Dissecting Protein–Protein Recognition Sites

Pinak Chakrabarti[1] and Joël Janin[2]*
[1]*Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Calcutta, India*
[2]*Laboratoire d'Enzymologie et de Biochimie Structurales, CNRS UPR9063, Gif-sur-Yvette, France*

ABSTRACT **The recognition sites in 70 pairwise protein–protein complexes of known three-dimensional structure are dissected in a set of surface patches by clustering atoms at the interface. When the interface buries <2000 Å^2 of protein surface, the recognition sites usually form a single patch on the surface of each component protein. In contrast, larger interfaces are generally multipatch, with at least one pair of patches that are equivalent in size to a single-patch interface. Each recognition site, or patch within a site, contains a core made of buried interface atoms, surrounded by a rim of atoms that remain accessible to solvent in the complex. A simple geometric model reproduces the number and distribution of atoms within a patch. The rim is similar in composition to the rest of the protein surface, but the core has a distinctive amino acid composition, which may help in identifying potential protein recognition sites on single proteins of known structures. Proteins 2002;47:334–343.**

Key words: molecular recognition; protein–protein interaction; interfaces; protein surface; residue clusters; amino acid composition

INTRODUCTION

Protein–protein recognition depends on the physical and chemical properties of the interfaces that form as two protein surfaces come in contact to form a specific complex. The structural basis of recognition resides in the atomic coordinates of the many protein–protein complexes for which an X-ray structure is available through the Protein Data Bank (PDB).[1] Protein–protein interfaces in these complexes have been characterized in terms of their geometry (size, shape, and complementarity) and of their chemical nature (the types of chemical groups and amino acids, hydrophobicity, electrostatic interactions, and hydrogen bonds);[2–14] for a recent survey, see the book edited by Kleanthous[15]. Most of the structural data and much of the biochemical data concern either enzymes, mostly proteases, interacting with protein inhibitors, or antibodies interacting with cognate antigens. In recent years, similar data have also become available for recognition processes involved in other physiological processes, such as signal transduction and the cell cycle. The proteins implicated in these processes are extremely diverse, but their recognition sites share some common properties. Subunit interfaces in oligomeric proteins are generally hydrophobic and similar in composition to the surface buried inside proteins on folding.[8,16,17] Instead, the interfaces found in protein–protein complexes resemble the remainder of the protein surface in their polar character and amino acid composition. As a consequence, generally it is difficult to identify potential protein recognition sites on the basis of their chemical composition, although this would be of great interest for functional prediction.[9,18,19]

In the present study, we dissect the protein–protein interfaces in a sample of 70 protein–protein complexes and show that they can be subdivided in at least two different ways. The larger recognition sites, such as the ones involved in signal transduction, comprise several patches on the protein surface, which can be identified by a simple geometric clustering algorithm. At least one of these patches is equivalent in size to the single surface patch that constitutes the recognition site of a protease inhibitor or the epitope of a protein antigen and forms with the cognate site on the protease or the antibody what Lo Conte et al.[12] called a "standard size" interface. In general, recognition sites can be viewed as being made of a standard size recognition patch, often augmented by adding smaller surface patches, and in a few cases, by duplication.

We then identify a core and a rim within recognition patches. The core region contains atoms that are buried on complex formation and is surrounded by a rim of atoms that remains partly accessible. The two regions differ in their amino acid composition: the rim is very similar to the remainder of the protein surface, whereas the core has a distinctive composition. Residue propensities for the core and rim are derived from the observed occurrences of the 20 amino acids on the protein surface and in the two regions of the recognition sites, and their relative contribution to the interface areas. These propensities can be used in identifying recognition sites on the protein surface.

METHODS AND RESULTS

Patches Within Recognition Sites

The 70 protein–protein interfaces studied here are listed in Table I. The sample was taken from the set of protein–

Grant sponsor: Indo-French Centre for the Promotion of Advanced Research (CEFIPRA).

*Correspondence to: J. Janin, Laboratoire d'Enzymologie et de Biochimie Structurales, CNRS UPR9063 91198, Gif-sur-Yvette, France. E-mail: janin@lebs.cnrs-gif.fr or P. Chakrabarti, Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Calcutta 700 054, India. E-mail: pinak@bic.boseinst.ernet.in

