Protein Science (1997), 6:1793–1805. Cambridge University Press. Printed in the USA.

Structural motifs at protein–protein interfaces: Protein cores versus two-state and three-state model complexes

CHUNG-JUNG TSAI,[1] DONG XU,[2] AND RUTH NUSSINOV[2,3]

[1]Laboratory of Mathematical Biology, NCI-FCRF, Building 469, Room 151, Frederick, Maryland 21702
[2]Laboratory of Mathematical Biology, SAIC, NCI-FCRF, Building 469, Room 151, Frederick, Maryland 21702
[3]Sackler Institute of Molecular Medicine, Tel Aviv University, Tel Aviv 69978, Israel

(RECEIVED January 13, 1997; ACCEPTED May 7, 1997)

Abstract

The general similarity in the forces governing protein folding and protein–protein associations has led us to examine the similarity in the architectural motifs between the interfaces and the monomers. We have carried out extensive, all-against-all structural comparisons between the single-chain protein structural dataset and the interface dataset, derived both from all protein–protein complexes in the structural database and from interfaces generated via an automated crystal symmetry operation. We show that despite the absence of chain connections, the global features of the architectural motifs, present in monomers, recur in the interfaces, a reflection of the limited set of the folding patterns. However, although similarity has been observed, the details of the architectural motifs vary. In particular, the extent of the similarity correlates with the consideration of how the interface has been formed. Interfaces derived from two-state model complexes, where the chains fold cooperatively, display a considerable similarity to architectures in protein cores, as judged by the quality of their geometric superposition. On the other hand, the three-state model interfaces, representing binding of already folded molecules, manifest a larger variability and resemble the monomer architecture only in general outline. The origin of the difference between the monomers and the three-state model interfaces can be understood in terms of the different nature of the folding and the binding that are involved. Whereas in the former all degrees of freedom are available to the backbone to maximize favorable interactions, in rigid body, three-state model binding, only six degrees of freedom are allowed. Hence, residue or atom pair-wise potentials derived from protein–protein associations are expected to be less accurate, substantially increasing the number of computationally acceptable alternate binding modes (Finkelstein et al., 1995).

Keywords: hydrophobic effect; motifs; protein folding; protein–protein interfaces; protein–protein recognition; structural comparison

Although it is still not understood how a given sequence folds specifically into a particular protein conformation, the concept of recurring structural motifs such as the 4-helix bundles and the β-barrels has implicitly helped in outlining the principles of protein folding. In general, a folded globular protein always possesses a hydrophobic core (Bowie et al., 1990; Dill, 1990) enclosed by interacting secondary structure elements. Because the number of spatial arrangements of secondary structure elements to form a hydrophobic entity is limited, the number of unique protein folds is believed to be limited as well (Wang, 1996). Hence, the task of the classification of the motifs has provided an invaluable insight into the understanding of protein structure and in relating the biological function of the protein to a unique fold (Rossman et al., 1975; Lesk & Chothia, 1980; Miller, 1989; Grindley et al., 1993; Orengo et al., 1993; Slingsby et al., 1993; Holm & Sander, 1994; Fischer et al., 1995; Murzin et al., 1995).

Owing to the general similarity that is observed between protein binding and protein folding (Argos, 1988; Janin et al., 1988; Janin & Chothia, 1990; Jones & Thornton, 1996), we can reasonably anticipate that the forces active in folding the polypeptide chain are also those responsible for protein–protein associations. Even a cursory look at protein–protein associations suffices to illustrate that the arrangements of secondary structure elements, well known in protein monomers, typically recur at the interfaces as well. Yet in the one case where a detailed comprehensive comparison of one motif, the four-helix bundle has been conducted, Lin et al. (1995) have observed that whereas resemblance unquestionably exists, significant differences between the bundle configurations in protein–

Reprint requests to: R. Nussinov, NCI-FCRF, Bldg. 469, Room 151, Frederick, Maryland 21702; e-mail: ruthn@ncifcrf.gov.

protein interfaces and in monomers (Harris et al., 1994) occur as well. Furthermore, as the hydrophobic effect is the dominant force in protein folding, it may be expected that it would be equally critical for protein–protein association. However, statistical analysis of the hydrophobic effect at the interfaces (Tsai et al., 1997a) has revealed that whereas the hydrophobic effect plays an important role in protein–protein association, it is not as dominant as its effect in protein folding. On the other hand, the role of hydrophilic bridges is more important in protein–protein interfaces compared to protein cores (Xu et al., 1997).

These observations, and their underlying rationale, raise the question of how similar are the architectures in protein cores compared to those observed at protein–protein interfaces. On the practical side, such an investigation may hold clues to several questions: (1) is an interface manifesting a structural motif, which recurs in the monomers, an inherently stable interface? (2) Can recurring motifs be used as structural templates in protein–protein recognition? And, (3) in monomers, recurring structural motifs have frequently been referred to as either building blocks or biological functional units in proteins. For example, the strand-helix-strand motif repeats itself to form a tim barrel fold, and the helix-turn-helix motif has been recognized as a calcium binding site. Do the motifs that recur at the interfaces correlate with a particular biological function?

To address the type and extent of architectural similarity between protein cores and interfaces we conduct a comprehensive structural comparison between a recently compiled dataset of protein–protein interfaces (Tsai et al., 1996a) and a dataset of single-chain proteins (Fischer et al., 1995). Owing to the nature of protein–protein interfaces, which are composed of two chains, with each contributing unordered fragments as well as isolated residues, an amino acid sequence order-dependent method is likely to fail in searching for structural similarities between these and protein cores. Utilizing our computer-vision–based structural comparison technique, which views protein structures as collections of points ($C\alpha$s) in 3D space (Nussinov & Wolfson, 1991; Bachar et. al., 1993; Fischer et al., 1994; Tsai et al., 1996b), such a comparison is, however, feasible. Our results reveal that, as expected, overall the structural motifs recurring at the interfaces are similar to those well known to exist in protein cores. Nevertheless, differences are consistently observed as well. It is these differences, within the framework of the general similarities, which are particularly illuminating.

Protein–protein associations can be divided into two classes. The first of these consists of complexes belonging to the so-called "two-state" model. The second class are associations demonstrating a "three-state" model behavior. The two-state model includes those chains that exist either unfolded or folded together in a complex. These chains fold cooperatively, in much the same way as a single-chain protein would typically fold. The cooperative folding-association displayed by these two-state model complexes resembles protein folding. On the other hand, in the three-state model, each of the chains folds separately. The next stage consists of the binding of the already folded monomers. Hence, three states are involved here: the unfolded chains, the native folded monomers, and the bound configuration. This mode of association is also referred to as rigid body binding. While conformational rearrangements may occur to maximize inter-molecular interactions and to stabilize the complex, these generally involve only minor changes.

Interestingly, inspection of the similarities and of the differences between the architectures at the interfaces and in the chains, indicates that interfaces belonging to the two-state model complexes are very similar to the motifs found in the monomers. However, interfaces belonging to complexes displaying the three-state model are, in general, similar only in outline. The arrangements of the secondary structures is similar. However, the details vary. Thus, while the geometric superpositioning of the $C\alpha$s of the cores and of the two-state model interfaces illustrates a good fit, that is not the case for the three-state model interfaces, constituting the bulk of the interfaces.

We conclude that the architectures of interfaces displaying a two-state model kinetics are inherently much more similar to protein cores than the three-state model interfaces. This is entirely logical and consistent with the lack of backbone freedom in the process of association of a three-state interface. On the other hand, the two-state interface has full backbone freedom, similar to the case of protein folding.

The similarities and the differences between protein–protein interfaces and monomers are not surprising. The specific arrangements of secondary structure elements, forming the limited set of folding patterns, i.e., the motifs, have been selected during evolution owing to their favorable stabilizing effect. Hence, an arrangement conferring stability on the monomer will also exert a similar effect on the associating molecules. Nevertheless, there is a major difference between protein cores and protein–protein interfaces. In the former, the polypeptide chain folds maximizing its hydrophobic interactions. Practically all degrees of freedom are available to the backbone to adopt the arrangement in which the lowest energy is achieved. That is not the case for the associating monomers. The monomers are already folded. Very little freedom is available to the folded chains. Hence, for the case of rigid protein binding, only a local maximization of the interactions is feasible. Only six rotational and translational degrees of freedom are available to the monomers in their conformational search. The same rationale applies to the reason for the lower extent of the hydrophobic effect in the three-state model interfaces compared to cores of monomers. As globular proteins are inherently stable in solution, too large an exposed non-polar surface area is unfavorable. In the already folded chains most hydrophobic residues are buried, a priori limiting the potentially attained hydrophobic effect at the three state model interfaces (Tsai et al., 1997a).

