Development of an Automated Identification Program for Three-Dimensional Protein Motifs

Hiroaki Kato hiro@cilab.tutkie.tut.ac.jp
Yoshimasa Takahashi taka@mis.tutkie.tut.ac.jp
Hidetsugu Abe abe@cilab.tutkie.tut.ac.jp
Department of Knowledge-based Information Engineering, Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi 441-8580, Japan

1 Introduction

Structural feature analysis or similarity analysis of proteins provides us a lot of useful information for better understanding of molecular biological science and the related areas. In our preceding works, the authors investigated the approaches to the 3D motif search of proteins [2, 3] and also reported the automated finding of new motif candidates or 3D common structural features of proteins in which it doesn't require any query pattern (or substructure) specified in advance [4]. However, the latter is only for a pair of proteins. In this paper, we describe the extension of our approach for three or more proteins.

2 Method

In principle, maximal common structural feature searching among three or more proteins can be done by the simple extension that for a pair of protein molecular graphs, which is based on clique finding algorithm [1] with a docking graph technique reported in our previous work [4]. In the case, however, the docking graph to be considered is expected to have a large number of vertexes. Because, the number of vertexes of the docking graph increases in proportional to R^n: R is the number of amino acid residues or the polypeptide segments to be considered, and n is the number of proteins. The matter means that it requires a large amount of memories, and it easily reaches to combinatorial explosion in the clique finding process. To avoid this difficulty, in this work, a heuristic approach is employed for finding motif candidates in such cases. Here, at first a reference protein and another chosen from the database are submitted for finding clique candidates of the entire set of proteins, and then the candidates obtained are used as queries in the following process of 3D pattern matching. For every candidate pattern it is examined one-by-one by means of 3D substructure searching technique whether the pattern is hit in all of other proteins. If the pattern failed to match with all the proteins remained then it removed from the candidates. The candidates that include the failure pattern are also ignored in the process followed by. All the patterns survived in this trial are resulted that they are the candidates of 3D motifs of these proteins. These processes were fully automated with a computer program, called AIM, which includes WWW-based interface for displaying the results obtained.

3 Results and discussion

To test the performance of our program AIM, *alcohol dehydrogenase* (1CDOA), *lactate dehydrogenase* (9LDTA) and *glyceraldehyde-3-phosphate dehydrogenase* (1CERO) were used in a search trial for the automated identification of 3D common structural features. These proteins are known as NAD-dependent dehydrogenases and have typical Rossmann-fold motif [5]. The search trial was carried out under the searching conditions that the different kinds of secondary structure elements (SSEs) were distinguished, the directions of each SSE segment on the primary sequence were considered,

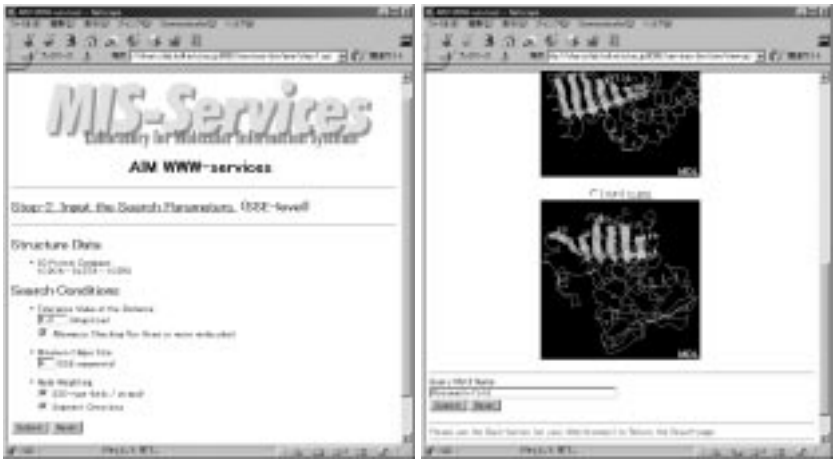

Figure 1: Snapshots of the WWW interface of AIM.

and the tolerance value of the distance was set at 6.0Å. For these three proteins, AIM found five distinct patterns as their common structural features that are maximal in terms of the number of SSEs included. Six parallel β-strands were identified as one of the maximal common substructures of these proteins (1CDOA: T195-F199, R219-V223, D240-V242, F265-E268, V289-L292, T313-G316, 9LDTA: L93-I96, K134-V137, V160-G162, K23-V27, E48-V52, K77-G81, 1CERO: V89-E94, K115-I118, I143-S145, K2-N6, V27-N31, K69-T74). These sites correspond to part of the polypeptide 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 found by the AIM. The searching was carried out by the use of a 3D substructure search program SS3D-P2 [3]. The database that contains 521 proteins taken from the PDB was used for this trial. The geometry of the 3D query pattern was based on that of the site identified for 1CDOA. Our program successfully found the corresponding sites in *d-glycerate dehydrogenase* (1GDHA), *malate dehydrogenase* (2CMD) and others that are known to have a 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 partially supported by a Grant-in-Aid for Encouragement of Young Scientists, from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Bron, C. and Kerbosh, J., Finding all cliques of an undirected graph, *Commun. ACM*, 16:575–577, 1973.

[2] KatoH. and Takahashi, Y., Three-dimensional structural feature search of proteins, *Bull. Chem. Soc. Jpn.*, 70:1523–1529, 1997.

[3] KatoH. and Takahashi, Y., SS3D-P2: a three-dimensional substructure search program for protein motifs based on secondary structure elements, *Comput. Applic. Biosci.*, 13:593–600, 1997.

[4] KatoH. and Takahashi, Y., Automated identification of three-dimensional common structural features of proteins, *Genome Informatics 1997*, 296–297, 1997.

[5] Rossmann, M. G., Moras, D. and Olsen, K. W., Chemical and biological evolution of a nucleotide-binding protein, *Nature*, 250:194–199, 1974.