Received 27 August 2001; Accepted 30 November 2001

TABLE I. Structural Features of Recognition Sites*

PDB file	Interface	Patches	Area (Å^2)	Residues	Atoms	Relative patch size (fraction of atoms)		
Protease-inhibitor (18)								
2ptc	Trypsin-PTI	1	1430	48	167			
1mct	Trypsin-bitter gourd inhibitor	1	1510	47	177			
1avw	Trypsin-soybean inhibitor	1	1740	55	202			
3tpi	Trypsinogen-PTI	1	1600	64	215			
1tgs	Trysinogen-PSTI	1	1720	54	193			
1cho	Chymotrypsin-ovomucoid	1	1470	45	166			
1acb	Chymotrypsin-eglin C	1	1540	49	183			
1cbw	Chymotrypsin-PTI	1	1460	46	160			
1ppf	Elastase-ovomucoid	1	1320	44	151			
1fle	Elastase-elafin	1	1770	47	176			
2kai	Kallikrein-PTI	1	1340	42	150			
1hia	Kallikrein-hirustatin	1	1740	51	169			
3sgb	*S. griseus* protease B-ovomucoid	1	1270	38	147			
1cse	Subtilisin-eglin C	1	1490	45	155			
2sic	Subtilisin-SSI	1	1620	48	180			
2sni	Subtilisin-CI2	1	1630	48	181			
1stf	Papain-stefin	1	1690	50	181			
4cpa	Carboxypeptidase A-inhibitor	1	1360	41	147			
Large protease complexes (5)								
1bth	Thrombin E192Q-PTI	1	2240	72	240			
4htc	Thrombin-hirudin	2	3310	95	343	0.57	0.43	
1tbq	Thrombin-rhodniin	2	3470	104	357	0.61	0.39	
1toc	Thrombin-ornithodorin	2	3500	94	365	0.55	0.45	
1dan	Factor VIIA-soluble tissue factor	3	3180	86	316	0.50	0.26	0.24
Antibody-antigen (18)								
1jhl	Fv D11.15-lysozyme	1	1250	38	147			
1vfb	Fv D1.3-lysozyme	1	1380	42	155			
1mlc	Fab D44.1-lysozyme	1	1390	43	155			
3hfl	Fab HyHEL5-lysozyme	1	1710	48	183			
3hfm	Fab HyHEL10-lysozyme	1	1610	48	180			
1fbi	Fab 9.13.7-lysozyme	1	1690	49	184			
1mel	Camel H chain-lysozyme	1	1690	46	177			
1dvf	Fv D1.3-Fv E5.2	1	1630	51	173			
1nfd	Fab H57-N15 T cell receptor	1	1620	50	184			
1ao7	T cell receptor-HLA A2	1	1990	56	222			
1jel	Fab Jel42-HPR	1	1360	42	141			
1nca	Fab NC41-flu neuraminidase	1	1950	59	214			
1nmb	Fab NC10-flu neuroaminidase	2	1290	40	132	0.55	0.45	
1nsn	Fab N10-Staph. nuclease	2	1780	60	195	0.50	0.50	
1osp	Fab-*Borrelia* OSP A	2	1470	43	157	0.60	0.40	
1qfu	Fab BH151-flu HAX31	2	1840	58	195	0.68	0.32	
1iai	Fab 730.1.4-Fab 409.5.3	2	1890	61	214	0.75	0.25	
1kb5	Fab Désiré-1-TCR Fv	3	2320	66	236	0.50	0.29	0.21
Enzyme complexes (8)								
2pcc	Peroxidase-Cytochrome c	1	1140	33	106			
1gla	Glycerol kinase-Factor IIIGlc	1	1300	38	130			
1brs	Barnase-barstar	1	1560	43	177			
1udi	Uracil DNA glycosylase-inhibitor	1	2020	61	208			
1dhk	α-Amylase-bean inhibitor	1	3020	95	333			
1fss	Acetylcholinesterase-fasciculin	2	1970	56	220	0.60	0.40	
1ydr	Protein kinase A-inhibitor	2	2000	54	202	0.70	0.30	
1dfj	RNase A-RNase inhibitor	3	2580	90	291	0.56	0.26	0.18
G-proteins, cell cycle, signal transduction (11)								
1a0o	CheA-Che Y	1	1130	33	116			
1gua	Rap1A-cRaf1	1	1290	37	135			
1a2k	Ran-NFT2	1	1650	48	168			

TABLE I. (Continued)

PDB file	Interface	Patches	Area (Å^2)	Residues	Atoms	Relative patch size (fraction of atoms)		
1agr	$G_{1\alpha}$-RGS4	1	1630	52	181			
1tx4	Rho-Rho GAP	2	2280	61	229	0.70	0.30	
1gg2	$G_{i\alpha1}$-$G_{1\beta1\gamma2}$	2	2330	74	243	0.66	0.34	
1got	Transducin $G_{t\alpha}$-$G_{t\beta\gamma}$	2	2500	76	254	0.65	0.35	
2trc	$G_{t\beta\gamma}$-phosducin	2	4430	121	468	0.54	0.46	
1fin	CDK2-cyclin A	2	3400	94	355	0.76	0.24	
1aip	EFtu-EFts *T. thermophilus*	3	2880	88	288	0.38	0.34	0.28
1efu	EFtu-EFts *E. coli*	4	3630	108	357	0.30	0.28	0.23
Miscellaneous (10)								
1ak4	Cyclophilin-HIV capsid	1	930	27	102			
1igc	Protein G-Fab MOPC21	1	1130	28	117			
1efn	Fyn SH3 domain-HIV Nef	1	1250	33	131			
1fc2	Protein A-Fc fragment	1	1300	35	140			
1seb	HLA DR1-enterotoxin B	1	1340	41	145			
1atn	Actin-DNase I	1	1770	55	147			
1ycs	p53 core-53BP2	2	1500	43	163	0.69	0.31	
2btf	Actin-profilin	2	2060	60	221	0.55	0.45	
1hwg	HGH receptor-human growth hormone	2	4200	117	462	0.60	0.40	
1dkg	Grep E-DNA K	3	1970	59	205	0.44	0.32	0.24
All 70 interfaces		Average	1906	57	204			
		SD	759	22	78			
46 single-patch		Average	1560	47	170			
		SD	340	11	39			
18 two-patch		Average	2510	73	217	0.63		
		SD	960	25	102	0.08		
6 three and four-patch		Average	2760	83	258	0.45		
		SD	600	18	137	0.09		

*Protein–protein recognition sites may form one or several patches on the protein surface. The interface areas and the number of interface atoms and residues cited here are for both protein components of each complex. When the interface is multipatch, the fraction of the number of atoms belonging to each patch also is cited.

protein complexes used by Lo Conte et al.,[12] omitting the erythropoietin receptor–peptide complex (1ebp) and four complexes for which atomic coordinates are not in the PDB. It comprises 23 protease-inhibitor interfaces, 18 antibody–antigen interfaces, and 29 interfaces in complexes of other kinds. Table I cites their interface area and the number of atoms and residues present at the interface. The interface area is the area buried in the interaction, measured as the sum of the solvent accessible surface area (ASA) of the two component proteins, less that of the complex.[2] Recognition sites on the two protein components of a complex contain equivalent numbers of atoms, and they contribute almost equally to the interface area, with a few exceptions where the surfaces are strongly curved and the convex side contributes 10–15% more area than the concave side.[12] ASA values were computed with program ACCESS,[20] which implements the algorithm of Lee and Richards.[21] The group radii and probe size (1.4 Å) were the same as in Ref. 12. However, we omitted from the calculation all atoms with temperature factors of 0 or occupancy factors < 1, which explains minor differences with published values. All atoms or amino acid residues that lose >0.1 Å^2 ASA in the complex were counted as interface atoms or residues.

