BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 20 2010, pages 2549–2555
doi:10.1093/bioinformatics/btq490

Structural bioinformatics

Advance Access publication September 6, 2010

Identification of protein binding surfaces using surface triplet propensities

Wissam Mehio[1], Graham J.L. Kemp[2], Paul Taylor[1] and Malcolm D. Walkinshaw[1,*]

[1]Institute of Structural and Molecular Biology, School of Biological Sciences, The University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JR, UK and [2]Department of Computer Science and Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The ability to reliably predict protein–protein and protein–ligand interactions is important for identifying druggable binding sites and for understanding how proteins communicate. Most currently available algorithms identify cavities on the protein surface as potential ligand recognition sites. The method described here does not explicitly look for cavities but uses small surface patches consisting of triplets of adjacent surface atomic groups that can be touched simultaneously by a probe sphere representing a solvent molecule. A total of 455 different types of triplets can be identified. A training set of 309 protein–ligand protein X-ray structures has been used to generate interface propensities for the triplets, which can be used to predict their involvement in ligand–binding interactions.

Results: The success rate for locating protein–ligand binding sites on protein surfaces using this new surface triplet propensities (STP) algorithm is 88% which compares well with currently available grid-based and energy-based approaches. Q-SiteFinder's dataset (Laurie and Jackson, 2005. *Bioinformatics*, **21**, 1908–1916) was used to show the favorable performance of STP. An analysis of the different triplet types showed that higher ligand binding propensity is related to more polarizable surfaces. The interaction statistics between triplet atoms on the protein surface and ligand atoms have been used to estimate statistical free energies of interaction. The ΔG_{stat} for halogen atoms interacting with hydrophobic triplets is -0.6 kcal/mol and an estimate of the maximal ΔG_{stat} for a ligand atom interacting with a triplet in a binding pocket is -1.45 kcal/mol.

Availability: Freely available online at http://opus.bch.ed.ac.uk/stp. Website implemented in Php, with all major browsers supported.

Contact: m.walkinshaw@ed.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 27, 2010; revised on July 29, 2010; accepted on August 18, 2010

1 INTRODUCTION

Intermolecular interactions drive and control most biological functions including, for example, signal transduction, transcription and cell cycle events. An understanding of small molecule protein recognition is also key to more effective medicinal chemistry and design of tight binding drug-like ligands (Otvos, 2008; White *et al.*, 2008). The recent growth of information available in structural databases and also of protein-ligand binding data (NIH[1]) provides the required experimental data for the development of a range of *in silico* algorithms; for example predicting ligand binding sites (Capra *et al.*, 2009), predicting absolute ligand binding free energies (Mobley *et al.*, 2007) and predicting protein–DNA binding sites (Tsuchiya *et al.*, 2005).

Proteins frequently play a role in multimer complexes and communicate with a number of different partners, and there is a growing body of structural information on such assemblies as the spliceosome (Guo *et al.*, 2009), the stressosome (Marles-Wright *et al.*, 2008), ribosome (Matsumoto and Ishida, 2009; Taylor *et al.*, 2009), ubiquitination pathways (Falbo *et al.*, 2009), chaperone complexes (Alexander *et al.*, 2009) and transcription machinery (Shechner *et al.*, 2009). A picture is emerging in which even small proteins are likely to have multiple non-overlapping docking or binding sites which may be used in allosteric communication networks. The possibility of being able to identify such allosteric docking sites also opens up new opportunities in the field of drug discovery. Ligand specificity is a major problem in many drug discovery projects where an enzyme active site is highly conserved among a number of isoforms (e.g. protein kinases or phosphatases). By blocking allosteric binding pockets, it may be possible to target specific isoforms (or in the case of infection to target species-specific ortholog proteins). Despite the expanding experimental structural database of protein–ligand and protein–protein interactions, experimental data are still insufficient to identify all potential binding sites that may be of biological relevance to a particular protein target. Reliable computational prediction methods to identify potential protein binding sites are therefore of significant current interest.

Several approaches have been developed to search for protein binding surfaces. Geometry-based algorithms have been quite successful. For example, PocketDepth (Kalidas and Chandra, 2008) utilized the depth of pockets to predict the location of binding sites and report to successfully predict 55% of ligands as first rank predictions. SURFNET (Laskowski, 1995) searches for cavities on the surface of the protein and fills them in a GRID-like method (Goodford, 1985) and then predicts the largest cavity to be the ligand binding site. Searching for pattern similarity (sequence, spatial or orientation patterns) of pocket-forming residues is an alternative approach (Binkowski *et al.*, 2005; Kleywegt, 1999;

*To whom correspondence should be addressed.

