Protein Function Prediction by Matching 3D Structural Data

Chien-Cheng Chen[†] Jeng-Ting Tu[†] Pei-Ken Chang[†]
Bing-Yu Chen[‡] Rung-Huei Liang[†] Ming Ouhyoung[*]
National Taiwan University
[†]{ccchen, magictu, zick, liang}@cmlab.csie.ntu.edu.tw, [‡]robin@ntu.edu.tw, [*]ming@csie.ntu.edu.tw

Abstract

Active sites determine the activities and interactions of proteins and constitute the targets of most drugs. However, the exponential growth in the amount of structural data of proteins far exceeds the ability of experimental techniques to identify the locations and key amino acids of active sites. Several approaches have been applied to this problem, including analyzing gene expression patterns, wavelet transform analysis of protein three-dimensional structures, statistical analysis of protein one-dimensional amino acid sequence, and protein-protein interactions. In this paper, we develop an approach that is mostly based on analysis of the geometric shape of a protein in space. Unlike statistics-based methods, our approach exploits the shape of a protein to find its functions and possible locations of active sites. Aiming at predicting functions of proteins, we collect three-dimensional structures of active sites, which have already been found by biochemists to create a database containing three-dimensional structures of active sites. Since the shape of an active site determines the function of a protein, for each protein of interest, we compare it with all the site structures in our site database and select a site structure, to which we can find the most similar structure in the protein. The function of the protein is then assigned to the one of the selected active site. Moreover, the location where the protein matches the selected site can be considered as a possible position of an active site. To show that our approach can find significant features in a protein, we use the proteins belonging to enzyme class five as our test set. The result shows the proposed algorithm can successfully predict correct functions of 39 out of 43 proteins and the accuracy is about 90.7%.

1. Introduction

The structural genomics initiative (SGI) proposes to solve 10,000 protein 3D structures in this decade, however, many biological functions still remain unknown.

Protein function is highly related to the recognition of specific substrates, ligands or co-factors in a specific region, that is, a binding site. Binding sites in proteins are where the substrates or ligands interact with proteins to trigger some events such as chemical modifications or conformational changes [1]. As a result, proteins exihibiting related functions are likely to share some similiarities in their 3D structures [2].

The development of new drugs is an extremely expensive and time-consuming undertaking. Usually it takes ten to twelve years from initial lead discovery to completion of clinical trials and costs around $800 million [4]. Consequently, structure-based drug design (SBDD) methods have become more prevalent in recent years due to their efficient and fast capability in sieving out the possible drug candidates among a group of chemical compunds [3, 4].

Exploitation of 3D structural data is a key factor for SBDD being enhanced, and the prediction of protein functions and possible binding sites in proteins have become quite popular in SBDD, especially at front-ends to molecular docking [5, 6, 7, 8, 9] or alternative binding sites are sought otherwise.

In this paper, we propose a protein-function- prediction method which is mainly based on the geometrical features of proteins. Proteins with similar three-dimensional structures often have related functions even if their one-dimensional amino acid sequences are not alike. If the conformation of a protein is destroyed, its function will also disappear. Hence we focus on the geometrical features that belong to proteins and if the active sites of a protein with unknown function are similar to those of another protein with known function, then it is possible that two proteins share the same function [1, 7, 10, 11]. The proposed method utilizes the observation that proteins with related functions are likely to share certain similarities in their three-dimensional structures. Consequently, the structures of proteins with unknown functions are compared with those of active sites of which functions are already found by biochemists. Using the geometric hashing algorithm, we can search out structural similarities between protein structures and we predict the function of a protein as the one of its corresponding best-matched active site structure.

This paper is organized as follows. Some related works are discussed in section 2. The geometric hashing algorithm we use are detailed in section 3 while the experimental results are provided in section 4. Conclusion and our future directions are given in section 5 and 6.

