all residues belonging to a particular secondary structure type in the interface, respectively; $\boldsymbol{n}_{\mathbf{i,sse,total}}$ and $\boldsymbol{N}_{\mathbf{sse,total}}$ are the corresponding counts in the entire tertiary structure. Usually, the normalization is done such that the two factors in the denominator are for the whole database; by restricting these to a given SSE in the present definition, the preference of a residue for that SSE in the interface is compared to that for the same SSE in the overall structure. Thus a value greater (or less) than 1.0 indicates that the residue is observed more (or less) in that SSE when in the interface than in the rest of the structure.

### 2.3 Definition of secondary structural segments (SSSs)

Interface residues along the polypeptide chain were organized into secondary structural segments (SSSs) on the basis of secondary structure. Each segment consists of interface residues that are close in the primary sequence and located on the same secondary structural element—helix (a contiguous stretch of $\alpha$- and $3_{10}$-helices was assumed to constitute a single helix), strand and non-regular region (an element of which would encompass a continuous stretch devoid of any helix or $\beta$-strand residues). A segment could be an entire SSE or a part of it, being bounded by the two extreme interface residues on that element; there could be intervening non-interface residues in a segment. Each interface was thus divided into a series of helical, strand and non-regular segments, labeled H, S and NR, respectively; each numbered sequentially from the N-terminal onwards. The labels of the SSSs in the second chain had a ‘′’ symbol suffixed—thus the SSSs from the two subunits could be distinguished (H2 and H2′, for example).

### 2.4 Identification of SSS pairs and the calculation of surface area buried between them

We identified all interface atom–atom pairs that were within 4.5 Å (Saha *et al*., 2005) [as calculations using atom counts, rather than residues, provide more accurate results (Saha and Chakrabarti, 2006)]. An atom may have multiple interface contacts (within the threshold value) and the shortest one was selected. Tracing back to the secondary structures of the involved residues allowed us to assemble statistics on the number of contacts between SSS pairs. We also estimated the buried area between each SSS pair. When interface atom ‘A’ from chain 1 (belonging to SSS ‘X1’, for example) has atom ‘B’ from the chain 2 (a part of the SSS ‘X2′’, say) as its shortest contact, the buried area of atom ‘A’ was taken as contributing to the area buried between the SSS pair (‘X1-X2′’). This operation was performed sequentially for all the interface atoms (in both the chains). The surface areas buried between the different SSS pairs added up to the total interface area (or very close to it).

### 2.5 Calculation of packing angle between SSS pairs

The following algorithm computes the angles between two helices or two strands that are packed across the interface: the program takes as input the entire length of the two secondary structural elements containing the two SSSs. If, however, an SSE is kinked, for example, when a helix is a composite of $\alpha$- and $3_{10}$-helices (Pal *et al*., 2005) there is usually an asymmetry in the area buried on the two sides of the kink and the side having the maximum number of interface contacts was considered. A model helix/strand having its axis along the *z*-axis was superposed onto the input structures. The transformed *z*-axis provided the axes of each of the two SSEs, which were then used to calculate the angle.

### 2.6 Classifying interfaces according to secondary structural features

All the interfaces were distributed into four classes [$\alpha$,$\beta$, mixed $\alpha\beta$ and non-regular (NR)] according to the overall secondary structure composition of the interface residues. The following criteria were used: $\alpha$ interfaces must contain at least 40% interface residues in helix and $<10\%$ in strand; likewise, $\beta$ interfaces must contain at least 40% interface residues in strand with $<10\%$ participating in helices; mixed interfaces must possess at least 40% interface residues in helices and strands, with at least 10% in one of the groups; lastly, NR interfaces must have >60% residues with backbone conformations corresponding to turn, loop or other unstructured regions. This methodology ensured that we can cover all the interfaces and adequately represent what we visualize using a molecular graphics program.

## 3 RESULTS

### 3.1 Secondary structure composition of interfaces

Forty percent residues in homodimeric interfaces are helical, significantly higher than the 26% in heterocomplexes (Table 1). In homodimers the contribution of $\beta$-strands is low compared to helical residues (19% versus 40%), whereas they contribute comparably in complexes (24% versus 26%). Non-regular structures (including coils, turns and loops) appear in large numbers in both, but form the single largest group in heterocomplexes. Grouping helical and strand residues as ‘regular’ and the remainder as ‘non-regular’ structures, we find a statistically significant preference for the former in the homodimeric interfaces. We also decided to study the influence of interface size on the relative content of regular and non-regular structures (Fig. 1A). For the homodimers, the composition does not show variation with size. For the heteromers, however, regular secondary structures become more abundant as the proteins form larger interfaces (‘regular’ increases from ~40 to ~64% as the interface size increases 10-fold).

## 3.2 Secondary structure preferences of interface residues

Plots of the propensities of occurrence of all 20 amino acids in three secondary structure elements (SSEs) when located in the interface compared to the total protein structure are given in Figure 2. Arg and the aromatic residues are consistently observed more in all interface SSEs. Interface strands appear more hydrophilic than a buried strand in the protein interior with significant enrichment of Arg (and also Asp in heterocomplexes). Increased hydrophobicity of non-regular (NR) regions in the interface (to facilitate burial) is achieved by a higher percentage of aromatics, Met (and Cys in complexes); Met is a preferred residue in all the SSEs in homodimers. Ala, which has a high helical propensity in proteins, is used less in interface helices. Likewise, Val, Leu and Ile, which have high $\beta$-sheet propensities, are less prominent in the interface.

**Table 1.** Statistics on the distribution of regular and non-regular structures in interfaces

| Dataset | Frequency[a] | | $P$-values[b] (Regular versus Non-regular) |
|---|---|---|---|
| | Regular [Helix, Strand] | Non-regular | |
| Homodimers | 0.59 (0.1) [0.40, 0.19] | 0.41 (0.1) | 3.38 $E$−09 |
| Complexes | 0.50 (0.1) [0.26, 0.24] | 0.50 (0.2) | 0.49 |

[a] Standard deviations are in parentheses. Regular structures are separated [in square brackets] into individual contributions from helix and strand.

[b] For the Student’s $t$-test for paired samples, a $P$-value of $<0.01$ implies that the observed difference between the frequencies of regular and non-regular structures has a probability of $<1\%$ to occur by mere chance.

