

A New Algorithm to Predict the Active Sites Using Amino Acid Vectors and Biochemical features of Surface Patches

Sunshin Kim¹, Chung-Sei Rhee², Jung Do Choi³, Yong Je Chung⁴, Keun Ho Ryu⁵

^{1,5}Database/Bioinformatics Laboratory, Chungbuk National University, The Republic of KOREA

²Algorithm Laboratory, Chungbuk National University, The Republic of KOREA

Department of Biochemistry, Chungbuk National University

Email: csrhee@chungbuk.ac.kr

ABSTRACT: It is a very active research field to predict protein functions and active sites by protein surface patches. It is especially challengeable to predict the active sites by structural or biochemical characters. We, here, suggest a simple and efficient method, and at this time focus just on our strategy to solve the tough problems.

1 INTRODUCTION

To predict protein functions, it is reasonable to investigate and use the biophysical properties of protein surfaces. The reason is that it is in some cases very difficult to infer protein functional relationships from the features of sequences or structures[1, 2, 3, 4].

Schmitt *et al.*[5] developed a method to detect functional relationships among proteins independent of a sequence or fold homology, which based on the idea that protein function is intimately related to the recognition and subsequent response to the binding of a substrate or an endogenous ligand in a well-characterized binding pocket. Binkowski *et al.*[6] described an approach for inferring functional relationship of proteins by detecting sequence and spatial patterns of protein surfaces. The SURFACE database was built for functional annotation and comparison of protein surface patches[7].

It is also proper to take advantage of biophysical or biochemical features of protein surface patches for prediction of protein-protein interactions or binding sites. We can guess that it is most important to extract information about the features from protein surfaces since the actions of proteins directly happen in the parts. The chemical and structural properties of active sites have been analyzed extensively. Looking at the distribution of amino acid residues, it was found that polar and aromatic residues are more abundant in interfaces[8, 9]. Neuvirth *et al.*[10] investigated and analyzed the chemical and structural properties of binding sites based on assumption that binding sites differ from the rest of the protein, from which they developed an interface prediction algorithm. They used thirteen attributes to get the combination of the final score. But this brought the problem of computer power limitations. Moreover, since the most attributes are dependent one another, the score combination was made by fixing the most important attributes, which seem to be intuitively least dependent.

Carugo *et al.*[11] presented a method to predict whether two proteins interact, based on their surface patch comparison. They used a very simple and fast approach which gets patch shapes by the eigenanalysis of the matrix drawn from the distribution of atoms. This is very efficient to discriminate interacting proteins from non-interacting ones.

As we can see from the above facts, there are close relationships between protein functions and protein surfaces. It can especially be very useful for prediction of protein functions to identify the structural and chemical characters of active sites on protein surfaces. We have also recognized that the most influential factors are, to discriminate active sites from inactive sites, atoms or amino acids in structural character and several chemical features. The attributes are x, y, z, and biochemical features, which are independent each other. If there are the other serious factors for it, this factor and the above mentioned ones may probably be dependent each other. So we do not consider the dependent factors in our method since using naïve Bayes classifier.

Here, we suggest a new method to predict protein active sites based on amino acid vectors and chemical features of surface patches. This method is so simple and efficient since we use just several independent attributes and naïve Bayes classifier.

2 METHODOLOGY

2.1 Overview

We make detailed plans for predicting active sites. That is, we want to suggest a simple and efficient method with high probability which will have similar accuracy or better in the experimental results than the existing one. Our schemes comprise 6 steps on the whole. First, select proper protein datasets used here. Second, identify where interfaces(active sites) are located on the protein surfaces. Third, get amino acid vectors and chemical atoms on the binary of protein surfaces. Fourth, take the distributions of pseudoatoms and chemical atoms from training datasets. Fifth, classify test datasets by naïve Bayes classifier. Sixth, assess the results.

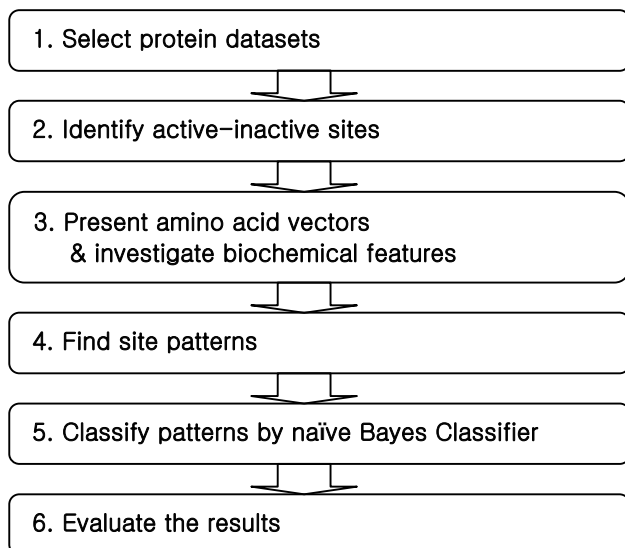


Figure 1: Flowchart for our method

2.2 Datasets of proteins

The first step in this method is to extract the datasets from the Protein Data Bank. In general, we can consider three separate datasets as enzyme-ligand binding, protein-protein dimerization, and antibody-antigen complexes[12]. The enzyme-ligand binding complexes will just be used in this work. The datasets are split into binary of interacting and noninteracting sites. The datasets consist of training and test sets. We will select the 62 proteins in the single-chain enzymes dataset from Protein Data Bank

2.3 Identification of active and inactive sites

To identify the active sites, the SURFNET[12] algorithm will be used. The largest and second largest clefts will be identified by the program. Whether the clefts are active or not can be recognized from Roman *et al.*'s[13]. By investigating the clefts, we can get the structural and chemical characters of the active or inactive sites.