2. Previous Work

As substantial protein structures are determined by high throughput machines, assigning functions to those novel proteins becomes a major task in recent years. The Protein Data Bank (PDB) [12] currently contains more than 19,500 structures and it is estimated the number of structures in the PDB may exceed 35,000 by 2005 [13]. Much effort has been put into finding the motifs, functions, binding sites of proteins and many approaches based on different techniques are developed.

The classical way of finding motifs in proteins is to find homologies of their one-dimensional amino acid sequences among proteins in a protein database using programs such as FASTA and PSI-BLAST. Waterman et al., Delcoigne and Hansen, and Needleman and Wunsch [14, 15, 16] find sequence motifs by minimizing a cost function which represents the edit distances between sequences. Multiple alignment of sequences [17] is a NP-hard problem and its computational time increases exponentially with the sequence size.

Using amino acid sequence data, Bill and Saman [18] extract protein motifs from sequences belonging to the same family by neuro-fuzzy optimization which involves a statistical method to find short patterns with high frequency and then uses neural network training to optimize the final classification accuracy.

Based on signal processing techniques, Kevin B. Murray et al. [19] use wavelet transform to detect and characterize repeating motifs in protein sequences and structural data.

G. Patric, Jr. and Pieter F.W. Stouten [1] use a geometry-based approach, PASS, which fills the cavities in a protein structure with a set of spheres and to identify a few of these spheres that most likely represent the centers of binding pockets.

Herein, we propose a geometry-based approach which is able to predict the function of a protein and identify possible locations of the residues of active sites in a protein molecule.

3. Algorithms

Amino acids are the building blocks of a protein which plays an important role in many reactions in living things. There are twenty different types of amino acids in nature, each of which has a different property. An amino acid has four main parts: the acid group(COOH), the amino group(NH_3), the carbon, the side chain(R-group). Figure 1 explains the relationship among them.

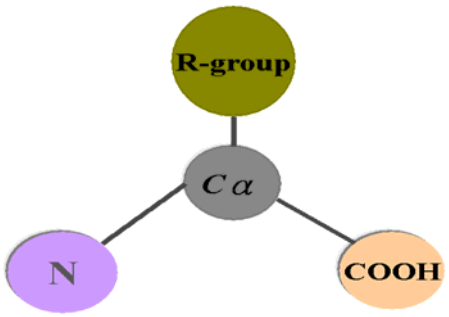


Figure 1. Amino acid structure

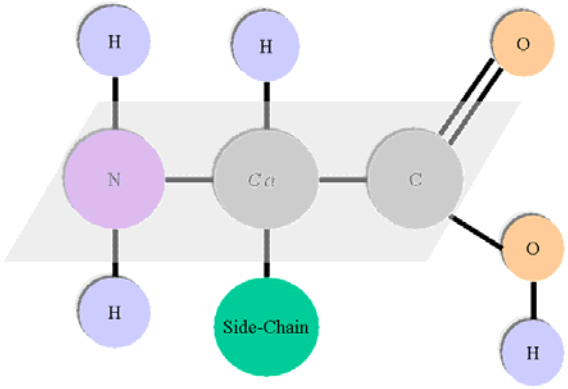


Figure 2. Peptide plane

When two amino acids link together, they use a peptide bond which is formed by a dehydration process. A protein is constructed by linking amino acids to form a sequence which folds in space to generate a complex three-dimensional structure. When binding, a ligand may induce structural changes in the receptor protein. Nevertheless, in most cases, changes in backbone structure are negligible and only side-chain reorientation occurs on ligand binding. Therefore we use the plane formed by the atom N, Cα, and C as our reference frame shown in Figure 2.