## 3.3 Pairing of interface secondary structures

Statistics on pairing of the SSEs (Fig. 3) show differences between the two datasets. Homodimer interfaces are mainly composed of helix–helix, helix–NR and NR–NR pairings.

Heterocomplex interfaces have reduced helix–helix packing, and instead, pairings involving NR regions are prominent. Helix–strand and strand–strand combinations are under-represented in both the categories. Suppression of helix–strand pairing can be attributed to their poor steric complementarity (Jiang *et al.*, 2003). Although fewer in number, strand–strand pairs dominate the interface architecture in individual cases (discussed later).

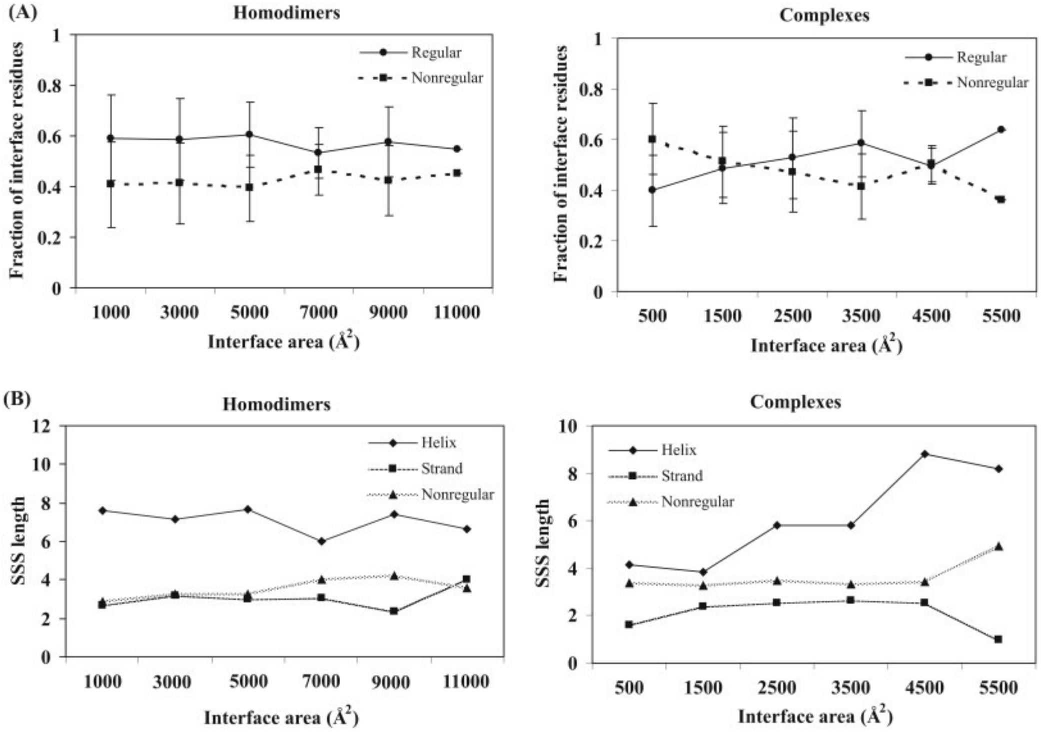

**Fig. 1.** Plots of (**A**) the fraction of interface residues occurring in regular (helix and strand) and non-regular (the rest) structures, and (**B**) the average lengths of the three different types of SSSs (helix, strand and non-regular) as a function of the interface area (considering the contribution of both the subunits). Interfaces are grouped according to their size into bins of 2000 Å$^2$ (homodimers) and 1000 Å$^2$ (complexes); the average values for each bin are then calculated.

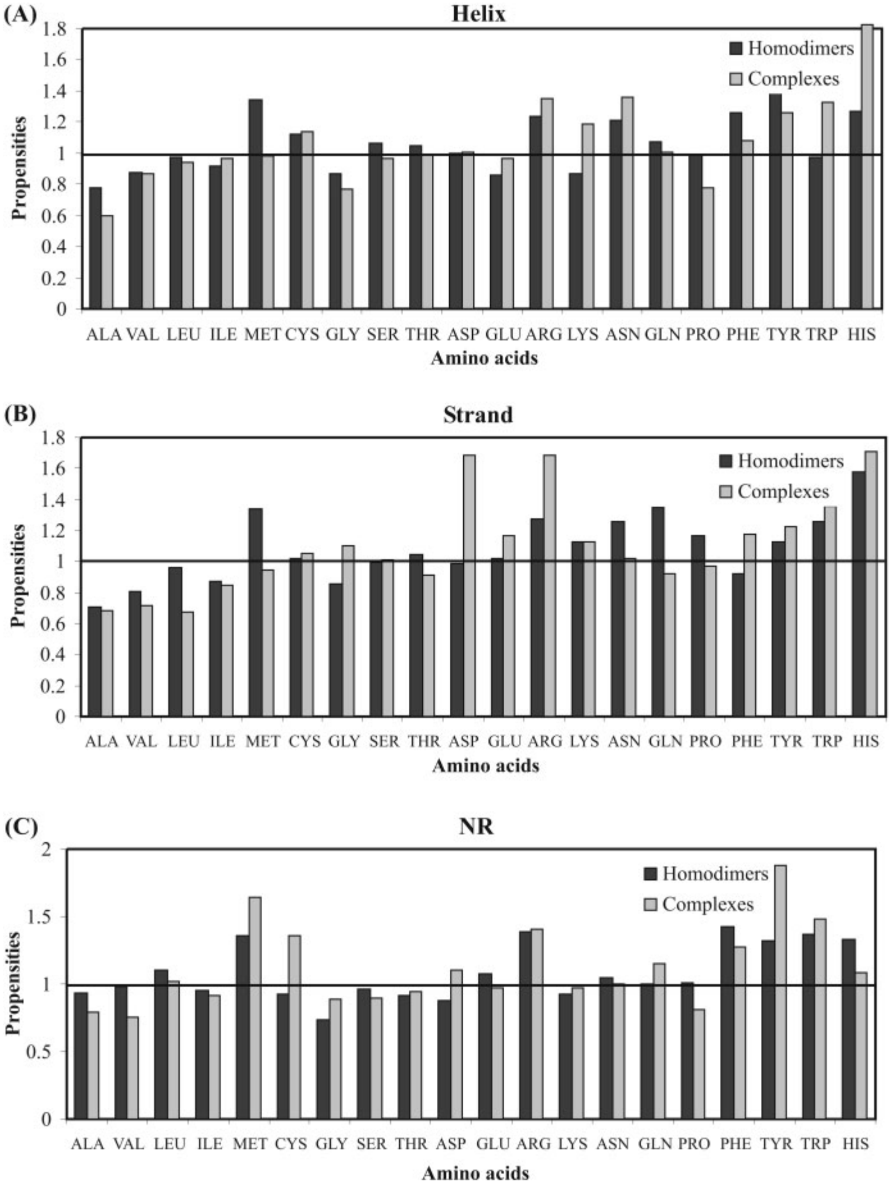

**Fig. 2.** Propensities of residues to occur in a particular secondary structure type ('Helix', 'Strand' and 'NR') in the interface. The residues are arranged according to the environment-based classification of amino acid residues (Guharoy and Chakrabarti, 2005).

