Predicting Protein Interaction Sites: Binding Hot-Spots in Protein-Protein and Protein-Ligand interfaces

Running Header: Binding Hot-Spots in Protein Interfaces

Nicholas J. Burgoyne, Richard M. Jackson*

Institute of Molecular and Cellular Biology,
Faculty of Biological Sciences,
University of Leeds,
Leeds,
LS2 9JT, UK

* To whom correspondence should be addressed
Tel: +44 (0)113 343 2592, Fax: +44 (0)113 343 3167
Email address: r.m.jackson@leeds.ac.uk
ABSTRACT

MOTIVATION: Protein assemblies are currently poorly represented in structural databases and their structural elucidation is a key goal in biology. Here we analyse clefts in protein surfaces, likely to correspond to binding “hot-spots”, and rank them according to sequence conservation and simple measures of physical properties including hydrophobicity, desolvation, electrostatics and van der Waals energies, to predict which are involved in binding in the native complex.

RESULTS: The resulting differences between predicting binding-sites at protein-protein and protein-ligand interfaces are striking. There is a high level of prediction accuracy (≥93%) for protein-ligand interactions, based on the following attributes: van der Waals potential, electrostatic potential, desolvation and surface conservation. Generally, the prediction accuracy for protein-protein interactions is lower, with the exception of enzymes. Our results show that the ease of cleft desolvation or “de-wetting” is strongly predictive of interfaces and strongly maintained across all classes of protein binding interface.
INTRODUCTION
Protein interactions are critical for all aspects of cellular function, and determining how they interact has become a major goal in the post-genomic era (Sali et al., 2003; Rhodes et al., 2005). An emerging new approach is to take advantage of structural information to predict physical binding. In particular, prediction of protein binding-sites can guide the structural elucidation of protein complexes, allowing function prediction for numerous unannotated structural genomics targets and the design of molecules that can modulate biological function at a systems-level (Russell et al., 2004).

Protein-protein interfaces are generally considered to be circular and relatively flat (Jones and Thornton, 1997), studies of their properties have also shown that hydrophobic area prevails while electrostatic residues and hydrogen-bonding groups are evenly spread across the surface (Xu et al., 1997; Lo Conte et al., 1999), suggesting a uniform distribution of properties across the binding interfaces. However, alanine-scanning mutagenesis has shown that the stability of a complex is determined by only a fraction of the interface residues. This was first shown in the complex between human-growth hormone and the human growth hormone binding protein (Clackson and Wells, 1995). They showed that a far greater loss in affinity was seen when two tryptophan residues were mutated to alanine when compared with other mutations made in the interface. Similar results have been collated for a number of other protein complexes (Bogan and Thorn, 1998; Thorn and Bogan, 2001). Analysis of this data shows that these “hot-spot” residues are usually found in the centre of the interface, and are surrounded by residues that have a lesser effect on stability (Bogan and Thorn, 1998). Structural analysis has shown that the clusters of
conserved, hot-spot, residues form clefts in the surface of the protein (Li et al., 2004). Enrichment of hot-spot areas with tryptophan, tyrosine and arginine residues have been shown based on protein family conservation (Hu et al., 2000; Ma et al., 2003; Keskin et al., 2005). However, other highly sequence conserved residues, including polar residues, may also be conserved (Hu et al., 2000). Solvent exclusion around polar or charged interactions lowers the effective dielectric constant thus strengthening the interaction. It has been suggested that the hydrophobic effect of hot-spot residues is almost double that of other residues in an antibody/antigen interface (Li et al., 2005). Multiple hot-spot regions can be found in a single interface (Ma et al., 2003), these tend to cluster at the centre of the interface, and interact with equivalent hot-spot residues in the binding partner protein (Halperin et al., 2004; Keskin et al., 2005). The clustering of conserved residues has been used to predict possible interacting partners based on alignment to a known pair of interacting proteins (Aytuna et al., 2005; Espadaler et al., 2005). Clefts in the protein surface are also important for the prediction of protein-ligand interactions. Indeed, ligand binding-sites can often be predicted by identifying the largest exposed cleft (Laskowski et al., 1996). Ligand binding-sites may also be detected as cleft regions with high predicted binding affinity, based on energetic contouring with interacting molecular probes (Laurie and Jackson, 2005).