The three atoms N, Cα and C in each amino acid form a triangle which uniquely defines the position and orientation of the amino acid in the three- dimensional structure of a protein. Therefore, all the N, Cα and C atoms of a protein molecule together act as a backbone or skeleton to which the side chains R are attached. Since the length of Cα - N and Cα - C are fixed, and N – Cα - C bond angle is also changeless, the skeletons corresponding to two

common substructures of two proteins will be exactly congruent. The correspondence between two triplets of points in three-dimensional space is sufficient to uniquely determine a rigid transformation (which would take one triplet onto the other). Making use of this fact, the geometry of the atoms attached to the Cα is perfectly determined. In particular, the three atoms N, Cα, C form a known triangle from which we can define a frame. It can uniquely determine the position and orientation of a residue in space. With this mechanism, we can choose a single residue as a basis.

We now create a reference frame, that is, a basis, for each residue in a protein. A basis is calculated by the following steps and is illustrated in Figure 3.

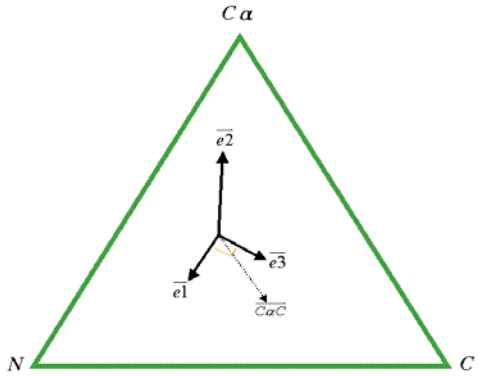


Figure 3. 3D reference frame

1. Normalize $\overrightarrow{C_\alpha N}$ to $\vec{e1}$ ..... (1)

2. Define $\vec{e2}$ as the following:

$$\vec{e2} = \frac{\vec{e1} \times \overrightarrow{C_\alpha C}}{\left|\vec{e1}\right| \cdot \left|\overrightarrow{C_\alpha C}\right|} \quad \text{..... (2)}$$

3. Calculate $\vec{e3}$ by the following equation similar to the way above.

$$\vec{e3} = \vec{e2} \times \vec{e1} \quad \text{..... (3)}$$

We calculate one basis for each residue, and then use the bases created to generate coordinates for each atom in a protein in the next step, geometric hashing algorithm.

Three-dimensional conformation of a protein in space usually dominates its function. Thereby, proteins with related functions usually share some similarities among their three-dimensional structures. According to the above description, common structures among proteins of the same family can be the representatives of possible binding pockets.

Herein we introduce geometric hashing algorithm [20] as our fundamental method to explore similarities among a set of proteins. The geometric hashing algorithm is a technique originally developed in computer vision for matching geometric features against a database of such features. In recognizing objects, geometric hashing is efficient and can easily be made parallel. Furthermore, it is not only especially attractive in model-based schemes, but also holds significant advantages in pair-wise object scene comparisons because of its ability to handle partially occluded objects.

To solve the problem that objects may appear based on different reference coordinates, coordinate information based on different reference frame of a model is encoded in the preprocessing step and stored in a large memory, in this case, a hash table. The contents of the hash table are independent of the scene and thus can be computed offline to reduce the time needed for recognition. Access to the memory is based on geometric information that is invariant of the object's pose and computed directly from the scene. During the recognition phase, the method accesses the previously constructed hash table using the indices of the encoded coordinate information of the input object and finds their common spatial features.

In short, the geometric hashing algorithm is composed of two stages: preprocessing and recognition. The basic idea is to store in a database at preprocessing time a redundant representation of the models by rigid transformation. By doing so, the representation of the query object processed at recognition time will present some similarities with that of some database models. Matching is possible even when the recognizable database objects have undergone transformations or when only partial information is present.

Common substructure matching of objects meets the problem of the transformation in their scale, that is, two same objects will be considered as different ones if they differ only in size. In our application, however, we won't encounter this problem in that the relative distance of each atom in a protein molecule won't change in nature. Hence it is quite suitable to use the geometric hashing algorithm in our algorithm.