### 3.4 Dissection of the recognition surface into secondary structural segments

The binding surface of each protein chain was divided into secondary structural segments (SSSs), demonstrated using a specific example in Figure 4. Each SSS comprises of a series of interface residues that are close in the primary sequence and also occur within the same SSE—helix, strand or non-regular (details in Methods section). Two interface SSSs can be contiguous (segments 5 and 6, for example) or separated by a non-interface region (segments 1 and 2). An SSS can also consist of just one residue (segment 8).

The SSSs are then characterized in terms of their numbers and lengths (Table 2), which are useful parameters for assessing their relative importance for protein–protein association. The SSSs are more numerous in homodimer interfaces than in heterocomplexes. However, as the former is on average twice the size of the latter (Bahadur *et al.*, 2004) upon normalization we find 8.7 SSSs per 1000 Å$^2$ of interface area for the homodimers versus 11 for heterocomplexes. Helices are significant contributors to homodimeric interfaces, with an average interface possessing nine helices, each having a length of 7.2 residues—significantly longer than the average lengths of both strands and unstructured segments (3.0 and 3.3). In contrast, none of the structural segment types are conspicuous by their lengths in heterocomplexes. Interestingly however, average lengths of helical interface segments increases

| | HELIX | STRAND | NR |
|---|---|---|---|
| HELIX | 22.4<br>(10.9) | 5.6<br>(7.1) | 26.0<br>(25.4) |
| STRAND | | 8.8<br>(8.4) | 14.3<br>(20.9) |
| NR | | | 22.8<br>(27.3) |

**Fig. 3.** Interface secondary structure pairing matrix. The values for homodimers are followed by those for heterocomplexes in brackets.

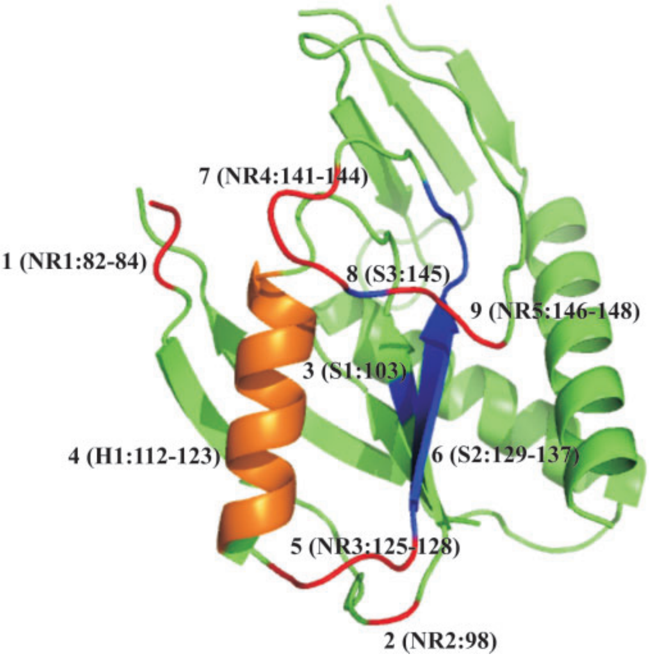

**Fig. 4.** Secondary structural segments (SSSs) (helix in orange, strand in blue and NR in red, with the rest of the structure in green) defining the interface of subunit A of the homodimeric structure with the PDB code, 1A3C. The serial number of the SSS, its identifying label and the residue range are provided.

**Table 2.** Statistics on secondary structural segments in interfaces

| Feature | Homodimers | Complexes | *P*-values |
|---|---|---|---|
| Number of SSSs | 32.0 (18.8) | 19.2 (8.0) | 1.71*E*−11 |
| Number of SSSs per 1000 Å$^2$ interface area | 8.7 (2.9) | 11.0 (3.3) | 2.52*E*−11 |
| Number of helices | 9.0 (5.9) | 4.2 (3.4) | 7.97*E*−15 |
| Helix length | 7.2 (5.0) | 4.8 (4.4) | 3.71*E*−05 |
| Number of strands | 7.4 (7.4) | 5.7 (4.0) | 0.01 |
| Strand length | 3.0 (1.5) | 2.4 (1.1) | 7.39*E*−05 |
| Number of NR segments | 15.7 (9.5) | 9.4 (4.0) | 5.65*E*−11 |
| Length of NR segments | 3.3 (1.2) | 3.3 (1.1) | 0.78 |

Average values of the number and the length of SSSs, including both the subunits forming the interface. The SDs are in parentheses.

**Table 3.** Statistics on the contribution of pairs of regular and non-regular SSSs to the interface

| Dataset | % of SSS pairs | | *P*-values (Regular versus Non-regular) |
|---|---|---|---|
| | Regular (H-H, S-S, H-S) | Non-regular (H-NR, S-NR, NR-NR) | |
| (A) Based on number | | | |
| Homodimers | 46 (35) | 54 (35) | 0.17 |
| Complexes | 25 (23) | 75 (23) | 1.9*E*−40 |
| (B) Based on buried surface area | | | |
| Homodimers | 25 (24) | 23 (21) | 0.54 |
| Complexes | 19 (20) | 56 (25) | 1.8*E*−32 |

Only major SSS pairs are considered in this analysis (H, S and NR stand for helix, strand and non-regular region). Two sets of values are provided, (A) gives the fraction of regular and non-regular SSS pairs out of the total number of major SSS pairs; (B) is on the basis of area buried only by the major SSS pairs relative to the total interface area (and as such, the two values for a given dataset do not add up to 100—the difference is contributed by the non-major SSS pairs). The SDs are in parentheses.

significantly as bigger interfaces are formed by the heterodimers (Fig. 1B), a variation not observed in homodimers.

