BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 12 2012, pages 1579–1585
doi:10.1093/bioinformatics/bts182

Structural bioinformatics

Advance Access publication April 11, 2012

Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles

Zhong-Ru Xie[1,2] and Ming-Jing Hwang[1,2,]*

[1]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 112 and [2]Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Knowledge about the site at which a ligand binds provides an important clue for predicting the function of a protein and is also often a prerequisite for performing docking computations in virtual drug design and screening. We have previously shown that certain ligand-interacting triangles of protein atoms, called protein triangles, tend to occur more frequently at ligand-binding sites than at other parts of the protein.

Results: In this work, we describe a new ligand-binding site prediction method that was developed based on binding site-enriched protein triangles. The new method was tested on 2 benchmark datasets and on 19 targets from two recent community-based studies of such predictions, and excellent results were obtained. Where comparisons were made, the success rates for the new method for the first predicted site were significantly better than methods that are not a meta-predictor. Further examination showed that, for most of the unsuccessful predictions, the pocket of the ligand-binding site was identified, but not the site itself, whereas for some others, the failure was not due to the method itself but due to the use of an incorrect biological unit in the structure examined, although using correct biological units would not necessarily improve the prediction success rates. These results suggest that the new method is a valuable new addition to a suite of existing structure-based bioinformatics tools for studies of molecular recognition and related functions of proteins in post-genomics research.

Availability: The executable binaries and a web server for our method are available from http://sourceforge.net/projects/msdock/ and http://lise.ibms.sinica.edu.tw, respectively, free for academic users.

Contact: mjhwang@ibms.sinica.edu.tw

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

Received on December 16, 2011; revised on March 13, 2012; accepted on March 31, 2012

1 INTRODUCTION

The number of sequences and structures of protein of unknown function is increasing rapidly as a result of genome research. These proteins can usually be assigned a putative function if their ligand-binding sites can be predicted. As such, ligand-binding site prediction (LBSP) is now a competition category in critical assessment of protein structure prediction (CASP) experiments (López *et al*., 2009). LBSP is also often the starting point for protein–ligand docking, the core methodology for virtual screening in drug design and development (Laurie and Jackson, 2006; Leis *et al*., 2010).

Ligands usually bind at specific sites of target proteins, and it is, therefore, generally believed that ligand-binding sites can be distinguished from other parts of the protein surface by certain specific features, for example, the electrostatic potential and size of a cavity formed on the protein surface. Many LBSP algorithms based on this line of thinking have been proposed in the past decade. Generally, the main difference between these algorithms is whether they perform geometric computations to identify cavities, i.e. binding site pockets, on the protein surface.

Many methods that compute binding site pockets (Brady and Stouten, 2000; Hendlich *et al*., 1997; Huang and Schroeder, 2006; Laskowski, 1995; Levitt and Banaszak, 1992; Liang *et al*., 1998; Weisel *et al*., 2007; Yu *et al*., 2010; Zhu and Pisabarro, 2010), also known as geometry-based methods (Laurie and Jackson, 2006; Leis *et al*., 2010), have been developed based on the assumption that a ligand always binds in the largest cavity in its target protein. This is, of course, not 100% correct but is a reasonable starting point, and the assumption has been shown to be correct ∼70% of the time in an analysis of a set of 210 known protein–ligand complex structures (Huang and Schroeder, 2006). However, how to precisely identify empty spaces (i.e. cavities) on the protein surface and determine their size is not trivial, as different cavity-computing algorithms do not always give the same result for the same protein structure (Huang and Schroeder, 2006; Laskowski, 1995; Yu *et al*., 2010; Zhu and Pisabarro, 2010).

To avoid relying on the inherently incorrect assumption that the largest cavity is the binding site, a different type of method searches for energetically favored regions for ligand binding using energy calculations (An *et al*., 2005; Goodford, 1985; Kinoshita and Nakamura, 2003; Laurie and Jackson, 2005; Morita *et al*., 2008). To do so, these energy-based methods distribute probes (usually a sphere representing a small molecule or a chemical group with a particular function, e.g. hydrophobicity) around the target protein and calculate the interaction energy between the probes and the protein atoms with which they are in contact. Probes with favorable interaction energies are then clustered, and the most energetically favored region is identified and predicted to be the ligand-binding site. As these methods do not aim at identifying the largest cavity, they are more likely to succeed where geometry-based methods often fail, e.g. in cases in which the ligand binds in a shallow cavity on

*To whom correspondence should be addressed.

the protein surface; on the other hand, energy calculations may miss the largest cavity, and they are more expensive to compute and may require a sophisticated clustering scheme. Furthermore, in LBSP applications, the ligand molecule is usually not known, and it is difficult to devise a probe system that can simulate a variety of chemical properties to cover the large possible number of unknown ligands.