Here we analyse the ability of different key physicochemical attributes and other binding surface properties, such as surface conservation, to predict the binding interface in protein-protein and protein-ligand complexes. Our approach attempts to define binding hot-spots on the protein surface. These are defined by clefts on the protein surface that are further described by their key attributes. We predict clefts on
the protein surface using Q-SiteFinder (Laurie and Jackson, 2005) an energy-based method for the prediction of protein binding clefts. The properties of these predicted clefts are then investigated and compared. Given that the key attributes we use to describe clefts are the same for both protein-protein and protein-ligand interactions, the resulting differences we find between their binding-sites are striking. This has important implications for understanding the nature of protein interactions, for identifying the suitability of sites as drug targets, and for identifying critical regions for docking and structure-based drug design.

METHODS

Interaction Datasets

Enzyme-inhibitor and antibody-antigen complexes of the protein-protein docking benchmark 1.0 (Chen et al., 2003) were supplemented with a selection of other available protein complexes to create a protein-protein interaction dataset of 97 pairwise non-obligate hetero-complexes. These additional complexes satisfied the following criteria: 1) No two complexes had a sequence-identity of more than 60% in an alignment over more than 80% of the sequence of both proteins involved in the interaction. 2) In keeping with the other complexes of benchmark 1.0 no complex had a resolution of greater than 3.25Å. 3) No complex had large disordered regions in the interface. Oligomeric interfaces for all obligate protein-protein interactions of the dataset were obscured by using the appropriate multimeric complex as the interacting monomer. This was achieved by analysis of the protein quaternary structure database (Henrick and Thornton, 1998). As such, monomers mentioned in the text may refer to protein structures of more than one peptide chain. Occasionally this left several independent interfaces between interacting proteins. In the cases where more than one
independent interface to identical proteins exists, all interfaces of the protein are considered. Where proteins have only one interface only that interface was assessed. Cognate ligands required for the function of the protein were retained, while those not indicated as such in the literature were discarded. The final dataset consisted of 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, and 56 protein complexes that could not be included in either classification (Other complexes). Details of the proteins used can be found in the supplementary material.

134 ligand-protein complexes from the Gold dataset were used (Nissink et al., 2002), and prepared as described by Laurie and Jackson (2005). These constitute a subset of the full Gold-dataset where proteins of high structural similarity are removed. The dataset was further classified into protein-ligand interactions in enzymes (95 complexes) and non-enzymes (39 complexes). The binding-site of a single ligand is assessed for each protein. Additional ligands covalently bound to the protein are treated as protein while other solvent molecules are removed. As with the protein-protein interaction dataset the interacting proteins are assessed in the absence of the binding molecule.

**Pocket Detection and Occupancy using Q-SiteFinder**

Clefts were identified in the protein-surface using Q-SiteFinder (Laurie and Jackson, 2005). The method is briefly described. A grid is built around each protein, such that the spacing between grid points is 0.9Å in all orthogonal-directions and the entire protein is covered. The non-bonded interaction energy is then calculated using the GRID forcefield (Goodford, 1985) parameters for every grid-point that is sufficiently far from the protein to allow a methyl (-CH₃) group to be positioned without steric
overlap with any protein atom (Jackson, 2002). Probes with calculated interaction energies of less than a -1.3 kcal/mol are retained and clustered according to spatial proximity whereby no probe in any cluster has a centre that is more than 1.0Å from the centre of another probe in the same cluster. The total non-bonded interaction energy is then calculated across all probes in the cluster and serves as the means by which the clusters are ranked. The highest ranking site is the cluster with the highest cumulative interaction energy. Occupancy of a cleft is defined as the percentage fill of the cleft by atoms of the binding protein or ligand; a threshold of 25% is applied to define a successful cleft. Proteins with no occupied clefts were excluded from further analysis. These consisted of the larger proteins of two protein-protein interactions 1HE8 and 1DE4, and six protein-ligand complexes (1CGY, 1DR1, 1HDY, 1LDM, 1PBD and 2PDM).