## 3.5 Association between SSSs and their relative contribution to interface formation

The SSSs described above pack across the interface to form SSS pairs and the extent of their interaction can be quantified in terms of the accessible surface area (ASA) buried. Higher this value, greater is the contribution of the SSS pair to the interface and concomitantly, more important its role in interface formation and stabilization. We chose a cutoff of 5% of the total interface area to identify the major SSS pairs. The contributions of regular and non-regular SSS pairs to the two different types of interfaces were enumerated (Table 3). In homodimers, the regular SSS pairs contribute almost half of the interface on average, whereas for the heterocomplexes, they contribute significantly less (about one-third) compared to pairs involving non-regular segments. This is true both in terms of numbers and the area buried.

The contribution made by the SSS pairs to the interface area is plotted in Fig. S1. Most (85–90%) helix–helix and strand–strand pairs bury <20% interface area, though there are instances (mostly in homodimers) where a single HH or SS pair contributes more (Table S1A). Helix–strand pairs burying >10% interface area are common in complexes, but extremely rare in homodimers. Non-regular SSS pairs contributing >20% area is almost non-existent in homodimers, but occur often in heterocomplexes (Table S1B).

## 3.6 Packing geometry of the SSS pairs

The packing angles between helix–helix and strand–strand pairs were calculated. Due to geometrical differences between the two types of helices ($\alpha$ and $3_{10}$), we separated $\alpha$-helix pairs from those involving at least one $3_{10}$-helix. Furthermore, we selected only those $\alpha$-helical pairs where each helix had at least eight residues. This method retains the pairs that bury moderate to large extents of the total interface and usually, only these are important for binding. Almost all the $3_{10}$ helices were three residues long and only a few extended upto four residues.

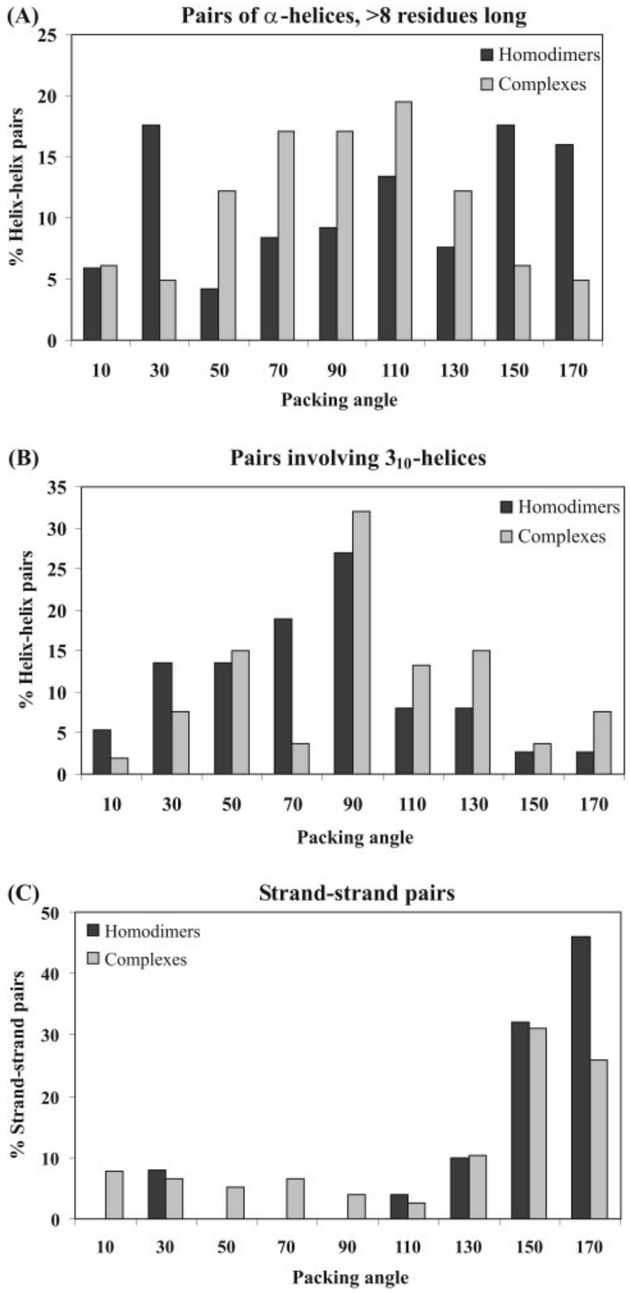

**Fig. 5.** The distribution of angles in (°) between (A,B) helix axes and (C) interacting strands. In (A) only $\alpha$-helices that are at least 8 residues long are considered, in (B) at least one of the helices is of the type $3_{10}$-, the other could be $\alpha$- or $3_{10}$-.

The packing angle distributions have reasonably clear distinctions (Fig. 5). $\alpha$–$\alpha$ pairs in homodimers have a preference for parallel or antiparallel orientations (angle $<40°$ or $>140°$, respectively). For heterocomplexes, there is a reversal of the above trend, with the peak occurring at ~90°. Pairs involving $3_{10}$-helices show a large preference to pack around 90° in both datasets. The preferential angle for packing between two interface strands (Fig. 5C) indicates antiparallel orientation, which is almost exclusively observed in homodimers.

We also investigated the relative usage of $\alpha$ and $3_{10}$ helix–helix pairs across the interface. The homodimer and heterocomplex datasets contain 174 and 170 pairs, out of which those having one or both $3_{10}$ helices are 37 (21.3%) and 55 (32.4%), respectively. This indicates a possibly greater role of $3_{10}$-helix pairs in heteromeric interfaces.

## 3.7 Classifying interfaces based on the prevalence of secondary structural elements

Analogous to protein structural class assignments, we grouped interfaces based on the proportion of interface residues belonging to helix, $\beta$-strand or non-regular (NR) regions. Four classes are identified: $\alpha$, $\beta$, mixed ($\alpha\beta$) and NR. While $\alpha/\beta$ and $\alpha+\beta$ are distinct protein classes, we have used just one $\alpha\beta$ class for the interfaces. Figure 6 shows the distribution of the interfaces among the different classes.