As more and more protein structures are deposited in the PDB (Berman *et al.*, 2000), homology-based methods (Brylinski and Skolnick, 2008; Laskowski *et al.*, 2005; Lopez *et al.*, 2011; Roche *et al.*, 2011; Wass and Sternberg, 2009; Wass *et al.*, 2010; Zhang, 2008), in which one finds structures resembling that of the target protein and predicts the ligand-binding site by inferring homology through structural alignments, have become viable; these can be more reliable than other methods but require the availability of homologous structures with known ligand-binding sites (Wass *et al.*, 2011).

Another strategy that can be applied, thanks to the increase in structure data, is to statistically identify features that may have been evolutionarily imprinted in ligand-binding sites. Such methods are propensity-based methods. For example, Soga *et al.* (2007) found that the amino acid composition at ligand-binding sites is significantly different from that at non-binding site protein surfaces and that the difference between the two could be used to rescore the binding site pockets reported by others to improve the success rates of LBSP.

As these different strategies have different strengths and weaknesses, methods, such as Fpocket (Le Guilloux *et al.*, 2009; Schmidtke *et al.*, 2010), that combine different strategies show potential to perform better than single-strategy methods. Indeed, using consensus of results of other methods, i.e. performing meta-predictions, MetaPocket 2.0 (MPK2) reports the best LBSP success rates to date (Zhang *et al.*, 2011). Finally, residue conservation is a critical property for increasing the success rate of LBSP and is commonly incorporated in LBSP methods (Capra *et al.*, 2009; Gutteridge *et al.*, 2003; Huang and Schroeder, 2006; Wass *et al.*, 2011).

In this study, we developed a novel algorithm for predicting protein's ligand-binding sites. This is an extension of our previous work (Xie and Hwang, 2010), in which we developed a novel docking score function, MotifScore, based on network motifs of 3D protein–ligand contacts (i.e. interactions). During the derivation of MotifScore, we observed that some of the 3D protein–ligand interaction motifs contributed much more significantly to MotifScore than others and that their three constituent protein atoms, which form a triangle on the protein surface, were often seen together, i.e. were concomitantly enriched, at the ligand-binding sites. This new LBSP method was named LISE on the basis that these protein triangles are *L*igand *I*nteracting and binding *S*ite *E*nriched. Benchmark predictions using the same set of structural data and the same evaluation criteria showed that, in cases where comparisons were made, LISE outperformed other methods on the whole, often with a significantly better accuracy for the first predicted site (i.e. top1 prediction). Furthermore, in a prediction for 19 recent CASP targets with a 'full' ligand molecule, as opposed to an ion or small solvent molecule, LISE correctly identified 17 (89%) as the top1 site, a level of accuracy similar to that achieved for the benchmark predictions, suggesting that LISE can be an accurate and reliable LBSP tool in helping to annotate proteins of unknown function.

2 METHODS

2.1 Protein triangles, binding-site enrichment factors and triangle scores

As described previously (Xie and Hwang, 2010), MotifScore was constructed by counting the occurrences of motifs of 3D protein–ligand interaction networks extracted from 6276 protein–ligand complex structures. These motifs capture interactions between three protein atoms and two ligand atoms, those contributing significantly to MotifScore being found to occur preferentially at ligand-binding sites. The three protein atoms of a motif thus derived form a triangle bounded by the constraints that all three sides of the triangle must be longer than 2 Å and shorter than 13 Å and at least two sides must be shorter than 10 Å (Xie and Hwang, 2010). The triangle, referred to as a protein triangle, is actually an atom-type triangle, because different atoms of similar chemical nature are assigned to the same atom type (all the protein atoms are represented by 1 of 13 atom types, see Table 1 in Xie and Hwang (2010); throughout this article, hydrogen atoms are ignored when protein or ligand atoms are mentioned). Each protein triangle is associated with a binding site enrichment factor, F_b, which is derived from a statistical analysis of protein triangles found in both ligand binding and non-ligand binding parts of the protein and reflects the propensity of a protein triangle to occur at the ligand-binding site (Xie and Hwang, 2010).

To cut down computational cost, in this work, we only considered those protein triangles with a non-zero solvent accessible surface area for all their three atoms. In addition, those with an $F_\mathrm{b} < 1$ were ignored; these omissions constituting a large percentage (>70%) of the total number of protein triangles found at the protein surface but only making a marginal contribution to the triangle score (S_t) defined below (see Supplementary Figs S1 and S2).

$$S_\mathrm{t} = F_\mathrm{b} + w \sum_{i=1}^{3} C_i \tag{1}$$

where C_i is the conservation score of the residue to which the *i*th ($i = 1,2,3$) atom of the triangle belongs, and w is a weighting factor, which, in this work, was determined in preliminary tests to be 1.7 for optimal performance (data not shown). The conservation score was taken directly from PSI-BLAST's PSSM (position-specific scoring matrix; Altschul *et al.*, 1997). The conservation score made a minor contribution to LISE, increasing its success rates by several percentages in some of the benchmark predictions (see Section 3 and Supplementary Table S1).

