Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Computational analyses of the surface properties of protein–protein interfaces

Jan Gruber,‡ Alexander Zawaira,‡ Rhodri Saunders, C. Paul Barrett and Martin E. M. Noble*

Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 3QU, England

‡ These authors contributed equally to this work.

Correspondence e-mail: martin.noble@biop.ox.ac.uk

Received 18 July 2006
Accepted 6 November 2006

Several potential applications of structural biology depend on discovering how one macromolecule might recognize a partner. Experiment remains the best way to answer this question, but computational tools can contribute where this fails. In such cases, structures may be studied to identify patches of exposed residues that have properties common to interaction surfaces and the locations of these patches can serve as the basis for further modelling or for further experimentation. To date, interaction surfaces have been proposed on the basis of unusual physical properties, unusual propensities for particular amino-acid types or an unusually high level of sequence conservation. Using the *CXXSurface* toolkit, developed as a part of the *CCP*4*MG* program, a suite of tools to analyse the properties of surfaces and their interfaces in complexes has been prepared and applied. These tools have enabled the rapid analysis of known complexes to evaluate the distribution of (i) hydrophobicity, (ii) electrostatic complementarity and (iii) sequence conservation in authentic complexes, so as to assess the extent to which these properties may be useful indicators of probable biological function.

1. Introduction

In recent years, substantial effort has been directed towards characterizing the properties that distinguish the parts of a protein that are involved in molecular recognition (*e.g.* Jones & Thornton, 1996; Lo Conte *et al.*, 1999; Ma *et al.*, 2003; Teichmann, 2002). The reasons behind this are twofold. Firstly, there is the scientific goal of understanding the physical principles that underlie the exquisite molecular-recognition processes that permit fidelity in processes such as signal transduction. Secondly, there is the more applied goal of using structures to contribute to the functional annotation of genomic projects.

For this latter goal, the role of analysing the properties of structurally characterized protein–protein interaction sites is to inform the analysis of newly determined structures so as to permit identification of parts of the molecule that are likely to be involved in protein–protein (or other) interactions. A further application of this approach is to exploit knowledge of authentic interfacial properties in validating hypothetical models in which proteins have been docked together.

The drive to characterize protein–protein interactions has given rise to a number of computational tools that may be used either in analysing a newly determined protein structure or in using three-dimensional structural information to assist in studying a novel gene sequence. Application of such tools is a way to maximize the benefit that may be derived from structural data, particularly when used to generate functional

hypotheses that can be subsequently tested by experimental techniques such as site-directed mutagenesis.

This article reviews a subset of analytical tools that characterize a protein interaction site by mapping quantitative descriptors of a property of that protein onto a triangulated surface representation. This approach effectively filters the possible set of descriptors of a molecule so as to focus on that part of the molecule that is responsible for its interactions, namely the molecular surface. We find that this approach can 'sharpen' the signal that demonstrates that properties (such as hydrophobicity, electrostatic complementarity and sequence conservation) are more typical of a protein–protein interface than of a protein surface in general.

2. Conservation

Descriptive studies in which the extent of conservation of interface residues is compared with the extent of conservation of the rest of the protein surface have led to the generally accepted view that active-site and ligand-binding site residues are more conserved than general surface residues across many different protein families (Grishin & Phillips, 1994; Ouzounis *et al.*, 1998; Bartlett *et al.*, 2002; Caffrey *et al.*, 2004). This result is perhaps not surprising considering that the precise arrangement of residues required for catalysis and ligand binding is expected to impose strong constraints on the evolution of sequences and structures (Shakhnovich *et al.*, 2003; Torrance *et al.*, 2005).

Furthermore, predictive studies have shown that clusters of residues that make up active sites or ligand-binding sites are invariably more conserved than clusters of residues defined elsewhere on the surface of a protein. These results show that conservation analysis is of predictive value in the identification of active sites and ligand-binding sites using sequence-based (Watson *et al.*, 2005), structure-based (Laskowski *et al.*, 2005*a*,*b*) or mixed strategies (Watson *et al.*, 2005).

On the other hand, the role of conservation is less clear for protein–protein interfaces (Grishin & Phillips, 1994; Valdar & Thornton, 2001; Caffrey *et al.*, 2004). The generally accepted model for the variation of the rate of evolution of proteins is