We now expound our idea in finding common parts within proteins and a formal methodology of the geometric hashing algorithm will be given.

There are two steps, preprocessing and recognition, in the algorithm. Preprocessing step calculates new coordinates of atoms in a protein with respect to each basis mentioned previously and recognition step intends to find the structural similarities among a set of proteins. We now discuss each step respectively.

Preprocessing: Each residue in a protein can be regarded as a 3D reference frame. With each frame, we have three orthonormal vectors as the three new coordinate axes, and then calculate the new coordinate for each atom in the protein with respect to the new reference

frame. Based on the basis, the three-dimensional positions of all the residues are the features, which are inserted into the hashing table with an index. This step is performed without any knowledge of the database objects to be matched and hence can be done once for all.

Recognition: Choose a reference frame of the query protein. For each different reference frame of a protein in the hashing table, we accumulate the number of the three-dimensional features matched, which is called voting. The number of matched features will be the match score of these two frames. The process is repeated with each frame of the query protein until all the reference frames of these two proteins have been tested. We keep the match with the highest score. Figure 4 shows how we complete the recognition process.

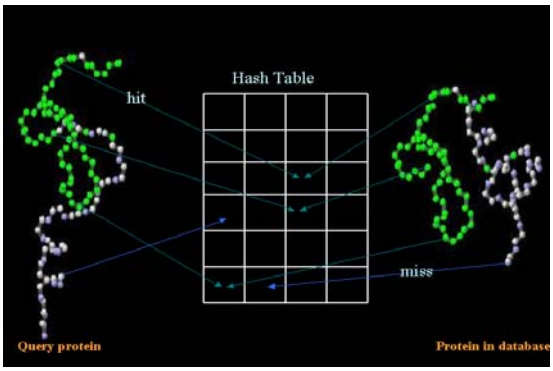


Figure 4. Geometric hashing

The hash function which calculates the index for each atom of a protein in the hash table uses the distance from origin to itself and the orientation of the atom as its hash value. For example, let A1 be an atom with the coordinate (x, y, z) with respect to a reference frame, we calculate its hash table index in the following way.

$$\text{Hash value of A1} = \sqrt{x^2 + y^2 + z^2} \quad \text{..... (4)}$$

However, some other issues need to be considered here. Due to the low resolution of protein structural data from the Protein Data Bank, there might be inaccuracy in recognition among proteins. Moreover, it is also possible to incorrectly match geometrically similar proteins with different chemical properties. In order to enhance the performance of the method proposed, we adopt a frequently used similarity matrix, Dayhoff PAM250, which is originally used for sequence alignment problems. The PAM250 similarity matrix is listed in Table 1.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Table 1. PAM250 table

When two reference frames match with each other, we accumulate the similarity score by looking up the similarity matrix. We set a threshold distance 1Å, beyond which atoms will not be considered as a match. If no atoms can be matched within the threshold distance, we assign the score to the minimal score, which is –8, of the similarity matrix. The final score is normalized for perfect matching to have a unity score.

The following shows the overall algorithm step by step.

Preprocessing phase:

For each protein structural data in the database do the following:

1. Extract the 3D data of each atom in a protein. Assume that *the data size is n*
2. For each reference frame, or basis, do the following:
   (a) Compute the coordinates (u, v, w) of the remaining atoms in the coordinate frame defined by the basis.
   (b) After proper quantization, use the hash index for each atom into a hash table data structure and insert in the corresponding hash table by the information of an atom.

Recognition phase:

When presented with an input structure of a protein, do the following:

1. Extract the atoms' data of interest. Assume that S is the set of the interest atoms found.
2. Choose an arbitrary reference frame of interest atoms in the structure.
3. Compute the coordinates of the remaining atoms of interest in the new coordinate system .
4. Appropriately quantize each such coordinate and ac-

cess the appropriate hash table; for every entry found there, cast a vote for the protein and the reference frame.

