doi:10.1016/j.jmb.2004.10.077 *J. Mol. Biol.* (2005) **345**, 1281–1294

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues

Ozlem Keskin[1,2]*, Buyong Ma[2] and Ruth Nussinov[2,3]*

[1]*Koc University, Center for Computational Biology and Bioinformatics, and College of Engineering, Rumelifeneri Yolu 34450 Sariyer Istanbul, Turkey*

[2]*Basic Research Program SAIC-Frederick, Inc., Laboratory of Experimental and Computational Biology NCI-Frederick, Frederick MD 21702, USA*

[3]*Sackler Inst. of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University Tel Aviv 69978, Israel*

**Corresponding authors*

Structurally conserved residues at protein–protein interfaces correlate with the experimental alanine-scanning hot spots. Here, we investigate the organization of these conserved, computational hot spots and their contribution to the stability of protein associations. We find that computational hot spots are not homogeneously distributed along the protein interfaces; rather they are clustered within locally tightly packed regions. Within the dense clusters, they form a network of interactions and consequently their contributions to the stability of the complex are cooperative; however the contributions of independent clusters are additive. This suggests that the binding free energy is not a simple summation of the single hot spot residue contributions. As expected, around the hot spot residues we observe moderately conserved residues, further highlighting the crucial role of the conserved interactions in the local densely packed environment. The conserved occurrence of these organizations suggests that they are advantageous for protein–protein associations. Interestingly, the total number of hydrogen bonds and salt bridges contributed by hot spots is as expected. Thus, H-bond forming residues may use a "hot spot for water exclusion" mechanism. Since conserved residues are located within highly packed regions, water molecules are easily removed upon binding, strengthening electrostatic contributions of charge–charge interactions.

Hence, the picture that emerges is that protein–protein associations are optimized locally, with the clustered, networked, highly packed structurally conserved residues contributing dominantly and cooperatively to the stability of the complex. When addressing the crucial question of "what are the preferred ways of proteins to associate", these findings point toward a critical involvement of hot regions in protein–protein interactions.

Keywords: protein–protein interactions; hot spots; residue conservation; residue cooparativity; residue networks

Introduction

An ultimate goal in molecular and cellular biology is the ability to predict the preferred mode of protein associations.[1–6] Detecting and mapping protein interactions on a genomic scale would bring the field of structural and functional genomics a step closer to the design of effective drugs.[7–12] Similar protein structures can associate in different ways[13–15] and perhaps even more striking is the observation that proteins with globally different structures can associate in similar ways. Many biological processes are driven by the formation of protein–protein complexes.[16] Numerous studies have addressed protein–protein binding, yet the principles governing protein interactions are not fully understood. Although binding sites are mainly hydrophobic, planar, globular and protruding[17–19] and composed of relatively large surfaces with good shape and electrostatic complementarity,[20–22] no general patterns are observed. Binding sites are dynamic, facilitating binding to

Abbreviations used: PDB, Protein Data Bank; ASA, accessible surface area; CN, coordination number; rmsd, root mean square deviation.

E-mail addresses of the corresponding authors: ruthn@ncifcrf.gov; okeskin@ku.edu.tr

different proteins with diverse compositions and shapes.[13,23–25] On the other hand, solvent effects are crucial. Ringe[26] discusses the effect of water molecules often occupying enzyme binding sites rendering them flexible and polar.

Predicting binding sites based on physical and chemical characteristics is important for drug and protein design. A recent review by Salwinski & Eisenberg[7] summarized the high-throughput experimental approaches and protein interactions databases. These cover a broader range of functions than earlier and facilitate distinguishing between common features of all interaction types *versus* specific features in certain complexes such as enzymes/inhibitors, antibodies and membrane proteins.[27] On the other hand, studies of protein binding on a residue level might be very informative. For example, different characteristics dominate binding processes of antibody/antigen complexes *versus* enzymes/inhibitors. Different types of complexes were compared in terms of interface size and hydrophobicity,[28,29] and predictive algorithms were developed based on these.[30,31] Skolnick and coworkers introduced a threading-based prediction of protein–protein interactions on a genomic scale.[10] Chakrabarti & Janin[32] analyzed about 70 protein complexes for the sizes of binding regions. When the buried surface area is under 2000 Å^2, the recognition sites usually form a single patch, in contrast to larger multi-patch interfaces. Interfaces display distinctive amino acid compositions compared to the rest of the protein. Analyzing 153 protein–protein complexes, we found that residue-specific pairwise potentials are similar for protein cores and interfaces especially under solvent-mediated conditions.[33] Interactions between the proteins can be different based on the type of the complex. In serine protease–inhibitor complexes, the interactions are between backbone atoms, whereas antibody–antigen complexes are mainly stabilized through side-chains.[34]

Alanine scanning mutagenesis[35] allowed an exploration of the energetic contributions of individual side-chains in protein binding. Using this technique, Wells and his colleagues have discovered that single residues can contribute a large fraction of the binding free energy.[35,36] Their studies have indicated that the free energies are not uniformly distributed across the interfaces. Instead, there are certain critical residues contributing the most to the binding free energy. These are called hot spots.[35–37] Bogan & Thorn[37] compiled a dataset of alanine mutants for which the change in free energy of binding upon mutation to an alanine was measured, and analyzed the hot spots' anatomy. According to their definition (followed in this work), a hot spot is a residue that when mutated to alanine, gives rise to a distinct increase in the absolute binding energy of more than 2 kcal/mol. Hu *et al.*[38] and Ma *et al.*[39] analyzed 11 clustered families of interfaces, showing that although overall binding sites are hydrophobic, specific conserved polar residues at specific locations, serve as energy hot spots.

We have extracted and iteratively clustered all interfaces between two protein chains in dimers, trimers and higher complexes from the Protein Data Bank (PDB),[40] leading to 3799 clusters.[41] Further filtering decreased this number to 44 unique interface clusters in which all member complexes belong to a single family and have similar functions. The list of the 44 structurally non-redundant clusters is given in the Supplementary Data, Table A. They include 292 member interfaces. All members in a single cluster are structurally aligned with MultiProt, a multiple structure alignment software.[42] Analysis of residue identities in these structural alignments shows that some residues are conserved at certain positions. Conserved residues across all interface clusters may have specific functional roles (e.g. catalysis, recognition, binding). However, in addition these residues are likely to be important for interface stability.

We compared the structurally conserved interface residues and the experimental enrichment energies from alanine scanning data.[37,43] The very high correlation prompted us to investigate the organization and contribution of the computational hot spots. Our analysis indicates that the computational hot spots are not randomly spread along the protein–protein interfaces; rather, they tend to be clustered. The assemblies of hot spots are located within densely packed regions. Within an assembly, the tightly packed hot spots form networks of interactions. We have named these assembly regions hot regions. As expected, these regions further contain residues that are moderately conserved. An interface may contain a single, or a few hot regions. The total number of hydrogen bonds and salt bridges involving the hot spots is as expected. However, the highly packed nature of the hot spots within the hot region facilitates removal of water molecules upon binding, strengthening the contributions of charge–charge interactions. This mechanism is in agreement with the earlier prediction made by Bogan & Thorn[37] in their insightful O-ring proposition. Thus, overall, packing emerges as a dominant factor in binding, as in protein folding. Moreover, this tight, networked hot spot organization implies that the contributions of the hot spots to the stability of the protein–protein complex within a hot region is cooperative; however, the contributions of independent hot regions are additive. This binding site organization rationalizes how a given protein molecule may bind to different ligand partners and has implications for scoring functions in docking. It further implies that summation of the single residue hot spot contributions will overestimate the binding energetics.

