Three-Dimensional Motif Search of Proteins Using Abstract Representation of Secondary Structure Segment

Hiroaki KATO　　Yoshimasa TAKAHASHI

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

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

Abstract

This paper describes an approach to three-dimensional(3-D) motif search of proteins, which is based on a graph-theoretical clique finding algorithm. In this implementation, higher abstract representation of a protein structure has been also investigated for the description of secondary structure information such as α-helix and β-strand. The algorithms and the implementations are discussed with a couple of execution examples of the 3-D motif search using protein structure database.

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 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 our previous work, an approach to 3-D substructure search using graph-theoretical algorithms was developed and applied to the analysis of 3-D structural features of proteins using an abstract representation of amino acid residue[1]. In the present paper, further abstract representation of the protein structure which involves the higher structures of α-helix and β-strand has been investigated to establish the 3-D motif search.

2 Method

In the present work, the 3-D structure of a protein molecule is regarded as a set of pseudo-atoms of which the 3-D coordinates are approximated with those of α-carbon(Cα) atoms of the main chain in the same way as the previous work. In addition to this, higher abstract representation has been devised here in order to describe the secondary structure information. Each of the secondary structure segments is represented by two residues which are located in N-

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

and C-terminal side of the segment, and their coordinates are also approximated by those of Cα-atoms of the residues, respectively. In this representation, thus, a secondary structure segment is described as a pair of points (the start-point for N-terminal side and the end-point for C-terminal side) which are located in the terminals of the segment. The assignment of secondary structure is determined using the program DSSP[2]. Two types of secondary structures are considered in the present work: helix(DSSP class: H, G and I) and strand(E). Any residue which is assigned to the other classes of DSSP is regarded as one that belongs to random coil, and it is ignored in this treatment.

To distinguish the start-point and the end-point on the identical segment from those on the other segment, a pseudo-bond between the points on the identical segment is defined. The pseudo-bond is referred as helix-bond for a helix segment or strand-bond for a strand segment. It is possible to distinguish between the pair of points which describes a secondary structure segment and other pairs of points by means of the sign of the corresponded array of the distance matrix, because that each of secondary structure segments is characterized by the value with negative sign. This sophistication, at the matching time, allows us to specify the tolerance of the simple distance between the amino acid residues and the tolerance of the sign of secondary structure segments to be compared, independently.

3 Result and Discussion

We prepared a 3-D structure database that contains 521 proteins taken from PDB files using the abstract representation mentioned above. This dataset was originally selected according to the list of representative protein dataset by Hobohm et.al.[3]. Then we screened the dataset by the following criteria: (i) the structures of all entries in the dataset were determined by X-ray crystallography, (ii) they had the resolution of 2.8Å or better, and (iii) the total residues of each protein was 500 or less. The search trial with the query of 'Crystallins beta and gamma Greek-key' motif [4] that consists of four strands in gamma-b crystallin (eye lens protein; 4GCR: K2-D8, Q12-C18, S34-S39, G60-Y62) correctly found the similar motif sites on other eye lens proteins (2BB2: K2-D8, Q12-C18, S34-S39, G60-Y62 and 2GCR: R89-R95, R99-I105, S123-E128, G149-Y151). And it also identified the similar structural features which consist of four anti-parallel β-strands, in phosphotransferase (1PTF: D66-D60, H7-E2, Y37-D32, N43-K40), hormone/receptor (3HHRC: Y107-S113, G116-F123, L66-R70, K81-E82), and electron transport (2BBKH: A148-D153, A158-D164, T137-Q142, T129-L131). 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, Sports and Culture of Japan.

References

[1] H.Kato and Y.Takahashi : *Proc. Genome Informatics Workshop 1994*, 162-163(1994)

[2] W.Kabsch and C.Sander : *Biopolymers*, 22, 2577-2637(1983)

[3] U.Hobohm, M.Scharf, R.Schneider and C.Sander : *Protein Sci.*, 1, 409-417(1992)

[4] G.Wistow : *J. Mol. Evol.*, 30, 140-145(1990)