Protein Engineering, Design & Selection vol. 18 no. 2 pp. 65–70, 2005
Published online March 30, 2005 doi:10.1093/protein/gzi006

A novel statistical ligand-binding site predictor: application to ATP-binding sites

Ting Guo[1,2], Yanxin Shi[2,3] and Zhirong Sun[1,4]

[1]Institute of Bioinformatics, MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology and [3]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[4]To whom correspondence should be addressed.
E-mail: sunzhr@mail.tsinghua.edu.cn

[2]These authors contributed equally to this work

Structural genomics initiatives are leading to rapid growth in newly determined protein 3D structures, the functional characterization of which may still be inadequate. As an attempt to provide insights into the possible roles of the emerging proteins whose structures are available and/or to complement biochemical research, a variety of computational methods have been developed for the screening and prediction of ligand-binding sites in raw structural data, including statistical pattern classification techniques. In this paper, we report a novel statistical descriptor (the Oriented Shell Model) for protein ligand-binding sites, which utilizes the distance and angular position distribution of various structural and physicochemical features present in immediate proximity to the center of a binding site. Using the support vector machine (SVM) as the classifier, our model identified 69% of the ATP-binding sites in whole-protein scanning tests and in eukaryotic proteins the accuracy is particularly high. We propose that this feature extraction and machine learning procedure can screen out ligand-binding-capable protein candidates and can yield valuable biochemical information for individual proteins.

Keywords: ATP-binding site/binding site prediction/Oriented Shell Model/protein–ligand interaction/support vector machine

Introduction

High-throughput projects in structural genomics, aimed at exhaustively ‘covering’ the genome with protein structural data, are leading to an increasingly large databank of protein three-dimensional (3D) structures. It is likely, however, that many of these emerging structures will be relatively poorly understood in terms of exact biological or biochemical function (Kinoshita and Nakamura, 2003). On the other hand, the rapid accumulation of tertiary structures aptly represents a foundation for subsequent functional and mechanistic characterization (Burley *et al*., 1999; Vitkup *et al*., 2001). The detection of ligand-binding sites, in particular, has been the target of considerable research effort as it can provide hints about protein function and also facilitate the drug design process.

Ligand-binding sites, or functional sites, can be recognized by a variety of different cues (Campbell *et al*., 2003). One intuitive way is to trace the conservation of amino acid residues in protein families for functionally important sites (Lichtarge and Sowa, 2002). Recent developments of the evolutionary tracing method include mapping conserved residues on to a protein surface (Pupko *et al*., 2002) and analysis of inter-family conservation consistency (Kunin *et al*., 2001; Friedberg and Margalit, 2002).

Alternatively, functional site predicting can be approached from an energetic point of view. Molecular docking exploits statistical mechanics and quantum chemistry calculations of binding energies in view of molecular force fields (Goodford, 1985), hydrogen bonding (Wade *et al*., 1993a,b), hydrophobic interaction (Kellogg *et al*., 1991) and/or solvation energy (Pitt and Goodfellow, 1991). Considering the chemistry of protein–ligand interactions, docking is probably the most natural, simulative approach to functional site prediction. Nonetheless, molecular docking is usually very computationally costly and as a result its application to genome-wide ligand-binding site screening is only at a pioneering stage (Pang *et al*., 2001; Jackson, 2002).

Since distinguishing a functional site from a ‘non-site’ is essentially a two-class classification problem, statistical pattern recognition methods have also been introduced. Work of this type focused on forging a statistical 3D template via machine learning of known binding sites. For instance, Di Gennaro *et al*. (2001) developed a ‘fuzzy functional form’ descriptor for disulfide oxidoreductase and applied it to the functional annotation of the *Bacillus subtilis* genome. For recognition from structural clues, earliest efforts were devoted to discovering conserved patterns in peptide sequences, but the accuracy was a concern (Devos and Valencia, 2000). Rantanen *et al*. (2001) divided atoms from both the ligand and its receptor into many classes in terms of their chemical environment and modeled their probabilistic spatial relations, leading to a reduced prediction error. This model, however, did not take into account heterogeneity of functional sites at the level of atom types. Wei and Altman (2003) looked at a collection of physicochemical properties, scoring them with structures in the PDB in a spherically symmetrical fashion by summing scores associated with atoms at various distances from the site center. In this protocol, orientation relationships of features are lost, which probably leads to a sensitive although unspecific predictor. Studies that used neural networks to predict active sites have also been presented (Gutteridge *et al*., 2003).