2.4 Amino acid vectors & aromatic atoms

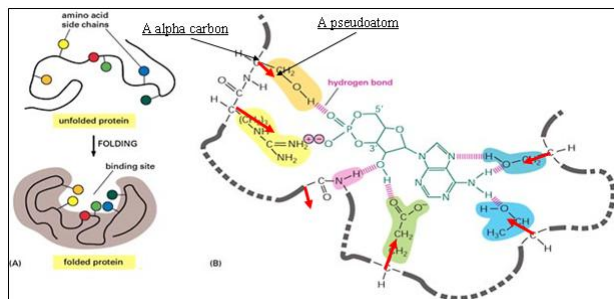


Figure 2: Amino acid vectors on the active sites
http://www.accessexcellence.org/RC/VL/GG/prot_Bindg.html

To represent the distribution of amino acids and the direction of atoms of amino acids, the amino acid vectors are introduced in joining the alpha carbons and the pseudoatoms calculated as the average coordinate of residue

side-chain atoms. Each arrow in Figure 2 represents arbitrary vectors we drew properly from alpha carbons to pseudoatoms to help understand what amino acid vectors are.

In contrast with the structural features, the chemical property of atoms is considered since it is a main point to discriminate binding sites from non-binding sites. We introduced the atoms with chemical features under assumption that the chemical feature of atoms are almost independent on each other as well as the amino acid vectors

2.5 Site patterns

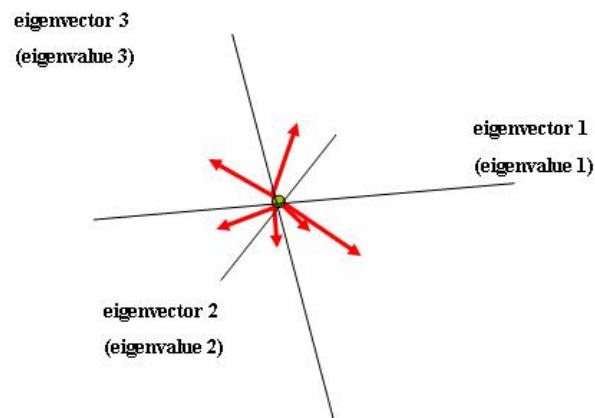


Figure 3: A site pattern description

To find the binary patterns of active-inactive sites, we use the eigenanalysis of the 3×3 matrix from amino acid vectors and get probability values. That is, we can get 6 amino acid vectors from active-inactive sites in Figure 2. The vectors have the x, y, and z components. So, we can get the 3×3 matrix $M = A^T A$ when given a 6×3 matrix A as Eqs. 1 and 2. The eigenanalysis of the M gives the three eigenvectors and eigenvalues. Since the eigenvectors are orthogonal each other, we can get three-independent probability values.

In Figure 3, we describe the amino acid vectors represented in Figure 2 along the principal axes. We can take the spread of the pseudoatoms along the direction of each eigenvector. We can also see the eigenvectors orthogonal to each plane defined by two eigenvectors. Distances of the pseudoatoms from the planes orthogonal to the three eigenvectors are computed and the distributions are determined.

We suppose that the chemical characters of atoms is, without mentioning the relations among them, not related with the structural features described above. That is, the probability values of the chemical atoms are each other independent of the values of structural amino acids as well as themselves. We finally have several independent probability values in the binary parts.

Now let's consider a simple example as Figure 4. The four independent features are chosen and each frequency of those is calculated from t proteins. We can get the sample mean and the standard deviation from each feature.

Eigenvector1	Active	Inactive	Eigenvector2	Active	Inactive
Protein 1	e_a11	e_i11	Protein 1	e_a21	e_i21
Protein 2	e_a12	e_i12	Protein 2	e_a22	e_i22
...
Protein t	e_at	e_it	Protein t	e_a2t	e_i2t
mean	e_a1p	e_i1p	mean	e_a2p	e_i2p
std dev	e_a1s	e_i1s	std dev	e_a2s	e_i2s

Eigenvector3	Active	Inactive	A biochemical feature	Active	Inactive
Protein 1	e_a31	e_i31	Protein 1	a_a21	a_i21
Protein 2	e_a32	e_i32	Protein 2	a_a22	a_i22
...
Protein t	e_at	e_it	Protein t	a_a2t	a_i2t
mean	e_a3p	e_i3p	mean	a_a2p	a_i2p
std dev	e_a3s	e_i3s	std dev	a_a2s	a_i2s

Figure 4: Frequency of four attributes on active or inactive sites of t proteins

2.6 Pattern classification

To classify a new protein part into interacting versus non-interacting sites, we use naïve Bayes classifier which is a simple but highly effective when the attributes are independent each other. When dealing with numerical attributes, we assume that attributes have a Normal or Gaussian probability distribution. The probability density function for the normal distribution is defined by two parameters. The class of a new instance is predicted as in Eqs. 4.

