doi:10.1016/j.jmb.2004.02.040 *J. Mol. Biol.* (2004) **338**, 181–199

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

ProMate: A Structure Based Prediction Program to Identify the Location of Protein–Protein Binding Sites

Hani Neuvirth[1,2], Ran Raz[2] and Gideon Schreiber[1]*

[1]*Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100 Israel*

[2]*Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100 Israel*

**Corresponding author*

Is the whole protein surface available for interaction with other proteins, or are specific sites pre-assigned according to their biophysical and structural character? And if so, is it possible to predict the location of the binding site from the surface properties? These questions are answered quantitatively by probing the surfaces of proteins using spheres of radius of 10 Å on a database (DB) of 57 unique, non-homologous proteins involved in heteromeric, transient protein–protein interactions for which the structures of both the unbound and bound states were determined. In structural terms, we found the binding site to have a preference for β-sheets and for relatively long non-structured chains, but not for α-helices. Chemically, aromatic side-chains show a clear preference for binding sites. While the hydrophobic and polar content of the interface is similar to the rest of the surface, hydrophobic and polar residues tend to cluster in interfaces. In the crystal, the binding site has more bound water molecules surrounding it, and a lower *B*-factor already in the unbound protein. The same biophysical properties were found to hold for the unbound and bound DBs. All the significant interface properties were combined into ProMate, an interface prediction program. This was followed by an optimization step to choose the best combination of properties, as many of them are correlated. During optimization and prediction, the tested proteins were not used for data collection, to avoid over-fitting. The prediction algorithm is fully automated, and is used to predict the location of potential binding sites on unbound proteins with known structures. The algorithm is able to successfully predict the location of the interface for about 70% of the proteins. The success rate of the predictor was equal whether applied on the unbound DB or on the disjoint bound DB. A prediction is assumed correct if over half of the predicted continuous interface patch is indeed interface. The ability to predict the location of protein–protein interfaces has far reaching implications both towards our understanding of specificity and kinetics of binding, as well as in assisting in the analysis of the proteome.

Keywords: bioinformatics; protein–protein interactions; transient hetero-complexes; binding-site prediction

Introduction

Protein–protein interactions play a pivotal role in the organization of life. While some interactions form stable complexes resulting in permanent, multi-protein structures, others are of a transient nature. The latter are abundant in signal transduction, protein–inhibitor complexes, antibody–antigen interactions and others.

Structural knowledge on a residue and atom level is one of the keys in achieving a better understanding of these processes. X-ray crystallography

Supplementary data associated with this article can be found at doi: 10.1016/j.jmb.2004.02.040

Abbreviations used: MS, molecular surface; DB, database; AA, amino acids; NR2St, non-regular secondary structures length; ASA, accessible surface area; TF, temperature factor; CE, combinatorial extension; CII, Connolly interface index; SD, surface dots; HI, hole-index; MI, mutual information.

E-mail address of the corresponding author: gideon.schreiber@weizmann.ac.il and NMR are without doubt the best methods to obtain such information. However, they are too demanding to be used to cover the proteome, even for a relatively primitive organism such as yeast, which already shows many thousands of protein–protein interactions.

Computational methods are therefore needed to assist the finding of potential binding sites for a deeper understanding of protein–protein interactions even if no structural data are available for the complex. If the location of protein–protein binding sites is imprinted in the structures of the proteins, the *in silico* work of building a virtual proteome would be greatly facilitated. Experimental evidence supports the hypothesis that this information can be extracted even without the knowledge of the protein-partner. Wells *et al.* showed that random peptides consistently bind the same site on the Fc fragment of human immunoglobulin G.[1] Strynadka *et al.* have shown that two different β-lactamase inhibitors (BLIP) bind exactly the same site on TEM1.[2] These examples suggest the possibility that binding surfaces share common properties which distinguish them from non-binding surfaces. According to this hypothesis, not the whole surface is amenable to be engaged in protein–protein interactions, but only specific areas.

The chemical and structural properties of binding sites have been analyzed extensively. Looking at the distribution of amino acid residues, it was found that polar and aromatic residues are more abundant in interfaces.[3–8] Clusters of hydrophobic residues were also found to assist binding.[4,9] In 90% of the cases examined by Argos *et al.*,[4] the largest or second largest hydrophobic patch overlapped the interface. In addition to hydrophobic interactions, electrostatic interactions between the monomers are formed through hydrogen bonds and salt-bridges; hydrogen bonds appear to be more abundant in non-permanent complexes.[6] Although rare, disulfide bonds have a large stabilizing effect when occurring on interfaces.[10] From a structural point of view, interfaces usually appear in between domains, particularly in large proteins.[7,11,12] Regarding the secondary structure, loops usually appear on the edges of interfaces, contributing about 40% of the interfacial contacts.[13] The shape of the interface is approximately circular.[10]

The evolutionary conservation of amino acid residues is an important property that contributes to the identification of interfaces, albeit not to our understanding of their nature.[14–16] Some studies specifically referred to the conservation of polar amino acid residues, claiming that they provide hot spots and specificity for binding.[14,16]

The analysis of binding sites is complicated by the diverse repertoire of binding partners of proteins, including DNA, small molecules, peptides and other proteins. Protein–protein complexes can be further divided into homo and hetero-complexes. Homo-complexes are found primarily as complexes. Hetero-complexes can be divided into permanent (structural) complexes and transient complexes. Among all protein–protein complexes, the transient ones are maybe the most interesting, as they exist both in the bound and unbound states, with binding having a functional role in regulating biological function. Therefore, it is not surprising that a large spectrum of kinetic and thermodynamic behaviors have been attributed to different transient interactions, ranging from very weak interactions between electron transfer partners to extremely tight ones in enzyme–inhibitor complexes. Other transient hetero-complexes include protein–receptor complexes, antibody–antigen complexes, signal transduction partners, etc.

The varying nature of these interactions is expected to be expressed through the different interface properties. Permanent interfaces are usually larger and more hydrophobic compared to transient interfaces, and homo-dimers are more densely packed than hetero-dimers (in particular antibody–antigen complexes).[10] Therefore, interface properties of each of these sub-classes have to be evaluated separately.

If binding sites indeed differ from the rest of the protein, the development of an interface prediction algorithm is called for, as the ability to map the location of binding sites has many applications both *in silico* and for the experimentalist. Thornton *et al.*[5] divided the protein's surface into patches and ranked them by their probability of forming protein–protein interactions according to their chemical and structural parameters. The parameters applied include the solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area (ASA). The prediction was considered successful for 66% of the proteins. Three other groups tried to predict the amino acid residues that construct the interface, basing their algorithm mostly on sequence information. Shan *et al.*[17] used a neural network to predict the interface residues based on the sequence profile and solvent exposure data. The fraction of interface residues according to their interface definition is 29%; 65% of them were identified as interface. Out of all the residues that the predictor found to belong to the interface, 70% were correct. Casadio *et al.*[15] again used a neural network and a multiple sequence alignment to predict the interface residues. The predictor identified correctly 73% of the interface residues using a relatively generous interface definition with which the interface fraction is 40% of the total surface. Yao *et al.*[18] identified clusters of evolutionarily important residues. Expecting these clusters to overlap with protein-binding sites, a success rate of 69% to 91% was declared, depending on the measure used. Unfortunately, objective comparison between all of these algorithms is difficult, as each study used different interface definitions and criteria for success; further, the predictions were done using different databases (DBs).

Table 1. Summary of the results for all the unbound DB

Protein No.	Protein name	PDB ID	Equivalent bound	Reliability of largest patch	Fraction of interface detected (sensitivity)	Patches predicted
1	Barstar	1a19A	1brsD	1	0.29	1
2	Barnase	1a2pA	1brsA	0.9	0.19	1
3	Tumor suppressor p16ink4a	1a5e-	1bi7B	0.88	0.1	1
4	Acetylcholinesterase	1acl-	1fssA	0.24	0.14	1
5	Plastocyanin	1ag6-	2pcfA	0.7	0.16	1
6	cdc42hs	1aje-	1am4D	0.72	0.3	1
7	rhogdi	1ajw-	1cc0E	0.73	0.24	1
8	fkbp-rapamycin-binding domain	1aueA	1fapB	0.9	0.35	1
9	Trypsin inhibitor	1avu-	1avwB	1	0.29	2
10	Human procarboxypeptidase a2	1aye-	1dtdA	0.54	0.24	1
11	Hydrolase angiogenin	1b1eA	1a4yB	0.69	0.24	1
12	Bifunctional trypsin/alpha-amylase inhibitor (rbi)	1bip-	1tmqB	1	0.27	1
13	Cytochrome *f*	1ctm-	2pcfB	1	0.12	1
14	Granulocyte colony stimulating factor	1cto-	1cd9B	0.36	0.29	1
15	Receptor chey mutant	1cye-	1eayA	0	0	1
16	Calcium-free equine plasma gelsolin	1d0nA	1c0fS	0.67	0.03	2
17	Hydrolase inhibitor	1d2bA	1ueaB	0.92	0.31	1
18	Transferase	1ekxA	1d09A	0	0	1
19	Bovine chymotrypsinogen a	1ex3A	1cgiE	1	0.29	1
20	Neuronal t-snare syntaxin-1a	1ez3A	1dn1B	1	0.06	1
21	Amino-terminal domain of enzyme i from *Escherichia coli*	1eza-	3ezaA			0
22	rgs4	1eztA	1agrE	0.54	0.13	1
23	Enteropathogenic *E. coli* intimin	1f00I	1f02I	0	0	1
24	Coxsackie virus and adenovirus receptor	1f5wA	1kacB	1	0.06	1
25	fk506 binding protein	1fkl-	1b6cA	1	0.2	1
26	Uracil-DNA glycosylase	1flzA	1euiA	0.52	0.19	1
27	Neuronal sec1	1fvhA	1dn1A			0
28	Hydrolase	1g4kA	1ueaA	0.78	0.21	1
29	Radixin ferm domain	1gc7A	1ef1A	0.78	0.06	1
30	Granulocyte colony stimulating factor (rhg-csf)	1gnc-	1cd9A	0.06	0.02	1
31	N-terminal region of p67phox	1hh8A	1e96B	0.5	0.02	1
32	Lipase (EC 3.1.1.3)	1hplA	1ethA	0.07	0.03	1
33	p53 core DNA-binding domain	1hu8A	1ycsA	0.05	0.02	1
34	Interleukin-1 beta	1iob-	1itbA	0.31	0.06	1
35	Actin	1j6zA	1c0fA	0	0	1
36	α-Amylase lysozyme	1jae-	1tmqA	0.5	0.13	1
37	(EC 3.5.1.28) mutant	1lba-	1aroL	0.6	0.24	1
38	Knob domain from adenovirus serotype 12	1nobA	1kacA	0.07	0.03	1
39	Nitric oxide synthase oxygenase domain	1nos-	1nocA	0	0	1
40	Porcine pancreatic procolipase b	1pco-	1ethB	0.6	0.12	1
41	Profiling	1pne-	1hluP	0	0	1
42	Phosphotransferase (hpr)	1poh-	1ggrB			0
43	Papain (EC 4.3.22.2)	1ppp-	1stfE	0.91	0.3	1
44	Streptokinase domain b	1qqrA	1bmlC	0.85	0.32	1
45	Rhogap	1rgp-	1am4A	0.5	0.05	1
46	Selenosubtilisin	1selA	1cseE	0.61	0.27	1
47	Cyclin a	1vin-	1finB			0
48	p120gap	1wer-	1wq1G			0
49	β-Lactamase tem1	1xpb-	1jtgA			0
50	Ribonuclease inhibitor	2bnh-	1a4yA	1	0.04	1
51	Cyclophilin a	2cpl-	1ak4A	0.76	0.23	1
52	Glucose-specific phosphocarrier	2f3gA	1ggrA	1	0.12	1
53	Negative factor (fprotein)	2nef-	1avzB	0.57	0.24	1
54	RalGEF-rbd streptomyces	2rgf-	1lfdA	0.2	0.05	1
55	Subtilisin inhibitor cytochrome *c* peroxidase	3ssi-	2sicI	1	0.24	2
56	(EC 1.11.1.5) mutant	6ccp-	2pcbA	0	0	2
57	BLIP	Personal communication	1jtgB	0.94	0.22	1