In this paper, we report a novel 3D descriptor of ligand-binding sites in proteins. This Oriented Shell Model (OSM) takes into consideration both the distance and the orientation information of a variety of physicochemical properties around a functional site. These properties are aimed at exhaustively extracting useful information around a binding site. Via the use of the support vector machine (SVM), irrelevant properties are spontaneously ignored in the final prediction process. Using ATP-binding sites as a case study, our results show relatively high sensitivity and specificity, as evidenced in a set of whole-protein search tests. Moreover, different taxonomic groups seemingly have their own preferred prediction parameters, opening up the possibility of a more refined genome-scale interpretation of structural data.

Materials and methods

Theory and model

For any given type of ligand-binding site, a center atom and two reference atoms are arbitrarily chosen from the ligand molecule to set up an unequivocal *xyz*-coordinates reference frame. In the ATP-binding site, C1* is designated as center and PG, C4 as reference atoms (Figure 1). A coordinate system is set up with reference to these three atoms. Next, a series of gradually enlarging, concentric spheres are defined, all centered at C1* and equally spaced by 1.25 Å (Wei and Altman, 2003). The outmost sphere in this sphere set should at least fully encompass the ligand molecule (in the case of ATP, this means 12 shells in all with the largest radius of 15 Å). The volume enclosed between every two neighboring spheres thus specifies a 'shell' in which atoms can be considered roughly equidistant from C1*. Next, each shell is further subdivided into six 'blocks', each block occupying a different direction in the *x*, *y* or *z* axis (Figure 1). In this way, the vicinal space around ATP-binding site is partitioned into 72 bonnet-like blocks contained in nested shells.

Blocks could overlap each other and some atoms could belong to two or three blocks. This allows for some flexibility in the machine-learned standard template for functional sites. A block can be regarded as the part of a shell intersected by a sphere with a radius *r*, where *r* specifies the size of a block given a shell radius. *R* is the radius of the shell; in our implementation, it is approximated as the arithmetic average of the radii of the inner and outer spheres. To cover a shell fully, *r* should be >0.9194*R*. To avoid oppositely positioned shells from overlapping, *r* has an upper limit of 1.4142*R*. Deciding the value of *r* represents a means for controlling the stringency of prediction.

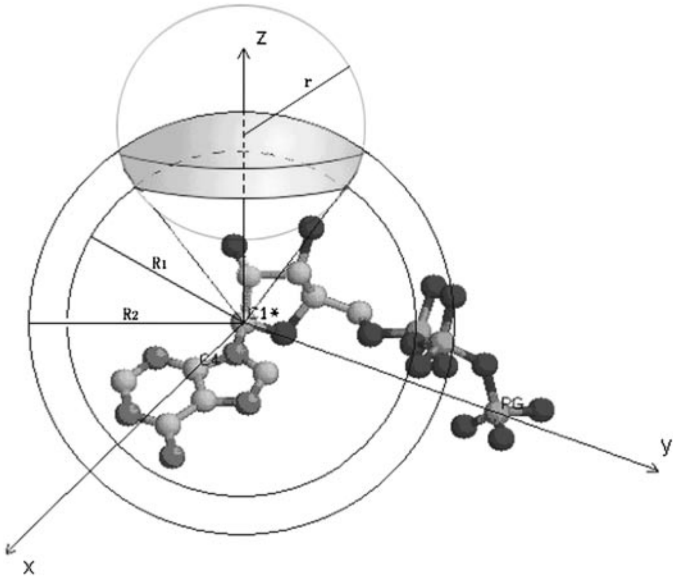

Fig. 1. Schematic representation of the shell-block system. The site-proximal space is partitioned by a series of enlarging shells, all centered at C1*. A shell with radius *R* is illustrated. A block (colored gray) can be regarded as the part of a shell intersected by a sphere with a radius *r*; where *r* specifies the size of a block.

We use a group of physicochemical properties largely as reported previously (Wei and Altman, 2003), which included atom types, amino acid residue types and chemical group type, partial atomic charge, hydrophobicity, van der Waals radii of atoms, peptide backbone mobility and secondary structure read from DSSP (Kabsch and Sander, 1983). *B*-factors are not used because they are comparable only intrastructurally. Scores are assigned to atoms and are summed over all atoms within a block for each property. The ligand molecule itself, if present, was removed prior to data extraction for the training set.

