Automatic cDNA Classification System for Mouse Genome Project

Yoshifumi Fukunishi [12] **Hideaki Konno** [12] **Yoshihide Hayashizaki** [1]
fukunisi@rtc.riken.go.jp hkonno@rtc.riken.go.jp yosihide@rtc.riken.go.jp

[1] Genome Science Lab, RIKEN Life Science Tsukuba Center
3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan
[2] CREST, Japan Science and Technology Corporation (JST)

1 Introduction

Classification of cDNA/protein sequence data is necessary for analysis and further application of sequence data, i.e. understanding of pathway of proteins in a cell and signal transduction. Since rapid increase of the number of cDNA/protein sequences is expected, the procedure of the classification must be systematic and automated to avoid human eye inspection. In this study, the sequence data are classified based on two terms; subcellular location and function. Again, false positive/false negative results are always expected in the use of prediction method, and the human inspection is necessary to avoid this problem. We proposed a method to reduce the false positive for PROSITE motif search by using a structural information.

2 Method

2.1 Classification by subcellular location

The classification is done by two methods; one is a database search and another is an analysis of signal peptide. The first step is a FASTA homology search [5] based on the data files in which the entry proteins were classified by subcellular location. The data files for nuclear consists of the sequences of 2353 nuclear localized proteins, that for cytoplasm consists of the sequences of 1101 cytoplasmic proteins, etc. The same proteins from different species are removed from the data files to reduce the redundancy of proteins. The second step is the searching signal peptide and the checking physical property of peptide sequence in question (Fig. 1). PSORT II [4] and GCG are used to check the signal peptide, and TopPred [3] is used to check the existence of the transmembrane domain. The final result is given by the linear combination of the scores calculated by these programs, and the parameters for weighting were optimized to give the best prediction result for 275 known proteins.

2.2 Classification by function: PROSITE motif search

PROSITE [1] is one of the most useful database to find the functional domain in protein, and the number of entries is more than 1300. Some rare motifs consist of 20 amino acids while the most frequently found motif consists of only 2 amino acids. We will find false positive PROSITE motifs every 10 amino acids in a protein in question, even if the protein includes only a few real motifs. One method to avoid the false positive motifs is using a subset of PROSITE which consists of only rare motifs. This method always ignores some motifs of small number of amino acids. We assume that the motif can function only in the suitable structural environment. The ATP-binding P-loop motif can bind an ATP only when the secondary structure of ATP-binding site is loop and the site is accessible for solvent molecules. We prepared a data file of secondary structure and accessibility of amino acid for each entries of PROSITE. After the PROSITE motif search, GOR4 [2] is used to predict the secondary structure of query protein. If the secondary structure of the predicted region of the query protein matches the secondary structure of the motif in PROSITE database, we can expect that the motif is functional (Fig. 2). Otherwise the motif is false positive.

Acknowledgements

This study has been supported by Special Coordination Funds and a Research Grant for the Genome Exploration Research Project from the Science and Technology Agency of the Japanese Government, CREST (Core Research for Evolutional Science and Technology) of Japan Science and Technology Corporation (JST), and a Grant-in-Aid for Scientific Research on Priority Areas and Human Genome Program from the Ministry of Education and Culture, Japan to Y.H.

References

[1] Bairoch, A., Bucher, P., and Hofmann, K., The PROSITE database, its status in 1997, *Nucleic Acids Res.*, 24:217–221, 1997.

[2] Garnier, J., Gibrat, J.F., and Robson, B., Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *J. Mol. Biol.*, 120:97–120, 1978.

[3] Heijne, G., Membrane Protein Structure Prediction, Hydrophobicity Analysis and the Positive-inside Rule, *J. Mol. Biol.*, 225:487–494, 1992.

[4] Horton, P. and Nakai, K., Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier, *Intelligent Systems for Molecular Biology*, 5:147–152, 1997.

[5] Pearson, W.R. and Lipman, D.J., Improved tools for biological sequence analysis, *Pro. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.

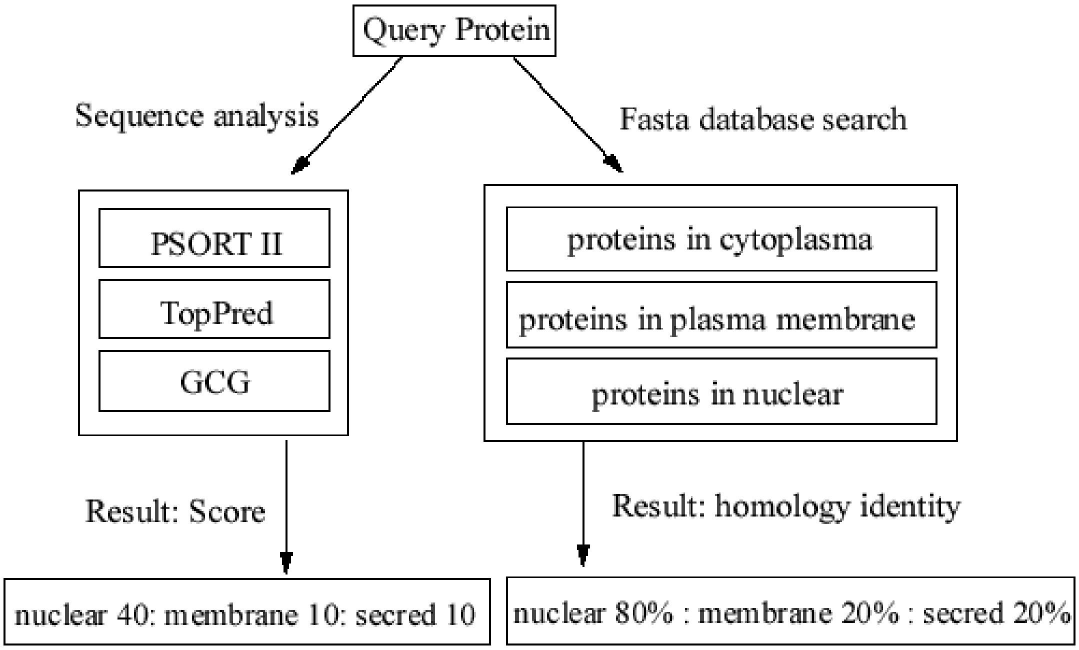

Figure 1: Automatic classification by protein subcellular location.

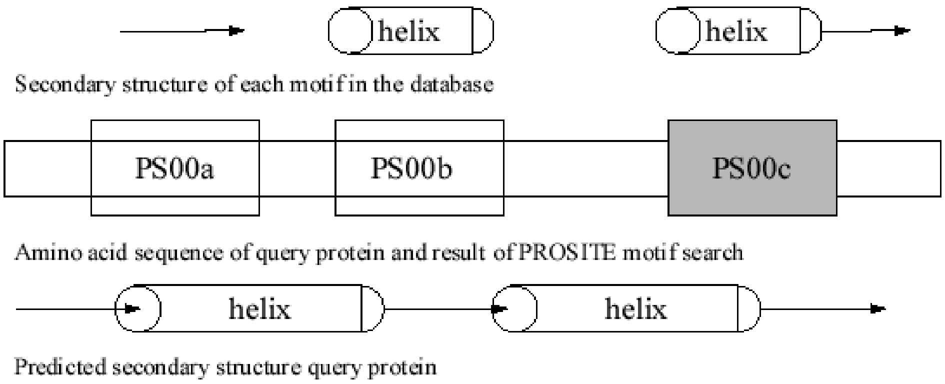

Figure 2: Schematic representation of functional-motif selection method. PS00c is chosen as a functional motif while PS00a and PS00b are false positive.