2.2 Grids, sphere of grids and sphere (binding site) scores

To identify hot spots for ligand binding using the triangle scores, we created a regular 3D grid with a step size of 1 Å surrounding the target protein and labeled each grid point as either 'protein-occupied' or 'empty', depending on whether a protein atom could be found within 2.7 Å of the grid point. For each empty grid point, a grid point score, S_g, was calculated as the sum of the triangle scores, S_t [Equation (1)], of the surrounding protein triangles:

$$S_\mathrm{g} = \sum^{n} S_\mathrm{t} \tag{2}$$

where n is the number of protein triangles in which the geometric center falls within 4 Å of the empty grid point under consideration. A sphere 11 Å in diameter and centered on each empty grid point was then created, and a sphere score, S_s [Equation (3)], was calculated as the sum of the grid point scores, S_g [Equation (2)], for all the empty grid points found within the sphere, and the number of empty grid points found was m:

$$S_\mathrm{s} = \sum^{m} S_\mathrm{g} \tag{3}$$

Finally, spheres were ranked according to their sphere score, S_s, and the first (top ranked) predicted site was assigned to the sphere with the highest S_s, the

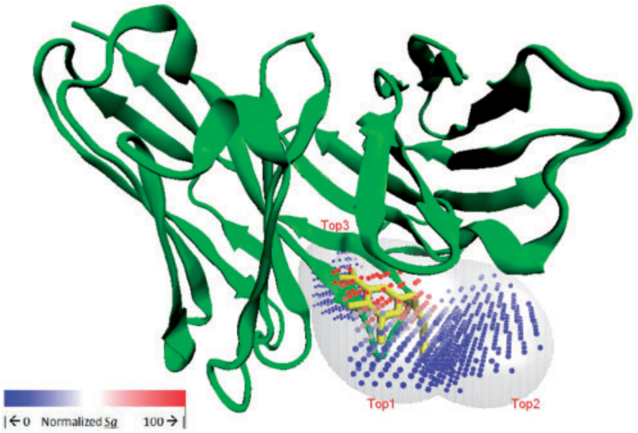

Fig. 1. The top three ligand-binding sites of the immunoglobulin B1–8 fv fragment (1a6w) predicted by LISE. The ligand is depicted by yellow sticks, and LISEs top three predicted ligand-binding sites are labeled and shown as a transparent sphere containing grid points, represented by dots colored from blue to red to indicate the normalized value (see the color spectrum at bottom left) of their grid score. This figure was created using the VMD program (Humphrey *et al.*, 1996).

second was assigned to the sphere with the highest S_s for all spheres centered at least 8.25 Å (1.5 times the sphere's radius of 5.5 Å) from the center of the first sphere, the third was assigned following the same procedure and had to be located at least 8.25 Å from both the first and second site and so on for subsequent ranked spheres. Note that, although all the above distance parameters may seem arbitrary, they are in line with considerations of protein and atom structures and are either identical to, or within the range of, those used in related studies (Capra *et al.*, 2009; Gutteridge *et al.*, 2003; Huang and Schroeder, 2006; Laskowski, 1995; Muegge and Martin, 1999; Ruvinsky and Kozintsev, 2005; Xie and Hwang, 2010). Figure 1 shows an example in which the first three top-ranked ligand-binding sites predicted by LISE for the immunoglobulin B1–8 fv fragment (PDB 1a6w) are labeled, with each site (sphere) filled with empty grid points colored according to the value of their score, S_g, as indicated by the accompanying color spectrum.

2.3 Criteria for assessing binding site predictions

Different criteria have been proposed to evaluate whether an LBSP is a success or a failure. Of these, the most widely used are a distance criterion and a precision criterion. The distance criterion checks the distances between a single point, usually the geometric center of the predicted ligand-binding site, and all the atoms of the ligand and dictates that at least one of those distances must be <4 Å for the prediction to be considered a success (Huang and Schroeder, 2006). The precision criterion checks what percentage of the grid points within a predicted binding site are within 1.6 Å of any of the ligand atoms and dictates that the percentage should be greater than a specified percentage (e.g. 25% or 0%) for the prediction to be considered a success (Laurie and Jackson, 2005). In this work, we used both criteria to compare our results with the results generated using other methods reported in the literature.

3 RESULTS

3.1 Comparison with other methods

Different datasets and different criteria have been used to evaluate the performance of LBSP methods. For comparison with other methods, LISE was evaluated on two benchmark datasets, using the same evaluation criteria used by the other methods.

Table 1. Comparison of the top1 and top3 success rates for various LBSP methods using 210 ligand-bound structures[a]

Methods	Top1 (%)	Top3 (%)
LISE (this work)[b]	83	94
MPK2 (Zhang *et al.*, 2011)	81	95
MPK1 (Huang, 2009)	75	93
Q-SiteFinder (Laurie and Jackson, 2005)[c]	70	90
LIGSITE^{csc} (Huang and Schroeder, 2006)	75	–
LIGSITE^{cs} (Huang and Schroeder, 2006)	70	86
PASS (Brady and Stouten, 2000)[c]	51	80
SURFNET (Laskowski, 1995)[c]	42	57

[a] The success rates were computed as the percentage of the 210 structures for which the best (top1) or any one of the best three (top3) predicted binding sites satisfied the 4 Å distance criterion (see Section 2).