The aim of this work is to focus entirely on the analysis of transient protein–protein heterocomplexes and to use the information obtained to develop an interface prediction program. The expression of different properties is compared over binding and non-binding surfaces and how these are manifested in the structure of the unbound proteins, *versus* the structures of the same proteins solved in complex. All properties are defined in a quantitative manner that enabled

Table 2. Summary of the results for the disjoint bound DB

Protein No.	Protein name	PDB ID	Reliability of largest patch	Fraction of interface detected (sensitivity)	No. of patches predicted	Reliability of best patch[a]
1	HIV-1 capsid	1ak4D		0		
2	T7 RNA polymerase	1aroP	0	0	3	0.14
3	Cyclin-dependent kinase 6	1bi7A	0.56	0.08	2	
4	Son of sevenless-1	1bkdS	0	0	3	1
5	Interleukin-1 beta convertase	1bmqB	0.88	0.15	1	
6	α-1,4-Glucan-4-glucanohydrolase	1bplA	0.72	0.19	1	
7	β2-Bungarotoxin	1bunA	0	0	1	
8	Ubiquitin yuh1-ubal	1 cmxA	0.76	0.16	1	
9	Succinyl-CoA synthetase α chain	1cqiA	0.95	0.17	1	
10	Succinyl-CoA synthetase β chain	1cqiB	1	0.08	1	
11	Fibroblast growth factor receptor 1	1cvsC	0.8	0.07	1	
12	PKN	1cxzB	0.33	0.05	1	
13	Aspartate carbamoyltransferase regulatory chain	1d09B	1	0.46	2	
14	Bean lectin-like inhibitor	1dhkB	0.75	0.38	1	
15	Naphthalene 1,2-dioxygenase α-subunit	1eg9A	0.34	0.1	2	
16	Naphthalene 1,2-dioxygenase β-subunit	1eg9B	0.69	0.15	2	
17	Colicin e9	1emvB	1	0.07	1	
18	Elongation factor eef1ba	1f60B	1	0.03	1	
19	Flavocytochrome *c*	1fcdA	1	0.02	1	
20	Sulfide dehydrogenase	1fcdC	0.67	0.2	1	
21	Hydrogenase	1frvA	0.37	0.14	1	
22	Hydrogenase	1frvB	0.92	0.15	2	1
23	β-Lactamase inhibitor protein ii	1jtdB	0.39	0.48	1	
24	Type 1 chloramphenicol acetyltransferase	1nocB	0.05	0.03	1	
25	Chaperone protein papd	1pdkA	0	0	2	1
26	Protein papk	1pdkB	0.87	0.16	1	
27	Karyopherin beta2	1qbkB	0.07	0.01	5	
28	Nuclear pore complex protein nup358	1rrpB	0.25	0.06	2	1
29	*Erwinia chrysanthemi* inhibitor	1smpI			0	
30	Stefin b	1stfI	0.75	0.09	1	
31	Transcription initiation factor tfiid	1tbaB	0.62	0.09	2	1
32	Rabphilin-3a	1zbdB	0	0	2	1
33	*Klebsiella aerogenes* urease	2kauC	0	0	1	
34	Cytochrome *c*	2pcbB	0.55	0.15	1	
35	Human growth hormone receptor	3hhrB	1	0.17	2	

[a] For cases where the largest patch is not the best patch.

us to use them for the computational prediction of binding sites, without any prior knowledge of the binding partner.

Results

The work presented here is divided into two sections. In the first, we characterize quantitative differences between protein surfaces that are involved in protein–protein interactions, and the remaining protein surface. In the second section, we use the information gained to develop a computer algorithm that predicts the location of a protein–protein binding site on the structure of an unbound protein. This work focuses entirely on transient hetero-complexes, excluding antibody–antigen interactions because of their specific nature. The extraction of the different interface properties was executed independently on three DBs. The unbound DB containing 57 structures, the bound DB containing 92 proteins and the disjoint bound DB containing 35 structures. The disjoint DB consists of proteins from the bound DB, which have no homologous structure in the unbound DB. The identity of the proteins in the three DBs is given in Tables 1 and 2. The set of bound structures that have an analogous form in the unbound DB is referred to as the homologous bound DB. The comparison of the three DBs verifies the stability of the results relative to differences in the DB, and demonstrates the differences between the bound and unbound states.

For the statistical analysis of binding *versus* non-binding surfaces, a protein's surface was sampled using circles with a radius of 10 Å around anchoring dots, which are uniformly distributed over the monomer's surface (0.1 dot/Å^2). Circles with Connolly interface index, $CII \geq 0.7$ were considered interface, and those with $CII = 0$ as surface. The rest (boundary) were not used for data retrieval. It is important to note that only surface residues as determined using Connolly's MS dots program† were used for the analysis and later for the prediction.

† http://www.biohedron.com/msp.html

The chemical composition of binding sites

Amino acid propensities in binding sites

The amino acid preference of protein–protein interfaces has been analyzed previously.[3,6,7,17,19] Therefore, this property is a good indicator to validate the data extraction method used by us in comparison to other methods. The amino acid propensity shown in Figure 1 agrees well with those presented by Thornton[19] and by Janin,[7] although they evaluated the ASA contribution of amino acid residues while we counted them. Tyr, Met, Cys and His are the most favored on the interface, Thr, Pro, Lys, Glu, and Ala are least commonly found on the interface. Janin also found Arg to be abundant in interfaces, while we did not. No significant difference was found between the bound *versus* the unbound DBs for this property.

In addition to the analysis of surface residues, we determined the distribution of individual atoms on the protein's surface (normalized to the amino acid distribution). The difference between this analysis and the former is that here only exposed atoms (and not amino acid) were analyzed. The results shown in Figure 2 emphasize the significant role of aromatic functional groups in the interface. Almost all of the atoms displaying a significant difference in the interface are the benzene carbon atoms of Trp, Phe and Tyr. Interestingly, all atoms with differential interface propensities are carbon atoms.

A third method for analyzing the protein's surface composition is to group all surface atoms according to their chemical character. All atoms (including backbone) were grouped into five categories: positively or negatively charged, aromatic, hydrophobic or polar. Analyzing the data according to chemical character shows little significant differences between the binding site and the rest of the protein (Figure 3). The only unequivocal conclusion that can be drawn is regarding the higher frequency of atoms belonging to aromatic functional groups, which are preferred in binding sites.

Charged atoms seem to have a preference for non-interfaces. However, this is more significant at the amino acid level. An interesting conclusion is

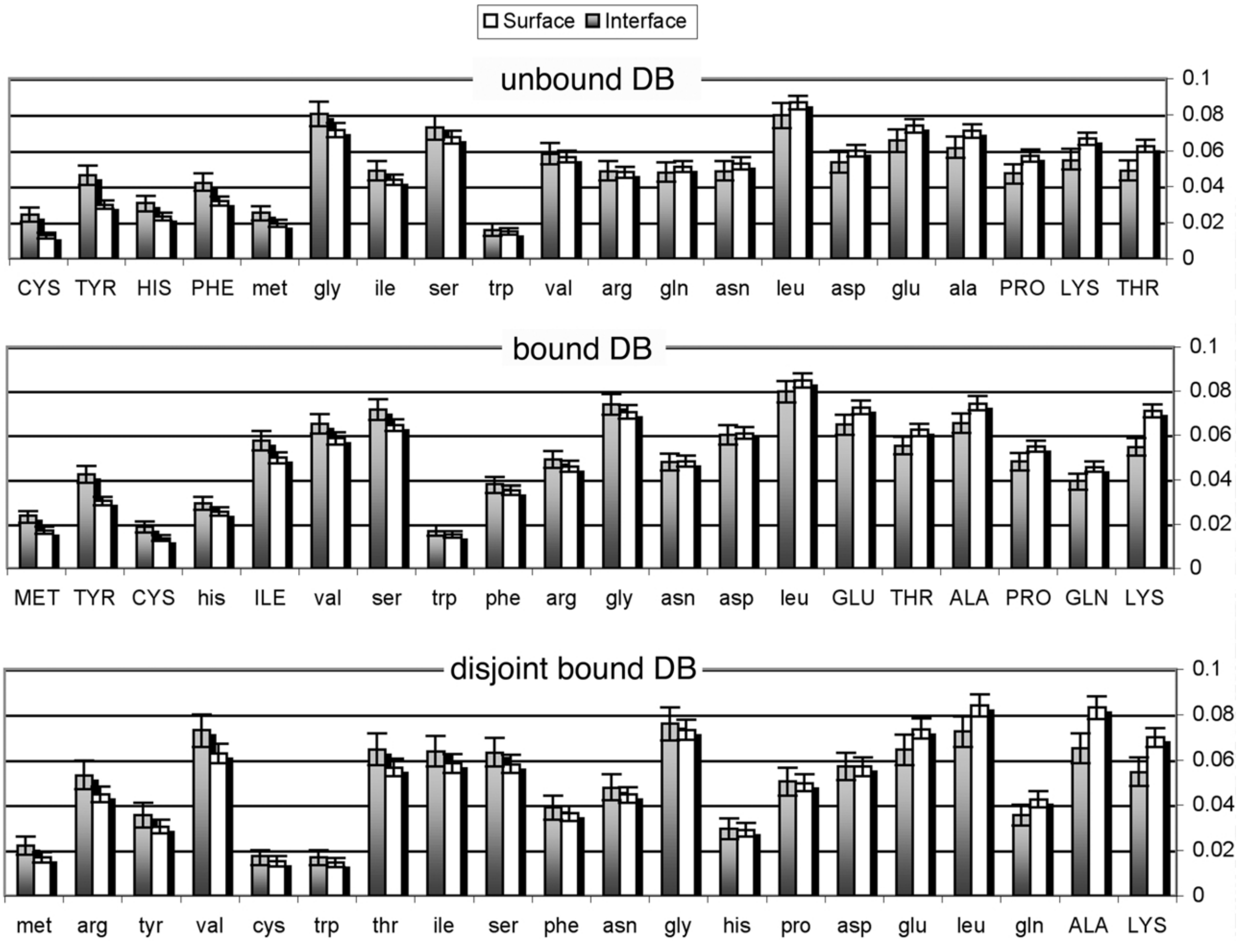

Figure 1. The amino acid distribution in the unbound, bound and disjoint bound DBs. The error bars are the 70% confidence intervals. The names of the amino acids with non-overlapping error bars are marked in upper case. The bars are sorted from right to left in an increasing preference for binding sites. The distribution is stable over the three DBs. Pro, Lys, Thr, Ala and Glu appear to be preferred on the regular surface while Met, Tyr and Cys are favored on binding surfaces.

