PROTEIN RETRIEVAL BY MATCHING 3D SURFACES

Shann-Ching Chen and Tsuhan Chen

Dept. of Electrical and Computer Engineering, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
{shanncc, tsuhan}@andrew.cmu.edu

ABSTRACT

Folding into complex 3D structures, protein molecules are responsible for carrying out nearly all of the essential functions in living cells by properly binding to other molecules with a number of chemical bonds connecting neighboring atoms. Locations of these atoms are called the binding sites. To help biologists identify the functions of unknown proteins and to discover new functions of known proteins, it is desirable to retrieve common binding sites among proteins. We propose to use the geometric hashing method to perform protein surface matching to identify similar binding sites. Furthermore, two techniques, *α-hull* and 3D reference frames, are adopted to reduce the complex computation.

1. INTRODUCTION

It is the geometrical shape that determines if a protein can bind to another molecule. Consequently, proteins sharing resembling structures may perform similar functions. It is desirable to find common substructures among proteins rather than the whole structures. To perform substructure matching, Fischer et al. [1] have exploited the geometric hashing paradigm previously introduced in computer vision [2]. Their method is based on preprocessing and recognition algorithms of complexity $O(n^3)$, where n is the number of residues of interest. Later, Pennec and Ayache [3] introduced a 3D reference frame attached to each residue, which drastically reduces the complexity of recognition. Andrew et al. also use the geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases[4]. They investigated those enzymes with the His-based "catalytic triad" active site.

Other work deals with the protein structure representation. Edelsbrunner proposed a α-hull theory [5], which interprets a set of discrete points as a shape. The α-hull is able to identify cavities and protrusions on the protein surfaces to model the protein shape.

With increasing 3D biological molecular structures derived and deposited at the Protein Data Bank (PDB) [6], interesting similarities among proteins can be found based on the 3D geometry. However, the geometric hashing algorithm is very computational expensive. Base on the existing method, our work is trying to reduce the computation and to find more general cases for enzyme active binding sites. From a PDB file we can get the protein binding site information, which is the position of certain molecules that actually interact with the substrates or ligands. The binding site is represented as several residues in the protein. Since most of the proteins' binding sites are on the surface, in order to simplify computation only those residues on the surface need to be considered as 3D reference frames to generate the geometric hashing table. Then the residues on the surface can be extracted with the help of surface triangulation generated by the α-hull algorithm. After this, geometric hashing algorithm is performed to find the surface matching of proteins.

This paper is organized as follows. In Section 2 we discuss the algorithm. Some experiment results with the Enzyme Classification Database are provided in Section 3. Conclusions and extension of our work are in Section 4.

2. PROTEIN MODELING AND MATCHING STRATEGY

2.1 Geometric hashing algorithm

The geometric hashing algorithm was introduced [2] for model-based recognition in computer vision. It is composed of two stages: pre-processing and recognition. The basic idea is to store in a database at pre-processing time a redundant representation of the models by rigid transformation, based on local features to allow for occlusion. By doing so, the representation of the query protein computed at recognition time will present some similarity with that of some database proteins. Figure 1 shows the flowchart of our algorithm.

Preprocessing. Local features are extracted from each protein structure. Each residue in the protein can be treated as a 3D reference frame, which we will discuss later. With this frame, we have three orthonormal vectors to describe other residues of the protein in this particular 3D coordinate system. Based on the basis, the 3D positions of all the residues are the features, which are inserted into the hashing table with an index (protein, residue). This step is performed without any knowledge of the

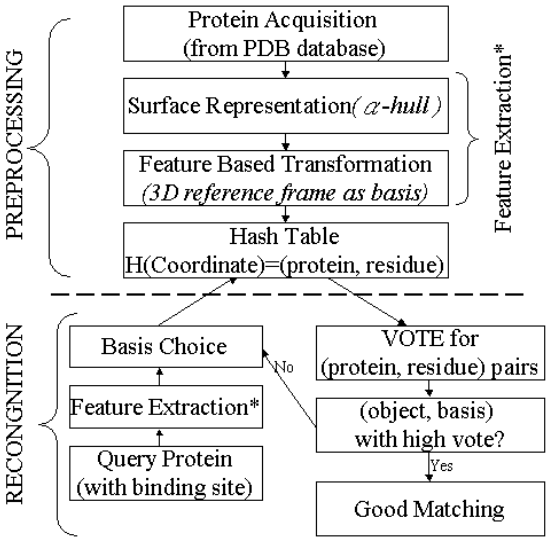


Figure 1. Flowchart of the proposed algorithm

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 frame in the hashing table, we accumulate the number of compatible 3D features, which is called voting. This will be the matching score of these two frames. The process is repeated with each frame of the query protein taken as the reference frame. The output is the list of matching reference frames of query and database proteins with their associated score. We keep the matches with high scores.