Results and Discussion

A protein–protein interface consists of two

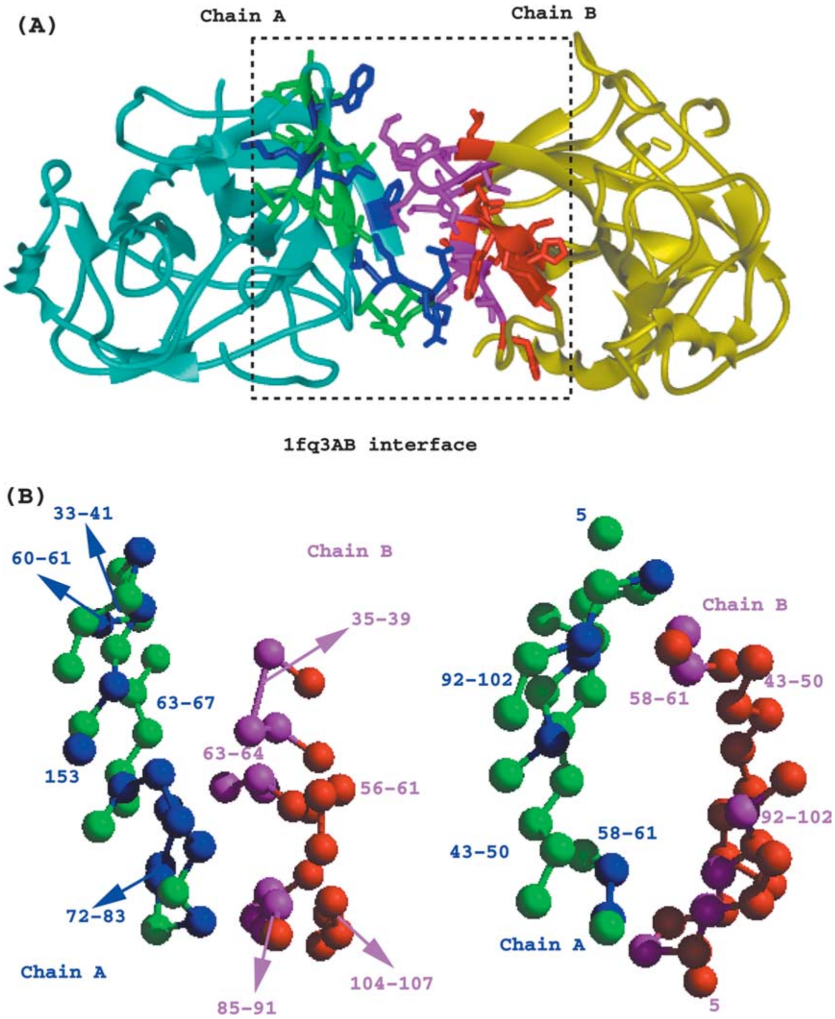

Figure 1. Protein–protein interfaces: why constructing a structurally non-redundant dataset is difficult. (A) An example of a two-chain interface between chains A and B (PDB code 1fq3). Residues which interact across the interface (called contacting residues) are in magenta and dark blue for the two chains. Residues in their spatial vicinity (called nearby residues) are in red and green. The remaining residues in the chains are in yellow and cyan for chains B and A, respectively. As can be seen, an interface consists of bits and pieces of each of the chains, and some isolated residues. There are nine contacting and 20 nearby residues in the chain A interface. The chain B side of the interface consists of 14 contacting and 14 nearby residues. Here, we use only the contacting residues.

polypeptide chains forming the two interface sides. Residues on both sides interact with each other. Two residues are defined to be contacting if the distance between any two atoms of the two residues from different chains is less than the sum of their corresponding van der Waals radii plus 0.5 Å.[41,44] On the other hand, a residue is defined to be "nearby" if the distance between its C$^{\alpha}$ atom and a C$^{\alpha}$ atom of any contacting residue is less than 6 Å. Nearby residues are important in structural clustering, since they provide the interface scaffold. Contacting residues are those responsible for the interactions across the interfaces. Here, the interface residues used in the statistical analyses of conserved residues, hydrogen bonds and the packing ratios are solely the contacting residues. An example of an interface is given in Figure 1(A), displaying both contacting and nearby residues in the interface between chains A and B of the human homodimer granzyme B. In Figure 1(B), only the C$^{\alpha}$ atom positions of the contacting and nearby residues are shown. The dataset we have used is described in Methods.

Structurally conserved and hot spot residues in the interfaces

All members of an interface cluster are superimposed structurally (using C$^{\alpha}$ atoms only). We calculate the number of matched residues among the superimposed interfaces and the number of identical or highly conserved residues across the matched residues. The conservation propensities of the interface residues to be conserved are calculated as detailed in Methods.

Figure 2 gives the correlation between experimentally determined amino acid enrichments from alanine scanning mutagenesis experiments[43] and our computed conservation propensities. On the x axis are the experimental enrichment values,[43] and the y axis has the computed values from this work. The experimental enrichments are calculated from the ASEdb† by dividing the number of a given residue type with $\Delta\Delta G \geq 2$ kcal/mol by the count of that amino acid in the whole database (in August 2003 the total number of residues with $\Delta\Delta G \geq 2$ kcal/mol was 199, and the total number

† http://140.247.111.161/hotspot/index.php

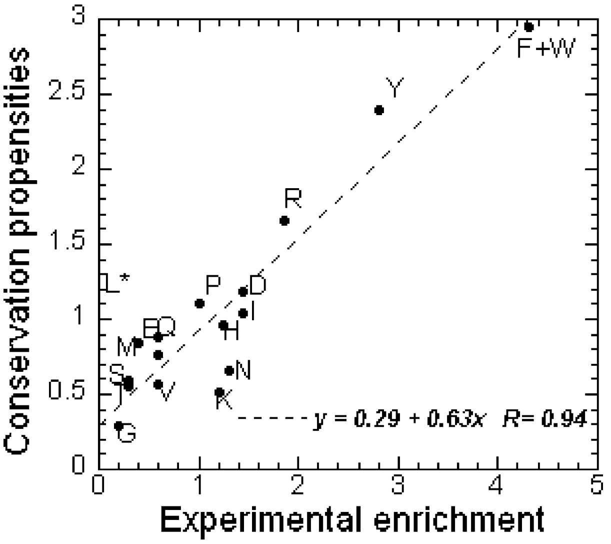

Figure 2. Correlation of residue conservation propensities obtained from 292 non-redundant interfaces *versus* the experimental enrichment of hot spots. L* relates to Leu, which is an outlier.

of all residues in the database was 2915).[43] We have adopted 2 kcal/mol as the cutoff value to define hot spots as suggested in the original study[37] and in our previous studies.[38,39] A higher cutoff would give the residues that are extremely important in binding; however, the number of such residues would be too low for our statistical analyses. The residue conservation propensities (in equation (1)) are calculated from the frequencies of occurrence of the residues compared with the rest of the chain. We define a residue to be structurally conserved if its conservation score threshold value is at least 0.5, which will provide a general picture for the dataset used here. These residues are "computational hot spots". We have further multiplied each residue propensity by its average side-chain accessible surface area (ASA) and normalized it by the average surface areas of all residues.[45] In Figure 2 we present the correlation for clusters containing proteins from the same functional family. Including all residues, the correlation coefficient with alanine scanning mutagenesis is 0.90. Here, we combine the contributions of Trp and Phe. With conservation thresholds of 0.4, 0.6 and 0.7, the correlations are 0.86, 0.91, and 0.92, respectively. If we exclude one outlier residue (Leu), the correlation coefficient increases from 0.90 to 0.94 with the 0.5 threshold. Thus, the threshold appears to have little effect on the conserved residues *versus* experimental enrichment, with only a slight improvement observed with higher thresholds.

We have also studied residue similarities. As in the substitution matrices,[46] amino acids are grouped by the chemistry of their side-chains. S, T, P, A, and G are in the small hydrophilic group. N, D, E, and Q are acid, acid amide and hydrophilic residues. H, R, and K are in the basic group. M, I, L, and V are in the small hydrophobic and F, Y, and W are the aromatics. If an amino acid is substituted by a similar amino acid, it is weighted with a factor of 0.4 as explained in Methods. Using residue similarity matrices, the correlations over all residues, excluding Leu, are 0.94 or 0.95 for all thresholds. These results suggest that using Blossom-like amino acid substitution matrices slightly improves the correlations with the experimental data. Thus, residue hot spots are of a similar but not necessarily of a specific type. We note that these statistics relate to single family interface clusters.