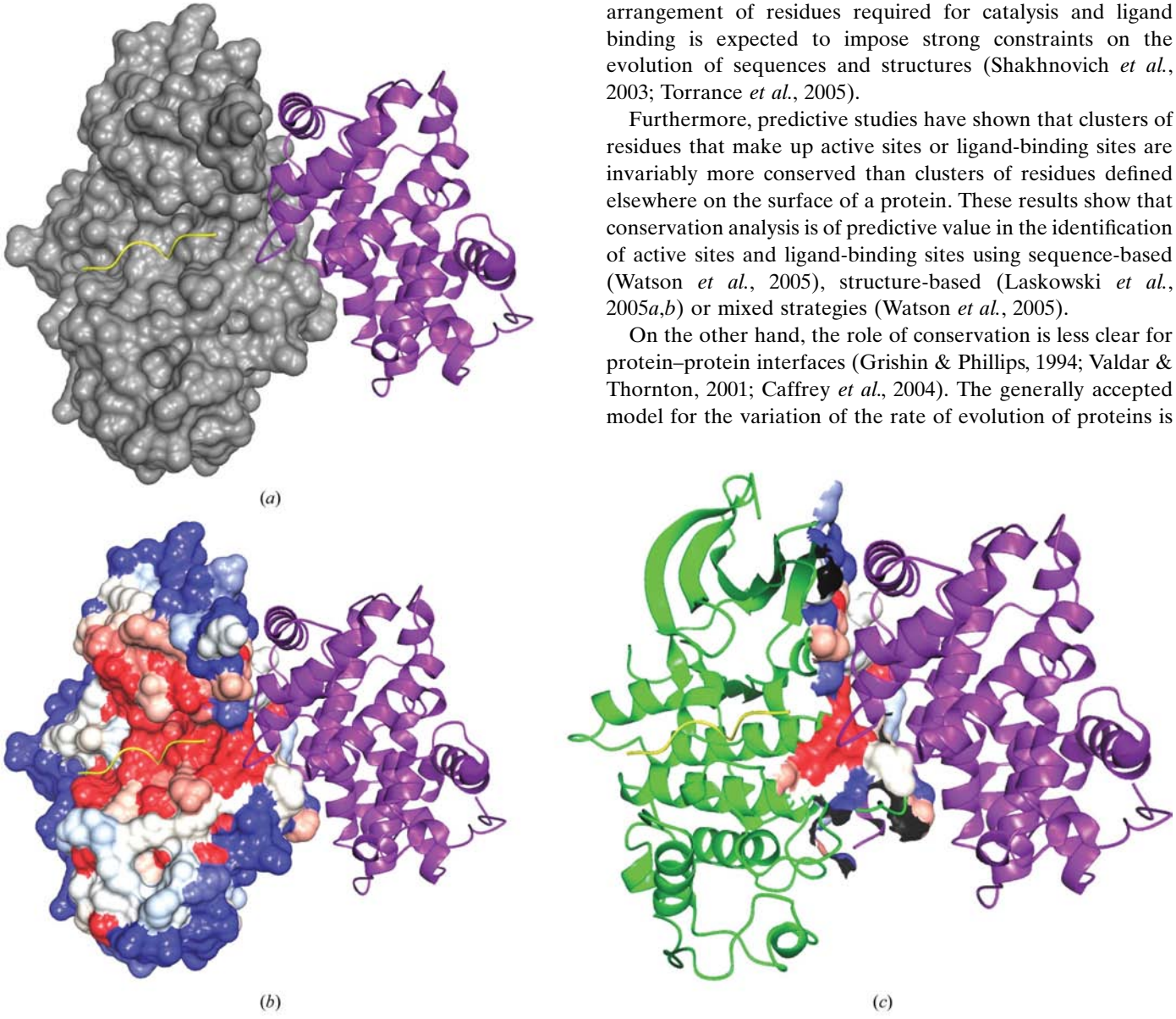


Figure 1
Example of a conservation-mapped molecular surface and an interfacial subset. (*a*) A molecular surface is generated from the CDK2 chain in a structure of the CDK2–cyclin A complex (PDB code 1qmz; Brown *et al.*, 1999). The cyclin molecule is shown in purple and the molecular surface in grey. The peptide substrate is shown in yellow. (*b*) In the next step of the analysis, conservation scores calculated from a multiple sequence alignment of cdc2 functional homologues are projected onto the molecular surface. The molecular surface is now coloured in shades of red (high conservation), white (intermediate conservation) and blue (low conservation, *i.e.* high variability). (*c*) The CDK2–cyclin A interface is extracted by identifying that part of the CDK2 molecular surface that is buried by the cyclin A molecule upon complex formation.

one in which the rate of evolution increases (*i.e.* conservation decreases) from the catalytic site to the protein core, to substrate-specificity sites and finally to surface regulatory regions (Valencia, 2005).

Conservation-analysis studies usually perform separate statistical analyses of homodimeric and heterodimeric protein–protein interfaces. Using sets of diverse homologues, Caffrey and coworkers have found that in homodimers, while the interface residues generally have higher conservation scores than the total surface residues (in 17/42 cases), the result is not statistically significant (P value 0.388; Caffrey *et al.*, 2004). Grishin & Phillips found similar results for homodimeric enzyme interfaces (Grishin & Phillips, 1994). Caffrey and coworkers also found that in heterodimers interface residues have larger conservation scores (in 11/12 interfaces) than residues making up the total protein surface. This result was also found to be statistically significant (P value = 0.0387).

Structure-based evolutionary methods seek to predict functional sites by sampling the evolutionary histories of gene families and projecting the information onto the structure of a single (presumably representative) structure of a member of the gene family. The methods involve a step in which a phylogenetic lineage of the sampled gene family is constructed. Evolutionary-trace (ET) analysis (Lichtarge *et al.*, 1996) is the most widely implemented form of such evolutionary methods. ET exploits a phylogenetic tree or sequence-identity dendogram to rank residues by evolutionary importance. For example, the colour red is assigned to those conserved in all groups and green to those conserved in at least one (but not all) groups. The ranks are then mapped onto a representative structure (*e.g.* the structure of one of the sequences in the analysis). It has been found that the highest ranked residues often cluster together and can be used to identify interaction sites (Yao *et al.*, 2003).