This may be a prime reason for the multiple potentially feasible conformations in which protein molecules can associate. It may explain the difficulty in predicting one prevailing docked configuration. In particular, this suggests that unlike the case of protein folding where there is an energy gap between the native and the non-native conformations, this may not be the case for binding, where many potential binding configurations exist, with the energy difference between them significantly smaller. We shall come back to this point and its implications in the Discussion.

Despite these basic differences between protein folding and binding, and despite the lack of chain connectivity, the same gross architectures are still observed, although differing in detail. This argues that these arrangements of secondary structures are highly favorable (Finkelstein & Ptitsyn, 1987). At the same time, they are likely to be more variable in the interfaces compared to their recurrence in the chains, as the associated chains are constrained in attaining their most favorable complex conformations by their already folded structures. In this sense, the repertoire of motifs at the interfaces is likely to be less informative than that observed in protein monomers.

Below we describe the comparisons we have conducted between the datasets of the interfaces and of the monomers. The obtained

similarities and differences are displayed and discussed. To constitute a complete analysis of the interface architectures, we have further explored the prevailing secondary structures at the interfaces compared to the chains. This analysis has been carried out both with respect to residues, and to the polar and non-polar, side-chain and backbone accessible surface areas. These analyses have been carried out for the α-helices, β-strands, and coils. The observed distribution is quite similar to the expected one.

Interface dataset used in the analysis

Two independently compiled interface datasets are used in this study. The first is a dataset of 187 stable interfaces, picked from 376 representative, non-redundant interfaces. The second is a representative dataset of 57 symmetry-related oligomeric interfaces generated from PDB files (Bernstein et al., 1977), containing coordinates of only a single chain. The detailed procedure describing the generation of the so-called symmetry-related proteins around a particular monomer is given in Appendix A. Following an all-against-all comparison between these two datasets, 19 symmetry-related oligomeric interfaces are found to be similar to one of the 187 stable interfaces, and hence, are not included in the final dataset used in the analysis of the secondary structures at the interfaces. A total of 225 protein–protein interfaces (187 stable interfaces + 38 oligomeric interfaces) are utilized in this study. These are listed in Appendix B. The details of the calculations differentiating between stable and unstable interfaces have already been described (Tsai et al., 1997a). Briefly, we consider the extent of the absolute buried surface area, the fraction of the buried surface area in the interface with respect to the surface area available when the molecule is in the uncomplexed state, and the ratio of the buried surface area in the interface with respect to other crystal symmetry-generated interfaces. An insight into stable versus crystal interfaces has been provided by Janin and Rodier (1995), who have particularly addressed this important issue.

A detailed description of the generation of the dataset, using the Geometric Hashing technique (Nussinov & Wolfson, 1991; Fischer et al., 1994, 1995; Tsai et al., 1996a), has already been given. The parameters employed in the generation of the dataset of interfaces and in its comparison with the dataset of monomers have been given in Tsai et al. (1996a). The non-redundant monomer dataset utilized here has also been generated from the PDB by the *Geometric Hashing*, iterating through the same steps followed in the generation of the dataset of interfaces.

Secondary structure composition at the interface

To answer the question of whether there is a preference of a particular type of secondary structure that is involved in protein–protein associations, two statistical analyses of the secondary structure composition are performed. The first is based on a residue as a counting unit; the second utilizes the solvent accessible surface area.

Secondary structure assignment

A residue can be assigned to one of three secondary structure types: α-helix, β-strand, and random coil. We use the secondary structure assignment specified in the PDB file. If not given, the secondary structure assignment is determined by a procedure utilizing hydrogen bond patterns to determine the assignment, in a way similar to that performed by the DSSP (Kabsch & Sander, 1983) algorithm. Although several types of helices have been defined in the PDB, in this study only normal and 3_{10} helices are considered.

Accessible surface area (ASA)

The solvent accessible surface area (ASA) of a protein is calculated following the Lee and Richards' definition (1971), with a probe ball radius of 1.4 Å. The algorithm of Shrake and Rupley (1973) is adopted in our implementation, similar to that described by Miller et al. (1987). The atomic radii used in the calculation of the accessible surface area have been taken from CHARMM (Brooks et al., 1983). The accessible surface area of an atom is represented by discrete surface points of a sphere, each with an associated surface area. In our implementation, starting with an icosahedron geometry, a series of 12, 42, 162, and 642 spherical points are calculated in a quasi-uniform distribution. We have adopted the 162-point icosahedron to represent the surface of a sphere, because the difference between the surface area calculated by using 162 points and that calculated by using 642 surface points is always within 0.1% for most of the proteins we have examined.

Definition of interacting interface residues

In this study, we use the ASA to determine whether a residue is an interacting interface residue. First, let us define some terms used in the calculation. An interface contains two chains, chain A and chain B. A "reference" ASA, ASA_{ref}, refers to the available ASA of a residue. That ASA is the surface area that is excluded, i.e., not buried, either by atoms from the same residue, or by the linking backbone atoms (C– and –N). ASA_A of a residue refers to its reference ASA, buried only by its own chain A. ASA_B refers to its reference ASA, buried only by chain B. $ASA_{A,B}$ refers to the ASA that is buried simultaneously by both chains, A and B. ASA_{expos} refers to the ASA that is not buried, either by chain A or by B. Note that the reference ASA of a residue equals the sum of ASA_{A}, ASA_B, $ASA_{A,B}$, and ASA_{expos}. The fractions of each of the above four ASAs from the reference ASA_{ref}, are denoted as P_{A}, P_B, $P_{A,B}$, and P_{expos}, respectively. A residue of chain A is defined as an interface residue if $P_B > 0.05$ or if $P_{A,B} > 0.10$, except for the case where $(P_A + P_{A,B}) > 0.95$ and $P_{A,B} < 0.15$. A residue of chain A is defined as a surface residue if $(P_B + P_{\text{expos}}) > 0.25$.

Two hundred twenty-five interfaces, containing 450 chains (a total of 116,565 residues) are used in the calculation of the secondary structure composition. A percentage of 40.2 are defined as surface residues and 15.7% are interface residues. The secondary structure composition in terms of residues either from the whole chain, from the chain surface, or of the interacting interface residues, is summarized in Table 1.

Secondary structure composition based on ASAs

By using ASA, the secondary structure composition at the interface can be calculated directly. The expected secondary structure composition of a chain A in terms of ASA is defined as the sum of ASA_B and ASA_{expos}, which is equal to its available ASA when isolated. The observed secondary structure composition is based on the area buried by chain B alone, ASA_B. In addition to the overall ASA analysis, the ASA is further separated into the contribution from the backbone, non-polar, and polar atoms. A per-

Table 1. *The secondary structure composition of 225 two-chain interfaces*[a]

	α-Helix	β-Strand	Random coil
Whole chain	0.365	0.228	0.407
Chain surface	0.343	0.131	0.525
Interface	0.358	0.170	0.472

[a]The secondary structure composition of chains and interfaces. Two hundred twenty-five two-chain interfaces have been used in the calculation of the secondary structure composition for all residues of the entire chains (whole chain), for residues on the chain surface (chain surface), and for the interacting interface residues (interface). The ASA of each residue (solvent accessible surface area) was utilized to define whether it is a surface residue or an interface residue. See text for detailed definition.

centage of 15.2 of the overall available ASA is buried in the 225 interfaces. Separated into the contributions from backbone, non-polar, and polar atoms, the buried percentages are 13.5, 18.4, and 12.2, respectively. The results of the secondary structure composition based on ASAs are given in Table 2.

These results illustrate that the distribution of the interacting residues into the secondary structure types is roughly as expected. Inspection of the distribution of the residue-based secondary structure content indicates that in the entire chains, on their surfaces and in their interfaces, the content of the α-helices is constant, around 35%. On the other hand, more residues belonging to β-strands are found in the chains as a whole than on their surfaces. The interfaces contain more β-strand residues than the protein surfaces, although less than in the entire chains. The situation is reversed for the loops. The fact that the helical content of proteins is higher than of strands has been recognized for a considerable time. Regarding the strands, their lower content on the surfaces compared to the interior is a reflection of the geometry of the β-sheet, which contributes less to the surfaces, compared to the helices. On the other hand, the higher content of the loops on the surfaces is also well known. The fact that β-sheet residues are found at a higher concentration in the interfaces than on the surfaces, compared to the reversed phenomenon of the random coils, is straightforwardly understood, because loops typically contain a higher proportion of hydrophilic residues. Inspection of the secondary structures as calculated by accessible surface area reflects these trends directly, which are entirely within the expected ranges. Furthermore, the observed division of the accessible surface area into backbone, polar, and non-polar moieties are all as expected. We conclude that the secondary structure distribution of the residues that interact across the protein–protein boundary reflects the division observed in protein monomers.

