Prediction of Secondary Metabolism Pathways from Genomic Information and Compounds Data

Atsushi Ogiwara
ao57473@GlaxoWellcome.Co.UK
Bioinformatics Unit, Scientific Computing Department
Tsukuba Research Laboratories, Glaxo Wellcome K.K.
43 Wadai, Tsukuba 300-4247, Japan

1 Introduction

Genome sequencing projects have determined complete genomic sequences of more than 20 organisms. Now we obtain catalogs of complete gene sets of these species. But we really want to know is not only catalogs of each gene or protein. What we want to know is how they express biological functions. Biological function can be understood only by the overall behavior of an organic system. Thus we must understand how biological system is constructed from elemental genes and molecules and the interaction of these molecules.

Pathway informatics or process informatics is a methodology to understand the construction and the behavior of the living system based on informatics. The aim of pathway informatics includes the construction of pathway databases like KEGG [5], the development of pathway reconstruction method *in silico*, and the simulation of biological system using above databases and methods.

To achieve the third goal, our current knowledge is insufficient. Limiting to the metabolic system, we relatively well know about primary metabolism, but we have little knowledge about secondary metabolism. Drug metabolism is one of the example we want to reveal, for we must develop drugs with good efficacy and least toxicity.

Here we propose a method to predict uncharacterized metabolic pathways from genetic information and compounds data.

2 Method

To predict unknown metabolic pathways, we use three types of data:

1. Differential gene expression (DGE) profiles by adding target compounds to a cell

2. Sequence information of genes detected by the DGE experiment

3. Structure of compounds used by the DGE experiment

The basic ideas are as follows:

1. From DGE data, we can pick up related genes to the target metabolic pathway. Also there would be other kind of genes concerned to regulation, signal transduction, and transport pathways. Since genes of the same functional category tends to have similar expression patterns [2], we would be able to roughly classify these genes according to expression patterns

2. Sequences of genes detected by the DGE experiment are analyzed using homology, motifs and profiles. Candidate enzymes can be selected and roughly classified according to the types of enzymatic reactions.

3. Once a list of all possible enzymes and structure of a target compound are given, we can enumerate all possible products by tracing the correspondence of atoms between substrates and products. The structures of substrate and product must be exactly overlaid only except for the group that must be transfered or modified by one of the given enzymes.

Before starting the DGE experiment, we are making software tools to process the data generated by the above-mentioned strategy.

3 Characterization of related genes from gene expression profiles

We made a program to cluster expression patterns by average linkage clustering method like Brown *et al.* [3]. Unlike their method, our program can treat not only the correlation coefficient but also the Euclid distance or other kind of measures. We applied the program to the published data of the diauxic shift of yeast [2], and found the Euclid distance was better in some cases especially the expression looked relatively stationary.

4 Characterization of related genes from sequence information

As the first step of the classification of enzymes that are characteristic in drug metabolism, we tried to classify cytochrome P450 [4]. We firstly performed phylogenetical analysis using well classified P450 sequences in SWISS-PROT, and then made alignment blocks for each subfamilies. These blocks are used to classify uncharacterized P450 candidate genes.

5 Inference of metabolism pathways from compound structures

We adopted Arita's method [1] to trace atom-to-atom correspondence between the initial substrate and candidate products. We extracted structural data of already published compounds from our in-house compounds database, and tried to trace the chain of substrate - product relationship.

Acknowledgments

This work is supported by a "Genome Frontier Pioneering Program" from Science and Technology Agency in Japan.

References

[1] Arita, M., *Automatic Metabolic Reconstruction*, *PhD thesis*, The University of Tokyo, 1999.

[2] DeRisi, J.L., Iyer, V.R., and Brown, P. O., Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.

[3] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[4] Lewis, D. F., Watson, E., and Lake, B.G., Evolution of the cytochrome P450 superfamily: sequence alignments and pharmacogenetics, *Mutation Research*, 410:245–270, 1998.

[5] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto encyclopedia of genes and genomes, *Nucl. Acids Res.*, 27:29–34, 1999.