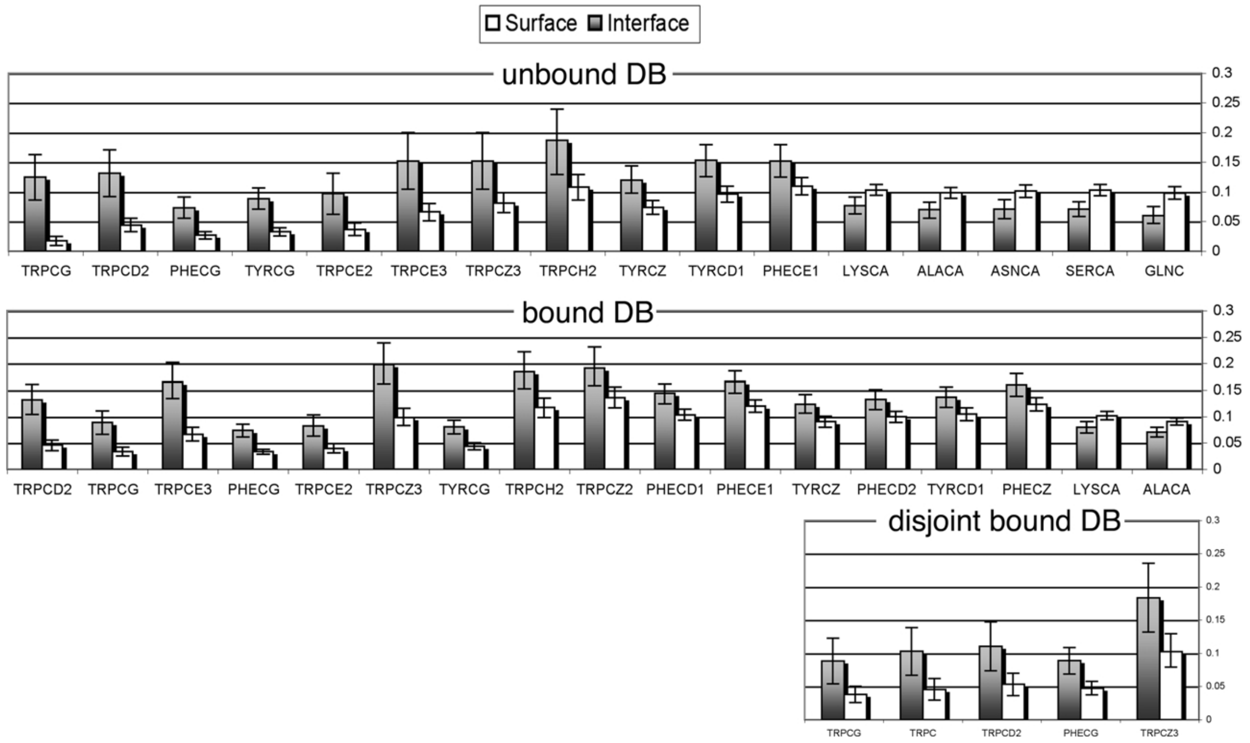

Figure 2. The distribution of atoms on protein surfaces. This is similar to the amino acid distribution, but the surface properties of interfaces are analyzed in relation to property of the individual atoms displayed on the protein's surface. Only atom types with a significantly different propensity for interfaces are shown. Results are given for the unbound, bound and disjoint bound DB. The frequency of the atoms is normalized by the relevant amino acid frequency. Atoms that participate in aromatic rings seem to play an important role in interfaces.

that hydrophobic atoms have no apparent preference for interfaces.

Pairwise AA distribution

Cooperativity between different attributes of the protein's surface is believed to be important for binding. For some complexes, "networks" of interactions between the monomers have been observed; i.e. interactions that involve more than a single atom from each monomer (charged polar or hydrophobic).[20,21] For those, we would expect to find repeating patterns of atoms on the binding surface. We focused on patterns of spatially close, neighboring amino acid residues, which exhibit

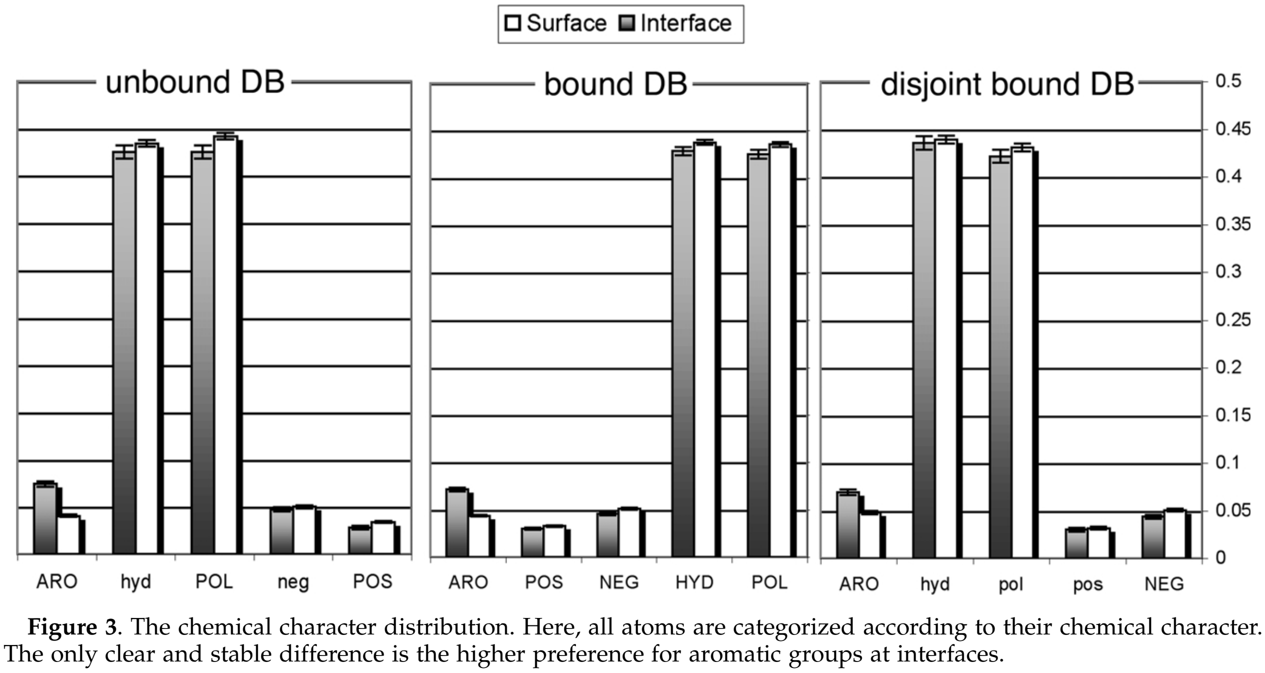

Figure 3. The chemical character distribution. Here, all atoms are categorized according to their chemical character. The only clear and stable difference is the higher preference for aromatic groups at interfaces.

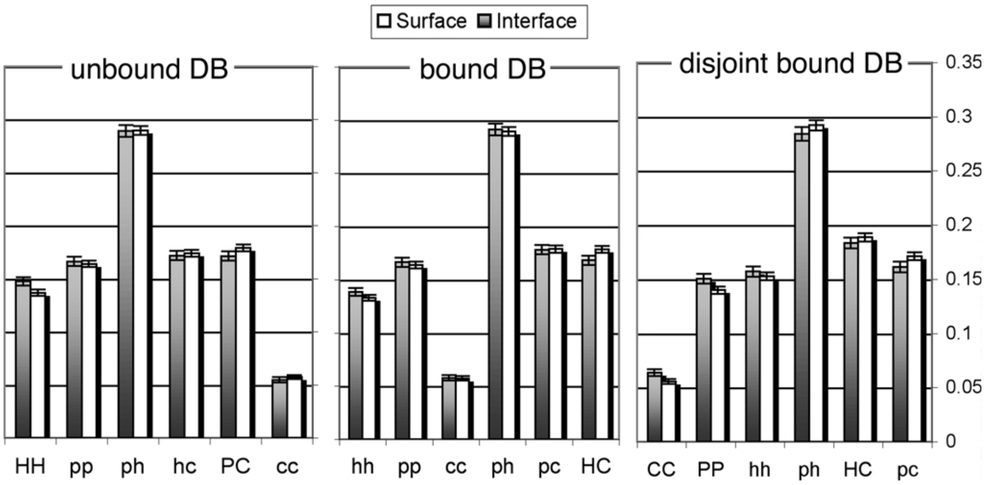

Figure 4. Amino acid pair distribution. A pair of amino acid residues is defined as any two residues with a C^α distance of under 6 Å residing within a circle. Then, all pairs were categorized according to chemical character, charged (C), hydrophobic (H) and polar (P). Capital letters are used for statistically significant pairs.

certain correlated properties. Two amino acid residues were defined to be neighbors if the distance between their C^α atoms was smaller than 6 Å. A pair of neighboring amino acid residues was considered to be part of the interface if the C^α of both residues in the pair appeared within an interface circle. The possible pairs were divided into six distinct categories, according to the hydrophobic/polar/charged nature of the amino acid residues in the pair (Figure 4).

Though the differences seem to be minor, there is a stable preference for pairs of hydrophobic amino acid residues. This appears to be an expression of the fact that interfaces tend to overlap large hydrophobic patches.[21] The same pairing tendency also appears for polar amino acid residues, suggesting the existence of polar patches, similar to the hydrophobic ones. Physically, this would suggest a non-random distribution of residues in interfaces, with a preference to cluster hydrophobic and polar residues separately.

Evolutionary conservation

It has been long suggested that functional residues tend to be evolutionarily conserved. Shan *et al.*[17] have suggested this to be true also for protein–protein interfaces. Here, we used a simplified version of this algorithm. Each amino acid was set to the value that appears in the diagonal of the PSI-BLAST output matrix and the distribution of these values was explored. Using our interface definition, the higher degree of conservation compared to the rest of the protein's surface is still apparent.

Geometric properties

Secondary structure

Proteins having both β-strands and α-helices were selected from our DBs (46,79 and 30 for the unbound, bound and disjoint bound DBs, respectively) and their secondary structure extracted using the program: PROMOTIF.[22] The statistical distributions of secondary structures are displayed in Figure 5. Most striking is the preference of β-strands for interfaces, and at the same time, the disfavoring of α-helices. This preference has major implications towards the structure of interface regions, as β-strands form flat areas with three-dimensional connectivity, while α-helices form cylindrical perturbing surface structures with the three-dimensional structure following the one-dimensional sequence. The preferences for β-strands and α-helices is the reverse of those found by Thornton *et al.*[6] This may be the result of the basically different DBs used. While we completely excluded homodimers and antibody–antigen complexes, these types of complexes constitute about two-thirds of Thornton's DB. Furthermore, the constraint demanding both sheets and helices to appear in each of the proteins analyzed is unique to this work.

The length of non-regular secondary structures

The flexibility of polymers varies with their length. Below a certain length, termed the persistence length, a polymer appears to be rigid. The persistence length of polypeptides is usually in the range of five amino acid residues (AA) to 10 AA, depending on their specific composition.[23] Therefore, there should be a difference in the nature of the unstructured regions of the protein as a function of their length.

