Optimal Docking Area: A New Method for Predicting Protein–Protein Interaction Sites

Juan Fernandez-Recio,[1] Max Totrov,[2] Constantin Skorodumov,[2] and Ruben Abagyan[1]*
[1]*Department of Molecular Biology, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California*
[2]*Molsoft, LLC, 3366 Torrey Pines Court, La Jolla, California*

ABSTRACT **Understanding energetics and mechanism of protein–protein association remains one of the biggest theoretical problems in structural biology. It is assumed that desolvation must play an essential role during the association process, and indeed protein–protein interfaces in obligate complexes have been found to be highly hydrophobic. However, the identification of protein interaction sites from surface analysis of proteins involved in non-obligate protein–protein complexes is more challenging. Here we present Optimal Docking Area (ODA), a new fast and accurate method of analyzing a protein surface in search of areas with favorable energy change when buried upon protein–protein association. The method identifies continuous surface patches with optimal docking desolvation energy based on atomic solvation parameters adjusted for protein–protein docking. The procedure has been validated on the unbound structures of a total of 66 non-homologous proteins involved in non-obligate protein–protein hetero-complexes of known structure. Optimal docking areas with significant low-docking surface energy were found in around half of the proteins. The 'ODA hot spots' detected in X-ray unbound structures were correctly located in the known protein–protein binding sites in 80% of the cases. The role of these low-surface-energy areas during complex formation is discussed. Burial of these regions during protein–protein association may favor the complexed configurations with near-native interfaces but otherwise arbitrary orientations, thus driving the formation of an encounter complex. The patch prediction procedure is freely accessible at http://www.molsoft.com/oda and can be easily scaled up for predictions in structural proteomics. Proteins 2005;58:134–143.**

Key words: protein–protein association; desolvation energy; atomic solvation parameter (ASP); binding site prediction; docking hot-spots

INTRODUCTION

Better understanding of the energetics and mechanism of protein–protein association is a matter of great scientific and practical interest, as it can improve modeling and prediction of protein interactions (ranging from 1:1 complexes to complete interaction networks) and may help to identify new therapeutic targets for drug design. Although much experimental and theoretical work on molecular recognition already exists, the energy determinants of protein–protein interaction are still a subject of much debate. Kinetic experimental and theoretical studies indicate that in some cases electrostatics could drive the formation of an 'encounter' complex[1–3] formed during protein–protein association, while more specific interactions, such as hydrogen bonding, salt bridges and interaction between hydrophobic patches, could account for the specificity of the final orientation. Often, only a few residues contribute extensively to the binding energy ('hot spots'), but no common physical-chemical pattern has been found in them.[4–8] In addition, continuum electrostatic calculations suggest that the relative contributions of electrostatic and hydrophobic forces to complex formation vary widely among different complexes.[9] However, it is widely accepted that the hydrophobic effect is the major contributor to the affinity of the association.[10,11] Indeed, statistical analyses of known protein–protein complex structures have clearly shown the hydrophobic character of protein interfaces in obligate complexes (where the individual subunits do not present functionality when they are separated).[12–18] Similarly, it has been shown that large hydrophobic patches found on protein surfaces[19] correlate with protein–protein interfaces in obligate complexes.[20] However, in non-obligate complexes, despite noticeable preferences in residue content at protein interfaces,[18,21–24] these preferences are not strong enough to unambiguously predict the location of the protein interface.

Recently, it has been shown that differently oriented rigid-body docking poses from computer simulations accumulate around the known protein–protein interaction sites.[25] Favorable desolvation around the native protein–protein interfaces is not highly dependent on the specific mutual orientations, and thus it could be one of the main contributors to the rigid-body docking funnels found dur-

Present address for Juan Fernandez-Recio is Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, UK.

*Correspondence to: Dr. Ruben Abagyan, Department of Molecular Biology, TPC-28, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. E-mail: abagyan@scripps.edu.

Received 9 March 2004; Revised 29 June 2004; Accepted 8 July 2004

Published online 19 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20285

ing simulations. Indeed, theoretical analysis of protein–protein binding energy landscapes has also shown the relevant role of desolvation in the formation of encounter complexes around the native binding sites during protein–protein association.[26,27] This makes it necessary to characterize the desolvation properties of protein surfaces and further analyze their role in protein–protein binding. Previous electrostatic desolvation calculations,[28] accessible surface area (ASA)-based calculations[14,19,29] and small-molecule probing[30,31] have led to the identification of apolar patches on unbound protein surfaces that are expected to be important for binding. However, although these apolar patches correlate quite often with the location of small-molecule binding sites and/or the location of obligate protein–protein interfaces,[20] it has always been difficult to show their importance (if any) for non-obligate protein–protein association.

Here we present a method for the identification of optimal docking areas (ODAs). Our algorithm generates protein surface patches of different sizes and analyzes their docking surface energy, based on atomic solvation parameters previously derived from octanol/water transfer experiments and adjusted for protein–protein binding.[25] These areas of low docking surface energy would correspond to regions likely to be buried in the interaction with other proteins, and, as shown below, the patches correlate well with known protein interaction sites in transient hetero-complexes. The method can be used as a fast predictor of non-obligate protein interaction sites. A web server of the method is publicly available (http://www.molsoft.com/oda), and its performance is of high potential interest for increasing numbers of proteomics and 'interactomics' projects.

METHODS AND METHODS

In order to map low desolvation areas on protein surfaces, we need to evaluate different arrangements of surface residues that are able to form surface patches. The common approach involves dividing the protein surface into equal-area patches. However, such patches may poorly represent the real protein interaction sites that vary significantly in size. We propose here an alternative method. First, we generated N surface points evenly distributed along the protein surface. Then, we calculated the ODA that could be generated from each of these surface points (see the following section). Finally, the surface points that generated the ODAs with significant low-energy values were used to define a region over the protein surface most likely to be involved in protein–protein binding.

