Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A simple and efficient method for predicting protein-protein interaction sites

R.H. Higa[1,2] and C.L. Tozzi[1]

[1]Faculdade de Engenharia Elétrica e de Computação,
Universidade Estadual de Campinas, Campinas, SP, Brasil
[2]Centro Nacional de Pesquisa em Informática Agropecuária,
Empresa Brasileira de Pesquisa Agropecuária, Campinas, SP, Brasil

Corresponding author: R.H. Higa
E-mail: roberto@cnptia.embrapa.br

Genet. Mol. Res. 7 (3): 898-909 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 23, 2008

ABSTRACT. Computational methods for predicting protein-protein interaction sites based on structural data are characterized by an accuracy between 70 and 80%. Some experimental studies indicate that only a fraction of the residues, forming clusters in the center of the interaction site, are energetically important for binding. In addition, the analysis of amino acid composition has shown that residues located in the center of the interaction site can be better discriminated from the residues in other parts of the protein surface. In the present study, we implement a simple method to predict interaction site residues exploiting this fact and show that it achieves a very competitive performance compared to other methods using the same dataset and criteria for performance evaluation (success rate of 82.1%).

Key words: Interaction site prediction; Binding site prediction; Binding sites; Interaction sites; Protein structure

INTRODUCTION

Motivation

Protein-protein interactions are important for many biological processes, playing a key role in living cells. Information about them improves our understanding about these processes and may support the development of new technologies, for instance, knowledge about a disease process and the development of a corresponding new therapeutic approach. This information encompasses not only the identification of the interactions among proteins, but also the molecular recognition process that underlies them.

However, even considering only a fraction of the whole proteome of an organism, the number of protein-protein interactions involved is so high that experimental methods to find and analyze them all are impractical. In this context, there is a need for computational methods for predicting different aspects of protein-protein interactions.

In this paper, we consider the particular problem of the prediction of protein-protein interaction sites (or interface regions) from a three-dimensional structure. In this case, information about interactions at the atomic level, obtained from experimental methods such as X-ray crystallography and nuclear magnetic resonance spectroscopy, is used to predict interaction sites for proteins where there are no data revealing the structural details about how a protein interacts with its partner protein. Such a computational method could be used to identify potential interaction sites for both experimental verification and docking algorithms as well as to assist in the identification of potential function for the increasing number of protein structures having unknown function deposited in the Protein Data Bank - PDB (Berman et al., 2000).

In trying to reveal those properties, which discriminate between interaction sites and the rest of protein surface, different studies have analyzed chemical and structural properties (Chothia and Janin, 1975; Jones and Thornton, 1997a; Larsen et al., 1998; Madabushi et al., 2002). In their study, Jones and Thornton (1997b) suggested that no isolated property is capable of predicting interaction sites. Therefore, different studies have attempted to predict interaction sites by using a combination of chemical and structural parameters (Jones and Thornton, 1997b; Neuvirth et al., 2004; Bradford and Westhead, 2005; Bradford et al., 2006). Others, including Zhou and Shan (2001), Fariselli et al. (2002), Koike and Takagi (2004), Chen and Zhou (2005), and Bordner and Abagyan (2005) have relied on evolutionary conservation information, using multiple sequence alignment profile, sometimes in association with solvent accessible surface (SAS) and a residue conservation score, to identify residues belonging to interaction sites. All these methods are characterized by a level of accuracy between 70 and 80%.

On the other hand, Lo Conte et al. (1999), Chakrabarti and Janin (2002) and Bahadur et al. (2003) showed that the interaction site can be modeled as a core interface region surrounded by a rim interface region. A residue in the interaction site is considered as belonging to the core interface region if it is accessible to solvent in the unbound state but has zero accessible surface area in the complexed state. Otherwise, a residue in the interaction site that still has some accessible area in the complexed state is considered as belonging to the rim region. In addition, when comparing the amino acid compositions of these regions with those of the protein interior (hydrophobic core), it was found that the composition of the core interface region was closer to that of the protein interior while the composition of the rim region was closer to the protein surface composition.

