RETRIEVAL OF 3D PROTEIN STRUCTURES

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

A protein molecule is made of a long chain of amino acid sequences that fold into a complex three-dimensional structure. It is often the geometrical shapes that determine the protein functions. In molecular biology, researchers use sequence alignment and structure matching to compare the similarity among proteins. Considering proteins as 3D structures, we propose a novel algorithm to identify geometry-based features to retrieve similar proteins without having to deal with complex chemical characteristics and biological properties. A web-based Dali server, which performs well in three-dimensional structure matching, is used as the ground truth to evaluate our algorithm. Our system performs close to the ground truth with much simplicity and efficiency.

1. INTRODUCTION

In the living cells, proteins carry out nearly all of the essential functions by properly binding to other molecules. Therefore, the protein structures are significant to their functions [1]. Proteins with similar structures typically have the same functions. To help biologists to identify the functions of unknown proteins and to discover new functions of known proteins, it is desirable to find the similarities between protein 3D structures. Protein classifications and function predictions can be made by proteins structure comparisons.

The Protein Data Bank (PDB) [2] is the worldwide repository for the processing and distribution of three dimensional biological molecular structure data. From the Research Collaboratory for Structural Bioinformatics (RCSB) [3] website we can access the PDB database, in which biologists use X-ray crystallography and nuclear magnetic resonance spectroscopy to analyze the structure of protein molecules. From the PDB file of each protein we can understand its amino acid sequences and atoms with 3D coordinates and connectivity. Figure 1 shows an example.

```
HEADER    OXIDOREDUCTASE             19-AUG-97              2ATJ
ATOM    1  N   MET A  0    -17.410  2.471   5.878  1.00 34.70        N
ATOM    2  CA MET A  0    -17.162  1.046   5.517  1.00 35.11        C
ATOM    3  C   MET A  0    -16.558  0.282   6.691  1.00 32.97        C
ATOM    4  O   MET A  0    -16.012  0.882   7.618  1.00 32.19        O
•
•
•
HETATM 4779 FE    HEM A 350 -1.868   0.604  31.288  1.00  8.71 FE
HETATM 4780 CHA HEM A 350  1.435   1.584  31.604  1.00  3.76   C
HETATM 4781 CHB HEM A 350 -2.725   2.618  33.944  1.00  7.16   C
HETATM 4782 CHC HEM A 350 -5.152  -0.067  30.732  1.00  4.45   C
```

Figure 1. PDB file example

The HEADER record contains the classification for the entry, the date of deposition of the file, and the PDB identification code. The ATOM records present the atomic coordinates for standard groups. Column 3 represents the atom name, while the amino acid sequence to which the atom belongs is in Column 4. The x, y, z coordinates of each atom in angstroms are in Column 6, 7, and 8, respectively. The HETATM record, which contains heterogeneous atoms, presents the atomic coordinate records for atoms within "non-standard" groups. The columns have the same meanings as those of the ATOM record. The connectivity of atoms is stored implicitly according to the atom order and the amino acid it belongs. Another record called CONNECT, which is not shown here, contains additional connectivity explicitly. A snapshot of a portion of a protein structure is in Figure 2.

