PROTEINS: Structure, Function, and Genetics 39:331–342 (2000)

Conservation of Polar Residues as Hot Spots at Protein Interfaces

Zengjian Hu,[1] Buyong Ma,[1] Haim Wolfson,[2] and Ruth Nussinov[1,3,4]*
[1]*NCI-FCRDC, Laboratory of Experimental and Computational Biology, Frederick, Maryland*
[2]*Computer Science Department, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel*
[3]*Sackler Institute of Molecular Medicine, Tel Aviv University, Tel Aviv, Israel*
[4]*SAIC, Laboratory of Experimental and Computational Biology, Frederick, Maryland*

ABSTRACT **A number of studies have addressed the question of which are the critical residues at protein-binding sites. These studies examined either a single or a few protein–protein interfaces. The most extensive study to date has been an analysis of alanine-scanning mutagenesis. However, although the total number of mutations was large, the number of protein interfaces was small, with some of the interfaces closely related.**

Here we show that although overall binding sites are hydrophobic, they are studded with specific, conserved polar residues at specific locations, possibly serving as energy "hot spots." Our results confirm and generalize the alanine-scanning data analysis, despite its limited size. Previously Trp, Arg, and Tyr were shown to constitute energetic hot spots. These were rationalized by their polar interactions and by their surrounding rings of hydrophobic residues. However, there was no compelling reason as to why specifically these residues were conserved. Here we show that other polar residues are similarly conserved. These conserved residues have been detected consistently in all interface families that we have examined. Our results are based on an extensive examination of residues which are in contact across protein interfaces. We utilize all clustered interface families with at least five members and with sequence similarity between the members in the range of 20–90%. There are 11 such clustered interface families, comprising a total of 97 crystal structures. Our three-dimensional superpositioning analysis of the occurrences of matched residues in each of the families identifies conserved residues at spatially similar environments.

Additionally, in enzyme inhibitors, we observe that residues are more conserved at the interfaces than at other locations. On the other hand, antibody–protein interfaces have similar surface conservation as compared to their corresponding linear sequence alignment, consistent with the suggestion that evolution has optimized protein interfaces for function. Proteins 2000;39:331–342.

Key words: protein–protein interfaces; hot spots; molecular recognition; protein folding and binding; residue conservation; polar residues

INTRODUCTION

Proteins, like other types of molecules in vivo or in vitro, function through their binding. Intermolecular associations are involved in practically all biological functions. A number of studies have investigated the principles of protein–protein interactions, interface size and shape, number and type of amino acids involved, hydrophobicity, hydrophilic interaction, segmentation, and secondary structure analysis.[1–9]

Investigations of a large number of protein–protein interfaces have shown them to comprise a recurring set of structural motifs.[10] Hence, it has been suggested that, analogous to the situation in protein chains, there are also a limited number of ways for proteins to interact. Although it is generally agreed that the hydrophobic effect is the driving force in protein folding,[11] the chain does not fold in a way which optimizes its buried nonpolar surface area. Similarly, an examination of the non-polar buried surface area between two chains illustrates that although it may be substantial, there are frequently alternate ways for the two molecules to interact, yielding a stronger hydrophobic effect.[12] These facts complicate the prediction of folding and binding[13,14] and contribute to the difficulty in the differentiation between crystal and biological interfaces.[15–17] Analysis of the frequencies of residue-pairs across the interfaces has shown these to be similar to those within protein monomers. An empirical set of solvent-mediated potential functions has been derived both for intramolecular and intermolecular interactions and has been shown to hold equally well for both.[18] Consistently, it has been shown that there are large patches of hydrophobic residues at protein-binding sites.[19,20] On the other hand, because monomers are stable in solution, they have a larger fraction of polar surface area on their surfaces as

Grant sponsor: National Cancer Institute, National Institutes of Health; Grant number: NO1-CO-56000; Grant sponsor: Binational Science Foundation, Israel; Grant number: 95-00208; Grant sponsor: Israel Science Foundation; Grant sponsor: Ministry of Science; Grant sponsor: Tel Aviv University; Grant sponsor: Center of Excellence, Israel Academy of Sciences.

*Correspondence to: R. Nussinov, NCI-FCRDC, Building 469, Room 151, Frederick, MD 21702. E-mail: ruthn@ncifcrf.gov

Received 30 September 1999; Accepted 7 February 2000

compared to the surface area buried within protein chains.[15,21]

As in folding, intermolecular binding has also been shown to be relatively robust to mutational events. What is often observed is that side chains would rotate, to occupy the space vacated by the altered residue.[22] Nevertheless, some residues have larger effect than others, and some are absolutely essential for preserving binding specificity. The functionally more important residues form the binding epitope, i.e., the energetic or functional epitope of the molecule.

Clackson and Wells have probed the energetic contributions of individual side chains to protein binding through alanine-scanning mutagenesis.[23] Recently, a computational alanine scanning was also performed to probe protein–protein interactions by calculating binding free energies.[24] Interestingly, Clackson and Wells found that despite the large size of the binding interface, single residues can still contribute a large fraction of the binding free energy.[22,25] Consequently, the critical components in a functional epitope have been defined as those sites where alanine mutations caused a significant change in binding affinity. Bogan and Thorn[8] have compiled and analyzed a dataset of 22 protein interfaces. Their goal was to study the contributions of individual amino acid side chains to protein–protein binding. Their analysis has shown that at the level of the side chains, there is little correlation between buried surface area and free energy of binding. Interestingly, they found that the free energy of binding is not distributed evenly across protein–protein interfaces. Instead, there are hot spots of binding energy consisting of a small subset of residues at the dimer interface. These hot spots are enriched in tryptophan, tyrosine, and arginine and are surrounded by a shell of energetically less important residues that most likely serve to occlude bulk solvent from the hot spot.[8]

Recently, we have surveyed exhaustively the structures of protein–protein interfaces.[4] We have carried out structural comparisons of the interfaces, in a manner that is independent of both the sequence at the interface and the entire fold of the protein chain to which the interface belongs. Based on the interface architectures, we have clustered the interfaces into 351 families. These clustered, structurally similar interface families provide a rich dataset, enabling an examination of the extent of residue conservation and variability within the family. However, at the same time, the families themselves differ from each other, representing a structurally diverse set.

Here we utilize this set to carry out a systematic structural analysis of protein–protein interfaces. Out of the entire dataset, we select 97 protein–protein interfaces which fulfill our criteria with respect to the minimal number of members in a family, and the extent of the homology between any member and the corresponding family representative. Our final dataset is listed in Appendix A. These interfaces are divided into 11 families according to the extent of the geometric similarities of their crystal structures (taken from the Protein Data Bank, PDB).[26] Each interface of a family is superimposed structurally (C_α atoms only) on the interface that represents the family. We calculate the number of matched residues in the two superimposed interfaces and the number of identical or strongly conserved residues among the matched residues. We probe which residues in these protein interfaces are conserved, which are the major determinants in the binding (hydrophobicity or other), and which amino acids are preferred, avoided, or neutral in the interface.