In this study, we exploit this core and rim interaction site model, which suggests that residues in core interface region are more dissimilar to the rest of the protein surface than the rim interface region, in order to predict protein-protein interaction sites based on structural information. A predictor is proposed and using a previously validated dataset we show that the proposed method achieves a quite competitive level of accuracy compared to other methods using the same dataset and performance criteria.

General description of the proposed method for prediction of protein-protein interaction sites

The proposed method for predicting protein-protein interaction sites consists of a two-stage process. In the first stage, the core interface residues are detected by a Bayesian classifier, assuming that core interface residues can be more easily discriminated from the rest of the surface residues. A reject option is used such that residues close to the decision boundary are temporarialy assigned to the rejection class. The final decision about these residues is postponed until the next stage, when additional information is considered.

The idea in the second stage is to extend the set of interface residues detected in the first stage by using a less restrictive decision rule. Assuming that the residues detected in the first stage are clustered in spatial groups, the set of rejected residues in the first stage is re-evaluated through an empiric process, which considers as restriction basically the distance of the candidate residue to the closest spatial group of core interface residues.

Residues having SAS greather than zero (SAS >0) are considered exposed at protein surface. Each of them is labeled as either interface or non-interface, according to the difference between its SAS value calculated in unbound and bound states ($\Delta SAS = SAS_{unbound} - SAS_{bound}$). If ΔSAS is higher than zero, the residue is labeled as interface residue, otherwise as non-interface residue. In addition, interface residues having $SAS_{bound} = 0$ are also labeled as core interface residues, otherwise as rim interface residues. Notice that although our aim is to detect interface residues, the Bayesian classifier in the first stage is adjusted to discriminate only between core interface residues and non-interface residues.

A set of 28 chemical and structural properties (Tablel 1) is computed in order to compose the feature vector used by the Bayesian classifier. It consists of a 56-dimensional vector, where 28 elements correspond to the set of properties computed on the residue and 28 to the average of the same set of properties computed on the residue neighborhood. All them but amino acid type are normalized per protein by calculating the corresponding z-score.

MATERIAL AND METHODS

Dataset

For training and testing our prediction method, we used an already validated dataset (Bradford and Westhead, 2005). Its aim is to be representative of the diversity of types of protein-protein interactions, including both obligatory and non-obligatory interactions. In an obligatory protein-protein interaction, the protomers are not found as stable structures on their own *in vivo* while in non-obligatory ones the protomers exist independently (Nooren and Thornton, 2003). Even though there are larger datasets built by automated processes (Keskin et al., 2004), the

former presents the advantage that each complex was manually checked in the literature for evidence that it occurs *in vivo* (Bradford and Westhead, 2005). From the original 180 proteins, we removed eight for which we could not measure residue conservation, and therefore, the dataset we effectively used contained 172 non-redundant protein structures. It consisted of 37,758 solvent-exposed residues, from which 8642 (22.8%) were labeled as interface residues while the other 29,116 (77.2%) residues were labeled as non-interface residues. In addition, from the total interface residues, 5103 were core interface residues and 3539 were rim interface residues.

Usually, each method in the literature presents results based on its own dataset, which makes it difficult to compare them. Our choice allowed us to not only save the time of building our own dataset but also, as a secondary benefit, to compare directly our results to those reported by Bradford and Westhead (2005) and Bradford et al. (2006).

Table 1. List of properties measured over the protein surface.

List of properties		
Amino acid type	1	Aaindex 1
	2	Aaindex 2
Solvent accessibility	3	Solvent accessible surface (SAS)
	4	Relative SAS
	5	Molecular surface
	6	Residue depth - Cα
	7	Residue depth - average
	8	Half-sphere exposure - up
	9	Half-sphere exposure - down
Packing	10	Coordination number
Solvation energy	11	According to Eisenberg and McLachlan
	12	According to Wesson and Eisenberg 1
	13	According to Wesson and Eisenberg 2
	14	Solvation energy by area - Eisenberg and McLachlan
	15	Solvation energy by area - Wesson and Eisenberg 1
	16	Solvation energy by area - Wesson and Eisenberg 2
Geometry	17	Principal curvature - minor
	18	Principal curvature - major
	19	Gaussian curvature
	20	Mean curvature
	21	Shape index
	22	Curvedness
	23	Index of planarity
Electrostatic	24	Electrostatic potential
	25	Electrostatic solvation energy
Residue conservation	26	Relative entropy
	27	Information content
	28	Evolutionary pressure

Bayesian classifier with reject option

A Bayesian classifier relies on a probabilistic framework to assign labels to objects. Using a feature vector derived from a set of measurements over the object and a set of specified cost values involved in assigning a wrong label to an object, it tries to minimize the expected cost of classification (Duda et al., 2001).

