Malaysian Journal of Mathematical Sciences 2(2): 205-213 (2008)

Reference Frame for Protein Structure Recognition

[1]Fazilah Othman, [1]Rosni Abdullah and [2]Jamaludin Ali
[1]School of Computer Sciences, Universiti Sains Malaysia,
11800 USM, Pulau Pinang
[2]School of Mathematical Sciences, Universiti Sains Malaysia,
11800 USM, Pulau Pinang
E-mail: fazot@cs.usm.my

ABSTRACT

Data representation is an initial important issue in any procedures. For any raw data being input to a procedure, the researcher will first extract the required features from the raw data. These features must be embodied into a meaningful representation to be inserted in the procedure. This paper will discuss the data representation called reference frame used in geometric hashing algorithm for protein structure matching. Matching 3D structure needs special care so that the important and unique information can be encapsulated and differentiated between one another. The reference frame is generated from backbone fragment i.e. N-Cα-C. This paper will first show the calculation of the reference frame RF1 and second, given a single coordinate for atom S (x,y,z), we want to find new coordinate for S $(\hat{x}, \hat{y}, \hat{z})$ in terms of the reference frame RF1. These new coordinates will be used as the matching features between two structures. Add result here.

Keywords: Reference frame, geometric hashing and tertiary structure recognition

INTRODUCTION

Protein structure matching is an important step towards protein function determination because proteins with similar 3-dimensional (3D) structure may impose the same function [4], [9], [5]. Prior to efficient matching, the tertiary structure must be well embodied to keep the unique characteristics of the structures, thus the similarity and difference can be discovered. From previously published works on structure matching; the terms structure approximation, distance matrices [11] and vector representations are usually used. Reference frame is originally named as *object representation* in computer vision by Lamdan et. al. [1].

In molecular biology, backbone fragment (N-C_α-C) is significant enough to characterize the geometrical features of a protein [4] and [5]. Therefore, the matching in this work will based on the geometrical features of backbone fragment.

This paper will explain the basic flow of the geometric hashing algorithm for structure matching. The calculation to derive reference frame is shown and later, the matching results are represented.

GEOMETRIC HASHING ALGORITHM

Geometric hashing algorithm is divided into two phases; pre-processing and recognition phase [8]. Pre-processing phase can be simplified as below:

a. Extract geometrical information from model structure A;
b. Choose a reference frame;
c. Calculate 3D orthonormal basis associated with this reference frame;
d. Compute the new coordinate of all the other points in the structure as referred to this particular reference frame;
e. Store the reference frame at the hash table by using the coordinates as the address;
f. Repeat for each model reference frame.

Pre-processing step can be implemented in advance without prior knowledge on the query protein. In certain application, more than one model structures are processed here. In order to not affect the processing time, generally the pre-processing phase is done off-line. The reference frames populated in the hash table will be used for matching in recognition phase.

Recognition phase includes the same steps as in pre-processing phase, except that the input to this phase is the query structure.

a. Geometrical information is extracted from query structure B;
b. Choose a reference frame; For each reference frame:
c. Compute 3D orthonormal basis associated with this reference frame;
d. Compute the new coordinate of all the other points in the structure as referred to this particular reference frame;
e. Use each coordinate as an address to the hash table, and retrieve all entries at the hash table address;
f. Vote the reference frame that will be aligned to the current reference frame;
g. Repeat the above steps for each query reference frame.

The model structure with the highest votes will be the structure highly similar to the query structure. The main framework has been published in [3]. From the algorithm, two aspects will be described. First is the calculation to derive reference frame, and second is the computation of new coordinates of each point in the structure as referred to as a related reference frame. Next sub-section will provide explanation on protein tertiary structures.

Protein tertiary structure

Protein data can be divided into primary, secondary, tertiary and quaternary structure. Primary structure is represented by string or sequence representation which composed of 20 amino-acids. Protein secondary structure is the local geometry along the sequence, usually in a form of helices, sheets and turns [4]. Tertiary structure retrieved from PDB consists of atomic coordinate information such as 3-D coordinate information. Protein tertiary structure data can be obtained from publicly available online databases such as Protein Data Bank (PDB), CATH, PPSP, SCOP and VAST. For this work we will use PDB as the main data source. PDB is available at http://www.rcsb.org/pdb/. We chose to match the backbone fragment because it is significant enough to characterize the geometrical features [5], [7], and [8].

Structure comparison can be done at different levels such as at residues or secondary structure element (SSE) level. In this work, data used is at the residue level where for each residue we extract only the backbone fragment N-C_α-C. Each atom is represented by a set of coordinates i.e. (x, y, z). The atom and its connection to other atoms can be visualized as point and edge. From Figure 1, the initial 3 atoms form a known triangle which defines a frame. With this frame, we can determine the position and orientation of a residue in space.