Table 2. *The expected and observed secondary structure composition of 225 interfaces*[a]

	α-Helix	β-Strand	Random coil
Reference ASA	0.377	0.229	0.397
Exp. (all ASA)	0.343	0.140	0.517
Obs. (all ASA)	0.356	0.146	0.498
Exp. (backbone)	0.233	0.114	0.653
Obs. (backbone)	0.215	0.133	0.653
Exp. (non-polar)	0.366	0.140	0.494
Obs. (non-polar)	0.380	0.150	0.470
Exp. (polar)	0.373	0.154	0.474
Obs. (polar)	0.393	0.147	0.460

[a]The secondary structure composition of 225 two-chain interfaces calculated directly based on individual ASA of every atom in the protein. The expected secondary structure composition of a chain is based on its available ASA when isolated, while the observed one is based on the area buried by its partner. In addition to the overall statistics (all ASA) of the interface ASAs, the secondary structure composition has been further divided into the contributions from the backbone, non-polar, and polar atoms, respectively. See text for detailed description.

Interface structural motifs

To detect automatically structural motifs that recur both at the interfaces and in protein cores we have performed extensive, all-against-all comparisons between the single-chain dataset with 361 proteins and the protein–protein interface dataset, containing 376 two-chain interfaces and 38 oligomeric interfaces. In these comparisons both the interacting residues and the ones in their vicinity (i.e., those whose $C\alpha$s are within 6 Å of a $C\alpha$ of an interacting residue, see Tsai et al., 1996a) are included. The Geometric Hashing algorithm (for a description, see Nussinov & Wolfson, 1991; Bachar et al., 1993; Fischer et al., 1994) has been utilized as the tool for these structural comparisons. While, as discussed by Tsai et al. (1996a), being amino acid sequence order independent, the Geometric Hashing is uniquely suitable to carrying out such a task, still owing to the nature of a comparison between an interface—composed of two chains—and a structure of a single chain, from the monomer dataset, there are some potential, practical difficulties that might arise. To address these, in the comparisons only matches fulfilling the following three criteria have been considered as candidates for constituting a motif. First, the relative connectivity score of the interface should be higher than 0.5. The connectivity score is a measure of the quality of the superposition between two structures. It takes into account the matched condition of the two residues bordering a matched residue pair. The score has been designed in a way such that when an interface is compared with itself, its connectivity score is equal to its residue size. The relative connectivity score is the connectivity score divided by the size of the interface (see Tsai et al., 1996a, for further details.) Hence, this criterion ensures that at least 50% of the interface is matched with the protein monomer. Second, at least 25% of the match arises from each of the chains of the interface. This criterion is designed to exclude cases where only one chain in the interface is dominantly involved in the match, with no significant participation from the second chain. Third, the connectivity score should be higher than an absolute, pre-defined value, 25 residues. This criterion has been designed to filter out matches having no significant secondary structure content. The results are summarized in Table 3. Fifty-three interfaces are described by the spatial arrangements of their secondary structure elements. Note that in the list of Table 3, none of the cases is from the 38 oligomeric interfaces. This suggests that the "symmetric" interactions observed in oligomeric proteins are not found in single-chain monomer folding.

The results obtained in the matching of the monomers and the interfaces are divided to three categories. The first group includes perfect matches, with a large number of matched residues, a high connectivity score (>85% and >75%, respectively), and a low RMSD (<1.5 Å). The second group includes good matches, with a significant number of matched residues and connectivity score

Table 3. *A list of interfaces and of their secondary structure motifs that partially match motifs found in monomers*[a]

	Interface	Description in terms of secondary structural elements
1	1aarAB	six-stranded barrel
2	2afnBC	quite complicated
3	1babAC	two parallel helices
4	1bbbAB	six helices
5	1bbbAC	two long helices with four short helices
6	1bbhAB	four-helix bundle like
7	1bovAE	four-stranded sheet with two helices
8	1bbrHE	like a chain with cleavage
9	1bgsFG	four-stranded sheet with two helices plus two short helices
10	1cdtAB	four-stranded sheet with two loops
11	1colAB	10 helix heads
12	1cosAB	two intertwined helices
13	1ctdAB	four helices
14	1d66AB	two short helices
15	1dfnAB	six-strands barrel-like motif
16	1fc1AB	eight-stranded compressed barrel-like with two short helices
17	1fc2CD	two helices with two short helices
18	1fvdAC	eight-stranded barrel-like motif
19	1ggaRA	two helices with two short helices
20	1gp1AB	two helices and loops
21	1hgeAC	four-stranded sheet with loop
22	1rprAB	four-helix bundle
23	1hrhAB	four-stranded sheet with two helices
24	1hviAB	interlaced beta sheet and loops
25	1ithAB	two helices with four short helices
26	1ltaDC	helix with helix + strand + helix
27	1ltsFC	single strand with single helix
28	1mlpAB	two very long intertwined helices
29	1molAB	an open eight-stranded (3/5) sandwich
30	1rhgAC	an open four-helix bundle
31	1rhgBC	four-helix bundle like
32	1rtp23	five short helices and loops
33	1plfAB	four-stranded sheet with two helices
34	1sltAB	parallel eight-stranded (4/4) sandwich
35	1sosFE	open beta sandwich
36	1trzBD	two helices with two strands
37	1vfaAB	eight-stranded barrel-like motif
38	2ccyAB	four-helix bundle like
39	2dh1AB	two intertwined helices
40	2hf1HY	an open seven-stranded barrel-like sheet enclosed by two loops
41	2mltAB	two bent helices
42	2msbAB	four-stranded sheet with four short helices
43	2pcbAB	several terminal helices
44	2pccAC	two long helices with two short helices
45	2rs1AB	two intertwined helices with loops
46	3hhrBC	six-stranded barrel-like motif
47	3inkCD	three-helix bundle
48	3insAB	three helices
49	3monCD	five-stranded sheet with one helix
50	3sc2AB	like a chain with cleavage
51	3sdhAB	four-helix bundle like
52	4azuAB	loops
53	4rubAB	six short helices

[a]A list of the interfaces and a description of their secondary structure motifs that partially match motifs from the monomer dataset.

(>60%). The third group consists of interfaces with a fair match with the monomers (>50% of the relative connectivity score). The first group (listed in Table 4) corresponds to either an interface from a folded single-chain protein being cleaved into two chains or a single-chain protein actually composed of two chains (possibly an error in the PDB). Figure 1 depicts an example taken from this table. As a result of the above concern, there are four interfaces—1cauAB, 1hleAB, 1srnAB, and 2ltnCD—which have not been included in Table 3 of the 53 interfaces. One interface, 1vfaAB, which is listed in Table 4, has also been included in the third group, as described below. Most of the interfaces of the second group (listed in Table 5) correspond to either interfaces described by the two-state model or to a very simple motif (e.g., two intertwined helices). These latter motifs recur frequently in protein monomers. Figure 2 illustrates several examples of interface structural motifs matching well protein monomers. These structural motifs, taken from Table 5, include an open eight-stranded (3/5) sandwich, a four-stranded sheet with two helices, two intertwined helices, an eight-stranded barrel-like motif, two helices plus two strands, a six-stranded barrel-like motif, and a four-helix bundle. An interface is assigned to belong to a two-state model, if it contains a compact hydrophobic folding unit, with both monomers contributing to it equally (Tsai & Russinov, 1997c). Performing a thermal unfolding experiment, Steif et al. (1993) have reported a strict two-state behavior for the ROP dimeric protein from *Escherichia coli* (1rpr), one of the interfaces in the second group. Interfaces included in the third group (listed in Table 6) have similar spatial arrangements of the secondary structural elements as those of the monomers, however, differing in detail. Hence, only a fair superposition with single-chain proteins has been obtained for this group. Two examples from the third group are given in Figure 3.

Functional motifs

Inspection of Tables 4 and 5 reveals that two matches having a large number of matched pairs, 2afnBC (the interface between

Table 4. *A list of perfect matches (group I) between the dataset of interfaces and that of the monomers*[a]

Interface	Chain	RMSD	Number of matched pairs	Percentage of match
1bbrHE	4ptp	1.10	174	86
1cosAB	1bgc	1.10	55	98
1cauAB	1phs	1.04	134	87
1hleAB	1attA	1.05	179	97
1srnAB	7rsa	0.85	74	100
1vfaAB	1mfa	1.44	87	98
2ltnCD	1scs	1.07	201	95
3insAB	6insE	1.27	44	86
3sc2AB	1ysc	1.37	271	90

[a]A list of perfect matches between the dataset of interfaces and that of the monomers. The percent of the match is the number of matched pairs divided by the size of the interfaces. Note that the matches are a function of the thresholds used in the structural comparisons. The thresholds used in the comparisons have been described in Tsai et al. (1996a). The rationale adopted has been outlined both in the text, and, is further detailed in the reference cited above. An example taken from this table is depicted in Figure 1. Note that interface 1vfaAB is listed both here and in Table 6, matching different chains.