Lo Conte et al.[12] noted that most of the protein–protein complexes bury a surface area in the range of 1200–2000 Å^2 and defined their interfaces as "standard size." The present sample of 70 complexes includes 47 standard-size interfaces. All 18 antibody–antigen interfaces but one (1kb5), 18 of the 23 protease-inhibitor interfaces, and 12 of the other types are standard size. The sample also includes four small interfaces burying <1200 Å^2 and 19 large interfaces burying more than 2000 Å^2. The five protease-inhibitor interfaces listed as "large" in Table I have interfaces above that limit. Three of them involve the blood serine protease thrombin and its inhibitors hirudin, rhodniin, and ornithodorin. The inhibitors are unrelated to each other, but they all bind at the same two sites on the protease: the active site and a secondary site specific for fibrinogen, the natural substrate of thrombin. They differ in their binding mode from other protease inhibitors, which recognize only the active site. In the thrombin–ornithodorin complex,[22] for instance, two distinct patches of the protease surface assemble with two patches on the inhibitor surface. The interface buries 3500 Å^2, twice the area of the pancreatic trypsin inhibitor (PTI)–trypsin interface.[23] In the family of blood serine proteases, Factor VIIA also makes a large interface with soluble tissue

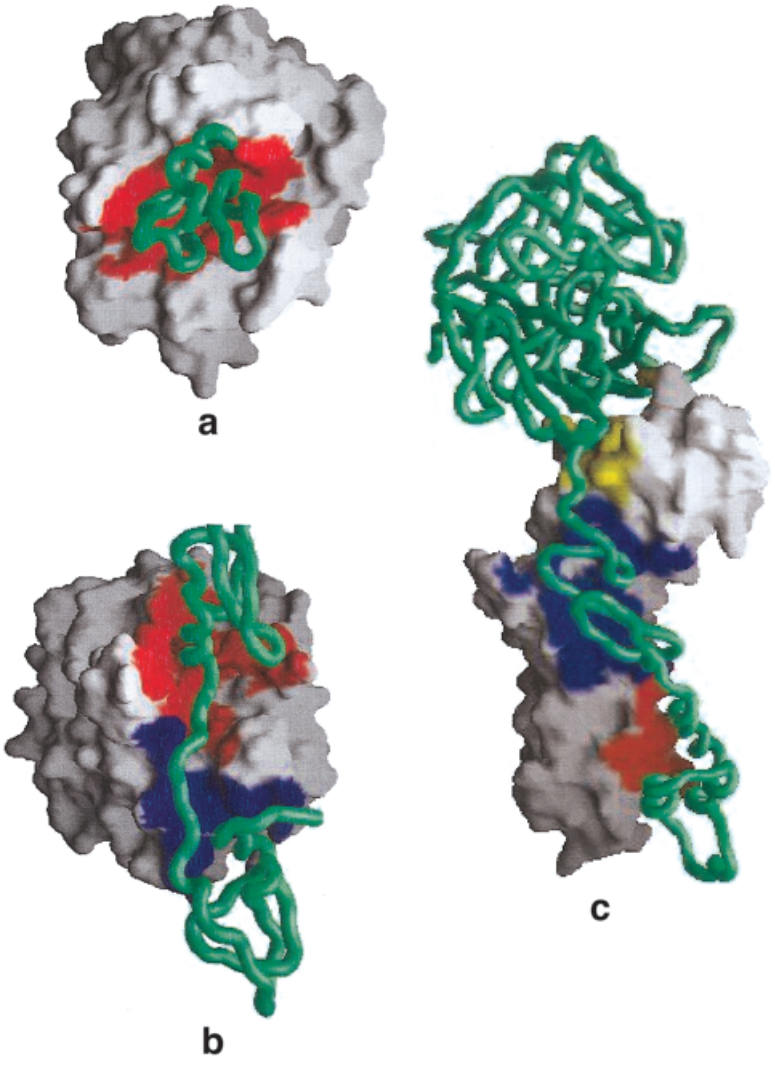

Fig. 1. Recognition patches in protease-inhibitor complexes. Recognition sites are colored on the molecular surface of one protein component, the other component being imaged as a backbone tube. **A:** The surface of trypsin in red in contact with PTI in green forms a standard size single-patch interface (2ptc).[23] **B:** Ornithodorin in green has two PTI-like domains and forms a two-patch interface with thrombin (1toc)[22]; the red and blue patches are equivalent in size to the trypsin patch in (A). **C:** Distinct domains of Factor VIIA in green make contact with the three patches in red, blue, and yellow, of the soluble tissue factor surface (1dan).[24] Drawn with GRASP.[41]

factor, a protein that activates rather than inhibits its protease activity, and the interaction involves three sites carried by separate domains of Factor VIIA.[24] These interfaces are illustrated in Figure 1.

We assumed that other recognition processes than protease regulation may involve more than one patch on the protein surface. To identify such patches, we clustered interface atoms by the average linkage method.[25] A threshold distance of 15 Å was selected on the basis of the maximum distance between two atoms of a standard size interface, which is about 30 Å. Very similar results were obtained with thresholds in a range of 13–18 Å. The clustering algorithm was run separately on each polypeptide chain of each component protein of a complex. In a few cases, it found a different number of clusters on the two components. The difference was never more than one, and it could be brought to zero by adjusting the threshold distance to a value between 15 and 20 Å. In this way, recognition patches always occur in pairs, one on each component protein, and the number of patches reported in Table I is the number of pairs present in each complex.

Table I shows that the recognition site comprises a single pair of patches in 46 of the 70 interfaces (single-patch interfaces), two patches in 18, three in 5, and 4 in 1. The algorithm finds all protease-inhibitor interfaces to be single-patch, except the three thrombin complexes mentioned above, where it correctly finds two pairs of patches, and the Factor VIIA-soluble tissue factor complex, where it finds three, also in agreement with the structure.

The data indicate a strong correlation between the size of the interface and the number of recognition patches. Forty-three of 46 single-patch interfaces bury <2000 Å^2