5. Histogram all hash table entries that received one or more votes during step 4. Proceed to determine those entries that received more than a certain number, or threshold, of votes: Each such entry corresponds to a potential match.
6. For each potential match discovered in step 5, recover the transformation T that results in the best least-squares match between all corresponding proteins.
7. go back to step 2 and repeat the procedure using a different reference frame.

4. Experimental Results

The International Union of Biochemistry and Molecular Biology (IUBMB) has classified enzymes into six classes.

EC1 Oxidoreductase
EC2 Transferase
EC3 Hydrolase
EC4 Lyase
EC5 Isomerase
EC6 Ligase

They are classified according to their different chemical reactions. In the following paragraphs, we will discuss how and why we choose a certain data set as the input to our system. Furthermore, a list of results will be given to show the capability of our system and the proposed algorithm.

We classify every protein in enzyme class five according to their functions and find the possible candidates of active sites of proteins belonging to this class. An active site of a protein determines the function of this protein. Thus proteins with similar functions share some similarities in their shapes in the vicinity of their active sites. Hence we collect protein files containing site information from the Protein Data Bank and then extract site three-dimensional structures from within these collected files to create our site database. Every site structure in the site database is related to a specific function and has an enzyme class number (E.C. No.) representing the class to which this active site belongs.

In the followings, we describe how we choose test protein files to be compared with site structures in the site database.

First, protein files belonging to enzyme class five are collected and we then keep those files, which have enzyme class numbers in order to verify the results and do not contain site information to prevent a bias in our experiment since site structures are extracted from protein files containing site information in creation of site database. If we select these files in as our test data, they will always match site structures in the site database perfectly, thus bias the results of our experiments.

The way we conduct our experiment is explained in this section and the result will be shown. For every test protein, we use our algorithm to find common substructures between its three-dimensional structure and every site structures in the site database. The structural similarity between a test protein structure and a site structure is calculated for every site structure in the site database and these similarities are then sorted in descending order. We compare the enzyme class number of a test protein with the enzyme class number of the site with highest structural similarity between the test protein structure and its own structure. If these two enzyme class numbers are the same, then we claim that we have correctly found the function of this test protein and their common substructures are possible locations of active sites of this test protein. The results are shown in Table 2.

The result shows that we have correctly classified 39 proteins out of 43 trials and the accurate rate is about 90.7%. As we know, the E.C. No. is a four-number code. A test protein is not included when calculating the accurate rate, if there exists a “99” in the E.C. No. of this test protein because an E.C. No. with a “99” is a miscellaneous collection of enzymes and there might be more than one function in that E.C. entry.

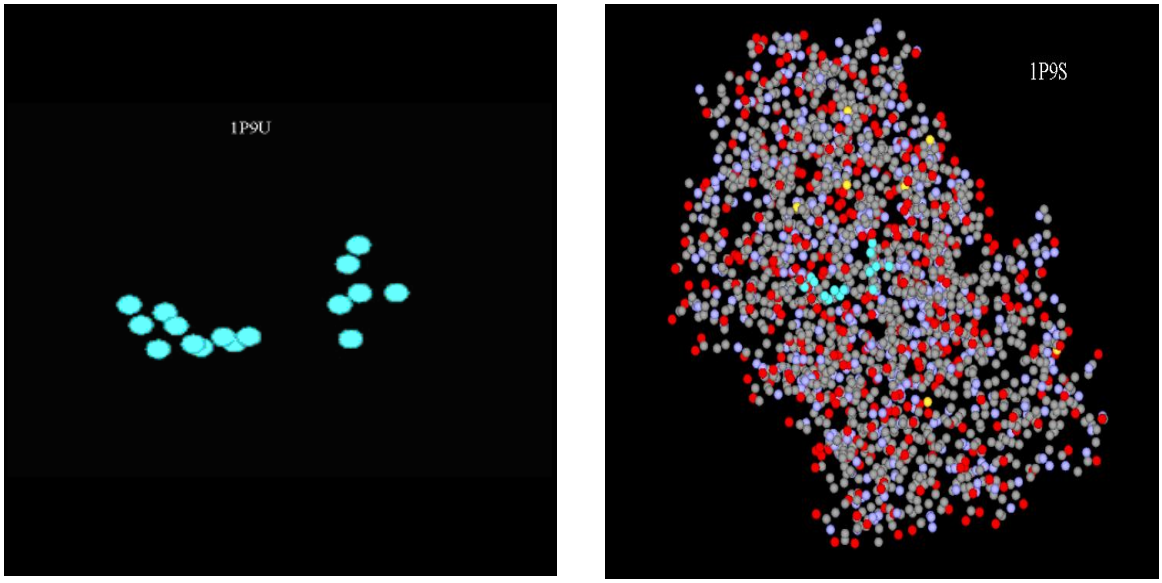
Moreover, proteins with correct matches may have active sites locating on the positions where their common substructures between themselves and the corresponding best-matched site structures are.

