

A method for comparing microarray data of different species

Kenji Hagimoto **Jun Miyazaki** **Shigehiko Kanaya**
kenji-h@is.naist.jp miyazaki@is.naist.jp skanaya@gtc.naist.jp

Naotake Ogasawara **Toshiyuki Amagasa** **Shunsuke Uemura**
nogasawa@bs.naist.jp amagasa@is.naist.jp uemura@is.naist.jp

Graduate School of Information Science, Nara Institute of Science and Technology

Keywords: DNA microarray, gene expression, comparative genome, time series data

1 Introduction

DNA microarray is now frequently used to collect an amount of time series data, for example to monitor gene expressions during the cell cycle. However DNA microarray method does not compare gene expressions between different species, but rather gene expressions between two statuses in one species are compared. In order to understand universality and diversity of different species, it is important to develop methods for comparing microarray data of different ones.

In this paper, we propose a method for comparing microarray data for two species by focusing time series data which is obtained by similar experiments. To directly compare such data, we correct distortions caused by differences in experiments and species. We describe correction function $k(\alpha)$ which corrects the distortions. It is calculated based on gene expressions of gene pairs in two species which have similar gene expression profiles in time series. Therefore we extract such gene pairs by two following procedure. First, using BLAST[1], we extract gene pairs with a high level of sequence homology. Second, by applying Dynamic Time Warping algorithm (DTW) which enables to find similar sequential data even when they are of different time duration, we moreover extract them with similar gene expression profiles in time series. For the purpose of calculating correction function $k(\alpha)$, both approaches are combined to screen gene pairs with similar gene expressions.

2 Proposed Method and Result

In order to calculate correction function $k(\alpha)$, we extract gene pairs which have similar gene expression profiles by the following procedure. At this time, we applied our proposed method to time-series DNA microarray data of *Bacillus subtilis* (8 points) and *Escherichia coli* (11 points) which were provided by our research groups.

2.1 Extracting gene pairs with a high level of sequence homology by using BLAST

Using BLAST, we extracted 730 gene pairs with a level of sequence homology under e-value 0.5 and without lack of time-series DNA microarray data between all genes of two species.

2.2 Screening gene pairs with similar gene expression profiles by applying DTW

For further screening gene pairs with similar gene expression profiles, we applied DTW to time-series gene expression data as follows. We could think of time-series data as wave by applying DTW.

- (1) Normalization for both gene expression profiles

We normalized time-series gene expression profiles so that the value of gene expression intensity might be settled between -1 and 1.

(2) Creating distance matrix and Computing similarity score between two waves

We created a distance matrix $d(i, j)$ based on difference of gene expressions $(X_1, X_2, \dots, X_b, Y_1, Y_2, \dots, Y_j)$ using the following equations.

$$X_i = X_{i+1} - X_b \quad Y_i = Y_{i+1} - Y_b$$

$$d(i, j) = |X_i - Y_j|$$

We computed distance score $g(i, j)$ which indicated degree of similarity between two waves (i.e. gene expression profiles) by using the distance matrix as following equations. At the same time, when only one wave crossed x-axis and the other did not, we imposed a penalty ($p=0, 0.5, 1, 2, 4$) on the similarity score. As a result, we obtained gene pairs ranked by the score.

$$g(i, j) = d(i, j) + \min \left\{ \begin{array}{l} g(i-1, j) \\ g(i, j-1) \\ g(i-1, j-1) \end{array} \right\}$$

2.3 Creating a scatter plot and Calculating coefficients of correlation

We created scatter plots and calculated coefficients of correlation as shown in Table 1, using the result described in Sect. 2.2. It is known that *Bacillus subtilis* has 58 genes related to ribosomal proteins[2] which are well conserved among all living thing. When we imposed $p=0$ for the penalty, 10 genes related to ribosomal proteins were observed in the top 50. Note that when we imposed $p=0.5, 1, 2, 4$ for the penalty, 15 genes related to ribosomal proteins were observed in the top 50.

Table 1: coefficients of correlation

rank	p=0	p=0.5	p=1,2,4
10	0.648	0.623	0.623
30	0.563	0.649	0.649
50	0.572	0.618	0.626

3 Conclusion

We have proposed a method for comparing microarray data of different species by focusing time series data which is derived from similar experiments. In the future, we plan to improve the accuracy of selecting gene pairs with similar gene expression profiles and calculate correction function $k(\alpha)$. We will also verify practical effectiveness of our proposed method.

Acknowledgments

I would like to thank Prof. Hirotada Mori and Prof. Naoki Ogasawara, Nara Institute of Science and Technology, for providing experimental data of DNA microarray. This work was supported in part by a Grant-in-Aid for 21st COE Research Program from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] <http://www.ncbi.nlm.nih.gov/BLAST/>

[2] <http://genolist.pasteur.fr/SubtiList/>