Fig. 1. A stereo view of an example taken from group I (high-quality matchings, as judged by a large number of matched pairs, high connectivity score, and low RMSD). Here the figure depicts the superposition of interface 2ltnCD (pea lectin) with the perfectly matched monomer 1scs (concanavalin A). The two chains of the interface are shown in red and green, respectively, with the darker color highlighting the interface region. The monomer chain is colored blue.

chains B and C of *Alcaligenes faecalis* nitrite reductase) with 1aozA (ascorbate oxidase) and 1hviAB (the interface of HIV-1 protease) with 1mpp (pepsin), are particularly interesting. These two unusual matches between a protein–protein interface and a

Table 5. *A list of good matches (group II) between the dataset of interfaces and that of the monomers*[a]

Interface	Chain	RMSD	Number of matched pairs	Percentage of match
2afnBC	1aozA	1.96	169	84
1babAC	256bA	1.70	41	95
1bbbAC	2ctc	2.16	56	85
1bovAE	1tplA	1.81	58	85
1ctdAB	4icb	1.70	58	91
1d66AB	1sesA	1.25	34	85
1dfnAB	2tbs	1.80	42	93
1fvdAC	2stv	1.89	53	88
1rprAB	1bgc	1.83	101	89
1hrhAB	1attA	1.97	42	91
1ltaDC	1brd	1.77	42	93
1ltsFC	1cpcB	1.70	44	94
1molAB	1aizA	1.79	43	84
1rhgAC	1huw	1.54	85	82
1rhgBC	1gmfA	1.80	62	81
1sltAB	1stvA	1.85	60	95
1trzBD	1ede	2.04	49	94
2ccyAB	1hmcA	2.13	67	88
2dhlAB	1pgd	1.54	51	98
2mltAB	1aco	1.97	48	92
2pccAC	1bmdA	1.74	43	90
3hhrBC	1ofv	2.21	45	87
3monCD	1cew	1.83	71	80

[a]A list of good matches between the dataset of interfaces and that of the monomers. The percent of the match is the number of matched pairs divided by the size of the interfaces. For further details see the legend to Table 4. Examples taken from this table are depicted in Figure 2.

Table 6. *A list of fair matches (group III) between the dataset of interfaces and that of the monomers*[a]

Interface	Chain	RMSD	Number of matched pairs	Percentage of match
1aarAB	1byb	2.06	61	88
1bbbAB	2hpdA	1.91	75	71
1bbhAB	1aep	1.97	86	85
1bgsFG	1pec	1.90	61	86
1cdtAB	2tmdA	1.75	44	79
1colAB	2gstA	1.92	55	66
1fc1AB	1hplA	1.88	89	73
1fc2CD	1bnh	2.07	51	88
1ggaRA	2btfA	2.08	51	91
1gp1AB	1tpfA	2.12	58	84
1hgeAC	1scs	1.75	52	81
1hviAB	1mpp	2.04	95	77
1ithAB	2sas	2.13	55	82
1mlpAB	1vsgA	1.62	79	68
1plfAB	1alkA	1.75	66	83
1rtp23	1ysc	2.25	56	86
1sosFE	1amp	1.89	53	88
1vfaAB	2bat	2.05	65	73
2hflHY	1pgd	2.12	56	93
2msbAB	1php	2.09	57	86
2pcbAB	1wsyB	2.04	48	81
2rslAB	2lbp	2.00	58	85
3inkCD	1amp	1.87	49	89
3sdhAB	1colA	2.02	72	69
4azuAB	1pha	1.89	45	78
4rubAB	1aep	2.11	51	80

[a]A list of fair matches between the dataset of interfaces and that of the monomers. The percent of the match is the number of matched pairs divided by the size of the interfaces. For further details see the legend to Table 4. Examples taken from this table are depicted in Figure 3.

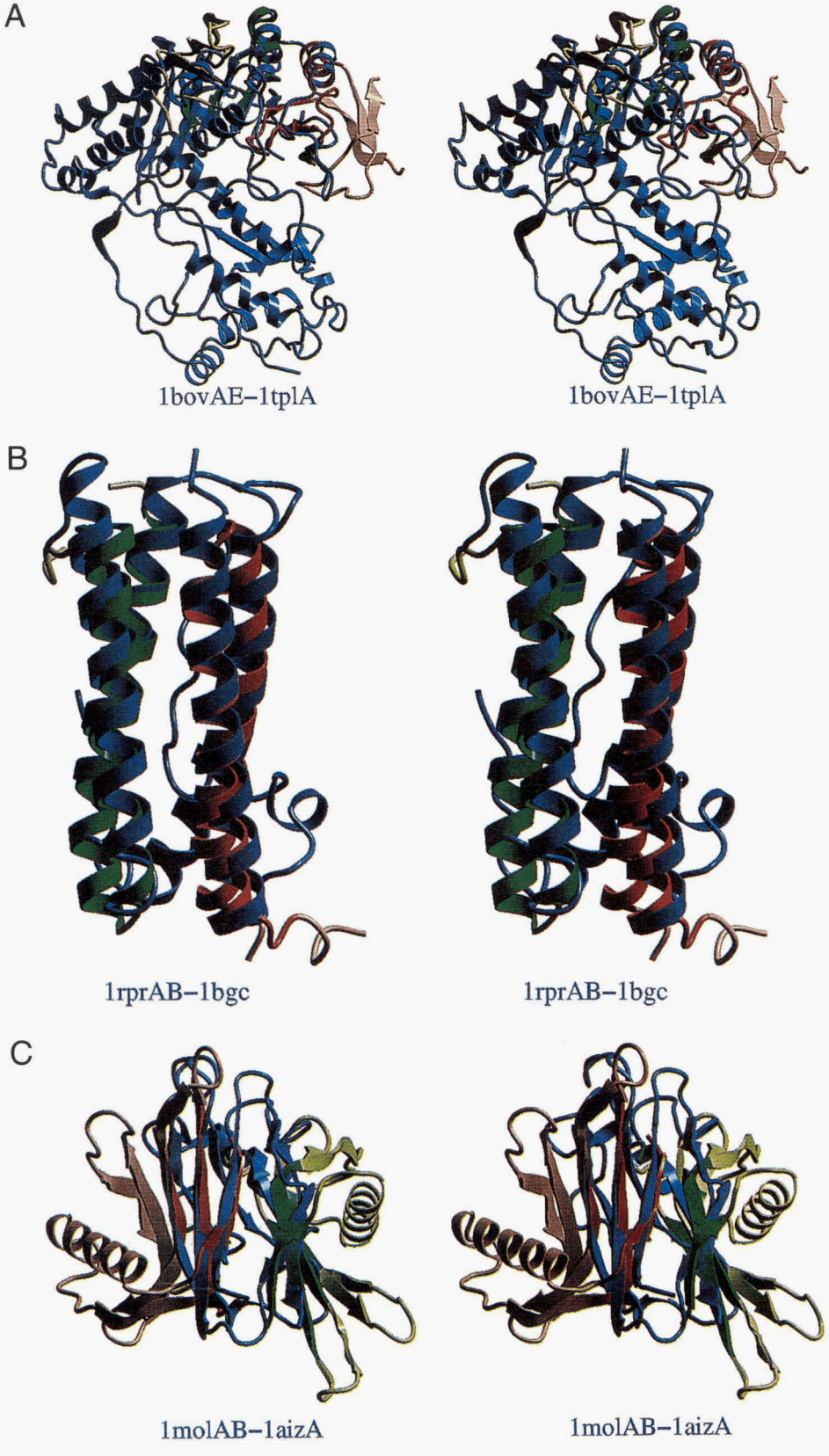

Fig. 2. Stereo views of the superposition of selected interfaces from group II, with well matched monomers. These are (**A**) interface 1bovAE (verotoxin-1) with monomer 1tplA (tyrosine phenol-lyase), (**B**) interface 1molAB (monellin) with monomer 1aizA (azurin), and (**C**) interface 1rprAB (ROP) with monomer 1bgc (granulocyte colony-stimulating factor). The two chains of the interface are depicted in red and green colors, respectively, with the darker color highlighting the interface region. The monomer chain is colored blue.