[1]National Institutes of Health: http://www.ncbi.nlm.nih.gov

Schmitt *et al.*, 2002) which is generally faster since no energy calculations are required. Molecular mechanics approaches (Huang, *et al.*, 2006) or other energy-based algorithms are also used as exemplified by the program eF-site (Kinoshita and Nakamura, 2003). A third class of algorithm analyses the chemical composition of the protein assigning 'Interface Propensities' to indicate how frequently particular residues appear in binding sites (Jones and Thornton, 1996; Soga *et al.*, 2007a, b).

We present an algorithm, surface triplet propensities (STP), which is based on a score table giving the propensities of atom types of surface atoms that appear in ligand binding sites. We have compared the performance of STP with that of Q-SiteFinder (Laurie and Jackson, 2005) and the method proposed by Morita *et al.* (2008), and have shown that STP is able to predict the known binding sites of 88.2% of the proteins under study (as one of the top three predictions).

2 THE STP ALGORITHM

2.1 Classification and triplet grouping of surface atoms

Protein atoms are classified into 13 atomic group types (Tsai *et al.*, 1999). This classification is based on heavy atom types (carbon, nitrogen, sulfur and oxygen), the number of covalently attached hydrogen atoms and the number of all covalently attached atoms (Table 1). Atomic groups on the surface of the protein are identified by rolling a probe sphere (radius 1.4 Å to simulate a water molecule) over the protein. A triplet (triangle) is defined as a group of three surface atomic groups that can be simultaneously touched by the rolling probe sphere. Neglecting handedness, there are a total of 455 distinct 'triplet-types' that can be generated from combinations of the 13 different atom types.

2.2 The STP dataset

A dataset of 309 protein–ligand complex structures were selected from the PDB, all of which were refined using data to better than 1.7 Å (Supplementary Table 1). For this study, ligands were defined as molecules having more than 10 carbon atoms which made at least 4 van der Waals interactions with the protein. The size of the ligands in the dataset was also restricted such that the maximum separation of any two atoms in the ligand was <23 Å. No two proteins in this test dataset have a sequence identity >50%.

Table 1. The 13 atomic groups according to the classification of Tsai *et al.* (1999) and their occurrence in the dataset

Atom type	Example	Occurrence in the dataset
N3H0	Proline N	1737
N3H1	Amide N	60661
N3H2	Arginine NH1	21820
N4H3	Lysine NZ	7321
O1H0	Carbonyl O	111685
O2H1	Serine OG	15539
C3H0	Carbonyl C	68498
C3H1	Tyrosine CD1	31897
C4H1	Alanine CA	72955
C4H2	Proline CB	124494
C4H3	Alanine CB	66433
S2H0	Methionine SD	1516
S2H1	Cysteine SG	517

For each of these 309 proteins, sets of triplets that are part of the ligand binding surface were identified by calculating the surface triplets of the 'ligand removed' protein structure and comparing them with the protein triplets calculated using the protein–ligand complex. Those triplets that are hidden from the probe sphere in the protein–ligand complex are classified as belonging to the binding site. This resulted in a total dataset of 1 223 008 surface triplets (average of 3958 triplets per structure), of which 34 288 are binding site triplets (average of 111 triplets per structure).

2.3 Interface propensities

Interface propensities are used to quantify the relative abundance of triplet types in binding sites. As a null hypothesis, we expected that any given triplet type will occur in a binding site with the same frequency as it would occur anywhere on protein surfaces. STP uses propensity scores as a score table of occurrence of triplet types in binding sites. Triplet propensities are calculated using Equations 1, 2 and 3 below and the values lie in the interval [−3.54, 5.16] (Supplementary Tables 2 and 3). A logarithmic function is used to calculate propensity values, so for example, a positive value of 3 for a particular triplet means that it is 2^3 times more abundant in binding sites than anywhere on the surface.

$$\mathrm{InterProp}(\alpha) = \frac{\mathrm{InterCount}(\alpha)}{\sum_{i=1}^{455} \mathrm{InterCount}(i)} \quad (1)$$

$$\mathrm{SurfProp}(\alpha) = \frac{\mathrm{SurfCount}(\alpha)}{\sum_{i=1}^{455} \mathrm{SurfCount}(i)} \quad (2)$$

$$\mathrm{Propensity}(\alpha) = \log_2\left(\frac{\mathrm{InterProp}(\alpha)}{\mathrm{SurfProp}(\alpha)}\right) \quad (3)$$

where α designates a certain triplet type; InterProp(α) is the proportion of all ligand binding site triplets that are of type α; SurfProp(α) is the proportion of all surface triplets that are of type α; InterCount(α) is the count of occurrences of triplet type α in ligand binding interfaces in the dataset; SurfCount(α) is the count of occurrences of triplet type α on protein surfaces in the dataset; and i spans the 455 triplet types.