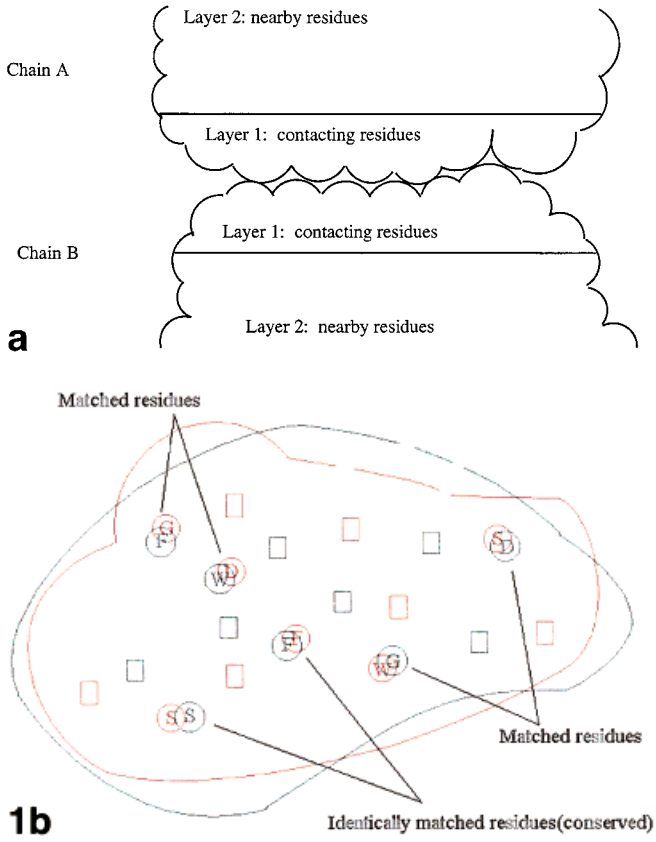

Fig. 1. **a:** An illustration of the definition of a protein–protein interface as used in this study. **b:** An illustration of an interface alignment to locate matched residues.

Whether in folding or in binding, it is universally agreed that if residues have a tendency to be conserved, they are likely to fulfill some role. In binding, conserved residues may contribute to structural stability, catalysis, or recognition. Similarly, in protein folding, it has been suggested that a folding core may be predicted as a set of conserved stabilizing residues without an apparent functional role, other than for folding.[27] A particularly interesting recent study of RNA-binding proteins found that solvent-exposed aromatic residues on a β-sheet surface are highly conserved among RNA-binding domains.[28] Here we examine the functional and stabilizing roles of conserved residues at protein–protein interfaces. In the future we plan to analyze conserved residues in the partners, across the interfaces.

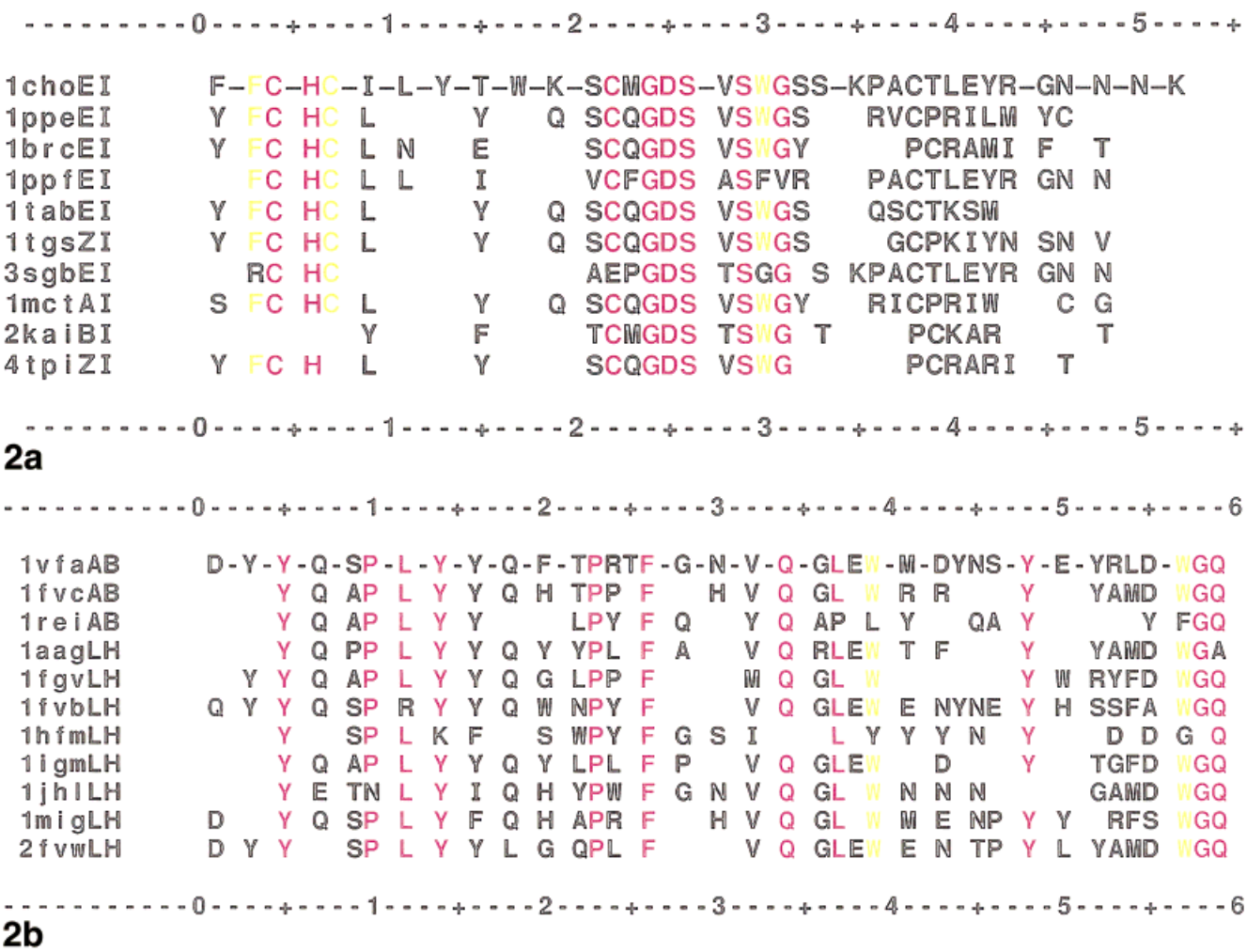

Fig. 2. Multiple interface alignments to identify conserved residues. Only first-layer interface residues are shown. If the residues listed here are separated by a line or a space, they are not sequentially connected. Red letters indicate absolutely conserved residues, and yellow letters refer to strongly conserved (above 80% conservation) residues. **a:** The 1choEI family, an example of conserved residues within the same segment of the chain. **b:** The 1vfaAB family, an example in which conserved residues are distributed over different segments of the interface.

