Development of HocDB (Homology-Based Clustering DataBase), a Sequence Classification System for Database Searching

Motokazu Ishikawa¹ Yoshinori Sato¹
m.ishikawa@cbo.mss.co.jp jin-sato@cbo.mss.co.jp

Hiroyuki Toh²
toh@kuicr.kyoto-u.ac.jp

¹ Mitsubishi Space Software Co., Ltd., 5-4-36 Tsukaguchi-honmachi, Amagasaki, Hyogo 661-0001, Japan
² Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: sequence database, clustering, BLAST

1 Introduction

Recent growth of amino acid sequence databases is remarkable. The increase of the entries provides us enormous amount of information. At the same time, however, the growth of databases decreases the efficiency in prediction of structure and function of proteins by database searching. In other words, redundancy in the database generates the problem. For example, the top 1,000 sequences in the output list of the searching through nr database with PSI-BLAST by using a HIV protease as a query are occupied by close relatives of the query HIV protease. The tertiary structures of the HIV protease and the distant relative, aspartic acid protease, are available now. But, let’s assume that we don’t know the tertiary structure or catalytic mechanism of HIV protease. We can infer the structure and catalytic mechanism of the viral protease if we can detect sequence similarity between the viral protease and the aspartic acid protease by database searching. As told above, however, the high redundancy of the viral proteases inhibits the detection of such sequence similarity. In addition, recent progress of genome projects have increased the entries of ORF products, whose functions have not been identified, in databases. Detection of such sequences is not useful for the prediction of structure and function of proteins.

To address the problems, we have developed HocDB, a web-based sequence classification system for efficient database searching. Through this system, gapped BLAST [1] output is edited in two processes; reducing the redundancy and adding secondary information.

2 Method and Results

2.1 System Architecture

HocDB system edits an output of gapped BLAST search as depicted in Figure 1. At first, BLAST search is executed against sequence database (Fig. 1-a). Next, the detected sequences are summarized into several sequence clusters, according to the classification databases which we have originally developed (Fig. 1-b). The readability of the edited output is higher than the original output list, due to the decrease of the redundancy. Moreover, the system fetches secondary information from Swiss-Prot (comments and feature table data), PIR, PDB and SCOP, and appends such information to the edited
The information appending system enables the users more efficiently to obtain the structural and functional information for their query sequences.

### 2.2 Sequence Classification Algorithm

To build a clustering database, we have classified whole Swiss-Prot sequence entries based on their sequence homology. At first, all-against-all gapped BLAST searches for Swiss-Prot sequences data are carried out. Then, if two HSPs (High-scoring Segment Pairs) corresponds to the same region of a sequence, and if the deviations of the start and end points between the two HSPs are smaller than 13 a.a., these HSPs are classified into a cluster (Figure 2). Then, all the entry sequences of SwissProt (Ver 41) have been classified into 39,948 clusters by using the criterion.

### 3 Acknowledgments

This study was performed through the IT-program of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

### References