2.4 Coloring the protein surface

Triplet propensity values can be mapped onto the surface of the protein to provide a useful way of visualizing predicted binding regions. Each surface atom is given a *PatchScore* which is defined as the average of all propensity scores of all the surface triplets whose centroids are found within a certain distance from this atom. This PatchScore would indicate the likelihood of this atom to belong to a binding site based on information from its surrounding atomic groups.

To give a visual output, PatchScores attributed to each of the atoms are scaled from 0 to 100 (internal atoms are given a dummy PatchScore) and stored in the B-Factor column of the PDB file. Most molecular viewers are capable of using B-Factors (converted from PatchScores) to color the protein structure from blue to red. An example of this color-coded representation is given in Figure 1.

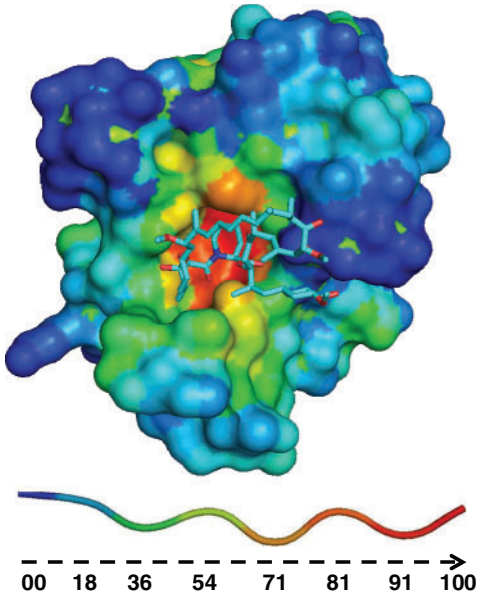

Fig. 1. FKBP12 (PDB id 2DG3) surface colored by STP: in this case, a correct prediction of the binding site as shown by stick representation of known ligand FK506.

The effect of different patch radii on the coloring pattern has been examined (see Supplementary Figs 1 and 2). Increasing the patch radius means including more triplets in the averaging process and might lead to losing the signal of small binding sites since their atoms are given scores incorporating triplets outside the binding site itself. A patch radius of 7.5 Å seems to give optimum results. This value of 7.5 Å is also close to a maximum on the bell-shaped curve showing the frequency of inter-triplet distances in protein ligand binding sites (Supplementary Figs 1 and 2). The scaling 0–100 gives the best visual representation for an individual protein; however, in order to compare groups of proteins, there is an option in the STP program to score multiple structures on the same scale.

3 RESULTS

3.1 Validation of the STP algorithm

A 10-fold cross-validation scheme was used. The training dataset was divided into 10 random and mutually exclusive subsets. The protein structures of each subset were tested against propensity score tables that had been generated using a combination of the other nine subsets. Each protein–ligand complex in the 10th (non-contributing) subset was given two attributes based on this score table; the first attribute being the average propensity of all triplets found in the binding site and the second attribute being the average propensity of all triplets found on the entire protein surface of that protein. This was carried out iteratively over each of the 10 subsets, giving the complexes of each subset two attributes based on a score table calculated from the union of the other nine. The distributions of these two (unbiased) attributes were compared to assess STP's capability of giving high scores to binding site atoms (Fig. 2). The average propensity of all triplets in the ligand-binding sites is higher (mean 0.32, SD 0.41) and can be distinguished from average propensities

