Automated Identification of Three-Dimensional Common Structural Features of Proteins

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-cho, Toyohashi 441, Japan

Abstract

This paper describes an approach to automated identification of three-dimensional (3D) common structural features of proteins. The structure of a protein was represented by a set of secondary structure elements (SSEs) in the same manner used in our previous work, where only α-helices and β-strands were considered. The maximal common subgraph matching algorithm, based on a graph theoretical clique finding approach, was used to identify the 3D common structural features for a pair of proteins. The program called AIM (Automated Identification of 3D Motif in proteins) was developed and tested by the execution trials for finding the secondary structure segments related to the Rossmann-fold motif as a 3D common structural feature between alcohol dehydrogenase and lactate dehydrogenase, both of which are known to have the motif site. The result of a substructure search for a protein structure database using the 3D structural feature that was identified will also be discussed.

1 Introduction

Structural feature analysis or similarity analysis of proteins can give us a lot of useful information for molecular biological science and related areas. With the rapidly increasing number of proteins of which 3D structures are known, efficient approaches are required for a systematic analysis of the 3D structural feature of proteins. In our preceding works, the authors have investigated the approaches to a 3D substructure search or motif search of proteins. A computer program, called SS3D-P, has been developed for a 3D substructure search of proteins [1]. The program allows us to identify all occurrences of a user defined 3D query pattern consisting not only of chain-based peptide segments but also of a set of disconnected amino acid residues at the residue level for the protein structure databases. The authors also reported another program called SS3D-P2 for a 3D structure motif search of proteins based on the SSEs [2]. In this paper, we extend these studies and describe a computer program, AIM, which can be useful for finding of new motifs or 3D common structural features for a pair of proteins without any query substructures 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 3D patterns of atoms that constitute a particular spatial arrangement of certain types of secondary structure. The identification of individual secondary structure segments was carried out using Kabsch and Sander's method [3]. In this manner, α-helix and β-strand secondary structure segments are described by vectors in 3D space. Further reduction of a structure representation is employed, in which each secondary structure segment described with a vector is reduced into a conceptual point that is labeled with the starting and the ending residues, a length of the segment (i.e. a distance between

the two residues) and a type of secondary structure. Thus, the whole structure of a protein can be represented by a set of the conceptual points that involve only the secondary structure segments identified within a protein. This set of points can be regarded as a graph in mathematical graph theory. The approximation described above allows us to highly reduce a protein structure which consists of thousands atoms or points in 3D space. The maximal common subgraph matching algorithm based on a graph theoretical clique finding procedure [4] was used for the search for the geometrical patterns which are common in the proteins under investigation. The basic procedure consists of two major parts: (i) the generation of the docking graph from two protein molecular graphs obtained by the method just mentioned; (ii) the identification of the maximal clique(s) of the docking graph. These processes were computerized and implemented on a program called AIM.

3 Results and discussion

To test the performance of our program AIM, alcohol dehydrogenase (1DHXA) and lactate dehydrogenase (9LDTA) were used for the trial of the automated identification of 3D common structural features. It is known that they are NAD-dependent dehydrogenases and have typical Rossmann-fold motif [5]. The 3D coordinate data were taken from the PDB (Protein Data Bank) file and represented using the reduced representation described above. The trial was carried out using the search conditions where the different kinds of secondary structure segments were distinguished, the directions of each secondary structure segment on the primary sequence were considered, and the tolerance value of the distance was set at 5.0Å. For these two proteins, AIM identified five distinct patterns as the common structural features that are maximal in terms of the number of SSEs. Six parallel β-strands within 1HDXA (T194-F198, R218-V222, E239-I241, F264-E267, T288-I291 and T313-G316) and those within 9LDTA (L93-I96, K134-V137, V160-G162, K23-V27, E48-V52 and K77-G81) were identified as one of the maximal common substructures of these proteins at the SSE representation level. These sites correspond to the segments that form a Rossmann-fold motif known as a NAD-binding domain.

Subsequently, we tried a substructure searching for the protein structure database using the 3D structural feature identified above. The search was carried out by the use of a 3D substructure search program, SS3D-P2, which was developed in our previous work [2]. The database that contains 521 proteins taken from the PDB file was used for this computational trial. The geometry of the 3D query pattern was based on the site identified for 1HDXA. In addition to 1HDXA and 9LDTA, the program successfully found the corresponding sites in d-glycerate dehydrogenase (1GDHA), malate dehydrogenase (2CMD) and others that are known to have the Rossmann-fold motif. These results show that the present approach is successfully applicable to finding the motif candidate as a 3D common structural feature of proteins.

Acknowledgments

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 of Japan.

References

[1] H.Kato and Y.Takahashi, *Bull. Chem. Soc. Jpn.*, **70**, 1523–1529 (1997).

[2] H.Kato and Y.Takahashi, *Comput. Applic. Biosci.*, in press.

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

[4] C.Bron and J.Kerbosh, *Commun. ACM*, **16**, 575–577 (1973).

[5] M.G.Rossmann, D.Moras and K.W.Olsen, *Nature*, **250**, 194–199 (1974).