Identification of Ortholog Groups in KEGG/SSDB by Considering Domain Structures

Masumi Itoh itoh@kuicr.kyoto-u.ac.jp
Akihiro Nakaya nakaya@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Keywords: ortholog, clustering, domain extraction

1 Introduction

Huge amount of genome information is stored in databases with the advent of recent genome projects. Although we can effectively predict protein sequences from these genomes, functions of most proteins are not experimentally determined. Therefore computational methods are most important for the function prediction, based on comparison and clustering of protein sequences. However, complications arise from the fact that the unit of conservation is not entire protein molecules but domains which are parts of the protein molecule. Hence a method to classify proteins according to their domain structures must be developed for use in functional predictions. Here, we propose a method for extracting domain information from a cluster of similar proteins obtained by all to all pairwise sequence comparisons of completely sequenced genomes.

2 Material and Method

The KEGG/SSDB database contains Smith-Waterman similarity scores of about 100,000,000 pairs from 350,000 proteins in 100 genomes of KEGG/GENES [2]. Our method performs domain extraction and fine protein clustering for a given group of similar proteins by the following procedures.

1. Construction of similarity profiles for each residues:

   Extract one protein (target protein) from the group and compare against all other proteins. Construct a bit vector for each residue of the target protein, in which the bit is one if the residue is aligned to the corresponding protein or zero if not aligned (Fig. 1).

2. Self-comparison of similarity profiles

   Calculate similarity scores, which is defined by the equation in bottom of Figure 1, for all amino acid pairs within the target protein (Fig. 2).

3. Extraction of domains

   Detect the position where the similarity score between the current position and the next position becomes higher than the similarity score between the current position and the precedng position along the amino acid sequence. This is considered as a boundary of domain candidates.

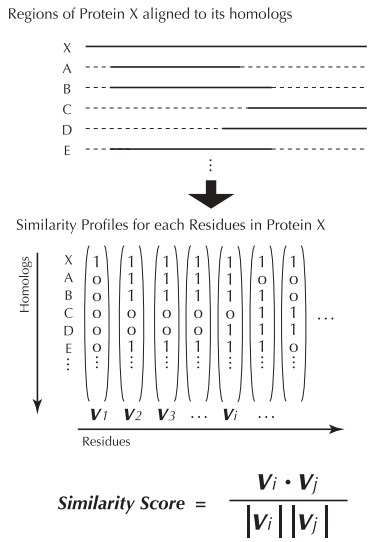

Figure 1: Similarity profile and similarity score.

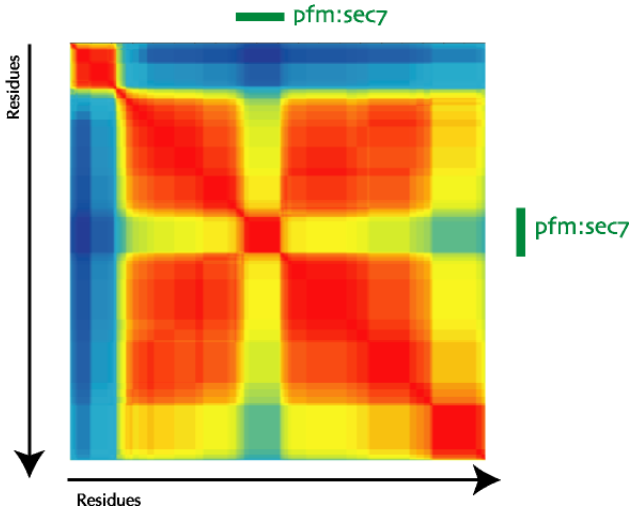

Figure 2: Example of Self-comparison of similarity profiles.

4. Clustering of domain candidates

   Repeat the above procedures for every other protein in the group. For all domain candidates identified, select a representative bit vector among all vectors constituting each candidate, and calculate the similarity score between domain candidates using the vectors. Then, perform clustering of domain candidates.

5. Classification of proteins in the group by domain organization

   Each protein in the group is now associated with a series of domain identifiers representing similar domain clusters. This information can be used to classify proteins in the group with various measures of similarity of domain structures.

3 Results and Discussion

Figure 2 shows an example of protein sce: YDL170C where five domains are extracted using our methods. Vertical and horizontal axes of the matrix show the amino acid position compared with itself, and coloring (tone) of the matrix shows the similarity score between two amino acid positions. This protein has SEC7 domain which functions as GTPase exchange factor (GEF) of Arf small GTPase. Lines on top and right of the matrix show a region of SEC7 domain annotated by Pfam [1]. We could also find another new domain (conserved unit), which is not included in Pfam, at upper left of the matrix. In summary, we could successfully classify domains and proteins using the method presented here. Since our method can be used to refine clusters of similar sequences, we are applying it to large scale date sets extracted from KEGG/SSDB.

References

[1] Bateman, A. *et al.*, The Pfam protein families database, *Nucleic Acids Res.*, 30:276–280, 2002.

[2] Kanehisa, M. *et al.*, The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30:42–46, 2002.