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

Raymond Wan
rwan@kuicr.kyoto-u.ac.jp

Hiroshi Mamitsuka
mami@kuicr.kyoto-u.ac.jp

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

Motivation

Microarray data sets (after the image is scanned) are generally noisy. Existing methods for dealing with this noise are mainly statistically-based.

In this work, we introduce a new method for cleaning noise in microarray expression profiles which draws on existing work from data mining and image restoration. The basis of our method is a probabilistic model called Markov random fields (MRFs).

Related Work

Some related work in this area include noise in microarray data, general noise cleaning, distance-based outlier detection, and image processing with MRFs.

Microarray data

- There are at least two sources of noise during microarray experiments: sample preparation and hybridization [Tu et al., 2002].
- Statistical methods and normalization can be used to adjust values to fit a distribution
  [Nadon and Shoemaker, 2002, Yang et al., 2002].

Related Work (cont.)

Data mining

- Existing data cleaning algorithms operate on general data [Kubica and Moore, 2003, Teng, 2004].
- Distance-based outlier detection employs a distance measure to compare each record (row) in a data set with every other record to find anomalies [Knorr et al., 2000].

Image processing

- In image restoration of a noisy picture with MRFs, each pixel's most likely value is cleaned using its neighboring pixels [Li, 2001].

Related Work (cont.)

Differences with previous work:

- Statistical methods and data mining compare each gene's expression level with every other genes'. In our work, the number of comparisons are reduced based on similarity between genes.
- In image processing, each pixel only has a handful of neighbors. Our method employs neighborhoods of varying sizes which range from 0 to every gene in the data set.

MRFs and Microarrays

In an MRF, each state only depends on its neighboring states. We define an MRF as a set of sites $\mathbb{S}$ and a set of labels $\mathbb{L}$. A mapping of labels to every site is a configuration f of the MRF. Each site $s \in \mathbb{S}$ has a set of neighbors $\mathbb{N}_s$ such that $s \notin \mathbb{N}_s$. So, an MRF is usually represented as an undirected graph.

A microarray data set is a two-dimensional data table of N genes and M experiments. At the intersection of a gene and an experiment is an expression level. An entire row of expression levels is the profile for that gene.

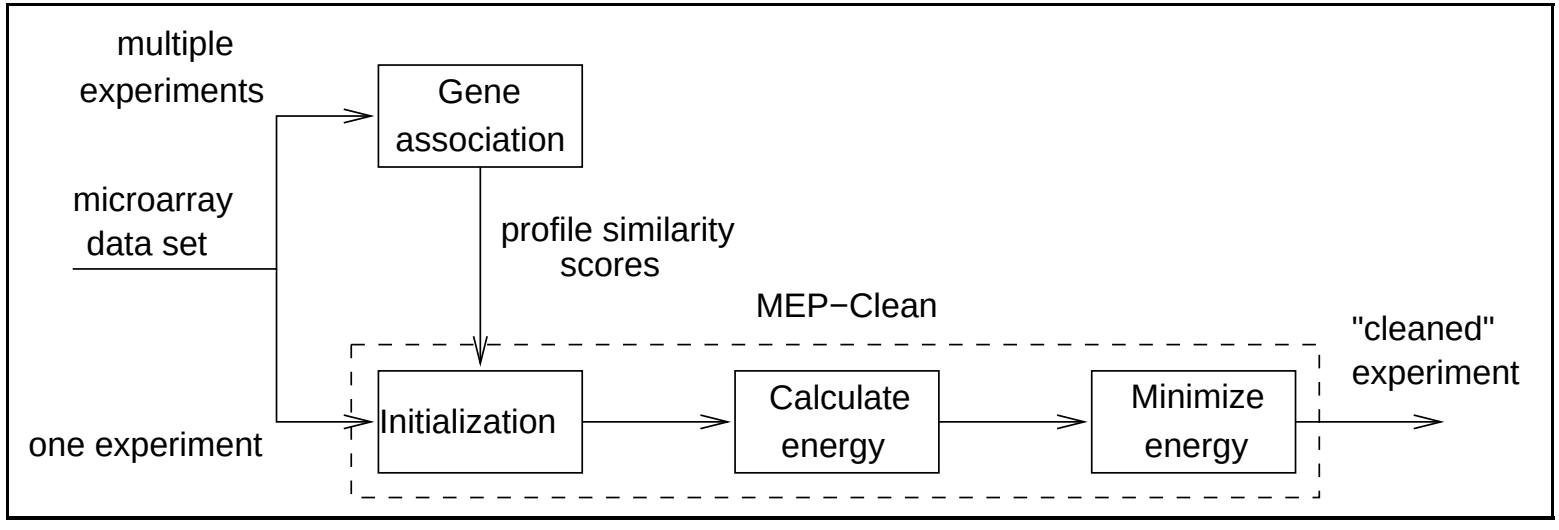
MRFs $\Longleftrightarrow$ Microarray Data