In addition to the above experiment, we also verify that the structures of active sites of coronavirus protein extracted from human being and pig are alike. Kanchan et al.[21] have found the site structure of the coronavirus protein from pig and claim that coronavirus protein from human being is likely to have an active site on sequence number 41 and 144 of its amino acid sequence, that is, His 41 and Cys 144. We then utilize the information to extract the site structure from coronavirus from pig and compare this site structure with the one from human coronavirus. Figure 5 illustrates the result.

Test protein pdb i.d.	E.C. No.	Site source protein pdb i.d.	E.C. No.	Match	Test protein pdb i.d.	E.C. No.	Site source protein pdb i.d.	E.C. No.	Match
1A31	5.99.1.2	1J5S	5.3.1.12	No	1G57	5.4.99.-	2CHT	5.4.99.5	Fair
8CHO	5.3.3.1	1OPY	5.3.3.1	Yes	1FZT	5.4.2.1	1E59	5.4.2.1	Yes
3GSB	5.4.3.8	1E7S	5.1.3.-	No	1FKK	5.2.1.8	1D7I	5.2.1.8	Yes
2SQC	5.4.99.-	1A7X	5.2.1.8	No	1FD9	5.2.1.8	1D7J	5.2.1.8	Yes

1TCD	5.3.1.1	1CT1	5.3.1.1	Yes	1F8A	5.2.1.8	1PIN	5.2.1.8	Yes
1QO2	5.3.1.1	1A7X	5.2.1.8	No	1DBF	5.4.99.5	2CHT	5.4.99.5	Yes
1QNG	5.2.1.8	1CWO	5.2.1.8	Yes	1D9T	5.3.1.1	1TPC	5.3.1.1	Yes
1OIS	5.99.1.2	1CWI	5.2.1.8	No	1D6M	5.99.1.2	1DYW	5.2.1.8	No
1N1A	5.2.1.8	1D7I	5.2.1.8	Yes	1CYO	5.99.1.2	1O99	5.4.2.1	No
1MUW	5.3.1.5	1GW9	5.3.1.5	Yes	1CLK	5.3.1.5	1GW9	5.3.1.5	Yes
1MO0	5.3.1.1	1TPH	5.3.1.1	Yes	1CD5	5.3.1.1	1TPC	5.3.1.1	Yes
1MNZ	5.3.1.5	1GW9	5.3.1.5	Yes	1C7H	5.3.3.1	1OPY	5.3.3.1	Yes
1M7O	5.3.1.1	1CT1	5.3.1.1	Yes	1C5F	5.2.1.8	1CWL	5.2.1.8	Yes
1M6J	5.3.1.1	1CT1	5.3.1.1	Yes	1BXB	5.3.1.5	1GW9	5.3.1.5	Yes
1M5Y	5.2.1.8	1E7S	5.1.3.-	No	1BTM	5.3.1.1	1TPH	5.3.1.1	Yes
1M1B	5.4.2.9	1PYM	5.4.2.9	Yes	1BKF	5.2.1.8	1D7J	5.2.1.8	Yes
1LOP	5.2.1.8	1CWF	5.2.1.8	Yes	1BHW	5.3.1.5	1GW9	5.3.1.5	Yes
1L6F	5.1.1.1	1CWH	5.2.1.8	No	1B6C	5.2.1.8	1D7I	5.2.1.8	Yes
1KOJ	5.3.1.9	1GZV	5.3.1.9	Yes	1B0Z	5.3.1.9	1GZV	5.3.1.9	Yes
1JVM	5.2.1.8	1A7X	5.2.1.8	Yes	1AW1	5.3.1.1	1TPC	5.3.1.1	Yes
1IHG	5.2.1.8	1CWL	5.2.1.8	Yes	1AMK	5.3.1.1	1TPH	5.3.1.1	Yes
1I8H	5.2.1.8	1A7X	5.2.1.8	Yes	1AK4	5.2.1.8	1CWH	5.2.1.8	Yes
1I7O	5.3.3.10	1GTT	5.3.3.10	Yes	1AG1	5.3.1.1	1HG3	5.3.1.1	Yes
1I45	5.3.1.1	1TPW	5.3.1.1	Yes	1A41	5.99.1.2	1GW9	5.3.1.5	No
1HOP	5.2.1.8	1BCK	5.2.1.8	Yes	8TIM	5.3.1.1	1TPH	5.3.1.1	Yes
1GYJ	5.3.2.1	1GYY	5.3.2.1	Yes					

Table 2. The first and second columns are the pdb id and E.C. No. of test proteins respectively, while the third and fourth columns are the pdb id and E.C. No. of the source proteins to which the sites belong. If the E.C. No. of the test protein is the same as the E.C. No. of the site source protein, it represents a correct prediction; otherwise a mismatch.