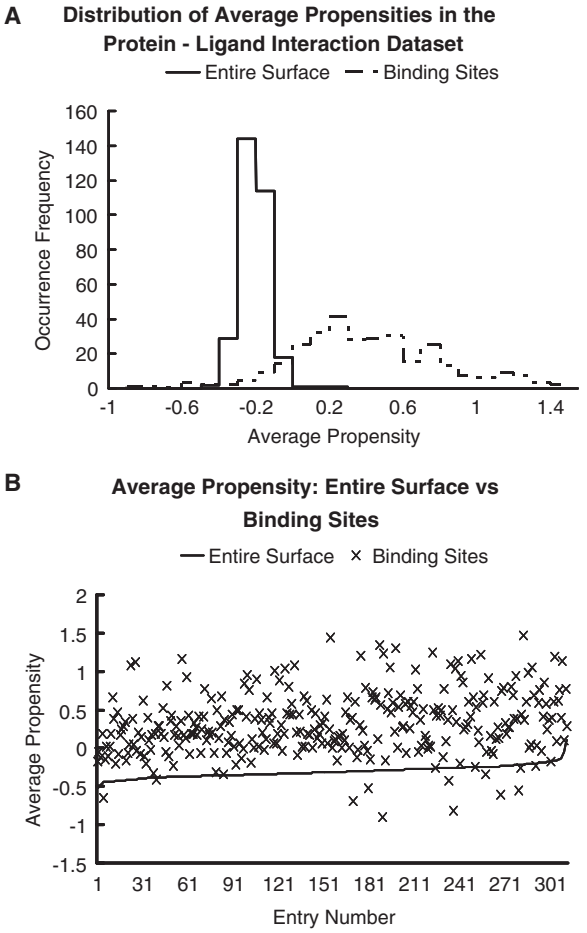

Fig. 2. Comparison of average propensities of triplets belonging to binding sites and triplets belonging to the entire protein surface. (**A**) Shows how the binding sites distribution is shifted to the right compared to the entire surface distribution while (**B**) shows that the average propensity of a binding site is higher than that of the entire surface in ∼96% of the cases tested.

for the entire surface (mean −0.3, SD 0.08). Figure 2B shows that the binding sites have an average propensity greater than the average propensity of all the triplets on the protein surface in ∼96% of the structures. The PDB structures in the dataset, along with their average propensity of interface triplets and average propensity of surface triplets, are given in Supplementary Table 1.

We examine the structures whose binding sites received an average propensity lower than the average propensity of the entire surface. In two cases (1GX5 and 1HYV), the structures were DNA-binding proteins and it is likely that the 'signal' from the DNA binding site was masking the small molecule binding sites used in the test set. In other cases (1IWH, 1MWQ and 1O7J), STP picked-out the main binding site (the active sites of 1MWQ and 1O7J and the Heme binding site of 1IWH) but failed to pinpoint the test binding site.

3.2 The chemical nature of the triplets

The triplets in the dataset were sub classified into four categories (Table 2). There are a total of 35 types of 'hydrophobic triplets'

Table 2. The TCP of the four triplet classes is calculated by dividing the percentage composition of binding site triplets that are of a certain class and dividing it by the percentage composition of all surface triplets that are of that type

Triplet type	Fraction composition of surface triplets	Fraction composition of binding site triplets	TCP
Hydrophobic	0.168	0.259	1.545
Mostly hydrophobic	0.497	0.415	0.834
Mostly polar	0.295	0.271	0.921
Polar	0.040	0.055	1.358

A value greater than one indicates a higher propensity for a certain triplet class to exist in binding sites. Results show an overrepresentation of 'hydrophobic and polar triplets' and an underrepresentation of 'mostly hydrophobic and mostly polar triplets'.

consisting of three hydrophobic (carbon) atomic groups. The 120 types of 'polar triplets' consist of permutations of three polar (N, O and S) atomic groups. The 120 types of 'mostly hydrophobic triplets' contain two hydrophobic atomic groups and the 180 types of 'mostly polar triplets' contain two polar atomic groups. The dataset comprised 1 223 008 surface triplets out of which 34 228 belonged to the binding sites. 'Polar triplets' show the highest average propensity of 1.4 while 'mostly hydrophobic triplets' exhibit the lowest average propensity of 0.4. 'Hydrophobic and mostly polar triplets' each have an average propensity of 0.7. The distribution of individual propensities around the average propensity (signaled by the standard deviation) is similar in all categories (Supplementary Table 4).

We also calculate the '*Triplet Class Propensity*' (TCP) by calculating the percentage composition of binding site triplets that are of a certain class and dividing it by the percentage composition of all surface triplets that are of that type (Table 2). 'Hydrophobic and polar triplets' show an overrepresentation with TCPs of 1.55 and 1.36, respectively, while 'mostly hydrophobic and mostly polar triplets' show an underrepresentation with TCPs of 0.83 and 0.92, respectively. The overrepresentation of 'hydrophobic and polar triplets' indicates the key role of these triplets in protein ligand binding, as they provide strong hydrophobic and polar interactions with the ligand atoms to ensure a strong binding.