MRF	$G(V, E)$	Microarray data
sites ($\mathbb{S}$)	set of vertices V	genes
labels ($\mathbb{L}$)	values at those vertices	expression levels at the genes
neighborhoods ($\mathbb{N}_s$)	given by the edges E	genes with similar profiles
configuration (f)	map values to vertices	map expression levels to genes

So, a configuration is an assignment of expression levels to genes based on the gene profile similarity.

Methodology

Our method consists of two phases: gene association and MEP-CLEAN.



Gene Association

Calculate the similarity between profiles using the Euclidean distance. Normalized by the number of experiments.

Exclude:

- Pair of experiments where at least one is missing.
- The most dissimilar pair of experiments.

Pairs of genes which are less than a distance threshold of d_T (i.e., genes which are similar) result in a edge added between them in the MRF.

MEP-Clean

MEP-CLEAN first calculates an energy for the graph G:

$$U(f) = \alpha \sum_{v \in V} (\tilde{v}_i - \tilde{v}_i^*)^2 + \beta \sum_{(e_{ij}) \in E} (\tilde{v}_i - \tilde{v}_j)^2 \,. \qquad (1)$$

Where a vertex in the graph is v_i, for some i, its expression level is $\tilde{v}_i$. An edge e_{ij} exists between vertices i and j. Two parameters α and β.

Energy is derived from the sum of how much an expression level has changed so far (first term) and how different it is with its neighborhood (second term).