Given a set of K classes $\Omega = \{\omega_1,..., \omega_k\}$, an n-dimensional feature vector, x, and a matrix, $C(\omega_i, \omega_j)$, expressing the cost involved in assigning an object to class ω_i when its true class is ω_j, the expected cost of assigning an object to class ω_i using x, $R(\omega_i|x)$, is given by

$$R(\omega_i|\mathbf{x}) = E\{C(\omega_i, \omega_k)|\mathbf{x}\} = \sum_{k=1}^{K} C(\omega_i, \omega_k)P(\omega_k|\mathbf{x}) \qquad \text{(Equation 1)}$$

where $P(\omega_k|x)$ is the probability of assigning an object to class ω_k using the feature vector x. Thus, the Bayes rule can be interpreted as: assign an object characterized by x to class ω_i if $R(\omega_i|x) \leq R(\omega_j|x)$ for all i,j = 1,...,K.

In order to consider the reject option and use the same Bayes rule for classification, we may introduce a new class to Ω, namely the reject class ω_0, such that the set of K + 1 classes is given by $\Omega^* = \{\omega_0,..., \omega_k\}$. Using the reject option, an object is assigned to class ω_i, $i \neq 0$, only if the associated expected cost is lower than a specific threshold. Otherwise, the object is assigned to the reject class, ω_0.

The second stage

The second stage is aimed at extending the set of interface residues detected in the first stage by using a less restrictive decision rule. It is assumed that the residues detected in the first stage belong to the core interface region, and as a consequence, they cluster in spatial groups. This way, this stage consists of an empirical process, which re-evaluates those residues assigned to reject class in the first stage by considering as restriction basically their distances to the closest spatial group of core interface residues. The idea is to capture additional core interface residues as well as rim interface residues close to the core interface region.

To identify groups of core interface residues, we use the k-means algorithm (Everitt et al., 2001), considering the corresponding xyz coordinates with the cluster centroids randomly initialized. To figure out the optimal number of clusters, we use the average silhouette value as criteria. Previously, we had determined that the optimal number of clusters for most of the proteins varies from 1 to 3. Therefore, we evaluated up to 3 clusters with the average silhouette value, considering the number of clusters corresponding to the minimal average silhouette value.

When evaluating the rejected residues, a residue is considered close to a cluster if the distance between its xyz coordinates and the cluster centroid is less than a threshold. To specify this threshold, the interaction site area for the test protein is estimated from its total surface area by a linear regression using the training dataset. The threshold corresponds to 47% of the radius of the circular area corresponding to the estimated interaction site area. This quite conservative threshold value is similar to the value used by Bradford and Westhead (2005), Bradford et al. (2006) and Jones and Thornton (1997b) to estimate patch size. Finally, to be accepted as an interface residue, a close residue needs to have at least two core residues among its neighbors.

Performance evaluation

The overall prediction performance is evaluated through the success rate, estimated by a leave-one(protein)-out cross-validation method. By this form of validation, we mean that when testing a protein, its residues are considered as testing data and those of all other proteins

are considered as training data. The success rate is defined as the percentage of successful predictions over the entire dataset. A prediction is defined as successful if at least one predicted cluster of residues shows coverage ≥20% and precision ≥50% (Bradford and Westhead, 2005), according to the following equations:

$$\begin{aligned} precision &= \frac{TP}{TP+FP} \\ coverage &= \frac{TP}{TP+FN} \end{aligned} \qquad \text{(Equation 2)}$$

where TP corresponds to the number of interface residues correctly classified, FN corresponds to the number of interface residues incorrectly classified and FP corresponds to the number of non-interface residues incorrectly classified. Using this criterion, our results can be directly compared to those reported by Bradford and Westhead (2005) and Bradford et al. (2006).