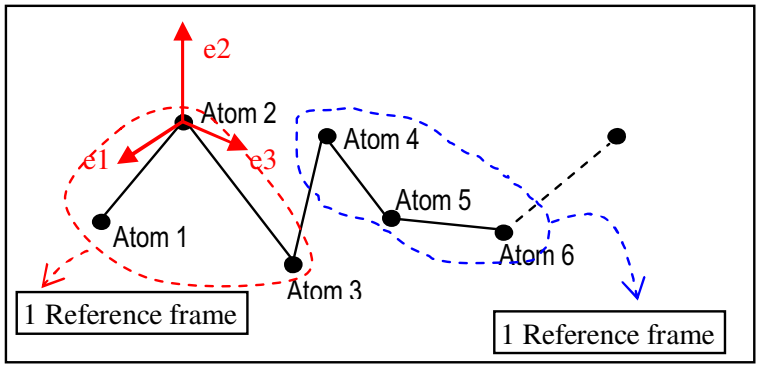

Figure 1: Reference frame formation

3-D REFERENCE FRAME

Reference frame is originally named as *object representation* in computer vision by Lamdan et. al. [1]. A 3D reference frame can be created from any three non-collinear points. One point will be used as an origin, and it links to the other two points used to create a set of three orthonormal vectors as the axes to describe other structure atoms in this particular 3D coordinate system.

This representation must be rich enough to allow reliable distinction between different objects in the database, which later permits efficient matching. Considering the protein backbone fragment, atom nitrogen (N), alpha-carbon (C_α) and carbon (C) will be used as the origin because it is located in the center and connecting to atom N and C. Three orthonormal vectors can be constructed from the links between the origin to its previous and subsequent atoms. This has been applied in previous research by Lamdan et. al [1] who used only C_α and Z. H. Huang et. al. [10] extracted only C_α and C_β to form the backbone. Our matching process will be based on the whole backbone fragment i.e. N-C_α-C from each residue in the structure as applied by [5], [7] and [8].

Calculation

In this section, we first show the calculation of the reference frame RF_1 and second, given the original coordinate S (a,b,c), find the new coordinate for S $(\hat{a},\hat{b},\hat{c})$ in terms of reference frame RF_1. The matching of structures will be based on these new coordinates calculated from each reference frame. Take the first 3 atoms to form a reference frame, namely atom P, Q and R. See Figure 2.

i. Normalize $\overrightarrow{PQ}$ to e_1. It is to change $\overrightarrow{C_\alpha}$ into a unit vector.

$$e_1 = \frac{\overrightarrow{PQ}}{\left\|\overrightarrow{PQ}\right\|}$$

ii. Define e_2 as:

$$e_2 = \frac{v}{\|v\|} \text{ where } \underline{v} = e_1 \times \overrightarrow{PR}$$

iii. $e_3 = e_1 \times e_2$

Once we have basis points $\{e_1, e_2, e_3\}$, the new coordinate for S $(\hat{a}, \hat{b}, \hat{c})$ can be calculated using the equations below:

i. $\hat{a} = \overrightarrow{PS} \cdot e_1$
ii. $\hat{b} = \overrightarrow{PS} \cdot e_2$
iii. $\hat{c} = \overrightarrow{PS} \cdot e_3$

Let say, we have calculated 1 basis for the first residue. Assume the coordinates for each atom are as below:

Atom 1 = (7.407, 11.245, 0.360)
Atom 2 = (7.457, 11.326, 1.841)
Atom 3 = (8.083, 10.146, 2.672)

Then we choose the atom 2 as the origin:

Origin = (7.457, 11.326, 1.841)

By using the calculation above, the value of e_1, e_2 and e_3 should be as below:

e_1 = <-0.0337, -0.0546, -0.9979>
e_2 = <-0.8974, -0.4379, 0.0542>
e_3 = <0.4399, -0.8974, 0.0342>

We calculate one basis for each residue, and then use the bases created to generate coordinates for each atom in a protein. 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. Using this reference frame, we hope to get an acceptable result at par with other matching works.