Here, each AA that is not part of a β-strand, α-helix or a 3,10-helix is considered to be a non-regular secondary structure (NR2St or "loop"). The distribution of the NR2St length is presented in Figure 6. There is a clear preference for interfaces to appear in regions of long loops. The mean loop length within the interface is 11.17 AA. To evaluate the significance of this value, we

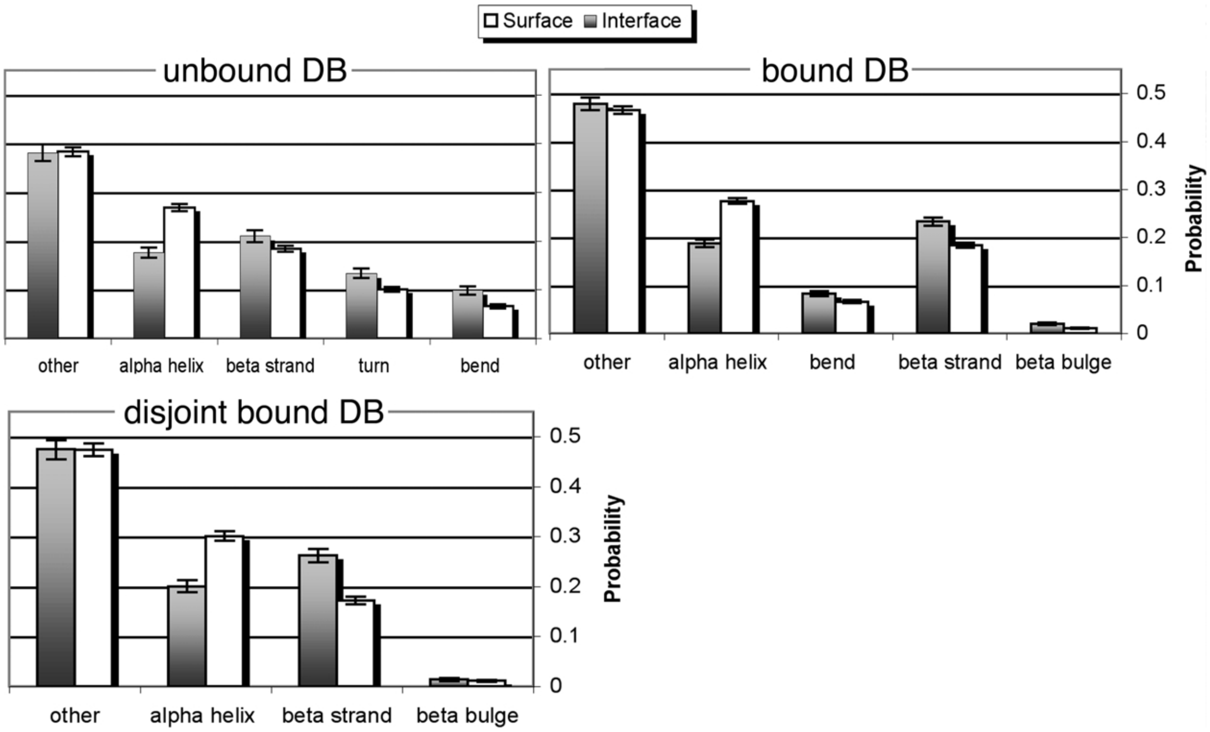

Figure 5. Secondary structure distribution. Only statistically significant bars are shown for each DB (unbound, bound, disjoint bound). This distribution was extracted only for proteins that contain both helices and sheets. The preference for β-strands over α-helices within interfaces is clearly the most outstanding result.

randomly selected 1000 samples equal to the number of interface samples, from the non-interface samples. For those, the mean loop length was found to be 8.75 ± 0.15. The *p*-value for that result is bellow matlab's precision. These results show that long, flexible NR2St are preferred to be located within binding sites, apparently giving greater flexibility for interfaces.

Sequence distance score

The amino acid distance in three dimensions is not directly related to the one-dimensional distance between residues along the peptide chain. Here, we wanted to evaluate whether the distribution of all the sequence distances within 10 Å circles on the protein's surface is similar to that of interface surfaces (Figure 7). The interface is clearly under-represented at the shortest sequence distances (<6 AA). No significant differences could be found at longer sequence distances. This result is in line with the above-mentioned regarding secondary structure preferences. A preference for β-strands in interfaces is expected to show a preference for longer distances. Contrary to all other graphs shown here, the confidence intervals in Figure 7 are the standard deviation of the frequency values when analyzing each protein separately.

Specific information obtained from crystal structures

The coordinates of an X-ray solved protein structure include information concerning the *B*-factor for each residue (also referred to as the

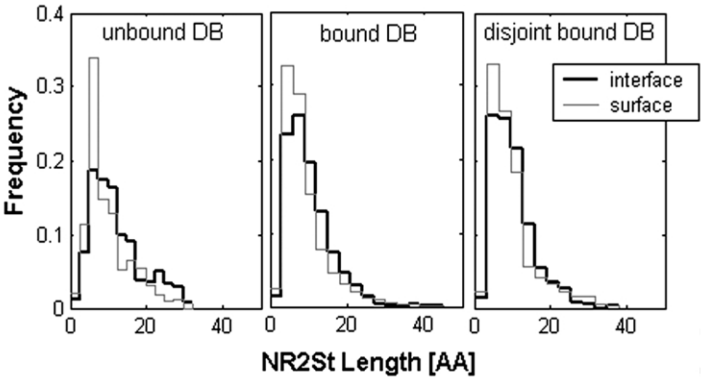

Figure 6. The distribution of non-regular secondary-structure lengths. NR2St are defined as regions of the protein's backbone that are not part of helices or strands. Longer regions are significantly preferred within interfaces.

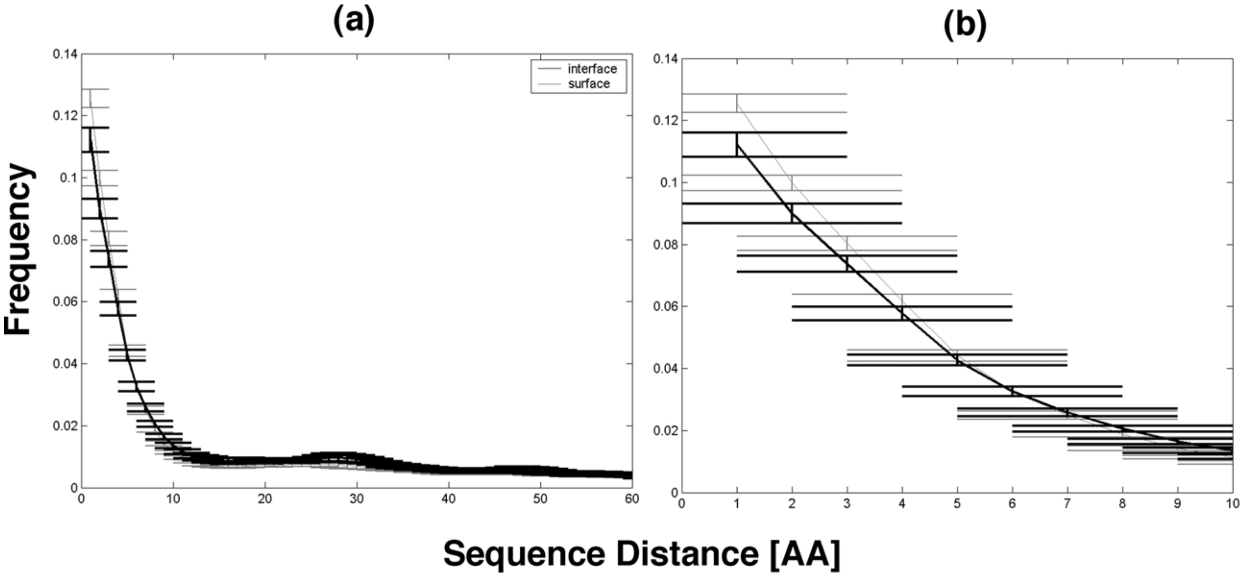

Figure 7. The distribution of the sequence distance. This is calculated as all distances along the polypeptide chain within a circle. The error bars are the standard deviation of the values over all the unbound proteins. A zoom into the shortest distances (up to 10 AA distance) is given in (b).

temperature factor (TF)), as well as the location of water molecules surrounding the protein. Both of these were found to differ significantly between binding and non-binding surfaces.

Temperature factor

The TF is a measure of the oscillations of an atom around its mean position. It has long been recognized that in a complex, interface residues have lower TFs than the exterior of the protein in general.[6] Here, we analyzed whether the same can be said for the unbound form (Figure 8). Indeed, the TF distribution is somewhat lower for the interface already in the unbound state. After complexation, the interface atoms become buried, and their *B*-factor drops further. Two statistical tests were performed to evaluate the measure of significance of this result. The Kolmogorov–Smirnov test comparing interfaces *versus* the rest of the surface on the unbound DB gave a *p*-value of 1×10^{-45}. The mean TF of a thousand samples randomly selected

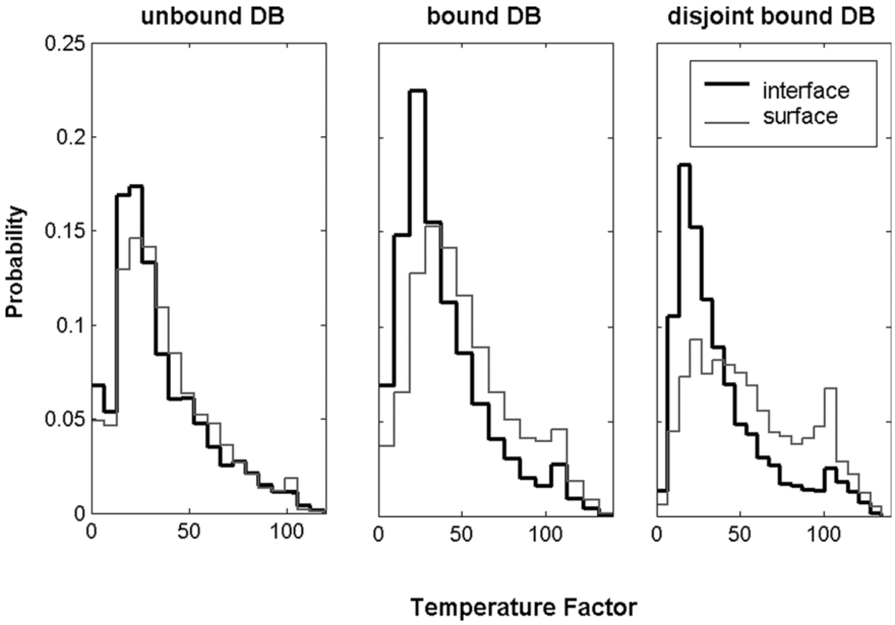

Figure 8. The *B*-factor (TF) distribution over the unbound, bound and disjoint bound DBs. *B*-factor values are retrieved from the structure coordinates. The interface atoms appear to have lower *B*-factors already in the unbound state.

from the non-interface surfaces was found to be 34.7 ± 0.2. In comparison, the mean TF of the interface sample is 31.5. The *p*-value for randomly selecting a sample with this TF is 2.2×10^{-45}. The number of proteins in the DBs for which TF information is available is 47 unbound, 88 bound and 34 disjoint unbound.