Generation of Surface Patches

Optimal docking surface patches were generated by an iterative method, as explained in Figure 1. First, a series of points ($i = 1, 2, \ldots, N$) was systematically generated on the protein solvent-accessible surface (expanded by 3 Å to overcome minor structural details) by dividing the surface into triangles with average sides of 5 Å [Fig. 1(a)]. This distribution of points around the molecule is insensitive to the atomic details of the molecular surface but reflects its overall shape (including sizeable concave and convex elements). These surface points were subsequently used to generate series of surface patches. As Figure 1(b) shows, different-sized surface patches were generated by selecting all surface residues at different distances ($d = 1, 2, \ldots, 20$ Å) from a given surface point i. The docking surface energy of these surface patches, based on the atomic ASA of their component residues, was then calculated (described below). Among all surface patches generated from surface point i, the patch with the lowest energy value was called an ODA, and its optimal energy value was assigned to the surface point i. As Figure 1(c) shows, the process was repeated for all N surface points, and the energy values of the calculated ODAs were thus mapped onto all the surface points (hereafter also referred to as 'ODA points'). The ODA points can be graphically displayed in varying color and size according to their docking surface energy values, so the ones with the most favorable values for protein binding can be easily spotted.

Calculation of Surface Docking Desolvation Energy

The docking desolvation energy of a given surface patch (defined as a set of surface residues) for its transfer from water to a protein–protein interface was calculated using an atomic ASA-based model (i.e. as a sum of per-atomic contributions proportional to the solvent-ASA)[32,33] according to eq. (1).

$$E_{\text{desolv}} = -\sum_i \sigma_i \, \text{ASA}_i \qquad (1)$$

where σ_i is the atomic solvation parameter (ASP) for atom type i (i.e. the contribution to the solvation energy per unit of ASA). The ASPs for the different atom types are shown in Table I. Its values were calculated from linear fitting to octanol/water transfer energies of N-acetyl aminoacid amide derivatives,[34] and finally adjusted for optimal rigid-body docking energy landscapes by weighting the contributions of polar, aromatic and aliphatic atoms, as previously described.[25]

Database Generation

We used three different datasets. Database A was formed from 38 unbound structures of proteins involved in complexes of known structure, which we previously compiled as a benchmark for protein–protein docking (Table II). Database B is an external dataset based on the one compiled by Schreiber's group[35] and consisting of 51 non-homologous unbound proteins that have a highly homologous complexed form (>70% sequence identity) in the Protein Data Bank (Table III). Some of the cases in the original Schreiber database were not considered here, because either the unbound or the bound structure was incomplete (PDB codes of disregarded cases: 1aye, 1cto, 1d0n, 1ez3, 1nos, 1qqr). Finally, database C was created from all X-ray structures (i.e. no NMR structures) in database B, along with the X-ray structures in database A that had low sequence identity with the proteins in database B (a threshold of pP [-log(Probability of Struc-