Data set

A total of 174 structures deposited before July 2004 in the Protein Data Bank (PDB) (Berman *et al*., 2000) are complexed with ATP. After eliminating a few low-quality or obsolete structures, 230 ATP-binding sites remained, 10 of which were reserved as the test set. The remaining 220 binding sites were fully employed as the 'site' training set. The 'non-site' training set arose from two sources. From those structures used in the training set, we randomly picked up 94 non-site positions as training samples. In addition, we deliberately picked up another 410 non-sites randomly, all within 12 Å but more than 5 Å further from the center of an ATP-binding site, so as to sharpen the classifier against the nuances between true sites and their surroundings, which often bear a site-like chemistry. Nonetheless, the choosing of these 'para-site' non-sites was through a random procedure. Non-site samples were read in a random reference frame.

The non-redundant structure set referred to in the Discussion was generated by removing from the training set homologous structures of sequence identity >30% with the Blastclust program of the BLAST package (Altschul *et al*., 1990). One representative structure was picked out from each homology cluster while intentionally preserving most test set samples for the sake of comparison. This leads to a training set comprising 66 sites and 485 non-sites.

Classifier: the support vector machine

We utilized the support vector machine (SVM) method (Vapnik, 1995, 1998) in the two-class classification problem of identifying ligand-binding sites. SVM seeks an optimal separating hyperplane (OSH) in a transformed high-dimensional Hilbert space in which training and test samples are presented. We used in our study a software tool for SVM classification developed by Chang and Lin (2001).

The order of specifying blocks in feature vectors follows a definite spatial route, ensuring the correct spatial register of features in the vector. The kernel function of SVM was the Radial Basic Function (RBF) kernel:

$$K\left(x_i, x_j\right) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$$

where γ is a coefficient to be optimized. To define a SVM classifier, yet another parameter, *C*, which controls the trade off between margin and misclassification error, must be determined. *C* and γ of the kernel function were experimentally tuned to achieve best performance.

Whole-protein scanning for ligand-binding site

In classifying a query position in a protein, a feature vector is read as for generating the training set, but with a random reference frame. Next, a systematic 24 coordinate systems transformation is performed to check for possibilities in other orientations. These 24 orientations are obtained by rotating the original reference frame to cover the full sphere surface while keeping furthest apart from each other. Only when all these 24 systems gave negative results does the classifier regard the query as a non-site, analogous to the lock-and-key model of ligand binding. Probabilistic estimation shows that for 12 shells each containing up to 10 characteristic 'trait points' that distribute randomly within the shell, the probability that at least one of the 24 transformations still retains >90% traits in original shell-blocks is about 47% when r is equal to R. Considering the chemical similarity and the fact that we often observe hits in clusters, the mathematical expectation of hits reported around a site is well above one. In our studies, predicting only the 24 coordinate systems indeed worked fairly well.

Beginning with a protein structure, we first build a 3D grid with grid spacing 2.5 Å. We tested a group of four proteins each with 10 independently generated random grid origins and in only three out of the 40 cases did the number of true positives or false negatives differ. Hence we believe that using a random grid origin will not significantly affect the prediction result. Then, for each grid point that was inside the protein or within a reasonable distance from its surface, we applied the 24 coordinate transformations to read 24 data for a single point, which were subsequently processed by the SVM classifier. Our current implementation takes about 1 h to scan a protein structure for ATP-binding sites on a Pentium IV 2.3 GHz PC. We experimented with a wide range of (C, γ) value sets in each protein we tested to its best performance.

Cross-domain prediction accuracy

As an attempt to analyze the potential divergence between eukaryotic binding site structures and their prokaryotic counterparts, the complete training set was split into two subsets according to the two taxonomic domains, namely, the structures from the Eukarya and those from the Prokarya. Next, three SVMs were trained on the all, Eukarya and Prokarya data, respectively and the resulting classifiers were used to predict all the three groups of the training set to obtain the accuracy on training set.