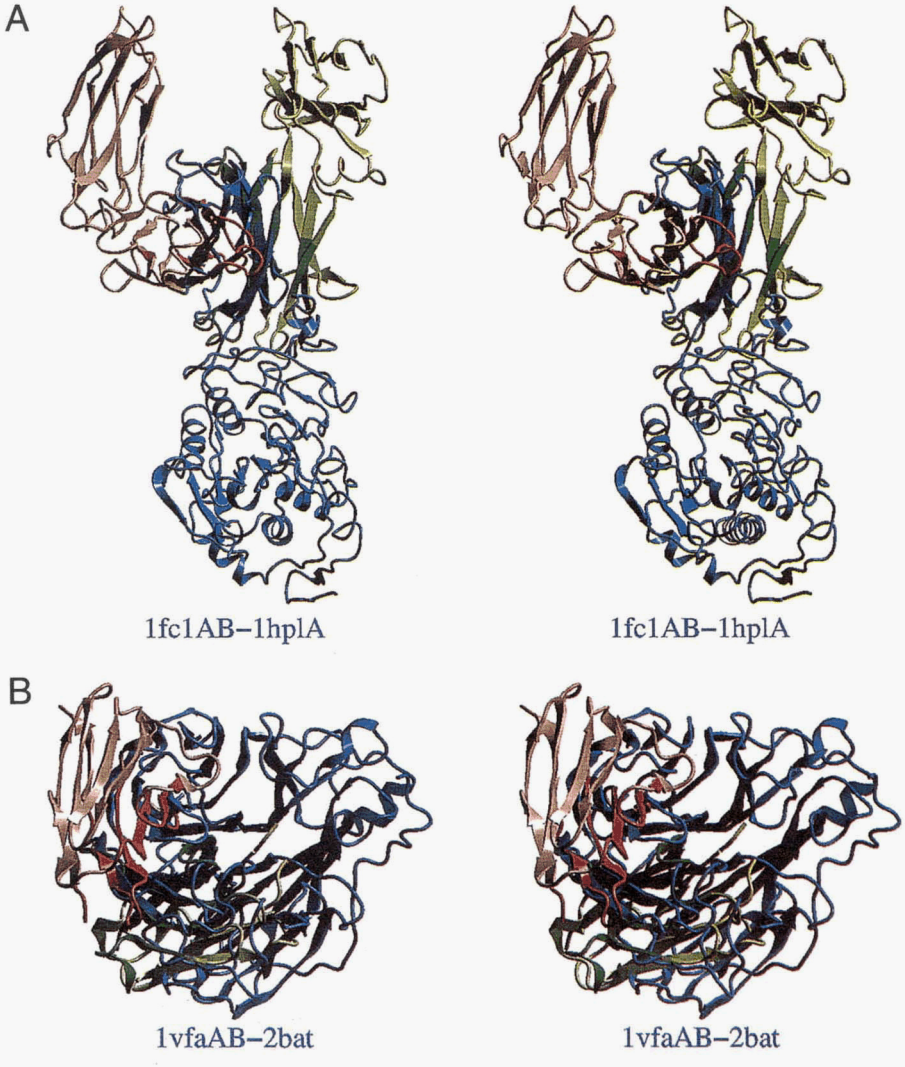

Fig. 3. Stereo views of the superposition of two selected interfaces from group III, with fair matching monomers. These are (**A**) interface 1fclAB (FC fragment) with monomer 1hplA (lipase), and (**B**) interface 1vfaAB (FV fragment) with monomer 2bat (neuraminidase N2). The two chains of the interface are shown in red and green, respectively, with the darker color highlighting the interface region. The monomer chain is shown in blue.

protein monomer are a direct outcome of the imposed functional requirement. Interfacial "functional motifs" differ from structural motifs, which are borne out strictly owing to hydrophobicity considerations. In this type of motifs, we observe a novel manifestation of evolution. Results obtained from detailed analyses of such matchings are expected to bear upon the mechanism of these proteins, their evolution, and the inter-relationship between structure and function.

Figure 4 depicts the superposition of the 2afnBC:1aozA match. The nitrite reductase (2afn) is a functional trimer (Godden et al., 1991) both in the crystal and in solution, with each monomer containing two similar β-barrel domains related to plastocyanin and azurin. On the other hand, the ascorbate oxidase (1aoz) is a dimer in solution and a tetramer in the crystal (Messerschmidt et al., 1992). In each monomer there are three clear-cut β-barrel domains resembling the domains found in nitrite reductase. The functional and structural similarities as well as the sequence alignment between these two oxidoreductases have been noted in the literature (Fenderson et al., 1991; Godden et al., 1991). The active sites in both oxidoreductases contain two copper sites: the first is a type I copper in both oxidoreductases, whereas the second is a type II copper in nitrite reductase, and a trinuclear copper center in ascorbate oxidase. In both oxidoreductases the type I copper site acts as an electron acceptor and the other site is the reducing center. In nitrite reductase, the type I copper accepts an electron from a type I copper of an attached pseudoazuin (a substrate). It then donates it via an intramolecular electron transfer pathway to the type II copper, where nitrite is reduced to nitric oxide. On the other hand, the ascorbate oxidase, through a similar electron transfer pathway, reduces an oxygen to water with a concomitant one-electron oxidation of the substrate. Figure 4 illustrates the match between the intersection of the three domains of ascorbate oxidase with the two-chain interface of nitrite reductase. This type of matching provides functional evidence that the mechanism of electron transfer requires a unique spatial orientation of the two copper sites with respect to each other.

The superposition of the 1hviAB:1mpp match is shown in Figure 5. The good match is not surprising given that both the HIV-1

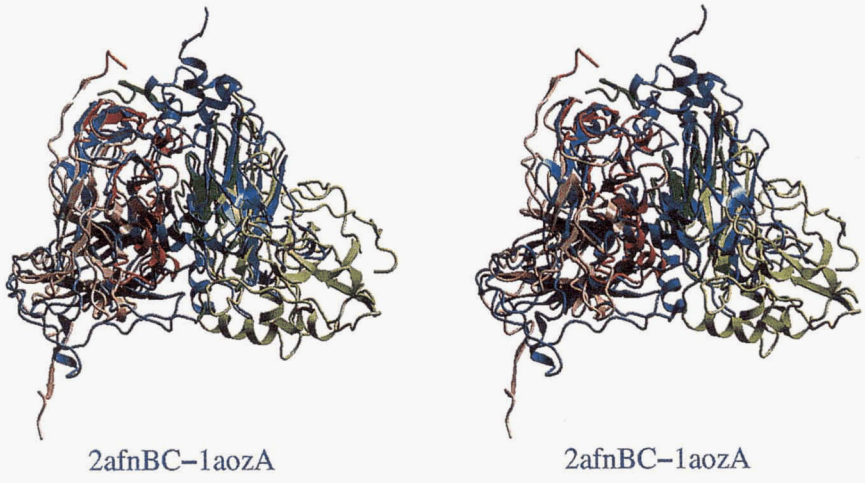

Fig. 4. A stereo view of the superposition of the interface (2afnBC) and the monomer (1aozA). The two chains of the interface are shown in red and green colors, respectively, with the darker color highlighting the interface region. The monomer chain is shown in blue. The nitrite reductase (2afn) is a functional trimer, with each monomer containing two similar domains. On the other hand, the ascorbate oxidase (1aoz) is a functional monomer containing three domains. In both oxidoreductases the active sites contain two copper sites. The occurrence of this type of match between the interface of three domains in the ascorbate oxidase and the interface of the two chains of nitrite reductase indicates that the mechanism of electron transfer requires a unique spatial orientation between the two copper sites. One site acts as an electron acceptor and the other at the reaction center is the reducing agent. See text for further details.

protease (1hvi, a functional dimer) and the pepsin (1mpp, a monomer with two domains), belong to the aspartic protease family. Hence, the similarity is a straightforward outcome of their biological function, reflecting the necessity of a unique spatial arrangement of the catalytic triad as well as the associated enzymic environment.

