

Cleaning Microarray Expression Data with Markov Random Fields Based on Profile Similarity

Raymond Wan¹ Hiroshi Mamitsuka¹
rwan@kuicr.kyoto-u.ac.jp mami@kuicr.kyoto-u.ac.jp

¹ Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan

Keywords: data cleaning, microarray expression data, Markov random fields, expression profile similarity

1 Introduction

Microarray technology enables the expression levels of thousands of genes to be measured simultaneously. However, the expression profiles produced are known to be noisy due to various stages of the experiment. Both statistical methods and normalization address this problem [2, 3]. In this abstract, we describe an alternative called MEP-CLEAN which “cleans” the noise in microarray expression profiles.

MEP-CLEAN expands on related work in the field of image restoration with Markov random fields, or MRFs. Potential anomalies in a data set are identified and replaced with more suitable expression levels. The “correctness” of an expression level is assessed using other similar genes in the data set. Preliminary experiments show that our technique is able to clean 90% of the noise when 10% artificial noise is introduced.

2 Method and Results

Our technique consists of two phases, as shown in Figure 1. In the first phase, expression profiles are used to associate similar genes. Similarity is judged using a Euclidean distance measure where the most dissimilar pair of experiments are removed.

Genes which are more similar than a given distance threshold are used to construct an MRF for the second phase, MEP-CLEAN. MRFs extend Markov chains to higher dimensions. In an MRF, the probability of each event is conditioned on its neighboring events. In the context of expression profiles,

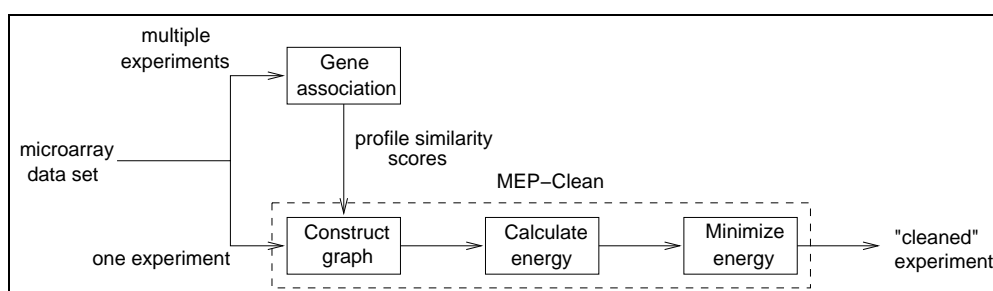


Figure 1: The application of MEP-CLEAN is shown along the bottom of the figure. Prior to using MEP-CLEAN, an initial “gene association” phase is required, shown at the top.

each gene’s expression level depends on the values of its neighboring genes. A gene’s neighborhood is formed by the genes in the data set which are the most similar.

MEP-CLEAN creates a separate MRF for each experiment. An MRF is represented as an undirected graph where each node contains a gene and its corresponding expression level. Edges are added between genes with similar expression profiles.

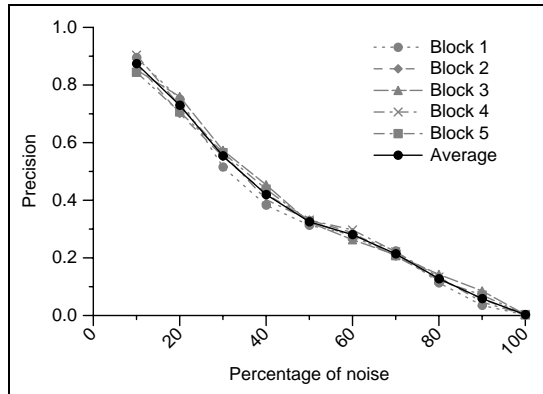


Figure 2: Five-fold cross-validation with GDS465.

Experiments were conducted within the framework of five-fold cross-validation using the data set

GDS465 from the Gene Expression Omnibus (GEO) [1]. This data set consists of 7,085 genes and 90 experiments. For each fold, 80% of the experiments is used for the first phase, while the second phase is repeated for each of the remaining experiments.

In the experiments, MEP-CLEAN is applied twice. The first execution takes the data set to a known baseline state to compare to. Then, artificial constant noise of +0.20 is added at probabilities ranging from 10% up to 100%. MEP-CLEAN is applied next and a comparison is made with the baseline levels. Figure 2 plots the precision of MEP-CLEAN, calculated from the number of noisy expression levels cleaned divided by the total number of noisy values. As the graph shows, MEP-CLEAN cleans 90% of the noise when 10% of the values are noisy. Effectiveness decreases as more noise is added, but even with 30% noise, 50% of the expression levels are successfully cleaned.

3 Discussions

In contrast to statistical methods and normalization, MEP-CLEAN takes into account the biological significance between genes in order to clean microarray expression profiles. Some potential future work include examining other means of determining gene similarity or other types of energy formulas.

Acknowledgements This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (C) “Genome Information Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

- [1] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. 30(1):207–210, January 2002.
- [2] R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics*, 18(5):265–271, May 2002.
- [3] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. 30(4):e15, February 2002.

$$U(f) = \alpha \sum_{r \in V} (\tilde{r} - \tilde{r}^*)^2 + \beta \sum_{(r,s) \in E} (\tilde{r} - \tilde{s})^2 \quad (1)$$

Next, an energy level $U(f)$ is calculated for the field using Equation 1. This energy is subsequently minimized by visiting each node and altering its expression level. The order in which nodes are visited is based on decreasing average similarity with its neighbors. The formula consists of two terms, where the first is summed over every vertex V and the second is summed over every edge E . The first term looks at the difference between the original and the current expression level of gene r . The second term considers the difference in expression levels between genes at the ends of each edge.