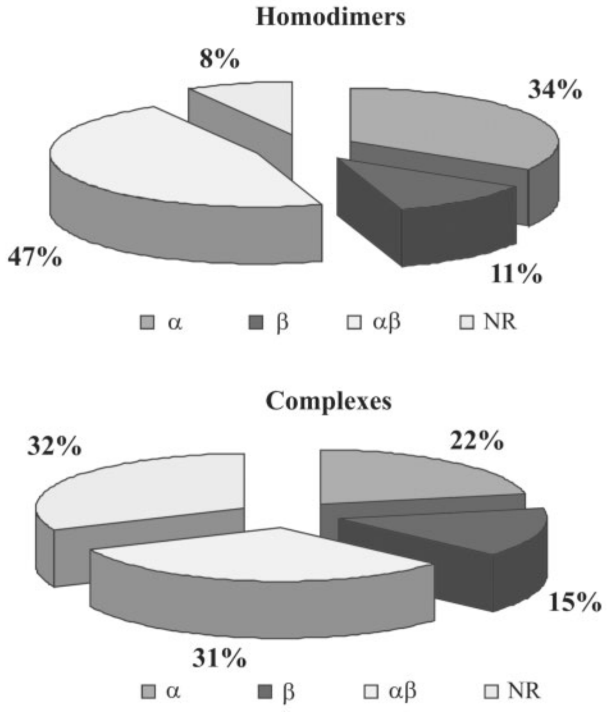

**Fig. 6.** Pie-charts showing the distribution of four classes of interfaces: $\alpha$, $\beta$, $\alpha\beta$ and NR.

Both $\alpha$ and $\alpha\beta$ interfaces are more abundant in homodimers (34 and 47%, respectively) when compared to heterocomplexes (22 and 31%, respectively). $\beta$ interfaces are almost equally abundant in both the datasets. However, more protein–protein complex interfaces belong to the NR-type (32%) compared to the homodimers of which only a mere 8% are of this type.

A pertinent question here is whether the interface structural class is dependent upon the protein tertiary structural class. The degree of correspondence between the protein and interface classes is shown in Table 4. The identity is maximum (91%) for $\alpha$ proteins in heterocomplexes, and somewhat lesser (79%) for homodimers. The match is mediocre (55–59%) for the mixed classes. For $\beta$ it is comparable (67%) only for homodimers, but poor (26%) for heterocomplexes (47% of which use non-regular regions for complexation). Another interesting question that can be addressed relates to the equivalence of the interface classes of the two interacting chains in heterocomplexes. Results shown in Table S2A indicate that when the binding region of one protein chain is of the class $\alpha$, $\beta$, $\alpha\beta$ or NR, the interface class of the partner would be identical in only 29, 17, 22 and 26% cases, respectively. When the equivalence of the interface classes from enzyme-inhibitor and antigen–antibody complexes was analysed separately (Tables S2B and C), mostly NR regions from both the partners were found in the interface.

## 3.8 Conservation of residues in different interface classes

Interfaces can be dissected into core and rim regions (Chakrabarti and Janin, 2002) and the residues belonging to the core are usually more conserved than those in the rim, as indicated by the mean sequence entropy values obtained from an alignment of homologous proteins ($\langle s\rangle_{\mathrm{core}} < \langle s\rangle_{\mathrm{rim}}$)(Guharoy and Chakrabarti, 2005). We compared these values between interface classes (Table 5). The interfaces belonging to $\alpha$ class are more conserved than the rest, based on both the average values ($\langle s\rangle_{\mathrm{core}}$ and $\langle s\rangle_{\mathrm{rim}}$), as well as their ratio. The overall trend of the mean sequence entropy of the core being less than that of the rim is maintained even when the interfaces are split into classes, except for the $\beta$ class where the rim region seems to be as conserved as the core making the ratio close to 1. Even from the distribution of sequence entropies of individual interfaces (Fig. S2) it can be seen that both the core and rim regions in $\alpha$ class have lower values (60 and 72% of the cases are $<0.80$ in homodimers and complexes, respectively) compared to NR (90 and 54% $>$ 0.80) and mixed interfaces (60 and 67% >0.80).

**Table 4.** The match between SCOP (Andreeva et al., 2004) class of individual chains and the corresponding interface class

| SCOP class (Tertiary structure) | Interface class | | | |
|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\alpha\beta$ | NR |
| (A) HOMODIMERS (113 cases)[a] | | | | |
| $\alpha$ (29 cases) | 23 | 0 | 6 | 0 |
| $\beta$ (15) | 1 | 10 | 1 | 3 |
| $\alpha/\beta$, $\alpha+\beta$ (65) | 18 | 5 | 38 | 4 |
| Others (4) | 0 | 2 | 2 | 0 |
| (B) PROTEIN–PROTEIN COMPLEXES (396 cases)[a,b] | | | | |
| $\alpha$ (58 cases) | 53 | 0 | 3 | 2 |
| $\beta$ (115) | 1 | 30 | 30 | 54 |
| $\alpha/\beta$, $\alpha+\beta$ (152) | 20 | 18 | 84 | 30 |
| Others (71) | 14 | 10 | 9 | 38 |

[a] In a few cases, the interface is formed by more than one domain having different SCOP assignments, which precludes a direct comparison; 9 homodimers and 12 heterocomplexes were thus excluded.

[b] For protein–protein complexes, the comparison was carried out for both the subunits separately; for homodimers, this was not necessary because of the identical nature of the two associating chains.

## 3.9 Interface architectures

The SSS pairs combine to form recurring super-structures. The packing of major SSSs in individual classes was inspected visually to identify interface motifs. Interfaces that are classified as helical contain at least a pair of interacting helices, but very often contain two (and sometimes more) pairs of helices. Depending on the number and interaction patterns of the helices, and analogous to what is observed in tertiary structures, we identified four distinct motifs: single helix–helix pair (Figs 7A and B), 4-helix bundle, $\alpha$-sandwich and coiled-coil (Table S3). Six types of sub-geometries are observed in four-helix bundles occurring in protein interiors and interfaces (Harris *et al.*, 1994; Lin *et al.*, 1995). In homodimers and heterocomplexes, the numbers observed in the various types of bundles are: square (13 and 3), splinter (7 and 2), X (17 and 4) (Fig. 7C), unicornate (18 and 5), bicornate (18 and 10) and splayed (10 and 11). While unicornate and bicornate are the favoured arrangements in homodimers, the preference is for splayed geometry [opposite to what is statistically expected (Lin *et al.*, 1995)] in heterocomplexes, again showing the subtle differences in the two interface categories. When more than two pairs of helices occur side-by-side in aligned orientations the motif is termed $\alpha$-sandwich (Fig. 7D). The intertwined helices in coiled-coil motifs are typically long and often these alone make up the entire protein chain (Fig. 7E).

