CARPATHIAN J. MATH.
20 (2004), No. 2, 261 - 266

Detection Method of Function Site of Proteins by Using a Graph-Theoretic Algorithm

SHIN-ICHI NAKAYAMA, YOSHIHIRO KAWASAKI, SHIHO MATSUDA, AKIKO KAMIYA, TETSUYA MAESHIRO and MASAYUKI YOSHIDA

ABSTRACT. We present a new detection method of amino acid residues, which play the role of the protein functional activities. The principle of the method is that proteins with the same function have similar amino acid residues at the similar 3D positions. Thus we assume that maximal amino acid residues existing in the similar 3D positions of proteins are their function site residues. The method first constructs the graph describing relations of 3D positions of all possible pairs of amino acid residues in two proteins and then extracts its maximal complete subgraph, which means the maximal amino acid residues at the similar three dimensional positions, by the algorithm of Carraghan and Pardalos.

The method was tested using electron transport proteins: azurin and plastocyanin, and 6 reasonable amino acid residues as the function site residues of those proteins, were obtained. The method was also applied to acid proteases: porcine pepsin and *Rhizopus* pepsin. In this investigation, we restricted the formation of amino acid pairs only to the same ones, because of the limitation of computer memory. The result gave 11 amino acid residues, consisting in some active site residues, as the function residues of acid proteases. The results indicate the effectiveness of this new method.

1. INTRODUCTION

The detection of active sites of proteins is of importance for protein studies. As the active sites have been determined by interaction studies between inhibitors and proteins[1] (Ringe, 1995), the method is seen time and costs expensive. To solve the problem, a number of approaches, using protein 3D databases, have been developed. These methods find the binding sites on the surface of proteins able to attach ligands (for example, GRID and MCSS)[2,3] (Goodford, 1985; Miranker and Karplus, 1991). Among these, LIGSITE automatically recognize pockets and/or cavities of protein surface.[4] (Hendlich *et al*., 1997). The procedure of Rosen *et al*.[5] (1998) search the active site by using the similarity of already decided active sites. Recently, Schmitt *et al*.[6] (2002) proposed a method to find active site amino acid residues using motif sequences, obtained from large protein sequence database, and folding similarity.

Those methods could not be applied to the proteins without active site pocket and/or original active site structures. Here, we propose an approach according to which, in two proteins with the same function, and having some amino acids residues in the same 3D positions with respect to the function place, the 3D overlap is merely possible. Thus, if the maximal 3D overlapped amino acid residues in two of the same function proteins were detected, we could find the amino acid residues constructing the function site. In this paper, we report some results by using this hypothesis.

Received: 26.09.2004; In revised form: 17.01.2004
2000 *Mathematics Subject Classification*. 92E10
Key words and phrases. *protein similarity, 3D overlap algorithm, protein functional activity*

2. DETECTION METHOD

We selected a method using graphs to find out quickly the 3D overlap of amino acid residues. The usefulness of the method is known in the studies of protein structure matching[7] (Grindley et al., 1993), protein docking problem[8] (Gardiner et al., 2000) and so on. In this study, the position of each amino acid residue was approximated by that of alpha carbon atom in the residue. The maximal 3D overlapped alpha carbon atoms of two proteins are extracted by using labeled graph. The nodes of the graph are possible pairs of alpha carbon atoms in two proteins, and the edges of the graph are the allowance distance between them. A maximal complete subgraph (clique) obtained from the constructed graph identifies the pairs of the maximal 3D overlapped alpha carbon atoms. We used the algorithm of Carraghan and Pardalos[9] (1990) for fast finding the clique, in previous studies.

3. EXPERIMENTAL DETAILS AND DISCUSSION

3.1 SYSTEM

For the following studies, we develop two systems. One is System-1, which finds maximal 3D overlapped amino acid residues of two protein chains considering all possible amino acid pairs, and the other is System-2, which makes a node only if two amino acid residues are same ones. As the protein chain is obtained from PDB file, two file names and those chain codes are given as system parameters. The edge of the graph is constructed if the difference between the distances for the pairs of alpha carbon atoms is below a user-defined threshold. The system requires the threshold value as a parameter.