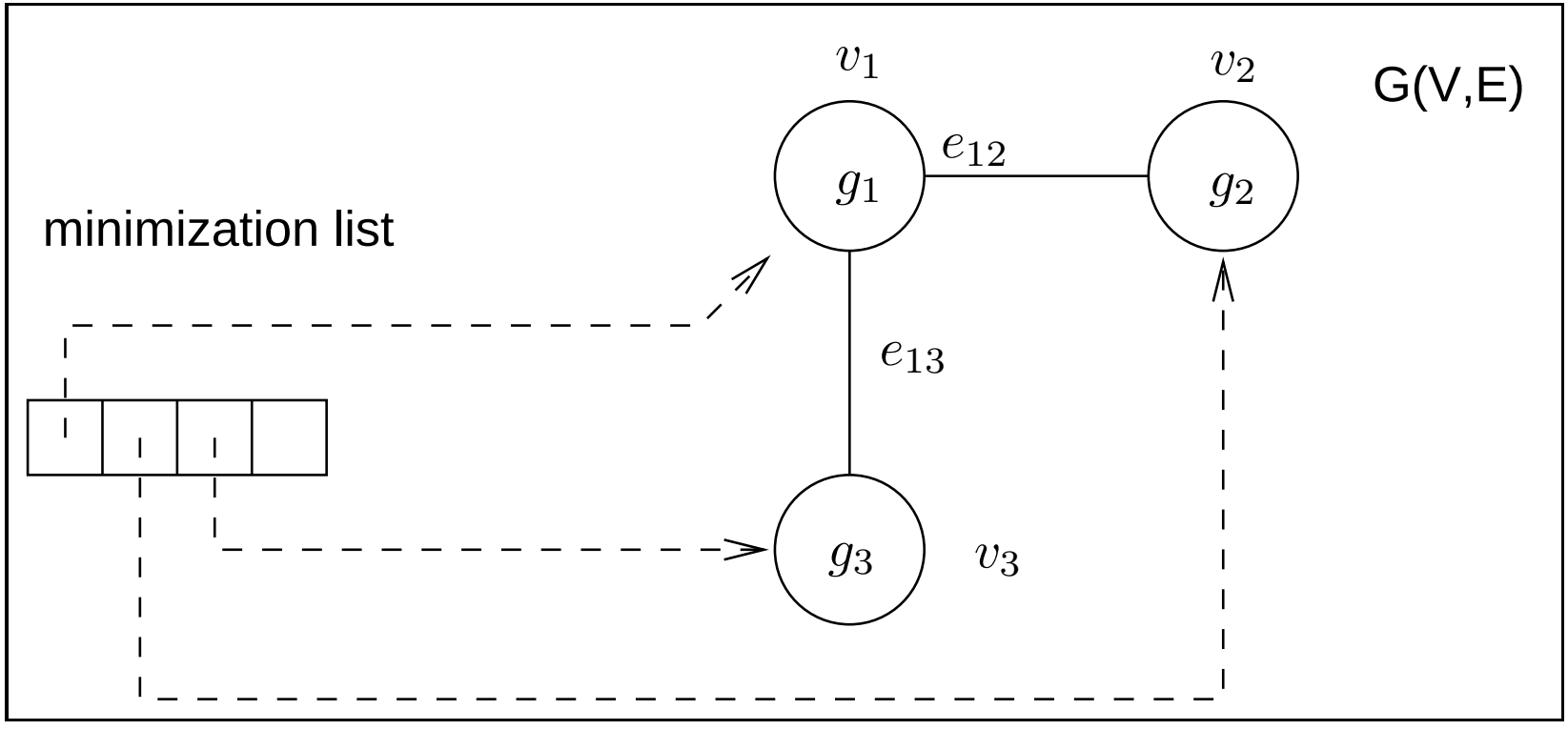
MEP-Clean (cont.)

Then, clean the expression levels by applying the following:

1. Create a minimization list ordered by decreasing average similarity of each node.
2. Visit each node and change its expression level until its contribution to the global energy is locally minimized.
3. Continue passing over the minimization list until a single pass yields no changes to the graph.

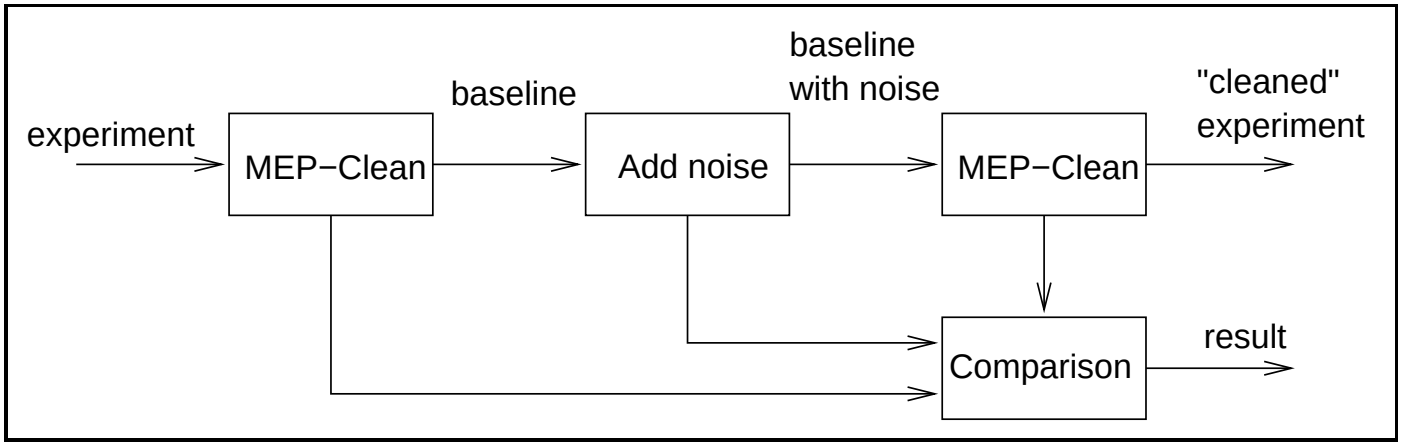
MEP-Clean (cont.)