**Pocket Ranking**

**Hydrophobicity**

The atomistic Solvent Accessible Surface Area (SASA) covered by each cleft was calculated by NACCESS (Hubbard and Thornton, 1993). The solvent accessibilities of all protein atoms were calculated in both the presence and absence of the probes of each cluster. The difference between these values represents the covered area of the cleft. Hydrophobic area was defined as the area exposed by atoms that are either carbon or sulphur (as in NACCESS). The cleft with the highest proportion of hydrophobic accessible atom area was ranked first.

**Desolvation**
The entropy associated with the removal of water from the surfaces covered by the probe clusters was estimated using coefficients for the transfer of N-acetyl derivative amino-acids between water and octanol. The total transfer energies have previously been converted to atomistic-based values by linear fitting (Fauchere and Pliska, 1983), and optimized for protein-protein docking (Fernandez-Recio et al., 2004). The coverage of each atom was calculated as described above, allowing the desolvation energy of the cleft to be calculated by multiplying the area by appropriate solvation parameter taken from (Fernandez-Recio et al., 2004). Clefts were ranked such that the cleft most easily desolvated was ranked first.

Electrostatics

The peak, average, and total electrostatic potential for each cleft was calculated using the DelPhi v.4 package (Rocchia et al., 2001; Rocchia et al., 2002). Protons were added to the protein as described by Q-SiteFinder (Laurie and Jackson, 2005). A grid of 101 points in all dimensions was built around each protein such that the molecule occupied 50% of the cubic grid’s volume. Amber charges and radii were used to describe the protein and associated cognate ligands (protein-protein interface analysis only) (Giammona, 1984; Weiner et al., 1984; Schneider and Suhnel, 1999; Antony et al., 2000; Meagher et al., 2003). A dielectric boundary condition was used, with the dielectric constants of 4 within and 80 outside the molecular surfaces defined by a water probe radii of 1.4Å. Salt concentration in the solute was set at 0.15M and a 2Å exclusion radius was applied around the protein. The finite difference Possion-Boltzmann calculations were performed until the potentials at the grid points converged to 0.0001 kT e⁻¹, after which the potentials were extrapolated to the centre of the individual probes. The total cleft potential was defined as the sum of the
modulus of individual probe potentials that form the cluster, while the average potential was this total value divided by the number of probes in the cluster. The peak values for each cleft were defined as the highest modulus individual probe potentials of each cluster. The same calculations were repeated with all proteins with a unit charge applied to all atoms. The same Amber atom radii were used as described, as were all other variables bar the convergence criteria. Unit charge calculations were run until they converged to within 0.01 kT e⁻¹.

**Conservation**

The surface conservation for each cleft was calculated by extracting the protein sequences of each chain in the protein from the associated ATOM records of their Protein Data Bank coordinates. For each chain of the protein, close-homologues were found by performing a PSI-BLAST search over the non-redundant Swiss-Prot database release 47.6 (Altschul et al., 1997; Bairoch et al., 2005). Three iterations are performed and the search refined using sequences with similarity to the query sequence defined at an e-value of less than 0.001. Redundant sequences were removed and the remaining full-length sequences were then aligned by Muscle (Edgar, 2004). The conservation of each position in the alignment relative to the initial protein sequence was calculated by Scorecons (Valdar, 2002). Conservation scores are based on the sum of the weighted pairwise exchange of residues, for each pair of sequences, at each position in the alignment. Values lie between 0 (no conservation) and 1 (completely conserved). A Solvent Accessible Surface Area was calculated for the protein structure of the same chain by MSMS with vertices spread evenly across the protein surface at a density of 1 vertex per Å² (Sanner et al., 1996). The conservation scores for each residue were then mapped to the appropriate
vertices. Cleft conservation was defined as the average conservation score of vertices within 2.0 angstroms of the centre of a probe belonging to the cluster.

