

A Protein Structure Retrieval System Using 3D LRA

Chan-Yong Park¹, Sung-Hee Park², Dae-Hee Kim³,
Seon-Hee Park⁴, Chi-Jung Hwang⁵

Keywords: protein structure retrieval, protein structure comparison, 3D Locally Relative Angle

1 Introduction.

Since the functions of protein may come from its structure, the method of measuring structural similarity between two proteins can infer their functional closeness [1]. Fast structural comparisons and retrieval methods are necessary to deal with the increasing number of protein structural data via high-throughput structural genomics research.

Many structural comparison methods of proteins have been proposed [2,3,4,5]. Among them, distance matrices, approximation of structure, and vector representation are the most commonly used. The distance matrices, also called distance plots or distance maps, contain all the pair-wise distances between alpha-carbon atoms, i.e. C α atoms of each residue [3]. A matrix is a two-dimensional (2D) representation of a three-dimensional (3D) structure. This representation has critical weak points in terms of computational complexity and sensitivity to errors in the global optimization of alignment. In recent research work, this representation was improved by taking into consideration average conformations, the average coordinate of a small number of contiguous residues, instead of alpha-carbon atoms [4]. The improved representation is an approximation of the protein structure for a quick comparison. It overcomes the weak points on sensitivity to errors and computational complexity.

In this paper, we designed a retrieval model using the 3D Locally Relative Angles(LRA) and described, in detail, all processes in the model. We, then, designed and implemented a protein structure retrieval system based on the 3D LRA. We showed that the retrieval system based on the 3D LRA was relatively fast and accurate in retrieval performance.

2 Protein Structure Retrieval Model Using 3D LRA

The protein structure retrieval using the 3D LRA is a 3D adaptation of the efficient protein structure representation. The model consists of two parts, indexing and retrieval. Indexing and retrieval have almost the same processes that perform 3D LRA. The indexing part stores the 3D LRA with all the proteins in its database while the retrieval part adds the comparison process that compares the 3D LRA of a query to the 3D LRA of all the proteins in the database.

¹ Electronics and Telecommunications Research Institute 161 Gajung-dong, Yusung-gu, Daejeon,305-350, Korea, E mail: cypark@etri.re.kr

² Electronics and Telecommunications Research Institute 161 Gajung-dong, Yusung-gu, Daejeon,305-350, Korea, E mail: sunghhee@etri.re.kr

³ Electronics and Telecommunications Research Institute 161 Gajung-dong, Yusung-gu, Daejeon,305-350, Korea, E mail: dhkim98@etri.re.kr

⁴ Electronics and Telecommunications Research Institute 161 Gajung-dong, Yusung-gu, Daejeon,305-350, Korea, E mail: shp@etri.re.kr

⁵ Dept. of Computer Science, Chung Nam University(CNU), Taejon, KOREA, E-mail: cjhwang@ipl.cnu.ac.kr

The process of generating 3D LRA has following steps. (1) We find all atoms that composed of main chain of protein like Ca, C and N atoms (2) We define these atoms as atom-1, atom-2, ... , and atom-n. (3) We calculate 3D LRA. A figure 1 shows an angle of θ and ϕ in detail. A figure 2 shows equation of these angles. We assume that length(r) of two atoms is 1. The comparison of 3D LRA is similar to string search algorithm. The angle of (θ and ϕ) for each atom is compared with angles of atoms of the other proteins.

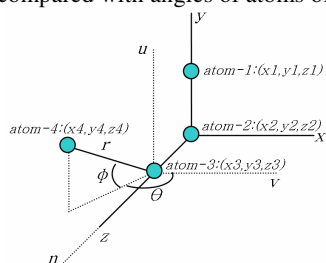


Figure 1: Definition of r , θ and ϕ

$$r = \sqrt{x^2 + y^2 + z^2} = 1$$

$$\theta = \tan^{-1}\left(\frac{y}{x}\right)$$

$$\phi = \sin^{-1}\left(\frac{\sqrt{x^2 + y^2}}{r}\right) = \cos^{-1}\left(\frac{z}{r}\right) = \cos^{-1}(z)$$

Figure 2: Equation of r , θ and ϕ

3 Implementation and Result

Our retrieval system was tested on the system environment Windows XP OS, Pentium 4 CPU 3.0GHz, 2GB main memory as a stand-alone application.

Indexing is a process that generates index files to be used in retrieval processes. The input data of this process are PDB files and the output data is a 3D LRA binary file. In our experiment, 25,000 protein PDB files were used as input data. The total indexing time was 10 minutes. Retrieval is a process that searches for protein structure(s) similar to query protein structures from the index file. We have exploited the Euclidean distance between two LRAs as a similarity measure.

Our empirical study of the protein structure retrieval system using the 3D LRA could lead to very encouraging results. We queried a substructure of protein 1gi1:A and residue index of query structure is between 136 to 153. Table 1 shows the retrieval result. We can also see in the Figure 3 that retrieval results show that they are relatively similar to the query. The average retrieval time for this system is within 10 second.

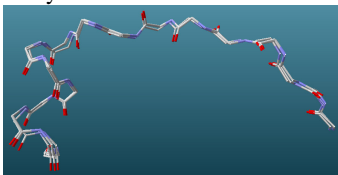


Figure 3. Retrieval result

PDB ID	Residue Index	Length
1gi1:A(query)	136	17
1a0j:A,B,C	136	17
1an1:E	136	17
1anc:A	136	17
1auj:A	136	17

Table 1. Retrieval result

References

- [1] Philip E. Bourne and Helge Weissig: *Structural Bioinformatics*, Wiley-Liss, 2003.
- [2] Taylor, W. and Orengo, C., "Protein structure alignment," *Journal of Molecular Biology*, Vol. 208(1989), pp. 1-22.
- [3] L.Holm and C.Sander, "Protein Structure Comparison by alignment of distance matrices", *Journal of Molecular Biology*, Vol. 233(1993), pp. 123-138.
- [4] Rabian Schwarzer and Itay Lotan, "Approximation of Protein Structure for Fast Similarity Measures", *Proc. 7th Annual International Conference on Research in Computational Molecular Biology(RECOMB) (2003)*, pp. 267-276.
- [5] Amit P. Singh and Douglas L. Brutlag, "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representation", *Proc. Intelligent Systems for Molecular Biology (1993)*.