The basic data structure:



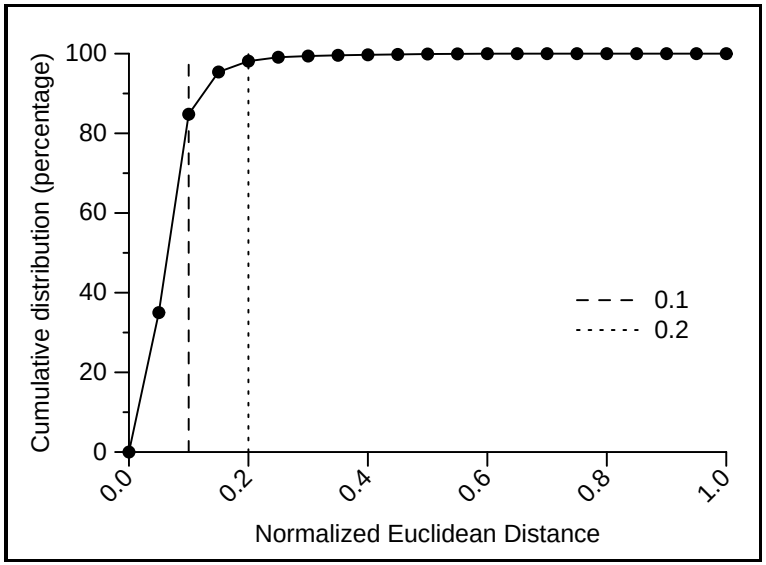
Experimental Framework

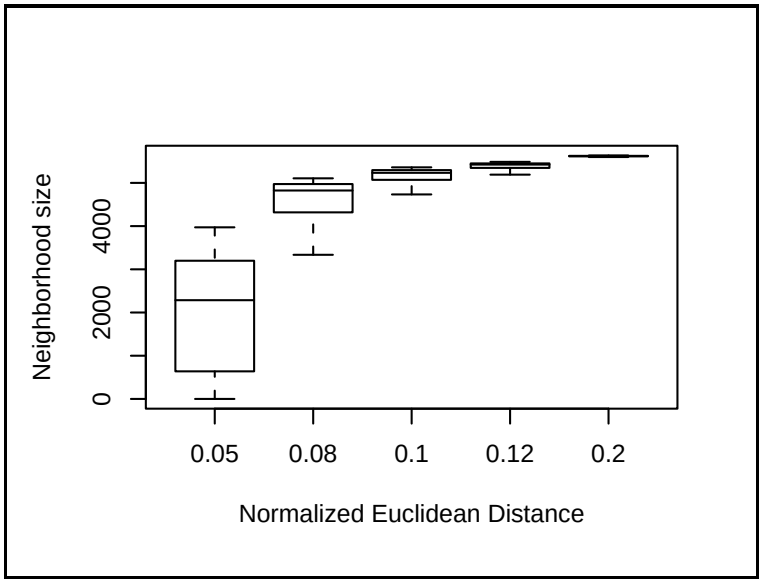
1. Use five-fold cross-validation by dividing the data set into five folds. Each fold has $M/5$ experiments. Gene association is performed with any four folds, while the last one is cleaned.
2. Apply MEP-CLEAN twice to each experiment with artificial constant noise of $+0.20$ at fixed percentages.