MATERIALS AND METHODS

Definition of Protein–Protein Interfaces

An interface is composed of at least two chains. An interface consists of residues that interact with each other across the binding interface and those that are in their vicinity in the supporting scaffold, within a certain distance threshold. The selection of interfacing residues in one chain of the interface is based on how close they are to the other chain. As illustrated in Figure 1a, the two-chain protein interface in our definition has two layers: The first layer is composed of residues that are in actual contact (contact residues), whereas the second layer includes residues that are "nearby," within a certain vicinity of residues that have been assigned to belong to the first layer. Specifically, if the distance between two atoms, each belonging to a different chain, is less than the sum of their van der Waals radii plus a proximity threshold of 0.5 Å, both residues to which the atoms belong are registered as interfacing residues (contacting residues, first layer). To enable an examination of the types of architectures at the interface, residues whose C_α atoms are within a distance of 6.0 Å from a C_α atom of an already assigned contacting residue are included as "nearby" residues (second layer). Further details are given in Tsai et al.[4] In this work, most of the analysis pertains to the contacting residues, i.e., first layer. Hence, unless explicitly stated differently, the results presented are with regard to this first layer.

Structural Comparison Algorithm and Definition of Conserved Residues

To be able to highlight the major structural features of protein–protein interfaces, we need a technique that enables carrying out structural comparisons between them, without taking into account the linear sequence alignment. Because protein interfaces are composed of at least two chains, and, possibly only discontinuous pieces of each,

TABLE I. Summary of the Dataset of 11 Protein–Protein Interface Families†

No.	Family representative	Number of members	Protein family	Number of interface residues (%)		
				Maximum	Minimum	Average
1	1babAD	5	Hemoglobin	29 (10.1)	18 (6.3)	20 (7.1)
2	1bbbAB	6	Hemoglobin	42 (14.6)	37 (12.9)	39 (13.9)
3	1choEI	10	Serine proteinase–inhibitor	47 (16.7)	28 (12.7)	37 (14.3)
4	1cseEI	6	Serine proteinase–inhibitor	47 (13.9)	37 (9.7)	40 (11.8)
5	1hhjAB	6	Histocompatibility antigen	64 (17.2)	52 (14.1)	57 (15.3)
6	1hilCD	30	Immunoglobulin	91 (20.9)	72 (17.2)	81 (18.9)
7	1hviAB	5	Protease complexes	76 (38.4)	60 (27.2)	69 (34.1)
8	1vfaAB	11	Immunoglobulin	40 (17.8)	31 (14.5)	35 (15.8)
9	2gstAB	5	Glutathione transferase	62 (14.0)	51 (11.8)	55 (13.0)
10	2plv12	6	Picornavirus and rhinovirus	113 (20.3)	56 (14.1)	94 (18.3)
11	3insAB	7	Hormone	34 (66.7)	24 (5.2)	29 (51.5)

†The numbers in parentheses are percentages based on total residues in chains A and B (Fig. 1a).

any structural comparison method which is dependent on the order of the residues in the chain would be inadequate. Here we utilize the Geometric Hashing algorithm, which considers protein structures in a manner that is divorced from their polypeptide chain connectivity. This sequence-order–independent, computer-vision–based structural comparison technique views protein structures as collections of points (atoms) in three-dimensions (3D) space, disregarding the order of the residues in the chains.[4,29–31]

We start by picking one model interface, where the interface includes both chains A and B, each with both first and second layers (Fig. 1a). A second, query interface (which also includes both chains A′ and B′, and both first and second layers) has been rotated and translated by the Geometric Hashing to achieve maximal overlap between the corresponding C_α atoms of this interface and the model interface. The result is shown in Figure 1b. The figure illustrates schematically a map of the two superimposed protein–protein interfaces from a clustered family, selected based on our previously described classification. Note that the superimposition is structure based and sequence-order independent. The model interface is the representative of the family, and the query interface is a family member. The superpositioning of the two interfaces, computed by the Geometric Hashing algorithm, yields a list of matching C_α atom pairs. The details of the matching algorithm and the measure of similarity are as reported previously.[4] Next, as indicated in Figure 1b, for interfaces A and B we obtain the number of matched residues and its percentage, based on the total number of surface residues. If the matched pair is of the same residue type (for example, both are alanines), we define an identical pair for interfaces A and B. For a given clustered interface family, we separately align the interfaces of all family members with the interface that represents the family, obtaining the statistics of identically matched residues. If an identical pair has been found in all superposed interfaces, for all family members, we define an absolutely conserved residue at this given interface position (Fig. 2). Some matched residue pairs are not absolutely conserved. If their occurrence is above 80%, they are defined as strongly conserved residues. In this work, the term conserved residue includes both absolutely and strongly conserved residues and only first layer residues.

Propensity

The propensity (P_i) of a residue i, $i = 1, 2 \ldots 20$ (20 residues), to occur at the interface is calculated as the fraction of the count of residue i in the interface as compared with its fraction in the whole chains,

$$P_i = (n_i/N_i)/(n/N)$$

where n_i is the number of residues of type i at the interface, N_i is the number of residues of type i in the chains, n is the total number of residues at the interface, and N is the total number of residues in the chains.

The propensity (P_m) of residue m to match identically in the interface is calculated by

$$P_m = (n_m/n)/(N_m/N)$$

where n_m is the number of identically matched residues of type m at the interface, n is the number of all identically matched residues at the interface, and N_m is the number of residues of type m in the chains.

The propensity (P_k) of residue k to be conserved in the interface is calculated by

$$P_k = (n_k/n)/(N_k/N)$$

where n_k is the number of conserved residues of type k at the interface, n is the number of residues at the interface, and N_k is the number of residues of type k in the chains.

A propensity greater than 1 indicates that the residue occurs more frequently in the interfaces than in other locations in the protein.

RESULTS AND DISCUSSION

Summary of the Datasets Used

We have used the dataset of 351 families with 1,629 protein–protein interfaces[4] to select a structurally non-redundant subset of 11 interface families. Our selection covers all families with at least five distinct interface members. Within each family, the sequence similarity (derived from sequence alignment) is below 90% (with a

TABLE II. Summary of Matched and Conserved Interface Residues of 11 Interface Families

No.	Family representative	Maximum (%)		Minimum (%)		Average (%)		Layer one residues			Layer two residues		
		Matched[a]	Identical[b]	Matched[a]	Identical[b]	Matched[a]	Identical[b]	Conserved[c]	Total[d]	%	Conserved[c]	Total[d]	%
1	1babAD	77.8	59.0	72.2	27.0	74.0	50.8	4	20	20.0	24	36	66.7
2	1bbbAB	94.6	82.2	90.8	47.2	93.0	71.4	2	39	5.1	25	58	43.1
3	1choEI	85.0	54.2	61.4	21.0	74.8	41.2	9	37	24.3	7	47	14.9
4	1cseEI	98.4	76.6	81.2	48.4	87.8	61.0	22	40	55.0	27	56	48.2
5	1hhjAB	95.2	87.2	91.8	69.2	93.6	76.8	40	57	70.2	48	78	61.5
6	1hilCD	90.0	86.0	59.8	30.0	77.8	54.6	24	81	29.6	42	107	39.3
7	1hviAB	99.2	69.4	77.6	34.6	88.2	55.2	28	69	40.6	10	41	24.4
8	1vfaAB	95.6	62.6	81.2	36.6	87.6	50.6	12	35	34.3	17	52	32.7
9	2gstAB	93.2	70.2	87.8	20.8	89.2	36.6	14	55	25.5	17	75	22.7
10	2plv12	89.2	43.4	62.2	20.6	73.2	29.8	21	94	22.3	12	78	15.4
11	3insAB	94.2	92.2	76.6	22.2	86.0	54.2	7	30	23.3	4	19	21.1

[a]The percentage of the matched residues normalized by the interface size (both layers, Fig. 1a).
[b]The percentage of the identical residues normalized by the interface size (both layers, Fig. 1a).
[c]Number of conserved residues.
[d]Total number of residues in the interface (both chains A and B, Fig. 1a).