Discussion

It has long been recognized that protein structures consist of specific geometric arrangements of their secondary structure elements. Some spatial combinations of the α-helices and of the

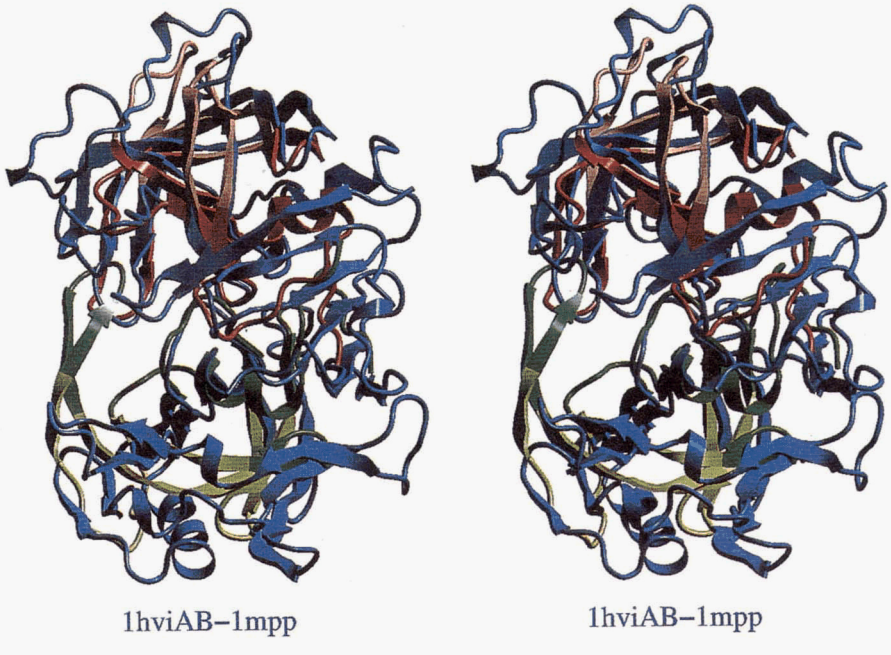

Fig. 5. A stereo view of the superposition of the interface (1hivAB) and the monomer (1mpp). The two chains of the interface are shown in red and green colors, respectively, with the darker color highlighting the interface region. The monomer chain is shown in blue. Both the HIV-1 protease (1hvi, a functional dimer) and the pepsin (1mpp, a monomer with two domains), belong to the aspartic protease family. Hence, the similarity is an outcome of the requirement of a unique spatial arrangement of the catalytic triad and the associated enzymic environment.

β-strands consistently recur, whereas others are not observed. Some of the motifs have specific associated biological functions. Others form parts of larger structural assemblies. The interior of these motifs is typically hydrophobic. The loop regions connecting them are typically hydrophilic, exposed to the solvent. Without exception, the cores of native globular proteins consist of predominantly hydrophobic side chains. Only a limited number of spatial arrangements of the secondary structures enable obtaining the most favorable optimal interactions. Because the forces acting at protein–protein interfaces are those responsible for protein folding, it is natural that the same types of architectures observed in the monomers would also be manifested in the interfaces (recently reviewed by Jones & Thornton, 1996), despite the absence of chain connectivity in the latter. Here we address the question of the extent of this similarity. Exploring the architectures of the interfaces compared to those at protein cores illuminates some of the basic similarities and differences between folding and binding.

To conduct such an investigation, three items are critically needed: a dataset of non-redundant monomer structures; a dataset of protein–protein interfaces, and a technique for their comparison. Here we have utilized a dataset of stable interfaces (Tsai et al., 1997a), picked from a larger dataset of non-redundant protein–protein interfaces, derived from the PDB (Tsai et al., 1996a). We have further generated automatically all symmetry-related oligomers, extensively compared these with the existing interface dataset, and included the unrelated ones in the set used in the analysis. Because interfaces are composed of unordered fragments of two polypeptide chains as well as isolated residues, this investigation necessitates a residue order-independent structural comparison tool. The availability of our computer vision-based structural comparison technique has enabled carrying out the comparisons both between the interfaces and between the interfaces and the monomers, examining the extent of recurring folding pattern arrangements.

Levitt and Chothia (1976) and Richardson (1977) have postulated already two decades ago that protein globules adopt folding patterns displaying recurring topologies. In a seminal review, Finkelstein and Ptitsyn (1987) have addressed the question of why do globular proteins fit the limited set of folding patterns. Proteins may differ substantially biochemically or philogenetically and yet manifest similar or identical folding patterns, indicating that the reason for this limited set is likely to be a physical limitation rather than evolutionary divergence or convergence. Their underlying simplifying assumption has been that the most favorable set of folding patterns is determined by the thermodynamic stability rather than by the protein folding pathways. Finkelstein and Ptitsyn further assumed that this stability can be evaluated without taking into account all the details of the atomic structures.

As expected, despite the lack of chain connectivity between the two monomers the interfaces display the same architectures as the cores, reinforcing the recognition that the number of potentially favorable ways that secondary structure elements can be arranged while still maintaining thermodynamic stability is limited. Furthermore, this overall architectural similarity is obtained despite the fact that the actual details display considerable variability. However, this variability is not uniform. When compared to protein cores, the complexes fall into two major classes, inherently different from each other. Interfaces belonging to the first type (derived from two-state complexes) manifest a high similarity to protein monomers. Hence, good matches are obtained for this class. On the other hand, no good geometric superposition is obtained for interfaces belonging to the second category (derived from three-state complexes). These interfaces resemble the monomers only in general architectural outline, exhibiting an appreciably larger extent of variability.

A single-chain protein can possess more than one domain or hydrophobic folding unit (Tsai & Nussinov, 1997b). The portion of a monomer exhibiting similarity to a two-chain interface may involve either a single hydrophobic folding unit, or an interface between domains in a single-chain protein. Almost all of the 53 examples detected and tabulated in this study belong to the former case. This type of similarity has been referred to as a "structural" motif, because hydrophobicity is its major determinant. In such a motif, the respective biological functions of the monomer and of its corresponding structurally similar two-chain complex interface, are most likely unrelated. On the contrary, in the second case, the similarity between a monomer and a two-chain complex interface has been imposed by the requirement of their similar biological functions. This type of similarity has been termed a "functional" interfacial motif. Two particularly interesting cases illustrating a similarity between a two-chain interface and an interface between domains within a monomer have been automatically discovered in this study. Clearly, such a similarity can provide a clue to the functional mechanism of the protein. Furthermore, the two cases observed here provide a unique insight into the evolutionary advantage exhibited by an oligomeric protein. That is, for a particular biological function, encoding a single domain protein within a gene is sufficient to enable the protein to conduct its essential prescribed function. There is not necessarily a need to encode a protein with two or more domains.

Focusing on the class of oligomeric proteins whose interfaces demonstrate a high architectural similarity to the interior of the proteins shows them to be relatively small if they encompass only one hydrophobic core. These proteins are unstable as monomers. Depending on the conditions, in solution the monomers are either unfolded or folded in a complex. The chains fold cooperatively and, hence, the structures at the two-chain interface resemble conformations typically recurring at the interior of proteins. Comparisons of these interfaces with monomers yield a relatively accurate superposition, reflecting a good fit between the respective C_α atoms. The conformations of these complexes reflect the lowest free energy. On the other hand, inspection of the second class illustrates a larger deviation in the positions of the respective C_αs of the interfaces with respect to similar folding patterns of the monomers. Interfaces belonging to this class are derived from complexes whose monomers fold separately with subsequent association as relatively rigid bodies. These so-called three-state model interfaces arise from already folded monomers, each at its free energy minimum. Although some conformational re-arrangement is likely to take place, maximizing side-chain interactions, in essence only six degrees of freedom, are available to the associating monomers. That is unlike the case of the two-state model interfaces, resembling protein folding, with the backbone possessing all degrees of freedom to attain its most stable configuration. As in rigid-body binding, the monomers are already folded in solution the absence of very large patches of hydrophobic surfaces exposed to the solvent certainly makes sense. Hence, the extent of the hydrophobicity at the three-state model interfaces is not as large as either in the two-state or in protein cores.

This difference between protein cores and two-state model binding compared to the three-state model suggests that the gap between the native complex configuration and alternate binding modes for rigid binding is likely to be appreciably smaller than that ex-

pected for the two-state binding or for folding. Figure 6 illustrates this difference between folding and two-state binding on the one hand versus three-state binding on the other. This schematic diagram serves to illustrate two points. First, the gap between a native folded protein and its unfolded states (or, a native folded two-state complex with respect to its two separate unstable monomers) is much greater than the gap between a three-state complex and the two separate stable monomers. This is understandable. It is manifested in the dissociation process of a three-state complex: the complex separates into two chains before each chain initiates its unfolding process. Second, the number of mis-folded states of either the polypeptide chain in protein folding or of the two unstable monomers in two-state binding is substantially larger than the number of ways the two monomers can associate in three-state model binding. This again can be rationalized by the flexibility of the backbone in the folding of the former and its relative rigidity in the latter. With only relatively minor movements of side-chain optimization enabled in rigid-body binding, the requirement of surface complementarity reduces appreciably the number of potential mis-bound states. Based on this latter fact, solving the protein docking problem should be substantially easier than solving the protein folding problem. However, this favorable fact is offset by the first point noted above, namely, that the energy gap is narrower between the native three-state complex and the mis-docked complexes compared to the gap between the native protein and its mis-folded conformations.

