Automated Identification of Three-Dimensional Motif in Proteins

Hiroaki Kato
hiro@mis.tutkie.tut.ac.jp

Yoshimasa Takahashi
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 automated identification of three-dimensional (3-D) motif in proteins. Here, the structure of a protein was reduced into abstract representation which consists of the α-helix and β-strand secondary structure elements, these being described by vectors in 3-D space rather than the point-like atoms that are used in the simple Cα approximation. The algorithms and the implementations are discussed with a couple of execution examples of the identification of the 3-D motif candidates using well known motifs.

1 Introduction

In our preceding works, the authors reported the algorithm for the three-dimensional substructure search for ordinary organic molecules, and it's application to the 3-D motif search of proteins[1]. That was validated by computational search trials using several query motifs that are well known[2]. In the present work, we have investigated an approach to automated identification of 3-D motifs (or motif candidates) in proteins based on the maximal common subgraph finding algorithms rather than the 3-D motif search that was based on 3-D substructure search with a certain query substructure specified in advance.

2 Method

In the present work, to avoid the need to consider the thousands of atoms of proteins, the Cα approximations have been used for the implementation of a geometrical searching for the identification of 3-D patterns of atoms that constitute a particular spatial arrangement of certain types of secondary structure. The identification of individual secondary structure segment was carried out with Kabsch & Sander's method[3]. The staring and terminating residues were used for the description of each of the secondary structure segments of a protein. In this manner, α-helix and β-strand secondary structure segments are described by lines in 3-D space. And further reduction of the structure representation was employed, in which each secondary

structure segment described with the line is reduced into a point that is labeled with staring and terminating residues, the length of the segment (the distance between the two residues) and the type of secondary structures. Thus, the whole structure of a protein can be represented by a set of points in 3-D space that involves just secondary structure segments identified within the protein. A maximal common subgraph matching algorithm based on the graph theoretical clique finding procedure[4] was used for the search for the geometrical patterns common to a group of proteins to be tested. The approximation described above allows us to highly reduce the thousands of atoms or points to be searched in the 3-D space.

3 Result and Discussion

Two calcium-binding proteins, Troponin C (PDB code: 1TOP) and Parvalbumin (1PAL), were used for the trial of the automated identification of 3-D motif candidates. The 3-D coordinate data were taken from PDB file and represented using the abstract representation mentioned above. The trial was carried out under the search condition which (1) different kinds of secondary structure segments are distinguished and (2) the tolerance value of the distance is 3.0A. Two orthogonal α-helices within 1TOP (E131-S141 and F151-E159) and those within 1PAL (D79-G89 and V99-I106) were identified one of the maximal common substructures between these two proteins. These detected sites are well known as the two helical segments that forms a EF-hand motif. In the case of the tolerance value of 4.5A, it identified the motif candidate consisting of two anti-parallel β-strands and two α-helices (1TOP: D36-S38, L42-M48, K55-V65 and T72-D74, and 1PAL: F57-I58, F66-N69, D79-G89 and I97-G98).

Next, the 3-D motif search developed in our previous work was carried out with the query of the motif candidate obtained above. The substructure taken from 1TOP according to the second result (4.5A) was used for the search trial with the 3-D protein structure database consisting of 521 proteins. As the result, the system found the corresponding sites in Calmodulin (1CLL), Recoverin (1REC) and other proteins. Some of the detected sites may be related to the calcium-binding sites. These results show that the present approach is successfully applicable for the automated identification of 3-D motif in proteins.

Acknowledgment

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas '*Genome Science*', from The Ministry of Education, Science, Sports and Culture in Japan.

References

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

[2] H.Kato and Y.Takahashi : *Proc. Genome Informatics Workshop 1995*, 132-133(1995)

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

[4] C.Bron and J.Kerbosh : *Communications of the ACM*, 16, 575-577(1973)