Hot spots are buried and tightly packed

Since hot spots contribute dominantly to protein–protein interactions, the question arises as to

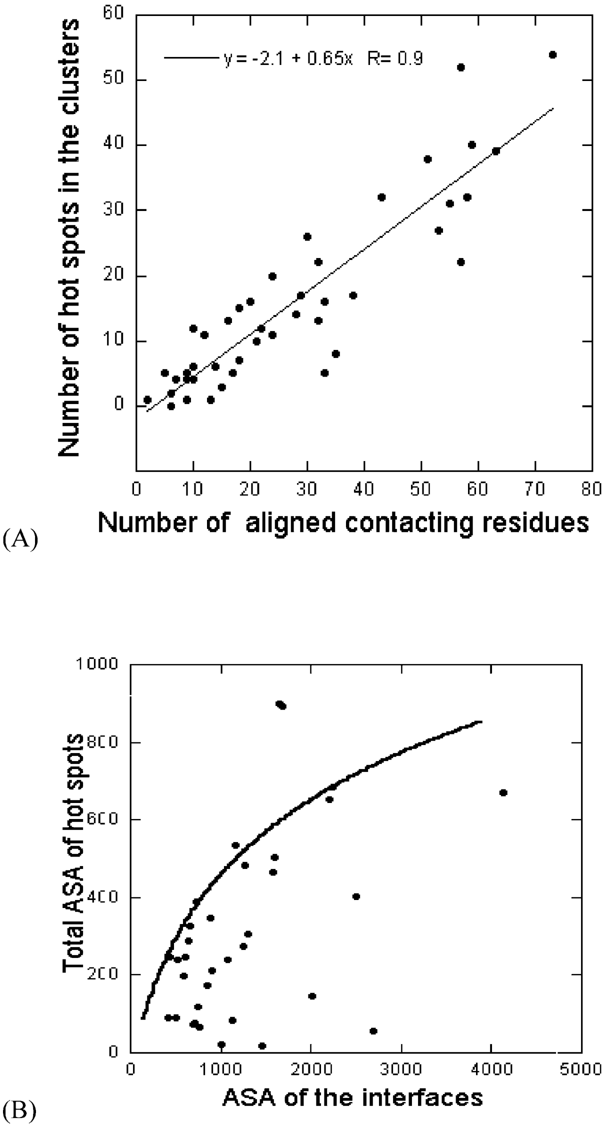

Figure 3. (A) Correlation between the number of contacting residues in the interfaces with the number of hot spots. (B) Correlation between the accessible surface areas (ASA) of interfaces with the accessible surface areas of the hot spot residues.

whether their number increases with the interface size, or is limited. Figure 3(A) indicates a correspondence between the number of hot spots and the number of contacting residues in the interfaces. There is a linear relationship between the number of hot spots and the interface size (with a correlation coefficient of 0.9). Thus, as the interfaces get bigger, the number of computational hot spots increases.

When we analyze the accessible surface areas rather than the number of hot spots, we observe more buried hot spots. Figure 3(B) provides the cumulative sum of the ASAs of hot spots of the interface cluster representatives *versus* the ASAs of the interfaces they belong to. The calculations are detailed in Methods. The hot spot ASAs are obtained from the complex form of the interfaces. A logarithmic equation leads to the best fit curve, although it still appears as if the tail has not reached a smoother plateau region. The graph suggests that the hot spots are more buried, since the ASA does not increase linearly with the interface for large interfaces. In the Figure, three of the complexes which appear in the flatter region are 1cov13 (coxsackie virus coat protein), 1hri14 (a rhinovirus coat protein) and 1as4AB (cleaved antichymotrypsin complex). It is noteworthy that these do not belong to obligate proteins. The mid region contains enzymes such as ligase, protease, and synthase. The lower part of the curve is mostly metal-transport (1ji5AC), enzyme inhibitor, enzyme/small peptide complexes and photosynthetic proteins. This suggests that there might be differences in the number of hot spots (and in the binding strength) depending on the biological function of the complexes.

Thus, the number of hot spots increases with increasing interface sizes, but not their total ASA. This suggests that hot spots are buried in protein pockets.[47] Yet, overall (Figure 3(A) and (B)) the number of contacting residues is almost linearly proportional to the ASAs of the interfaces (with a correlation coefficient of 0.76). Thus, interface packing and the extent of residue solvent exposure are not homogeneous across the interface. Similarly, Lo Conte *et al.*[27] find that the number of interface atoms scales linearly with the interface areas. We note that the number of the contacting residues in Figure 3(A) is the number of residues that are matched simultaneously among the members of the interface clusters by MultiProt. The actual interface could be larger than the number of the matched residues.

We further analyze residue packing densities around the conserved hot spots *versus* the rest of the interfaces. To study packing, we investigate the number of non-bonded neighbors (coordination number, *CN*) around each residue where residues are represented by their C^{α} atom positions. Figure 4 shows the histogram of the coordination numbers around the hot spots (Figure 4(B)) and the rest of the interface residues (Figure 4(A)). The average coordination number of the hot spots is 7.0, decreasing to 5.6 for the rest of the interface residues. Thus, packing around hot spots is significantly tighter than in the rest of the interface. Such a correlation reveals what differentiates hot spots from the rest of the interface. These high density motifs are reminiscent of the densely packed protein cores where the conserved nucleation residues reside.[48–51] Indeed, the *CN* of hot spots is very similar to *CN* of protein cores.[52] Such a similarity is suggestive, validating the proposition that binding and folding are similar processes.[53] It further explains why these residues are conserved and have larger energetic contributions.

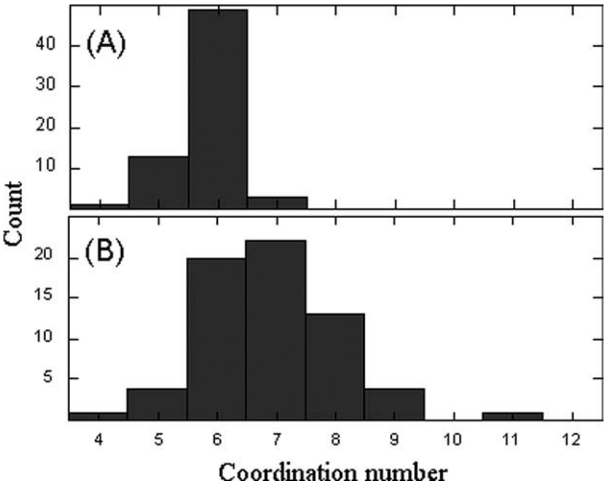

Figure 4. A histogram of the coordination number of the interface residues (A) for all interface residues, and (B) for the hot spots.

Furthermore, densely packed regions are less mobile, thus allowing proteins to associate with a smaller entropy penalty from their unbound form, indicating the importance of entropy in protein binding. On the other hand, the lesser packed regions may be critical for protein binding site flexibility. We also observe that protein–protein complexes bury more aromatic and bulky hydrophobic residues[39,41] than the rest of the proteins, although these residues are more difficult to place and orient at the interfaces.

Organization of hot spots: hot regions

We further analyze the organization of the hot spots in the interface. We find that hot spots are not homogeneously distributed in the interfaces. Rather, they are gathered locally in one or more regions, depending on the interface size. This results in densely packed clusters of networked hot spots, or hot regions. Bogan & Thorn[37] already realized in their study that experimental hot spots from alanine scanning mutagenesis data were clustered in the center of the interfaces rather than at the rims. Our study further indicates that depending on the size of the interface, there might be more than one hot spot cluster. Figure 5(A) is a

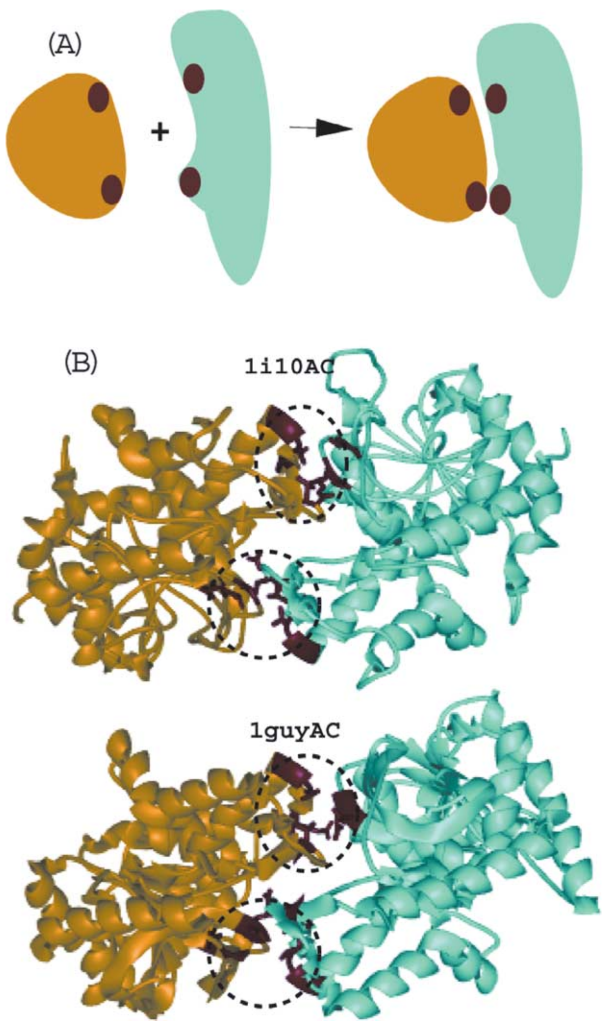

Figure 5. The organization of the hot spots in the interfaces. (A) The top panel provides a schematic representation of two proteins before and after complex formation. The red regions are the hot spot clusters. The Figure illustrates that the hot spot clusters interact with each other across the two-chain interface. (B) The lower panel depicts ribbon diagrams of the two members of oxidoreductase family. The red residues are the hot ones. The hot spot residues are not randomly distributed along the interfaces; rather they are clustered within densely packed regions. Here, there are two hot spot clusters (hot regions) among these members, circled in black.

schematic illustration of two monomers forming a complex. The red regions are the hot spot clusters. Hot spot clusters in one monomer face the clusters on the second monomer. This is in agreement with Halperin *et al*., who observed that hot spots tend to couple across binding interfaces.[54] Figure 5(B) is an example showing the hot spot organization in one of the interface clusters. The upper protein in the Figure is the representative of the cluster muscle L-lactate dehydrogenase (1i10AC), and the lower protein is a member in the cluster, malate dehydrogenase (1guyAC). All proteins in this interface cluster are oxidoreductases. The common interface consists of 64 residues and there are 14 hot spots. Here, these hot spots are grouped into two clusters, illustrating that hot spots are unevenly distributed along the interface.