Since the random initialization of the centroids in k-means algorithm, used in the second stage, may introduce some variation in the prediction, we reported the average performance after repeating the leave-one(protein)-out cross-validation process 10 times.

Implementation

All procedures concerning classification were implemented using the Matlab environment 7.0 (Mathworks, 2004), the pattern recognition toolbox PRTools (Duin et al., 2004) and the statistics toolbox for multivariate analysis.

For estimating the parameters of the Bayesian classifier in the first stage, a Gaussian distribution with a common covariance matrix was considered, such that we obtained a minimum Mahalanobis distance classifier. For the cost matrix, the following values were considered: the cost of assigning a residue to its correct class is equal to zero, the cost of assigning a core interface residue to the non-interface class is equal to 2, the cost of assigning a non-interface residue to the core interface class is equal to 3, and the cost of assigning a residue to the reject class is equal to 1. All these values were determined empirically.

The 56-dimensional feature vector used by the Bayesian classifier encompasses 28 chemical and structural properties corresponding to the residue's set of properties and 28 to the average of the same set of properties computed for its neighborhood, where the neighborhood is formed by a set of residues (neighbors) obtained from the Alpha Shapes data structure (Liang et al., 1998), used to calculate SAS and molecular surface (MS).

The following 28 properties (see Table 1) were computed in order to form the feature vector:

• To represent the 20 standard amino acid types, we used two indexes (Hagerty et al., 1999) derived from the Aaindex database (Kidera et al., 1985). These two indexes summarize a collection of more than 400 indexes describing each of the 20 standard amino acids. In particular, the two indexes that we used are strongly correlated to residue size and hydrophobicity on one hand and residue preference for being in a loop or strand on the other (Hagerty et al., 1999).

• To compute SAS and MS, we used the "volbl" program, included in the package Alpha Shapes software (Liang et al., 1998), considering a probe radius of 1.4 Å and the set of atom radii provided in the package. In order to avoid the possibility of discontinuity in the interaction site due to cavities introduced in the complex formation, as may happen when

arbitrary distance or ΔSAS threshold is used, the "outside fringe" option was used. In addition, relative SAS was calculated from the SAS by using the values of SAS for each residue in extended state (Ala-X-Ala), as reported by Ahmad et al. (2004).

• To compute residue depth (Chakravarty and Varadarajan, 1999), half-sphere exposure (HSE) (Hamelryck, 2005) and coordination number, we used the Bio.PDB Biopython toolkit (Hamelryck and Manderick, 2003). The residue depth associated with a residue is defined as the average of the atom depth associated with each of its atoms, where atom depth is defined as the shortest distance between the atom and the MS. HSE is defined using a sphere (radius = 13 Å) with its center in the Cα atom position. Two half-spheres are defined by a plane orthogonal to the line connecting Cα and Cβ atoms, namely the side-chain half-sphere and the main-chain half-sphere. HSE is defined by two numbers, the number of Cα atoms inside the side-chain half-sphere, called HSE up (HSEu), and the number of Cα atoms inside the main-chain half-sphere, called HSE down (HSEd). The packing parameter coordination number is computed as HSEu + HSEd.

• Solvation energy per atom, in cal/mol · $Å^2$, was computed by multiplying the atomic solvation parameter (ASP) (Eisenberg and McLachlan, 1986) and its respective SAS. Three different sets of ASPs were considered (Eisenberg and McLachlan, 1986; Wesson and Eisenberg, 1992). To calculate solvation energy per residue, additive contribution was assumed such that it was calculated by adding the solvation energy of the corresponding set of atoms. In addition, for each set of ASP considered, we calculated the solvation energy per area.

