Protein Structure Alignment by Principle Component Analysis

Sung-Hee Park, Soo-Jun Park, Seon-Hee Park[1]

Keywords: alignment, protein structure, principle component analysis

1 Introduction.

One of the most difficult problems in structural bioinformatics is aligning two or more protein structure geometrically. Our structure alignment method involves the principle component analysis (PCA). PCA is well-known as one of the statistical methods for multivariate analysis. In general, PCA is used to reduce dimensionality of multi dimensional data. However, in this research work, PCA is used to align the two protein structure.

Many protein structure alignment methods have been proposed to compare with protein structures. Among them, one is a method using distance matrix for *alpha carbons of proteins structure*[1] while another is a method using distance matrix for *the mass center of the alpha carbons consisting of a segment* which consist of several residue[2]. That is, the former was improved into the latter. And the other research approach aligns protein structure by vectors of secondary structures.[3] Then it measures a similarity with the vector representation. A recently proposed precise method aligns by incremental combinatorial extension(CE) of the optimal path[4].

Methods mentioned above maximize similarity function to optimize the alignment. They may perform optimization process repeatedly. This is a weak point of above methods.

But, it is the advantage of proposed method that it does not need to optimize similarity function to align structure as other methods.

2 Method.

PCA is a popular statistical method for dimensionality reduction. However, we approach the point of geometrical view of PCA. Geometrically, PCA transforms all atom coordinates on the original coordinate space to those on a new one where the variant with largest variation of covariance matrix is first principle component. Two protein structures aligned by PCA is used to measure similarity.

3 Application: nearest neighbor search.

Here, we used bond line distribution as similarity measure. We named bond line distribution the 3 dimensional edge histogram. 3D edge histogram is a distribution of 10 edge patterns that stands for atom bond.

The flowchart of protein structure comparison by PCA alignment is shown in Fig.1

[1] Bioinformatics Team, Electronic and Telecommunication Research Institute, Daejeon, South Korea. E-mail: {sunghee, psj, shp}@etri.re.kr

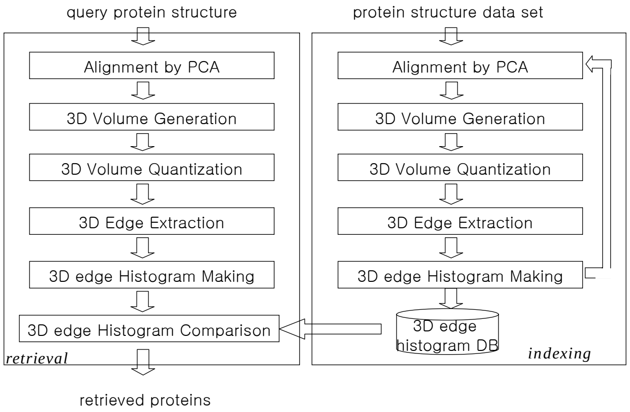

Figure 1: Flowchart of protein structure comparison by PCA alignment.

4 Discussion.

We verified the effectiveness of alignment by PCA through application to nearest neighbor search. We propose a protein structure alignment method of new concept by PCA. The proposed method does not need to optimize similarity function to align structure, while other existing methods may perform optimization process repeatedly. This reduces the time cost for alignment and makes the nearest neighbor search from huge database rapid.

References

[3]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

[1]Lholm and C.Sander, “Protein Structure Comparison by alignment of distance matrices”, *Journal of Molecular Biology*, Vol. 233, pp. 123-138, 1993

[2]Rabian Schwarzer and Itay Lotan, “Approximation of Protein Structure for Fast Similarity Measures”, *Proc. 7th Annual International Conference on Research in Computational Molecular Biology(RECOMB2003)*, pp. 267-276, 2003

[4]Ilya N.Shindyalov and Philip E.Bourne, “Protein structure alignment by incremental combinatorial extension(CE) of the optimal path”, *Protein Engineering* vol.11 no.9 pp.739-747, 1998