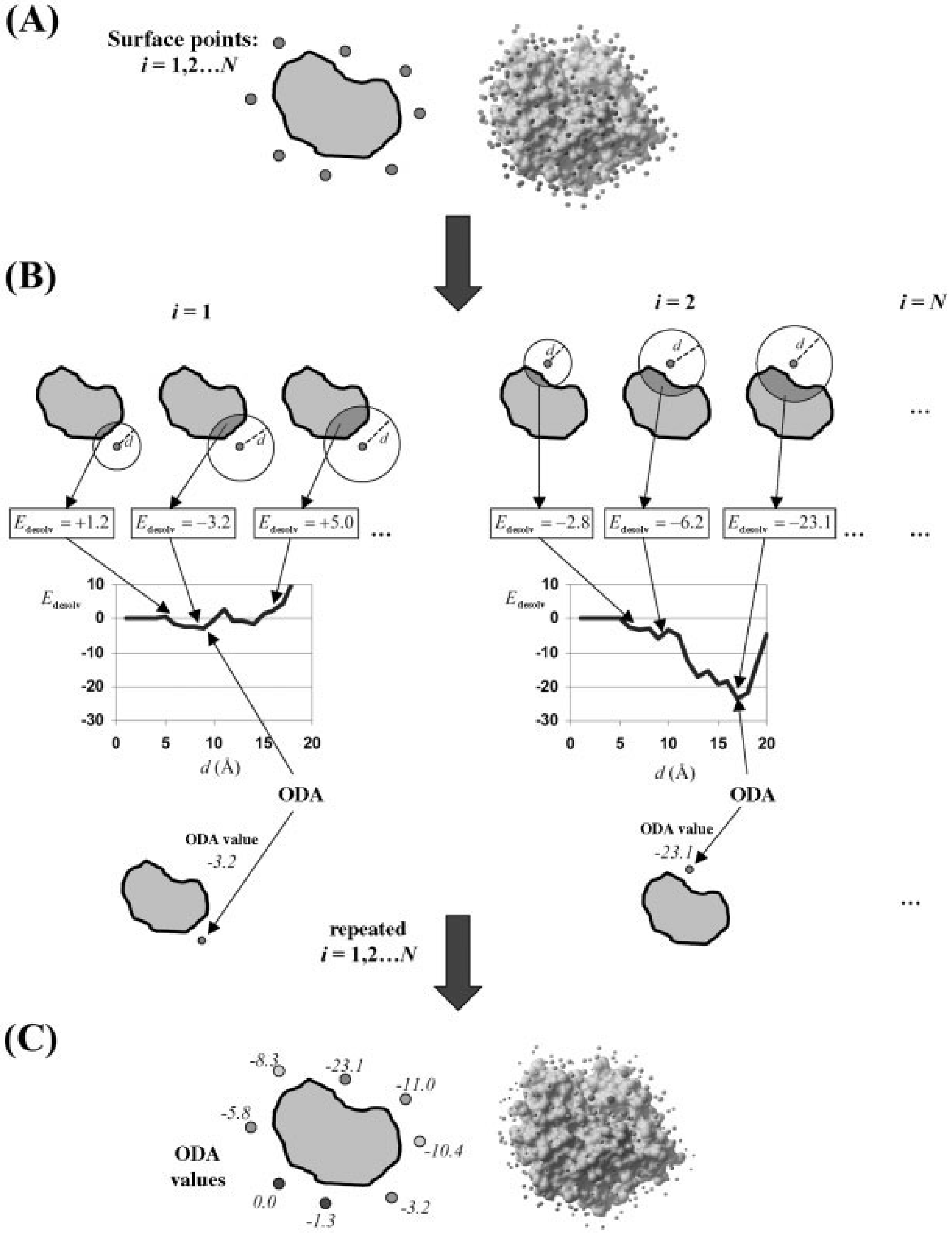

Fig. 1. (a) Representation of surface points generated from the structure of the unbound protein. (b) For each surface point, different surface patches were generated by selecting all residues within a sphere of increasing radius *d*. The docking desolvation energy of the surface patch is a function of this *d* value. For each surface point, the surface patch with the minimum energy was defined as the ODA, and its value was assigned to that surface point. (c) The process was repeated for all *N* surface points, and the calculated ODAs were thus mapped onto these surface points, which can be represented in different colors and sizes according to their ODA values.

tural Dissimilarity)] = 5 for the ZEGA[36] sequence similarity significance was used). The list of PDB entries of the unbound structures in database C is as follows: 1a19, 1a2p, 1acl, 1ag6, 1aue, 1avu, 1b1e, 1ctm, 1d2b, 1ekx, 1ex3, 1f00, 1f5w, 1fkl, 1flz, 1fvh, 1g4k, 1gc7, 1hh8, 1hpl, 1hu8, 1iob, 1j6z, 1jae, 1lba, 1nob, 1pne, 1poh, 1ppp, 1rgp, 1sel, 1vin, 1wer, 1xpb, 2bnh, 2cpl, 2f3g, 3ssi, 6ccp, BLIP, 1rge, 1aap, 2ci2, 1fsc, 1hpt, 2ugi, 1fxa, 1que, 1mlb, 1buy, 1ern, 1b1z, 1bec.

RESULTS

The identification of low-energy ODA spots on a protein surface, as described in the ‘Methods’ section, defines a spatial region for which docking surface energy would be favorable upon protein binding. We studied the presence of these favorable docking areas on proteins involved in known protein–protein interactions in order to evaluate the ability of the ODA predictor to identify protein–protein

TABLE I. Atomic Solvation Parameters[a]

σ (cal mol^{-1}Å^{-2})	Radius (Å)	Atom Type
19.18	1.95	C aliphatic
110.80	1.8	C aromatic
−39.10	1.7	N uncharged
−126.04	1.7	N_ζ in Lys^+
−62.56	1.7	$N_{\eta}1$, $N_{\eta2}$ in Arg^+
−42.55	1.6	O hydroxyl
−31.28	1.4	O carbonyl
−68.77	1.4	O^- in Glu, Asp
25.76	2.0	S in SH
5.06	1.85	S in Met or S-S

[a] Based on octanol/water transfer energies optimized for protein–protein binding.[25]

interaction sites. To ensure realistic predictions, we used the three-dimensional coordinates of the unbound subunits in a completely automatic way and assumed no previous information about the location of the protein–protein interfaces.

Analysis of ODAs in Unbound Proteins

We applied the ODA procedure (described in 'Methods') to two different datasets (A and B; see 'Database Generation' in 'Methods') of unbound protein molecules known to be involved in transient protein–protein heterocomplexes. The total number of evaluated surface patches (20 × number of surface points) varied from 2,160 (ovomucoid) to 11,580 (neuronal sec I). The computational time of the overall procedure ranged from 1 s (ovomucoid) to 6 min (neuronal sec I) on a single 2.4 GHz P4 CPU running Linux. The surface points of some of the molecules analyzed in this work are represented in Figure 2, colored by the docking surface energy value of their corresponding ODAs. It can be seen that the ODA points with the lowest energy values are located near the known protein–protein interfaces. The ODA value represents the hypothetical gain in energy if the corresponding surface patch were buried upon protein binding. Therefore, significant low-energy ODA points (we established an arbitrary energy cutoff of −15.0 kcal/mol) define regions in which desolvation effects upon binding are expected to be most favorable. Such significant low-energy ODA points, representing hot spots for protein–protein docking, were detected in 58% of the molecules in database A and in 63% of the molecules in database B. The number of low-energy ODA points per molecule varied widely, ranging from 1 in some cases (e.g. chymotrypsin) to 163 [rhg-csf (recombinant human granulocyte colony-stimulating factor)].

ODA Hot Spots as a Predictor of Protein Interaction Sites: Benchmark Results

We evaluated whether the predicted protein interaction sites defined by the significant low-energy ODA points ('ODA hot spots') were located in the known protein–protein interfaces. In some cases, more than ten ODA hot spots were found, which defined surface areas much larger than the average protein–protein interfaces. In those cases, only the top ten ODA hot spots, as sorted by their energy values, were considered for the predictions. For a given unbound molecule, the prediction rate was defined as the percentage of predicted protein interaction sites (i.e. up to ten top ODA hot spots) that were correctly located in a known protein–protein interface (i.e. at a distance of <5 Å from any heavy atom of the partner molecule, after superimposing the structure of the unbound protein onto the equivalent molecule in the protein–protein complex). In 89% of the cases (Tables II and III), the prediction rate (100 × number of predicted sites correctly located / number of predicted interaction sites) was better than expected by random distribution (100 × number of predicted interaction sites / total number of ODA points), which indicates that location of the ODA hot spots correlates with the location of the known protein binding sites.