The position of water in crystal structures

Proteins reside and interact in aqueous solutions. Thus, a reasonable way for a binding surface to advertise its position is by altering the structure of the surrounding water-cage. If this indeed is the case, some of that information might be reflected in the static location of water molecules found in crystal structures. Therefore, we analyzed the amount of bound water surrounding the different regions of the protein. As the number of water molecules is also a reflection of the resolutions of the X-ray structure, the water score is the normalized and not the absolute number of water molecules within a circle. A standard normalization procedure was applied: reducing the mean of all the circles of a protein and dividing them by their standard deviation.

Examining Figure 9 reveals that both in complex and as unbound proteins, interfaces are solvated by more bound water. In the complex, these molecules are probably caught in the interfacial space. More interesting is the meaning of this finding for the unbound proteome. If in fact the water's structure or dynamics is different near interfaces, this would suggest a means to transmit the location of an interface over the first water shell. In other words, this information would be available to an incoming protein partner.

The number of proteins for which the information of the location of water was available is 40, 67 and 22 for the unbound, bound and disjoint-bound DBs, respectively. Using the Kolmogorov–Smirnov test to measure significance, the *p*-values were 2×10^{-35}, zero (within matlab's precision) and 3×10^{-25}, respectively. The three *p*-values extracted from the estimation of the mean were within matlab's precision.

The high water content near binding sites may be rationalized by the higher degree of order found for the amino acid distribution, forming areas of polar and hydrophobic patches, with the polar patches providing the coordination points for water. Intuitively, having a higher number of ordered water molecules near the interface is disadvantageous for complex formation, as the water has to be removed. However, comparing the bound and unbound DBs reveals that the relative amount of water in the interface is higher also in the bound form, suggesting a specific role for these water molecules also in the bound state.

Since both a preference for crystallographic water, and for a low *B*-factor were found at interfaces, one may wonder whether this is not due to the potential of binding sites to be involved in crystal contacts. If this is the case (which we cannot prove or disprove for the DB as the amount of work involved would be enormous), then we measure a secondary effect. Nevertheless, the conclusion would be similar, as the formation of crystal contacts at a certain position would suggest physico-chemical differences of the involved surface patches. In such a case, these two properties would clearly be dependent.

ProMate: predicting the interface location

Interface characterization was done bearing in

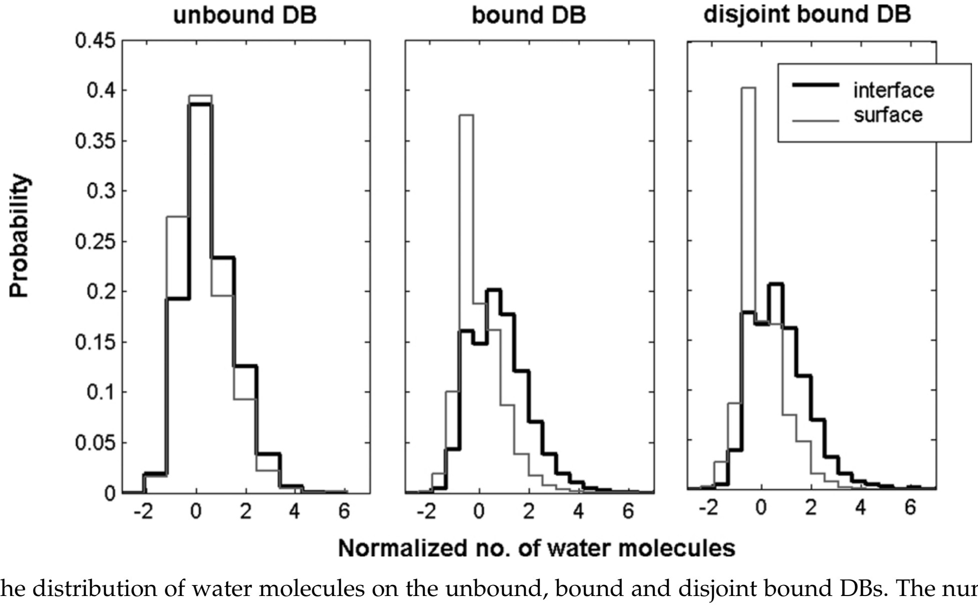

Figure 9. The distribution of water molecules on the unbound, bound and disjoint bound DBs. The number of water molecules is counted for interface and surface circles, and normalized to the mean and standard deviation of all the water molecules around the protein. Both in the bound and in the unbound state there is more bound water at the interface.

mind the goal of computing a protein interface prediction program. For a detailed description of how ProMate works, see Methods. In summary, the prediction process starts with a stage of preprocessing, where the training DB is analyzed and the specific properties are extracted. Then, for a given new protein, a set of surface dots (SD) at a density of 0.1 Å^2 is extracted. The interface probability of each surface dot is estimated according to the distribution of each property within a 10 Å radii circle around it. All of these values are combined to give a final score. These scores go though a smoothing process that fixes the circle's score according to the scores of its neighbors. From these, a set of amino acid residues that received the highest 10% of the scores is selected, given that they scored at least 0.7. The final stage is the extraction of predicted interface patches. A patch is a set of such predicted interface amino acid residues, where each residue is not more than 13 Å C$^\alpha$-distant from all the others. The minimum patch size is 2 AA. The predicted interface is the largest of all the patches.

Choosing the combination of the final score

Significant scores were extracted for 13 different properties that were shown to be relevant for the distinction between binding and non-binding surfaces. Namely, the amino acid and atom distribution, chemical character of atoms, pairs of amino acid residues, evolutionary conservation, sequence distance within a circle, secondary structure and length of NR2Sts. Crystallographic data were also used whenever available (distribution of *B*-factor and bound water). In addition, several other scores that are not mentioned here in detail were tested. Such was the hydrophobic patch score, using the program QUILT.[24] This score is based on the finding that the largest hydrophobic patch tends to overlap the interface. Here, we used both the size distribution of the hydrophobic patches as well as its rank. Another score tested was the domains' score that is based on the assumption that, for large proteins, interfaces tend to appear between domains. This score did not prove to be very useful.

With some of the scores being mutually dependent, an optimization procedure is required to choose the combination of scores that produces the best interface predictions. However, one has to be careful not to over-fit the data. Therefore, the success of a certain combination of scores was evaluated using a cross-validation method. The DB containing all unbound proteins was divided into 11 samples. Each sample defines a test set containing five or six proteins, and a training set containing all the other proteins from the unbound DB. All the 11 test sets were disjoint and altogether they covered the whole unbound DB. For each of the 11 samples, the probabilities were extracted for the training set, normalized regarding the specific scores combination and run on the test set for prediction. A prediction was considered successful if it was reliable. Specifically, if at least half of the amino acid residues that were declared as interface by the predictor were truly so. Only the biggest patch was considered as the predicted interface (see Methods). The success rate of the score combination is described by the mean, and standard deviation of the success fraction over the 11 samples. This optimization method is problematic, in the sense that the same set of 11 samples is used for all combinations, and thus the optimized combination might be biased for it. This was done due to computational resource limitations, but we do not believe that it would affect the final results significantly.

Enumeration over all possible combinations was not feasible due to computer power limitations. Thus only a subspace of the 2^{13} possible combinations can be scanned thoroughly. The goal of the following analysis was to increase the chances that the best combination is in this subspace. The general scheme is presented here, and is described in detail in Methods. In the first stage, four scores that from their separate analysis seemed most important, and intuitively least dependent were fixed, while all other scores where enumerated. A comparison of the success rate of each of the non-fixed scores (Figure 10(a)) helped us to evaluate its contribution. From this analysis, we can objectively conclude on which subspace we should focus. The chosen subspace is the one keeping the scores that had the highest contribution in Figure 10(a) fixed. For comparison, the same analysis of the contribution of each score is presented in Figure 10(b).

Explicitly, at the first stage, four scores were fixed (atom distribution, water, sequence distance and hydrophobic patch rank), and all the possible combinations of the remaining nine scores were enumerated (512 runs). The different combinations of scores overlapped with one another, as with each run only one score was altered. The many pairing combinations that were produced (these are combinations that are identical except for a single score) were used to learn about the contribution of each score to the predictor's success. The relative contribution of all the variable scores towards the final score is summarized in Figure 10(a). The contribution of each score at all combinations is plotted using the matlab boxplot function. The box is formed between the lower and upper quartile, with a line showing the median. The notches are a robust estimate of the uncertainty about the mean. All the data values that fall outside the box are displayed using a plus (+) symbol. From this plot, the scores that appear to be most constructive for the goal of interface prediction are the NR2St, evolutionary conservation, chemical character and, to a lesser extent, the amino acid pairs distribution. At the second stage, the first three scores were fixed, and all the rest were enumerated. The results for the second step of optimization are shown in Figure 10(b). The

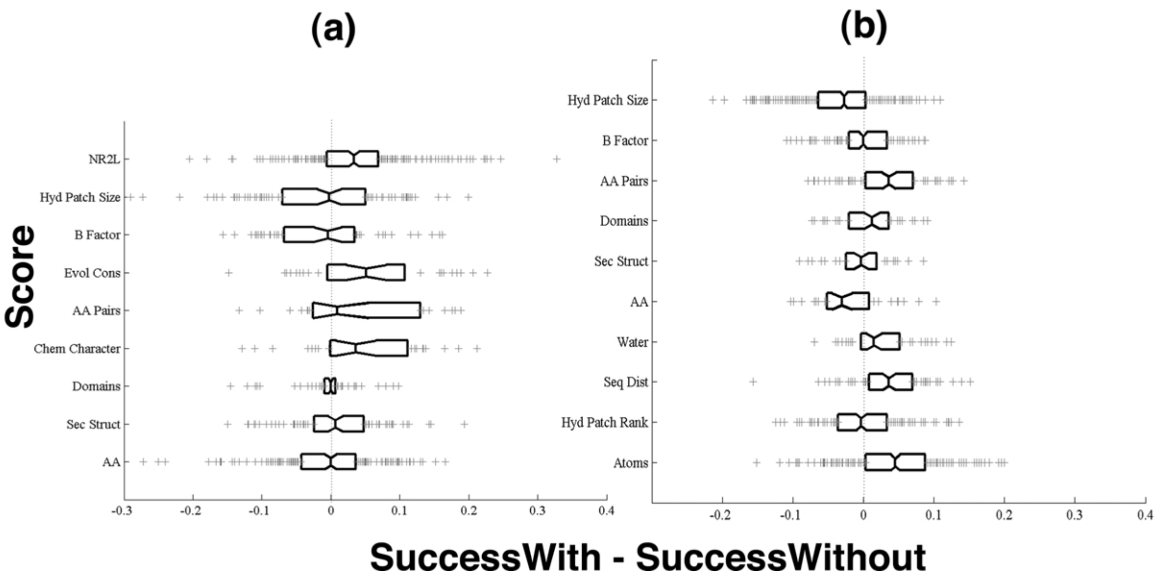

Figure 10. Finding the best combination of scores. The DB containing all unbound proteins was divided into 11 samples, such that the test set containing five or six proteins each is not used for training. For each of the 11 samples, the probabilities were extracted for the training set, normalized regarding the specific scores combination and run on the test set for prediction. A prediction was considered to be successful if it was reliable (over half the amino acid residues predicted as interface are correct). Only the biggest patch was considered as the predicted interface (see Methods). In the first step, the predictor was run over all score combinations shown on the *y*-axis of (a), with four scores being fixed (hydrophobic patch rank, sequence distance, atoms and water). The difference in the mean success rate over the 11 test sets was calculated for each pairing combination. The distribution of these values is presented using a boxplot. The vertical lines of the box are the lower and upper quartiles, and the median. The notches are the estimated range of the mean. All samples that fall outside the box are showed using a plus (+) sign. The scores that consistently improve the predictions according to this plot are evolutionary conservation, length of NR2St chemical character and amino acid pairs. The first three were chosen as fixed for the second phase of enumeration. Its results are shown in (b). Scores that are consistently improving at this phase are amino acid pairs, water, sequence distance and atoms. The best scores from the first and second step were chosen to be used for the predictor.

scores giving the highest overall contributions are amino acid pairs, sequence distance, water, and atoms distribution. An interesting case is the score for amino acid pairs, which appears to have a significant contribution only in the second stage (Figure 10(b)). This suggests that it has dependencies with some of the fixed scores of the first stage (presumably the atoms scores), which makes its contribution less significant. The domains score, which was not used at the first step of optimization, seems also to have some contribution. Surprisingly, the amino acid distribution does not contribute much. This may be due to the high dependency of this score on both the evolutionary conservation and chemical character scores. The optimization method clearly shows that certain scores, which gave statistically significant results in the interface definition section, are seemingly mutually dependent, and thus lose their importance upon optimization. Choosing the atoms score in the first stage probably forced the convergence to a certain subset that does not contain the amino acid distribution score. Starting from the amino acid score may have affected the final combination, but the quality is not expected to improve significantly.