These observations suggest that significant functional insight can be derived from visualization of the distribution of sequence conservation when mapped onto representations of protein structure. Glaser and coworkers have implemented this approach in the *ConSurf* server (Glaser *et al.*, 2003) by assigning a conservation score to sequence positions of a protein of known structure so that a VDW representation can be used to identify more conserved patches that.

We have mapped sequence conservation onto triangulated molecular-surface representations of a protein. In addition to permitting visualization of the conservation of amino acids that form the molecular surface (*e.g.* Fig. 1*b*), this approach has allowed us to re-evaluate the extent to which sequence conservation is a statistically significant property of protein–protein interfaces using a scoring system in which the conservation score of a residue is weighted by the surface-area contribution made by that residue to the surface of interest.

This is a departure from traditional conservation-analysis studies (Grishin & Phillips, 1994; Valdar & Thornton, 2001; Caffrey *et al.*, 2004; Nimrod *et al.*, 2005), in which analyses are performed on whole residues (surface and interface) without taking into account the relative size of the contribution that the residue may make to a surface.

To assess which sequences should be included in the multiple sequence alignment (MSA) from which conservation scores are calculated, a strategy was used in which an MSA for the protein class of interest was manually compared with a structure-based (*HOMSTRAD*; Mizuguchi *et al.*, 1998) alignment of members of that family. If the MSA was significantly poorer than the *HOMSTRAD* alignment, the most divergent sequence was deleted and the MSA was recalculated. This procedure was iterated until the MSA closely resembled the *HOMSTRAD* alignment.

The conservation scores of the columns of the final MSA were calculated by a scheme that uses a normalized BLOSUM62 (Henikoff & Henikoff, 1992) substitution matrix to quantify the total pairwise similarity of all possible residue pairs that can be assembled within each column of the alignment. The conservation score $[C(x)]$ of a column within a set of aligned sequence was calculated using

$$C(x) = \frac{\sum_{i}^{N}\sum_{j>i}^{N} M[s_i(x), s_j(x)]}{\left\{(N-1) \times \sum_{j}^{N} M[s_j(x), s_j(x)]\right\}}, \qquad (1)$$

where $s_i(x)$ is the amino acid at column x in the ith sequence and N is the total number of sequences in the alignment. $M(a, b)$ is the similarity between amino acids a and b. The similarity matrix M is derived from the BLOSUM62 substitution matrix $[m(a, b)]$ using the transformation

$$M(a, b) = \frac{m(a, b)}{[m(a, a) \times m(b, b)]^{1/2}}. \qquad (2)$$

This transformation has been suggested to permit the use of a substitution matrix to measure amino-acid similarity (Valdar, 2002). These conservation values are passed on to the corresponding residue in the sequence of the member of the gene family whose structure is known (also present in the alignment).

Each vertex of the triangulated surface inherits the conservation score of the atom that immediately underlies it. Each triangle (T) of the surface can thus be assigned a conservation score $[C(T)]$, calculated as the mean conservation score of the triangle's three vertices. A surface or surface patch (S) can be assigned a mean conservation score $[C(S)]$ by forming the sum

$$C(S) = \frac{\sum_{T\in S} C(T) \cdot A(T)}{\sum_{T\in S} A(T)}, \qquad (3)$$

where $A(T)$ is the area of triangle T.

Fig. 1 shows an example of the generation of spatial conservation patterns of molecular surfaces and of the extraction of the interfacial subset of the surface. This subset corresponds to those triangles of the molecular surface, generated from the atoms of one of the chains in the complex, that become buried upon formation of a complex with its partner. Apparent immediately is the increased level of

Table 1
Analysis of the relative conservation of interface and non-interface surfaces in a set of homodimers and heterodimers.

Interface type	Total No. of interfaces studied	No. in which interface is more conserved than non-interface molecular surface	*P* value
Homodimer	18	10	0.3927
Heterodimer	27	25	4.32×10^{-4}

conservation of surface residues involved in binding between the two proteins and at the protein–substrate interface.

We have evaluated $C(S)$ for the interface and non-interface surfaces of a set of homodimers and heterodimers. The results we obtained from analysis of spatial conservation patterns of heterodimers and homodimers are summarized in Table 1. In agreement with other workers (Caffrey *et al.*, 2004; Grishin & Phillips, 1994; Valdar & Thornton, 2001) our results show that (i) in heterodimers the interface is more conserved than the molecular surface and (ii) in homodimers the interface is no more conserved than the molecular surface. The *P* value we obtained in the analysis of heterodimers (0.0004322) indicates greater significance than that obtained by Caffrey and coworkers (0.0387; Caffrey *et al.*, 2004). This can be explained by a number of factors: (i) our sample is larger than that used by Caffrey and coworkers, (ii) we have used different criteria in selecting sequences for building the MSAs and (iii) our approach includes consideration of three-dimensional structural information, namely the fractional surface area of each residue involved in the interface.

Irrespective of which factor contributes most significantly, these results may be rationalized in terms of the comparison between the conceivably more complicated co-evolutions of two independent proteins that are constrained to maintain an heterodimeric interface and the conceivably simpler evolution of a single sequence constrained to maintain the homodimeric interface.