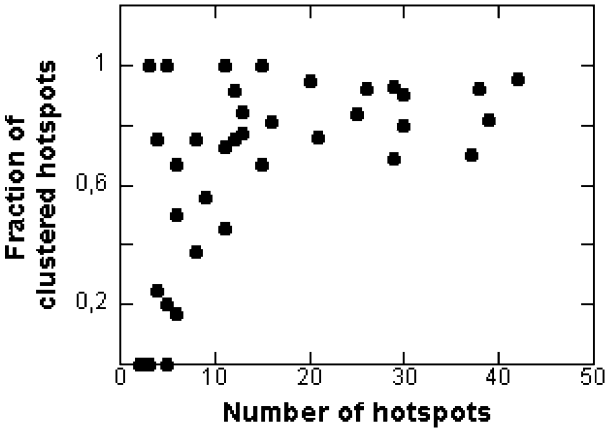

Figure 6. Fraction of hot spots involved in hot regions. The fraction is the number of hot spots inside the hot regions divided by the total number of hot spots. The x-axis is the number of hot spots in different interfaces, and the y-axis provides the fraction of the hot spots in the hot regions for each of the individual interfaces.

We have carried out a systematic analysis over the 44 interface clusters with 568 hot spots to study their organization in hot regions. We have calculated the number of hot spots that are in the hot regions and the number of the outliers (see Methods). 79% (449 out of 568) of the hot spots were found to be in the hot regions. Figure 6 displays the fraction of hot spots in hot regions. The x-axis is the number of hot spots in different interfaces, and the y-axis displays the fraction of the hot spots in the hot regions for each of the individual interfaces. Since our definition indicates that a hot region should have at least three hot spots (Methods), interfaces with two hot spots failed to be clustered. As expected, the Figure suggests that if the number of hot spots is higher, they are more likely to be clustered. The fact that the plot indicates reaching a plateau at seven residues probably reflects the number of interacting neighbors that the hot spot may have. As shown above (Figure 4) the average coordination number (*CN*) of the hot spots is 7.0. Thus, hot spots interact with other hot spots, leading to an interaction network within the clusters. As discussed below, moderately conserved residues further contribute to these interaction networks.

Hot spots are surrounded by moderately conserved residues

Since a hot spot has a significant energetic contribution to protein associations, the residue identity, size and charge, and the interactions it establishes with its neighboring residues should be crucial. Thus, although the residues that interact with hot spots are not always hot, nevertheless they may be expected to have high conservation ratios. We analyze the conservation of residues around all hot spots in our top level 44 single-family interface clusters. A residue is identified as a "hot spot-neighbor" if the distance between its C^α and a C^α of a residue is under 6.5 Å. Figure 7 displays schematically the hot spots (red dots) and their neighbors (the dots within the broken circles). Yellow and blue dots represent the moderately conserved and non-conserved residues, respectively. The average conservation ratio of these neighboring residues is 0.47 (within the circles) as compared to a 0.26 average conservation ratio for the rest of the interface residues (outside the circles). In most cases, there is one of two reasons for mutations of neighboring residues: (1) two companion neighboring residues are mutated in accord with each other, i.e. they are coupled. In one such cluster example, one of the members has a Gln that interacts with an Arg. In another member, these two residues are observed to be Glu and Lys. (2) There is a specific ligand in the interface that closely interacts with the neighboring residue. For example, Glu and Asp occur in one cluster member, which interacts with lactoylglutathione and Gln and Asn in another member, which interacts with epoxy.

Figure 8 displays the distribution of the hot spots in eight representative interfaces. The Figure presents examples of the projections of hot regions in the interfaces. Since interfaces are three-dimensional, it is not straightforward to present the organization of the hot spots. Here, we compute the best fit plane through the three-dimensional coordinates of C^α atoms forming the interface using principal component analysis. The coordinates of all C^α atoms are projected onto the plane formed by the first and second principal axes. In the Figure, the x-axis is the first principal axis. The highest deviations of the atomic coordinates are along this axis. Thus, it aligns along the longest dimension of the interface, crudely giving the length of the interface. The second principal axis represents the width of the interface. The blue and red dots in the Figure are the hot spots in the two complementary chains of the interfaces, red from one chain and blue from the other. Similarly, yellow and green are the moderately conserved residues (with conservation ratios higher than 0.35) from the two chains, respectively. Gray is the projection of all interface residues. With these thresholds, the results show that interfaces have one to three hot regions. Most of the hot spots are surrounded also by moderately conserved residues, further suggesting the conservation of interactions in these regions. In some cases (one case is shown in Figure 8(D)), outlier hot spots are surrounded by moderately conserved residues, so these may also form clusters of conserved regions. Only six of the 568 hot spots are totally isolated. In some of the interfaces, hot spots cover almost 70% of the interfaces (seven cases out of 40; four interface clusters have less than three hot spots, so they are not considered in the calculations). In these cases, either the conservation threshold should be higher than 0.50 or these may reflect the presence of a single hot region in the interface. Figure 9 presents three-dimensional illustrations of several complexes. Some portions of the complexes are not displayed in order to show more clearly the interface residues. One chain of the complex is colored green and the other yellow. The interface atoms are shown as hard spheres with their corresponding vdW radii. The hot spots from the two chains are colored red and cyan, respectively. The hot regions are further highlighted by broken black circles. Two of the proteins in Figures 8 and 9 are the same: a transferase (1axdAB) and a lectin (1qmoAB). These are chosen to illustrate the conversion from 3D spatial organizations (Figure 9) into 2D projections (Figure 8). Three blue hot spots present in Figure 8(H) of lectin are not visible in the 3D representation because they are positioned at the back side of the interface shown in Figure 9.

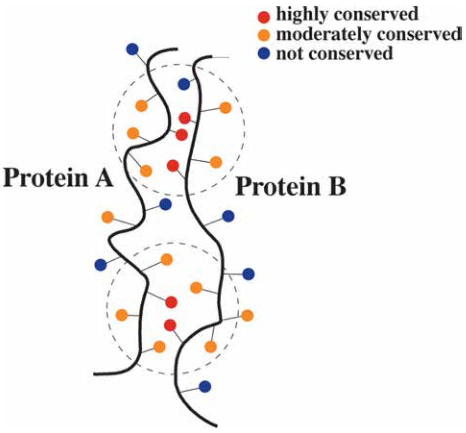

Figure 7. Interface packing and residue conservation: a hypothetical example. The red dots correspond to the computational hot spots (1.0 > conservation ration > 0.50), yellow dots indicate residues moderately conserved (0.50 > conservation ration > 0.35), and blue stands for the non-conserved residues. There are mostly yellow (rather than blue) dots around red dots. The dotted circles indicate the first coordination shell of the hot spots, i.e. all the residues within a radius of 6.5 Å. The average conservation ratio within the shells is 0.47 whereas it decreases to 0.26 outside the shell. We also observe that the packing is higher around hot spots (on average 7.0 residues in the shell) and lower at the other regions (5.6 residues outside the shell). For the packing calculations, residues whose C^α are closer than the cutoff distance are defined to be in contact, excluding the two bonded sequential neighbors. This Figure illustrates the hot regions.