Results

Cross-validation of SVM classifiers

We performed 5-fold cross-validations of SVM classifiers for ATP-binding sites with two empirically determined outperforming (C, γ) pairs (Table I), one suitable for eukaryotes and the other for prokaryotes. Cross-validation accuracy was defined as the overall percentage of correctly classified training samples over the training set. As eukaryotic and prokaryotic ATP-binding sites apparently exhibit a certain degree of difference (Table III and IV), the cross-validation accuracy might have been undermined by random partitioning of the training set. However, the overall accuracy from two categories is still ~85%, significantly higher than a random classifier. The accuracy exhibited in cross-domain prediction (Table II) is much higher. This indicates that with this feature-extraction and machine learning scenario, ATP-binding sites and non-sites were indeed mapped into two recognizably separate regions in the high-dimensional space.

Table I. Cross-validation accuracy with two sets of empirical parameters

C	γ	Cross-validation accuracy (%)
0.15	0.0078125	88.2434
0.52	0.0078125	84.5090

Whole-protein ATP-binding site scanning

We next tested our algorithm on an array of 11 whole-protein functional site searches. A 2.5 Å spacing grid was built superimposed on each protein and each grid point was subjected to SVM classification. The block size r was set to R and γ for the RBF kernel was 0.0078125. C was optimized in each case. We did not include a non-ATP-binding protein as a control because the presence of 'non-site' query positions inherently served as numerous negative controls.

Table III summarizes the results. Apparently, there is a distinct tendency for optimal C values favored by the eukaryotes (~0.15) and prokaryotes (~0.52). Further, the prediction system was highly accurate and precise, especially for eukaryotes.

Using the two empirical optimal C values, we re-tested the whole-protein scanning power of the predictor on the same set of proteins (Table IV). In 69% of the cases the predictor was able to identify the binding site correctly (the ADP-binding protein was not counted). The precision for eukaryotic proteins was fairly high (60%) but in prokaryotes there were more false positives, leading to lower precision, similar to what happened with an influenza-derived viral protein (PDB code 1JJV).

Figure 2 shows one visualized instance of prediction results. In the human Aurora-A protein kinase (PDB code 1OL6; Bayliss *et al*., 2003), the SVM recognized two query positions as binding site in close proximity to C1*, in addition to a false positive found in a surface cleft. In our study, the SVM almost always picked out 'clusters' of a few closely associated hits

Table II. Cross-domain prediction accuracy (%)

Values of parameters	Train test	All	Eukaryotic	Prokaryotic
C = 0.52, γ = 0.0078125	All	99.0541	86.8919	94.3243
	Eukaryotic	99.6644	99.3289	92.2819
	Prokaryotic	98.3108	79.0541	98.6486
C = 0.15, γ = 0.0078125	All	95.7950	74.9014	68.3311
	Eukaryotic	98.6577	88.9262	72.8188
	Prokaryotic	93.5811	72.6351	75.6757

Table III. Searching for optimal C in whole-protein scanning

Proteins	PDB ID	Predicted/all sites	False positives	Optimal C	Source
Eukaryotic	1ol6	1/1	1	0.13	Human
	1nsf	1/1	0	0.18	Hamster
	1phk	1/1	0	0.16	Rabbit
	1ql6	1/1	0	0.14	Rabbit
	1hck	1/1	0	0.14	Human
	1am1 (ADP)	1/1	0	0.52	Yeast
Prokaryotic	1dy3	1/1	1	0.52	*E.coli*
	1jjv (viral)	1/1	2	0.52	*H.influenzae*
	1a82	1/1	1	0.52	*E.coli*
	1ji0	1/1	3	0.50	*Thermotoga*
	1b0u	0/1	6	0.52	*Salmonella*
	1mjh	2/2	3	0.52	*Methanococcus*

Table IV. Summary of whole-protein scanning results

Proteins	PDB ID	Predicted/all sites	False positives	*C* value	Source
Eukaryotic	1ol6	1/1	1	0.15	Human
	1nsf	0/1	0	0.15	Hamster
	1phk	0/1	0	0.15	Rabbit
	1ql6	1/1	0	0.15	Rabbit
	1hck	1/1	1	0.15	Human
	1am1 (ADP)	1/1	0	0.52	Yeast
Prokaryotic	1dy3	1/1	1	0.52	*E.coli*
	1jjv (viral)	1/1	2	0.52	*H.influenzae*
	1a82	1/1	1	0.52	*E.coli*
	1ji0	0/1	5	0.52	*Thermotoga*
	1b0u	0/1	6	0.52	*Salmonella*
	1mjh	2/2	3	0.52	*Methanococcus*