Furthermore, from a visual survey of the conservation patterns for the data set used, it is evident that the structure of conservation patterns in interfaces is complex. Areas of high conservation are embedded within areas of high variability, *i.e.* the distribution of conservation within interfaces is generally uneven, with higher levels of conservation generally found at the centre of an interaction site. Other workers have noted similar results in analyses where conservation scores are projected onto residues (Caffrey *et al.*, 2004; Valdar & Thornton, 2001).

3. Physical properties

3.1. Hydrophobicity

Water molecules in bulk water form a network of short-lived hydrogen bonds participating on average in 3.5 hydrogen bonds at any one time. An intuitive model of hydrophobicity proposes that water molecules close to a hydrophobic surface are unable to form hydrogen bonds with the nonpolar atoms of the solute and therefore have a restricted choice of orientations. Water close to the surface is therefore more ordered than in bulk solvent. The ordering of water molecules in this way would lead to a local decrease in entropy and hence an unfavourable free energy of solvation (Frank & Evans, 1945; Dill, 1990). For locations close to extended hydrophobic surfaces it might be impossible to form the maximum number of hydrogen bonds, making these positions enthalpically unfavourable as well. It is the lack of opportunities for hydrogen bonding that renders surfaces hydrophobic.

The importance of hydrophobicity for protein stability was postulated in 1945, when it was recognized that protein molecules contain nonpolar groups that would be exposed to solvent in an unfolded protein (Southall *et al.*, 2002). This intuition was confirmed by the first protein structure showing that hydrophobic residues are indeed preferentially buried in the protein interior (Kendrew *et al.*, 1958). Walter Kauzmann introduced the term 'hydrophobic bonding' to describe interactions driven by exclusion from water (Kauzmann, 1959).

While continuum electrostatics provides a relatively well understood framework for the analysis of charge–charge interactions, there is still no consensus about an appropriate treatment of the hydrophobic effect.

Traditionally, amino acids have been roughly classified as either hydrophobic, aliphatic or charged/hydrophilic. There have been a number of attempts to make this classification more quantitative by introducing continuous scales of hydrophobicity (*e.g.* Nozaki & Tanford, 1971; Radzicka *et al.*, 1988). Hydrophobicity scales are commonly derived from the partitioning of model compounds between an aqueous and an oil-like phase. The use of such scales is based on the assumption that the relative hydrophobicity is independent of the apolar phase used. However, analysis of 36 hydrophobicity scales for the 20 common amino acids reveals that the relative hydrophobicity is strongly dependent on the apolar phase used.

A related approach for the characterization of protein surfaces assigns hydrophobicity scores based on the hydrogen-bonding capacity of atoms or functional groups; for example, classing all surface elements formed by carbon as hydrophobic. Intuitively, however, the true hydropathy of a given point close to the surface of a protein depends on the total opportunity for hydrogen bonding at that point. This can only really be assessed by summing over all possible hydrogen-bonding partners for a hypothetical water molecule at that point, taking into consideration the highly directional character of hydrogen bonding. For instance, not all points close to an oxygen atom provide equal hydrogen-bonding opportunity.

Our approach for the assignment of hydrophobicity is based on empirical hydrogen-bonding potential parameterized in the program *GRID* (Goodford, 1985; Wade & Goodford, 1993). *GRID* parameterizes hydrophobicity by evaluating the summed pairwise interactions of a water probe molecule with all surrounding atoms.

For each position close to a protein surface, *GRID* determines the energy ($E_{\rm HB}$) of hydrogen bonds that could be formed by a water molecule at that position. At every point

the Lennard–Jones potential (E_{LJ}) for a water molecule is also calculated and added to the hydrogen-bonding potential. The effect of the Lennard–Jones term is to distinguish between protein interior, protein surface and bulk solvent. If the hydrophobic potential is evaluated for a point within the VDW radius of a protein atom, the VDW term will result in a large positive (repulsive) term. Outside this volume, the VDW potential provides the behaviour of a weakly attractive force that falls off sharply with separation (being proportional to r^{-6}).

The enthalpic component to the solvation is then corrected by an entropy offset calculated under the simplified assumption that water molecules close to a surface can only participate in three hydrogen bonds to bulk solvent instead of 3.5. Thus, displacement of these waters gives rise to a constant entropy offset at each point of the surface. The energy returned by the GRID hydrophobic probe is calculated as

$$\varphi = E_{LJ} + W_{ENT} - E_{HB}. \quad (4)$$

Close to hydrophobic patches, where few hydrogen bonds can be formed, the value of φ is dominated by the entropy offset. Close to charged or polar groups, in contrast, it is dominated by the hydrogen-bonding energy term. The Lennard–Jones potential term contributes a small favourable component close to surfaces of any character, but dominates as a strongly repulsive potential at probe positions that are too close to or within protein atoms.

The *GRID* hydrophobic potential generated in this way can be interpreted as a measure of the energy that would be required to remove a water molecule from a given point on the protein surface into bulk solvent. Owing to the detailed hydrogen-bonding function employed in *GRID*, this hydrophobic scoring function is very sensitive to both the position and orientation of groups at the protein surface. This quality allows the generation of high-resolution maps of surface hydrophobicity.