The results for our initial database A (Table II) show that, in 86% of the cases, at least half of the predicted protein interaction sites were located in the known interface (i.e., at a distance of <5.0 Å from any non-hydrogen atom of the partner molecule). Moreover, in around half of the cases, the prediction rate was as high as 90–100%.

This initial database A was previously compiled to test our protein–protein docking methods. It only contained examples for which the structure of both the complex and the two unbound subunits were known. In order to validate our predictions on a larger and more diverse dataset, we applied the method to database B (Table III), based on a set of unbound structures recently compiled by the Schreiber group.[35] This dataset represents a broader sample of unbound molecules. The results on database B show that ODA predictions were correctly located (>50 % accuracy) in 65% of the cases. This success rate is lower than in our initial test. The main reason seems to be the large number of NMR structures in database B (we noticed that the success rate on the NMR structures was particularly poor: predicted sites were correctly located in only 33% of the cases). Indeed, if we disregard NMR structures, the success rate improves to 77%. The remaining discrepancy between the results on the two sets may have a dual explanation. First, some of the protein families with good predictions (e.g. trypsin family) may be overrepresented in database A, which might artificially increase the average success rate. Second, the predictions for database B in many cases were compared with experimental data from complexes formed by highly homologous but not necessarily identical proteins. Thus, there is a chance that the two homologous proteins (the unbound one and the complexed one) might not have exactly the same interface with respect to a partner protein, which would adversely affect the evaluation of the predictions.

When we combined all non-homologous X-ray structures (i.e. no NMR structures) of databases A and B into a single final database C (see 'Methods'), the predicted interaction sites were correctly located in 80% of the cases.

Figure 3 shows the location of these predicted interaction sites on some proteins, compared to the structures of

TABLE II. Analysis of Optimal Docking Areas (ODAs) in Unbound Database A

				Prediction Rate	
Unbound Protein	PDB (Chain ID)	Total ODA Points[a]	ODA Hot Spots[b]	Hits (%)[c]	Complex Partner[d]
Enzymes					
Acetylcholinesterase	2ace	437	27	10	1fss (b)
α-Amylase	1pif	406	58	60	1bvn (t)
Bamase	1a2p (a)	152	—	—	—
Chymotrypsin	5cha (a)	241	1	100	1ca0 (d)
Chymotrypsinogen	1chg	243	10	90	1cgi (i)
Kallikrein A	2pka (a,b)	251	33	80	2pka (x,y)
Ribonuclease Sa	1rge (a)	144	—	—	—
Subtilisin BPN	2stl	240	—	—	—
Subtilisin Carlsberg	1sbc	239	1	100	1cse (i)
TEM-1 β-lactamase	1xpb	265	—	—	—
Thermitase	1thm	240	—	—	—
Trypsin (bovine)	5ptp	225	2	100	1taw (b)
Trypsin (rat)	1ane	230	—	—	—
UDG	1akz	256	35	50	1ugh (i)
Inhibitors					
APPI	1aap (a)	106	4	50	1ca0 (a,b,c)
Barstar	1a19 (a)	133	9	89	1bgs (a)
BLIP	BLIP[e]	212	82	100	1jtg (a)
BPTI	1bpi	120	—	—	—
CI-2	2ci2 (i)	123	—	—	—
Eglin C	1egl[f]	138	57	0	1acb (e)
Fasciculin II	1fsc	116	—	—	—
HPTI	1hpt	114	16	100	1cgi (e)
Ovomucoid	1omu[f]	108	—	—	—
Subtilisin inhibitor	3ssi	160	—	—	—
Tendamistat	2ait[f]	126	40	90	1bvn (p)
UGI	2ugi (a)	137	—	—	—
Electron-transfer					
Cytochrome c	1hrc	148	—	—	—
Cytochrome c Peroxidase	1ccp	305	—	—	—
Cytochrome f	1ctm	323	8	63	2pcf (a)
Ferredoxin	1fxa (a)	136	6	100	1ewy (a)
Ferredoxin-NADP+ reductase	1que	336	—	—	—
Plastocyanin	1ag6	138	—	—	—
Antibodies					
Fab D44.1	1mlb (a,b)	432	12	80	1mlc (e)
Fv D1.3	1vfa (a,b)	255	11	90	1vfb (c)
Others					
Erythropoietin	1buy (a)	241	6	100	1eer (b)
Erythropoietin receptor	1ern (a)	283	8	0	1eer (a)
SpcA	1blz (a)	251	7	57	1l0x (a)
TCR-β	1bec	323	45	70	1tcr (a)

[a] Total number of surface points used to generate ODAs, as described in 'Methods.'
[b] Number of ODA points with energy <-15 kcal/mol.
[c] Percentage of ODA hot spots (up to a maximum of ten) correctly located in a known protein–protein interface (<5 Å from any non-hydrogen atom of the partner molecule in the complex, after the unbound protein is superimposed onto the equivalent molecule of the protein–protein complex).
[d] PDB code of the complex structure used to evaluate the ODA predictor.
[e] No coordinates deposited in PDB: the structure was kindly provided by its authors.[40]
[f] Structure solved by NMR.

known protein–protein complexes in which these proteins are involved.

As a test, we also applied the described protocol to the three-dimensional structures of the complexed subunits taken from the corresponding protein–protein complexes. The results are practically identical (data not shown), which indicates that the ODA predictor is not sensitive to small conformational changes in the binding interfaces.