[b] Without considering residue conservation [Equation (1)], the success rates were 76% (top1) and 90% (top3).

[c] The success rates of these LBSP methods were taken from (Huang, 2009).

Table 1 presents a comparison of the top1 and top3 success rates for LISE for a set of 210 ligand-bound protein structures with those reported in the literature on four geometry-based methods [LIGSITE^{cs} and its improved version LIGSITE^{csc}, which takes residue conservation into consideration (Huang and Schroeder, 2006), PASS (Brady and Stouten, 2000) and SURFNET (Laskowski, 1995)], the energy-based method Q-SiteFinder (Laurie and Jackson, 2005) and two meta-predictors MPK1 (Huang, 2009) and MPK2 (Zhang *et al.*, 2011), which make their predictions based on a consensus of those of four and eight other methods, respectively. These success rates were evaluated using the 'distance' criterion (see Section 2). The results showed that LISE was superior to all the non-meta predictors, particularly as regards performance on the top1 success rates, and was comparable with MPK2.

Another widely used dataset for evaluating LBSP performance is a set of 48 proteins for which PDB structures for both the ligand-bound and unbound forms are available (Huang and Schroeder, 2006). There is some overlap between this dataset and the 210 bound-only structure set (29 bound structures in common), but results for more methods are available for comparison using this dataset. As presented in Table 2, these include six additional geometry-based methods, CAST (Liang *et al.*, 1998), PocketPicker (Weisel *et al.*, 2007), MSPocket (Zhu and Pisabarro, 2010), VICE (Tripathi and Kellogg, 2010), DoGSITE (Volkamer *et al.*, 2010) and POCASA (Yu *et al.*, 2010), as well as Fpocket (Le Guilloux *et al.*, 2009), which incorporates structural and physicochemical properties for its prediction. In general, predictions for the bound form are more likely to succeed than those for the unbound form, indicating that, as for docking computations [e.g. Shih and Hwang (2012)], binding-induced conformational changes present some difficulties for LBSP, as noted previously (Huang and Schroeder, 2006; Laurie and Jackson, 2005). Table 2 shows that, for the bound forms, LISE again produced better top1 success rates than the other methods, though, as with the results shown in Table 1, the improvement going from top1 to top3 was less significant than other methods. For the unbound forms, LISE obtained similar results with MPK2 and VICE. However, it should be mentioned that MPK2 is a meta-predictor that built on results of other methods (Zhang *et al.*, 2011), and VICE used

Table 2. Comparison of the top1 and top3 success rates for various LBSP methods using 48 bound/unbound structures[a]

Methods	Bound		Unbound	
	Top1 (%)	Top3 (%)	Top1 (%)	Top3 (%)
LISE (this work)[b]	92	96	81	92
MPK2 (Zhang *et al*., 2011)	85	96	80	94
VICE (Tripathi and Kellogg, 2010)[c]	85	94	83	90
MPK1 (Huang, 2009)	83	96	75	90
DoGSite (Volkamer *et al*., 2010)	83	92	71	92
Fpocket (Le Guilloux *et al*., 2009)	83	92	69	94
LIGSITE^{cs} (Huang and Schroeder, 2006)	81	92	71	85
LIGSITE^{csc} (Huang and Schroeder, 2006)	79	–	71	–
MSPocket (Zhu and Pisabarro, 2010)	77	94	75	88
POCASA (Yu *et al*., 2010)	77	90	75	92
Q-SiteFinder (Laurie and Jackson, 2005)[c]	75	90	52	75
PocketPicker (Weisel *et al*., 2007)	72	85	69	85
CAST (Liang *et al*., 1998)[c]	67	83	58	75
PASS (Brady and Stouten, 2000)[c]	63	81	60	71
SURFNET (Laskowski, 1995)[c]	54	78	52	75

[a] The success rates were computed as the percentage of the 48 bound or unbound structures for which the best (top1) or any one of the best three (top3) predicted binding sites satisfied the 4 Å distance criterion (see Section 2).

[b] Without considering residue conservation [Equation (1)], the success rates were 88% (top1) and 96% (top3) for the bound structures and 79% (top1) and 90% (top3) for the unbound structures.

[c] Data reported in Huang (2009) and Huang and Schroeder (2006).

Table 3. Comparison of the top1 and top3 success rates for various LBSP methods for 48 bound/unbound structures using different precision criteria[a]

Methods	Precision > 0%				Precision > 25%			
	Bound (%)		Unbound (%)		Bound (%)		Unbound (%)	
	Top1	Top3	Top1	Top3	Top1	Top3	Top1	Top3
LISE	92	96	85	92	83	85	67	77
Q-SiteFinder	79	85	52	77	77	83	46	69
POCASA	69	85	73	88	54	75	52	65
Fpocket	19[b]	48[b]	56	88	17[b]	35[b]	44	73

[a] Data were computed from the output of these methods run on their respective server using default parameters (POCASA: http://altair.sci.hokudai.ac.jp/g6/service/pocasa/; Q-SiteFinder: http://www.modelling.leeds.ac.uk/qsitefinder/; Fpocket: http://bioserv.rpbs.univ-paris-diderot.fr/cgi-bin/fpocket).