3.2 EXTRACTION FROM ELECTRON TRANSFER PROTEINS

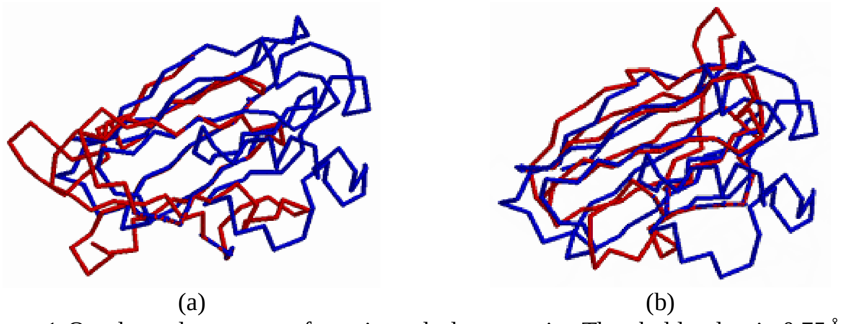
Azurin and plastocyanin have electron transfer function and have similar structures. To extract the function amino acid residues from the two proteins, the System-1 was applied because the number of amino acid residues was small; 128 and 99, respectively. The threshold was set at a value of 0.5, 0.75 or 1.0 Å. The obtaining number of nodes for each clique and the calculation times are shown in Table 1. As the threshold value became large, the number of nodes increases and the calculation time increases suddenly because of increase in number of edges.

Table 1. Results of electron transfer proteins

Threshold value(Å)	Number of nodes	Number of nodes for same amino acids	Calculation time*(hr)
0.5	13	1	7.5
0.75	17	1	23.5
1.0	20	6	78

* Pentium III Xeon 550MHz

Overlapped structures of two proteins by the cliques are shown in Figure 1. The style of overlap is different between cliques by small and large threshold values. The latter overlap seems good for consideration of Cu position. Table 2 shows position and amino acid type of the pairs. In that overlapping, 6 same amino acid residues are positioned in same 3D places.



(a) (b)

Figure 1 Overlapped structure of azurin and plastocyanin: Threshold value is 0.75Å (a) and 1.0 Å (b)

Table 2. Estimated active site amino acid residues (pairs of same amino acid residues in bold)

Azurin		Plastocyanin	
Amino acid position		Amino acid position	
N	**18**	**N**	**17**
D	23	G	22
F	29	K	73
G	45	P	36
H	**46**	**H**	**37**
N	**47**	**N**	**38**
L	86	Y	62
G	88	N	64
K	92	Q	68
M	109	G	81
F	110	V	82
T	113	D	85
A	119	G	91
K	**122**	**K**	**93**
G	123	M	94
T	**124**	**T**	**95**
L	125	I	96
T	**126**	**T**	**97**
K	128	Q	99

Those 6 residues on azurin are shown in Figure 2. As two residues (histidine-46 and asparagine-47) are placed near Cu atom, it seems that those residues have important role for electron transfer function. The remaining 4 residues place one side of the molecule and all of them are polar residues.

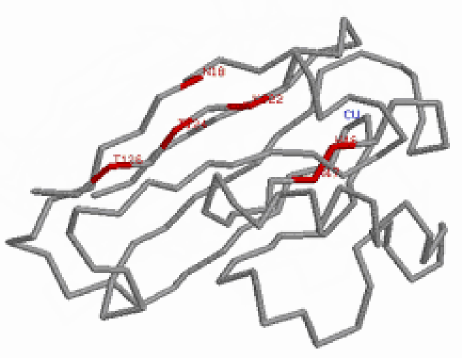


Figure 2. Estimated active site amino acids on azurin.

Azurin has a tetramer structure and the residues place a cross-shaped structure as shown in Figure 3.

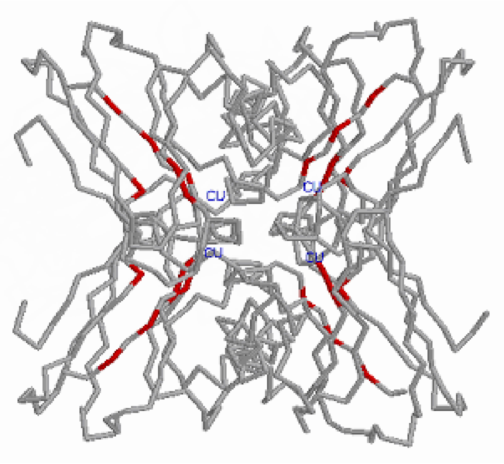


Figure 3. Tetramer structure of azurin.