The next three architectural motifs involve $\beta$ structures and are observed in $\beta$ interfaces and also to some extent

**Table 5.** Average sequence entropy values in the core and rim regions of different interface classes

| Class | $\langle s\rangle_{\mathrm{core}}$ | $\langle s\rangle_{\mathrm{rim}}$ | $\langle s\rangle_{\mathrm{core}}/\langle s\rangle_{\mathrm{rim}}$ |
|---|---|---|---|
| (A) Homodimers (121 cases) | | | |
| $\alpha$ | 0.54 | 0.68 | 0.82 |
| $\beta$ | 0.65 | 0.64 | 1.0 |
| $\alpha\beta$ | 0.65 | 0.84 | 0.85 |
| NR | 0.84 | 0.92 | 0.94 |
| (B) Complexes[a] (364 cases) | | | |
| $\alpha$ | 0.40 | 0.52 | 0.82 |
| $\beta$ | 0.60 | 0.69 | 0.95 |
| $\alpha\beta$ | 0.72 | 0.87 | 0.86 |
| NR | 0.62 | 0.77 | 0.81 |

[a] Excluding antibody–antigen complexes (Guharoy and Chakrabarti, 2005).

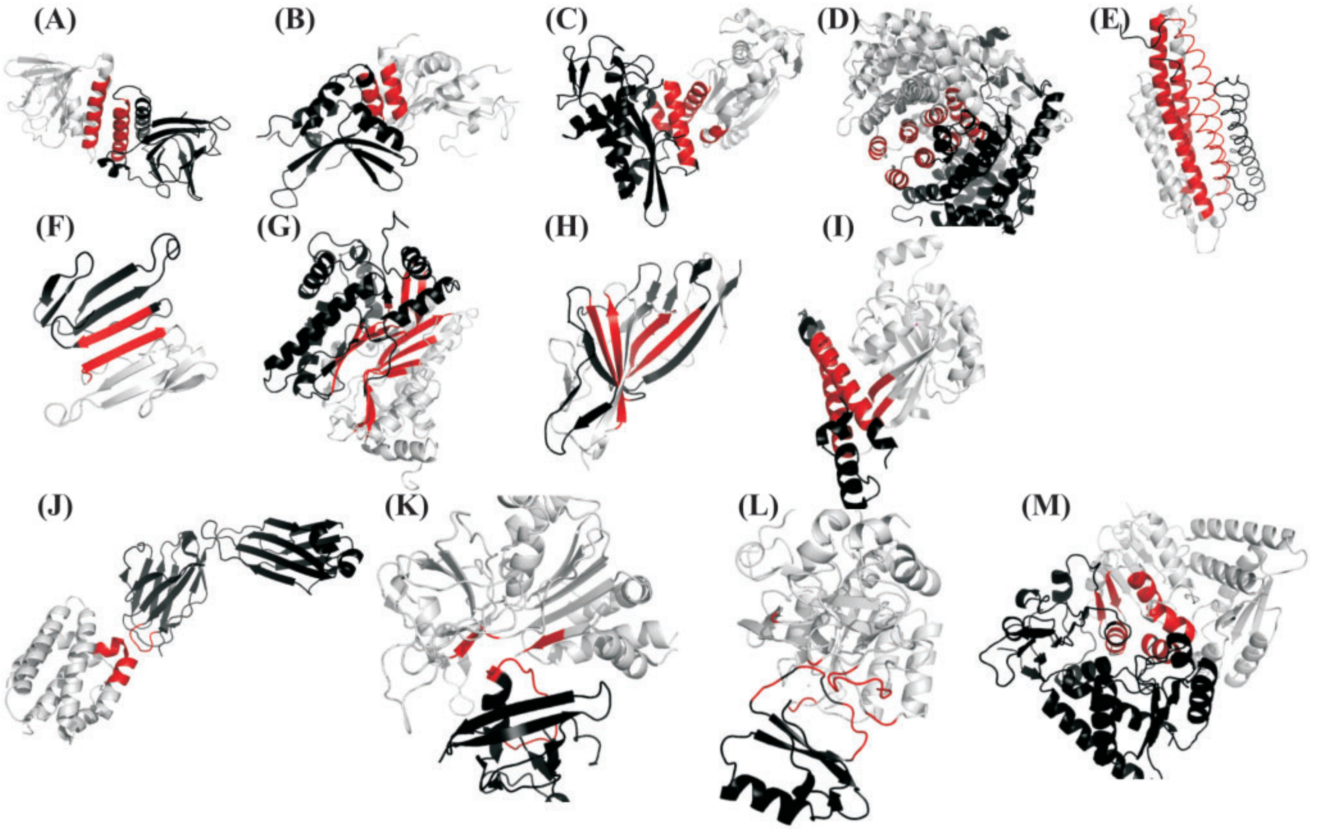

**Fig. 7.** Examples of interface motifs and different modes of packing of the SSSs. Single helix–helix pair with (**A**) antiparallel orientation in 2ARC (PDB code), and (**B**) parallel orientation in 1AF5. (**C**) 4-Helix bundle in 1BAM; (**D**) $\alpha$-sandwich in 1CSH; (**E**) coiled-coil in 2LIG; (**F**) continuous $\beta$-sheet in 1KBA; (**G**) $\beta$-sandwich in 1B5E; (**H**) mixed $\beta$ in 1CDC; (**I**) helix-sheet in 1CXZ; (**J**) helix-NR in 1LK3; (**K**) strand-NR in 1EWY and (**L**) NR–NR in 2TEC. An example of an interface with two distinct motifs (continuous $\beta$-sheet and 4-helix bundle) is shown in (**M**) for 1A4I. Different levels of shading are used to distinguish the two interacting subunits, with the motif of interest shown in red. Structures shown in (A,B,C,D,E,F,G,H and M) are homodimers and the rest are heterocomplexes.