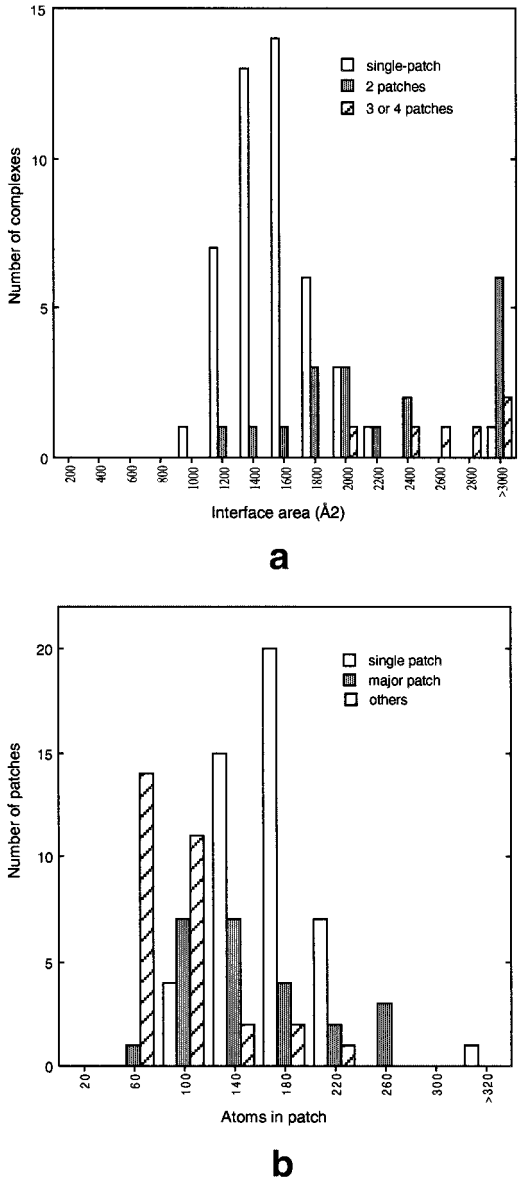

Fig. 2. Interface and patch size. **A:** Histograms of the interface area in 46 single-patch interfaces (empty bars), 18 two-patch interfaces (filled bars), and 6 three or four-patch interfaces (strips); patches are defined by clustering interface atoms as described in the text. **B:** Histograms of the number of interface atoms in the 46 single-patch interfaces (empty bars), in the major pair of patches of the 24 multipatch interfaces (filled bars), and of other pairs in the same interfaces (stripped).

and, therefore, are standard size or smaller [Fig. 2(A)]. Conversely, all but 8 of the 47 standard size interfaces are single-patch. Single-patch interfaces > 2000 Å^2 occur in three enzyme-inhibitor complexes. The first involves uracil-DNA glycosylase (1udi) and is just above the limit. The second is thrombin-PTI (1bth), where the inhibitor binds at the active site of the protease as it does in trypsin-PTI, but loop movements on thrombin bury additional surface. The largest single-patch interface involves α-amylase (1dhk), which has large mobile loops that interact with a lectin-like inhibitor in addition to its active site.[26] The loops are not identified as distinct patches by the clustering algorithm.

Seven two-patch and one three-patch interfaces bury <2000 Å^2. They occur in five antibody–antigen complexes, p53 core-53BP2 (1ycs), acetylcholinesterase-fasciculin (1fss), and Grep E-DNA K (1dkg). The interaction with the p53 core protein involves two distinct regions of 53BP2: a SH3 domain and an ankyrin repeat.[27] The total buried surface area is only 1500 Å^2, but some contacts may be missing because neither of the two components of the crystalline complex is the complete protein. The acetylcholinesterase–fasciculin and the Grep E–DNA K interfaces are just below the 2000 Å^2 limit. The presence of two recognition patches at the antibody-combining sites is a consequence of the clustering procedure, which we ran separately on the heavy and light chains. The algorithm finds three patches at the antibody–antigen interface between Fab Désiré-1 and the Fv fragment of the T-cell receptor[28] (1kb5). This is the largest antibody–antigen interface in our sample and the only one above standard size.

Geometric Properties of the Patches

Size of the patches

The 70 interfaces contain 101 pairs of recognition patches. On average, a pair of patches buries 1320 ± 520 Å^2 (mean ± SD, here and below) of protein surface. The 46 single-patch interfaces bury 1560 ± 340 Å^2, or about 800 Å^2 per recognition patch. The mean value of the single-patch interface area and its standard deviation are close to the mean (1600 Å^2) and range (400 Å^2), which defines a standard size interface. The 18 two-patch interfaces bury an average of 2510 Å^2; the 6 three- and four-patch interfaces bury 2760 Å^2. Thus, patches in multipatch interfaces tend to be smaller than a single-patch interface. However, Figure 2(B) shows that these interfaces generally contain at least one pair of patches that is equivalent in size to a single-patch interface. This figure is a histogram of the number of interface atoms rather than of areas, but the two quantities are proportional (see Fig. 4 below). Table I gives the fractional distribution of atoms between patches in multipatch interfaces. On average, a single-patch interface includes 170 ± 39 surface atoms, or 85 atoms per recognition patch. In a two-patch interface, one pair of patches is often significantly larger than the other pair. On average, this major pair represents 63% of the interface and includes 165 ± 60 atoms, almost the same number as in a single-patch interface. Thus, it would by itself constitute a standard size interface. The rule that multipatch interfaces contain a pair of patches that is at least standard size applies to all two-patch interfaces in our sample, except three antibody–antigen interfaces. In four cases, a second pair is also standard size: the three thrombin-inhibitor complexes mentioned above, and the complex of the human growth hormone with its receptor.

Among the six interfaces with three and four pairs of patches, those of the Factor VIIA-soluble tissue factor (1dan), Fab Désiré–TCR Fv (1kb5), and RNase A–RNase inhibitor (1dfj) complexes contain one standard size and

two small pairs of patches. In the other cases, the patches are all smaller than standard size.

Flatness

Argos[29] and Jones and Thornton[8] have noted that the interfaces of protein–protein complexes tend to be flat. We used the same criterion as Jones and Thornton[8] to analyze deviations from planarity in an interface or recognition patch: the RMS distance Δ of the atoms to a least-square plane drawn through them. The mean value of Δ was 2.8 Å in their sample of protein–protein complexes, which comprises almost exclusively single-patch interfaces. Antibody–antigen interfaces are more planar than average, protease-inhibitor interfaces, less planar[8,12]: Δ is near 2.5 Å in the first case, 3.5 Å in the second, and 3.0 ± 0.6 Å for all 101 pairs of patches of our sample. Most (52%) of the patches are like antibody–antigen interfaces and have Δ in the range 2–3 Å. Most of the remainder (42%) is like standard size protease-inhibitor interfaces with Δ in the range of 3–4 Å. Five pairs of patches are less planar than those with Δ in the range of 4–4.7 Å. They occur in the thrombin–rhodniin, thrombin–ornithodorin, Factor VIIA–soluble tissue factor, $G_{t\beta\gamma}$–phosducin, and HGH receptor–human growth hormone complexes. Lastly, the α-amylase-inhibitor interface, which is by far the largest single-patch interface in our sample, is clearly nonplanar with Δ = 5.6 Å.