3.3 Extraction from Acid Proteases.

Porcine pepsin and *Rhizopus* pepsin are same functional proteins from different origin. They have similar 3D structures and active sites. As the number of amino acid residues of these proteins is 327 and 325, inter-node number is more than 5 billions. Thus we apply System-2 to extract function amino acid residues from these two proteins. In this system, nodes are constructed only the pairs of same amino acid residues, thus inter-node number is reduced to about 14 millions.

First, we tried the threshold value of 1.0 A. The obtaining number of nodes for the clique is too much (60). Thus the threshold value is reduced to 0.5 and 0.2 A. The obtaining number of nodes and computing times are shown in Table 3. In the case of threshold value of 0.2 A, the obtaining nodes are 11 and computed only 1 hr. As the threshold value increased, the number of nodes and time for calculation increase, respectively.

Table 3 Results of acid proteases

Threshold value(Å)	Number of nodes	Calculation time*(hr)
0.2	11	1
0.5	30	2
1.0	60	41

*Pentium III 450MHz

Overlapped structures of two proteins by the cliques from the threshold value 0.2 Å are shown in Figure 4. The style of overlap looks reasonable. The positions and amino acid type of the pairs are shown in Table 4. The aspartate-215 was detected but aspartate-32 was not found although the next amino acid residue, glycine-33, was estimated. The 11 residues on porcine pepsin are shown in Figure 5.

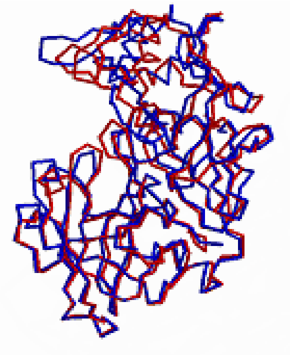


Figure 4. Overlapped structure of porcine pepsin and *Rhizopus* pepsin.

Table 4 Estimated active site amino acid residues of acid proteases.

No	Amino acid	Position of porcine pepsin	Position of *Rhizopus* pepsin
1	T	33	36
2	G	34	37
3	S	35	38
4	S	62	64
5	S	72	74
6	I	83	86
7	L	84	87
8	D	87	90
9	Q	99	102
10	L	123	125
11	D	215	218

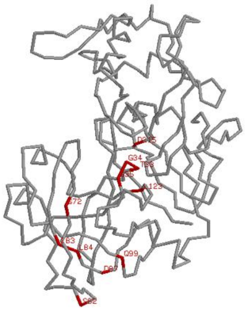


Figure 5. Estimated active site amino acids on porcine pepsin.

As 5 amino acid residues appeared in the bottom of active site cleft, our method correctly found the active site of acid proteases. The remaining 6 amino acid residues are placed on the same side of the leaf. It suggests that the leaf structure has important role for a ligand recognition.

ACKOWLEDGEMENTS. This work was supported in part by the funds of the University of Library and Information Science Research Projects.

References

[1] Ringe, D., Curr. Opin. Structural Biol. **5** (1995), 825-829.

[2] Goodford, P. J., J. Med. Chem. **28** (1985), 849-857.

[3] Miranker, A. and Karplus, M., Proteins **11** (1991), 29-34.

[4] Hendlich, M., Rippmann, F. and Barnickel, G., J. Mol. Graph. Model. **15** (1998), 359-363.

[5] Rosen, M., Lin, S. L., Wollfson, H. and Nussinov, R., Protein Eng. **11** (1998), 263-277.

[6] Schmitt, S., Kuhn, D. and Klebe, G., J. Mol. Biol. **323** (2002), 387-406.

[7] Grindley, H., Artymiuk, P. J., Rice, D. W. and Willett, P., J. Mol. Biol. **229** (1993), 707-721.

[8] Gardiner, E. J., Willett, P. and Artymiuk, P. J., J. Chem. Inf. Comput. Sci. **40** (2000), 273-279.

[9] Carraghan, R. and Pardalos, P. M., Operations Res. Lett. **9** (1990), 375-382.

Research Center for Knowledge Communities,
University of Tsukuba,
1-2 Kasuga, Tsukuba, Ibaraki, 305-8550,
Japan

University of Library and Information Science,
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550,
Japan
E-mail: nakayama@slis.tsukuba.ac.jp