TABLE III. Distribution of Conserved Amino Acid Types in the 11 Interface Families

No.	Family representative	Number of conserved residues[a]																			
		G	A	V	F	P	M	I	L	D	E	K	R	S	T	Y	H	C	N	Q	W
1	1babAD			1		1			1						1						
2	1bbbAB								1				1								
3	1choEI	2								1				2			1	3			
4	1cseEI	6	1		1	1		1	2					3	1	1	1		3	1	
5	1hhjAB	2	2	2	3	2	1	1	2	4	1		3	2	2	3	2		1	4	3
6	1hilCD			1	2	2	1		4		1			6	1	1	1		1	3	
7	1hviAB	6				2			8	4			4		4						
8	1vfaAB	1			1	2			2							3				2	1
9	2gstAB		2							4			4						2	2	
10	2plv12	2		1		3					2		1		2	3	2	1	3	1	
11	3insAB								2									5			
Total		19	5	5	7	13	2	2	22	13	4	0	13	13	11	11	7	8	10	13	4

[a]After we superimpose interfaces for a given clustered interface family, we look up identically matched residues. If an identical pair has been found in all superposed interfaces with more than 80% occurrence, it is defined as conserved residues.

range of 20–90%). The dataset, listed in Table I and in Appendix A, contains a total of 97 protein–protein interfaces, and includes oligomeric proteins, enzyme–inhibitor complexes, antibody–antigen complexes and hormone–receptor complexes.

The clustering of the protein–protein interfaces is based on the residue-connectivity score as the measure of similarity.[4] In the present study, we are concerned with the mapping of specific residues at the interfaces. Hence, it is the number of matched residue pairs which is relevant for our purpose. Table II lists the geometric similarities of the interface families studied here. This table shows that our datasets include interfaces with a compromise between similarity and diversity in terms of the number of matched residues. The number of residues in the interfaces ranges between 18 and 113, i.e., 6.3–66.7% of the total number of residues from both chains (chains A and B, Fig. 1a). As for the entire interfaces, the ratio of the first layer (interacting residues) to the second layer (residues belonging to the second shell) ranges between 1:1 and 1:25, i.e., the second layer is generally larger.

Mapping of Interface Families: Conserved Interface Residues

For all families studied here, a high percentage of matched residues has been found, owing to the clustering process. Even though the root-mean-squared deviation (RMSD) has not been considered in measuring the similarity between the two interfaces, the RMSDs of both C_α and C_β atoms are generally very small. Conserved residues may serve as fingerprints characterizing the interface family. Table III shows that all families have their own set of conserved residues, which are in contact across the interface. With few exceptions, the percentages of the conserved residues range from 20–50% of all contacting residues.

Here we focus on two interface families from our dataset, 1choEI and 1hhjAB. We chose these two families because they have long chains and because large portions of their structures are not included in the interface regions. They represent two different types of proteins, the serine proteinase inhibitor (1choEI) and the histocompat-

TABLE IV. Geometric Similarities of the Interface Families Versus Sequence Similarity

				Number of identical residues		Percentage of identical residues (%)	
Interface	Size[a]	Interface	Size[a]	Interface[b]	Chain[c]	Interface[d]	Chain[e]
1choEI	94	3sgbEI	82	35	78	39.8	29.6
1choEI	94	1ppeEI	93	39	97	41.8	35.8
1choEI	94	1brcEI	83	35	98	39.6	34.4
1choEI	94	1ppfEI	87	49	123	54.2	43.6
1choEI	94	1tabEI	82	39	98	44.4	35.8
1choEI	94	1tgsZI	98	47	110	49.0	38.4
1choEI	94	1mctAI	80	35	100	40.2	37.0
1choEI	94	2kaiBI	59	16	50	21.0	20.0
1choEI	94	4tpiZI	73	34	96	40.8	33.6
1hhjAB	142	2mhaAB	127	94	235	69.8	63.2
1hhjAB	142	1hocAB	134	97	258	70.2	79.8
1hhjAB	142	1hsaAB	133	120	334	87.2	89.0
1hhjAB	142	2vaaAB	141	98	265	69.2	70.8
1hhjAB	142	1hsaDE	133	120	334	87.2	89.0
1vfaAB	89	1fvcAB	87	47	132	53.4	58.2
1vfaAB	89	1reiAB	86	32	98	36.6	44.6
1vfaAB	89	1aagLH	95	45	116	49.0	51.6
1vfaAB	89	1fgvLH	77	47	139	56.6	61.6
1vfaAB	89	1fvbLH	89	47	116	52.8	51.6
1vfaAB	89	1hfmLH	84	37	126	42.8	56.8
1vfaAB	89	1igmLH	82	45	132	52.6	56.4
1vfaAB	89	1jhlLH	89	46	118	51.6	52.6
1vfaAB	89	1migLH	90	56	153	62.6	67.4
1vfaAB	89	2fvwLH	96	44	116	47.6	50.8

[a]Total number of interface residues (both layer 1 and layer 2, Fig. 1a).
[b]Number of identical residues in the interface (both layers, Fig. 1a).
[c]Number of sequence identities.
[d]Percentage of identical residues in the interface (both layers, Fig. 1a).
[e]Percentage of sequence identity.

ibility antigen (1hhjAB) systems. We compare the percentages of the conserved residues which are in contact across the interface versus the percent of sequence identities in the entire protein chains. Table IV shows that in the 1choEI family, residues are slightly more conserved at the interfaces than overall. On the other hand, although the 1hhjAB family has a larger percent of conserved interface residues (Tables II and III), it is mainly due to the larger overall sequence similarity of the chains in the family. When compared with the overall sequence alignment, the alignment of the 1hhjAB family does not show a preference for more conserved interface residues.