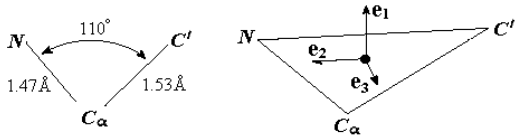
At the recognition stage, some other issues need to be considered. Due to the conformation changes when a protein binds to its ligands and the low resolution of protein structure determination, there may be certain variation between matching reference frames. On the other hand, it is also possible to incorrectly match geometrically similar proteins with totally different chemical properties. In order to enhance the matching performance, we adopt a frequently used similarity matrix Dayhoff PAM250 [7] for sequence alignment. When two reference frames match each other, we accumulate the similarity score by looking up the similarity matrix. We set a threshold distance 2Å, beyond which residues will not be considered. The matching residue is chosen as with the maximum of similarity score divided by the distance while distance is greater than 1Å. While smaller then 1Å, the distance does not affect the matching score. If no residues can be matched within the threshold distance, we accumulate it with the minimal score, which is –8, of the similarity matrix. The final score is normalized for perfect matching to have a unity score. In addition, in order to simplify the heavy computation of geometric hashing, the α-hull algorithm is applied before feature extraction. We will discuss it in the next subsection.

2.2 Alpha hull algorithm

In molecular biology, the van der Waals surface [8] is often used for molecular modeling. The α-hull theory can be applied to model the molecular surfaces with the van der Waals surface. With a public domain α-hull algorithm [8], we can extract the surface atoms and, further, the surface residues from a given PDB file. This step can reduce the computation in geometric hashing algorithm. In addition, since in PDB files the crystallized proteins usually have ligands attached at the binding site, we should remove the ligands first before we apply this algorithm.

2.2 3D Reference frame

Proteins are composed of possibly several chains of residues linked to each other by peptide bonds. The backbone of the protein is composed of Carbon (C_α) atoms. The geometry of the atoms attached to the C_α is perfectly determined. In particular, the three atoms N, C_α, C form a known triangle from which we can define a frame (a point and a trihedron; see Figure 2). It can uniquely determine the position and orientation of a residue in space [5]. With this mechanism, we can now choose a single residue as a basis. Within each protein, all the surface residues obtained by the α-hull algorithm are used as the basis to generate the hashing table at the pre-processing step of the geometric hashing algorithm.



$$e_1 = \frac{C'N \times C'C_\alpha}{\left\| C'N \times C'C_\alpha \right\|} \qquad e_2 = \frac{C'N}{\left\| C'N \right\|} \qquad e_3 = e_1 \times e_2$$

Figure 2. Geometry of a residue and the definition of a basis

3. EXPERIMENT RESULTS

An Enzyme Structure Classification Database is available [9]. There are 8390 PDB enzyme entries in the PDB (as of January 2002) including 8059 separate PDB files. Some entries have no corresponding files since for these entries only protein sequences and functions are known but not the 3D structures, while some other entries may have more than 30 files. Proteins in the same entry share the same function. One protein can have multiple PDB files since different crystallizations of this protein can be derived under different environments or while combining

E.C. number	PDB ID	Function Description	SWISS-PORT Code	No. of Residues	Score
E.C.5.2.1.8	1bck	Peptidylprolyl isomerase	CYPH_HUMAN	165	1.000000
E.C.5.2.1.8	1cyn	Peptidylprolyl isomerase	CYPB_HUMAN	178	1.000000
E.C.5.2.1.8	1dyw	Peptidylprolyl isomerase	CYP3_CAEEL	172	1.000000
E.C.5.2.1.8	1ihg	Peptidylprolyl isomerase	CYP4_BOVIN	364	0.743590
E.C.5.2.1.8	1a7x	Peptidylprolyl isomerase	FKB1_HUMAN	214	-0.492383