**Calculation of True and False Positive Rates**

The True Positive Rate (TPR), or sensitivity, was calculated as the number of genuine interface clefts that are ranked in the top \( k \) clefts (true positives, TP) divided by the total number of interface clefts identified for this monomer (TP plus false positives, FP), where \( k \) increases by one each time. The False Positive Rate (FPR), one minus specificity, was the number of clefts ranked in the top \( k \) clefts that are not in the interface (FP) divided by the total number of non-interface clefts (FP plus true negatives, TN). Equal TPR and FPR values at all values of \( k \) indicate no discrimination by the ranking method.

The sensitivity (TPR = TP/(TP+FN)) and false positive rates (FPR = FP/(FP+TN)) were calculated for the clefts ranked by the above attributes. An attribute that is a successful predictor of interface clefts will show high sensitivity and a low error rate. Plotting TPR against the FPR gives a Receiver Operating Characteristics (ROC) curve.

In order to give a single measure of prediction accuracy, which is independent of any decision threshold, we have calculated the ROC integral or Area Under the Curve (AUC) (Hanley and McNeil, 1982). A value of 0.5 indicates no correlation, a value less than 0.5 indicates negative correlation and a value approaching 1.0 indicates the theoretical maximum or perfect prediction.
RESULTS

Analysis of protein Clefts

Q-SiteFinder was used to define the clefts found on the protein surface (see methods). Q-SiteFinder was run on both bound monomers of each hetero-protein complex in the non-obligate protein-protein interaction data sets and on all proteins in the protein-ligand data set. Following evaluation at different probe interaction thresholds a value of -1.3kcal/mol was chosen, based on the success of the method in protein-ligand complexes around this value (Laurie and Jackson, 2005) and also based on visualisation of the clefts involved in protein-protein interactions, since it also best describes the clefts of the protein that were filled by the sidechains of the opposing binding protein. We found that altering the threshold did not greatly affect the predictive power for protein-protein interactions.

Q-SiteFinder was used to calculate 99 clefts on the protein surface. Occupied or true cleft predictions for each protein are defined as those occupied to more than 25% of their volume by atoms of the interacting molecule. Figure 1a shows that for all 194 monomers of the protein-protein dataset there is a slightly skewed distribution of the number of true interface clefts per protein monomer with a peak around six clefts. The separate distributions of smaller and larger monomers show differences. The peak number of successful interface clefts for the smaller (ligand) monomers, at six, is higher than the peak, at 2-4 clefts for larger (receptor) monomers, however, the latter has a broad peak spanning 2-9 clefts. Figure 1b shows there are approximately equal percentages of interface areas covered in both sets of monomers. Protein-ligand receptors show far fewer occupied clefts with a clear peak at one. Also true interface clefts have a greater coverage than do the non-interface regions in contrast to protein-
protein interfaces. However, given the large standard deviations it is not possible to conclude this observation is significant. Below we attempt to separate those clefts that are occupied from those that are not, based on properties that have been suggested to be important for the prediction and stability of intermolecular interactions.

Figure 1a. Distribution of the number of occupied interface clefts across the 97 protein-protein complexes and 134 protein-ligand complexes, as generated by Q-SiteFinder.
Figure 1b. Average coverage of the protein surface area; by the predicted clefts generated by Q-SiteFinder in 97 protein-protein complexes and 134 protein-ligand complexes. Error bars represent the standard deviations of each distribution.

Analysis of Cleft Properties

Intermolecular interactions are stabilised by a number of factors, these include the burial of areas of hydrophobicity, the formation of hydrogen-bonds and electrostatic complementarity (Chothia and Janin, 1975). We have found clefts in the protein surfaces using Q-SiteFinder and re-ranked them according to simple measures of the cleft properties (see methods). All the different properties were used to rank the clefts to assess their ability to predict which ones are involved in binding in the native complex. Receiver Operating Characteristics (ROC) curves are used for this purpose (see methods). Whilst, most of the properties used (van der Waals, electrostatic, hydrophobicity, desolvation, and conservation) are readily interpretable, the Unit electrostatic properties involve the calculation of electrostatic potential from a uniform charge density on all protein atoms, as opposed to the standard atom-specific charges. The method typically generates large potentials in larger or enclosed clefts (Bate and Warwicker, 2004), and does not reflect atom-specific electrostatic properties of the site.
Figure 2. Receiver Operating Characteristics graphs for the prediction of interface clefts in protein-protein (a-e) and protein-ligand (f) complexes. See text for a description.
**Enzyme-Inhibitor Complexes**