Recognition Patches Have a Core and a Rim

Another way to split recognition sites into distinct regions is based on their accessibility to solvent. Lo Conte et al.[12] define three categories of atoms present at protein–protein interfaces in approximately equal numbers: type A remains accessible and has a non-zero residual ASA in the complex; type B is fully buried and has zero ASA; type C is accessible, but it also has zero ASA if water molecules listed in the PDB file are included in the calculation. Type C is defined only in those X-ray structures that report solvent positions, and we consider here only two types: accessible for A and C, buried for B. The recognition site of the CI2 inhibitor forms a single-patch, standard size interface with the active site of subtilisin[30] (2sni). It is shown in Figure 3 next to a model recognition site adapted from Lo Conte.[12] In the model, 29 buried atoms form a core surrounded by 50 accessible atoms. On average, a single-patch interface has 31 ± 14 buried and 56 ± 30 accessible atoms. The number of interface and buried atoms is plotted in Figure 4(A) against the interface area. The two quantities are lineally correlated with a correlation coefficient $R^2 = 0.98$ and a slope of 9.5 $Å^2$ per interface atom. If the larger interfaces were single-patch, the proportion of buried atoms should increase linearly with their size. Instead, the relative number of buried atoms remains constant at 34% irrespective of the size of the interface, whereas the model predicts 36%. Thus, the model applies to patches in multipatch interfaces as well as to single-patch ones.

We analyzed the spatial distribution of buried and accessible atoms within the recognition patches. Using the suite of programs SURFNET,[31] we drew a least-square plane through all atoms of each patch, projected the atoms on that plane, and measured the distance d of the projections to the center of mass of the patch. We then compared the average value of d for buried atoms, d_B, to that d_{all}, for all interface atoms. In the model of Figure 3(A), these distances are expected to be near d_{all} = 8.0 Å and d_B = 4.8 Å, the patch being approximately 20 × 26 Å in size. The ratio d_{all}/d_B is 1.65 instead of 1 if buried, and accessible atoms were randomly distributed within the patch. In comparison, the CI2 recognition site has d_{all} = 7.8 Å, d_B = 6.5 Å, and a ratio of 1.2. Figure 3(B) shows that it contains only buried atoms at its center, but the rim of accessible

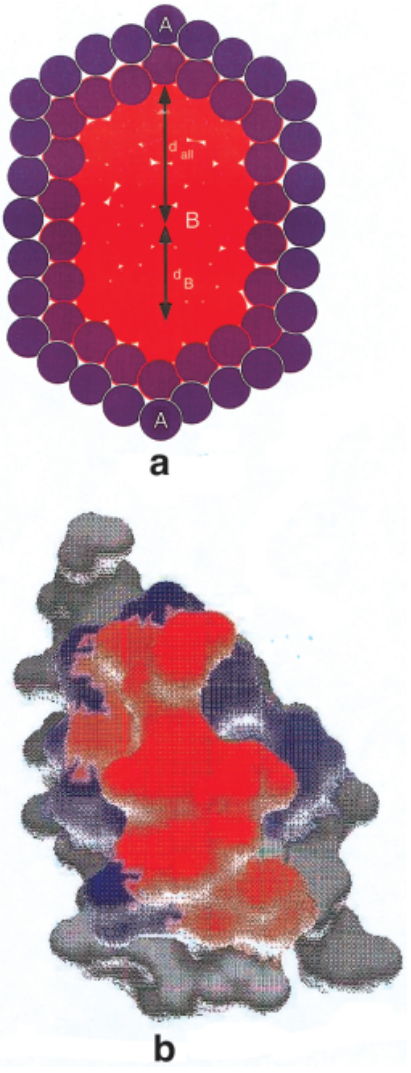

Fig. 3. Accessible and buried interface atoms. **A:** Sketch of the recognition patch in a standard size interface; the patch has a core made of 29 buried atoms (red) and a rim of 50 atoms in two shells that remain accessible to solvent in the complex. Atoms of the inner shell (purple) would be buried if crystallographic water molecules were taken into account; those of the outer shell (blue) are accessible to bulk water. d_{all} is the mean distance to the center of all the atoms in the patch; d_B is that of the buried atoms. Adapted from Lo Conte et al.[12] **B:** The recognition site of the CI2 inhibitor in complex with subtilisin (1sni).[30] The molecular surface is colored in red for buried interface atoms, and in blue for accessible atoms.

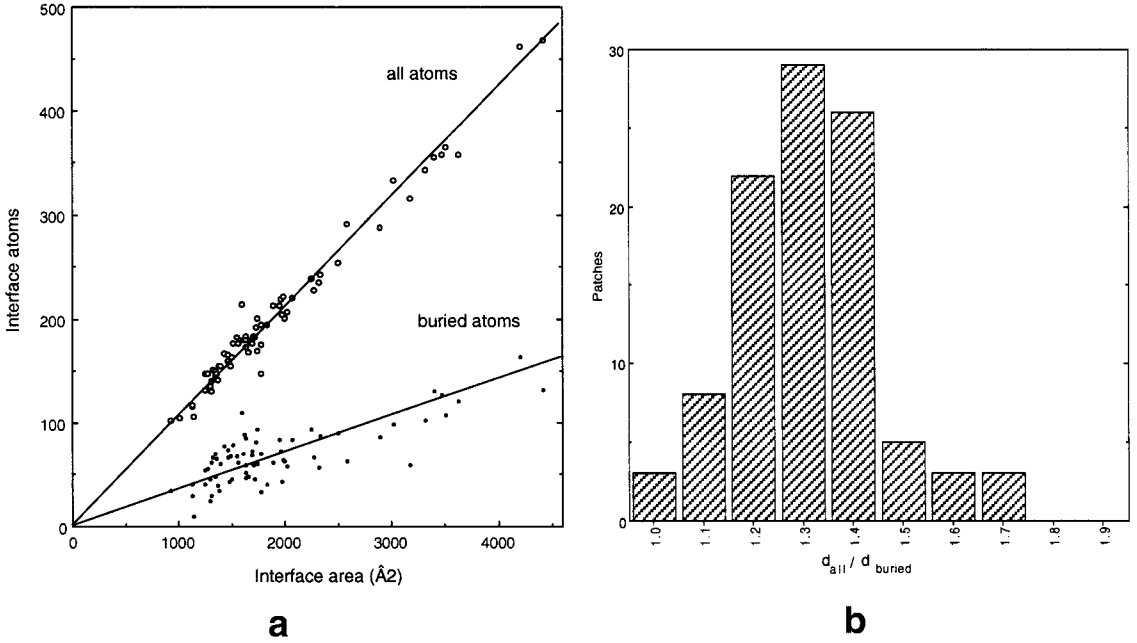

Fig. 4. Number and distribution of the interface atoms. **A:** The number of interface atoms and that of buried atoms is proportional to the area of the 70 interfaces. **B:** Histogram of the ratio d_{all}/d_B in the 101 recognition patches; d_{all} and d_B are defined in Figure 3.

atoms is incomplete and the ratio is lower than in the model. Over the 101 patches, d_{all} is 8.0 ± 1.7 Å, d_B, 6.5 ± 1.7 Å. A histogram of the values of the d_{all}/d_B ratio [Fig. 4(B)] indicates that 70% are in the range of 1.2–1.6. Thus, most of the patches are at least as close to the model as in CI2.