in $\alpha\beta$ interfaces. The most common is the continuous $\beta$-sheet, formed by interface strands coming side-on from both subunits. The total number of strands in the complete $\beta$-sheet varies from 2 (one strand from each chain) to a maximum of 16 with an average of 7 and 6.2 in the two categories; an example with 6 strands is provided in Fig. 7F. The two interface strands are usually hydrogen bonded in an antiparallel fashion, with only two exceptions in homodimers and ten in heterocomplexes. The second $\beta$-motif has face-to-face packing of $\beta$-sheets and is termed the $\beta$-sandwich (Fig. 7G). In some homodimers this motif constitutes the entire interface. When the above two motifs exist simultaneously (two continuous $\beta$-sheets packing against each other to form a $\beta$-sandwich) we have the mixed $\beta$-motif (Fig. 7H). The helix-sheet motif (Fig. 7I) is more prevalent in heterocomplexes. Mostly one helix is involved from one side and the other side may have just one strand, an entire sheet belonging to a single chain or a continuous $\beta$-sheet motif (discussed above). The remaining three motifs (helix/strand/NR–NR) are more numerous in heterocomplexes (Fig. 7J–L), often comprising the entire interface. All the important contacts are provided by these motifs, whereas in homodimers these play a subordinate role to the more dominant regular motifs.

## 4 DISCUSSION

### 4.1 Relative contribution of secondary structural elements

Different types of protein–protein interfaces may exhibit differences in physicochemical features (Bahadur *et al*., 2004; Ofran and Rost, 2003; Saha *et al*., 2005), which can also be seen in the two datasets used here. Data presented in Table 1 indicate that homodimers (with obligate interfaces) have more helices at the interface, with a percentage composition that essentially reproduces what has been reported (Tsai *et al*., 1997); in contrast, non-obligatory interfaces (in heterocomplexes) have a higher participation of non-regular regions, as noted recently (Ansari and Helms, 2005; De *et al*., 2005). Of greater interest however, is the fact that in heterocomplexes the involvement of regular secondary structures tends to increase with interface size (Fig. 1A). This is due to the presence of longer helical segments (Fig. 1B).

In homodimers, the contribution of regular SSS pairs are almost the same as that of the non-regular SSS pairs (Table 3). They have prominent pairs of regular SSSs with non-regular pairs stabilizing them. Heteromeric interfaces switch between exposed and buried states and must closely mimic the properties of a generic protein surface patch (otherwise the monomeric form will be unstable in solution), and are thus enriched in non-regular SSS pairs. Also their smaller size affects the choice of SSS pairs and their relative orientation (for helices in particular, Fig. 5A). Often, interactions involving NR segments are the only SSS pair types that can be detected in these complexes (Figs 7J–L).

### 4.2 Propensities of residues to be in an SSE in the interface as opposed to that within the tertiary structure

Figure 2 indicates that aromatic residues and Arg are enriched in interfaces. There are subtle differences between the two datasets. For example, Met is found more in interface helices and strands in homodimers than in heterocomplexes. Asp is found more in the interface strands in the latter—this result is in variance to a recent study (Hoskins *et al*., 2006) that found Asp to be underrepresented in interface strands. The presence of charged side-chains, such as Asp and Arg, in what would ordinarily be the hydrophobic side of the edge $\beta$ strand, may be a feature of negative design to avoid undesirable edge-to-edge aggregation (Richardson and Richardson, 2002). It is interesting to note that the three most common hot-spot (contributing more than 2 kcal/mol to the binding interaction) residues, Trp, Arg and Tyr (Bogan and Thorn, 1998), are also found more in interface SSEs. Asp is enriched in hot spots and also occurs with high propensity in interface strands—thus it may be worthwhile to see if Asp residues providing a large fraction of the binding free energy are actually located in strands.

### 4.3 Protein class versus interface class and functional implications

The first three interface classes ($\alpha$, $\beta$, $\alpha\beta$) are self-explanatory and are analogous to the ones found in protein domain classification databases [SCOP (Andreeva *et al*., 2004), CATH (Pearl *et al*., 2003)]. However, the inclusion of the NR-type interface has important connotations. Unlike protein 3D structures where unstructured regions are mainly responsible for linking the regular secondary structures, there are many complexes where the interface consists of pairs of interacting non-regular structural elements. For example, enzyme–inhibitor complexes favour using NR interface from both (23%) or at least one (48%) of the two partners, while 30% antibody–antigen complexes are of the NR–NR type and a further 39% use an NR interface from only one of the two participating protein components (Table S2B and C). On the contrary, a larger fraction (57%) of signalling complexes do not involve any NR interface on either side and there is no instance of an NR–NR combination (Table S2D). Thus the functionality of a molecule may have some influence on the interface class.

Keskin *et al*. (2004) divided interfaces into three broad types depending on the degree of similarity of the interfaces vis-à-vis their parent chains. Here, we ask whether interface class is likely to be the same as protein structural class? $\alpha$ classes of proteins are most likely to use helices in the binding region. Otherwise, the correlation is not very strong (Table 4); indeed, one striking mismatch can be seen in Fig. 7A, where a mainly $\beta$ protein (2ARC—classified by SCOP as a $\beta$-protein containing a double-stranded beta-helix fold) has $\alpha$ interface class. The oligomerization interface contains three helices from each of the two subunits. An interesting difference between the two datasets is that a large number of heteromers (except the ones having $\alpha$-protein class) form 'NR' interfaces. Antibody molecules are very good examples of $\beta$-class proteins forming NR interfaces while binding. Fifty-four percent of the 'Others' class proteins and nearly 20% mixed-class protein complexes use 'NR' interfaces for specific binding.