The *GRID* approach succeeds in predicting some aspects of protein-surface hydrophobicity that do not emerge from a simple categorization of underlying atoms. For example, while tryptophan is considered to be a hydrophobic amino acid, it does have the capacity to form polar interactions, particularly in the indole plane, through its N^{ε} atom (Fig. 2*a*) and thus this part of the residue should properly be characterized as hydrophilic. The *GRID* analysis captures this intuitive behaviour (Fig. 2*b*), clearly demonstrating hydrophilic patches on the surface resulting from the polar N^{ε} atom. Complementary intuitive behaviour is seen for arginine, a predominantly polar amino acid (Fig. 2*c*). In addition to the generally hydrophilic periphery of the guanidino group, *GRID*'s hydrophobic probe identifies both the hydrophobic aliphatic part of the side chain and a partly hydrophobic surface that is parallel to the plane of the guanidino group (Fig. 2*d*). This latter behaviour arises from the directional dependence of hydrogen bonds, which form preferentially in the plane of the guanidino moiety (Singh & Thornton, 1992).

An example of the insight that can be gained from the *GRID*-type hydrophobic analysis is illustrated in Fig. 3. SH3 domains generally bind a proline-rich peptide motif. From an analysis of the fold of an isolated SH3 domain (Fig. 3*a*), relatively few insights into the peptide-binding mechanism could be derived (Musacchio *et al.*, 1992). However, the hydrophobic surface potential reveals a striking correlation between the binding pattern of the naturally occurring ligand, as seen in the crystal structure of an SH3–peptide complex (Musacchio *et al.*, 1994), and the local surface hydrophobicity (Fig. 3*b*).

A further benefit of using this approach to characterizing hydrophobicity is demonstrated by posing the question of whether hydrophobicity can be statistically demonstrated to be preferentially expressed at protein–protein interfaces. Whereas this behaviour has been suggested, the statistical significance, as evaluated using atom- or residue-based methods of evaluating hydrophobicity, has not been strong (Lo Conte *et al.*, 1999). We have addressed this question by assigning to each triangle (T) that is part of a protein surface a hydrophobic potential [$\varphi(T)$] equal to the mean hydrophobic potential of a probe in contact with each of its three vertices. The mean hydrophobicity of a surface [$\Phi(S)$] is therefore the sum over all constituent

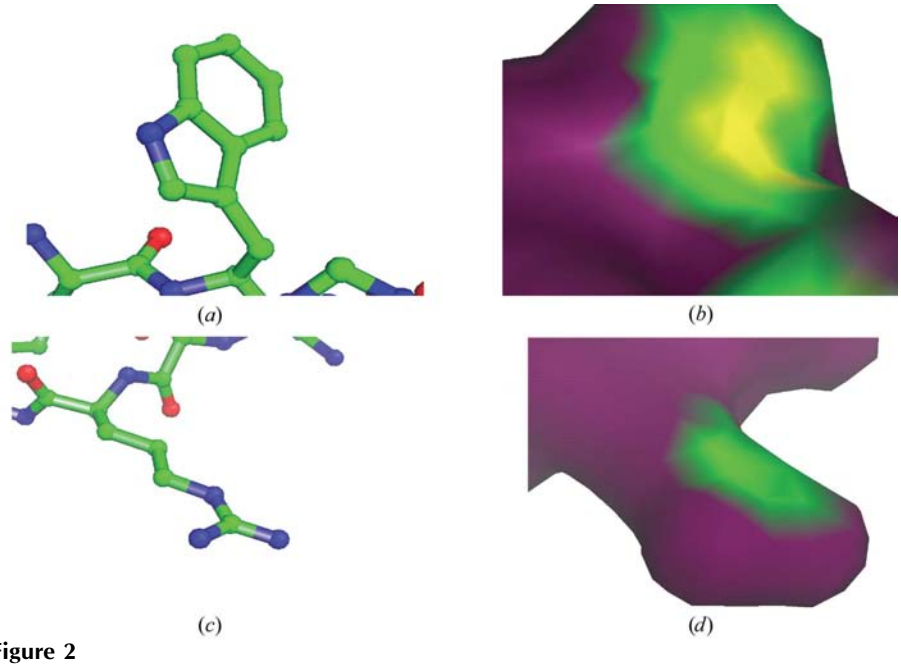


Figure 2
Distribution of hydrophobicity around tryptophan and arginine residues. The side chains of tryptophan (*a*, *b*) and arginine (*c*, *d*) are shown in either ball-and-stick (*a*, *c*) or molecular-surface (*b*, *d*) representation. Ball-and-stick representations are coloured by atom type, whereas the surface representation is coloured by *GRID*-assigned hydrophobic potential. Here, yellow indicates regions with high local hydrophobicity, while purple indicates nonhydrophobic surface patches.

triangles of that surface of the area-weighted hydrophobicity of that triangle, divided by the total area of the surface,

$$\Phi(S) = \frac{\sum_{T \in S} \varphi(T) \cdot A(T)}{\sum_{T \in S} A(T)}, \quad (5)$$

where $A(T)$ is the area of triangle T.