Thus, in the complex structures our hot spots tend to form hot regions. In order to probe the organization of these residues in the unbound state, we focus on two complexes for which both conformations exist: carbonyl reductases (PDB codes for bound: 1cyd, unbound: 1edo) and glutathione S-transferases (PDB codes for bound: 1axd, unbound: 1aw9). For the transferases, the root mean square deviation (rmsd) between the crystal

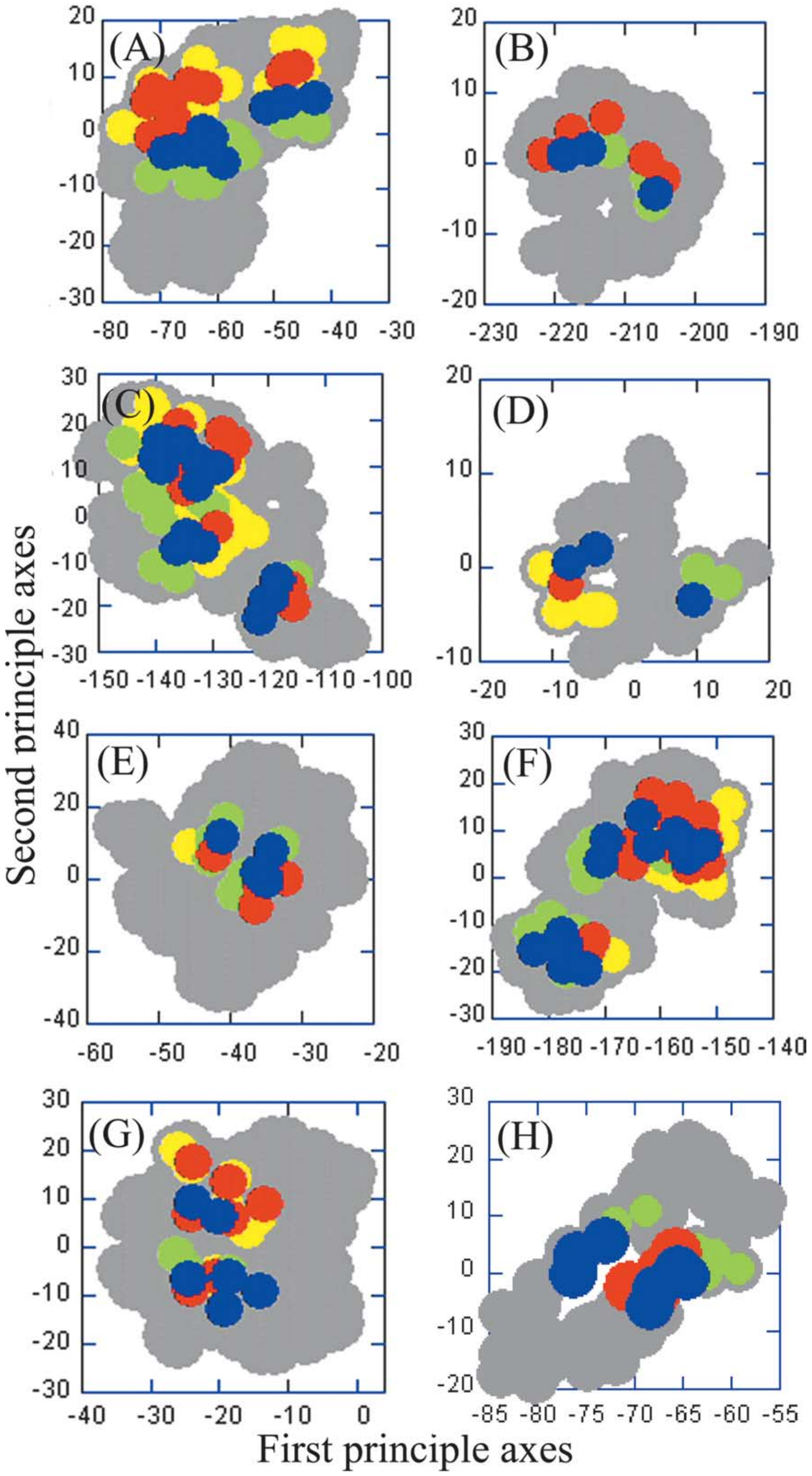

Figure 8. Examples of the hot spot clusters (hot regions) in the interfaces. The x and y axes are the first and second principal axes, respectively, for each of the representative interfaces. These two axes represent the best-fit plane of the interfaces. The coordinates of the C^{α} atoms are then projected onto the best fit plane. The blue and red dots are the hot spots in the complementary chains of the interfaces, red from one chain and blue from the other. Yellow and green are the moderately conserved residues (with conservation ratios higher than 0.35) from the two chains. Gray is for all of the interface residues. The representative names displayed in the Figure are as follows: (A) synthase (1rvv12); (B) hydrolase/activator (1fntHV); (C) coat protein (1al223); (D) adaptor protein/peptide (1azeAB); (E) oxidoreductase (1cydAB); (F) protease (1pmaAC); (G) transferase (1axdAB); and (H) lectin (1qmoAB).

structures is 5.61 Å and this rmsd decreases to 1.21 Å among the nine interface hot residues. For the reductases, the rmsd between 1cyd_A and 1aw9 is 8.77 Å and the rmsd between the two structures for the six hot spots is 2.29 Å. The hot residues in the unbound form are obtained by performing structure and sequence alignments between the bound and unbound structures. Since the rmsd is measured for fewer residues, a lower rmsd is expected. Thus, we perform the principal component analysis as previously. Figure 10 displays these two examples of the hot spot clusters (hot regions) in the unbound forms. The Figure shows the projections of the hot regions in the putative

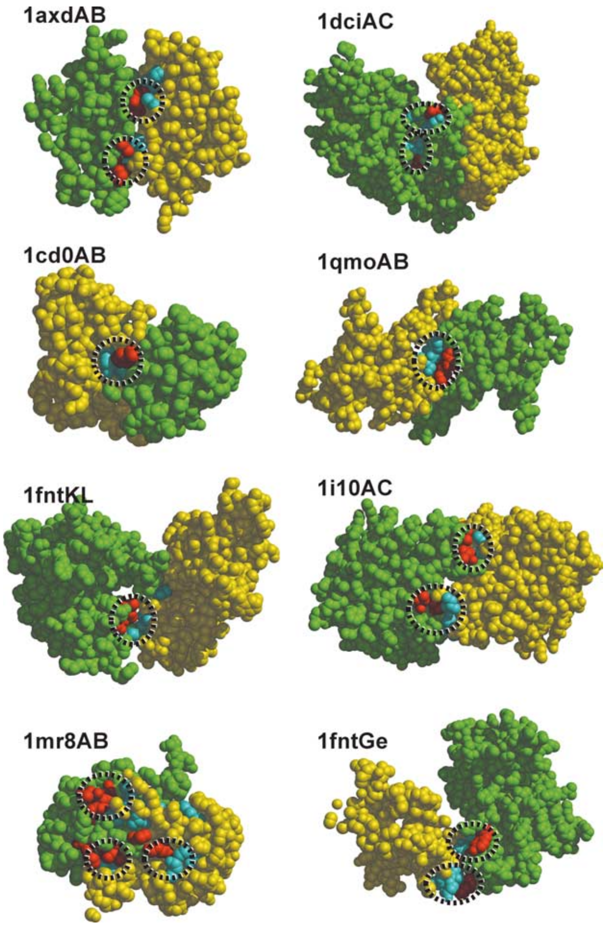

Figure 9. Examples of interfaces with hot regions. One chain of each complex is colored green and the other yellow. The interface residues are displayed as balls with their corresponding vdW radii. The hot spots from two chains are colored red and cyan in the Figure. The hot regions are further shown with broken black circles. For clarity, some interface portions of the proteins are not displayed. The PDB codes of the proteins shown in the Figure together with their chain IDs: 1axdAB (transferase), 1dciAC (lyase), 1cd0AB (immunoglobulin), 1qmoAB (lectin), 1fntKL (hydrolase), 1i10AC (oxidoreductase), 1mr8AB (metal transport), 1fntGe (hydrolase/activator).

binding region on the surface. The red dots in the Figure are the hot spots. Yellow are the moderately conserved residues (with conservation ratios higher than 0.35) of the unbound protein. Gray is the projection of all interface residues. With these thresholds, the results show that interfaces have one to three hot regions. These Figures resemble Figure 8. Similar to the bound form, here the hot spots appear to be pre-organized in clusters in hot regions in the unbound forms.

Hydrogen bonding and salt bridges *versus* hot spots

We investigate the involvement of hot spots in hydrogen bonding across the interfaces. Two atoms from each side of the interface are defined to form a hydrogen bond if the distance between the donors and acceptors is less than 3.5 Å. We further carry out the same type of analysis for salt bridges across interfaces. When a positively charged atom and a negatively charged atom from each of the two chains are closer than 4.5 Å, they are said to form a salt bridge. The listing of positively and negatively charged residues and H-bond donor and acceptors[55] is provided in Methods.