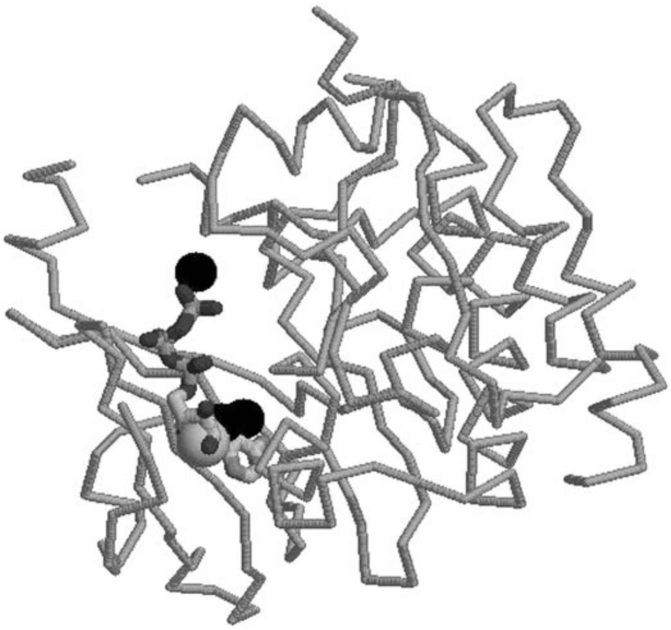

Fig. 2. Visualized result of whole-protein scanning for 1OL6. The backbone of the human Aurora-A kinase (PDB code 1OL6) is represented as ribbons. The ATP molecule has been added back and is shown as a ball-and-stick model and C1*, the arbitrarily defined center atom of ATP, is space-filled and colored gray. Hits are colored black.

rather than scattered, sporadic single hits. This probably reflects the local similarity in physicochemical features of surface crevices or clefts. We regard such a clearly shaped hit cluster as a predicted binding site.

Cross-domain prediction accuracy

To tentatively address the potential discrepancy in binding site structure between taxonomic domains further, we next calculated the cross-domain prediction accuracy. Three SVMs were trained on the all, Eukarya and Prokarya data, respectively, using the two optimal parameter sets, followed by calculation of training error of all the three groups of the training set (Table II). In both cases, SVM trained with prokaryotic samples exhibited lower and even unacceptable accuracy in predicting eukaryotic queries and vice versa. Nevertheless, when *C* is 0.15, the value preferred by eukaryotes, SVMs trained on both all and eukaryotic data yielded very high accuracy. On the other hand, when *C* is 0.52 (the prokaryotic penchant), all- and prokaryote-derived SVMs again showed comparably high accuracy. The reason for this differential response to eukaryotic and prokaryotic ATP-binding sites is unknown, although it is possibly due to statistical differences in physicochemical feature distribution. Nevertheless, it is advisable to apply different *C* values when treating a protein with a known organismic source.

Discriminating power of the oriented shell model

Most, if not all, ATP-binding sites are simultaneously catalytic sites which after hydrolysis reaction and conformational changes can bind ADP. The specificity or discriminating power of this prediction system was therefore assessed in two experiments. In one, two proteins complexed with ADP were subjected to prediction. In the other, two GTP-binding sites were examined to test cross-predictability.

Our prediction system identified the ADP-binding site in yeast Hsp90 molecular chaperone (PDB code 1AM1p; Prodromou *et al*., 1997) as an ATP-binding site. The bovine F1-ATPase structure (PDB code 1NBM) has three active sites caught in an ATP-binding conformation and three others in an ADP-binding state. In an experiment in which the 12 Å-proximal regions of active sites were searched, both types were recognized (data not shown). Therefore, it seems that this prediction system was capable of recognizing ATP-hydrolysis active site in both conformations.

We then tested the system on two GTP-binding sites, those of an *Escherichia coli* Moba protein (1FRW; Lake *et al*., 2000) and a mouse adenylosuccinate synthetase (1LOO; Iancu *et al*., 2002). In both cases, the SVM again responded positively around each site (data not shown). The classifier apparently failed to distinguish a GTP-binding site from its ATP-binding counterpart.