Amino Acid Composition of the Core and Rim

We count as interface residues all amino acid residues on the protein surface that lose ASA in the complex. Table I indicates that there are 57 ± 22 residues per protein–protein interface or about 29 per recognition site. Single-patch interfaces contain 47 ± 11 residues or 23 residues per recognition site. Like the number of atoms, the number of interface residues is proportional to the interface area, the linear correlation coefficient being $R^2 = 0.96$ and the slope 36 Å^2 per residue.

Residues are attributed to the core of the interface if they contain at least one buried interface atom and to the rim if they contain only accessible atoms. The sample contains 3973 interface residues, 53% of which belong to the core and 47% to the rim. On average, each protein component of a complex provides 15 residues to the core and 14 to the rim of the interface. In a standard size interface, or in the major patch of a multipatch interface, these numbers are 12 and 11. Core residues contribute 72% of the interface area; rim residues contribute only 28%. The amino acid compositions of the core, rim, and complete interfaces are given in Table II. They can be compared to the composition of the 23,063 residues present on the surface of all the proteins in the sample. A surface residue is any residue with an accessibility > 5%, defined as the ratio of its ASA in the protein to that in a model Ala-X-Ala tripeptide.[32] Two fractional compositions are cited in Table II: by number of occurrences of each residue type and by the area they contribute to the protein surface or interface. The difference between the number-based and area-based compositions can be significant. Arg and Thr residues are present at protein–protein interfaces in equivalent proportions (about 6.3%), but Arg contributes 10% of the average interface area, but Thr only 5%.

An Euclidian metrics can be defined in the 19-dimensional space of amino acid compositions by computing Δf between the compositions f and f′ as:

$$(\Delta f)^2 = 1/19 \sum_{i=1 \text{ to } 20} (f_i - f'_i)^2 \qquad (1)$$

With that metrics, the protein surface and interior differ in composition by Δf = 3.9%, setting an extreme value for the distance between two amino acid compositions. The average subunit interface in oligomeric proteins is similar in composition to the protein interior and remote from the surface (Δf = 1.3% and 3.7%, respectively[12,16]), but protein–protein interfaces are equally distant from the surface and the interior (Δ = 2.6‰). Figure 5, which summarizes the data of Table II, indicates that this results from averaging a core that is slightly more like the interior than the surface (Δf = 2.7% vs 3.4%) with a rim that is very close to the surface (Δf = 1.2%). This difference between core and rim is also apparent in the number-based compositions: the rim is only at Δf = 1.1% of the average protein surface, whereas the core is at 2.2%.

The core and the rim of a recognition site contain about the same number of amino acid residues, typically 12 in a standard size interface, but they do not have the same

TABLE II. Amino Acid Composition of Protein-Protein Interfaces

	Number (a)			Area (b)			Propensities (c)		Lo Conte	Jones and
Residue	Interface	Core	Rim	Interface	Core	Rim	Core	Rim	et al. (d)	Thornton (e)
All	100.0	100.0	99.9	99.9	100.0	100.0				
Ala	3.9	4.0	3.8	2.8	2.7	3.1	−0.40	−0.26	−0.43	−0.17
Arg	6.4	5.9	7.0	10.1	10.1	9.9	0.13	0.11	0.13	0.27
Asn	5.9	5.4	6.4	5.7	5.4	6.4	−0.14	0.03	−0.12	0.12
Asp	6.6	5.4	8.0	5.1	4.5	6.6	−0.46	−0.07	−0.31	−0.38
Cys	3.5	4.7	2.1	1.7	1.9	1.3	1.00	0.62	0.76	0.43
Gln	3.7	3.7	3.8	4.3	4.3	4.2	−0.34	−0.36	−0.36	−0.11
Glu	6.5	4.6	8.6	6.0	4.4	10.0	−0.80	0.02	−0.47	−0.13
Gly	8.1	7.5	8.7	4.8	4.2	6.4	−0.08	0.35	0.02	−0.07
His	3.4	4.4	2.3	3.8	4.4	2.4	0.84	0.23	0.64	0.41
Ile	3.6	4.1	3.1	4.6	4.9	3.5	0.71	0.38	0.56	0.44
Leu	5.0	5.5	4.5	5.7	5.8	5.3	0.34	0.25	0.29	0.40
Lys	5.7	3.7	8.0	6.5	5.2	9.7	−0.82	−0.20	−0.57	−0.36
Met	2.0	2.6	1.4	3.2	3.7	2.0	1.13	0.51	0.98	0.66
Phe	3.5	5.1	1.7	4.1	5.5	1.1	1.01	−0.60	0.79	0.82
Pro	3.8	3.4	4.2	3.6	3.5	4.1	−0.38	−0.22	−0.25	−0.25
Ser	7.9	7.8	8.1	5.4	4.8	7.3	−0.56	−0.14	−0.42	−0.33
Thr	6.2	5.7	6.8	5.0	4.7	5.9	−0.44	−0.21	−0.35	−0.18
Trp	2.8	4.1	1.3	4.2	5.3	1.6	1.41	0.21	1.25	0.83
Tyr	6.8	8.1	5.4	9.4	10.9	5.3	1.22	0.50	1.04	0.66
Val	4.5	4.3	4.7	3.8	3.8	3.9	0.08	0.11	0.09	0.27

(a) Number-based compositions: percent of residues present in the 70 interfaces, their core, or their rim; (b) Area-based compositions: percent contributed to the area of the 70 interfaces, their core, or their rim; (c) the propensity for a residue to be part of the core or the rim is $p_i = \ln(f_i/f_i^A)$, where f_i is the area-based composition of the core or rim, f_i^A, the area-based composition of the protein accessible surface reported in Table 4 of Lo Conte et al[12]; (d) propensity for a residue to be part of a protein-protein interface derived from the area-based compositions reported in the same Table; (e) area-based propensities reported in Table 2 of Jones & Thornton.[9]