Table 1 summarizes the hydrogen bonds across the interface originating from the hot spots. The first column gives the H-bond interaction type. SS indicates H-bonds between side-chain–side-chain groups; BB is for hydrogen bonds between backbone–backbone groups, and SB for hydrogen bonds formed between a side-chain group and a backbone group. The second column gives the total number of hot spots in the interfaces. The third is the number of total hydrogen bonds formed across the interface and the fourth column (match) is the number of hot spots involved in hydrogen bond formation. The hot spot ratio is the fraction of the match to the total number of hot spots, and the H-bond ratio is the fraction of H-bonds to the overall interface residues. The seventh column is the ratio of the fifth column to the sixth, which gives the enrichment of hot spots in hydrogen bond formation.

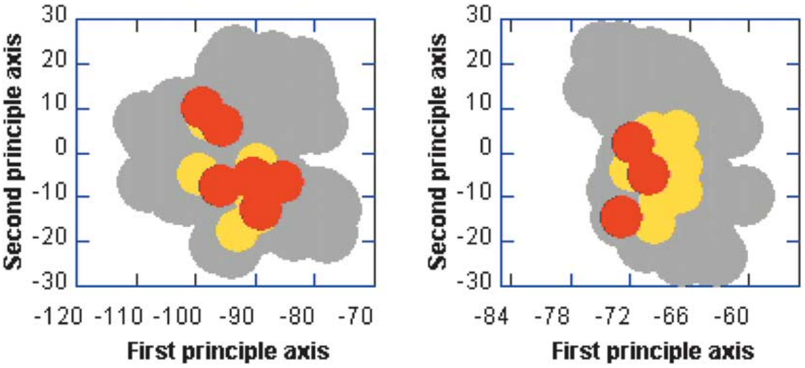

Figure 10. Examples of hot spot clusters (hot regions) in the unbound state. The Figures display the projections of the hot regions in the putative binding region on the surface. The x and y axes are the first and second principal axes, respectively, for each of the representative interfaces. These two axes represent the best-fit plane of the interfaces. The coordinates of the C^{α} atoms are then projected onto the best fit plane. The left-hand side is the carbonyl reductases (PDB code: 1edo). On the right-hand side glutathione *S*-transferases (PDB code: 1aw9). The red dots in the Figure are the hot spots. Yellow are the moderately conserved residues (with conservation ratios higher than 0.35) of the unbound protein. Gray is the projection of all interface residues. The bound structures for these proteins are in our clustered interface dataset.

Table 1. Hot spot and hydrogen bond distributions across interfaces (interface size = 1893)

Interaction[a]	Hot spots (I)[b]	Interaction (II)[c]	Match (III)[d]	Hotspot ratio (III/I)	Interaction ratio (II/1893)	Enriching Score[e]
H-Bond (SS)	582	267	91	0.16	0.14	1.14 (1.25)
H-Bond (BB)	582	192	63	0.11	0.10	1.10 (1.17)
H-Bond (SB)	582	220	64	0.11	0.12	0.92 (1.21)
H-Bond ALL	582	679	218	0.38	0.36	1.06 (1.21)
Electrostatic	582	100	31	0.05	0.05	1.00

[a] The H-bond interaction type, i.e. SS indicates the H-bonds between side-chain–side-chain groups, BB is for hydrogen bonds between backbone–backbone groups, and SB is for hydrogen bonds formed between a side-chain group and a backbone group.
[b] Total number of hot spots in the interfaces.
[c] Total hydrogen bonds formed across the interface.
[d] The number of hot spots involved in hydrogen bond formation.
[e] Ratio of the fifth column (Hotspot ratio) to the sixth (Interaction ratio), which gives the enrichment of hot spots in hydrogen bond formation.

Our results indicate similar contributions of backbone and side-chain atoms to H-bond formation (Table 1). This result reflects our database that includes both homodimers with high packing and shape complementarity and transient complexes whose overall packing is not as good. Polar residues are observed to contribute both through their side-chain and backbone atoms. As expected, side-chain atoms form H-bonds with other side-chain atoms and backbone atoms. Interestingly, interactions involving hot spot residues and other interface residues are similar. Hot spots are not favored to form H-bonds, and the enrichment score for all types of H-bonds is 1.06 based on residue count. This enrichment factor increases to 1.21 if we perform the statistics on atoms rather than on residues. The numbers in parentheses listed in the Table are the scores for atoms. We also observe that most non-polar residues contribute to H-bonding by their backbone carboxyl and amino groups.[55]

Electrostatic interactions are not very frequent in our database. It can be seen from Table 1 that hot spots show no preference to be involved in charged electrostatic interactions, and that there is a slight preference of non-polar residues in hotspots. Thus, in sheer numbers, electrostatic interactions are not favored in hot spots in protein–protein interfaces. This result is in agreement with the observation that coupling of conserved charged residues or of polar residues is unfavorable.[54] On the other hand, conserved Gly coupling with other residues such as aromatics and small hydrophobics is favored. Gly lacks a side-chain, and more easily packs against other residues. Since, however, electrostatic interactions at the interfaces (H-bonds and salt bridges) are buried within the highly packed hot regions the solvent is easily removed upon binding, suggesting why their contribution to protein–protein interactions can be crucial.

Conclusions

Recently, we have derived a new, large and diverse non-redundant dataset of protein–protein interfaces from the PDB, leading to 44 single-family clusters. Here, we first compare the conserved residues with hot spot ($\Delta\Delta G \geq 2$ kcal/mol) residues taken from the alanine scanning mutagenesis database. Experimentally, hot spot residues are major contributors to the stability of the protein–protein complex. We find that there is a correspondence between the experimentally identified energy hot spots and the structurally conserved residues. We focus on the spatial organization of the conserved hot spot residues within the interfaces in an effort to understand the origin of their stabilizing contributions to protein–protein associations. We observe that the hot spots are located within densely packed regions, explaining their conservation and their large energy contributions to the stability of the complex. Further, we find that the hot spots are not homogeneously distributed along the protein binding sites; rather, they are clustered. As expected, they are further surrounded by residues that are moderately conserved. Combined, these observations lead us to conclude that protein binding sites consist of one or a few densely packed hot regions. The hot spots are clustered within these, forming networks of conserved interactions. Thus, we propose that within a hot region, the contributions of the hot spots to the stability of the protein–protein association are cooperative. On the other hand, the contributions of independent hot regions are additive. The number of hot regions varies, and appears to depend on the interface size.

Moreover, we do not observe a higher than expected number of hydrogen bonds and charge–charge interactions involving the hot spots. Yet, electrostatic interactions and hydrogen bonds are well known to be crucial to the stability of protein–protein complexes.[55–58] The computational prediction of hot spots based on these terms has well demonstrated this fact.[59,60] This apparent contradiction may be rationalized. As already proposed by Bogan & Thorn in their landmark paper,[37] the densely packed hot regions make it easier to remove water molecules upon protein–protein binding, strengthening the charge–charge interactions.

Hence, consistent with the Bogan and Thorn proposition, our observations point toward the local organization of the hot spots as a critical factor in stabilizing protein–protein interactions.

Overall, we observe that packing, a well-known major contributor to the stability of globular proteins,[61–64] is also extremely important in binding. Thus, protein–protein interactions are optimized locally. The non-optimal packing along the interfaces between the hot regions, allows more freedom and tolerance for the rest of the interface. This might be one reason for the diversity of protein binding, where a given binding site can accommodate ligands with different sizes, shapes and composition.[23] This organization allows sharing similar motifs between binding regions of different protein pairs. We note that a cooperative behavior of hot spot residues within hot regions suggests modifications to docking scoring schemes, although the parameterization is expected to present major hurdles. We further suggest that similar principles of organization may exist in nucleic acid–protein interactions and in small molecule binding sites.

Methods

Dataset

We have applied the criteria for the definition of interfaces to all multi-chain PDB entries[40] in the database. On July 18, 2002 there were 18,687 entries in the PDB, which included 35,112 single chains.[41] PDB entries that contain more than two chains were used to obtain two-chain combinations. These included all two-chain interfaces from dimers, trimers and higher complexes. As a result, 21,686 two-chain interfaces were obtained. The interfaces were renamed as follows: if the PDB code of a protein is 1fq3 and it has two chains A and B, the interface is named 1fq3AB (see Figure 1), indicating that there is an interface between chains A and B of protein 1fq3. All interfaces were structurally compared by the Geometric Hashing sequence order-independent structural comparison algorithm.[44,65] We used a heuristic iterative clustering procedure that assigned an interface to a cluster if its similarity to the cluster representative was below predefined thresholds; otherwise it was assigned as a new cluster representative. Six clustering cycles were performed, gradually relaxing the thresholds. Overall, 3799 clusters were obtained. Sequences within each cluster were compared using CLUSTALW[66] and the BLOSSUM90 substitution matrix.[46] Any one of two entries in the same cluster sharing more than 50% similarity was eliminated. Clusters with less than five members were removed. The final set of clusters contains members as diverse as enzymes, antibodies, viral capsids, etc. Members of these clusters perform unique functions. They have similar chains and similar interfaces. Thus, the entire complexes are well aligned and within a cluster, all members belong to the same functional family. When constructing the dataset of interfaces, we also obtained clusters whose members belong to different functional families. However, in this study we neglected the second type of clusters, since here our purpose is to investigate the functionally important conserved hot residues.