Results

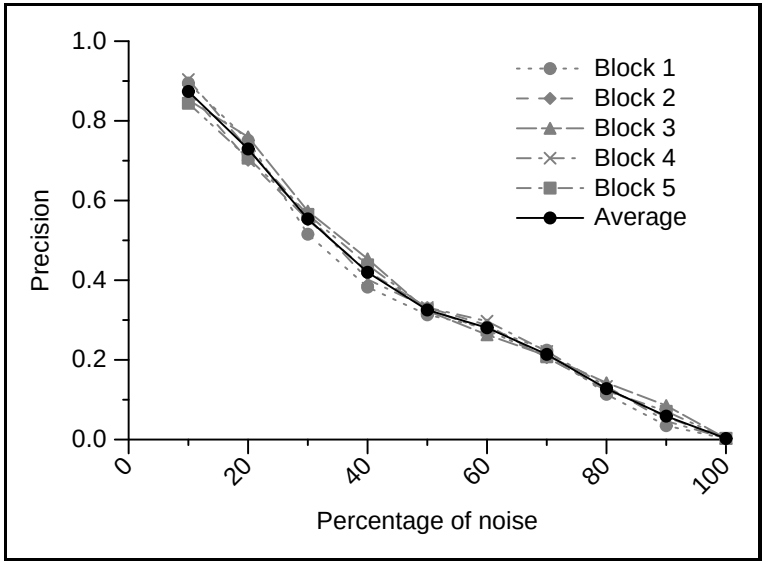
The test data is GDS465 (from GEO) of 7,085 genes and 90 experiments. Neighborhood sizes are shown below as the number of edges (left) and the neighborhood size (right) with respect to d_T.

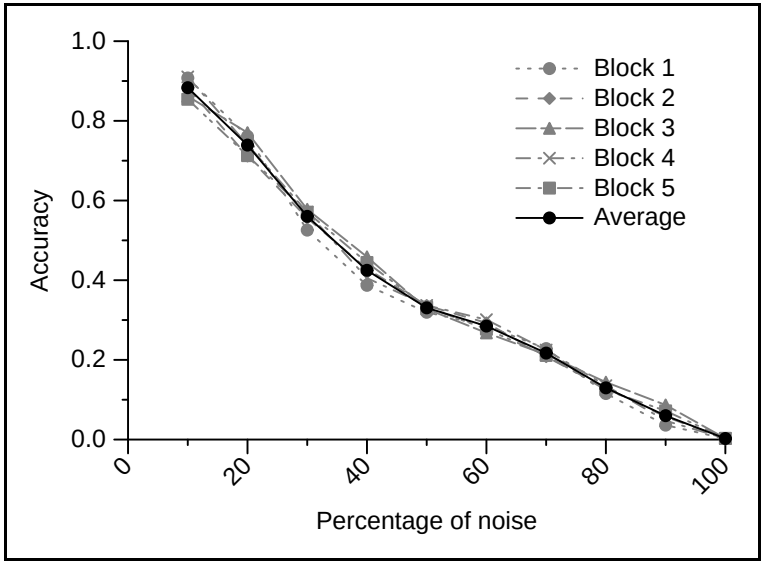




Results (cont.)

Define precision as $\frac{\text{\# of noisy values cleaned}}{\text{\# of total noisy values}}$ and accuracy as $\frac{\text{\# of clean values}}{\text{\# of total values}}$. Then, for GDS465:





- So, up to 90% precision and accuracy with 10% noise.

Future work

- Apply other types of artificial noise (Poisson, Gaussian, etc.).
- Instead of applying artificial noise, evaluate with real-world data that can be assumed to be “perfect” as the baseline.
- Vary the energy functions, parameters, and number of edges added to the graph.
- Use sequence similarity instead of profile similarity to create the edges in the graph.

References

E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *Special Issue on the Best Papers of VLDB '98, VLDB Journal*, 8(3–4):237–253, February 2000.

J. Kubica and A. Moore. Probabilistic noise identification and data cleaning. In X. Wu, A. Tuzhilin, and J. Shavlik, editors, *Proc. 3rd IEEE International Conference on Data Mining*, pages 131–138, November 2003.

S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Computer Science Workbench. Springer-Verlag, 2001.

R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics*, 18(5):265–271, May 2002.

C. M. Teng. Polishing blemishes: Issues in data correction. *IEEE Intelligent Systems*, 19(2):34–39, March/April 2004.

Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proc. National Academy of Sciences of the United States of America*, 99(22):14031–14036, October 2002.

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. *Nucleic Acids Research*, 30(4):e15, February 2002.