The difference between the two-state and the three-state interfaces has some direct implications toward understanding and predicting protein associations. While prediction approaches have made a significant progress over the last years, they are still faced with a major hurdle, namely distinguishing the native from the non-native docked configurations. The number of geometrically feasible docked configurations can still be substantial, especially for the real-life cases, where one docks molecules whose structures have been determined separately, and hence, considerable surface variability can be expected to exist. Furthermore, as reasoned above, for most systems multiple binding conformations are expected, given that the energy gap between alternate bound associations can be quite small.

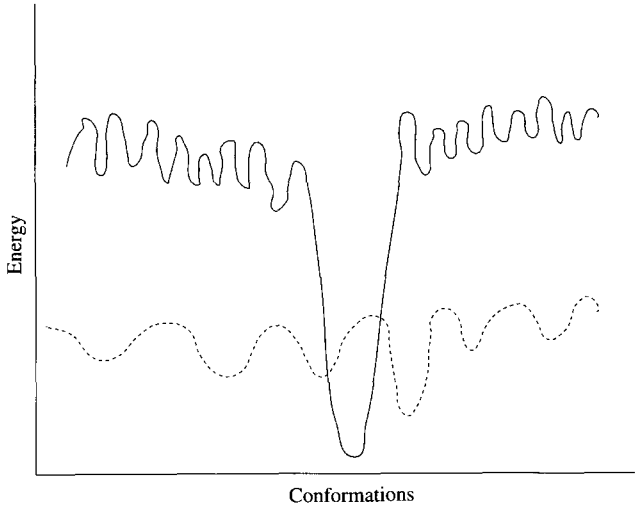

Fig. 6. A schematic diagram illustrating the relative energy gap and the relative number of local minima with respect to the global minimum (native conformation). These are shown for protein folding or a two-state model binding complex (solid line) compared to a three-state model complex (dotted line). The figure illustrates two points. First, the gap between a native protein and its mis-folded states (or a native two-state complex and the two mis-folded monomers) is much greater than the gap between a native three-state complex and the non-native complexes. Second, the mis-folded conformations of a protein or of the two unstable monomers from a two-state complex, outnumbers the non-native complexes from two folded monomers of a three-state complex.

It is therefore not surprising that developing a scheme for scoring the docked binding modes, which would be applicable to all protein–protein complexes, and successfully and uniquely discriminate "correct" from "incorrect" docked configurations, has been proven to be an exceedingly difficult problem. One frequently used approach is modeled after schemes for assessing protein folds, that is to utilize protein–protein interfaces for obtaining statistics on occurrences of residues or atoms that are in contact across the interface. These residue–residue, or atom–atom statistics are subsequently employed in evaluating docked conformations. Although inherently logical, owing to the nature of the rigid binding constraints, deriving and applying a set of statistically based potential functions for docking molecules can be anticipated to encounter difficulties.

In a particularly insightful work, Finkelstein et al. (1995) address the dilemma associated with the utilization of energy minimization as a reliable tool in the prediction of protein structure. Finkelstein et al. argue that even a relatively small uncertainty in the energetic parameters that are employed to assess the stability of predicted conformations can lead to an exponential increase in the number of calculated potential native folds. As Shortle et al. (1996) note, this argument is of particular significance because the energy function utilized is always an approximation of the "true" energy. ΔE, the difference between the energy of the "correct" structure and that obtained for the lowest energy structure represents the error in the calculations. As ΔE increases, the number of alternate, incorrect, conformations increases exponentially. Clearly, errors increase the uncertainty of the energy, and hence, the number of predicted structures grows exponentially with the energy. Furthermore, Finkelstein et al. argue that for a given level of error, the listing of the candidates would be particularly long when the energy gap between the native and alternate conformations is small.

Conclusions

Inspection of the motifs in the monomers with respect to those at protein–protein interfaces illustrates similarities between these two categories. This has been expected. The statistics of secondary structural composition indicates that there is no preference for a particular secondary structure element at protein–protein interfaces. In addition to the frequently observed "structural" motifs containing a hydrophobic core, two "functional" interfacial motifs have been discovered via an all-against-all structural comparison between the dataset of the single chain monomers and that of the interfaces. Unlike the "structural" motifs, the "functional" motifs involve two or more domains (or hydrophobic folding units) of a single-chain protein. This type of similarity between an interface and a chain is imposed by their common biological function.

The number of potential favorable arrangements of interacting secondary structure elements is limited (Finkelstein & Ptitsyn, 1987), and similar arguments can be advanced to protein–protein

associations as those that have been noted for single protein chains (Chothia, 1992; Alexandrov & Go, 1994; Crippen & Maiorov, 1995). The fact that connectivity does not determine the type of fold is consistent with studies of single chain proteins where loops have been cleaved and yet similar motifs have been observed. Hence, while chain connectivity increases the effective concentration of the units, it does not specify the packing architecture.

Nevertheless, the absence of chain connections across the interface on the one hand, and the difficulty in maximizing the side-chain (and backbone) interactions in recognition and binding on the other hand, indicates that some differences should be expected as well. It is those differences, within the framework of the similar architectures (Lin et al., 1995), which are particularly illuminating. Although the interfaces of three-state model complexes, representing rigid-body binding, resemble in outline the folding patterns observed in protein cores, they illustrate a considerable extent of variability. On the one hand, it rationalizes both the multiple potentially feasible alternate binding modes, and the difficulty in distinguishing between the native and non-native configurations, on the other hand it also suggests the feasibility and attractiveness of using structural templates in protein recognition.

Acknowledgments

We thank Drs. Lynn Young, Robert Jernigan, and Jacob Maizel, for helpful discussions, encouragement, and interest. We thank the personnel at the Frederick Cancer Research and Development Center for their assistance. The research of R. Nussinov has been sponsored by the National Cancer Institute, DHHS, under Contract No. 1-CO-74102 with SAIC, and in part by grant No. 95-00208 from the BSF, Israel, and by a grant from the Israel Science Foundation administered by the Israel Academy of Sciences and by a grant from the Recanati foundation. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

References

Alexandrov NN, Go N. 1994. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci 3*:866–875.

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

Bachar O, Fischer D, Nussinov R, Wolfson H. 1993. A computer vision based technique for 3-D sequence independent structural comparison of proteins. *Protein Eng 6*:279–288.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein databank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: Tolerance to amino acid substitutions. *Science 247*:1306–1310.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J Comp Chem 4*:187–217.

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

Crippen GM, Maiorov VN. 1995. How many protein folding motifs are there? *J Mol Biol 252*:144–151.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry 31*:7134–7155.

Fenderson FF, Kumar S, Adman ET, Liu MY, Payne WJ, Legall J. 1991. Amino acid sequence of nitrite reductase: A copper protein from achromobacter cycloclastes. *Biochemistry 31*:7180–7185.

Finkelstein AV, Gutin AM, Badretdinov AY. 1995. Perfect temperature for protein structure prediction and folding. *Proteins 23*:142–150.

Finkelstein AV, Ptitsyn OB. 1987. Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol 50*:171–190.

Fischer D, Lin SL, Wolfson HJ, Nussinov R. 1994. 3-D, sequence-order independent structural comparison of trypsin against the crystallographic database reveals active site similarities to subtilisin-like and sulfhydryl proteases: Potential implications. *Protein Sci 3*:769–778.

Fischer D, Tsai CJ, Wolfson HJ, Nussinov R. 1995. A 3D sequence-independent representation of the protein data bank. *Protein Eng 8*:981–997.

Godden JW, Turley S, Teller DC, Adman ET, Liu MY, Payne WJ, LeGall J. 1991. The 2.3 angstrom X-ray structure of nitrite reductase from achrombacter cycloclaste. *Science 253*:438–442.

Grindley HM, Artymiuk PJ, Rice DW, Willet P. 1993. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol 229*:707–721.

Harris NL, Presnell SR, Cohen FE. 1994. Four-helix bundle diversity in globular proteins. *J Mol Biol 236*:1356–1368.

Holm L, Sander C. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res 22*:3600–3609.

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

Janin J, Miller S, Chothia C. 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol 204*:155–164.

Janin J, Rodier F. 1995. Protein–protein interaction at crystal contacts. *Proteins 23*:580–587.

Jones S, Thornton JM. 1996. Principles of protein–protein interactions. *Proc Natl Acad Sci USA 93*:13–20.

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

Lee BK, Richards FM. 1971. The interpretation of protein structures. Estimation of static accessibility. *J Mol Biol 55*:379–400.

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

Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature 261*:552–557.

Lin SL, Tsai CJ, Nussinov R. 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.

Messerschmidt A, Ladenstein R, Huber R. 1992. Refined crystal structure of ascorbate oxidase at 1.9 Å resolution. *J Mol Biol 224*:179–205.

Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. *J Mol Biol 196*:641–656.

Miller S. 1989. Structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng 3*:77–83.

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

Nussinov R, Wolfson HJ. 1991. Efficient detection of motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci USA 88*:10495–10499.