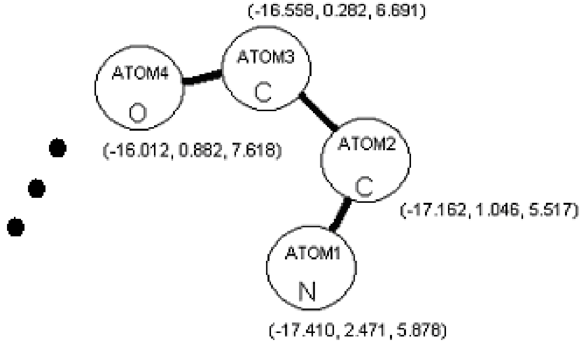


Figure 2. 3D structure of a portion of the PDB file

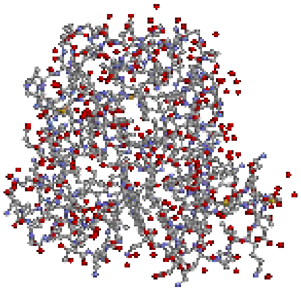


Figure 3. 3D rendition of PDB 103m

Atoms are connected together to form amino acid residues, which a protein is composed of. The visualization of the whole PDB file “103m” with the Rasmol program [4] is in Figure 3. Several algorithms have been proposed to perform 3D structure matching and have accessible websites. A popular used Dali server [5], presents a general approach to optimal pairwise alignment of protein structures. The three-dimensional coordinates of each protein are used to calculate residue-residue distance matrices. Another algorithm involves a combinatorial extension (CE) [6] of an alignment path defined by aligned fragment. Since there is no universal agreement of the similarity of proteins, it is not easy to assess the results of the similarity retrieval systems to tell which one is the best [7]. We use the Dali server as the ground truth to make assessment of our retrieval results.

We have used moments and the mesh representation for 3D model retrieval [8]. In this paper, we extend that work to perform retrieval of 3D protein structures. Here we mainly consider the 3D structure of the protein. The position of each atom was processed and weighted by the atomic weights in the periodic table. Some features as the primary and secondary structures extracted from the PDB files are also taken into account, but with less weight. The primary structure is the amino acid sequence, while the secondary structure refers to certain common repeating structures found in proteins like alpha helix and beta sheet.

The paper is organized as follows. In Section 2 we discuss the proposed matching algorithm. Some experiment results are provided in Section 3. Conclusions and other issues are given in section 4.

2. FEATURE EXTRACTION AND MATCHING STRATEGY

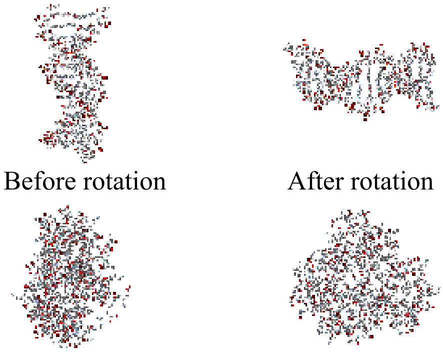
Since the position of each atom is of most interest, the original protein structure comparison problem can be reduced to comparing the similarity of the spatial relationship and overall appearance of two given sets of points. We extract geometry-based features of a set of points, such as the number of atoms, the aspect ratios and the moments, and then perform the matching and retrieval.

First of all, we would like to find the best alignment of two proteins before we extract the features for matching. Atoms that proteins are mainly composed of are Carbon(C:12.01), Nitrogen(N:14.01), Oxygen(O:16.00), Sulfur(S:32.07), and Hydrogen(H:1.00). We weight these atoms by the atomic weights (in the parentheses) and weight other sparse atoms as 1.00. The center of mass is then computed and each point is translated so that the new center of mass is at the origin. In order to align the 3D protein structures with rotation, a 3x3 matrix is constructed by the second-order moments [8]:

$$S=\begin{bmatrix} M_{200} & M_{110} & M_{101} \\ M_{110} & M_{020} & M_{011} \\ M_{101} & M_{011} & M_{002} \end{bmatrix}$$

where $M_{abc}=\sum_k m_k \times X_k{}^a \times Y_k^b \times Z_k{}^c$ and m_K is the atomic weight, (X_K, Y_K, Z_K) is the coordinates of the atom.

The principle axis is obtained by computing the eigenvectors of the matrix S, which is also known as the principle component analysis (PCA). The eigenvector corresponding to the largest eigenvalue is the first principal axis. The next eigenvector corresponding to the secondary eigenvalue is the second principal axis, and so on. The best alignment of two proteins [8][9] is found by rotating the point sets to their own principal axes. We also make sure that M_{300} and M_{030} are positive after the rotation to avoid future ambiguity. Figure 4 shows the results of this algorithm.



Before rotation

After rotation

Figure 4: Protein models before and after rotation

We discard those atoms belonging to the HETATM record (see Figure 1) in the PDB file, which presents the atomic coordinate records for atoms within "non-standard" groups. Those heterogeneous atoms will be some outliers of the proteins and contribute badly to the moments if they are far away from the center of mass.

