An Approach to Identify Unknown Gene Function through Higher Order Correlations for Gene Clusters

Akira Funae\textsuperscript{1} \texttt{funae@hal.rcast.u-tokyo.ac.jp}
Daisuke Komura\textsuperscript{2} \texttt{komura@hal.rcast.u-tokyo.ac.jp}
Hiroshi Nakamura\textsuperscript{1,2} \texttt{nakamura@hal.rcast.u-tokyo.ac.jp}
Sigeo Ihara\textsuperscript{2} \texttt{ihara@genome.rcast.u-tokyo.ac.jp}

\textsuperscript{1} Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
\textsuperscript{2} Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8904, Japan

Keywords: gene expression, 2nd-order correlation, prediction of gene functions

1 Introduction

Previous analyses of microarray data are focusing on similarity of gene expression pattern to classify genes into same function. However, among genes involved in the same pathway, many genes do not show similar expression profiles. Therefore, previous analyses can not collect genes with same function in such cases. Zhou et al. \cite{1} tried to overcome the above difficulty by calculating correlation of expression pattern of every pairwise genes in each data set and calculating correlations of those correlations across data sets measured under different conditions. They termed this approach as '2nd-order analysis', which identify genes of the same function even though genes don’t have similar expression patterns.

In this study, to predict unknown gene functions \textit{ab initio} more efficiently, we introduce an approach that captures 2nd-order relationships among gene clusters rather than pairs of genes. The complex functions of a living cell are carried out through the concerted activity of many genes and gene products. Therefore, it is useful to construct functional modules consisting of many genes, with gene clusters, to identify unknown gene functions. We propose an efficient approach that can identify functional modules by combining clustering and maximum clique method. Applying our approach to \textit{Saccharomyces cerevisiae} data sets, we found that our approach can identify functional modules with high accuracy. These results indicates that our approach is promising.

2 Method

We propose a method of constructing accurate functional modules and extracting 2nd-order relationship between modules. The overview of our approach is shown in Figure 1.

3 Results and Discussion

We applied our method to 36 \textit{Saccharomyces cerevisiae} microarray data sets from SMD, NCBI Gene Expression Omnibus and Rosetta compendium data. Each microarray consists of 6178 genes and each data set consists of more than 8 arrays.

We evaluated quality of 1st-order analysis by determining the percentage of functionally homogeneous modules based on the Gene Ontology (GO) biological process annotation using DAVID software \cite{3}. We found 84% of modules constructed by our approach are functionally homogeneous.
Step 1 We calculate expression correlation of all pairwise genes in each data set and remove pairwise genes with low value across all data sets.

Step 2 We cluster pairwise genes based on their correlation pattern across data sets. Pairwise genes in a same cluster have similar 2nd-order correlation.

Step 3 We describe genes as node and pairwise genes in a same cluster are connected in undirected edge. We search quasi cliques in each cluster. We use Xl’s algorithm described in [2] to search quasi cliques.

Step 4 We set average correlation of all pairwise genes in a module as confidence score and calculate average of correlation pattern. Based on such values, we select accurate functional modules and modules with 2nd-order relationship.

Figure 1: Overview of our approach

To evaluate the quality of 2nd-order analysis, we generated Receiver Operating Characteristic (ROC) curve and calculated area under ROC curve (AUC). Figure 2 shows ROC curve - a graph of the true positive rate vs false positive rate as 2nd-order correlation threshold varied. True positive cases indicate that the two modules were correctly identified to share same function; on the other hand, false positive cases indicate that they were wrongly identified. The value of AUC was 0.70 and is far above 0.50 of that of rancom predictor. This result indicates that 2nd-order analysis is useful in making prediction of gene functions.

Moreover, we predicted functions of unknown genes involved in the modules we constructed. 78% and 89% of unknown genes are predicted by only 1st-order analysis and by both 1st and 2nd-order analysis, respectively.

4 Conclusion

We proposed an approach to predict gene functions through 2nd-order correlation for gene clusters. The result indicates that our approach is promising in predicting unknown gene functions.

References