Influence of allosteric effect

The structure of a bacterial Rad50 ATPase whose dimerization is induced by ATP binding is available in the PDB in both ATP-bound and ATP-free states (PDB code 1F2U and 1F2T; Hopfner *et al*., 2000). The ATP-binding site lies at the interface between the two monomers. We conducted a local search in the region. The complete ATP-binding site in 1F2U was identifiable with our prediction system. In contrast, the half site existing before dimerization was not identified (Figure 3). It has been reported that the binding of ATP γ-phosphates to opposing conserved signature motifs in two opposing Rad50cd molecules promotes dimerization that likely couples ATP hydrolysis to dimer dissociation and DNA release (Hopfner *et al*., 2000). In this respect, the 1F2T half site does not have a characteristic ATP-binding microenvironment and it is no surprise that our method failed to report this site. This result is an example that some proteins exhibit different 3D structure and fundamentally different affinity for their ligands in different conformations and annotating their structure in only one conformational state may lead to deceptive conclusions.

Discussion

Significance of sequence homology in training set

The feature extraction scenario in this work captures physicochemical properties that distribute three-dimensionally. Because proteins fold into 3D structures after which the

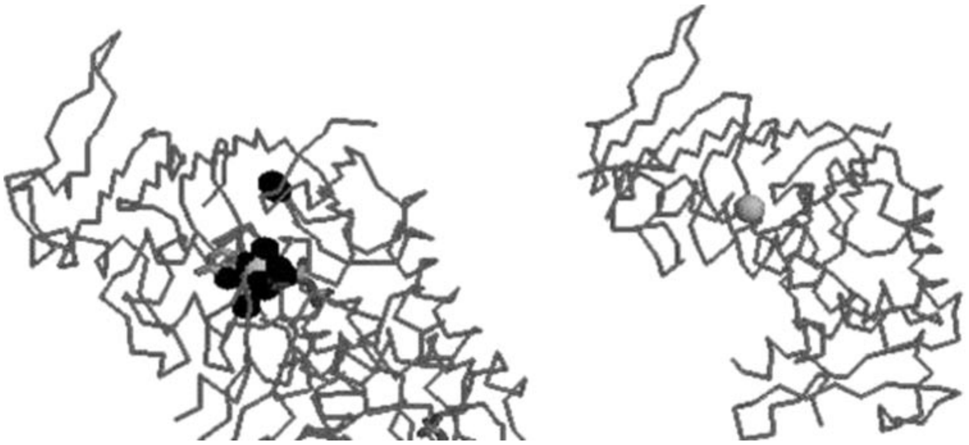

Fig. 3. Comparison of prediction results of an ATPase dimer and its constituent monomer. A local search in the region within 12 Å of the site center was conducted for both the bacterial Rad50 ATPase dimer (1F2 U) and monomer (1F2T). The color scheme is as in Figure 2. The complete ATP-binding site in 1F2U was identified with our prediction system when *C* was 0.52, but the half site in 1F2T was not, as expected.

Table V. Summary of whole-protein scanning results with non-redundant training set

Proteins	PDB ID	Predicted/all sites	False positives	*C* value	Source
Eukaryotic	1nsf	1/1	0	0.53	Hamster
	1phk	1/1	0	0.53	Rabbit
	1am1 (ADP)	1/1	5	1.72	Yeast
Prokaryotic	1jjv (viral)	1/1	6	1.72	*H.influenzae*
	1a82	1/1	3	1.72	*E.coli*
	1ji0	0/1	5	1.72	*Thermotoga*
	1b0u	0/1	6	1.72	*Salmonella*

distribution of residues does not correlate well with their order in primary sequence, sequence conservation of ATP-binding motifs of certain types is not likely to contribute significantly in our predictor. To test this, we generated a training set within which pairwise sequence identities are <30% and repeated the whole-protein ATP-binding site scanning experiments. Comparison of the results with the non-redundant training set (Table V) with the original training set (Table IV) reveals that there is indeed no decline in performance after removal of homologous sequences, as expected.

Sensitivity to conformational changes and structural minutiae

Conformational changes accompanying induced fit during ligand binding may pose a recognition problem to a classifier trained on a ligand-complexed state in identifying the same sites in apoproteins. A recent review (Gutteridge and Thornton, 2004) countered some of the suspicion where 11 enzymes were examined and in most of these enzymes only a relatively small amount of conformational change was observed. This is particularly true for residues directly involved in catalysis, with an r.m.s.d. of a C-α trace usually <1 Å.

Consistent with this observation, our prediction system correctly recognized the two ADP-binding sites in 1AM1 and 1NBM. In either case, the ADP-binding site is actually an active site that catalyzes the hydrolysis of ATP to ADP and is more properly termed an ATP/ADP-binding site. Therefore, it seems possible to predict ligand-binding sites from ligand-free apo-state structures as long as dramatic allosteric control is absent.