• Principal, Gaussian and mean curvatures were calculated per surface atom using an osculating quadric, as reported by McIvor and Valkenburg (1997), and considering the set of atoms in its neighborhood. From principal curvatures, we also computed the parameters curvedness and shape index, as proposed by Koenderink (1990). Finally, from the set of atoms in the neighborhood, we computed the index of planarity, defined as the reciprocal of the root mean square deviation of all atoms in the neighborhood relative to the least squares plane through them (Jones and Thornton, 1997a).

• Electrostatic potential, in kT/e, and electrostatic contribution for the solvation energy, in kJ/mol, were computed by using the APBS software (Baker et al., 2001) with hydrogen atoms added to the protein structure by using the pdb2pqr.py software (Dolinsky et al., 2004). It was also used to calculate most of the parameters used to run APBS, including the grid size optimized for each structure. A dielectric constant of 78.54 was used for the solvent and of 2.0 for the protein interior, with 150 mM as ion strength. Electrostatic potential per residue was calculated as the average of the electrostatic potential value corresponding to the grid dots closest to each of its surface atoms. Similarly, electrostatic solvation energy per residue was obtained by adding the corresponding atomic electrostatic solvation energy given by APBS.

• For measuring residue conservation degree, first we used the Blast software (Altschul et al., 1997), with substitution matrix BLOSUM62 and expect value = 0.1, against Swissprot/Uniprot knowledge base release 9.6 (Apweiler et al., 2004) in order to find similar protein sequences. Next, sequences in the blast result were filtered according to homology derived secondary structure of protein (HSSP) threshold (Rost, 1999) to keep only homologue sequences. Eight protein sequences in the original dataset (Bradford and Westhead, 2005) did not survive this filtering process (at least 5 homologue sequences). Consequently, we considered only 172 proteins in our experiment. Afterward, we used the ClustalW software (Higgins et al., 1994), with substitution matrix series BLOSUM, gapopen = 3.0 and gapext = 0.1, using

the resulting set of homologue sequences to build an optimized multiple sequence alignment. Finally, three parameters reporting the degree of conservation for each residue were calculated, the information content, as implemented in the Bio.PDB Biopython toolkit, the relative entropy, as defined in HSSP, and evolutionary pressure, using the Rate4Site software (Pupko et al., 2002).

RESULTS AND DISCUSSION

The proposed method reports the predicted binding sites as a set of residues, grouped in 1, 2 or 3 clusters. A prediction is considered to be successful if at least one of the predicted clusters shows coverage ≥20% and precision ≥50%. Therefore, using a leave-one(protein)-out cross-validation process to estimate the overall performance of the method, we found a success rate of 82.1%. It also displayed an associated standard deviation of ±0.25%, which indicates a very robust result.

In addition, by ranking the predicted clusters according to the number of core interface residues detected in the first phase, we found that 67.5% of the successful instances, correspond to the top-ranked cluster. On the other hand, only 4.2% of the successful instances corresponded to the third-ranked cluster. That means that better-ranked clusters show a higher probability of corresponding to a successful prediction. Table 2 presents detailed results, considering the different types of interactions represented in the dataset.

Table 2. Detailed performance evaluation.

Interaction type	No. of examples	No. of successes	Rank		
			1	2	3
Non-obligatory					
Enzyme inhibitors	31	23	17	6	-
NEIT	29	24	16	6	2
Subtotal	60	47	33	12	2
Obligatory					
Homodimers	85	72	46	22	4
Heterodimers	27	23	17	6	-
Subtotal	112	95	63	28	4
Total	172	142	96	40	6

Numbers correspond to one among the leave-one(protein)-out cross-validation. NEIT = non-enzyme-inhibitor-transient. Predicted clusters of residues are ranked according to the number of predicted core interface residues they contain. The column Rank indicates the rank of the best-ranked cluster satisfying the condition of success of the prediction.

As expected by the composition of the dataset, the prediction of obligatory interaction sites showed a performance (84.8%) slightly better than that of non-obligatory interactions (78.3%).