The list of 3799 structurally non-redundant interface representatives together with members is available†.

Structural alignment of interfaces

MultiProt is fully automated software to simultaneously align multiple structures of proteins. It uses a residue-sequence order and directionality independent algorithm making it applicable to protein–protein interfaces.[42] Since interfaces are formed by two polypeptide chains, and most of the time the chains consist of discontinuous segments, it is advantageous to use an alignment algorithm that is capable of aligning residues belonging to discontinuous segments. The residues in the interfaces are children of two different chains physically; therefore an alignment algorithm should not ignore which chain a residue belongs to. We have used MultiProt to structurally align all non-redundant interface members in each cluster. The parameters used in the alignments are as follows: maximal RMSD for matching = 3.5 Å; minimal size of rigidly matched fragments = 3; maximal shift in indices of two matched fragments = 20; overlap ratio = 0.8; OnlyRefMol = 0; and FullSet = 1. The alignments are multiple, not pairwise. The ten best alignments are listed as the output of the program. In all cases, we have chosen the best scoring alignment, unless the second best alignment has residues from both chains whereas the best scoring alignment has residues from only one chain. MultiProt chooses one of the structures as the representative for the multiple alignment. This representative is the one most similar to all other members during the multiple alignment process. These representatives are not necessarily the same representatives from the previous hierarchical clustering.

The complete list of interfaces used in this study is given in the Supplementary Data, Table A. These proteins belong to 44 structurally different clusters. Each cluster contains at least five interface members. Each cluster is named by its representative, which is the interface that is structurally most similar to all other members. Cluster representatives are listed in the second column. The third column lists the number of simultaneously matched interface residues within each cluster. The last column lists the classification of the representatives.

Calculation of hot spots

The propensity of residue i to be conserved (P_i^*) in the interface is calculated by:

$$P_i^* = (n_i^*/N_i^*)/(n/N) \qquad (1)$$

where n_i^* is the number of conserved residues of type i at the interface, N_i^* is the number of residues of type i in the chains, n is the number of conserved residues at the interfaces, and N is the total number of residues in the chains. We have further multiplied each residue propensity by its average side-chain accessible surface area (ASA) and normalized it by the average surface areas of all residues according to $P_i^{\text{asa}} = P_i^*/(\text{ASA}_i/\Sigma\text{ASA}_i)$; where ASA_i is the accessible surface area of residue i, and ΣASA_i is the average ASA over all 20 residue types. The ASAs of the individual residues are as follows (in Å^2): G = 85, A = 113, C = 140, D = 151, E = 183, F = 218, H = 194,

† http://home.ku.edu.tr/~okeskin/INTERFACE/INTERFACES.html and http://protein3d.ncifcrf.gov/~keskino

I=182, K=211, L=180, M=204, N=158, P=143, Q=189, R=241, S=122, T=146, V=160, W=259, Y=229.[45]

MultiProt was used to align simultaneously the interface members within each cluster. The conservation of a residue is calculated by taking the ratio of residue i to be in a specific position in the structural alignment of the cluster members as:

$$\text{conservation ratio}_i = \frac{\sum_1^m \delta_i}{m} \tag{2}$$

m is the number of members in the aligned interface cluster. δ_i is 1 if residue i exists at the specific conserved position, and zero otherwise. If the ratio is higher than a threshold value, we define the residue as a "conserved residue", and count the number of conserved residues over the clusters. We have tested different thresholds for the conservation ratios, 0.4, 0.5, 0.6 and 0.7 and their effects. For example, in the Supplementary Data, Table B, a sequence alignment of nine interfaces (belonging to the transferases) is shown. The first row lists the names of the interfaces. The first column represents the positions along the interface alignment. Here we show a portion of the alignment with 30 residues. The bold and italic indicate that Leu (position 29) is an invariant residue at that position.

We have also studied residue similarities. As in the substitution matrices,[46] amino acids are grouped by their chemistry. S, T, P, A, G are small hydrophilics. N, D, E, and Q are acid, acid amide and hydrophilics. H, R, and K are in the basic group. M, I, L and V are small hydrophobics and F, Y and W are the aromatics. If an amino acid is substituted by a similar amino acid, it is weighted with a factor of 0.4 using the formula below:

$$\text{conservation ratio}_\text{i} = \frac{\sum_1^m \delta_\text{i} + 0.4\Delta_\text{i}}{m} \tag{3}$$

Here, Δ_i is 1 if at that position there is a similar residue, and zero otherwise. For example, for position 4, there are five Val residues (with a conservation score of 5/9=0.60) in the alignment. In this position, there are one Ile and one Leu (which are in the same classification). Each counts for 0.4 identities. Therefore, the total score of that position increases from 0.60 to 0.64 ((5.0+0.4+0.4)/9=0.64).

The hot spots here are obtained from structural alignments of interfaces. Thus, we need multiple structures for our analysis. This restricts our dataset of hot spots to a limited number of cases where the structures are available.

Accessible surface areas of the interfaces

The accessible surface areas (ASAs) were calculated using an implementation of the Lee and Richards algorithm,[67] with a probe sphere of radius 1.4 Å. The ASAs of the hot spots in each interface are calculated according to the formula:

$$\text{ASA}^{\text{hot}}_{\text{representative}} = \sum_{i=1}^{n\text{hot}} \text{ASA}^{\text{hot}}_i \tag{4}$$

where $n$hot is the number of hot spots in an interface cluster and $\text{ASA}^{\text{hot}}_i$ is the accessible surface area of the ith hot spot in the interface. The hot spot ASAs are obtained from the complex form of the interfaces.

Packing of the residues

Residue packing ratios are calculated by taking a single point (C^α) for each residue. We define a cutoff distance of 6.5 Å around each residue (the first coordination shell). Residues whose C^α atoms are closer than the cutoff are defined to be neighbors. Close neighbors along the chain (i.e. for the ith, the $i-1$ and $i+1$ residues) are not summed in the calculations. Thus a contact between ith and jth residues is defined using:

$$\begin{aligned} \text{Contact}_{i,j} &= 0 \qquad\qquad \text{if } |i-j| \le 1; \\ \text{Contact}_{i,j} &= 1 \quad \text{if } |i-j| > 1 \text{ and } d_{i,j} \le 6.5 \text{ A} \end{aligned} \tag{5}$$

where $d_{i,j}$ is the distance between two C^α atoms of the ith and jth residues.[52] The coordination number of a residue i (number of residues around it) is calculated by:

$$CN_i = \sum_{j=1}^{res} \text{Contact}_{ij} \tag{6}$$

where CN is the coordination number and res is the number of residues in the protein chains. Since interface residues have neighbors both from their own chain and the complementary chain, we count the neighbors from both chains.

Clustering of hot spots into hot regions

Interfaces with at least three hot spots were included in the analysis. A hot spot is assumed to be in a hot region if it has at least two hot spot neighbors. Each hot spot residue is assumed to be a perfect ball with a specific volume.[17] The C^α-atoms of the hot spot residues are the centers of these balls. The radii of the balls are extracted from their ball volumes. If the distance between the centers of two balls (two C^α-atoms of two hot spots) is less than the sum of the radii of the two balls plus a tolerance distance (2 Å), the two hot spot residues are flagged to be clustered and to form a network in a hot region. Hot spots from both chains are considered.

H-bond and electrostatic bond calculations

All backbone N atoms are assigned as H-donors. Lys N^ζ, Asn $N^{\delta2}$, Gln $N^{\varepsilon2}$, Arg N^ε, Arg $N^{\eta1}$, Arg $N^{\eta2}$, Trp $N^{\varepsilon1}$ are taken as donors. All backbone O atoms and Asp $O^{\delta1}$, Asp $O^{\delta2}$, Glu $O^{\varepsilon1}$, Glue $O^{\varepsilon2}$, His $N^{\delta1}$, His $N^{\varepsilon2}$, Met S^γ, Asn $O^{\gamma1}$, Gln $O^{\varepsilon1}$, Ser O^γ, Thr $O^{\gamma1}$, Tyr O^η, Cys S^γ are assigned as H-acceptors. His $N^{\delta1}$, His $N^{\varepsilon2}$, Ser O^γ, Thr $O^{\gamma1}$, Tyr O^η, Cys S^γ can act as both donor/acceptor.[54] If the distance between an acceptor and a donor is less than 3.5 Å across the interface, an H-bond is formed. We have counted the number of H-bonds in each interface representative, and calculated how many of these bonds are formed by a hot spot atom.