For instance, if we get, in Figure 4, the frequency of the Eigenvector1 of a new protein, the probability density function of the protein is represented as Eqs. 5. The class of active sites on the protein is predicted as Eqs. 6. and that of inactive sites predicted as Eqs. 7.

$$\text{The sample mean : } \mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

$$\text{The standard deviation: } \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2} \quad (2)$$

$$\text{The density function: } f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

$$\text{General equation: } P(C_l | X) = \frac{P(X | C_l)P(C_l)}{P(X)} \quad (4)$$

$$P(p_{i_E1} | \text{Active}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(p_{i_E1}-\mu)}{2\sigma^2}} \quad (5)$$

$$P(\text{Active} | p_i) = \text{Const.}P(\text{Active})P(p_{i_E1} | \text{Active})P(p_{i_E2} | \text{Active})P(p_{i_E3} | \text{Active})P(p_{i_A} | \text{Active}) \quad (6)$$

$$P(\text{Inactive} | p_i) = \text{Const.}P(\text{Inactive})P(p_{i_E1} | \text{Inactive})P(p_{i_E2} | \text{Inactive})P(p_{i_E3} | \text{Inactive})P(p_{i_A} | \text{Inactive}) \quad (7)$$

2.7 Assessment

To measure the quality of the prediction, we calculate sensitivity, specificity, and accuracy. The sensitivity, the specificity, and the accuracy range between 0 and 1. While the sensitivity and the specificity lean to be inversely proportional as in Eqs. 8 and 9, the accuracy checks the overall success rate as in Eqs. 10, where both true positives and true negatives are considered.

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (9)$$

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

3 CONCLUSION

As indicated in the relative works, protein functions and protein surfaces have close relationships. So it helps infer protein functional relations among proteins to investigate and identify the structural and biochemical features of protein surfaces. It can especially be very useful for prediction of protein functions to identify the structural and chemical characters of active sites on protein surfaces. But the prior method (by Neuvirth et al.) to identify the location of binding sites requires too much computing power without giving optimal solutions.

We, therefore, suggest a simple and fast method to solve the problem. First of all, we have, from prior work, recognized that the most effective elements are, to discriminate active sites from inactive sites, the distributions of amino acids and biochemical feature. The several independent attributes are drawn, which are x , y , z , and chemical features. We can use a very efficient method which finds out patch patterns by the eigenanalysis of the matrix drawn from the distribution of amino acid vectors. Next, the three distribution values due to eigenvectors are taken. In addition to that, the distribution of chemical atoms can be extracted by analyzing the chemical character on binary of protein surfaces (active versus inactive sites). Since the several attributes are independent each other, we can easily classify the test results into binary classes of active versus inactive sites using naïve Bayes classifier.

ACKNOWLEDGEMENT

This work was partially supported by the BIT research center of Chungbuk National University.

REFERENCES

- [1] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273: 595--603, 1996.
- [2] L. M. Kauvar and H. O. Villar. Deciphering cryptic similarities in protein binding sites. *Curr. Opin. Biotechnol.*, 9: 390--394, 1998.
- [3] A. Via, F. Ferre, B. Brannetti, A. Valencia and M. Helmer-Citterich. Three-dimensional view of the surface motif associated with the P-loop structure: *cis* and *trans* cases of convergent evolution. *J. Mol. Biol.*, 303: 455--465, 2000.
- [4] C. Wilson, J. Kreychman and M. Gerstein. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, 297:233--249, 2000.
- [5] S. Schmitt, D. Kuhn, and G. Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, 323: 387--406, 2002.
- [6] T. A. Binkowski, L. Adamian, and J. Liang. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J.Mol.Biol.*, 332: 505--526, 2003.
- [7] F. Ferrè, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich. SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids Res.*, 32: 240--244, 2004.
- [8] F. Glaser, D. M. Steinberg, I. A. & N. Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Struct. Funct. Genet.* 43: 89--102, 2001.
- [9] P. Chakrabarti, & J. Janin. Dissecting protein-protein recognition sites. *Proteins: Struct. Funct. Genet.* 47: 334--343, 2002.
- [10] H. Neuvirth, R. Raz, and G. Schreiber. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, 338: 181--199, 2004.
- [11] O. Carugo, and G. Franzot. Prediction of protein-protein interactions based on surface patch comparison. *Proteomics*, 4: 1727-1736, 2004.
- [12] R. A. Laskowski. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, 13: 323-330, 1995.
- [13] R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton. Protein clefts in molecular recognition and function. *Protein Science*, 5:2438--2452, 1996.