Discussion

Desolvation Energy Upon Protein Binding: ODAs

The method proposed here for the identification of low docking surface energy regions for protein binding is based on the selection of ODAs among a series of different-sized surface patches generated from a set of surface points around the protein. In practice, the method calculates the

TABLE III. Analysis of Optimal Docking Areas (ODAs) in Unbound Database B[35]

				Prediction Rate	
Unbound Protein	PDB (chain ID)	Total ODA Points[a]	ODA Hot Spots[b]	Hits (%)[c]	Complex Partner[d]
Barstar	1a19 (a)	133	9	89	1brs (a)
Bamase	1a2p (a)	152	—	—	—
Tumor suppressor p16ink4a	1a5e[e]	261	5	20	1bi7 (a)
Acetylcholinesterase	1acl	447	7	0	1fss (b)
Plastocyanin	1ag6	138	—	—	—
Cdc42Hs	1aje[e]	288	76	40	1am4 (a)
Rhogdi	1ajw[e]	231	—	—	—
Fkbp-rapamycin-binding dom.	1aue (a)	153	10	80	1fap (a)
Trypsin inhibitor	1avu	227	11	70	1avw (a)
Hydrolase angiogenin	1b1e (a)	193	14	70	1a4y (a)
Trypsin/α-amylase inhibitor	1bip[e]	201	28	100	1tmq (a)
Cytochrome f	1ctm	323	8	63	2pcf (a)
CheY	1cye[e]	159	7	29	1bdj (b)
Hydrolase inhibitor	1d2b (a)	188	28	100	1uca (a)
Transferase	1ekx (a)	318	9	22	1ekx (b,c)
Bovine chymotrypsinogen a	1ex3 (a)	242	9	100	1egi (i)
Enzyme I	1eza[e]	360	—	—	—
Rgs4	1ezt (a)[e]	198	3	0	1agr (a)
Intimin	1f00 (i)	349	—	—	—
Coxsackie/adenovirus receptor	1f5w (a)	178	1	100	1kac (a)
Fk506 binding protein	1fkl	155	—	—	—
Uracil-DNA glycosylase	1flz (a)	256	10	90	1eui (c)
Neuronal sec1	1fvh (a)	579	7	71	1dn1 (b)
Hydrolase	1g4k (a)	210	69	100	1uea (b)
Radixin ferm domain	1gc7 (a)	378	—	—	—
Rhg-csf	1gnc[e]	277	163	10	1cd9 (b,d)
P67Phox	1hh8 (a)	252	—	—	—
Lipase	1hpl (a)	428	—	—	—
P53 core DNA-binding dom.	1hu8 (a)	238	3	0	1ycs (b)
Interleukin-1 beta	1iob	198	—	—	—
Actin	1j6z (a)	372	6	100	1c0f (s)
α-Amylase	1jae	377	1	100	1tmq (b)
$T7^{-}$ lysozyme	1lba	202	—	—	—
Fiber knob protein	1nob (a)	216	—	—	—
Procolipase b	1pco[e]	169	18	0	1eth (a)
Profilin	1pne	177	—	—	—
Phosphotransferase	1poh	130	11	80	1ggr (a)
Papain	1ppp	230	20	30	1stf (i)
Rhogap	1rgp	231	6	100	1am4 (e,f)
Selenosubtilisin	1sel (a)	240	5	20	1cse (i)
Cyclin a	1vin	280	21	100	1fin (a,c)
P120gap	1wer	386	—	—	—
TEM-1 β-lactamase	1xpb	265	—	—	—
Ribonuclease inhibitor	2bnh	445	—	—	—
Cyclophilin a	2cpl	190	—	—	—
IIA-Glc	2f3g (a)	184	—	—	—
HIV-1 Nef	2nef[e]	242	115	90	1avz (c,a)
RalGEF-rbd	2rgf[e]	164	1	100	1lfd (a,c,d)
Subtilisin inhibitor	3ssi	160	—	—	—
Cytochrome c Peroxidase	6ccp	298	—	—	—
BLIP	BLIP[f]	212	82	100	1jtg (a)

[a] Total number of surface points used to generate ODAs as described in 'Methods.'

[b] Number of ODA points with energy < -15 kcal/mol

[c] Percentage of ODA hot spots (up to a maximum of ten) correctly located in a known protein–protein interface (<5 Å from any non-hydrogen atom of the partner molecule in the complex, after the unbound protein is superimposed onto the equivalent molecule of the protein–protein complex).

[d] PDB code of the complex structure used to evaluate the ODA predictor.

[e] Structure solved by NMR.

[f] No coordinates deposited in PDB: the structure was kindly provided by its authors.[40]