It is well known that the overall sequence and structural similarity among serine proteinases is limited.[31] This is consistently observed in a multiple sequence alignment of the six serine proteinase inhibitors whose structures are in the 1choEI family (1choEI, 4sgbEI, 1mctAI, 2kaiBI, and 4tpiZI). The RMSDs between the structures are also large (Table V; Figs. 2a and 3). On the other hand, the interfaces match well, with a 21% (8 out of 37) residue conservation (Fig. 2; Table IV). The conserved residues are located mainly around the active site, making it difficult to separate the catalytic function from the binding affinity. In contrast to the serine proteinases, the antibody family (represented by 1hhjAB) shows a random surface distribution. We have examined multiple sequence alignments and the overall versus interface structural superpositions of additional antibody–antigen cases (represented by 1vfaAB), obtaining similar results. Consistently, it has been proposed that during evolution, the homodimers, enzyme inhibitors, and heterocomplexes have evolved to optimize their interface interactions. In contrast, the antibody–antigen interfaces may be selected principally according to the strength of their binding, as measured by their binding constant, without being subject to evolutionary optimization.[3]

TABLE V. Comparison of RMSD From Interface Alignment and Chain Alignment (Å)[†]

	Multiple chain alignment				
Interface	1cho	2kai	1mct	3sgb	4tpi
1cho	0	6.1	7.7	7.7	7.7
	Interface alignment				
Interface	1cho	2kai	1mct	3sgb	4tpi
1cho	0	1.2	1.0	1.3	1.0

[†]The RMSDs of the chain alignment are calculated from the $C\alpha$ positions of identical residues in the superposition of whole protein structure. The RMSDs of interface alignment are calculated from the $C\alpha$ positions of identical residues in the superposition of protein interfaces.

Fig. 3. Multiple chain alignment for the 1choEI family. Red letters indicate superimposed residues. The box indicates a conserved sequence region.

Distribution of Interface Residues and Preferred Conservation of Polar Residues on Protein–Protein Interfaces

It is generally agreed that protein–protein interfaces are largely hydrophobic, with the hydrophobicity being a dominant force in protein–protein interactions.[19,32] Analysis of protein–protein interfaces has shown that the extent of hydrophobicity, as measured by the non-polar buried surface area between the two chains, may vary to a large extent between the interfaces.[5] Although the hydrophobic effect in protein–protein interfaces is extremely important, it is not as strong as that observed in the interior of the monomers. Consistently, there are also indications of the existence of hydrophilic regions.[1,32,33] Two recent studies have highlighted the importance of hydrophilic interactions in biological function.[34,35] NK cells are a fundamental component of the innate immune system and have the intrinsic ability to recognize and destroy certain virally infected and tumor cells. The interface of the C-type-lectin–like receptor family (Ly49 A) has been observed to be highly hydrophilic, dominated by charged interactions.[34] A similar situation has been observed for the human CD2 molecule,[35] a transmembrane cell surface glycoprotein found on virtually all T cells, thymocytes, and NK cells. In general, electrostatic interactions have been shown to play a more important role in binding than in folding.[21] With regard to specific residues, the reported preferences vary from case to case, mainly owing to the change in the systems studied. In studies of a specific system, it is difficult to differentiate between residue conservation conferring binding specificity and conservation owing to the role of the residues in constituting energy hot spots. The large dataset used in the present study (97 protein–protein interfaces) provides a more general picture.

Here we examine the distributions of all interface residues, of identically matched residues between the representative of the family and each of the family members, and of conserved residues in the whole family. For individual families, there are significant preferences of certain residues to be at the interfaces. However, when examining the counts of identically matched residues, the preferences become much less significant. Figure 4 illustrates the propensities of the different residue types at the interfaces. Residue propensities shift when examining all interface residues as compared to identically matched residues, and as compared to conserved residues.

Whereas the distribution of all interface residues yields simple statistics, the propensities of conserved residues provide clues to the more important driving forces. Table VI summarizes the propensities in terms of hydrophobic or hydrophilic residue classifications. Here aromatic residues are considered hydrophobic. Table VI shows that with

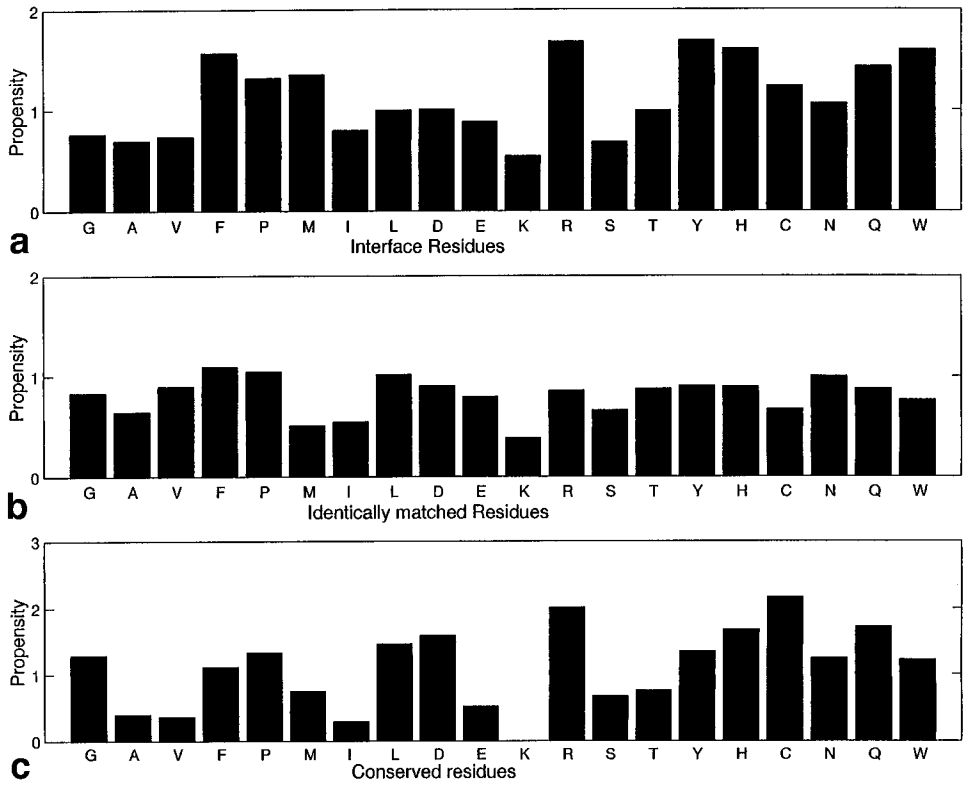

Fig. 4. Propensities of all interface residues (**a**), of identically matched residues (**b**), and of conserved residues (**c**).

TABLE VI. All Residues, Identical Residues, and Conserved Residues

	Total interface residues	Identically matched interface residues	Conserved interfaces residues
H	5.9	4.6	5.7
A	6.5	3.7	5.4
H + A	12.4	8.3	11.1
P	9.6	7.0	14.9

H, Hydrophobic residues, including Ala, Val, Pro, Met, Ile, Leu; P, polar residues, including Asp, Glu, Lys, Arg, Ser, Thr, Asn, Gln; A, aromatic residues, including Phe, Tyr, His, Trp.

respect to the total number of interface residues, hydrophobic residues have higher propensities than polar residues (12.4:9.6), consistent with previous, statistically based studies.[5,19] However, remarkably, the preference of hydrophobic residues decreases when analyzed in terms of the propensities of identically matched residues (8.3:7.0). Finally, in terms of conserved residues, polar residues dominate (11.2:14.9). Hence, we observe a trend of preferred conservation of polar residues in protein–protein interfaces.