(a) **(b)**

Figure 5. Similarity between functional site of coronavirus from pig and the whole structure of coronavirus from human. The blue atoms represent the three-dimensional common substructures between them. (a) The structure of the active site of coronavirus protein extracted from pig. (b) The structure of coronavirus from human.

5. Conclusion

In this paper, we have proposed a geometry-based algorithm which is able to predict the possible functions and the locations of active sites in protein molecules.

At first, we extract the active sites of all the proteins belonging to enzyme class five. For each protein with unknown locations of active sites, we compare it with all the site structures in our site database. In the process of comparing, the geometric hashing algorithm is used for common substructure matching. After the previous steps

are done, the common substructure where the protein matches can be considered as a possible position of an active site. Furthermore, it is probable that the protein shares the same function with the most similar active site in the site database.

In addition, a system is developed as well to demonstrate the feasibility of the proposed algorithm. The system has a visualization tool which displays the results visually on the screen and allows users to operate on them. We provide users a convenient way to control the behaviors of the system.

6. Future Work

Many steps in the proposed algorithm can be refined by more delicate approaches. This includes a new coordinate representation for the protein backbone structure, a new measurement of the dissolubility of residues in a protein [22, 23, 24] and a fault-tolerant ability to protein structural data. In addition to the improvement in the algorithm, we will also make efforts in accelerating the speed of prediction in order to handle massive structural data from structural genomics projects.

Though it is convenient to use the system on personal computer, it is still desirable that users can interact with the system via network. In next step, we will port our system to a web-based environment along with a database with extracted information. The database will be free for download and users around the world can use the prediction program by their web browsers. We'll also integrate more biological information such as solvent degree for each amino acid in a protein molecule into our algorithm to improve the precision of the proposed algorithm.

7. Acknowledgements

This research was supported in part by National Science Council 92-2622-E-002-002.

