Gene Set Enrichment Analysis for Time-series

Gene Expression Profile

Yuta Okuma, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Graduate School of Information, and Science, Technology, Osaka University

Abstract—Gene Set Enrichment Analysis (GSEA) is a method of analyzing microarray data that can be used to determine whether a microarray data set indicates significant biological changes in the expression of a biological pathway or an a priori-defined set of genes. However, GSEA cannot be applied to long-term time series data because of the multiple time points and seamless nature of such data, which prevent their division into discrete subsets.

Therefore, we have developed a new GSEA method of analyzing time series data that is based on a nonparametric statistical analysis model.

Using the method, we can detect periods of gene expression that cannot be found using the conventional GSEA, i.e., our method can identify the significant expression period for each gene set.

I. BACKGROUND

All living things have genes in their body, and control gene expression to generate RNA and proteins. The analysis of gene expression with DNA microarray is valuable for elucidating biological processes[1]. With respect to gene expression analyses, time series data can be meaningful because genes are expressed over time. Gene Set Enrichment Analysis(GSEA)[2] is one of the method to analyze gene expression, using gene annotations. The conventional GSEA assumes that the input data consist of a gene expression profile based on two phenotypes or two biological states. Even if time series data are input into the GSEA, only two groups can be analyzed. Some of GSEA-related methods can analyze a simple time series, e.g., gene sets whose expression increases monotonically. In this paper, we propose the method which enables us to analyze time series gene expression profile for each gene function using gene database.

II. METHOD

Proposed method is divided into 2 steps.

- Test for time significance
  - Test for genetic significance

In the first step, we decide the target (window) size and slide it from the start to the end of the input data. Then, we determine the ranks based on the rate of variability in the gene expression, and we compute the Enrichment score(ES) based on the rank.

In the next step, we confirm whether the score computed in the first step is significantly higher than the other scores. This step is divided into two parts because we must test two types of significance. Then, we perform two comparisons: one among the gene sets and the other among the time intervals. The intervals in which a gene set is significantly expressed compared with the other intervals and other gene sets are the output.

III. EXPERIMENT

To verify the effectiveness of the proposed method, we performed an experiment, using adipocyte differentiation profile. In this experiment, we execute proposed method and count the gene sets which are confirmed that they have significant timing of expression. It is known that adipocyte gene expression is divided into two parts. Early part is 4~24 hours, later part is 2~6 day[3]. Almost all gene sets in Figure 1 are in these parts.

IV. CONCLUSION

In this paper, our purpose is to decide the term which the gene set express significantly, and we show that we can find the term as the result of proposed method. At the chapter of experiment, we verify that proposed method can adapt to time-series data. Moreover, there are no contradictions in the result from the biological viewpoint.

ACKNOWLEDGMENT

The present study was supported in part by Grants-in-Aid for Scientific Research(Nos.22680023, 24650155, and 24700294) from the Japan Society for the Promotion of Science(JSPS) and by the SPIRE Supercomputational Life Science through the Ministry of Education, Culture, Sports, Science and Technology of Japan(MEXT).

REFERENCES