Only atoms in the ATOM record are considered. After alignment, we extract eight 3D structure features: the number of atoms, the render scale, two aspect ratios defined as height and depth divided by width, and 2nd and 3rd order moments including M_{200}, M_{210}, M_{102}, and M_{201}. There are three secondary attributes (HELIX, TURN, and SHEET) explicitly stored in the PDB files. The primary structure, i.e. the amino sequence, is also available of each protein. There are twenty kinds of amino acid residues. Some of them are hydrophobic and the others are hydrophobic. Excluding dependences, we can calculate nineteen residue ratios and the ratio of hydrophobic residues from the primary structure. Figure 5 shows the details of the features and their weights. Notice that we emphasize a lot on the geometric features, so most of the weights are put on 3D and secondary structures.

3D Structure		
Feature	Definition	Weight
Atom number	ATOM number excluding HETATMs	0.08
Render scale	$\max\left(\lvert X_{\max} - X_{\min}\rvert, \lvert Y_{\max} - Y_{\min}\rvert, \lvert Z_{\max} - Z_{\min}\rvert\right)$	0.08
Aspect ratio1	$\sqrt{M_{002}/M_{200}}$	0.08
Aspect ratio2	$\sqrt{M_{020}/M_{200}}$	0.08
Moment	M_{200} M_{210} M_{102} M_{201}	0.32
Secondary Structure		
Feature	Definition	Weight
HELIX	Number of HELIXs in PDB file	0.053
SHEET	Number of SHEETs in PDB file	0.053
TURN	Number of TURNs in PDB file	0.053
Primary Structure		
Feature	Definition	Weight
Residue ratio	Different Residue Ratio in the protein	0.001
Hydrophobic Residue ratio	Hydrophobic Residues Ratio in the protein	0.001

Figure 5: 31 features and their weightings

These 31 extracted features are first normalized by arctan function. The usage of arctan function is to make the largest value to be 1 and smallest value to be –1.

$m_k =$ mean of the k-th feature

$v_k =$ variance of the k-th feature

$F^{(0)}[k]$ = the normalized feature value

$F[k]$ = the original feature value

$$F^{(0)}[k] = \frac{2}{\pi} \times \arctan\left(\frac{F[k] - m_k}{\sqrt{v_k}}\right)$$

After this, all the features are normalized between 1 and –1. The similarity score is then calculated as follows:

$F_Q[k]$ = the k-th feature of the query model

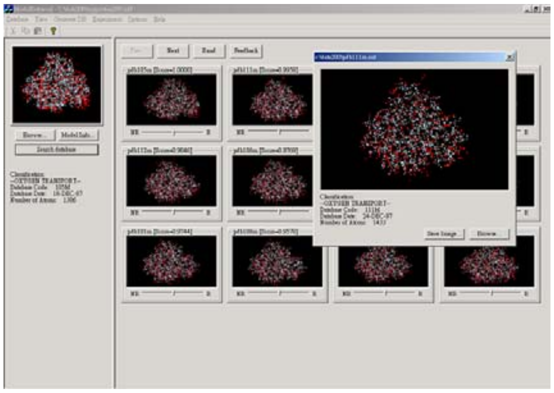
$F_R[k]$ = the k-th feature of the retrieval model

$$\text{Similarity score} = \frac{1}{1+\sum_{k=1}^{N} w_k \times \left(F_R[k] - F_Q[k]\right)^2}$$

where N is the number of features and weights w_k are as in Figure 5. With this similarity measurement, the most similar pair of proteins will have score one and less similar pairs will get a lower score. Then the retrieval results can be obtained by sorting the database with the similarity score with respect to the query. The higher the score, the more similar the model is to the query.

3. EXPERIMENT RESULTS

We have performed experiments based on our matching algorithm. There are around 18000 3D protein and nucleic acid molecular models stored in the Protein Data Bank (increasing rapidly every day). Our system collected 2500 protein 3D structures from them and performs 3D rendering for the retrieved results. (See Figure 6).



Classification:
–OXYGEN TRANSPORT–

Figure 6: Protein Model Retrieval System

We compare our system with the Dali server. The Dali server has a definition of representatives, which are proteins with some special characteristics [5] so that no two representatives have more than 25% one-dimension sequence identity. For each representative, a number of proteins similar to the representative are put together with the representative to form a group. There are 2920 groups in the Dali server. Figure 7 illustrates a group with 1a6m as the representative. Different proteins belong to different groups. Given a query protein, the number of retrieved proteins divided by the total number of proteins in the group gives us the retrieval ratio. From the retrieval ratio we can judge the retrieval performance.