The hydrophobicity of an interface is therefore the mean hydrophobic potential of that part of the surface that becomes buried upon complex formation, while the hydrophobicity of the non-interface surface is the mean hydrophobic potential of the rest of the surface.

Assigning hydrophobicity to contact and noncontact surfaces on this basis, we have evaluated the mean hydrophobicity of buried and exposed surface for 146 surfaces involved in the intermolecular interactions studied by Lo Conte and coworkers.

In testing whether hydrophobicity is significantly higher at protein–protein interfaces, the null hypothesis is that buried and exposed surfaces are subsets of the same population in terms of hydrophobicity. Under the null hypothesis, it is as likely that a buried surface should be more hydrophobic than an exposed surface as that an exposed surface should be more hydrophobic than a buried surface, *i.e.* that the probability of $\Delta\Phi$ [= Φ(buried) − Φ(exposed)] > 0 is identical to the probability of $\Delta\Phi < 0$. Therefore, the probability of finding x out of 146 values of $\Delta\Phi$ to be smaller than zero can be calculated using the binominal distribution on (p = 0.5, n = 144),

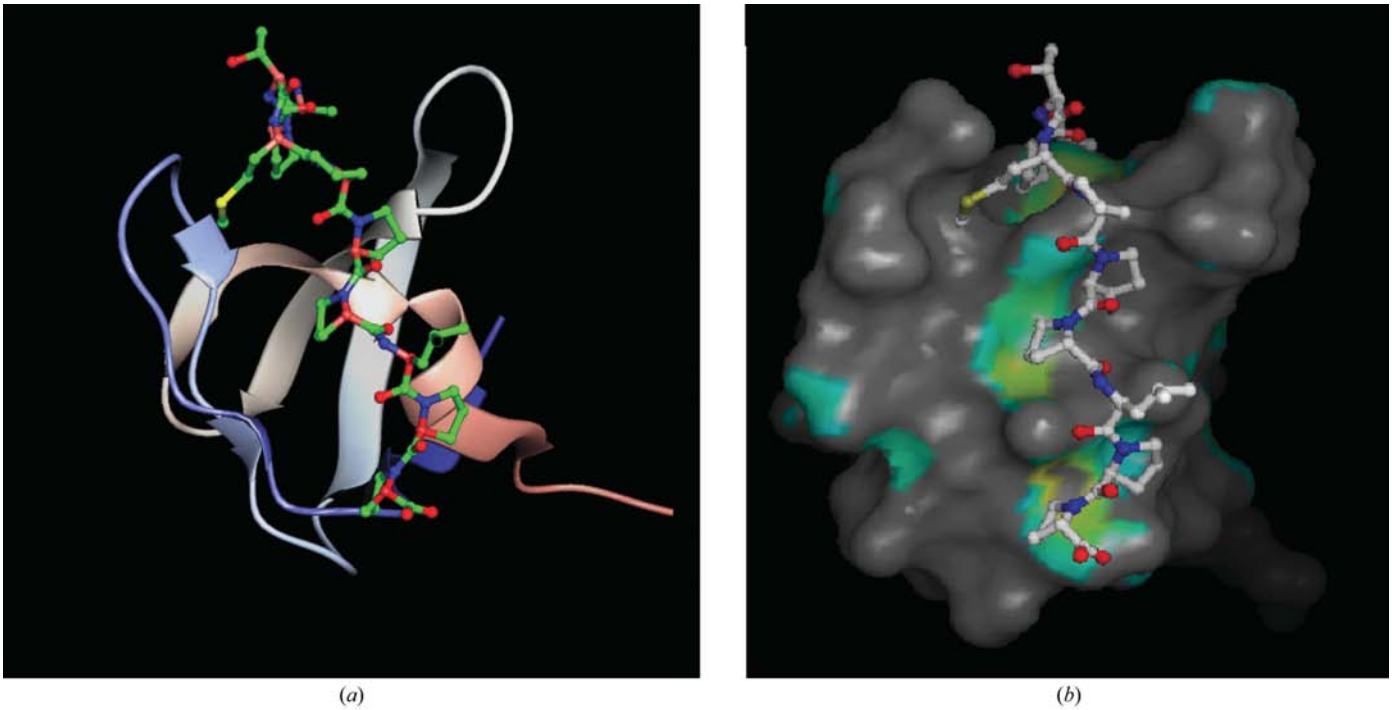
$$\mathrm{Bin}(144, 0.5) = \frac{144!}{x!(144 - x)!} 0.5^{144}. \quad (6)$$

As shown in Fig. 4, values for $\Delta\Phi$ are far from being distributed equally around zero, with only 15 examples of $\Delta\Phi$ being negative. Using the argument outlined above, the null hypothesis can therefore be rejected with a P value of 1.2×10^{-24}, unambiguously demonstrating that interfaces are significantly more hydrophobic than the rest of the protein surface.

3.2. Electrostatics

Whereas many biologically relevant protein–protein interactions derive their affinity from the burial of hydrophobic surface, electrostatics have been shown to play a key role in determining specificity and, in some cases, the thermodynamics and kinetics of macromolecular association (Honig & Nicholls, 1995). Evaluating the potential field around a protein is effectively a question of calculating the field around a group of fixed charges in a low-dielectric environment surrounded by a high-dielectric medium that contains freely diffusing ions.