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *J Mol Biol 6*:485–500.

Richardson JS. 1977. β-Sheet topology and the relatedness of proteins. *Nature 268*:495–500.

Rossmann MG, Liljas A, Branden CI, Banaszak LJ. 1975. Evolutionary and structural relationships among dehydrogenases. *Enzymes 11*:61–102.

Shortle D, Wang Y, Gillespie JR, Wrabl JO. 1996. Protein folding for realists: A timeless phenomenon. *Protein Sci 5*:991–1000.

Shrake A, Rupley JA. 1973. Environment and exposure to solvent of protein atoms. *J Mol Biol 79*:351–371.

Slingsby C, Bateman OA, Simpson A. 1993. Motifs in protein–protein interactions. *Mol Biol Rep 17*:185–195.

Steif C, Weber P, Hinz HJ. 1993. Subunit interactions provide a significant contribution to the stability of the dimeric four-helical-bundle protein ROP. *Biochemistry 31*:3867–3876.

Tsai CJ, Lin SL, Wolfson H, Nussinov R. 1996a. A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol 260*:604–620.

Tsai CJ, Lin SL, Wolfson H, Nussinov R. 1996b. Techniques for searching for structural similarities between protein cores, protein surfaces and between protein surfaces and between protein–protein interfaces. In: Marshak DR, ed. *Techniques in protein chemistry VII*. New York: Academic Press. pp 419–429.

Tsai CJ, Lin SL, Wolfson H, Nussinov R. 1997a. Studies of protein–protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Sci 6*:53–64.

Tsai CJ, Nussinov R. 1997b. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci 6*:24–42.

Tsai CJ, Nussinov R. 1997c. Hydrophobic folding units at protein–protein interfaces: Implications to protein folding and to protein–protein association. *Protein Sci.* Forthcoming.

Wang ZX. 1996. How many fold types of protein are there in nature? *Proteins 26*:186–191.

Xu D, Lin SL, Nussinov R. 1997. Protein binding versus protein folding: The role of hydrophilic bridges in protein associations. *J Mol Biol 265*:68–84.

Appendix A

To generate symmetry-related proteins around a particular protein chain is not a trivial task. To begin with, one needs to build symmetry operators for every space group appearing in the PDB (Bernstein et al., 1977) files, and to take care of all possible different notations in some space groups. In principle, two separate operations are needed to rebuild the crystal packing around a particular chain. First, crystal symmetry operations are applied to molecules in the asymmetric part of a unit cell in order to fill up the crystal packing inside a unit cell. Second, cell duplications ($3 \times 3 \times 3$ cells) along three crystal axes are utilized to generate the crystal packing between cells. In general, molecules in the asymmetric part of a unit cell are all stored in the PDB files. However, molecules in the asymmetric unit are sometime related by non-crystal symmetry. In that case, some PDB files might store the coordinates of only one molecule and give the corresponding transformation matrices in the "MTRIX" record. Hence, we begin by applying any transformation matrix given in "MTRIX" to generate all molecules occupying the asymmetric unit.

To be convenient to the general user, the stored coordinates in the PDB file have been converted from the deposited fractional coordinates to orthogonal Cartesian coordinates. Because crystal symmetry operators are only applicable to fractional coordinates, one has to re-convert the stored coordinates back to the original fractional coordinates. Fortunately, the transformation matrix is specified in the "SCALE" record. After the crystal symmetry operation plus cell duplication (by adding, or subtracting, one to/from the fractional coordinates), we apply an orthogonal matrix to convert the fractional coordinates back to orthogonal coordinates. The orthogonal transformation matrix has been calculated by the six lattice constants according to the specification in the PDB format manual. Unfortunately, for some PDB files, the interconversion between fractional coordinates and Cartesian coordinates are inconsistent. To fix this problem, we have made an extra transformation to the final Cartesian coordinates. When encountering an inconsistency, we superimpose the Cartesian coordinates that have been generated through the fractional–orthogonal conversion process onto the original stored Cartesian coordinates in order to find the "fix" transformation matrix. Although most PDB errors can be recovered by this extra "fix" operation, there are still a few PDB entries with inherent errors, owing to a substantial atomic overlap between the symmetry-related units. In such cases, the PDB entry was simply excluded from our calculations.

In a six-character notation of the interface generated by the symmetry operation, the first four characters represent the PDB code, and the fifth character is the chain identifier (if not given in PDB file, it has been replaced by a "0" character). The sixth character specifies how the symmetry-related chain was generated. If it is a number, it has been generated by the transformation matrix given in the "MTRIX" record, and the number corresponds to the order of the transformation matrices. We only consider the first 12 transformation matrices although some PDB entries give as many as 59 matrices. If the sixth character is a lower case letter (starting with "a" along the alphabet order), it has been generated by crystal symmetry operation. In our scheme, the identification of a nearby protein generated by the symmetry operation is specified by its distance to the protein, given in an asymmetric unit. The distance is calculated between the centers of the two molecules. The sixth letter refers to the symmetry-related protein as being specified by the order of its distance to the asymmetric protein, starting from the closest one.

Appendix B

A list of 187 stable protein–protein interfaces and 38 oligomeric interfaces (Tsai, et al., 1997a) generated by crystal symmetry operation.

1041AB	1aarAB	1abbCD	2afnBC	1alkAB	1atnAD	1atpEI
1babAD	1barAB	1bbbAB	1bbhAB	1bbpBD	1bbrLE	1bbrHE
1bbrKG	1bbt12	1bbt14	1bgsBF	1bmv12	1bovAE	1bsrAB
1cauAB	1caxCF	1chmAB	1choEI	1chrAB	1cmbAB	1colAB
1cosAB	1cpcAB	1cseEI	1csgAB	1ctdAB	1d66AB	1fbaAB
1fbaAD	1fc1AB	1fc2CD	1fcbAB	1fiaAB	1fod12	1gd1PQ
1gdhAB	1ggaAB	1glaFG	1gmaAB	1gp1AB	1hgeBF	1hgeEF
1hgtLH	1hgtHI	1hhhAC	1hhjAB	1hilCD	1hleAB	1hviAB
1hunAB	1ldnFG	1lldAB	1lmb34	1ltaEC	1ltsDE	1ltsAC
1mchAB	1mcoLH	1mec13	1minAD	1minBD	1minCD	1mioAB
1mioBD	1mypBD	1nipAB	1nscAB	1ovaAB	1ovoAB	1pdgAB
1pfkAB	1plfAB	1plfAC	1plfBD	1poxAB	1pp2RL	1prcCL
1prcCM	1prcLM	1prcLH	1prcMH	1psaAB	1ptsAB	1pyaAC
1pyaDE	1pyaEF	1pydAB	1pygAB	1r0913	1ragBD	1rfbAB
1rhgAC	1ribAB	1sacAB	1scmAB	1scmAC	1srnAB	1sryAB
1stfEI	1tbpAB	1tme12	1tme13	1tme34	1tnfAB	1tplAB
1trkAB	1troAC	1trzBD	1vfaAB	1vsgAB	1wsyAB	1ximAB
1ximAC	1ximAD	2aaiAB	2bbkHL	2bbkLJ	2bbnAB	2bpa13
2ccyAB	2cgrLH	2cwgAB	2fb4LH	2gstAB	2hhmAB	2hmzCD
2hppHP	2ltnAC	2ltnCD	2nckRL	2ohxAB	2pcbAB	2pccCD
2pfkAB	2plv12	2plv14	2plv23	2polAB	2scpAB	2spcAB
2tmdAB	2tprAB	2tscAB	2tunAD	2utgAB	3aahAC	3aahCD
3ecaAB	3ecaBC	3ecaBD	3gapAB	3hhrAB	3insAB	3ladAB
3mcg12	3monCD	3rubLS	3sc2AB	3sdhAB	3sgbEI	4chaAB
4ctsAB	4fbpAC	4fbpCD	4htcHI	4rubBC	4rubCV	4rveAB
4ts1AB	5cnaBD	5cnaCD	5cscAB	5rubAB	7aatAB	7timAB
8catAB	8fabCD	8rsaAB	9ldtAB	9rubAB		

1bam0a	1cae0b	1caf0c	1dgd0a	1dge0c	1dupA1	1grl01
1gsa0a	1hcy01	1idm0a	1lgr01	1lgr06	1lih0a	1mal01
1mpf0a	1msc0a	1pth01	1pyp01	1rpo0a	1svb0d	1trb0a
2bb20a	2gn50a	2mnr0a	2mnr0b	2pgd0a	2rhe0a	2sblB1
2ssi0a	2tct0a	2tgi0a	4enl0a	5ldh0b	6ldh03	7catAa
7catAb	7icd0a	8ldh02				