3.3 Recognition of ligand atoms by protein surface triplets

We now study the interaction between the surface triplets and ligand atoms, and search for preferences that some triplets might have for different ligand atom types. Ligand atoms in the dataset are classified according to the Tripos forcefield definitions (Clark *et al.*, 1989). The protein–ligand dataset contains 20 ligand atom types (Supplementary Table 5) grouped into the four classes: halogens, hydrophobic, polar and water.

The interaction between surface triplets and ligand atoms is quantified by measuring the distance between atoms and triplet centroids. For each surface triplet, the closest ligand atom or water molecule is recorded. If there are no atoms within a distance of 4 Å, the closest atom type is recorded as 'Empty'. The interaction between each triplet class (hydrophobic, mostly hydrophobic, mostly polar and polar) and each atom class (halogen, hydrophobic, polar and water, empty) is studied. The observed

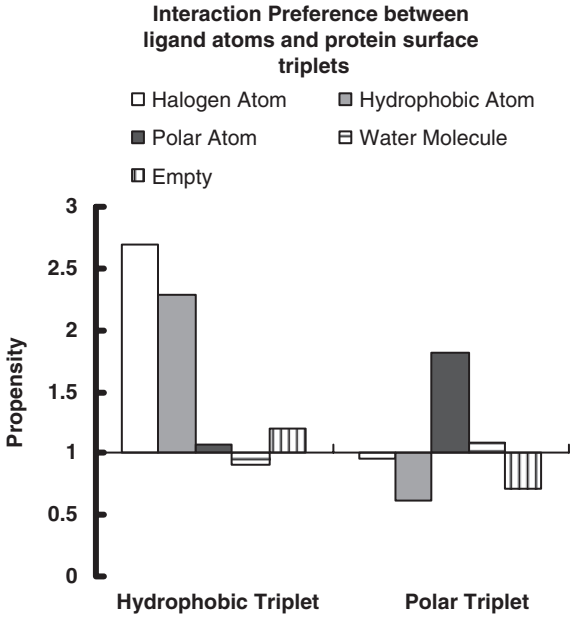

Fig. 3. The interaction preference of four ligand atom classes by hydrophobic and polar triplet classes as calculated in Supplementary Table 6. An interaction preference value greater than 1 indicates a favored interaction while a value less 1 indicates a disfavored interaction. Empty class designates a triplet that has no ligand atoms within 4 Å from its centroid. Results show a strong affinity for hydrophobic atoms and halogens to interact with 'hydrophobic triplets', as well as a high interaction preference for polar atoms to interact with 'polar triplets'.

frequencies of each interaction are recorded and compared with the expected frequency (EF). The EF of a certain interaction depends on the availability of a certain triplet class and a certain atom class in the dataset. For example using data from Supplementary Table 6, for the 206 halogen atoms and the 204 987 'hydrophobic triplets' in the database, the EF of 'hydrophobic triplet': halogen interactions = $(206 \times 204987)/(1223008)$ (where 1 223 008 is the total number of interactions). A final attribute (interaction preference) is calculated by dividing the observed frequency (OF) by the EF. An (OF/EF) value greater than 1 indicates a favored interaction while a result less than 1 indicates a disfavored interaction (Fig. 3 and Supplementary Table 6).

The triplet:ligand atom interaction data (Fig. 3 and Supplementary Table 6) indicates that water molecules and polar ligand atoms have a high propensity for 'mostly polar and polar triplets'. Hydrophobic ligand atoms and halogens have a high propensity for 'hydrophobic triplets', but not for 'mostly hydrophobic triplets'.

3.4 Comparison of STP and other ligand-binding site prediction methods

For this set of validation studies, STP was used to score and rank ligand binding sites of the 309 test structures. The binding site is defined as the set of atoms shielded from the water probe upon binding to the ligand. PatchScores (Section 2.4) were calculated for each atom associated with these binding sites. To make sure these PatchScores avoid any bias, the 10-fold cross-validation scheme (Section 3.1) was used again. In both experiments described below,

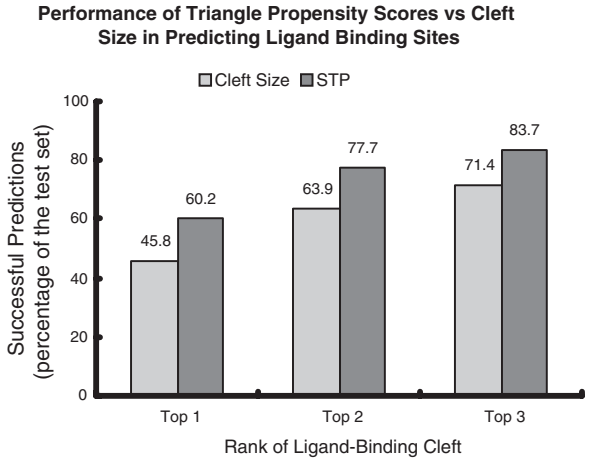

Fig. 4. Whether considering the top 1, top 2 or top 3 ranks (ranked by size of cavity or by the STP score), STP is more successful in locating ligand binding sites.