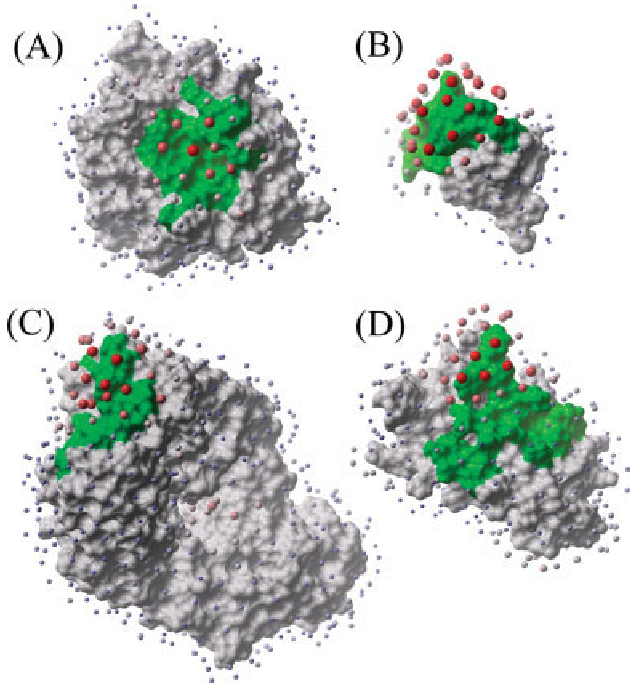

Fig. 2. Surface points around unbound proteins, colored according to the energy values of their respective ODAs (red are the lowest energy values). The size of the ODA points is proportional to their energy values (larger size represents lower energy value). The surface of the unbound molecule is represented in white. Known interface residues (i.e. residues within 5 Å from any non-hydrogen atom of a protein partner in a known complex) are shown in green: (a) chymotrypsin (protein interface with APPI, PDB 1ca0); (b) HPTI (protein interface with chymotrypsinogen, PDB 1cgi); (c) Fab (protein interface with lysozyme, PDB 1mlc); (d) EPO (protein interface with EPO receptor, PDB 1eer).

optimal size of the surface patch that would yield the best-possible docking surface energy on different regions of the protein surface. We can thus identify which regions (if any) would have a favorable energy change when buried by a hypothetical partner protein. Our results indicate that, for more than half of the proteins, we can find a region in which unspecific binding of another protein would be highly favored.

The atomic solvation parameters used here to calculate the docking surface energy of a given surface patch (values listed in Table I) were previously obtained from optimization of rigid-body docking energy landscapes,[25] and they seem to be consistent overall with previously reported values, such as the parameters derived by Lomize and coworkers from mutation energetics data (when no explicit van der Waals interaction is included).[37] The high values for the desolvation of aromatic carbon atoms in our model might seem surprising. However, they probably account for the aromatic/aromatic van der Waals interactions, which are much stronger than interactions with aliphatic atoms implicitly represented in regular solvation parameters derived from octanol/water transfer. Thus, our parameters may well represent the 'water/interface' desolvation energy, the non-specific transfer of a solvent-exposed atom to the interior of a protein–protein interface. This docking desolvation energy may mean all binding energy contributions that can be ascribed to the surface and thus estimated without explicit pair-wise calculations.

We explored the possibility of using different ASP values from the ones adjusted for docking (Table I). For instance, since the contribution of aliphatic carbons to the docking desolvation energy is very small compared to that of most of the other atom types, we recomputed the ODAs, setting the ASP for aliphatic carbons to 0. In this case, the overall success rate slightly decreased to 74% (compared to 80%). Moreover, when we took the original parameters derived from octanol/water transfer experiments, before any adjustment for docking,[25] the success rate dramatically decreased to 43%. Other ASP values systematically yielded lower prediction rates. This indicates that the ASP values used here are optimally balanced to describe the water/interface transfer energy in protein–protein docking.

ODA Hot Spots: Correlation with Known Protein Interaction Sites

In more than half of the proteins analyzed, it was possible to find at least one significantly low docking surface energy region on the protein surface. The location of the surface points with the lowest significant ODA values (ODA hot spots) defined a region on the protein surface that, remarkably, was located in or in the vicinity of the known binding site in most of the complexes. For some cases, the protein is known to be involved in interactions with different proteins using the same binding site. This is the case for chymotrypsin, kallikrein A and APPI, where predicted ODA points were found to be correctly located in the interfaces, as can be seen in Figure 3(a–c). There are also examples of proteins involved in different interactions by using different binding sites, as this is the case for TCR-β. Figure 3(d) shows that the top ten ODA hot spots accumulated around one of its binding sites, indeed the stronger one, in an interaction with the TCR-α subunit.

There were also some proteins for which most of the predicted protein interaction points were found to be far from the known binding site. This is the case for Eglin C, whose structure in solution has been determined by NMR (PDB code 1egl), and shows significant backbone rearrangement with respect to the bound form in complex with chymotrypsin, subtilisin and thermitase (PDB codes 1acb, 1cse and 2tec, respectively). However, when we used the complexed conformation (taking Eglin C coordinates from the mentioned complexes), the predicted interaction points were located closer to the binding site (prediction rate = 50%). Thus the low prediction rate for unbound Eglin C may be explained by the dramatically different conformation of the flexible loop found in the NMR structure (see later below for more details about general poor predictions in NMR structures). Other X-ray structures with inaccurate predictions include acetylcholinesterase [Fig. 3(e)], EPO receptor, transferase, p53, papain and selenosubtilisin. The non-native predicted binding spots may well be artifacts of our method; however, we cannot completely disregard the possibility of alternative binding areas for which there are not yet experimental or structural evi-