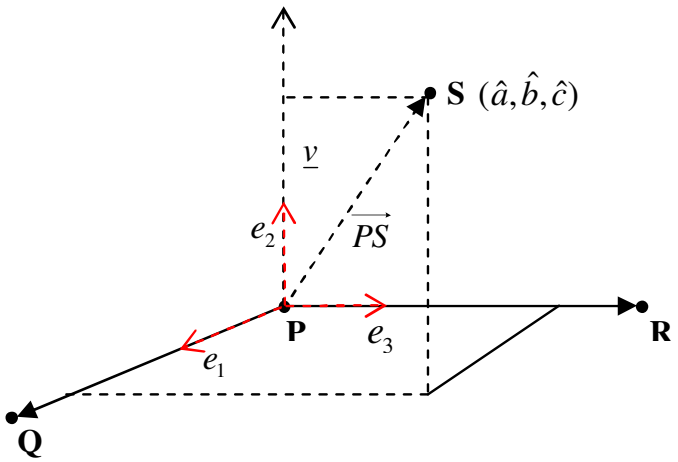

Figure 2: Description of reference frame

RESULTS

We have tested the program on datasets used by S. Canzar et.al. [13] as a benchmark to measure the correctness of the matching algorithm. The datasets are from *seryl* family members. The model structures are 1sry, 1set and 1ses. Meanwhile the target structure is 1ser. According to the results from S. Canzar et. al., when the model structures are matched to the target structures, the similarity score are ranked in the order of 1sry with the highest score, followed by 1set and 1ses. For this experiment we have extracted 60 atoms from each structure, and 20 reference frames are created from these atoms. Table 1 shows the matching result. By using 1ser as target, our program produced 1sry with the highest similarity followed by 1set and 1ses. It is interesting to note that this first benchmark attempted for this work has produced a similar trend to the work of S. Canzar et. al.

TABLE 1:Matching result for seryl family

PDB ID	Votes	Similarity percentage (%)
1ser (target)		
1sry	8915	67.41
1set	8781	66.40
1ses	8754	66.19

DISCUSSION

This is a brute-force style of testing the program, yet the preliminary experiments show promising results. However, the result can be improved by performing an extensive test on a more universal and larger protein data. This extensive test would help us determine whether the reference frame can represent the backbone fragment. Furthermore, the issues of hash table formation may also give an effect on the result especially on the type of hash function used to produce the key that act as the address to the hash table elements and also the selection of the size of the hash table [12]. Possible future work is to improve the processing time by parallelizing the geometric hashing algorithm.

ACKNOWLEDGEMENTS

Research reported here is pursued under the short-term Research Grant by Universiti Sains Malaysia for “Tertiary Protein Structure Matching Based on Geometrical Features” [304/PKOMP/636038].

REFERENCE

[1] Lamdan, Y., Wolfson, H.J., Geometric Hashing: A General and Efficient Model-Based Recognition Scheme”, *In Proceedings of the Second International Conference on Computer Vision*, Tarpon Springs (F1.), 1988, pp.238-249.

[2] Shann-Ching Chen and Tsuhan Chen, Retrieval of 3D Protein Structures, *Proceedings International Conference on Image Processing*, 2002, **3**, pp. 933-936.

[3] Fazilah Othman, Rosni Abdullah, Rosalina A. Salam, Geometric Hashing Algorithm for Protein Tertiary Structure Matching: A Preliminary Study", 1st International Symposium on Bio-Inspired Computing (BIC'05), Puteri Pan Pacific Hotel, Johor Bahru, Malaysia, 5th-7th September 2005.

[4] Bryan Bergeron, 2002, *Bioinformatics Computing*, Prentice Hall PTR, , Chapter 9.

[5] David R. Westhead, J. Howard Parish, and Richard M. Twyman, 2002. Instant Notes: Bioinformatics, The INSTANT NOTES Series, BIOS Scientific Publishers, New York

[6] Chien-Cheng Chen, Jeng-Ting Tu, Pei-Ken Chang, Bing-Yu Chen, Rung-Huei Liang and Ming Ouhyoung, 2004. Protein Function Prediction by Matching 3D Structural Data, *Proceedings of NICOGRAPH International 2004*, Hsinchu, Taiwan, 2004, pp. 113-120.

[7] Murray, K. B., Gorse, D., and Thornton J. M., 2002 . Wavelet Transform for the Characterization and Detection of Repeating Motifs", *Journal of Molecular Biology*, , **316**: pp. 341-363.

[8] Xavier Pennec and Nicholas Ayache, 1998. A Geometric Algorithm to Find Small but Highly Similar 3D Substructures in Proteins", *BIOINFORMATICS*, **14**,6, pp. 516-522.

[9] Marc. A. Marti Renome, Andres Fiser, M.S. Madhu Sudhan, Narayanan Eswar, Ursula Pieper, Min-yi Shen and Andrej Sali, 2003 Modeling Protein Structure from Its Sequence, *Current Protocols in Bioinformatics,* chap. 5.1.1 – 5.1.32.

[10] Zi H Huang, Xiaofang Zhou and Dawei Song, 2005. High Dimensional Indexing for Protein Structure Matching using Bowties", *in Proceedings of 3rd Asia Pacific Bioinformatics Conference (APBC 2005)*, Singapore, 17-21 January 2005, pp. 21-30.

[11] Liisa Holm and Chris Sander, 1993. Protein Structure Comparison by Alignment of Distance Matrices, *J. Mol. Biol.*, **233:** 123-138,.

[12] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein, 2001, *Introduction to Algorithms*, Second Edition, MIT Press and McGraw-Hill, , Chapter 11: Hash Tables, pp. 221-252.

[13] Stefan Canzar and Jan Remy, Shape Distributions and Protein Similarity, http://www.ti.ethz.ch/as/people/remy/preprints/ shape_2006.pdf, 2006.