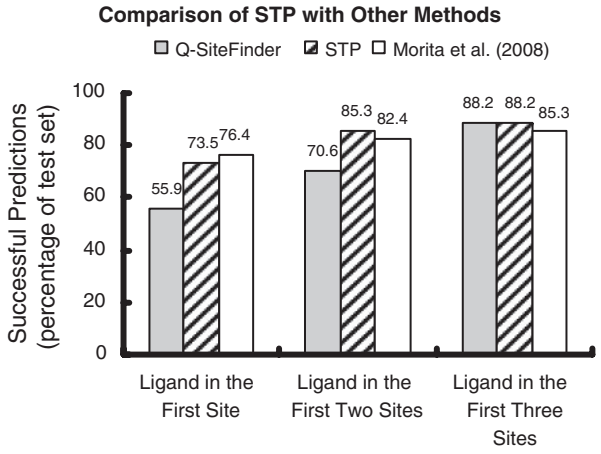

Fig. 5. Comparison of the performance of STP with Q-SiteFinder and the method by Morita *et al.* (2008) shows STP to be a competitive and successful binding site prediction program.

the PatchScores of surface atoms are scaled from 0 to 100 (zero being the minimum and 100 being the maximum). A binding site is then given an STP score equal to the number of atoms with a scaled PatchScores over 70. Such atoms would be colored orange to red and would be picked by eye as probable locations for a binding site (Fig. 1).

3.4.1 Comparison of STP with cavity size Cavities in protein surfaces are often associated with ligand recognition or enzymatic activity (Laskowski *et al.*, 1996; Liang *et al.*, 1998; Weskamp *et al.*, 2009) and several programs [SURFNET (Laskowski, 1995), Ligsite (Huang and Schroeder, 2006) and PocketFinder (An *et al.*, 2005; Hendlich *et al.*, 1997)] are available to identify such pockets. We used SURFNET to calculate cavities for the 309 structures in our dataset.

The atoms forming the cavities identified by SURFNET were used as input to STP and the cavities were ranked according to the number of high scoring atoms (PatchScore above 70 on a scale of 0–100) included. The performance of STP was then assessed by the percentage of cases where the ligand binding site was STP ranked in the top 1, 2 or 3 cavities. Figure 4 and Supplementary Table 7 show the results in comparison to ranking those cavities by size.

As shown in Figure 4, STP performs much better than just picking the largest cavity on the surface. On average, SURFNET found 29 cavities on the surface of each protein. Of total, 83.7% of the ligand binding sites were discovered in the top three cavities ranked by STP. In conclusion, ranking clefts with STP is a better indicator of the location of the binding sites than cavity size and the incorporation of STP with SURFNET as depicted in this experiment gives a better prediction of the binding site than using SURFNET on its own. In cases where the clefts identified by SURFNET are much larger than the ligand, information from STP can be used to help further pinpoint ligands (Supplementary Fig. 3).

3.4.2 Comparison with Q-SiteFinder and Morita et al. (2008) Several binding site prediction methods use GRID-like searches (Goodford, 1985) in which interaction energies are calculated between a probe atom and the surface of the protein. We compared the performance of STP against two such approaches; one implemented in the program Q-site finder and the other described in Morita *et al.* (2008). Both use high scoring probes as seeds for a clustering process that attempts to locate the most energetically favorable locus for a ligand.

A dataset of 34 structurally distinct proteins in the unbound state which share structural similarity with 34 proteins in the ligand-bound form was created by Laurie and Jackson (2005). This dataset was used to check the performance of STP on proteins in the unbound state. The unbound proteins were superimposed onto their bound homologues. Ligands were then extracted to mark the binding sites in the unbound proteins. The cavities on these proteins were extracted with SURFNET (Laskowski, 1995) and then ranked by STP. The binding site of one of these structures (1PHD) was an internal binding site and was omitted from the analysis.

