An Approach to Three-Dimensional Motif Finding in Proteins

Hiroaki KATO Yoshimasa TAKAHASHI

hiro@molout.tutkie.tut.ac.jp taka@molout.tutkie.tut.ac.jp

Laboratory for Molecular Information Systems,
Department of Knowledge-based Information Engineering,
Toyohashi University of Technology, Toyohashi, Aichi 441 JAPAN

Abstract

This paper describes an approach to three-dimensional(3-D) substructure search using graph-theoretical algorithms, and it's application to the analysis of 3-D structural features of proteins. An abstract representation of protein 3-D structures also devised from this analysis. The details of the approach will be discussed with a couple of illustrative examples that involve the motif search using Protein Data Bank(PDB) files.

1 Introduction

It is well known that 3-D structure of proteins is closely related to the function of itself. And especially, certain particular structural features called motifs which they have specific geometric arrangements within the protein molecules are considered that they are well-reserved sites in the genomic sequences. So that, to find such motifs or 3-D common structural features in more general sense is one of most important problems in genome informatic studies.

In this work, we have investigated an approach to 3-D structural feature search aiming to provide some basis for the analysis of higher structural features of proteins.

2 Method

In the present work, the 3-D structure of a molecule is treated as a set of points that correspond to constituent atoms in 3-D space. The set of points is described by a matrix representation of which each element involves the inter-atomic distance within the molecule. Thus the set of points can be regarded as an edge-weighted complete graph. In other word, we can represent

加藤博明、高橋由雅：豊橋技術科学大学 知識情報工学系、 〒441 豊橋市天伯町字雲雀ヶ丘 1-1

the structural information of a molecule including the 3-D geometry as a weighted (labeled) graph of which the nodes correspond to the constituent atoms. Therefore, the structural feature search can be treated as one of subgraph matching problems. The basic algorithm used here was referred to our previous work[1]. The details can be referred in the reference.

In order to apply the algorithm to structural feature analysis of proteins, an abstract representation of protein structures has been devised. In that way, each amino acid residue within a protein is regarded as a pseudo-atom(super-atom) and its coordinates are approximated by those of Cα of the residues. This approximation can make considerably decrease the size of the graph to be treated. The different amino acid residues are designated by the node-weights for the pseudo-atoms. Physical and chemical properties of the amino acid residues are also available in that way.

3 Result and Discussion

The algorithm was implemented into a 3-D substructure search program, *SS3D*. At first, we prepared a database that consists of ordinary organic molecules, and confirmed the abilities of this algorithm by preliminary experiments with the database. Obviously, the present approach doesn't require a connected substructure as the query of the substructure search, that is we can specify a set of atoms and/or atomic groups, as the query, which the elements have a particular special arrangement in 3-D space. This is a major advantage of the present approach.

Alternatively, we also prepared a 3-D structure database that contains 100 proteins taken from PDB files using the abstract representation mentioned above, and tested for the protein 3-D motif search. The search trial with the query of P-loop(ATP/GTP-binding site motif A; [AG]-x(4)-G-K-[ST]) [2, 3] that consists of eight amino acid residues(6Q21 A G10-S17) correctly found the similar motif structures on three proteins; 1EFM(G18-T25), 1ETU(G18-T25) and 1Q21(G10-S17). The result shows that the present approach is successfully applicable for the 3-D motif search of proteins.

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas '*Genome Informatics*', from The Ministry of Education, Science and Culture of Japan. And we used computer resources at Human Genome Center, Institute of Medical Science, The University of Tokyo.

References

[1] Y.Takahashi, S.Maeda and S.Sasaki : *Anal.Chem.Acta.*, 200, 363-377(1987)

[2] M.Saraste, P.R.Sibbald and A.Wittinghofer : *Trends Biochem.Sci.*, 15, 430-434(1990)

[3] J.E.Walker, M.Saraste, M.J.Runswick and N.J.Gay : *EMBO J.*, 1, 945-951(1982)