This statistical descriptor apparently failed to discriminate between GTP- and ATP-binding sites. This is not surprising. GTP and ATP differ by only two substitutions on the purine ring at C-2 and C-6, but otherwise share a similar overall geometry. The distance from GTP C1* to 6-O is 5.03 Å, so the block containing 6-O is roughly 4.3 Å in radius. In our statistical descriptor, a block that large is unlikely to reveal such structural details as other atoms contained in the block could have overwhelmed the difference.

Parameter setting for the SVM

We observed that a larger block size, translatable to a larger overlapping area between neighboring blocks, should lead to lower stringency in prediction and thus a higher occurrence of false positives and vice versa (data not shown). One explanation is that larger blocks cannot tell minor displacements of features and hence tolerate more structural heterogeneity. Although no theoretical model is available to prescribe an optimal value, within the 0.9194–1.4142 *R* range, we empirically chose $r = R$ throughout the experiments described in this paper.

The classifier was not sensitive to changes in γ. However, the SVM prediction stringency behaved differently toward fluctuations in *C*, as was supposed for such a non-linear SVM classification. Figure 4 shows the dependence of the number of hit points (but not positive clusters) on the value of *C* for an *E.coli* pyrophosphorylase (1DY3). The hit points drop dramatically with decreasing *C*.

Fortunately, in our study, the optimal *C* values—the smallest *C* when the genuine site is still identified but false positives are minimized, listed in Table III—show a clear tendency to hold for different proteins. For the one protein that we tested, we did notice that interestingly yeast fits the prokaryotic *C* value. It is reasonable to hypothesize that for ATP-binding sites, using these two suggested values is very likely to yield valuable binding site candidates.

Limitations on applicable ligand types

The shell-block division of site-proximal space that we described assumes that the ligand has a complicated 3D architecture. The ATP molecule, however, is almost planar with a rod-like triphosphate tail. Conceivably, a large proportion of

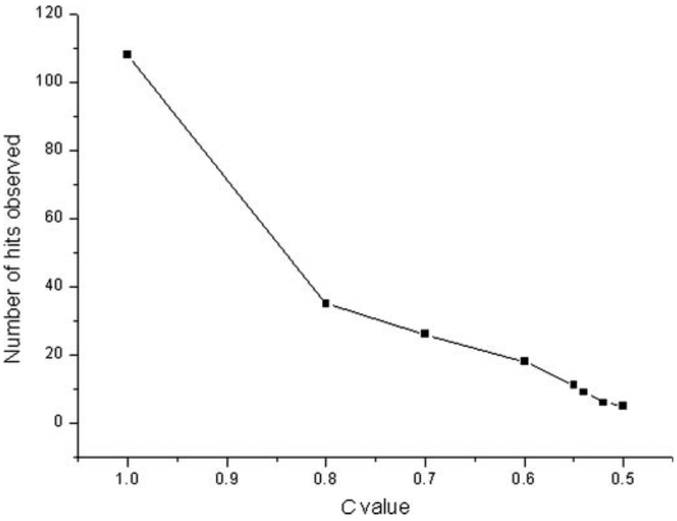

Fig. 4. Influence of *C* value on the number of hits observed in 1DY3 whole-protein search. *C* acts as a stringency-controlling factor in the prediction system. With very large *C*, a large number of false positives occur. When *C* is lowered to a certain level, usually even the true sites disappear from the prediction results. However, the true sites are almost always the last ones to disappear and the optimal value for *C* consistently parallels the domain origin of organisms, namely whether eukaryotic or prokaryotic.

the blocks would be ‘wasteful’ in terms of information content. Future improvements could be directed towards the creation of a ‘shape template’ for each different type of ligand where important blocks (i.e. those close to the ligand backbone) in the three-dimensional shell-block system are earmarked, whereas the others are set aside from consideration.

Another limitation is that not all binding sites are large and asymmetric enough to make the division into shell-blocks meaningful. For instance, when it comes to sites that recognize ions, e.g. a Ca^{2+}-binding site or very small molecules such as oxygen or CO_2, the prediction system is not expected to perform well.