Figure 5 and Supplementary Table 8 summarize the performance of STP in comparison with Q-SiteFinder, and the method created by Morita *et al.* (2008). On average, 30% of the atoms within 5 Å of an experimentally verified ligand have a PatchScore of 70–100. The binding site is located in the top predicted location by STP in 74% of the cases. This compares to 56 and 76% of the cases of the other two methods (Fig. 5). The binding site is located in the top two predictions in 85% of the cases (the comparable hit rates for the other two approaches are 70 and 82%, respectively (Fig. 5).

STP succeeds at identifying ligand binding sites in the structures 3APP and 1BYA where both other methods failed (Morita *et al.*, 2008). In both these cases, the ligands are large and long and this has hindered their prediction: the STP method is independent of the ligand size. We can define a false positive as an experimentally determined binding site which is not among the top three predicted STP sites for that protein. There are four such structures out of the total 34 in this test dataset (1NNA, 1PDY, 1HSI and 6INS). For 1NNA and 1PDY, their heteromerization sites dominated the signal and were predicted over their small molecule ligand binding sites. For 1HSI and 6INS, we can find no documented function for the predicted patches. Thus, the false positive rate from this dataset

is 2/34 = 5.9% (though we cannot exclude the possibility that these predicted patches play an as yet undiscovered role in ligand binding).

4 DISCUSSION AND CONCLUSIONS

The propensity of interaction of the different triplets with particular atom types can be used to give apparent binding energies named by Fersht and others (Fersht *et al.*, 1993) as statistical free energy (ΔG_{stat}) values. For example, a non-bonded interaction between a halogen-class atom and a 'hydrophobic triplet' has a propensity of 2.7, where the interaction preference is calculated as the OF/EF (Section 3.3 and Supplementary Table 6). This is equivalent to saying that the frequency of interaction of a halogen atom with a 'hydrophobic triplet' is about three times the expected value. The empirical free energy difference that accounts for this distribution can be calculated from: $\Delta G_{stat} = -RT \ln (OF/EF)$, where R is the gas constant = 1.9872 cal/deg/mol. This gives a $\Delta G_{stat} = -0.59$ kcal/mol at 298 K for the interaction of halogen atoms with 'hydrophobic triplets'. The other clear preference for atom environment (Supplementary Table 6) is the interaction between hydrophobic atoms and 'hydrophobic triplets' (interaction preference 2.28) which gives a ΔG_{stat} of -0.49 kcal/mol. The preference of polar atoms interacting with 'polar triplets' is less marked with a ΔG_{stat} of -0.35 kcal/mol.

The frequency of occurrence of the 455 different triplet classes occurring in the test dataset of 1.22 million triplets from the surfaces of the 309 proteins has been tabulated (Supplementary Tables 9 and 10). The most commonly occurring triplet types are the 'hydrophobic triplet' (C3H1, C3H1, C3H1) with 951 occurrences and the 'mostly polar triplet' (C4H2, N3H1, O1H0) with 927 occurrences. This analysis of protein surfaces provides a quantitative measure of the role played by different triplet subclasses in protein-ligand binding (Table 2). The highest propensity scores for frequently occurring triplets are those of N3H1, N3H1, N4H3 (4.1), C3H1, C3H1, S2H1 (3.5), C3H1, C4H2, S2H1 (3.2) and C3H0, C3H0, C3H1 (2.7). It is notable that in each of these high propensity triplets, the C and N atoms are mainly in the sp2 state, suggesting that higher ligand binding propensity is related to more polarizable surfaces.

The propensities defined in Equation (3) for the 455 triplet types (for being in a binding site) range from -3.54 to 5.16. These propensity values can also be converted to statistical free energy values: $\Delta G_{stat} = -RT \times (\text{propensity}) \times \ln(2)$, providing an empirical measure of energy of interaction of a particular triplet type with a ligand. These ΔG_{stat} values range between -1.45 kcal/mol and 2.12 kcal/mol (Supplementary Table 3). These statistical free energy values provide a measure of the average empirical interaction energy of a ligand (averaged over all ligand atoms and all atom types) with a particular class of atom triplet. Interestingly, the strongest ΔG_{stat} value of -1.45 kcal/mol is very close to the maximum affinity value of -1.5 kcal/mol per ligand atom which was estimated from an analysis of experimental binding data (Kuntz *et al.*, 1999).