Results of re-ranking the clefts for the enzyme-inhibitor monomers are shown in figure 2a and 2b. The results for the enzymes (receptors) show that surface residue conservation is the best predictor of true protein interface regions as well as for their protein inhibitors (ligands). However, it is a much better predictor for enzymes, in agreement with the study of Bradford and Westhead (2003). The unit electrostatic properties (unit peak and average) are also good predictors of interface regions in the enzymes but show no predictive power (or even slight anti-correlation) in the inhibitors. These results can be rationalized in terms of known protein function, in that the substrate/inhibitor binding cleft of the enzymes is conserved in sequence due to the conserved nature of the catalytic mechanism, and also active sites are often enclosed, pre-organised clefts, a property that may be important for enzyme catalysis. For example, the serine-protease family, are well represented in the enzyme class. The sequence conserved catalytic oxy-anion hole in serine proteases is a deep cleft which accommodates the conserved binding conformation of the protein inhibitor “canonical” loop (Jackson and Russell, 2000). Therefore it is unsurprising that unit electrostatic properties capture this characteristic for the enzymes only. Both of these cleft characteristics may capture the characteristics of the binding clefts of enzyme protein families very well but are not expected to generalize well to other protein-protein interactions. The only other strong predictor of interface regions is ranking clefts by desolvation energy. This is discussed further below.

**Antibody-Antigen Complexes**

The ROC curves for the predictions of the interface clefts of the 38 bound monomers of the 19 antibody-antigen complexes can be seen in figure 2c and 2d. Although the
results for the antigens (ligands) do not show any strong trends, those for the prediction of interface clefts in the antibodies (receptors) do. The highly significant anti-correlation of residue conservation in the antibodies is striking, yet entirely logical given the biological role of antibodies. This reflects antibody structure in which highly sequence diverse loops form the complementarity-determining regions (CDRs) and antigen binding interface. The CDRs determine specificity on what is otherwise a highly conserved protein framework. Both Q-SiteFinder and total unit electrostatics scores show some weak predictive capacity that may reflect their ability to define the size and position of clefts between the loops that form the CDRs of the antibodies. However, the results for the antibodies show that cleft desolvation energy is by far the best predictor of protein interface regions. It is interesting to see that it is also the most prominent of all attributes in the antigen proteins as well, albeit to a much lesser degree.

**Other Protein-Protein Complexes**

The ROC curves for the 112 bound monomers of the 56 ‘other’ complexes can be seen in figure 2e. Unlike antibody and enzyme results there are few strong correlations to be seen in either the ligand or receptor sets of proteins (see supplementary data). This is probably because this is a highly non-homogenous set of protein-protein complexes. This diversity almost certainly neutralizes protein sub-class specific characteristics such as those seen in the enzyme and antibody sets. The strongest correlation between interface cleft and high rank in both the receptor and ligand proteins is the desolvation energy. Conservation is only weakly predictive, and fails to re-rank the interface clefts highly for a more diverse set of protein complexes, consistent with the results of Caffrey et al, (2004). By changing the area over which
conservation is assessed (from a large patch in Caffrey et al’s investigation, to a small cleft in ours) it was hoped to improve these results in line with the findings of Li et al (2004). Li et al showed that the conserved residues of interface regions cluster around clefts in the protein surface. By re-ranking the clefts according to conservation score, we anticipated an improved relationship for the prediction of interfaces. However this was not the case. Consistent with the results of all the protein-protein interfaces analyzed in this piece of work, desolvation energy is the most effective common factor in the identification of interface clefts.