[b] For reasons unknown, the success rates for the bound structures were much worse than those for the unbound structures, though the correct answer could usually be obtained after the third predicted site.

a set of parameters different from default for this dataset (Tripathi and Kellogg, 2010).

We also evaluated the performance of LISE and three other methods on the same 48 bound/unbound structures shown in Table 2 but using a different evaluation criterion, the ‘precision’ criterion (Laurie and Jackson, 2005) (see Section 2). The three other methods were the geometry-based POCASA (Yu *et al*., 2010), the energy-based Q-SiteFinder (Laurie and Jackson, 2005) and the structure and physicochemistry-based Fpocket (Le Guilloux *et al*., 2009). These methods were chosen because their Internet server provides output data that can be used to compute the success rates using different precision criteria. As shown in Table 3, for all these methods, the success rates for a precision >0% were similar, or even identical, to those using the 4Å, distance criterion (Table 2), and they all decreased when a much more stringent criterion (precision >25%) was used, though the extent of the decrease varied from just a few percent for Q-SiteFinder to as much as 20% for the other methods. However, LISE again yielded the best success rates, especially for the top1 site. LISEs predictions could be made more precise and obtain better success rates for a precision >25% by removing grid points with very low grid point scores, but this resulted in poorer results obtained using the distance criterion shown in Tables 1 and 2 (data not shown), underscoring the different emphases of different evaluation criteria (see Section 4).

3.2 Analysis of unsuccessful cases

Further analysis by visual inspection of the prediction results revealed that, in the 13 cases (6%) in which LISE failed to correctly predict the ligand-binding site within the top three best-scored sites for the 210 bound structures set, it did so because of one of four reasons, summarized below and also in Supplementary Table S2a:

(i) The pocket was identified but not the precise ligand-binding site. Many of LISEs unsuccessful predictions placed the sphere of the top-ranked site (see Section 2) in the same pocket as the ligand-binding site, but the sphere, although close to where the ligand bound, only covered a small part, or even none, of the ligand. For example, of the 13 unsuccessfully predicted cases in the 210 structure set, 6 were kinases with a similar 3D structure, and, for these kinase structures, the first predicted sphere was at the deeper end of a large binding site pocket, entirely missing 5 of the 6 ligands (see Supplementary Fig. S3). The shortest distance between the center of the sphere and the one ligand overlapped a little by the sphere was 4.5 Å, slightly greater than the threshold for success. The same applied to the prediction for yeast hexokinase b (2yhx), for which the shortest distance was 4.2 Å.

(ii) The use of an incomplete biological unit that did not contain all subunit chains. The functional unit for some proteins may consist of more than the single chain presented in the PDB file; when the correct biological unit, as reported in PDBsum (Laskowski *et al*., 1997), was constructed and used for prediction, LISE could correctly identify the ligand-binding site, which often occurs at a subunit interface, as illustrated by the example of acyltransferase (1cla and 3cla; Fig. 2). However, using biological units according to PDBsum would not necessarily improve the success rates, because the added pocket space at the subunit interface would create new spheres, which could alter LISEs sphere ranking; for this study, the number of the cases moved up to within top3 was smaller than those moved out (3 versus 6), resulting in up to 4% decrease in the success rates (details in Supplementary Table S3).

(iii) A non-biological ligand. Some of the PDB structures used for LBSP may contain ligand molecules that are not the biological ligand of the protein but are agents used to reduce disulfide bonds or assist crystallization. For example, the ligand beta-mercaptoethanol in the PDB structure of bacteriophage T4 lysozyme (PDB 1l82) binds at a site different from the enzyme's catalytic site (Hurley *et al*., 1992), and LISEs first predicted sphere overlapped with the catalytic site but not the ligand in the PDB structure (Supplementary Fig. S4).

(iv) A small ligand interaction area. Two unsuccessful cases for the 210 bound structures set, lectin (2msb) and glucanotransferase (1cdg), shared a common feature in having a very small ligand-binding site, about a quarter or a half, respectively, of that in a complex (e.g. 1d3h) with a ligand of comparable size (see Supplementary Fig. S5).