Conserved Interface Residues and Energetic Hot Spots

A comparison of our conserved interface residues with experimentally identified energetical hot spots illus-

TABLE VII. Comparison of Experimentally Determined Enrichment in Hot Spots and Conserved Residue Propensities

Residue	Enrichment in hot spots[a]	Propensity to conservation[b]
Trp	3.91	1.22
Arg	2.47	2.01
Tyr	2.29	1.35
Leu	0.01	1.45
Ile	1.79	0.29
Asp	1.67	1.59
His	1.49	1.68
Pro	1.25	1.33
Lys	1.17	0
Asn	0.93	1.24
Glu	0.68	0.52
Gln	0.58	1.73
Phe	0.56	1.12
Met	0.54	0.74
Gly	0.45	1.29
Thr	0.28	0.75
Ser	0.21	0.67
Val	0	0.36

[a]Enrichment in hot spots gives the fold enrichment of that residue type in hot spots ($\Delta\Delta G \geq 2$ kcal/mol in alanine mutations) over the whole database of 2,325 alanine mutations.[8]

[b]This work. The propensity of conservation (P_k) of a residue to occur at the interface is calculated as the fraction of the count of residue k in the interface as compared with its fraction in the whole chains. $P_k = (n_k/n)/(N_k/N)$, where n_k is the number of conserved residues of type k at the interface, n is the number of residues at the interface, N_k is the number of residues of type k in the chains, and N is the total number of residues in the chain.

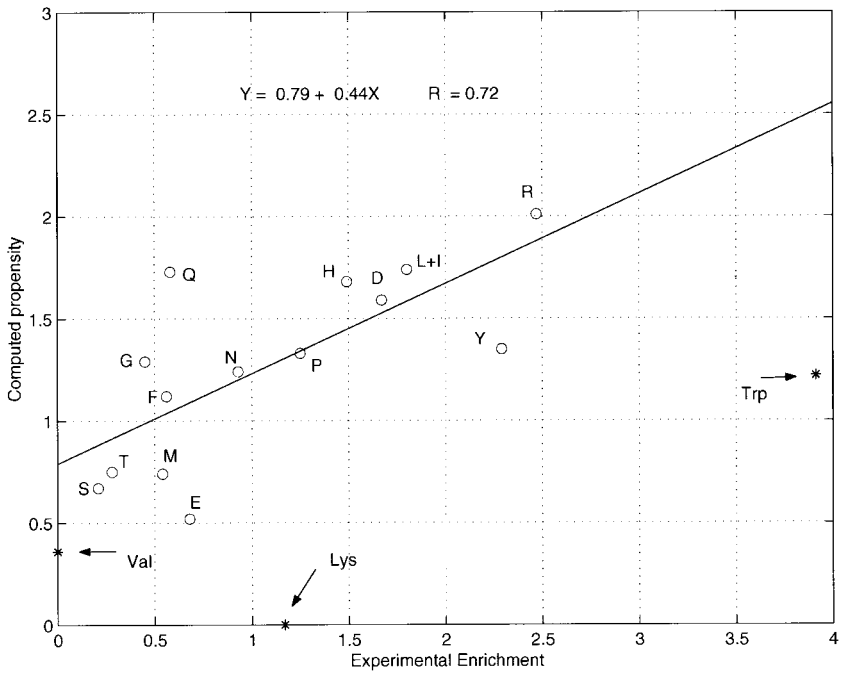

Fig. 5. A correlation of experimentally determined amino acid enrichment and our computed conservation propensity. Residues are indicated by their one-letter code. Outliers are indicated. They are not included in the least-squares fitting. Note that Val is classified as an outlier because of its zero value in the experimental results.

trates that the two most conserved residues are cysteine and arginine (Fig. 4). The high propensity of cysteine stems from the disulfide bond conservation, either to maintain the interface structure (for example, in the 1cho family) or to chemically bind interfaces together (e.g., in the hormone protein 3insAB family). The identification of cysteine validates our conserved residue propensity computations.

Table VII and Figure 5 show that our selection of conserved residues corresponds to experimentally identified hot spots. There is a surprising match between the propensities of most of our conserved residues and the residue enrichment in hot spots compiled from the database of alanine-scanning mutagenesis.[8] Except for a few outliers, the correlation coefficient of the experimentally determined amino acid enrichment and our computed conservation propensity is 0.72. Of the unmatched residues, alanine is incomparable. Cysteine is favored in our database, owing to its forming covalent bonds, and is thus also incomparable. Tryptophan has a lower propensity in our analysis, probably the outcome of its rareness. Serious disagreement is found for lysine. Lysine is well represented in our interface database. However, although we failed to identify conserved lysine residues, lysine has a good enrichment in the alanine-scanning data. A possible reason for this failure may relate to our high criteria of cut-off percentage (80%) for a conserved interface residue. Alternatively, lysine is a highly flexible surface residue, projecting its tail into the solvent. It is quite possible that it is not detected as a matched residue pair in the superpositioning owing to slightly larger shifts in its position. At first sight it appears that there is also a serious disagreement between our results and the alanine scanning with respect to leucine. However, if we combine the contributions of leucine and isoleucine, a good agreement is obtained (1.80 from alanine scanning, 1.74 from conservation propensity). Furthermore, this agreement also addresses the Bogan and Thorn's question as to why isoleucine has a higher frequency than leucine in their database.[8] Taken together, the solution may be that one should consider the sum of leucine and isoleucine, as we did in Figure 5. There are also disagreements in the lower third of the table, particularly in Gln, Phe, and Gly. Whereas the difference in Gly can be attributed to flexibility, we have no plausible explanation for the more frequent occurrence in Gln and Phe in our interface families as compared to the alanine scanning, apart from the limited number of dissimilar interfaces examined by the alanine scanning. In general, we see a higher preponderance in conserved polar residues (His, Asn, Gln, Thr, Ser), or partially polar (Phe, Met), as compared to the alanine scanning. This is consistent with the general observation and conclusion that have been reached by Bogan and Thorn.[8] Considering that the two methods are so different and that essentially there is no overlapping between the protein structures used in the two databases, the convergence of the results is remarkable. Further, Bogan and Thorn[8] have examined the location of the hot spots across interfaces. They found that the hot spots are usually located around the center of the interfaces, and hence protected from bulk solvent. This feature is also observed in our location of conserved interface residues.

Hence, whereas overall interfaces manifest a higher frequency of occurrence of hydrophobic residues,[5,19,32] the polar residues are those that are preferentially conserved. Within these, Bogan and Thorn[8] find a special enrichment of Trp, Arg, Tyr, Asp, Pro, and His. Analysis of the residue conservation in a larger number of interfaces shows Arg, Gln, His, Asp, Pro, and Asn to be especially enriched.