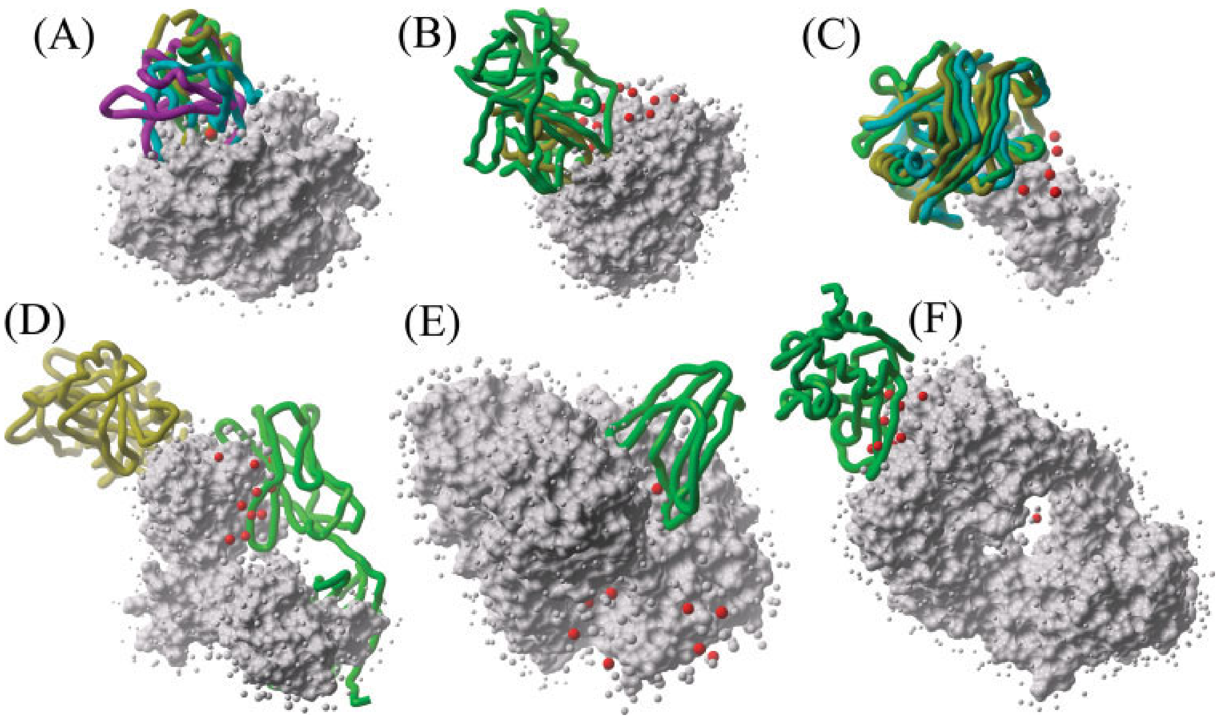

Fig. 3. Protein interface prediction based on the ODA points calculated in unbound protein molecules (represented in white surface). Predicted protein interaction sites (up to ten top ODA hot spots) are represented in red. The size of the ODA points is proportional to their energy values (larger size represent lower energy values). The partner molecule(s) in the bound conformation after superimposition of the corresponding molecule in the complex is represented in ribbon. (a) Unbound chymotrypsin; the structure of several bound proteins is represented: APPI in green (PDB 1ca0), BPTI in yellow (PDB 1cbw), Eglin C in magenta (PDB 1acb), ovomucoid in cyan (PDB 1cho). (b) Unbound kallikrein A; bound BPTI is shown in yellow (PDB 2kai); kallikrein A dimer is shown in green (PDB 2pka). (c) Unbound APPI; the structure of several bound proteins is shown: chymotrypsin is shown in green (PDB 1ca0); rat trypsin is shown in yellow (PDB 1brc); bovine trypsin is shown in gray (PDB 1taw). (d) Unbound TCR-β; bound TCR-α is shown in green (PDB 1tcr); SpeA is shown in yellow (PDB 1l0x). (e) Unbound acetylcholinesterase; bound fasciculin is shown in green (PDB 1fss). (f) Unbound Fab; bound lysozyme is shown in green (PDB 1mlc).

dence. In some cases, the ODA hot spots were located not very far from the known interface, so a possible explanation is that these low docking surface energy regions might be important in attracting the ligand during the association process, which could migrate later to the final bound conformation. This is highly speculative, and obviously future studies are needed in order to analyze the role of these low docking surface energy regions located outside the binding sites.

Proteins with No Significant ODAs

In about half of the analyzed molecules, we were not able to find any significant low-energy ODA (<-15.0 kcal/mol). In those cases, the physico-chemical character of the solvent-exposed surfaces seemed not to favor burial upon protein binding. For example, no significant ODAs were found on structurally homologous proteases subtilisin BPN, subtilisin Carlsberg and thermitase, as opposed to other proteases such as chymotrypsin and trypsin. The ODA analysis also showed that other homologous proteins such as bovine and rat trypsin (74% sequence identity) seem to have different surface properties, an interesting conclusion that would be difficult to draw from a mere analysis of their interfaces. Similar differences in surface properties were found between protease inhibitors. For instance, low-energy ODAs were found on Amyloid beta Precursor Protein trypsin Inhibitor (APPI), but not on Bovide Pancreatic Trypsin Inhibitor (BPTI). The ODA analysis could thus complement other structural physical-chemical studies in search of energy determinants for protein binding.

The absence of low-energy ODAs on a protein molecule could also indicate that desolvation is not important for its binding mechanism. This is the case for electron-transfer proteins, where no ODA hot spots were found. It has been proposed that the formation of this type of complex is diffusion controlled and driven by electrostatics.[1,38,39] The lack of low docking surface energy areas in these electron transfer complexes would be in accordance with the smaller contribution of desolvation to the complex formation. The ODA method could thus be used to identify whether association between two given proteins is likely to be driven by desolvation (if not, the complex formation would probably be driven by electrostatics).

Poor Predictions in NMR Structures

We found that prediction rates on NMR structures were poor: predicted sites were correctly located in only 36% of all NMR structures (databases A and B), as compared to the success rate of 80% for all X-ray structures (database C). Simultaneously, the number of cases with detected low-energy ODAs in NMR structures (79%) was significantly larger than the detection rate in X-ray structures (57%). This suggests that our method generated an excessive number of false positives in the former case. Surface side-chain packing in the NMR structures tends to be looser than in X-ray structures due to the inherent flexibility and/or the poorer resolution of surface side-chains (which have fewer contacts, and hence fewer NOEs, even if they are in fact quite rigid). Therefore, there may be portions of the solvent-exposed surface in NMR structures that would be buried in the X-ray structures. This overexposed surface could contribute to artificially lowering the surface energy of the ODA patches. Perhaps different ASP values would be needed to correctly detect ODAs in NMR structures, although the current set of structures is too small to find reliable values of general validity.