Of the 48 bound/unbound structures, LISE failed to correctly predict five pairs (within the top3 sites) for either the bound or the unbound form, or both (Supplementary Table S2b). Two were due to the reasons 'the pocket was identified, but not the precise ligand-binding site' for 1qpe (bound)/3lck (unbound), one of the kinase structures (Supplementary Fig. S3), and 'the use of an incomplete biological unit that did not contain all subunit chains' for the insulin complex, 3mth (bound)/6ins (unbound) (Supplementary Fig. S6). In the other three cases, LISE correctly predicted the bound form, but not the unbound form, resulting in a decrease in success rates of a few percentages for the unbound form (Table 2). Further examination showed that, in these three cases, the failure was caused by either a disordered loop (atom coordinates not available) at the ligand-binding site for the unbound form (dihydrofolate reductase, 5dfr) (Supplementary Fig. S7), or, in the case of HIV protease (4phv/3phv and 1ida/1hsi), a significant change in the conformation of a binding site loop, which resulted in a much larger site for the unbound form, and, consequently, some protein triangles formed in the bound form were lost in the unbound form, pushing LISEs sphere for this site out of the top three (Supplementary Fig. S8).

Another complication for LBSP is that some proteins may have more than one biologically relevant ligand-binding sites. For example, apart from the active site, HCV polymerase NS5B, which plays an essential role in viral replication and has attracted much attention for drug design (Kwong *et al*., 2008), has two other ligand-binding sites that can be targeted by small molecules to inhibit its polymerase function (Bressanelli *et al*., 2002; Hang *et al*., 2009; Le Pogam *et al*., 2006). As shown in Figure 3, LISEs top three spheres perfectly matched the three different binding sites of NS5B, attesting to its ability to suggest potential drug binding sites in addition to the first predicted site.

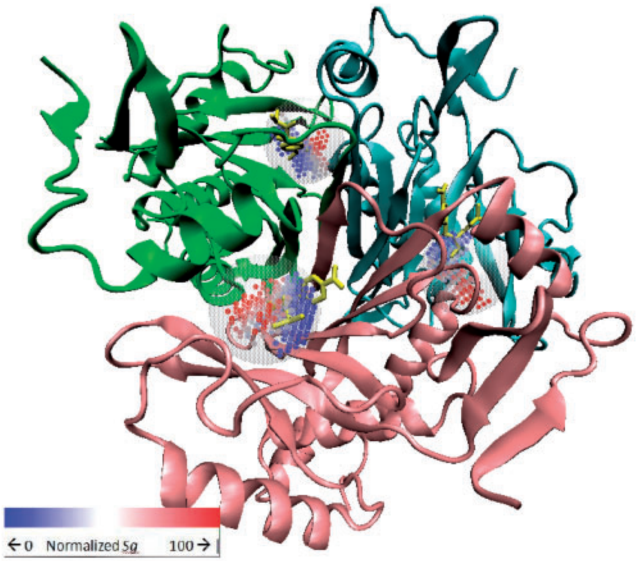

Fig. 2. LISE identifies the three symmetry-related ligand-binding sites at the subunit interface of the acyltransferase trimer. The three identical ligands, depicted by yellow sticks, fell, respectively, within 4 Å of the center of LISEs top three spheres. The three spheres had essentially the same sphere score [S_s, Equation (3) in Section 2] and did not form when the monomer structure of acyltransferase (1cla) was used. The transparent spheres contain grid points represented by dots colored from blue to red to indicate the normalized value (see color spectrum) of their grid score [S_g, Equation (2) in Section 2]. This figure was created using the VMD program (Humphrey *et al*., 1996).

3.3 Application to CASP8 and CASP9 targets

To further evaluate LISE, we also made predictions for targets in the LBSP section of CASP8 and CASP9. The ligand-bound structures of CASP8 and CASP9 targets have been recently determined, and most have been released in the PDB, making the evaluations possible. Of the total of 56 targets in these two CASP experiments, we omitted the 37 in which the ligand was either a metal ion or a solvent molecule, such as glycerol, or the 3D structures are not yet publicly accessible in the PDB. As shown in Supplementary Table S4, of the 19 targets (10 from CASP8 and 9 from CASP9) with a full ligand that were tested, LISE successfully predicted the ligand-binding site as the top-ranked site for 17 targets. The top1 success rate was therefore 89%, which is similar to those presented earlier for the 210 bound and 48 bound/unbound benchmark datasets, attesting to the robustness of LISEs predictions.

4 DISCUSSION AND CONCLUSION

In this study, we showed that LISE, a new LBSP method derived from 3D motifs of protein–ligand interactions (Xie and Hwang, 2010), achieved significantly better success rates,