**Protein-Ligand Interactions**

The ROC curves for the 134 protein-ligand complexes can be found in figure 2f. The difference between these results and those of the protein-protein interface clefts is striking. Unlike the protein-protein interfaces, several properties are excellent predictors of protein interface regions. The ranking of clefts based on Q-SiteFinder support both the relative merits of using this method for ligand binding-site prediction and the previous observation that this method has a 90% success rate in the top three predicted clusters when tested on the same dataset (Laurie and Jackson, 2005). Not so surprisingly the electrostatic total (based on amber charges) and unit electrostatic total show a very similar pattern, with the latter being slightly more successful than Q-SiteFinder. These two measures may be most dependent on the number of probe centres that make up the cleft, as defined by Q-SiteFinder, rather than the nature of the properties that define the cleft. Of all the attributes the unit electrostatics (average and peak) stand out and rank slightly above Q-SiteFinder overall, however, their initial prioritisation of clefts is in fact weaker. The finding that using unit charges (Laskowski et al., 1996) rather than the more physicochemically realistic amber
charges improves results (Bate and Warwicker, 2004) is also corroborated by our results. Other, strong predictors are cleft desolvation energy and surface residue conservation. The latter is not unexpected since active/ligand binding sites, have been shown to be sequence conserved across many different protein families (Bartlett et al., 2002) and consequently this property is useful in the identification of enzyme active sites (Greaves and Warwicker, 2005). This is further corroborated here by the difference in predictive power of conservation for enzymes versus non-enzyme protein-ligand complexes (see supplementary data).

With the exception of the enzymes, protein-protein interaction clefts have a poor correlation with each of the three measures for the unit electrostatics calculations, implying that clefts in protein-protein interfaces (defined by Q-SiteFinder) are fundamentally different to those in protein-ligand complexes. The observation that high electrostatics potential is indicative of a ligand binding-site seems at odds with the finding that the ease of desolvation of the cleft also correlates well. The ease of desolvation used in this analysis would be expected to favour non-polar surface with a high aliphatic or aromatic content rather than a charged surface. However, it is primarily the unit electrostatic properties that are highly predictive, and these may be most dependent on the shape, and depth of the cavity rather than true atom-base electrostatic properties. Furthermore, the poor predictive power of hydrophobicity in protein-ligand and most protein-protein interactions indicates fundamental differences between clefts defined by hydrophobicity and desolvation. This challenges the classical view that hydrophobicity, as defined by non-polar surface area, is a useful predictor of interaction interfaces (Chothia and Janin, 1975).
DISCUSSION

This study has assessed the applicability of using protein surface clefts for the prediction of protein-protein and protein-ligand interfaces. There are a number of multi-attribute methods for the prediction of protein-protein interfaces that use averages of properties across a patch of protein surface in order to determine whether that patch lies in an interface (Jones and Thornton, 1997; Zhou and Shan, 2001; Fariselli et al., 2002; Neuvirth et al., 2004; Bordner and Abagyan, 2005; Bradford and Westhead, 2005). Combinations of properties allow the prediction of protein-protein interactions with successes of up to 70%, despite treating the protein interface as an average of its properties. There is some evidence to suggest that the stability of a protein complex is not spread across the interface but instead localised around clefts in the protein surface. By identifying clefts in the protein surface we investigated whether any single property could discriminate those in the protein interface from those elsewhere on the protein surface.