Since our study is based on the dataset compiled by Bradford and Westhead (2005), it is convenient to discuss the similarities and differences between them in terms of methodology and results. Both methods rely on a set of structural and chemical parameters in order to predict interaction sites and report the result as a set of residues. Although there are differences in the set of parameters considered by each method, the main difference concerns the object described by the parameters. Bradford and Westhead (2005) follow a patch analysis method. Thus, they project all properties over

points distributed on the molecular surface, sample overlapping circular patches over this surface and calculate the average and standard deviation for all properties considering all surface points inside the patch. Afterward, the patches are classified as interaction sites or surface patches.

Our method, on the other hand, considers the surface residues as the basic units for prediction. Parameters are measured to characterize the residue and the neighborhood where it is located. Also, in the second stage of our method, the set of detected residues are divided into groups not necessarily displaying a circular shape. Furthermore, using the residue as the unit for prediction, our method can take advantage of the differences among residue properties inside the interface. These differences are reported in the literature as the core interface residues and rim interface residues. It also opens the possibility of exploring differences reported in the literature as hot spot residues (Bogan and Thorn, 1998).

Concerning pattern recognition techniques, our strategy was to keep them as simple as possible. This way, a classic Bayesian classifier with reject option is used in the first stage of our method. On the other hand, despite the fact that Bradford and Westhead (2005) use a theoretically more powerful method, support vector machine (Cristianini and Shawe-Taylor, 2000), the overall performance of their method (76%) was slightly lower than the one achieved by our method. Interestingly, in a later study Bradford et al. (2006) achieved a performance similar to ours (82%) using a simpler classifier, the naive Bayes classifier (Duda et al., 2001).

For illustrating the application of the method, two examples of prediction of protein-protein interaction sites are presented. The first example considers a subunit of the homodimer 3-hydroxy-3-methylglutaryl-CoA reductase (Figure 1), an enzyme involved in the biosynthesis of cholesterol in mammals (Tabernero et al., 1999). Two clusters are predicted for this example, one showing a coverage of 38.13% and precision of 79.1% and the other showing a coverage of 25.9% and a precision of 66.67%. Together, they cover most of the large interaction area of the molecule.

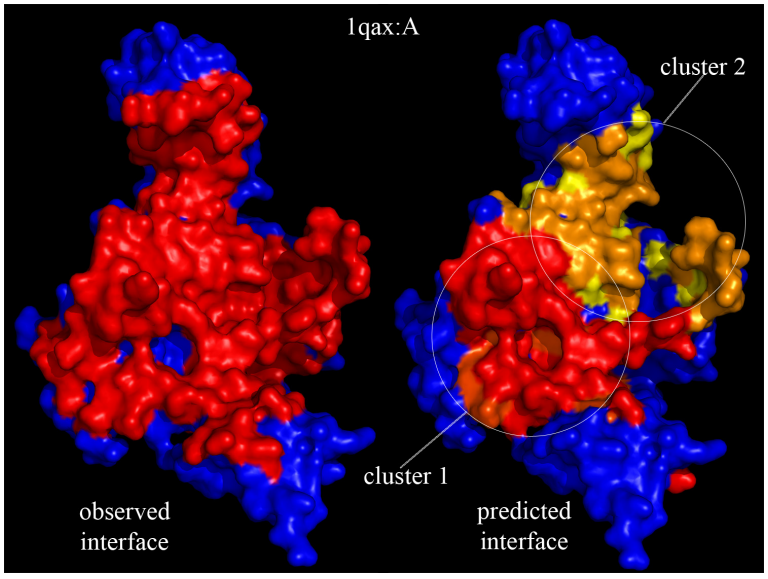

Figure 1. Observed (left) and predicted (right) interaction site for the homodimer 3-hydroxy-3-methylglutaryl-CoA reductase (PDB ID - 1qax:A).

The second example is barstar (Figure 2), the natural inhibitor of the RNase family member barnase, an extracellular enzyme of *Bacillus amyloliquefaciens* (Sevcik et al., 1998). In this case, three clusters are predicted but only the second ranked one satisfies the conditions to be considered as successful. It shows a coverage of 26.32% and a precision of 100%. In addition, the first-ranked cluster shows a coverage of 15.78% and a precision of 50% while the third-ranked one shows a coverage of 15.78% and a precision of 100%. However, if considered together, they cover most of the inhibitor interaction area. This example shows that even when the conditions to be considered successful are not satisfied, a predicted cluster of interface residues can provide a good indication of the interaction site as long as it is analyzed together with the other predicted clusters.