The incorporation of the propensity scores into the program STP provides a very fast method of analyzing and visualizing protein surface properties as a pattern of triplets with run-times for most PDB protein structures under one second on a standard home computer. The STP algorithm does not specifically search for protein cavities on the surface but rather uses the distribution of triplets of adjacent atomic groups. Such an approach gives STP the advantage of being able to identifying shallow binding sites (Supplementary Fig. 4). Most small molecule binding sites currently in the literature are located in deep pockets. The combination of the STP algorithm with more conventional cavity-finding programs may provide a new approach for finding as yet uncharacterized druggable pockets.

ACKNOWLEDGEMENTS

We are grateful to Yi-Gong Chen for creating the protein : ligand database. We also thank the Centre for Translational and Chemical Biology at the University of Edinburgh for use of facilities.

Funding: We acknowledge the Wellcome Trust and the Darwin Trust of Edinburgh.

Conflict of Interest: none declared.

REFERENCES

Alexander,L.D. *et al.* (2009) Evaluation of di-sansalvamide a derivatives: synthesis, structure-activity relationship, and mechanism of action. *J. Med. Chem.*, **52**, 7927–7930.

An,J. *et al.* (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics*, **4**, 752–761.

Binkowski,T.A. *et al.* (2005) Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.*, **332**, 505–526.

Capra,J.A. *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.

Clark,M. *et al.* (1989) Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.*, **10**, 982–1012.

Falbo,K.B. *et al.* (2009) Involvement of a chromatin remodeling complex in damage tolerance during DNA replication. *Nat. Struct. Mol. Biol.*, **16**, 1167–1172.

Fersht,A.R. *et al.* (1993) Protein stability: experimental data from protein engineering. *Philos. Trans. R Soc. Lond A*, **345**, 141–151.

Goodford,P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–857.

Guo,Z. *et al.* (2009) Single-molecule analysis of protein-free U2–U6 snRNAs. *Nat. Struct. Mol. Biol.* **16**, 1154–1159.

Hendlich,M. *et al.* (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.*, **15**, 359–363.

Huang,B. and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **24**, 6–19.

Huang,N. *et al.* (2006) Molecular mechanics methods for predicting protein–ligand binding. *Phys. Chem. Chem. Phys.*, **8**, 5166–5177.

Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Kalidas,Y. and Chandra,N. (2008) PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.*, **161**, 31–42.

Kinoshita,K. and Nakamura,H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.*, **12**, 1589–1595.

Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1897.

Kuntz,I. *et al.* (1999) The maximal affinity of ligands. *Proc. Natl Acad. Sci. USA*, **96**, 9997–10002.

Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph*, **13**, 323–330.

Laskowski,R.A. *et al.* (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438–2452.

Laurie,A.T. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.

Liang,J. *et al.* (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for drug design. *Prot. Sci.*, **7**, 1884–1897.

Marles-Wright,J. *et al.* (2008) Molecular architecture of the "stressosome," a signal integration and transduction hub. *Science*, **322**, 92–96.

Matsumoto,A. and Ishida,H. (2009) Global conformational changes of ribosome observed by normal mode fitting for 3D Cryo-EM structures. *Structure*, **17**, 1605–1613.

Mobley,D.L. *et al.* (2007) Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, **371**, 1118–1134.

Morita,M. *et al.* (2008) Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins*, **73**, 468–479.

Otvos,L. (2008) *Peptide-Based Drug Design*. Humana Press, Totowa, New Jersey.

Schmitt,S. *et al.* (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.

Shechner,D.M. *et al.* (2009) Crystal structure of the catalytic core of an RNA-polymerase ribozyme. *Science*, **326**, 1271–1275.

Soga,S. *et al.* (2007a) Identification of the druggable concavity in homology models using the PLB index. *J. Chem. Inf. Model*, **47**, 2287–2292.

Soga,S. *et al.* (2007b) Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model*, **47**, 400–406.

Taylor,D.J. *et al.* (2009) Comprehensive molecular structure of the eukaryotic ribosome. *Structure*, **17**, 1591–1604.

Tsai,J. *et al.* (1999) The packing density in proteins: standard radii and volumes. *J. Mol. Biol.*, **290**, 253–256.

Tsuchiya,Y. *et al.* (2005) PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. *Bioinformatics*, **21**, 1721–1723.

Weskamp,N. *et al.* (2009) Merging chemical and biological space: structural mapping of enzyme binding pocket space. *Proteins*, **76**, 317–330.

White,A.W. *et al.* (2008) Protein–protein interactions as targets for small-molecule therapeutics in cancer. *Exp. Rev. Mol. Med.*, **10**, e8.