Table 1: Part of high scoring proteins given query 1bck

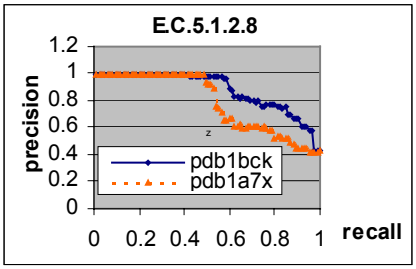


Table 2: Precision Recall Graph of two queries: 1a7x and1bck

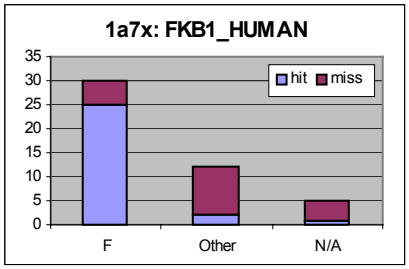


Table 3: Hit-and-miss bars for query protein 1a7x. F set has many successful hits.

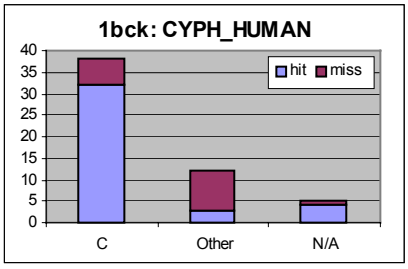


Table 4: Hit-and-miss bars for query protein 1bck. C set has many successful hits.

with different ligands. We can tell whether two proteins are the same by using the SWISS-PORT codes [10], which are the identifications of proteins. To get a query model, we select one of those proteins with binding site information. Currently, 20% (3122/16496) of the PDB files already have this information. We choose 200 proteins from Enzyme Database, with 85 of them are in Set E.C. 5.2.1.8., which contains peptidylprolyl isomerase proteins with similar functions. The reason to choose this set is that some enzymes here have clear and proper binding sites so that we can use them as queries. In our experiments, we would like to show that most of the proteins within the same set could be retrieved with high similarity.

In the first experiment, as shown in Table 1, we choose protein "1bck" as the query. This protein has one binding site with 13 residues "RFMQGANAQFWLH". At the first three rows, by the SWISS-PORT code and the number of residues we can tell that the target proteins (at 2nd and 3rd row) are not the same as the query protein, but they get perfect match with score 1! These are very good examples illustrating that different but similar proteins, with the same function and binding sites, can be retrieved by the proposed algorithm. Although there is one residue in the protein 1ihg cannot be matched with any of the query proteins' residues in the 3D space, the 4th row shows that, this protein still gets a high matching score 0.74.

Unfortunately, the protein at the 5th row can't match well. To study this further, we use 1a7x (FKB1_HUMAN) and 1bck (CYPH_HUMAN) as the query respectively and obtain the corresponding precision-recall plots as Table 2. There are 200 proteins in this experiment and we want to retrieve the 85 proteins with the same function. An interesting result shows that, 1a7x can retrieve those proteins with SWISS-PROT code starting from F very well, such as FKB2_HUMAN, FKBP_H, and FKBP_BOVIN. On the other hand, 1bck can retrieve those proteins with SWISS-PROT code starting from C very well, such as CYPB_ECOLI, CYP4_BOVIN, and CYPH_H. We manually group those SWISS-PROT codes into set C (starting from C), F (starting from F), N/A (without SWISS-PROT code) and Others (other miscellanary codes). Given the threshold t, a hit happens if the score is larger than the threshold. In this experiment t is 0.4. Hit-and-miss bars can be drawn in Table 3 and Table 4. Most of the proteins can be retrieved by these two queries. A reasonable explanation for this is that, proteins with different shapes may still perform the same function. Some proteins in F, C and Other sets are structurally quite different or have large conformational change. Moreover, if some proteins without the SWISS-PROT code can still be retrieved, it is possible that we have found their functions.