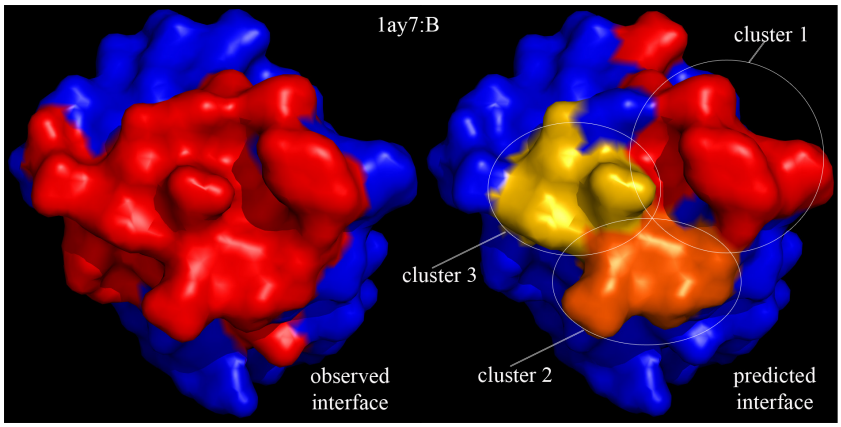

Figure 2. Observed (left) and predicted (right) interaction site for barstar (PDB ID - 1ay7:B).

CONCLUDING REMARKS AND FUTURE RESEARCH

A method to predict protein-protein interaction sites, consisting of two stages, was proposed. In the first stage, core interface residues are detected by using a Bayesian classifier with reject option. In the second stage, these detected core interface residues are then used to classify residues assigned to the reject class in the first stage. Despite its simplicity, the proposed method is successful in predicting interaction sites in 82.1% of the tested proteins, which is a very competitive performance (Bradford and Westhead, 2005; Bradford et al., 2006).

In the proposed method, the second stage only extends the region previously found. The interaction site is effectively identified by detecting core interface residues in the first stage. The results suggest that focusing only on the core interface residues could be a good strategy to identify interaction sites based on structural information. This is in agreement with experimental results reported in the literature showing that a large fraction of binding energy is associated with only a few key residues, called hot spots, which tend to form clusters in the center of the interaction site (Bogan and Thorn, 1998). In order to investigate similarities between the hot spot theory and our approach for classification focusing on core interface residues, we are working at the moment on extending the

feature vector with structural properties which were reported in the literature as useful in characterizing hot spot residues.

REFERENCES

Ahmad S, Gromiha M, Fawareh H and Sarai A (2004). ASAView: database and tool for solvent accessibility representation in proteins. *BMC Bioinformatics* 5: 51.

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Apweiler R, Bairoch A, Wu CH, Barker WC, et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32: D115-D119.

Bahadur RP, Chakrabarti P, Rodier F and Janin J (2003). Dissecting subunit interfaces in hodimeric proteins. *Proteins* 53: 708-719.

Baker NA, Sept D, Joseph S, Holst MJ, et al. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A*. 98: 10037-10041.

Berman HM, Westbrook J, Feng Z, Gilliland G, et al. (2000). The protein data bank. *Nucleic Acids Res.* 28: 235-242.

Bogan AA and Thorn KS (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* 280: 1-9.

Bordner AJ and Abagyan R (2005). Statistical analysis and prediction of protein-protein interfaces. *Proteins* 60: 353-366.

Bradford JR and Westhead DR (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* 21: 1487-1494.

Bradford JR, Needham CJ, Bulpitt AJ and Westhead DR (2006). Insights into protein-protein interfaces using a Bayesian network prediction method. *J. Mol. Biol.* 362: 365-386.

Chakrabarti P and Janin J (2002). Dissecting protein-protein recognition sites. *Proteins* 47: 334-343.

Chakravarty S and Varadarajan R (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* 7: 723-732.

Chen H and Zhou HX (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins* 61: 21-35.