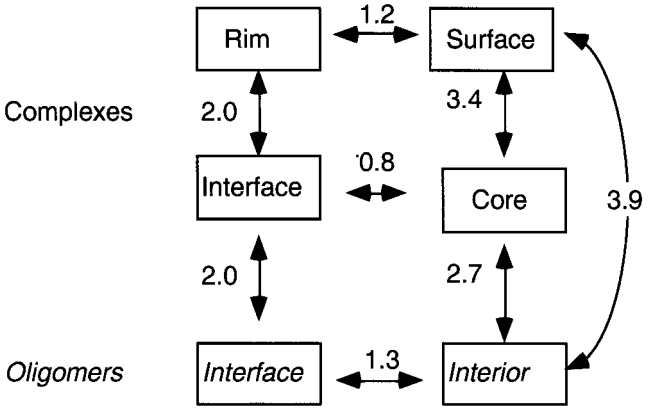

Fig. 5. Distances between amino acid compositions. Distances (in percent) are defined as in the text. The area-based amino acid compositions for the rim, core, and whole interfaces are listed in Table II. Area-based compositions for the protein surface and for the interface and interior of oligomeric proteins are taken from Table IV of Lo Conte et al.[12]

amino acid composition. The rim resembles the remainder of the protein surface; the core has a distinctive composition, with an excess of aromatic residues and a deficit in charged residues except Arg. The partition of the amino acid residues between regions of the protein surface can also be expressed as a set of propensities defined as:

$$p_i = \ln(f_i/f_i^A) \quad (2)$$

where the area-based composition f^A of the solvent accessible surface is taken as a reference (Table II). Propensities for the core are strongly marked, but not for the rim, where most of the p_i are not significantly different from zero.

DISCUSSION

Lo Conte et al.[12] noted two features that are present in most enzyme-inhibitor and antibody–antigen complexes. They form standard size interfaces, and they assemble as rigid bodies, that is, the component proteins undergo only minor changes of conformation on association. In contrast, the formation of complexes with interfaces larger than 2000 Å^2 generally induces major changes, which often play an essential role, as in signal transduction. In this respect, protein–protein recognition can be compared with protein–DNA recognition, which generally involves interfaces much larger than 2000 Å^2 and major conformation changes.[33,34] Most DNA-binding proteins are oligomers or tandem repeats of a simpler protein module that bear distinct recognition sites. In many protein–DNA complexes, individual recognition sites bury 1200–2000 Å^2 of protein and DNA surface, just like a standard size protein–protein interface.[33,35] Thus, individual protein–DNA and protein–protein recognition sites are often of comparable size.

The present study extends the comparison by showing that, like most protein–DNA interfaces, large protein–protein interfaces involve more than one recognition patch. These patches are identified by applying a simple geometric clustering algorithm to the interface atoms. Standard size interfaces are generally single-patch: the recognition sites form one patch on the surface of each component protein. Large interfaces, which are multipatch, generally

contain a pair of patches that, if alone, would make a standard size interface. Cases where the protein is a homodimer or a tandem repeat duplicating recognition sites are less common in protein–protein than protein–DNA recognition. In our sample, the HGH receptor is a dimer, and ornithodorin is a repeat of two PTI-like modules. Both proteins carry two-patch recognition sites, but their partners, the growth hormone and thrombin, are not themselves duplicates. In other cases, the standard size interface created by the larger pair is extended by the presence of smaller patches, but not duplicated.

Our analysis of patches in protein–protein recognition sites can be related to previous studies by Jones and Thornton.[8,9] We analyze how a recognition site is distributed over the surface of a protein. Jones and Thornton[8] consider how it is distributed along the polypeptide chain. They find that the average recognition site comprises five continuous chain segments in a sample of 27 protein–protein complexes, mostly protease-inhibitor and antibody–antigen with standard size interfaces. Except in the very rare case when there is only one segment, continuous chain segments do not constitute a recognition patch in our sense, and most of our patches (at least standard size ones) involve several segments. In their other study, Jones and Thornton[9] define patches on the protein surface like we do, but for a different purpose. They aim to identifying regions involved in recognition and are designed to cover the whole surface. The patch size is adjusted on each protein to the size of the known recognition site, which by implication is taken to be a single patch.

The amino acid composition of recognition sites has also been analyzed by Jones and Thornton[8,9] and, on a larger set, by Lo Conte et al.[12] The results of these two studies are included in Table II in the form of area-based propensity scales. The two scales are similar, with a linear correlation coefficient $R^2 = 0.90$. Propensity scales may be viewed as empirical single-body potentials, which in the present case, express the free energy gained by a residue in contact with protein atoms instead of water. Keskin et al.[36] and Glaser et al.[37] derived such potentials from sets of PDB structures that included both protein–protein complexes and oligomeric proteins. Keskin et al.[36] find their single-body potentials to be very similar for interface residues and for residues buried inside proteins. These potentials are poorly correlated to the propensities derived from the data of either Jones and Thornton[9] or Lo Conte et al.,[12] the correlation coefficients being $R^2 = 0.69$ and 0.58. Presumably, the discrepancy is due to merging of interfaces in oligomeric proteins, which are like the protein interior, with protein–protein interfaces, which are not. Glaser et al.[37] list pair potentials rather than propensities. They note that, although the overall amino acid composition of their interfaces is the same as in Keskin et al.,[36] there is a significant difference between the larger interfaces, essentially from oligomeric proteins, and the smaller ones, mainly from complexes.

These studies have considered interfaces as a whole. We go further here in splitting recognition sites into patches and defining a core and a rim within these. The amino acid compositions of the core and rim are different. The propensity scale for residues of the core cited in Table II is essentially the same as for whole interfaces, the correlation coefficient with the scale derived from the data of Lo Conte et al.[12] being $R^2 = 0.99$. In contrast, the propensity scale for the rim is uncorrelated with that for whole interfaces ($R^2 = 0.34$).