Reference

[1] Brady, Jr., G. P. and Stouten, P. F. W.: Fast prediction and visualization of protein binding pockets with PASS. Journal of Computer-Aided Molecular Design (2000) 14: 383-401.

[2] Chen, S.-C. and Chen, T.: Retrieval of 3D protein structure. Proceedings of ICIP 2002, 34-43.

[3] Günther, J., Bergner, A., Hendlich, M., and Klebe, G.: Utilising structural knowledge in drug design strategies: applications using relibase. Journal of Molecular Biology (2003) 326: 621-636.

[4] Rockey, W. M. and Elcock, A. H.: Progress toward virtual screening for drug side effects. Proteins. (2002) 48: 664-71.

[5] Smith, G. R. and Sternberg, M. J. E.: Prediction of protein-protein interactions by docking methods. Current Opinion in Structural Biology (2002) 12: 28-35.

[6] Totrov, M. and Abagyan R.: Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins. (1997) 29(S1): 215-220.

[7] Najmanovich, R., Kuttner, J., Sobolev, V., and Edelman, M.: Side-chain flexibility in proteins upon ligand binding. Proteins. (2000) 39: 261-268.

[8] Cappello, V., Tramontano, A., and Koch, U.: Classification of proteins based on the properties of the ligand-binding site: the case of adenine-binding proteins. Proteins. (2002) 47: 106-115.

[9] Bamborough, P. and Cohen, F. E.: Modeling protein-ligand complexes. Current Opinion in Structural Biology (1996) 6: 236-241.

[10] Ondrechen, M. J., Clifton, J. G., and Ringe, D.: a simple computational predictor of enzyme function from structure. Proceedings of the National Academy of Sciences of the United States of America. (2001) 98:12473-12478.

[11] Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F.: Prediction of protein function using protein-protein interaction data. Proceedings of CBS 2002, p.197.

[12] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M.: The Protein Data Bank: a computer-based archival file for macromolecular structures. Journal of Molecular Biology (1977) 112: 535-542.

[13] Hendlich, M., Bergner, A., Günther, J., and Klebe G.: Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. Journal of Molecular Biology (2003) 326: 607-620.

[14] Waterman, S., Arratia, R., and Galas, D.J.: Pattern recognition in several sequences: consensus and alignment. Bulletin of Mathematical Biology (1984) 45: 515-527.

[15] Delcoigne, A. and Hansen, P.: Sequence comparison by dynamic programming. Biometrika. (1975) 62: 661-664.

[16] Needleman, S. B. and Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology (1970) 48: 443-453.

[17] Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J. E.: Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology

in genome annotation and to protein docking Journal of Molecular Biology (2001) 311: 395-408.
[18] Chang, B. C. H. and Halgamuge, S. K.: Protein motif extraction with neuro-fuzzy optimization. Bioinformatics (2002) 18: 1084-1090.
[19] Murray, K. B., Gorse, D., and Thornton, J. M.: Wavelet transforms for the characterization and detection of repeating motifs. Journal of Molecular Biology (2002) 316: 341-363.
[20] Wolfson, H. J. and Rigoutsos, I.: Geometric hashing: an overview. IEEE Computational Science and Engineering (1997) 4: 10-21.
[21] Anad, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R., and Hilgenfeld, R.: Coronavirus main proteinase $3CL^{PRO}$ structure: basis for design of anti-SARS drugs. Science. (2003) 300: 1763-1767.
[22] Yeates, T. O.: Algorithms for evaluating the long-range accessibility of protein surfaces. Journal of Molecular Biology (1995) 249: 804-815.
[23] Yuan, Z., Burrage, K., and Mattick, J. S.: Prediction of protein solvent accessibility using support vector machines. Proteins. (2002) 48: 566-570.
[24] Pollastri, G., Baldi, P., Fariselli P., and Casadio R.: Prediction of coordination number and relative solvent accessibility in proteins. Proteins. (2002) 47: 142-153.