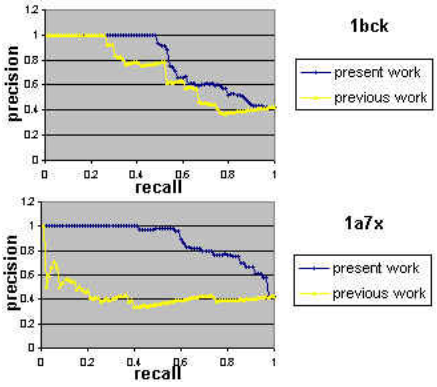


Table 5: Precision-Recall Plots

We also compare the proposed algorithmthis method with our previous work. [11] Previously we only consider the whole protein structure similarity. Table 5, the precision-recall graph, shows that considering the substructure matching is much better.

4. DISCUSSION AND CONCLUSIONS

There are several contributions of this work. First, due to the fact that the binding sites are on the protein surface, we use the α-hull algorithm to extract the surface points and surface residues to simplify the computation. Second, in [3] they proposed a 6D (3 translations and 3 rotations) index for the geometric hashing, while we used a 3D system where only translation components are considered. As a result, the computation time required for the matching can -be drastically reduced. Third, we adopted the Dayoff PAM250 similarity matrix to perform our 3D matching, which enhances the matching performance. Without using the similarity matrix, the matching is likely to yield geometrically similar proteins with totally different chemical properties. Finally, the algorithm can deal with protein structures that are not precisely determined, as long as the conformational deformations are small.

There are several possible ways to extend our work. First, due to the limited binding site information, we could only choose a small portion of proteins as the queries. According to many biologists, if the binding site information for a certain protein is not available, we can use the binding site information of the ligand for that protein as a close approximation. Second, when bonded with other molecules, proteins may experience conformation changes. Although our algorithm can handle proteins with small conformation changes, new theories should be developed to tackle proteins with large conformation changes. Furthermore, proteins totally different in shape can still perform same functions. Since our algorithm is suitable for local and precise matching of the binding sites, it is hard to handle large conformation changes and proteins with different shapes altogether. In addition, since people have classified proteins based on sequence alignments and 3D shape approximation, it is possible to introduce a new classification method using protein surfaces.

In molecular biology, to identify protein functions is essential to drug design and disease prediction. Biologists suggest that a large amount of protein structures will be crystallized without knowing their functions in the next decade. Our algorithm therefore is very promising to help identify the functions of unknown proteins and even discover new function of known proteins. It will also save time for biologists by reducing their search space in an effort to find good candidates to be used in lab experiments to identify protein functions. To achieve this, our algorithm can be applied to different E.C. sets and if some proteins get high similarity scores, it is possible that some new functions of the proteins have been discovered.

REFERENCES

[1] D. Fischer, O. Bachar, R. Nussinov and H. Wolfson. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. J. Biomol. Struct. Dynam., 9, 769–789, 1992

[2] Y. Lamdan and H. J. Wolfson. Geometric Hashing: A General and Efficient Model-Based Recognition Scheme: In Proceedings of the IEEE Int. Conf. on Computer Vision, 1988.

[3] X. Pennec and N. Ayache. An $O(n^2)$ Algorithm for 3D Substructure Matching of Proteins. Shape and Pattern Matching in Computational Biology, A. Califano, I. Rigoutsos and H.J. Wolson eds, Plenum Publishing, 1994.

[4] Wallace AC; Borkakoti N; Thornton JM. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: Application to enzyme active sites Protein Science, 6:2308-2323, 1997.

[5] H. Edelsbrunner and E. P. Mucke. Three-dimensional alpha shapes. ACM Trans. Graphics 13, 43-72, 1994. http://www.alphashapes.org/alpha/index.html

[6] Protein Data Bank, Quarterly Newsletter No. 71, Brookhaven National laboratory, 1995.

[7] R. M. Schwartz and M. O. Dayhoff. Atlas of Protein Sequence and Structure, National biomedical research foundation Washington DC, 5, 353-358, 1978.

[8] J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudharkar and S. Subramaniam. Analytic shape computation of macromolecules II: inaccessible cavities in proteins. Proteins: Structure, Function, and Genetics 33, 18-29, 1998.

[9] Enzyme Nomenclature, NC-IUBMB, Academic Press, New-York, 1992. http://www.biochem.ucl.ac.uk/bsm/enzymes/

[10] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its supplement TrEMBL, Nucleic Acids Research, 25, 31-36, 1997. http://www.expasy.ch/

[11] S. C. Chen and T. Chen, Retrieval of 3D Protein Structure, IEEE International Conference on Image Processing, 2002.