A very different approach led Bogan and Thorn[38,39] to propose a model of protein–protein recognition sites that resembles our core-and-rim model. These authors tabulate changes in binding energy ($\Delta\Delta G$) observed on deletion of individual side chains by mutation to alanine (alanine scanning[40]). They find that the "hot spots" where $\Delta\Delta G > 2$ kcal $\cdot$ mol^{-1} are in general buried residues that cluster at the center of the recognition sites. They are surrounded by accessible residues that have a lesser effect on affinity. Hot spots have a distinctive amino acid composition, with some common features with the composition we observe for the core of recognition sites. Trp, Tyr, and Arg are highly preferred at hot spots, Ser and Thr are disfavored, and Leu, Val (but not Ile) are essentially absent.[38,39] In our data, Trp and Tyr have the highest propensity for the core of recognition sites, and Ser and Thr have a negative propensity. Arg is the most abundant core residue but also generally abundant on the protein surface. Unlike the hot spots, the core has Leu and Val residues. Their abundance is low, but not much lower than elsewhere on the protein surface. In addition, we find that typically 12–15 residues contribute to the core of a recognition site. Hot spots are generally much fewer. Thus, although hot spots generally belong to the core, many core residues, especially aliphatic residues, do not show up as hot spots in alanine-scanning experiments. This is in part because these experiments tend to give more weight to polar than to nonpolar interactions and because they do not test interactions made by main-chain atoms, which are as frequent as side-chain interactions.[12]

CONCLUSION

We believe the present analysis of recognition sites to be relevant to the study of protein structure and function in the postgenomic era. Genome sequencing gives access to a very large number of protein sequences, and structural genomic programs are on their way to provide experimental models for many single-chain proteins. These will be further used for homology modeling, and reliable structural models may soon be available for most of the proteins produced by single genes. However, the function is determined as much by the interaction they make, with small molecule ligands, with DNA or other proteins, as by the molecular structure. Structural data on multimolecular assemblies are being produced at a much slower rate, and there is a growing interest in methods to predict which regions of the protein surface are involved in such interactions. Protein–protein recognition sites are remarkably diverse in size and nature, but we find that they all comprise at least one patch that covers of the order of 800 Å^2 on the protein surface. A recognition patch of this size includes about 85 atoms belonging to some 23 residues,

half of which are part of the core and have a distinctive amino acid composition. We show that a simple geometric model reproduces these characteristics. Procedures designed to identify potential sites on the protein surface should take these observations into account.

ACKNOWLEDGMENTS

We thank Drs. Debnath Pal and Francis Rodier for help in computation.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Chothia C, Janin J. Principles of protein–protein recognition. Nature 1975;256:705–708.
3. Janin J, Chothia C. The structure of protein–protein recognition sites. J Biol Chem 1990;265:16027–16030.
4. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. J Mol Biol 1993;234:946–950.
5. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein–protein recognition. Protein Sci 1964;3:717–729.
6. Janin J. Principles of protein–protein recognition from structure to thermodynamics. Biochimie 1995;77:497–505.
7. Janin J. Protein–protein recognition. Prog Biophys Mol Biol 1996;64:145–166.
8. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
9. Jones S, Thornton JM. Anaysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
10. Xu D, Tsai C-J, Nussinov R. Hydrogen bonds and salt bridges across protein–protein interfaces. Protein Eng 1997;10:999–1012.
11. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a staistical analysis of the hydrophobic effect. Protein Sci 1997;6:53–64.
12. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
13. Norel R, Perrey D, Wolfson HJ, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999;36:307–317
14. Scheinerman FB, Norel R, Honig B. Electrostatic aspects of protein–protein interaction. Curr Opin Struct Biol 2000;10:153–159.
15. Kleanthous C. Protein–protein recognition: frontiers in molecular biology. Oxford, UK: Oxford University Press; 2000.
16. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164
17. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 1991;9:37–55.
18. Zucconi A, Panni S, Paoluzi S, Castagnoli L, Dente L, Cesareni G. Domain repertoires as a tool to derive protein recognition rules. FEBS Lett 2000;480:49–54.
19. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbour list. Proteins 2001;44:336–343.
20. Hubbard S. ACCESS: a program for calculating accessibilities. Department of Biochemistry and Molecular Biology, University College of London; 1992.
21. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
22. van de Locht A, Stubbs MT, Bode W, Friedrich T, Bollschweiler C, Hoffken W, Huber R. The ornithodorin–thrombin crystal structure, a key to the TAP enigma? EMBO J 1996;15:6011–6017.
23. Huber R, Kukla D, Bode W, Schwager P, Bartels K, Deisenhofer J, Steigemann W. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. II Crystallographic refinement at 1.9 Å resolution. J Mol Biol 1974;89:73–101.
24. Banner DW, d'Arcy A, Chene C, Winkler F, Guha A, Konigsberg WH, Nemerson Y, Kirschhofer D. The crystal structure of the complex of blood coagulation factor VIIA with soluble tissue factor. Nature 1996;380:41–46.
25. Johnson RA, Wiechert DW. Applied multivariate statistical analysis. New Delhi: Prentice-Hall of India; 1996.
26. Bompard-Gilles C, Rousseau,P, Rougé P, Payan P. Substrate mimicry in the active center of mammalian α-amylase: structural analysis of an enzyme-inhibitor complex. Structure 1996;4:1441–1452.
27. Gorina S, Pavletich NP. Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. Science 1996;274:1001–1005.
28. Housset D, Mazza G, Grégoire C, Piras C, Malissen B, Fontecilla-Camps JC. The three-dimensional structure of a T-cell antigen receptor Vα-Vβ heterodimer reveals a novel rearrangement of the Vβ domain. EMBO J 1997;16:4205–4211.
29. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng. 1988;2:101–113.
30. McPhalen CA, James MNG. Structural comparison of two serine proteinase–protein inhibitor complexes, Eglin C-subtilisin Carlsberg and CI 2-subtilisin novo. Biochemistry 1987;27:6582–6598.
31. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. J Mol Graph 1995;13:323–330.
32. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
33. Nadassy K, Wodak SJ, Janin J. Structural features of protein–nucleic acid recognition sites. Biochemistry 1999;38:1999–2017
34. Jones S, van Heyningen P, Berman HM, Thornton JM. J Mol Biol 1999;287:877–896
35. Janin J. Geometric features in protein–protein and protein–DNA recognition. In: Vijayan M, Yathindra N, Kolaskar AS, editors. Perspective in structural biology, a volume in honour of GN Ramachandran. New Delhi: Indian Acad. Sciences Univ. Press; 2000.
36. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Residue frequencies and pairing preferences at protein–protein interfaces. Protein Sci 1998;7:2578–2586.
37. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 2001;43:89–102.
38. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
39. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effect on the free energy of binding in protein interaction. Bioinformatics 2001;17:284–285.
40. Wells JA. Systematic mutational analyses of protein–protein interfaces. Methods Enzymol 1991;202:390–411.
41. Nicholls A, Sharp K, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1992;11:281–296.