Conserved Interface Residues and Protein Function

Folding and binding are similar processes, manifesting similar types of physical interactions. Furthermore, both are part of an integrated series of steps toward protein function. In general, there is a high sequence homology within a given family of proteins. The observed sequence homology decreases in comparisons carried out across protein families. High sequence homology is expected for closely related proteins. However, sequence homology, or conservation, across protein families, even if a weak one, may suggest a biological role for a region, or residues. In protein folding, it has been proposed that the transition state conformations of fast-folding proteins share a small subset of common folding nucleus.[36] The residues forming the nucleus are highly conserved within families, with a few residues apparently sufficient to yield common folding patterns. Cytochrome *c* has been shown as one such example. The apparent conservation of 4 non-heme binding residues in highly diverged cytochromes *c* molecules has been interpreted as critically important for the protein folding nucleus of all subfamilies of type *c* cytochromes.[27] Examination of these residues illustrates that the most conserved ones are aromatic, apart from aliphatic hydrophobic residues. A similar situation has been observed for the globins.[37]

Interface residues are mainly conserved owing to their roles in contributing to binding affinities. However, tight binding between proteins may not be the sole determinant in protein association. Considerations of the energy landscapes of protein binding and folding suggest that conserved residues may have thermodynamic as well as kinetic roles. As in the proposed conservation of a folding nucleus, surface residues may be conserved to facilitate the binding process. To bind at the enzyme active site, a ligand needs to diffuse or be transported to, and across, the enzyme surface. It has been suggested that electrostatics may provide a steering force for the diffusion process. Comparisons of both the results of Brownian dynamics simulation and of an analysis of the electrostatic potentials for triosephosphate isomerases, superoxide dismutases, and β-lactamases from different species, Wade et al.[38] have identified the conserved features responsible for the electrostatic substrate-steering field. The conserved potentials are localized at the active sites and are the primary determinants of bimolecular association rates. Even in the absence of long-range electrostatic steering, conserved interface residues might affect energy gradients, or funnels, near the binding site. Such energy funnels can increase the association rate.[39] On the other hand, the hydrophobic rings around the polar residues may contribute to more mobile, easily displaced water molecules.[40] Rings around the polar residues may also have large entropic contributions. Recently, two important experimental results regarding entropic contributions to protein binding have been published.[41–43] In the calmodulin-peptide case, Lee et al. have observed clustered responses upon complex formation.[41] They have suggested that the cluster of increased rigidity (and hence decreasing in entropy) may indicate localization of significant binding enthalpy. Interestingly, Zidek et al.[42] have shown increased protein backbone conformational entropy upon hydrophobic ligand binding. Thus, polar residues have favorable enthalpic contributions, but at the entropic cost of rigidification.[41] Nevertheless, the hydrophobic rings around these polar residues may compensate the entropic terms by increasing both side-chain and backbone motions.[42]

A particularly exciting recent example of the non-energetic role of conserved residues is the RNA-binding proteins. The highly conserved, solvent-exposed residues (Tyr13, Gln54, and Phe56) in RNA-binding domain (RBD1) of human U1A protein were found to communicate through a network of interactions on the surface of a β-sheet.[28] There are cooperative interactions between the three residues in the free protein, indicated by thermodynamics and ^{15}N-backbone dynamics. However, the energy that is associated with these interactions is small. The presence of the interactions in the free protein is functionally significant, allowing them to respond to RNA binding through changes in the interaction energy among the three positions.[28] Furthermore, it is interesting that the residues conserved in the RNA-binding proteins are those observed to be conserved in our analysis of the families of interfaces. Whereas Tyr is not at the top of our residue conservation list (Table VII), both Phe and Gln have relatively high conservation, even though these two residues are not significant according to the alanine-scanning mutational analysis.

CONCLUSION

The current clustered dataset of protein–protein interfaces comprises 1,629 interfaces in 351 families. Each of the families was examined, and all sequences were aligned. Eleven families containing a total of 97 crystal structures passed our criteria, both with respect to the number of family members and the extent of homology within the families. To obtain the residue conservation in preserved 3D environments, we superimposed each of the members on its respective family representative. Our goal has been to detect binding epitopes shared by members of the same family. However, in particular, we sought to probe the potential existence of common principles, which hold in general for protein–protein binding sites. Several observations have been made in the literature. First, binding sites are generally hydrophobic, indicating that the hydrophobic effect plays a major role in protein–protein interactions. Second, the extent of the hydrophobicity is variable, with some interfaces considerably more hydrophobic that others. Consistently, hydrophilic side chains play a more important role in binding than in folding. Third, for some interfaces, it has been shown that electrostatics plays a major role, steering the ligand onto the binding site of the receptor. Fourth, the tendency of specific residues to occur at the binding site has been studied through systematic alanine mutagenesis. Although the number of substitutions was large, they were derived from a limited number of interfaces. Yet, the availability of a large number of

clustered interfaces, combined with the Geometric Hashing structural comparison tool, which enables superimposing the interfaces despite the fact that they are composed of discontinuous pieces of the chains, provide powerful tools. These allow us to re-examine this question on a larger set of data.

Our results confirm and extend the alanine-scanning mutagenesis. The interface families that we have examined show a preference for conservation of polar residues at their interfaces. In particular, conserved interface residues are strongly correlated with the experimentally identified "hot spots," compiled from the database of experimental alanine-scanning mutagenesis. Specifically, the more highly conserved residues in our analysis are Arg, Gln, His, Asp, Pro, and Asn, whereas Bogan and Thorn find a special enrichment of Trp, Arg, Tyr, Asp, Pro, and His. Bogan and Thorn[8] have argued that energetic hot spots are critically important for the affinity of a protein interface. The fact that these residues tend to be conserved at specific locations indicates that they may constitute binding epitopes (functional epitopes). As such, they may be utilized in searches for potential unknown binding sites. They may further be engineered in binding site design. Conserved interface residues may play a dual role in binding, both thermodynamic and kinetic.

Additionally, the surface alignment identified conserved residues in the 11 clustered protein–protein interfaces. In enzyme inhibitors, we find that the residues are more conserved at the interfaces than in other locations in the proteins, as shown by multiple sequence alignments. In contrast, antibody–protein antigen interfaces illustrate similar surface conservation as compared to linear sequence alignment, consistent with the proposition that evolution has optimized protein interfaces to achieve optimal function, because conserved residues are more likely to be optimized.