The optimization method used here ended up in an optimal scores combination that is probably a local optimum. Fixing different three scores in the beginning might have ended up in a somewhat different combination. Yet, repeating the optimization twice, while fixing different scores, examined the contribution of each score with respect to a group of other scores. Therefore, Figure 10 supplies a justification for each of the scores that appear in the final optimal combination. Some dependencies might still exist between two non-fixed scores, but they are reduced by the normalization to their actual distribution in the training DB.

Evaluating the success of the optimized ProMate score

The most successful score combination contains the following scores: NR2St, atom distribution, amino acid pairs, evolutionary conservation, chemical character, water, sequence distance, hydrophobic patches rank, and secondary structure. This combination successfully predicts the correct interface for 36 out of 51 proteins (Table 1, showing the largest patch). For the remaining six proteins in the unbound DB, no interface was found at all. One has to remember that, as explained above, the predictions are done without using the predicted protein (plus four or five others) in the training set. For four proteins, the predictor found two interface patches. For all the rest, only one patch was found. The results given in Table 1 are of the largest patch found. For comparison, it would be worth noting that predicting the interface using the four scores that were chosen to be fixed at the first optimization stage (Figure 10(a)) resulted with 44 predictions, of which 25 were successful. For the fixed combination of the second optimization stage (Figure 10(b)), 17 proteins were predicted successfully out of 29 predicted interface patches. All these numbers have to be compared to a random model to evaluate their significance. To create a random model we reshuffled the predicted scores over the amino acid residues before extracting the patches and then checked the success rate. Repeating this procedure ten times gave, on average, a prediction for 46.2 ± 1.3 proteins (the number of proteins for which a patch could be found). Out of these, the interface was predicted successfully for 13.0 ± 3.1 proteins (in the random model many predictions resulted in multiple patches, only the largest one was taken for the statistics).

The quality of the predictions is best appreciated on the structure of the relevant proteins. Four examples of the prediction outcome are shown on the relevant protein structures in Figure 11. From the Figure, one clearly sees that not all the interface was predicted, but that the predicted part fits the interface well. This is a result of the method used, which optimizes precision, rather than coverage of the whole interface. In Figure 11(d) we colored the protein also according to the full score range, from blue (low) to red (high). Doing so increases the size of the predicted interface, however, it may, in some cases be more confusing to the viewer. In the web version of ProMate we supply both the patch results as given in Figure 11(a)–(c) and the full scale score shown in Figure 11(d).

Using the probabilities extracted from the unbound DB, the predictor was run (using the same score combination as for the unbound DB) on the 35 proteins of the disjoint bound DB. For 20 proteins, a successful prediction was found, for three proteins, no interface was detected, and the rest failed. Taking the best patch (instead of the largest one) increases the success rate to 24 out of 32. The results for this DB are summarized in Table 2. The disjoint bound DB was not used during any stage of data extraction or optimization. Thus, the importance of this test is to show that the same rate of success is obtained in general, and is not a result of over-fitting the data. One might argue that this set contains bound proteins, and thus it is not an adequate test. Since this is only one of the proofs for the validity of ProMate, and since the differences between the bound and unbound DB were found to be minor for all the properties used in the optimized scores combination, we consider this set as a valid test.

Discussion

We initiated this project because we suspected that binding sites have some specific properties, which distinguish them from the rest of the protein's surface. Therefore, by identifying these properties, it may be possible to design an algorithm that is able to find these locations on the unbound proteome. The work focuses entirely on transient hetero-complexes, which are stable and functional, both in the unbound and bound forms. It was shown that the interface composition of the latter is significantly different from that found for permanent homo and hetero-complexes, which constitute the ternary structure of a protein. Antibodies were excluded entirely, because of their specific binding mode optimized through rapid evolution.

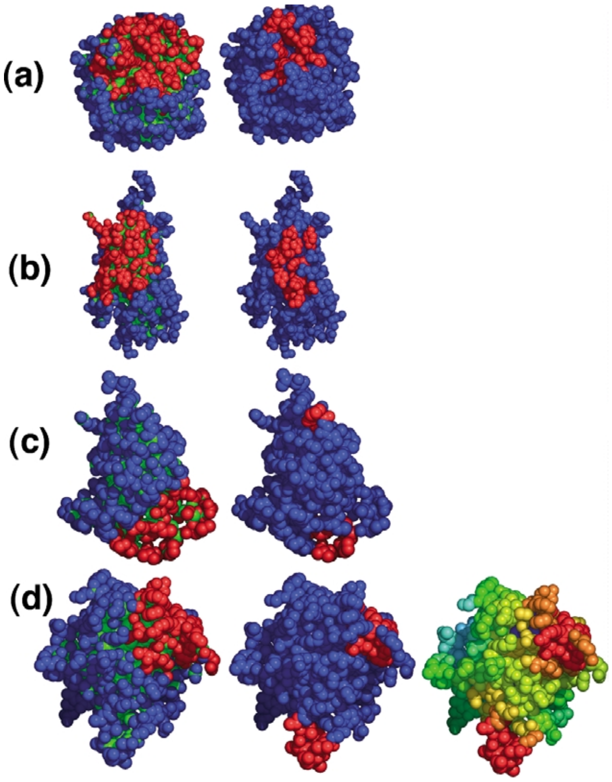

Figure 11. Four predicted interfaces from the unbound DB are drawn on the relevant protein structures. For each protein the true interface (red = interface, blue = surface, green = internal) is given on the left. The rightmost presents the predicted interface patch (in red). (a) 1ex3A, (b) 1ajw, (c) for 3ssi two interface patches were predicted. While the structure of the complex contains only one partner, mutagenesis studies have confirmed the second interface as well. (d) For 1avu, two patches were predicted, the larger one fits the experimentally determined. Here, we also added a third picture showing the full score given by ProMate before the circle-clustering step. The size of the predicted interface growth is considerable.

Starting from a well-defined problem

Two parts of this work, for which one would expect to find a consensus in the field, turned out to be rather inconsistent: how is an interface defined, and how to gauge its successful prediction. In simple words, what exactly are we looking for? Two common interface definitions are presently accepted. The first is based on the distance between the monomers in the complex. Examples can be found in the work by Zhou & Shan,[17] where a 5 Å distance between heteroatoms was used, or in the work by Fariselli *et al.*,[15] which used a 12 Å distance between C^{α} atoms. The second class of definitions is based on the change in the solvent-accessible surface upon complexation.[4,19] Both definitions are very similar in practice, and seem to suffer from the same bias. They mark a non-consecutive region of the protein's surface as interface. Some concave regions are not considered an interface even when they are buried within the interface. In Figure 12(a) (left), the surface of ribonuclease inhibitor (2bnh-) is colored using the ASA definition. The "holes" in the interface are seen clearly. On the right, the interface of the same protein is colored using a heuristic procedure to fill up the holes in the interface (see Methods). This result is a continuous interface definition. Indeed, any protein surface consists of protruding and indenting areas (see Figure 12(b)). We were intrigued to see whether the amino acid composition between the holes and the knobs on the protein surface are identical, and found that this is not the case. The graph presented in Figure 12(c) shows that the amino acid residues can be divided into four groups. The hydrophobic amino acid residues (Val, Cys, Ile, Met, Leu) have a monotonically decreasing preference for the different regions as the extent of solvent exposure increases. The hydrophilic amino acid residues (Ser, Pro, Asp, Asn, Glu, Glu, Lys) show exactly the opposite trend. The aromatic amino acids seem to prefer holes on the surface of the protein, where both the movements of the rings are facilitated, and their hydrophobic chain is protected from the water. Glycine is somewhat different from the rest, as it has a preference for the protein's