This is a complicated problem, but can reasonably be achieved by solving some form of the Poisson–Boltzmann equation (PBE; Sharp & Honig, 1990),



(*a*) (*b*)

Figure 3
The SH3 domain of Abl tyrosine kinase (PDB code 1abo) complex surface properties and function. (*a*) The secondary-structure representation of the Abl tyrosine kinase, showing a typical SH3 domain. (*b*) A molecular-surface representation including the proline-rich ligand peptide as a ball-and-stick model. The surface is coloured by local surface hydrophobicity, with strongly hydrophobic surfaces coloured yellow and weakly hydrophobic surfaces elements coloured green.

$$\nabla[\varepsilon(r)\nabla\varphi(r)] + 4\pi\rho(r) - \kappa_0^2\varphi(r) = 0, \quad (7)$$

where $\varepsilon(r)$ is the relative permittivity at r, $\rho(r)$ is the charge density arising from diffusible charges at r, $\varphi(r)$ is the electrostatic potential at r and κ_0 is the Debye–Huckel screening parameter.

We have implemented a finite-difference approach to this problem and made the resulting code available either as a stand-alone executable or as part of *CCP*4*MG* (Potterton *et al.*, 2004). Our implementation exploits a rapid FFT-based algorithm to define the protein interior, 'anti-aliasing' to distribute charges within the initial potential map and optimal over-relaxation, based on the spectral radius of the PBE map, to speed up convergence of the finite difference approach.

Following the approach of Nicholls *et al.* (1991), we have used the electrostatic potential maps that can be generated in this way to assign potentials to the vertices of a triangulated surface representation of a molecule. These representations often offer insight into the character of a molecular interaction, since electrostatic complementarity is a documented phenomenon at protein–protein and protein–ligand interfaces.

We have adapted the approach of McCoy *et al.* (1997) to explore whether the extended set of complex structures

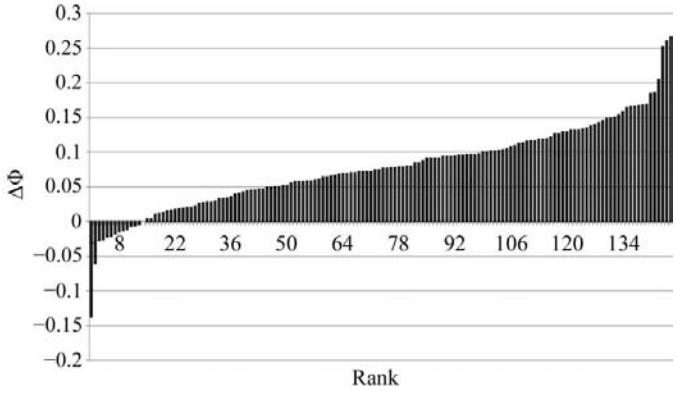


Figure 4
Rank-ordered distribution of $\Delta\Phi$. The calculated differences between the mean hydrophobicity of interface and non-interface surfaces are plotted in ranked order. In all but 15 cases, the interfacial surface is more hydrophobic than the non-interface surface.

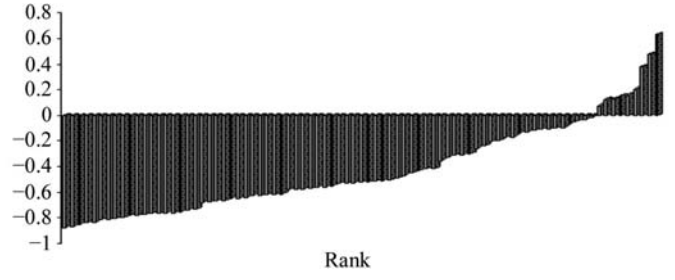


Figure 5
Rank-ordered distribution of electrostatic complementarity. The linear correlation coefficient of electrostatic potentials for different interacting partners is presented for 72 structures. In the vast majority of cases there is an anticorrelation of potential consistent with a marked electrostatic complementarity.

available to date confirms the previous observation that authentic protein–protein interfaces are characterized by experiencing a potential field generated by one binding partner that is complementary in character to the potential field generated by the other.

Briefly, for a complex of known structure containing proteins *A* and *B*, two potential maps are calculated. The first corresponds to the potential that would prevail with a solvent envelope defined by the *A*+*B* complex, but with only atoms of protein *A* charged, while the second corresponds to the potential that would prevail with a solvent envelope defined by the complex but with only atoms of protein *B* charged. The interfacial subset of the surface of protein *A* is isolated as described above for our analyses of conservation and hydrophobicity. Subsequently, two potentials are associated with each vertex of this surface: one derived from interpolating surface vertex positions into the first potential map and the other from interpolating into the second potential map. The electrostatic complementarity of the interface can then be evaluated by calculating the linear correlation coefficient of the two different potentials over the whole set of surface vertices that define the interface.

This calculation was performed for the Lo Conte set of protein–protein interfaces. From Fig. 5, it can be seen that the vast majority of complexes demonstrate a marked electrostatic complementarity.