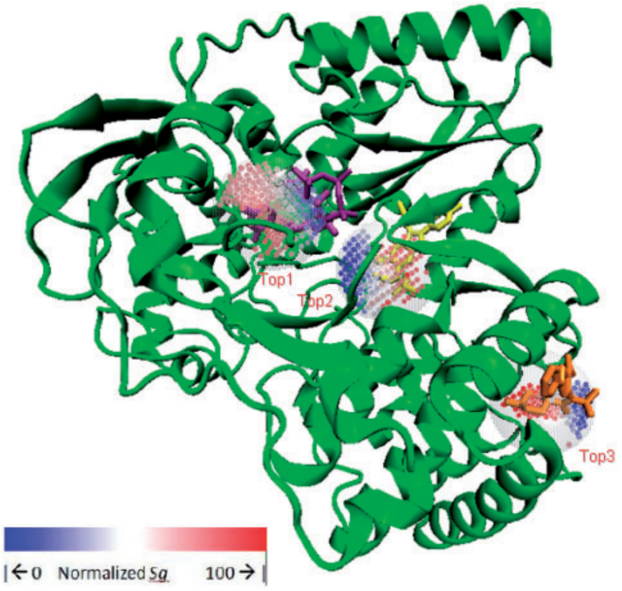

Fig. 3. LISE identifies the three different ligand-binding sites of HCV polymerase NS5B as its top three predicted sites. The protein–ligand complex structures of HCV polymerase NS5B for the three different ligands (depicted by sticks), HCV-796 (yellow), NNI-1 (orange), or UTP (purple), were solved separately (PDB 3fqk, 2gir and 1gx6, respectively); the three complex structures could be superimposed well, so, for clarity, the protein structures of 2gir and 1gx6 are omitted. LISEs top three predicted sites are shown and labeled, with their grid points (dots) colored according to the grid score, S_g [Equation (3) in Section 2], from blue (lowS_g) to red (highS_g), as indicated by the color spectrum. The figure was created using the VMD program (Humphrey *et al.*, 1996).

especially in predicting the top-ranked sites, than a number of previously reported methods on two benchmark datasets (Tables 1–3) and equally good success rates for 19 CASP targets (Supplementary Table S4). Individual examination suggested that LISEs performance, measured by these success rates, may in fact be underestimated, as, for example, many of the predictions identified the pocket of the ligand-binding site but did not meet the threshold used for success (Supplementary Table S2a and b and analysis in Section 3).

LISE is conceptually similar to the propensity-based method of Mehio *et al.* (Mehio *et al.*, 2010) in using a triangle of protein atoms enriched at ligand-binding sites. However, unlike the triplet of Mehio *et al.*, which composed of three mutually contacting surface atoms of the protein, LISEs triangles are part of a protein–ligand interacting motif, with the three protein atoms being simultaneously in contact with at least two ligand atoms and the size of the triangle being such that very close and chemically bonded protein atoms are excluded (Xie and Hwang, 2010). Furthermore, although the triplet propensities of Mehio *et al.* are mapped onto protein surface atoms and residues, LISEs triangle scores are mapped out into the empty space of the binding site (see Section 2), which eliminates the requirement in previously reported propensity-based methods (Mehio *et al.*, 2010; Soga *et al.*, 2007) of using binding site pockets predicted by other, geometry-based methods, thereby allowing evaluation to be performed independent of other LBSP methods.

Different evaluation criteria will affect whether a prediction is regarded as a success or not, which may lead to different conclusions about prediction results and performance of the prediction methods. Furthermore, proteins have evolved diverse binding sites to accommodate ligands of different sizes and chemical properties (Grkovic *et al.*, 2003; Ma *et al.*, 2002), and, adding to this complexity, a small ligand can bind to a large pocket (e.g. see Supplementary Fig. S9), and all these factors make it difficult to devise a single criterion to evaluate all the different LBSP methods objectively. Thus, for example, compared with the 'coverage' criterion (An *et al.*, 2005; Gutteridge *et al.*, 2003), the 'precision' criterion (Laurie and Jackson, 2005; Morita *et al.*, 2008) might allow a more precise assessment of how much of the predicted pocket is occupied by the ligand but cannot rule out the possibility that the predicted pocket, even using 100% precision, covers only a small part of the ligand (Laurie and Jackson, 2005). Alternatively, avoiding the difficulty of choosing a criterion to evaluate the success rates of LBSP using defined regions of empty space (i.e. pockets), one can evaluate whether an amino acid residue is in contact (e.g. within a certain distance) with the ligand for all the residues, allowing computation of the prediction's sensitivity and specificity (Capra *et al.*, 2009; Wass and Sternberg, 2009). However, there are cases in which different ligands, which may or may not be of similar size, bind to the same protein in the same pocket but in slightly different positions (e.g. Supplementary Fig. S10), such that a residue considered as a binding site residue for one ligand becomes a non-binding site residue for another, making the computed sensitivity and specificity values somewhat dependent on the structures containing the particular ligand used. Moreover, as the purpose of LBSP is usually to identify a pocket on which to carry out docking computations, a small number of false-positive or false-negative binding site residues will not seriously compromise the subsequent work, so long as a pocket, either revealed by clustering the predicted binding site residues or adopted from those predicted by other methods, can be identified.