The positively charged atoms are taken as Lys $N^{\zeta1}$, Arg $N^{\eta1}$, Arg $N^{\eta2}$, His $N^{\varepsilon2}$. The negatively charged groups belong to Asp $O^{\delta1}$, Asp$O^{\delta2}$, Glu $O^{\varepsilon1}$, Glu $O^{\varepsilon2}$. If two oppositely charged atoms are closer than 4.5 Å across the interface, a salt bridge is formed between them.

Acknowledgements

We thank particular Dr K. Gunasekaran for his help in accessible surface area calculations, and M. Shatsky for MultiProt. We thank Drs C. -J. Tsai, Y. Pan, D. Zanuy, H.-H. (G). Tsai, members of the

Nussinov-Wolfson group and Dr A. Gursoy and S. Aytuna from Koc University for their helpful discussions. We thank Dr Jacob V. Maizel for encouragement. The research of R.N. in Israel has been supported in part by the Center of Excellence in Geometric Computing and its Applications funded by the Israel Science Foundation (administered by the Israel Academy of Sciences. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. 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 organizations imply endorsement by the U.S. Government.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2004.10.077

References

1. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S. *et al.* (2003). CAPRI: a critical assessment of predicted interactions. *Proteins: Struct. Funct. Genet.* **52**, 2–9.
2. Gallet, X., Charloteaux, B., Thomas, A. & Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* **302**, 917–926.
3. Ehrlich, L., Reczko, M., Bohr, H. & Wade, R. C. (1998). Prediction of protein hydration sites from sequence by modular neural networks. *Protein Eng.* **11**, 11–19.
4. Ofran, Y. & Rost, B. (2003). Predicted protein–protein interaction sites from local sequence information. *FEBS Letters*, **544**, 236–239.
5. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002). Prediction of protein–protein sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356–1361.
6. Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27.
7. Salwinski, L. & Eisenberg, D. (2003). Computational methods of analysis of protein–protein interactions. *Curr. Opin. Struct. Biol.* **13**, 377–382.
8. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
9. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, **405**, 823–826.
10. Lu, L., Arakaki, A. K., Lu, H. & Skolnick, J. (2003). Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.* **13**, 1146–1154.
11. Lu, H., Lu, L. & Skolnick, J. (2003). Development of unified statistical potentials describing protein–protein interactions. *Biophys. J.* **84**, 1895–1901.
12. Spirin, V. & Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.
13. DeLano, W. L., Ultsch, M. H., de Vos, A. M. & Wells, J. A. (2000). Convergent solution to binding at a protein–protein interface. *Science*, **287**, 1279–1283.
14. Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
15. Tondi, D., Slomczynska, U., Costi, M. P., Watterson, D. M., Ghelli, S. & Shoichet, B. K. (1999). Structure-based discovery and in-parallel optimization of novel competitive inhibitors of thymidylate synthase. *Chem. Biol.* **6**, 319–331.
16. Kleanthous, C., ed. (2000). *Protein–Protein Recognition, Frontiers in Molecular Biology*, Oxford University Press.
17. Chothia, C. & Janin, J. (1975). Principles of protein–protein recognition. *Nature*, **256**, 705–708.
18. Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.
19. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
20. Jones, S. & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
21. Janin, J. (1995). Principles of protein–protein recognition from structure to thermodynamics. *Biochimie*, **77**, 497–505.
22. Janin, J. (1997). Specific *versus* non-specific contacts in protein crystals. *Nature Struct. Biol.* **4**, 973–974.
23. Ma, B., Shatsky, M., Wolfson, H. J. & Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **11**, 184–197.
24. DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **12**, 14–20.
25. Kuhlmann, U. C., Pommer, A. J., Moore, G. R., James, R. & Kleanthous, C. (2000). Specificity in protein–protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. *J. Mol. Biol.* **301**, 1163–1178.
26. Ringe, D. (1995). What makes a binding site a binding site? *Curr. Opin. Struct. Biol.* **5**, 825–829.
27. LoConte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
28. Ofran, Y. & Rost, B. (2003). Analysing six types of protein–protein interfaces. *J. Mol. Biol.* **325**, 377–387.
29. Janin, J. & Chothia, C. (1990). The structure of protein–protein recognition sites. *J. Biol. Chem.* **256**, 16027–16030.
30. Korn, A. P. & Burnett, R. M. (1991). Distribution and complementarity of hyropathy in mulisubunit proteins. *Proteins: Struct. Funct. Genet.* **9**, 37–55.
31. Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein–protein recognition. *Protein Sci.* **3**, 717–729.
32. Chakrabarti, P. & Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins: Struct. Funct. Genet.* **47**, 334–343.
33. Keskin, O., Bahar, I., Badretdinov, A. Y., Ptitsyn, O. B. & Jernigan, R. L. (1998). Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci.* **7**, 2578–2586.
34. Jackson, R. M. (1999). Comparison of protein–protein

interactions in serine protease-inhibitor and antibody–antigen complexes: implications for the protein docking problem. *Protein Sci.* **8**, 603–613.

35. Cunningham, B. C. & Wells, J. A. (1991). Rational design of receptor-specific variants of human growth hormone. *Proc. Natl Acad. Sci. USA*, **88**, 3407–3411.
36. Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.
37. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.
38. Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins: Struct. Funct. Genet.* **39**, 331–342.
39. Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. (2003). Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.
40. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
41. Keskin, O., Tsai, C.-J., Wolfson, H. & Nussinov, R. (2004). A new, structurally non-redundant, diverse dataset of protein–protein interfaces and its applications. *Protein Sci.* **13**, 1043–1055.
42. Shatsky, M., Nussinov, R. & Wolfson, H. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins*, **344**, 143–156.
43. Thorn, K. S. & Bogan, A. A. (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.
44. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1996). A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.* **260**, 604–620.
45. Miller, S., Lesk, A. M., Janin, J. & Chothia, C. (1987). The accessible surface area and stability of oligomeric proteins. *Nature*, **328**, 834–836.
46. Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
47. Li, X., Keskin, O., Ma, B., Nussinov, R. & Liang, J. (2004). Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that are pre-organized in the unbound state. *J. Mol. Biol.*. In the press.
48. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129.
49. Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophys. J.* **81**, 751–766.
50. Liwo, A., Pincus, M. R., Wawak, R. J., Rackovsky, S. & Scheraga, H. A. (1993). Prediction of protein conformation on the basis of a search for compact structures: test on avian pancreatic polypeptide. *Protein Sci.* **2**, 1715–1731.
51. Keskin, O. & Bahar, I. (1998). Packing of sidechains in low-resolution models for proteins. *Fold. Des.* **3**, 469–479.
52. Miyazawa, S. & Jernigan, R. L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.
53. Tsai, C. J., Xu, D. & Nussinov, R. (1998). Protein folding *via* binding and *vice versa*. *Fold. Des.* **3**, R71–R80.
54. Halperin, I., Wolfson, H. & Nussinov, R. Protein–protein interactions: coupling of structurally conserved residues and of hot spots across protein–protein interfaces. Implications for docking. *Structure*, **12**, 1027–1038.
55. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
56. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
57. Norel, R., Sheinerman, F. B., Petry, D. & Honig, B. (2001). Electrostatic contributions to protein–protein interactions: fast energetic filters for docking and their physical basis. *Protein Sci.* **10**, 2147–2161.
58. Fernández, A. & Scheraga, H. A. (2003). Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl Acad. Sci. USA*, **100**, 113–118.
59. Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
60. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
61. Lattman, E. E., Fiebig, K. M. & Dill, K. A. (1994). Modeling compact denatured states of proteins. *Biochemistry*, **33**, 6158–6166.
62. Bromberg, S. & Dill, K. A. (1994). Side-chain entropy and packing in proteins. *Protein Sci.* **3**, 997–1009.
63. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A relationship between protein stability and function. *Proc. Natl Acad. Sci. USA*, **92**, 452–456.
64. Xu, J., Baase, W. A., Baldwin, E. & Matthews, B. W. (1998). The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci.* **7**, 158–177.
65. Nussinov, R. & Wolfson, H. J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl Acad. Sci. USA*, **88**, 10495–10499.
66. Higgins, D., Thompson, J., Gibson, T., Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
67. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Edited by G. von Heijne

(Received 20 July 2004; received in revised form 18 October 2004; accepted 26 October 2004)