PDBid	Representative	Classification
101m	1a6m	myoglobin Mutant
102m	1a6m	myoglobin Mutant
103m	1a6m	myoglobin Mutant
104m	1a6m	myoglobin
105m	1a6m	myoglobin
106m	1a6m	myoglobin Mutant
• • •		
4mbn	1a6m	myoglobin (met)
5mbn	1a6m	myoglobin (deoxy)

Figure 7:a group with representative 1a6m

Here we choose three groups for the experiments and the precision-recall graph is presented in Figure 8. While the recall is not close to 1 (i.e. 100% recall), the missing proteins are typically those who have a large difference in the number of atoms than that of the query model such that our system considers them dissimilar. Some groups have low retrieval performance because the shapes and appearances of those proteins are quite different visually.

Group	1a6m	1l92	2cba
Number of Atoms of Representative	1336	1292	2079
Number of Protein Chains in the group	189	380	180

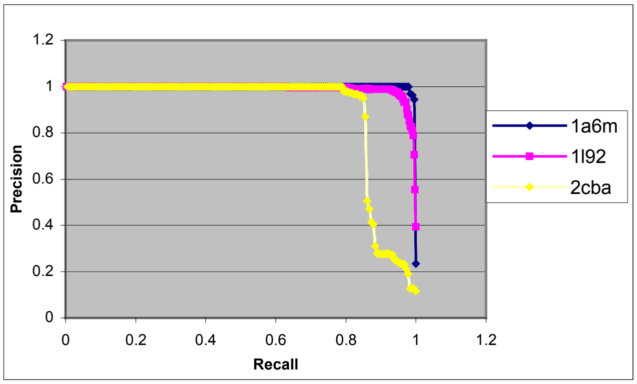


Figure 8. Different groups
and their Precision Recall Graph

By the way, the Dali server can be also used to retrieve only representatives in the database. Our algorithm is intended to deal with proteins with moments, number of atoms, and similar 3D shapes. Since the shape and features of a representative are quite specific because of the low 25% sequence identity constraint, with only representatives our retrieved results are quite different from those of the Dali server.

4. CONCLUSIONS

We developed a system to retrieve proteins with similar three-dimensional structures. Compared with the Dali server, within the same group we can achieve similar results with much simplicity and efficiency, since our algorithm is good at comparing the overall shapes among similar proteins. However, when comparing proteins with only representatives, the results are quite different from those of the Deli Servers'. Although it is not easy to tell which retrieval results are better because of the different considerations of similarity, our 3D model rendering system does present an efficient and intuitive way to recognize the similarity between proteins by their appearance.

There are several potential extensions for this work. First, higher order moments can be considered. Second, our work can be extended to the subregion-matching problem, about which molecular biologists concern a lot. Only some parts of a protein, which are called the binding sites, may have functions and interact with other proteins or DNA's. The subregion-matching problem gives us another point of view of similarity between proteins. Trying to find the common subset of atoms may lead us to discover unknown functions with existing proteins!

REFERENCE

[1] Dimitris Anastassiou, Genomic Signal Processing, IEEE SIGNAL PROCESSING MAGAZINE, July 2001.

[2] http://www.rcsb.org/pdb

[3] http://www.rcsb.org/index.html

[4] Sayle.R. and Bissel.A, RasMol: A program for fast realistic rendering of molecular structures with shadows. In Proceedings of the 10th Eurographics UK'92 Conference, 1992.

[5] Liisa Holm and Chris Sander, Protein Structure Comparison by Alignment of Distance Matrics, J. Mol. Biol. 233, 123-138, 1993. http://www2.ebi.ac.uk/dali/

[6] Shindialov, I.N. and P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(9): 739-747, 1998. http://cl.sdsc.edu/ce.html

[7] Ingvar Eidhammer, Inge Jonassen, William R. Taylor, Structure Comparison and Structure Patterns, Reports in Informatics No. 174, Department of Informatics, University of Bergen, Norway, July 1999.

[8] C. Zhang and T. Chen, "Efficient Feature Extraction for 2D/3D Objects in Mesh Representation", ICIP 2001, Thessaloniki, Greece.

[9] W.Kabsch. Acta Cryst.. A32:922-923, 1978.