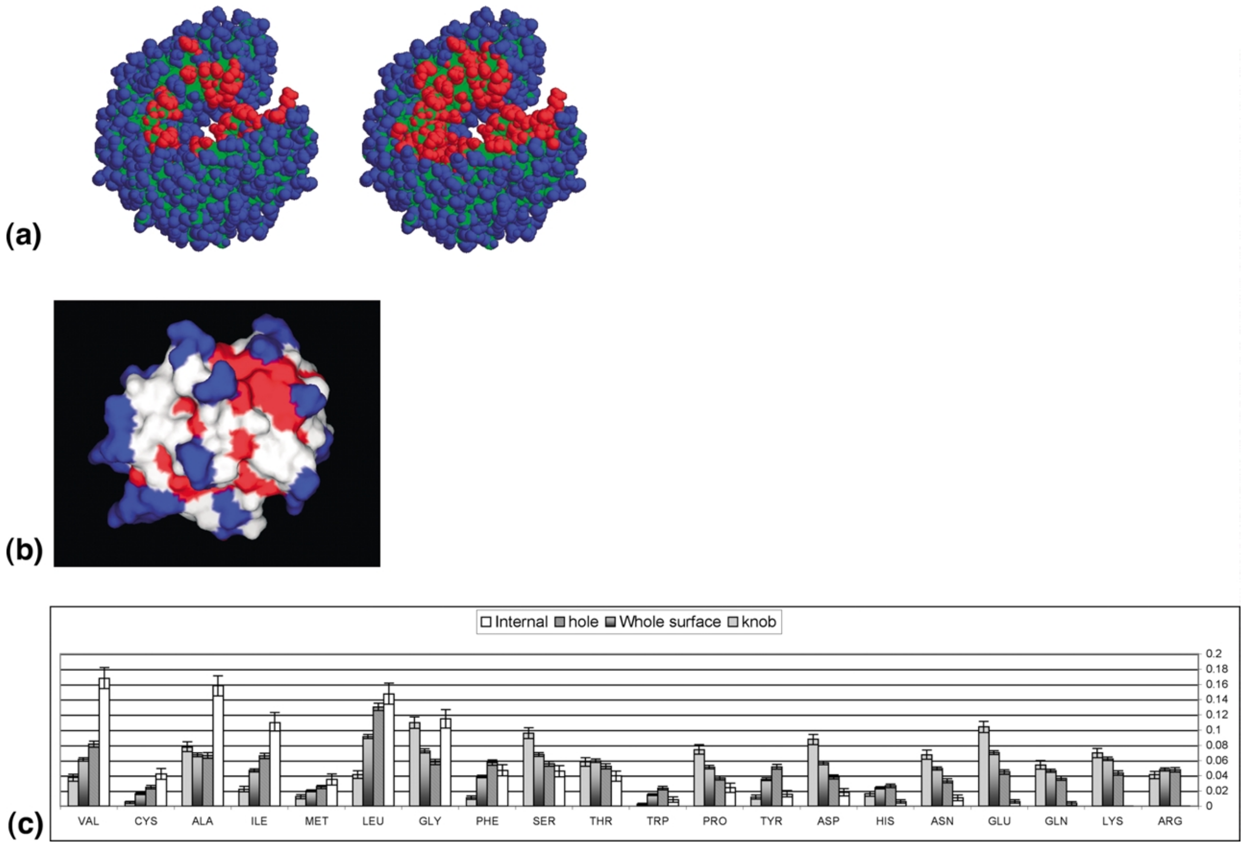

Figure 12. Amino acid distribution in concave and protruding surface areas (knobs and holes). (a) The Figure (left) is an example of an interface extracted for ribonuclease inhibitor (2bnh) using the simple ASA method (with Connolly's MS dots program). Interface atoms are colored in red, surface in blue and internal in green. Some non-interface atoms that are surrounded by interface atoms and therefore are buried within the interface can be seen. In the Figure (right), these atoms were added to the interface definition (see Methods). (b) Trypsin inhibitor (1avu) is colored according to its knobs (in blue) and holes (in red). The white surface is considered intermediate region ($0.2 < HI < 0.3$) and was not used to produce the graph in (c), which shows the amino acid distribution in protruding and concave surfaces in comparison to surface and buried amino acid residues. The bars appear (from left to right) in order of decreasing solvent exposure. Hydrophobic and hydrophilic amino acid residues seem to have a consistent trend with respect to their exposure. Aromatic amino acid residues prefer holes on the surface and glycine is more frequently seen either in internal regions or in knobs on the protein's surface.

interior, but if on the surface it is mostly found in knobs. Evaluating whether the amino acid distribution between knobs and holes is different for interfaces *versus* the rest of the surface did not result in a clear conclusion, as the data were too noisy.

Properties of binding surfaces

The analysis of the unbound and bound DBs was done in a quantitative, comparative way. Obviously, an interface prediction program should identify the location of the interface on the unbound structure. Therefore, we performed all of our analysis on the unbound protein DB. This gave us the opportunity to compare these results with an analysis of the same proteins in their bound form. Despite the structural changes accompanying complexation, all the interface characteristics examined were found to be similar for the bound and unbound DBs. These include the chemical composition, a lower *B*-factor, a higher water content and structural preferences (preference for β-strands and NR2St). Failing to find differences between the two DBs may be attributed to the method used for data collection. The summation of properties over 10 Å circles gives a low-resolution picture of the binding sites, one that seems to be constant for the bound and unbound forms.

The emerging picture of the structural–chemical character of a binding site can be summarized as follows: the binding site is stretched over multiple amino acid chains in two dimensions (β-sheets and longer NR2St), and less on α-helices or short consecutive stretches of amino acid residues. The strong preference for β-sheets may be explained by their ability to form densely packed structures when placed one against the other, thus having a higher potential for intermolecular bonds formation.

Chemically, the amino acid propensities were similar to those reported previously. The distribution of pairs of amino acid residues follows mostly the single distribution pattern, except for hydrophobic and polar pairs, which are more abundant. While the hydrophobic content of interfaces is similar to that of non-interface regions, hydrophobic residues tend to form larger clusters. Together, these data show that interfaces are characterized by clustering effects of polar and hydrophobic residues into patches, but not by a higher hydrophobic or polar content. Using crystallographic data, we learned that the *B*-factor of residues located in the binding surface is lower and that the water content is higher already in the unbound state of the protein. This might be an expression of a preference for crystal contact formation at interfacial regions, or a clever way for long-range advertisement of the location of the binding site (in addition to electrostatic steering). If so, this may help proteins to "feel" each other while in the encounter complex before short-range interactions are formed. This would result in acceleration of the association process.

The emerging picture is logical from both a structural and a chemical point of view. The secondary structure preferences indicate a priority for less rigid structures, which are more tolerant to adjust themselves toward a second protein, and are able to form close contacts across the interface. The tendency to form hydrophobic (and possibly polar) clusters exemplifies the importance of cooperativity between different groups on the same protein to achieve tight binding. Binding of two proteins is not achieved through a simple collection of many discrete adjacent bonds, but rather, a more complex network of interactions, webbed together to form a cooperative binding interface.

ProMate

How well does our prediction program perform? It is obvious that we do not succeed in all cases, but the large diversity of protein–protein interactions makes this task extremely difficult. Not all proteins use the same solutions for binding, and we obviously succeed in predicting only part of them. Still, the success rate is 70% (of interfaces for which ProMate made a prediction), with a success being defined rather stringently, as at least half of the predicted interface residues being indeed in the interface. Figure 11 shows that in those (and many other cases that are not shown) the predicted interface is actually located at the center of the real interface. This is not that surprising, taking into account the data acquisition method, which defines interface circles only as those that are over 70% interface, leaving much of the boundary outside the interface (boundary circles were discarded at the data acquisition state). Moreover, at the second step, ProMate uses only the top 10% of the scores, provided their score is over 0.7. Thus, both during data acquisition and during prediction, the center of the interface is preferred. Consciously, we prefer precision to sensitivity in predicting the location of the interface. An excellent test of the power of ProMate was the prediction of the location of interfaces in the disjoint DB. These proteins were not used at the data acquisition step, yet interface prediction worked nearly as well as for the unbound DB. This also shows that indeed the predictor is not sensitive to small changes during complexation. No significant differences were found between the predictions using NMR or crystal structures.

According to our results, four proteins from the unbound DB were predicted to have an additional binding region that does not appear in our DB of bound proteins (Table 1). With many proteins having multiple binding partners, one cannot exclude the validity of this prediction. A good example of such a protein is *Streptomyces subtilisin* inhibitor (3ssi, Table 1, Figure 11). This protein interacts with subtilisin BPN (PDB 2sic).[25] However, a second mode of interaction was reported with the zinc metalloproteinase ScNP.[26] Here, binding was suggested to occur at the opposite end of the inhibitor, fitting the second predicted binding site

(Table 1). Whether other additional binding sites predicted by ProMate are real is yet to be determined.

The bottom line

The emerging picture has far-reaching consequences in the way proteins bind. First, it would help explain the fast rate of protein–protein interactions, as potential binding sites are already probed during partially solvated intermediates formed en route to association.[27] If much of a protein's surface does not support binding, the search for specific interactions is reduced. A restricted binding surface would also suggest a way to answer the specificity paradigm. The first step in all protein–protein docking algorithms is to search for all possible bi-molecular conformations leading to a reasonable interface (in terms of shape complimentarity). Interestingly, many such potential conformations are found, albeit only one of them is correct. Reducing the available space for binding would actually filter out most wrong conformations. This is the basis of a new docking algorithm developed by us, which is performing surprisingly well.[28]

Future work

The web version of ProMate† enables to remove any currently used scores, and to add in external information. New properties that carry information regarding the interface location are bound to come up. Other properties might be within the expertise of other laboratories. Joining up as much information and experience as possible can significantly improve the interface prediction ability, which we believe would make a significant advance in the field of structural bioinformatics, and of docking in particular.

In the final ProMate version, we fixed the score combination to predict all interfaces, according to the overall best score as identified during the optimization step. However, we saw that different combinations fit better some individual cases. The goal here would be to identify a priori which combination to use for which case. Obviously, this would improve the predictive power of the program. As the dataset of protein complexes is currently very limited, this may be difficult to do before many more structures are determined.

Methods

Database construction

A DB of 67 structures of transient protein–protein heterodimers was derived from the PDB,[29] with at least one of the monomers being longer than 85 AA and both being longer than 50 AA. Antibodies were not included in the DB, since their evolutionary process is significantly more rapid than that of other proteins.

From this DB we derived a DB of 92 bound monomers that are longer than 85 AA. The minimum BLAST *p*-values between these monomers is 1×10^{-4}. A structural alignment was executed for each possible pair of monomers using the combinatorial extension method (CE).[30] For each pair that got a *Z*-score above 5, one of the proteins in the pair was removed from the DB. If the score was above 4, the complexes were aligned, and both were kept in the DB only if the interface location on the common monomer was different. The highest sequence identities according to the CE were 19.3%.

A DB of unbound structures was then derived from the bound DB. Using BLAST,[31] 57 of the monomers were found to have a highly homologous unbound form in the PDB, with more than 70% sequence identity.

Here, we perform a comparative analysis of two DBs: a bound DB containing 92 different monomers which structures have been determined in the bound form of the proteins, and an unbound DB containing 57 structures homologues to 57 of the bound monomers but solved in their unbound form. Most of the bound (87) and unbound (47) proteins were determined using X-ray crystallography and five bound and ten unbound were solved using NMR. Each amino acid on an unbound monomer was associated with an amino acid on the relevant bound monomer according to the BLAST's sequence alignment.

Surface analysis

The surface atoms of each monomer were extracted using Connolly's molecular surface dots program with a probe radius of 1.4 Å, and dot density of 1 dot/Å^2. Only the surface atoms were used throughout the analysis. For the statistical analysis of binding *versus* non-binding surfaces, the proteins surface was sampled using circles with a radius of 10 Å around anchoring dots, which are uniformly distributed over the monomer's surface (0.1 dot/Å^2). When examining the properties, each circle was given a score representing the level of the property within the circle. The surface examination code is using the EGAMB++functions library.‡

Interface definition

To extract the interface, we used a three-step calculation. First, all residues with atoms that are buried upon complexation were marked on the bound monomer using Connolly's molecular dot surface (MS) program. For the unbound proteins, these amino acid residues were projected onto the bound protein using the BLAST alignment. This definition leaves certain atoms unmarked, though they are surrounded by marked atoms and are located in the interface. We call these atoms "holes" in the description of the method. A heuristic procedure was used to find these holes and add them to the list of interface atoms. This procedure is executed for both bound and for unbound proteins. Finally, each circle was assigned a *CII* value, which is the fraction of interface atoms it possesses.

The "hole-removing" procedure looks at all triangles

† http://bip.weizmann.ac.il/promate

‡ gamba.math.tau.ac.il/group/GAMB++/GAMB.html

formed by any triplet of atoms that were found to be interface in the first stage (by the MS program). For each triangle, the algorithm projects all the atoms that are within a 15 Å distance from all the triangle's vertices to the triangle's plane. If the atom's projection falls inside the triangle, then the atom's score is increased. Finally, all the atoms with scores higher than the mean score are added to the interface atoms' list. The side effect of this procedure is that for non-convex interfaces the algorithm might add some of the atoms that lie on the margin of the interface. A pseudo code for the *CII* calculations is given as Supplementary Material.