Chothia C and Janin J (1975). Principles of protein-protein recognition. *Nature* 256: 705-708.

Cristianini N and Shawe-Taylor J (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge.

Dolinsky TJ, Nielsen JE, McCammon JA and Baker NA (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 32: W665-W667.

Duda RO, Hart PE and Stork DG (2001). Pattern Classification. 2nd edn. John Wiley and Sons, New York.

Duin RPW, Juszczak P, Paclik P, Pekalska E, et al. (2004). PRTools4, a Matlab toolbox for pattern recognition. Version 4.0. Delft University of Technology, Delft.

Eisenberg D and McLachlan AD (1986). Solvation energy in protein folding and binding. *Nature* 319: 199-203.

Everitt BS, Landau S and Leese M (2001). Cluster Analysis. 4th edn. Oxford University Press, New York.

Fariselli P, Pazos F, Valencia A and Casadio R (2002). Prediction of protein - protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* 269: 1356-1361.

Hagerty CG, Muchnik I, Kulikowski C and Kim SH (1999). Two indices can approximate four hundred and two amino acid properties. In: Proceedings IEEE Int. Simp. Intell. Control/Intell. Syst. and Semiotics, Cambridge, 365-369.

Hamelryck T (2005). An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins* 59: 38-48.

Hamelryck T and Manderick B (2003). PDB file parser and structure class implemented in Python. *Bioinformatics* 19: 2308-2310.

Higgins D, Thompson J, Gibson T, Thompson JD, et al. (1994). Clustal W: Improving the sensitivity of progressivemultiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Res.* 22: 4673-4680.

Jones S and Thornton JM (1997a). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* 272: 121-132.

Jones S and Thornton JM (1997b). Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* 272: 133-143.

Keskin O, Tsai CJ, Wolfson H and Nussinov R (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.* 13: 1043-1055.

Kidera A, Konishi Y, Ooi T and Scheraga HA (1985). Relation between sequence similarity and structural similarity in proteins. Role of important properties of amino acids. *J. Prot. Chem.* 4: 265-297.

Koenderink JJ (1990). Solid Shape. MIT Press, Cambridge.

Koike A and Takagi T (2004). Prediction of protein-protein interaction sites using support vector machines. *Protein Eng. Des. Sel.* 17: 165-173.

Larsen TA, Olson AJ and Goodsell DS (1998). Morphology of protein-protein interfaces. *Structure* 6: 421-427.

Liang J, Edelsbrunner H, Fu P, Sudhakar PV, et al. (1998). Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins* 33: 1-17.

Lo Conte L, Chothia C and Janin J (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* 285: 2177-2198.

Madabushi S, Yao H, Marsh M, Kristensen DM, et al. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* 316: 139-154.

Mathworks Inc. (2004). Matlab 7 - User's Guide. Mathworks Inc., Natick.

McIvor AM and Valkenburg RJ (1997). A comparison of local surface geometry estimation methods. *Mach. Vis. Appl.* 10: 17-26.

Neuvirth H, Raz R and Schreiber G (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* 338: 181-199.

Nooren IM and Thornton JM (2003). Diversity of protein-protein interactions. *EMBO J.* 22: 3486-3492.

Pupko T, Bell RE, Mayrose I, Glaser F, et al. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18 (Suppl 1): S71-S77.

Rost B (1999). Twilight zone of protein sequence alignments. *Protein Eng.* 12: 85-94.

Sevcik J, Urbanikova L, Dauter Z and Wilson KS (1998). Recognition of RNase Sa by the inhibitor barstar: structure of the complex at 1.7 A resolution. *Acta Crystallogr. D Biol. Crystallogr.* 54: 954-963.

Tabernero L, Bochar DA, Rodwell VW and Stauffacher CV (1999). Substrate-induced closure of the flap domain in the ternary complex structures provides insights into the mechanism of catalysis by 3-hydroxy-3-methylglutaryl-CoA reductase. *Proc. Natl. Acad. Sci. U. S. A.* 96: 7167-7171.

Wesson L and Eisenberg D (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* 1: 227-235.

Zhou HX and Shan Y (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* 44: 336-343.