4. Conclusions

Both the sequence conservation and the composite physical properties of protein–protein interfaces are significantly different from non-interface parts of protein surfaces. Mapping these properties onto surface representations offers a way to both visualize and statistically analyse them so as to produce insights into the collective properties of interfaces in general. This also permits functional hypotheses to be drawn. Further work is required to automate this process, so that any hypothesized interface on a newly determined structure has a confidence level associated with it. The results generated by the *ConSurf* server, which maps conservation scores onto surface residues, show promise for the applicability of such approaches and represent an available option for applying conservation analysis to protein molecules. This work suggests that an analysis of the distribution of conserved residues at the surface of a protein can contribute to both qualitative (Glaser *et al.*, 2003) and quantitative (Nimrod *et al.*, 2005) prediction of the functional sites on proteins.

Scripts and programs for generating and visualizing property-mapped molecular surfaces are being introduced into the *CCP*4 suite for visualization by *CCP*4*MG*.

The authors wish to thank Peter Goodford for many interesting discussions about hydrophobicity. This work was funded by an MRC studentship to JG, a Beit Trust Studentship to AZ and a BBSRC grant to CPB and MEMN.

References

Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). *J. Mol. Biol.* **324**, 105–121.

Brown, N. R., Noble, M. E., Endicott, J. A. & Johnson, L. N. (1999). *Nature Cell Biol.* **1**, 438–443.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. (2004). *Protein Sci.* **13**, 190–202.

Dill, K. A. (1990). *Biochemistry*, **29**, 7133–7155.

Frank, H. & Evans, M. (1945). *J. Chem. Phys.* **13**, 507–532.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E. & Ben-Tal, N. (2003). *Bioinformatics*, **19**, 163–164.

Goodford, P. J. (1985). *J. Med. Chem.* **28**, 849–857.

Grishin, N. V. & Phillips, M. A. (1994). *Protein Sci.* **3**, 2455–2458.

Henikoff, S. & Henikoff, J. G. (1992). *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Honig, B. & Nicholls, A. (1995). *Science*, **268**, 1144–1149.

Jones, S. & Thornton, J. M. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Kauzmann, W. (1959). *Adv. Protein Chem.* **14**, 1–63.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature (London)*, **181**, 662–666.

Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005*a*). *J. Mol. Biol.* **351**, 614–626.

Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005*b*). *Nucleic Acids Res.* **33**, W89–W93.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). *J. Mol. Biol.* **257**, 342–358.

Lo Conte, L., Chothia, C. & Janin, J. (1999). *J. Mol. Biol.* **285**, 2177–2198.

Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.

McCoy, A. J., Chandana Epa, V. & Colman, P. M. (1997). *J. Mol. Biol.* **268**, 570–584.

Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). *Protein Sci.* **7**, 2469–2471.

Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992). *Nature (London)*, **359**, 851–855.

Musacchio, A., Saraste, M. & Wilmanns, M. (1994). *Nature Struct. Biol.* **1**, 546–551.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). *Proteins*, **11**, 281–296.

Nimrod, G., Glaser, F., Steinberg, D., Ben-Tal, N. & Pupko, T. (2005). *Bioinformatics*, **21**, *Suppl.* 1, i328–i337.

Nozaki, Y. & Tanford, C. (1971). *J. Biol. Chem.* **246**, 2211–2217.

Ouzounis, C., Perez-Irratxeta, C., Sander, C. & Valencia, A. (1998). *Pac. Symp. Biocomput.*, pp. 401–412.

Potterton, L., McNicholas, S., Krissinel, E., Gruber, J., Cowtan, K., Emsley, P., Murshudov, G. N., Cohen, S., Perrakis, A. & Noble, M. (2004). *Acta Cryst.* D**60**, 2288–2294.

Radzicka, A., Pedersen, L. & Wolfenden, R. (1988). *Biochemistry*, **27**, 4538–4541.

Shakhnovich, B. E., Dokholyan, N. V., DeLisi, C. & Shakhnovich, E. I. (2003). *J. Mol. Biol.* **326**, 1–9.

Sharp, K. A. & Honig, B. (1990). *J. Phys. Chem.* **94**, 7684–7692.

Singh, J. & Thornton, J. (1992). *Atlas of Protein Side-Chain Interactions.* Oxford: IRL Press.

Southall, N. T., Dill, K. A. & Haymet, A. D. J. (2002). *J. Phys. Chem. B*, **106**, 521–533.

Teichmann, S. A. (2002). *Bioinformatics*, **18**, *Suppl.* 2, S249.

Torrance, J. W., Bartlett, G. J., Porter, C. T. & Thornton, J. M. (2005). *J. Mol. Biol.* **347**, 565–581.

Valdar, W. S. (2002). *Proteins*, **48**, 227–241.

Valdar, W. S. & Thornton, J. M. (2001). *Proteins*, **42**, 108–124.

Valencia, A. (2005). *Curr. Opin. Struct. Biol.* **15**, 267–274.

Wade, R. C. & Goodford, P. J. (1993). *J. Med. Chem.* **36**, 148–156.

Watson, J. D., Laskowski, R. A. & Thornton, J. M. (2005). *Curr. Opin. Struct. Biol.* **15**, 275–284.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M., Kavraki, L. & Lichtarge, O. (2003). *J. Mol. Biol.* **326**, 255–261.