Prospects for functional screening of structure libraries

We aimed to develop a technique that can initially screen raw structural data to give some idea of protein function. Our method outperformed other previous statistical models of this type, yielding higher accuracy and precision in whole-protein scanning tests. In some eukaryotic ATP-binding proteins, the classifier is almost capable of pinpointing the binding site and in prokaryotes only a small number of false positives appear.

Moreover, owing to the underlying physicochemical principle of this procedure, a reported site probably possesses a molecular microenvironment similar to that of a true functional site. A false positive, therefore, could be a potential target of cross-reactivity or toxicity that can be screened or verified by experimentation.

Acknowledgements

We thank Xuefeng Xia, Hu Chen and Wei Li for valuable discussions. We also acknowledge Shijun Shen for producing one of the illustrations. This project was supported in part by Grant 863 (2002AA231031, 2002AA234041) and by Grant 973 (2003CB715900) and NSFC Grants (90303017).

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Bayliss,R., Sardon,T., Vernos,I. and Conti,E. (2003) *Mol. Cell*, **12**, 851–862.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. and Swaminathan,S. (1999) *Nat. Genet.*, **23**, 151–157.

Campbell,S.J., Gold,N.D., Jackson,R.M. and Westhead,D.R. (2003) *Curr. Opin. Struct. Biol.*, **13**, 389–395.

Chang,C. and Lin,C. (2001) *LIBSVM: a Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Devos,D. and Valencia,A. (2000) *Proteins*, **41**, 98–107.

Di Gennaro,J.A., Siew,N., Hoffman,B.T., Zhang,L., Skolnick,J., Neilson,L.I. and Fetrow,J.S. (2001) *J. Struct. Biol.*, **134**, 232–245.

Friedberg,I. and Margalit,H. (2002) *Protein Sci.*, **11**, 350–360.

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

Gutteridge,A. and Thornton,J. (2004) *FEBS Lett.*, **567**, 67–73.

Gutteridge,A., Bartlett,G.J. and Thornton,J.M. (2003) *J. Mol. Biol.*, **330**, 719–734.

Hopfner,K.P., Karcher,A., Shin,D.S., Craig,L., Arthur,L.M., Carney,J.P. and Tainer,J.A. (2000) *Cell*, **101**, 789–800.

Iancu,C.V., Borza,T., Fromm,H.J. and Honzatko,R.B. (2002) *J. Biol. Chem.*, **277**, 26779–26787.

Jackson,R.M. (2002) *J. Comput.-Aided Mol. Des.*, **16**, 43–57.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Kellogg,G.E., Semus,S.F. and Abraham,D.J. (1991) *J. Comput.-Aided Mol. Des.*, **5**, 545–552.

Kinoshita,K. and Nakamura,H. (2003) *Curr. Opin. Struct. Biol.*, **13**, 396–400.

Kunin,V., Chan,B., Sitbon,E., Lithwick,G. and Pietrokovski,S. (2001) *J. Mol. Biol.*, **307**, 939–949.

Lake,M.W., Temple,C.A., Rajagopalan,K.V. and Schindelin,H. (2000) *J. Biol. Chem.*, **275**, 40211–40217.

Lichtarge,O. and Sowa,M.E. (2002) *Curr. Opin. Struct. Biol.*, **12**, 21–27.

Pang,Y.P., Perola,E., Xu,K. and Prendergast,F.G. (2001) *J. Comput. Chem.*, **22**, 1750–1771.

Pitt,W.R. and Goodfellow,J.M. (1991) *Protein Eng.*, **4**, 531–537.

Prodromou,C., Roe,S.M., O'Brien,R., Ladbury,J.E., Piper,P.W. and Pearl,L.H. (1997) *Cell*, **90**, 65–75.

Pupko,T., Bell,R.E., Mayrose,I., Glaser,F. and Ben-Tal,N. (2002) *Bioinformatics*, **18**, S71–S77.

Rantanen,V.V., Denessiouk,K.A., Gyllenberg,M., Koski,T. and Johnson,M.S. (2001) *J. Mol. Biol.*, **313**, 197–214.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.

Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) *Nat. Struct. Biol.*, **8**, 559–567.

Wade,R.C., Clark,K.J. and Goodford,P.J. (1993a) *J. Med. Chem.*, **36**, 140–147.

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

Wei,L. and Altman,R.B. (2003) *J. Bioinf. Comput. Biol.*, **1**, 119–138.

**Received October 6, 2004; revised January 26, 2005;
accepted February 4, 2005**

Edited by Valerie Daggett