ACKNOWLEDGMENTS

We thank Drs. Sandeep Kumar, Chung-Jung Tsai, and in particular, Jacob V. Maizel for numerous helpful discussions. The research of R. Nussinov and H. Wolfson in Israel has been supported in part by grant number 95-00208 from the Binational Science Foundation, Israel, by a grant from the Israel Science Foundation, by the Magnet grant, by the Ministry of Science grant, by Tel Aviv University Basic Research grants, and by the Center of Excellence, administered by the Israel Academy of Sciences. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

REFERENCES

1. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164.
2. Janin J. Elusive affinities. Proteins 1995;21:30–39.
3. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
4. Tsai CJ, Lin SL, Wolfson H, Nussinov R. A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 1996;260:604–620.
5. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;6:53–64.
6. Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels and protein function. Protein Sci 1999;8:1181–1190.
7. Stites WE. Protein–protein interactions: interface structure, binding thermodynamics, and mutational analysis. Chem Rev 1997;97:1233–1250.
8. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
9. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. Protein Eng 1999;12:713–720.
10. Tsai CJ, Xu D, Nussinov R. Structural motifs at protein–protein interfaces: protein cores vs two-state and three-state model complexes. Protein Sci 1997;6:1793–1805.
11. Dill KA. Dominant forces in protein folding. Biochemistry 1990;31:7134–7155.
12. Norel R, Petrey D, Wolfson H, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999;36:307–317.
13. Cherfils J, Duquerroy S, Janin J. Protein–protein interaction analyzed by docking simulations. Proteins 1991;11:271–280.
14. Cherfils J, Janin J. Protein docking algorithms: simulating molecular recognition. Curr Opin Struct Biol 1993;3:265–269.
15. Tsai CJ, Nussinov R. Hydrophobic folding units at protein–protein interfaces: implications to protein folding and protein–protein association. Protein Sci 1997;6:1426–1437.
16. Janin J, Rodier F. Protein–protein interaction at crystal contacts. Proteins 1995;23:580–587.
17. Lo Conte LL, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
18. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Empirical, solvent mediated potentials hold for both intramolecular and inter-molecular inter-residue interactions. Protein Sci 1998;7:2578–2586.
19. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein–protein recognition. Protein Sci 1994;3:717–729.
20. Vakser IA, Aflalo C. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins 1994;20:320–329.
21. Xu D, Lin SL, Nussinov R. Protein binding versus protein folding: the role of hydrophilic bridges in protein association. J Mol Biol 1997;265:68–84.
22. Clackson T, Wells JA. A hot spot of binding energy in a hormone–receptor interfaces. Science 1995;267:383–386.
23. Wells JA. Systematic mutational analyses of protein–protein interfaces. Methods Enzymol 1991;202:390–411.
24. Massova I, Kollman PA. Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies. J Am Chem Soc 1999;121:8133–8143.
25. Clackson T, Ultsch MH, Wells JA, de Vos AM. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. J Mol Biol 1998;277:1111–1128.
26. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein databank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
27. Ptitsyn OB. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes. J Mol Biol 1998;278:655–666.
28. Kranz JK, Hall KB. RNA recognition by human U1A protein is mediated by a network of local cooperative interactions that create the optimal binding surface. J Mol Biol 1999;285:215–231.
29. Nussinov R, Wolfson HJ. Efficient detection of motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci USA 1991;88:10495–10499.
30. Bachar O, Fischer D, Nussinov R, Wolfson H. A computer vision based technique for 3-D sequence independent structural comparison of proteins. Protein Eng 1993;6:279–288.
31. Fischer D, Lin SL, Wolfson HJ, Nussinov R. 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 1994;3:769–778.
32. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 1991;9:37–55.
33. Reference deleted in proofs.
34. Tormo J, Natarajan K, Margulies D, Mariuzza RA. Crystal structure of a lectin-like natural killer cell receptor bound to MNC class I ligand. Nature 1999;402:623–631.
35. Wang JH, Smolyar A, Tan K, Liu JH, Kim M, Sun ZJ, Wagner G, Reinherz EL. Structure of a heterophilic adhesion complex between the human CD2 and CD58(LFA-3) counterreceptors. Cell 1999;97:791–803.
36. Mirny LA, Abkevich VI, Shakhnovich EI. How evolution makes proteins fold quickly. Proc Natl Acad Sci USA 1998;95:4976–4981.
37. Ptitsyn OB, Ting KL. Non-functional conserved residues in globins and their possible role as a folding nucleus. J Mol Biol 1999;291:671–682.
38. Wade RC, Gabdoulline RR, Lüdemann SK, Lounnas V. Electrostatic steering and ionic tethering in enzyme-ligand binding: Insights from simulation. Proc Natl Acad Sci USA 1998;95:5942–5949.
39. Zhang C, Chen J, DeLisi C. Protein–protein recognition: exploring the energy funnels near the binding sites. Proteins 1999;34:255–267.
40. Ringe D. What makes a binding site a binding site? Curr Opin Struct Biol 1995;5:825–829.
41. Lee AL, Kinnear SA, Wand AJ. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. Nature Struct Biol 2000;7:72–77.
42. Zidek L, Novotny MV, Stone MJ. Increased protein backbone conformational entropy upon hydrophobic ligand binding. Nature Struct Biol 1999;6:1118–1121.
43. Forman-Kay JD. The 'dynamics' in the thermodynamics of binding. Nature Struct Biol 1999;6:1086–1087.

APPENDIX A. List of 11 Protein Interface Families Used in the Present Study

No.	Family representative	Members in the family					Protein family
1	1babAD	1hdsAD	1bbbBC	1bbbAD	1hdsBC		Hemoglobin
2	1bbbAB	1hdsCD	1hdaCD	1pbxAB	2mhbAB	2dhbAB	Hemoglobin
3	1choEI	3sgbEI	1ppeEI	4tpiZI	1brcEI	1ppfEI	Serine proteinase inhibitor
		1tabEI	1tgsZI	1mctAI	2kaiBI		
4	1cseEI	1sibEI	1meeAI	1sbnEI	2sniEI	5sicEI	Serine proteinase inhibitor
5	1hhjAB	2mhaAB	1hocAB	1hsaAB	2vaaAB	1hsaDE	Histocompatibility antigen
6	1hilCD	2cgrLH	2fb4LH	1bafLH	1bbdLH	1mfbLH	Immunoglobulin
		1dbjLH	1figLH	1gigLH	1ifhLH	2mcpLH	
		1mamLH	2fbjLH	2fgwLH	1bbjLH	1cbvLH	
		1cgsLH	1dfbLH	1faiLH	1fvdAB	1ggcLH	
		1ggiLH	1igfMJ	1igiLH	1indLH	7fabLH	
		1ncbLH	1tetLH	2iffLH	3hfmLH		
7	1hviAB	2rspAB	1hvpAB	2mipAB	1sivAB		Protease complexes
8	1vfaAB	1fvcAB	1reiAB	1aagLH	1fgvLH	1fvbLH	Immunoglobulin
		1hfmLH	1igmLH	1jhlLH	1migLH	2fvwLH	
9	2gstAB	1glqAB	1guhAB	1gssAB	1hncAB		Glutathione transferase
10	2plv12	1bbt12	1tme12	1r0912	1mec12	1rla12	Picornavirus and rhinovirus
11	3insAB	1izaAB	1trzAB	6rlxAB	1hitAB	6rlxCD	Hormone
		1izaCD					