Given these complications, in this work, we chose to predict the ligand-binding site as a sphere of a fixed size, 5.5 Å radius (see Section 2), which is about the average size of the ligand-binding sites surveyed by Gutteridge *et al.* (Gutteridge *et al.*, 2003). Further analysis also showed that, for most of the successfully predicted sites, a good percentage (>30%) of the triangles was in contact with ligand atoms, in comparison with very few (<10%), if any, for the unsuccessfully predicted sites (Supplementary Fig. S11). One obvious limitation of this simple representation is that it does not reflect well the generally non-spherical shape of a protein's ligand-binding sites; another drawback is that, for a large site, such as the examples shown in Supplementary Figures S3, S4 and S6, LISEs top three spheres were often located in essentially the same pocket, and this, by not allowing them to be located in other sites, partly explains the relatively smaller improvement achieved on going from top1 to top3 success rates using LISE compared with other methods (Tables 1–3). Nevertheless, the fact that this simple scheme did not prevent LISE from attaining excellent results is encouraging. The results also suggest that the protein triangles of the protein–ligand interaction motifs have, to a good extent, uncovered the footprint of evolution in protein structures that interact with small molecules. In principle, the general approach employed can be used to predict sites of other types of protein interactions, such as those between protein and protein or between protein and nucleic acid.

ACKNOWLEDGEMENTS

We thank the National Center for Genome Medicine at Academia Sinica, Taiwan, for helping to establish the LISE web server.

Funding: Support for this project (NSC 96-2627-B-001-004) and a post-doctoral research fellowship for Z.-R.X. (NSC 100-2811-B-001-067) were provided by the National Science Council of Taiwan.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

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

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Brady,G.P. Jr and Stouten,P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.*, **14**, 383–401.

Bressanelli,S. *et al*. (2002) Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides. *J. Virol.*, **76**, 3482–3492.

Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.

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

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

Grkovic,S. *et al*. (2003) Interactions of the QacR multidrug-binding protein with structurally diverse ligands: implications for the evolution of the binding pocket. *Biochemistry*, **42**, 15226–15236.

Gutteridge,A. *et al*. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330**, 719–734.

Hang,J.Q. *et al*. (2009) Slow binding inhibition and mechanism of resistance of non-nucleoside polymerase inhibitors of hepatitis C virus. *J. Biol. Chem.*, **284**, 15517–15529.

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, 389.

Huang,B. (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS*, **13**, 325–330.

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

Humphrey,W. *et al*. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38, 27–38.

Hurley,J.H. *et al*. (1992) Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J. Mol. Biol.*, **224**, 1143–1159.

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.

Kwong,A.D. *et al*. (2008) Recent progress in the development of selected hepatitis C virus NS3.4A protease and NS5B polymerase inhibitors. *Curr. Opin. Pharmacol.*, **8**, 522–531.

Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330, 307–328.

Laskowski,R.A. *et al*. (1997) PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.

Laskowski,R.A. *et al*. (2005) Protein function prediction using local 3D templates. *J. Mol. Biol.*, **351**, 614–626.

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.

Laurie,A.T. and Jackson,R.M. (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.*, **7**, 395–406.

Le Guilloux,V. *et al*. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.

Le Pogam,S. *et al*. (2006) Selection and characterization of replicon variants dually resistant to thumb- and palm-binding nonnucleoside polymerase inhibitors of the hepatitis C virus. *J. Virol.*, **80**, 6146–6154.

Leis,S. *et al*. (2010) In silico prediction of binding sites on proteins. *Curr. Med. Chem.*, **17**, 1550–1562.

Levitt,D.G. and Banaszak,L.J. (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.*, **10**, 229–234.

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

López,G. *et al*. (2009) Assessment of ligand binding residue predictions in CASP8. *Proteins*, **77**, 138–146.

Lopez,G. *et al*. (2011) Firestar—advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.

Ma,B. *et al*. (2002) Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.*, **11**, 184–197.

Mehio,W. *et al*. (2010) Identification of protein binding surfaces using surface triplet propensities. *Bioinformatics*, **26**, 2549–2555.

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

Muegge,I. and Martin,Y.C. (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.*, **42**, 791–804.

Roche,D.B. *et al*. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*, **12**, 160.

Ruvinsky,A.M. and Kozintsev,A.V. (2005) The key role of atom types, reference states, and interaction cutoff radii in the knowledge-based method: new variational approach. *Proteins*, **58**, 845–851.

Schmidtke,P. *et al*. (2010) Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.*, **38**, W582–W589.

Shih,E.S. and Hwang,M.J. (2012) On the use of distance constraints in protein-protein docking computations. *Proteins*, **80**, 194–205.

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

Tripathi,A. and Kellogg,G.E. (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins*, **78**, 825–842.

Volkamer,A. *et al*. (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.*, **50**, 2041–2052.

Wass,M.N. and Sternberg, M.J.E. (2009) Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins*, **77**, 147–151.

Wass,M.N. *et al*. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.

Wass,M.N. *et al*. (2011) Challenges for the prediction of macromolecular interactions. *Curr. Opin. Struct. Biol.*, **21**, 382–390.

Weisel,M. *et al*. (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*, **1**, 7.

Xie,Z.-R. and Hwang,M.-J. (2010) An interaction-motif-based scoring function for protein-ligand docking. *BMC Bioinformatics*, **11**, 298.

Yu,J. *et al*. (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, **26**, 46–52.

Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**, 40.

Zhang,Z. *et al*. (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**, 2083–2088.

Zhu,H. and Pisabarro,M.T. (2010) MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics*, **27**, 351–358.