ODA Predictor in Antibody/Antigen Complexes

Low docking surface energy areas found on antibodies corresponded well with their binding sites [e.g. Fig. 3(f)]. Although we have only shown the results obtained from two cases of unbound antibodies, the method was also applied to the bound coordinates of several antibodies extracted from complexes with protein antigens, and the prediction rates were always close to 90–100% (data not shown). The excellent correlation between ODA hot spots and binding sites in antibody molecules may reflect the fact that antibodies sometimes need to bind antigen surfaces that are not particularly 'protein-philic.' Thus, their binding surfaces must have evolved to develop especially strong general binding properties, which could be obtained by means of favorable desolvation. On the contrary, ODA hot spots found on antigens, if any, are not generally located in the regions where the antibody binds (data not shown). This was somewhat expected, as antigen surfaces have not evolved to bind antibodies. This also seems to indicate that their interaction with antibodies is not driven by desolvation of specific regions of the antigens.

CONCLUSIONS

A method of analyzing protein surfaces in search of ODAs that could play an important role in protein–protein association has been developed. The fact that, in a majority of the cases examined, the identified ODA hot spots correspond with known protein–protein interfaces has interesting implications. First, it emphasizes the role of desolvation in protein–protein association, which could be especially relevant during the formation of encounter complexes.[26,27] In encounter complexes, many conformations with different mutual orientations of the interacting molecules can coexist and, at least in some complexes, be explained by considering only desolvation effects. Second, the practical implication of our study is that we can predict the presence of a protein-binding site by identifying ODA hot spots on the three-dimensional structure of the unbound protein. This also proves that, at least in cases where desolvation seems to be important for the association, it is possible to predict protein–protein interfaces from surface residue analysis in non-obligate protein–protein interactions, a field on which much effort is being focused. Our method goes beyond the standard residue composition characterization, relying instead on the computation of the water/interface transfer energy by automatic generation of a series of surface patches. The method could be used to analyze the role of these optimal docking areas in other types of interactions, such as obligate homo-dimers and domain–domain interfaces. In addition, the speed and accuracy of this new binding site predictor makes it highly suitable for application to the coordinates of individual proteins from current structural proteomics projects, and it could therefore be an important tool to characterize protein interaction networks.

REFERENCES

1. Gabdoulline RR, Wade RC. On the protein-protein diffusional encounter complex. J Mol Recog 1999;12:226–234.
2. Gabdoulline RR, Wade RC. Protein-protein association: investigation of factors influencing association rates by brownian dynamics simulations. J Mol Biol 2001;306:1139–1155.
3. Gabdoulline RR, Wade RC. Biomolecular diffusional association. Curr Opin Struct Biol 2002;12:204–213.
4. Novotny J, Bruccoleri RE, Saul FA. On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. Biochemistry 1989;28:4735–4749.
5. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
6. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
7. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–386.
8. Clackson T, Ultsch MH, Wells JA, de Vos AM. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. J Mol Biol 1998;277:1111–1128.
9. Sheinerman FB, Honig B. On the role of electrostatic interactions in the design of protein-protein interfaces. J Mol Biol 2002;318:161–177.
10. Vakser IA, Aflalo C. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins 1994;20:320–329.
11. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–729.
12. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog in Biophys & Mol Biol 1995;63:31–65.
13. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
14. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
15. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;6:53–64.
16. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001;43:89–102.
17. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003;53:708–719.
18. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003;325:377–387.
19. Lijnzaad P, Berendsen HJ, Argos P. A method for detecting hydrophobic patches on protein surfaces. Proteins 1996;26:192–203.
20. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins 1997;28:333–343.
21. Chothia C, Janin J. Principles of protein-protein recognition. Nature 1975;256:705–708.
22. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990;265:16027–16030.
23. Conte LL, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.

24. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47:334–343.
25. Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 2004;335:843–865.
26. Camacho CJ, Weng Z, Vajda S, DeLisi C. Free energy landscapes of encounter complexes in protein-protein association. Biophys J 1999;76:1166–1178.
27. Camacho CJ, Vajda S. Protein docking along smooth association pathways. Proc Natl Acad Sci U S A 2001;98:10636–10641.
28. Scarsi M, Majeux N, Caflisch A. Hydrophobicity at the surface of proteins. Proteins 1999;37:565–575.
29. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
30. Kortvelyesi T, Dennis S, Silberstein M, Brown L, 3rd, Vajda S. Algorithms for computational solvent mapping of proteins. Proteins 2003;51:340–351.
31. English AC, Groom CR, Hubbard RE. Experimental and computational mapping of the binding surface of a crystalline protein. Protein Eng 2001;14:47–59.
32. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
33. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.
34. Fauchere JL, Pliska V. Hydrophobic parameters-pi of amino-acid side-chains from the partitioning of N-acetyl-amino-acid amides. Eur J Med Chem 1983;18:369–375.
35. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004;338:181–199.
36. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
37. Lomize AL, Riebarkh MY, Pogozheva ID. Interatomic potentials and solvation parameters from protein engineering data for buried residues. Protein Sci 2002;11:1984–2000.
38. Hart SE, Schlarb-Ridley BG, Delon C, Bendall DS, Howe CJ. Role of charges on cytochrome f from the cyanobacterium *Phormidium laminosum* in its interaction with plastocyanin. Biochemistry 2003;42:4829–4836.
39. Worrall JA, Reinle W, Bernhardt R, Ubbink M. Transient protein interactions studied by NMR spectroscopy: the case of cytochrome C and adrenodoxin. Biochemistry 2003;42:7068–7076.
40. Strynadka NC, Jensen SE, Johns K, Blanchard H, Page M, Matagne A, Frere JM, James MN. Structural and kinetic characterization of a beta-lactamase-inhibitor protein. Nature 1994;368:657–660.