An overall ROC integral or Area Under the Curve (AUC) gives a single measure of prediction accuracy, (Hanley and Mcneil, 1982) used in many scientific studies. A table of results for the AUC of all the studies presented, are given in the supplementary data. For all protein-protein interactions, conservation scores appear to be less effective than anticipated (AUC: 58%), confirming the results of Caffrey et al (2004) who concluded that protein-protein interfaces are not significantly more conserved than the rest of the protein surface. Only in the enzyme (AUC: 79%) and enzyme protein-ligand (AUC: 78%) interfaces is conservation of significant predictive value. Electrostatics also failed to show any general trends in the prediction of protein-protein interfaces on monomers (AUC: 45-54% for all interactions) other
than enzymes (AUC: 65%, electrostatic total) consistent with the theory that long-range electrostatics do not play an important role in the funnel concept of protein-protein interactions (Schlosshauer and Baker, 2004), but are rather used for orientation (Schreiber and Fersht, 1996). Of the properties studied in this investigation only desolvation energy of the clefts had any general predictive power across all protein-protein interaction types (AUC: 68%). Fernández-Recio et al. (2005) showed that desolvation scores over larger regions of protein surface area, as defined by circular patches, correlated with the interface in 58% of proteins. It appears that the cleft based approach employed here is more general. Charged interfaces, such as those between enzymes and inhibitors, and those of the antigens form the majority of interfaces that Fernández-Recio et al. failed to predict. Desolvation scores, as implemented by our cleft based analysis, correlate nearly as well with charged and antigen interfaces as they do with other types of interface, with the enzyme (AUC: 74%), inhibitor (AUC: 61%) and antigen (AUC: 62%) interfaces being predicted well. A similar high level of success is also seen in protein-ligand interfaces (AUC: 72%), which also have charged interfaces with a high electrostatic potential. The level of success may be due to the fact that in defining the protein surface in terms of clefts we are only sampling a portion of the protein interface, as opposed to the whole surface of a circular patch. It appears that sites of high electrostatic potential and favourable desolvation coexist independently in the same interface. This is supported by the fact that the correlation between sites ranked by these two attributes for both protein-protein and protein-ligand complexes is insignificant ($R^2 = 0.02-0.1$) and slightly negatively correlated (results not shown).
Protein-protein and protein-ligand interface clefts show several differences, as well as some similarities. The most obvious differences are the failure of electrostatic descriptors overall to identify protein-protein interface clefts. However, identification of enzyme active sites is possible by electrostatic methods, and the results were improved when using a unit charge applied to each atom in the protein (Bate and Warwicker, 2004). Our results confirm these observations. The power of electrostatic descriptors in the protein-ligand interactions, particularly those of the enzyme sub-set, where the ligands bind at active sites, is only maintained in the enzyme class of protein-protein interactions. In fact, the protein-protein enzyme class shows a high degree of correlation with protein-ligand interactions for other attributes, and both also show significant predictive power of Q-SiteFinder and residue conservation, in contrast to all other protein-protein interfaces. Therefore, it would appear that the binding-sites of enzymes are much more predictable than those of other protein-protein complexes. The very high predictive power of Q-SiteFinder (AUC: 88%) and all the unit electrostatic properties (AUC: 93%) for protein-ligand interactions is very encouraging and will allow the targeting of functionally important ligand binding-sites in functional genomics.

Overall, protein-protein and protein-ligand interface clefts do have similarities. Our results show that the ease of cleft desolvation or “de-wetting” of all classes of interface are strikingly similar and more indicative of interface regions than electrostatics, Q-SiteFinder, or residue conservation, which although highly predictive in some cases do not generalize well to all classes of protein interaction. Recent studies of the role of water in protein-protein association of the melittin tetramer (Barratt et al., 2005) and in protein-ligand binding (Liu et al., 2005) in mouse major
urinary protein (MUP) by Molecular Dynamics simulation illustrate the importance of the phenomenon of de-wetting in molecular interactions. In the former, the strongly hydrophobic surfaces induce the evaporation (drying) of water in the interface as the melittin protein subunits approach one-another. The authors conclude that sufficiently hydrophobic protein surfaces can induce a liquid–vapour transition providing the driving force towards protein association. In the later study, the MUP ligand binding-site cleft is pre-organised, hydrophobic and poorly solvated in the unbound form, which may explain the largely enthalpy (as opposed to entropy) driven thermodynamics of ligand binding. We have found that hydrophobicity of the surface does not correlate strongly with interface clefts in either protein-protein or protein-ligand interfaces. Hydrophobic, non-polar surfaces will also correlate closely with areas of low desolvation energy; however, in proteins that form non-obligate protein/ligand complexes (and hence must be independently stable in water) extended hydrophobic solvent exposed surfaces are unlikely to exist. Therefore, a more subtle phenomenon exists, a relationship between ease of desolvation (de-wetting) and interface propensity which may be an intrinsic feature of all protein binding surfaces.

ACKNOWLEDGEMENTS

The authors would like to thank Alasdair Laurie for help with Q-SiteFinder and the protein-small molecule dataset. NJB is funded by the Medical Research Council.
REFERENCES