### 4.4 Structural motifs in protein–protein interactions

Interface class usually guides the nature of the binding motif. $\alpha$ interfaces primarily contain helical motifs, with additional stabilization from helix–NR or NR–NR; however, it is highly unlikely that the motifs would involve strands. The opposite is true for $\beta$ interfaces. In the mixed interfaces, the motifs may contain helices or strands or both simultaneously. Lastly, the principal motifs in NR interfaces contain non-regular regions interacting with each other or with short segments of helices/strands from the other chain. In total, eleven binding motifs have been enumerated (Fig. 7 and Table S3). They are fairly broad, and one may use structural details for sub-classification. Interface motifs have been previously discussed in different contexts (Dou *et al*., 2004; Jones and Thornton, 1996; Keskin and Nussinov, 2005; Tsai *et al*., 1997), and some of these architectures are quite similar to those in the protein cores (Miller, 1989; Tsai *et al*., 1997). The motifs are not mutually exclusive and some interfaces may harbour more than one motif, as shown in Fig. 7M, which has a 4-helix bundle, as well as a continuous $\beta$-sheet. Functionally different proteins employing similar motifs for interface construction probably represent examples of convergent evolution, reinforcing the hypothesis that the existence of a limited number of folds in nature may be extended to the realm of protein–protein interactions as well.

## 5 CONCLUSIONS

The secondary structural elements have different importance in mediating interactions in the two types of interfaces formed by homodimers and heterocomplexes—helices are common in the former and the non-regular structures in the latter (Table 1). However, the complexes tend to switch the SSE preferences from non-regular to regular as larger interfaces are formed (Fig. 1). Helical segments in the two interface types show the largest distinction both in terms of average number per interface and average length (Table 2), contributing more towards homodimeric interfaces, in which helix–helix and helix–NR pairings are more prevalent, while NR–NR/H/S are observed more frequently in complexes (Fig. 3). The non-regular SSS pairs occupy three-quarters of an average hetero-interface (Table 3). The orientation of helix–helix pairs across the interface is surprisingly distinct, the homodimers showing a tendency for parallel or antiparallel packing, which is more near right angles in heterocomplexes (Fig. 5). However, the packed strand–strand pairs have similar features (the angle, >140°) in both the datasets. Classification of interfaces into four structural classes analogous to fold classification yields interesting results as well. The frequent use of helices in the construction of homodimer interfaces translates into a higher percentage of $\alpha$ and mixed ($\alpha\beta$) interfaces (Fig. 6). The primary use of non-regular regions in the hetero-interfaces manifests itself as higher proportion of NR interfaces compared to homodimers. It turns out that the structural classes of the interface and of the participating proteins do not have to be the same (Table 4). Residues in $\alpha$ class of interface show the highest degree of conservation (Table 5). The identification of recurring binding motifs (Fig. 7) indicates how simple patterns are used by nature to build large recognition surfaces. Lastly, aromatic residues and Arg have higher occurrences in the SSEs in interfaces relative to those within tertiary structures (Fig. 2).

## ACKNOWLEDGEMENTS

The work has been supported by a grant from the Department of Biotechnology, India. M.G. is a recipient of a fellowship from the CSIR, India.

*Conflict of Interest*: none declared.

## REFERENCES

Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.

Aloy,P. and Russell,R.B. (2004) Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.*, **22**, 1317–1321.

Andreeva,A. *et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Ansari,S. and Helms,V. (2005) Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, **61**, 344–355.

Argos,P. (1988) An investigation of protein subunit and domain interfaces. *Protein Eng.*, **2**, 101–113.

Bahadur,R.P. *et al.* (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.

Bahadur,R.P. *et al.* (2004) A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, **336**, 943–955.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.

Chothia,C. (1992) One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Davis,F. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.

De,S. *et al.* (2005) Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct. Biol.*, **5**, 15.

Dou,Y. *et al.* (2004) ICBS: a database of interactions between protein chains mediated by $\beta$-sheet formation. *Bioinformatics*, **20**, 2767–2777.

Douguet,D. *et al.* (2006) DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics*, **22**, 2612–2618.

Guharoy,M. and Chakrabarti,P. (2005) Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 15447–15452.

Harris,N.L. *et al.* (1994) Four-helix bundle diversity in globular proteins. *J. Mol. Biol.*, **236**, 1356–1368.

Hoskins,J. *et al.* (2006) An algorithm for predicting protein-protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. *Protein Sci.*, **15**, 1017–1029.

Janin,J. and Chothia,C. (1990) The structure of protein-protein recognition sites. *J. Biol. Chem.*, **265**, 16027–16030.

Jiang,S. *et al.* (2003) The role of geometric complementarity in secondary structure packing: a systematic docking study. *Protein Sci.*, **12**, 1646–1651.

Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Keskin,O. *et al.* (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

Keskin,O. and Nussinov,R. (2005) Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng. Design Selection*, **18**, 11–24.

Kundrotas,P.J. and Alexov,E. (2006) PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res.*, **35**, D575–D579.

Lin,S.L. *et al.* (1995) A study of four-helix bundles: investigating protein folding *via* similar architectural motifs in protein cores and in subunit interfaces. *J. Mol. Biol.*, **248**, 151–161.

Lo Conte,L. *et al.* (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

Miller,S. (1989) The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng.*, **3**, 77–83.

Neuvirth,H. *et al.* (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

Ofran,Y. and Rost,B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **325**, 377–387.

Pal,L. *et al.* (2005) $3_{10}$-helix adjoining $\alpha$-helix and $\beta$-strand: sequence and structural features and their conservation. *Biopolymers*, **78**, 147–162.

Pal,A. *et al.* (2007) Peptide segments in protein-protein interfaces. *J. Biosci.*, **32**, 101–111.

Pearl,F.M. *et al.* (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.

Richardson,J.S. and Richardson,D.C. (2002) Natural $\beta$-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA*, **99**, 2754–2759.

Saha,R.P. and Chakrabarti,P. (2006) Parity in the number of atoms in residue composition in proteins and contact preferences. *Curr. Sci.*, **90**, 558–561.

Saha,R.P. *et al.* (2005) Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface. *J. Proteome Res.*, **4**, 1600–1609.

Saha,R.P. *et al.* (2006) ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces. *BMC Struct. Biol.*, **6**, 11.

Saha,R.P. *et al.* (2007) Interaction geometry involving planar groups in protein-protein interfaces. *Proteins*, **67**, 84–97.

Stein,A. *et al.* (2005) 3DID: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **33**, D413–D417.

Tsai,C.-J. and Nussinov,R. (1997) Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. *Protein Sci.*, **6**, 1426–1437.

Tsai,C.-J. *et al.* (1996) Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.*, **31**, 127–152.

Tsai,C.-J. *et al.* (1997) Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci.*, **6**, 1793–1805.

Winter,C. *et al.* (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.