Using the *CII* value we divide the protein's surface into three areas:

Interface = {dots with $CII > 0.7$}
Non-Interface = {dots with $CII = \ = 0$}
Boundary = otherwise

Only the first two were used for gathering statistics.

Using this definition, the interface region occupies about 16% of the total surface (8383 SD), the non-interface is ~48% of the surface (25,672 SD) and the boundary is ~36% (19,314 SD).

Statistical significance

Error bars for categorical scores

The error bars are the 70% confidence intervals as they were estimated by the bootstrap resampling method using 1000 bootstrap samples.[32] In the graphs, categories where the error bars do not overlap appear in capital letters.

p-*Value for continuous scores*

Two statistical tests prove the significance of the result for continuous scores. First is by using the Kolmogorov–Smirnov test on the interface *versus* non-interface samples. The second test is a test for the mean, since, contrary to the distribution itself, this measure is expected to be distributed normally. A thousand samples of the length of the interface sample were randomly selected from the non-interface sample. The mean of these samples is normally distributed. The *p*-value is then estimated from the normal distribution using the mean and standard deviation values of these 1000 samples. These methods provides only a rough estimation of the random probability, since the random model here does not take into account the continuity of the interface. A more correct, but complicated model would need to use randomly generated interfaces. A possible way to produce such random interfaces is from false docking results of the monomers.

Knobs and holes

To measure the extent to which a surface dot is located inside a hole (concave area) on the protein surface we used a simplified version of the solid angle method suggested by Connolly.[33] The hole-index (*HI*) of a surface dot is the fraction of a 10 Å radius ball around this dot that is occupied by the protein. Thus, the possible *HI* values are in the range of (0,1), whereby a hole would have *HI* closer to 1. A knob was defined by having $HI \leq 0.2$ and a hole has $HI \geq 0.3$ (see Figure 12(b)).

ProMate: the interface predictor

The prediction process is constructed of three independent stages. First, the training set proteins are analyzed to produce the interface and surface histograms relevant to each property. Second, these data are used to estimate the interface probability of each circle of a test protein. Finally, neighboring circles, that were predicted to be interface, are grouped together into predicted interface patches. Each of the stages is explained in detail below.

Training (preprocessing): probability extraction

This stage only uses the interface and non-interface circles, ignoring the boundary ones. For categorical scores the probability of a certain category is simply its frequency of appearing in the interface and in the non-interface region of the training set. The histograms for continuous scores were constructed from the data, starting with a bin around each single sample, and merging bins closer than 0.005. Then, the bins were clustered using a greedy algorithm in order to reduce the noise. At each clustering iteration, two bins were merged into one using the following rules. First, the mutual information (*MI*) of the joint distribution of the examined property and the interface (a binary property having the value true for interface circles and false for non-interface ones) was calculated for each possible merge of two bins. Each such merge reduces the *MI* until at the extreme there is only one bin containing all the samples, which necessarily holds no information distinguishing the circles. At each iteration, the bins whose merge would least reduce the *MI* are chosen and merged. Of course, the algorithm does not go all the way to the extreme of a single cluster. Instead it stops when the decline in *MI* in a single step is higher than 5% of the original *MI* value. For categorical scores every two bins are examined, whereas for continuous scores only the merge of consecutive bins is considered.

Scoring a test protein

A test protein is initially processed as an independent set of circles. For every circle, each of the properties is examined and the likelihood of this circle to belong to the interface according to it is determined. All properties are divided into two classes: simple properties are those that produce a single value per circle, for example, the water score which is the (normalized) number of water molecules within a circle. For such properties the score is the observed frequency of the specific score in the interface of the training set divided by its sum of observed frequencies in the interface and non-interface. In other words, denoting interface by I and surface by *S*, when *O* refers to the observed frequency in the training set, for an input circle *c*:

$$\text{Estimated Pr}(c \in I|\text{property Val}(c) = v) = \frac{O(\text{property Val}(c) = v|c \in I)}{O(\text{property Val}(c) = v|c \in I) + O(\text{property Val}(c) = v|c \in S)} \qquad (1)$$

Effectively, this score is equivalent to the sum-of-log-likelihood scoring method ($x = \Sigma\log(S/I)$), transformed by $1/(1 + e^x)$ to produce a value in the range [0,1].

However, using this for categorical properties results in many values per circle. For example, examining a single circle's sample of the (multinomial) amino acid distribution results in a 20 entries vector containing natural numbers. To produce a similar single score per circle the amino acid residues (or in general, the dots in the circle) were regarded as independent. Thus, each dot is given a score according to equation (1), and the circle's score is the product of all the dots' scores in it. To correct for the dots independency assumption, the final scores are normalized according to the actual distributions of these scores in the training DB. For example, if 40% of the interface circles of the training set got a score of 1, but at the same time 20% of the non-interface circle got the same score, then the score 1 would be shifted to $0.4/(0.4 + 0.2) = 0.67$.

The explanation so far treats all the circles as equally likely to be interface or non-interface. There appears to be some increase in the fraction of protein's surface that is occupied by the interface as the protein gets longer. Therefore, each circle is also multiplied by the *a priori* probability of being an interface for a protein of that size. Then, for rescaling of the range of probabilities, it is multiplied by 1 − ("one minus") the average interface fraction over all proteins, which is 16% (in surface dots, not including boundary).

As for multinomial properties, the combined score is the product of all the scores resulted from the different properties. Here, this combined score is corrected according to the actual frequencies as they appear in the training set.

Considering neighboring circles

To further smooth the score of each dot the environment in which it resides is taken into consideration. After all the dots were assigned their scores, the score of each dot is again combined with the scores of all the adjacent dots in a 7 Å circle around it. This procedure is repeated for a small number of iterations.

Nevertheless, simply adding in the neighbors' scores is problematic, since the method presented assumes that there are no dependencies between the scores, whereas neighboring circles share up to 80% of their area, and thus their scores are highly positively dependent. Through the dependencies, the same information is being used more than once and by this the final scores become more extreme than they should be. This time, the dependency cannot be reduced by a simple normalization.

To reduce this effect, each term in the score is raised to the power of $\delta \in (0,1)$. Note, that this is only beneficial, since the dependency between the circles here is not stochastic, but a positive one. Due to the overlap between the circles, neighboring circles are expected to give similar scores, thus taking some root of the product would get us back to an averaged version of the original values. The score thus becomes:

$$\forall \text{circle } c,\ \ \mathrm{Sc}'(c) = \frac{\mathrm{Sc}(c)\left(\prod_{n \text{ neighbor of } c} \mathrm{Sc}(n)\right)^{\delta}}{\mathrm{Sc}(c)\left(\prod_{n \text{ neighbor of } c} \mathrm{Sc}(n)\right)^{\delta} + (1-\mathrm{Sc}(c))\left(\prod_{n \text{ neighbor of } c} (1-\mathrm{Sc}(n))\right)^{\delta}} \quad (2)$$

Explicitly, to the original score of the circle we add information from its environment, but since most of this information was already used in Sc(c), and since there is repetition of information between the neighbors themselves, the environment's effect is given a smaller weight by taking a small value for $\delta = 0.01$.

Clustering circles into predicted interface patches

The clustering step is calculated at the level of the amino acid residues. The score of each amino acid is an extrapolation of the scores of the neighboring dots to the location of its C^{α}. Interface amino acid residues are considered to be those with the 10% of the highest scores, but only if the score is above 0.7. If no amino acid residue fits this description, no interface is considered to have been found. In a second step, all these amino acid residues serve as graph nodes. An undirected edge is defined between two nodes, whose C^{α} are less than 13 Å apart. All the strongly connected components of this graph are identified, and components containing more than two nodes are considered as predicted interface patches. For the evaluation of the results only the biggest patch was considered (see Figure 11).

Acknowledgements

We thank Kay Gottschalk and Tal Peleg-Shulman for their critical reading of the manuscript. We thank Haim Wolfson for letting us use his EGAMB++functions library.

References

1. DeLano, W. L., Ultsch, M. H., de Vos, A. M. & Wells, J. A. (2000). Convergent solutions to binding at a protein–protein interface. *Science*, **287**, 1279–1283.
2. Lim, D., Park, H. U., De Castro, L., Kang, S. G., Lee, H. S., Jensen, S. *et al.* (2001). Crystal structure and kinetic analysis of betalactamase inhibitor protein-II in complex with TEM-1 beta-lactamase. *Nature Struct. Biol.* **8**, 848–852.
3. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Struct. Funct. Genet.* **43**, 89–102.
4. Lijnzaad, P. & Argos, P. (1997). Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins: Struct. Funct. Genet.* **28**, 333–343.
5. Jones, S. & Thornton, J. M. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
6. Jones, S. & Thornton, J. M. (1995). Protein–protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31–65.
7. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of proteinprotein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
8. Chakrabarti, P. & Janin, J. (2002). Dissecting

protein–protein recognition sites. *Proteins: Struct. Funct. Genet.* **47**, 334–343.

9. Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein–protein recognition. *Protein Sci.* **3**, 717–729.
10. Kleanthous, C. (2000). *Protein–Protein Recognition: Frontiers in Molecular Biology*, Oxford University Press, Oxford, UK.
11. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
12. 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.
13. Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng.* **3**, 77–83.
14. 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.
15. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356–1361.
16. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
17. Zhou, H. X. & Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Struct. Funct. Genet.* **44**, 336–343.
18. Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M. *et al.* (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261.
19. Jones, S. & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
20. Albeck, S., Unger, R. & Schreiber, G. (2000). Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *J. Mol. Biol.* **298**, 503–520.
21. Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.
22. Hutchinson, E. G. & Thornton, J. M. (1996). PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.* **5**, 212–220.
23. Bright, J. N., Woolf, T. B. & Hoh, J. H. (2001). Predicting properties of intrinsically unstructured proteins. *Prog. Biophys. Mol. Biol.* **76**, 131–173.
24. Lijnzaad, P., Berendsen, H. J. & Argos, P. (1996). A method for detecting hydrophobic patches on protein surfaces. *Proteins: Struct. Funct. Genet.* **26**, 192–203.
25. Takeuchi, Y., Satow, Y., Nakamura, K. T. & Mitsui, Y. (1991). Refined crystal structure of the complex of subtilisin BPN′ and *Streptomyces subtilisin* inhibitor at 1.8 Å resolution. *J. Mol. Biol.* **221**, 309–325.
26. Hiraga, K., Suzuki, T. & Oda, K. (2000). A novel double-headed proteinaceous inhibitor for metalloproteinase and serine proteinase. *J. Biol. Chem.* **275**, 25173–25179.
27. Selzer, T. & Schreiber, G. (2001). New insights into the mechanism of protein–protein association. *Proteins: Struct. Funct. Genet.* **45**, 190–198.
28. Gottschalk, K. E., Neuvirth, H. & Schreiber, G. (2004). A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *PEDS*, in the press.
29. 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.
30. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
31. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
32. Efron, B. (1982). *The Jackknife, The Bootstrap and Other Resampling Plans*, SIAM, Philadelphia, PA.
33. Connolly, M. (1986). Measurement of protein surface shape by solid angles. *J. Mol. Graph.* **4**, 3–6.

Edited by J. Thornton

(Received 8 May 2003; received in revised form 13 January 2004; accepted 4 February 2004)

Supplementary Material containing "A pseudo code for the *CII* calculation" is available on Science Direct