A new approach to analyzing DNA microarray data using principal component analysis

Jun-Ichi SAGARA\textsuperscript{1} Hideaki KOIKE\textsuperscript{2} Yasunobu TERABAYASHI\textsuperscript{2} 
Motoaki SANO\textsuperscript{1} Noriko YAMANE\textsuperscript{2} Junichiro MARUI\textsuperscript{2} 
Mitsuko DOHMOTO\textsuperscript{1} Ken ODA\textsuperscript{1} Eiji OSHIMA\textsuperscript{3} 
Kuniharu TACHIBANA\textsuperscript{3} Yoshimasu HIGA\textsuperscript{3} 
Shinichi OHASHI\textsuperscript{1} Masayuki MACHIDA\textsuperscript{12} 

\textsuperscript{1} Genome Biotechnology Laboratory, Kanazawa Institute of Technology, 3-1 Yatsukaho, Hakusan-shi, Ishikawa 920-1154, Japan
\textsuperscript{2} Research Institute for Cell Engineering, National Institute of Advanced Industrial Science and Technology (AIST), Higashi 1-1-1, Tsukuba, Ibaraki 305-8566, Japan
\textsuperscript{3} Sansho Seiyaku Co., Ltd., Fukuoka, Japan

Keywords: Aspergillus Oryzae, expression data mining principal component analysis

1 Introduction

DNA microarray experiments allow for the measurement of genome-wide expression data and present information on this genomic scale. Now that vast amounts of gene expression data have been collected, there are problems that the resultant mass of data are possibly irrelevant, insignificant or redundant. Therefore, we should carefully select methods for analyzing and representing microarray data.

Then, we have developed a method using the principal component analysis to detect weak correlation comparing low-level principal ingredients between gene and experimental conditions and applied to microarray data of Aspergillus oryzae RIB40 which are compared to two set of expression pattern and detected that genes which are related to producing kojic acid in Aspergillus oryzae.

Figure 1: LEFT: Microarray data profiles, RIGHT: Genes and conditions matrix.
2 Method and Results

Our method is based on principal component analysis and multi-dimensional scaling analysis. We defined alignment matrix $F$, each row of which is a gene vector $F^k$ for the $k$th gene of microarray data set. Each condition of expressions is encoded to a 3-bit binary number (conditions of expressions are respectively encoded to 100, 010, 001 and 000). A gene vector consists of 1s and 0s, and corresponds to a point in $4l$-dimensional space, where $l$ is the length of the sequence. The number $C^{kk'}$ of matched masses between genes $k$ and $k'$ can be expressed as the inner product of the gene vector. A comparison matrix $C$, each element of which is the number of matches for all pairs of genes can thus be expressed as the matrix product between alignment $F$ and its transpose $F^T$. The principal axes $\vec{u}_p$ are defined as $C\vec{u}_p = \lambda_p \vec{u}_p$, where $u_p$ is an eigenvector and $\lambda_p$ is the corresponding eigenvalue of comparison matrix $C$. Each genes is plotted on the two-dimensional plane called gene space. The coordinate $x^k_p$ of gene $k$ in dimension $p$ is given by $x^k_p = \lambda_p \vec{u}^k_p$. Genes are classified into one or more groups, according to the distance between the two-dimensional gene plots. The coordinates $y_p$ of condition in the sequence are given by $y_p = F^T \vec{u}_p$. The $i$th element of $y_p$ corresponds to a condition at position $i$ in the gene, and characteristic conditions of each group are detected by comparing the conditions with the group of genes.

We applied our method to microarray data of Aspergillus oryzae RIB40 which are compared to two set of expression pattern. We have detected two genes of unknow function in Aspergillus oryzae microarray data. These two genes are located close in the genome showing typical characteristics of genes involved in secondary metabolism.

References

