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

Raymond Wan **Hiroshi Mamitsuka** **Kiyoko F. Aoki**

Bioinformatics Center, Institute for Chemical Research,
Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
{rwan,mami,kiyoko}@kuicr.kyoto-u.ac.jp

Abstract

This poster introduces a method for cleaning noise found in microarray expression data sets using Markov random fields and genes with similar expression profiles. Preliminary results indicate potential for further developing our methodology into an alternative to existing techniques.

1. Introduction

Data sets produced from microarray experiments are inherently noisy. Statistical methods [Nadon and Shoemaker, 2002] and normalization [Yang et al., 2002] are two methods that can be used for this noise. This poster offers a third which combines a probabilistic model called Markov random fields (also used in image restoration [Li, 2001]) with data cleaning [Kubica and Moore, 2003].

In a Markov random field (MRF), the probability of an event depends on its immediate neighbors. The Hammersley-Clifford theorem [Hammersley and Clifford, 1971] showed how the probability of an MRF configuration can be calculated. One part of this calculation defines the field's energy which can be locally minimized if the ideal configuration is sought.

2. Method

In our work, an MRF is built for each microarray experiment. The MRF is represented as an undirected graph where each gene and its associated expression level is a node. Edges are drawn between genes which are similar using a user-provided parameter. An energy is assigned to the field and minimized by successively adjusting the expression levels at each node. This is analogous to cleaning the data set. These stages are shown in Figure 1.

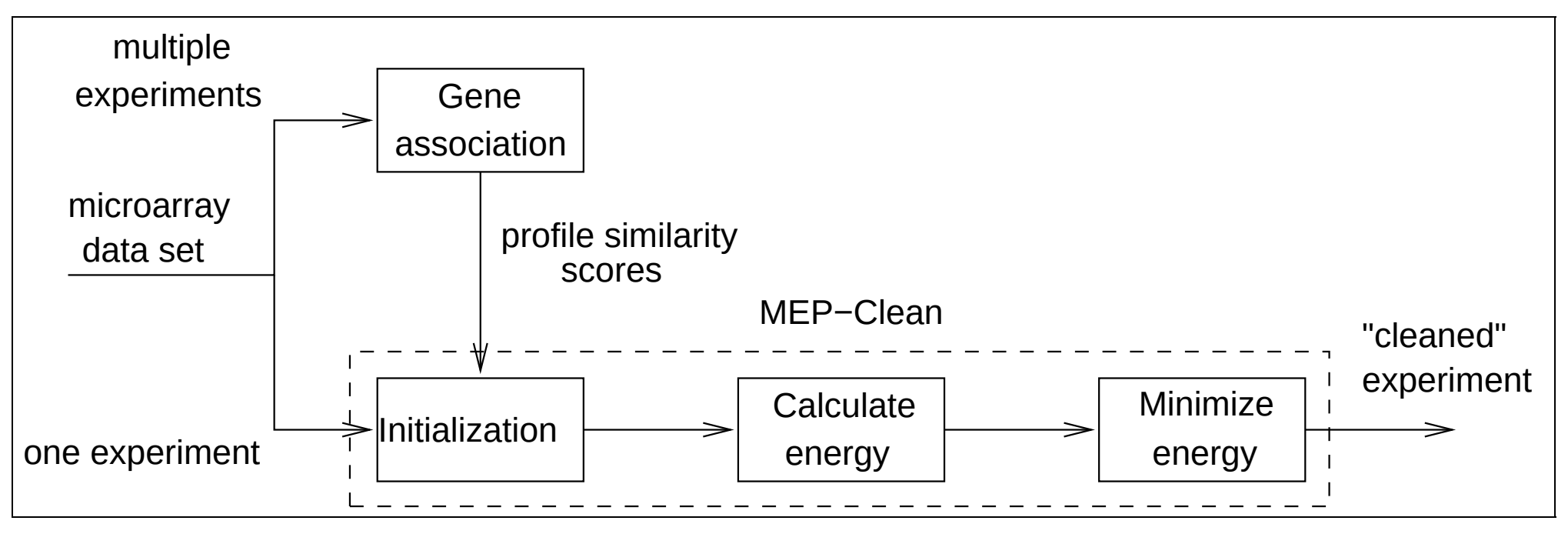

Figure 1: *The main stages of our method.*

The first step, *gene association*, determines the edges of the graph. In our work, the Euclidean distance between each *gene profile* is employed. All gene profile-pairs whose distance is less than a user-provided threshold d_T becomes a graph edge. Experiments were conducted with two data sets from the Gene Expression Omnibus (GEO) [Edgar et al., 2002]: `GDS91` and `GDS465`. Results are shown in Figures 2 and 3.

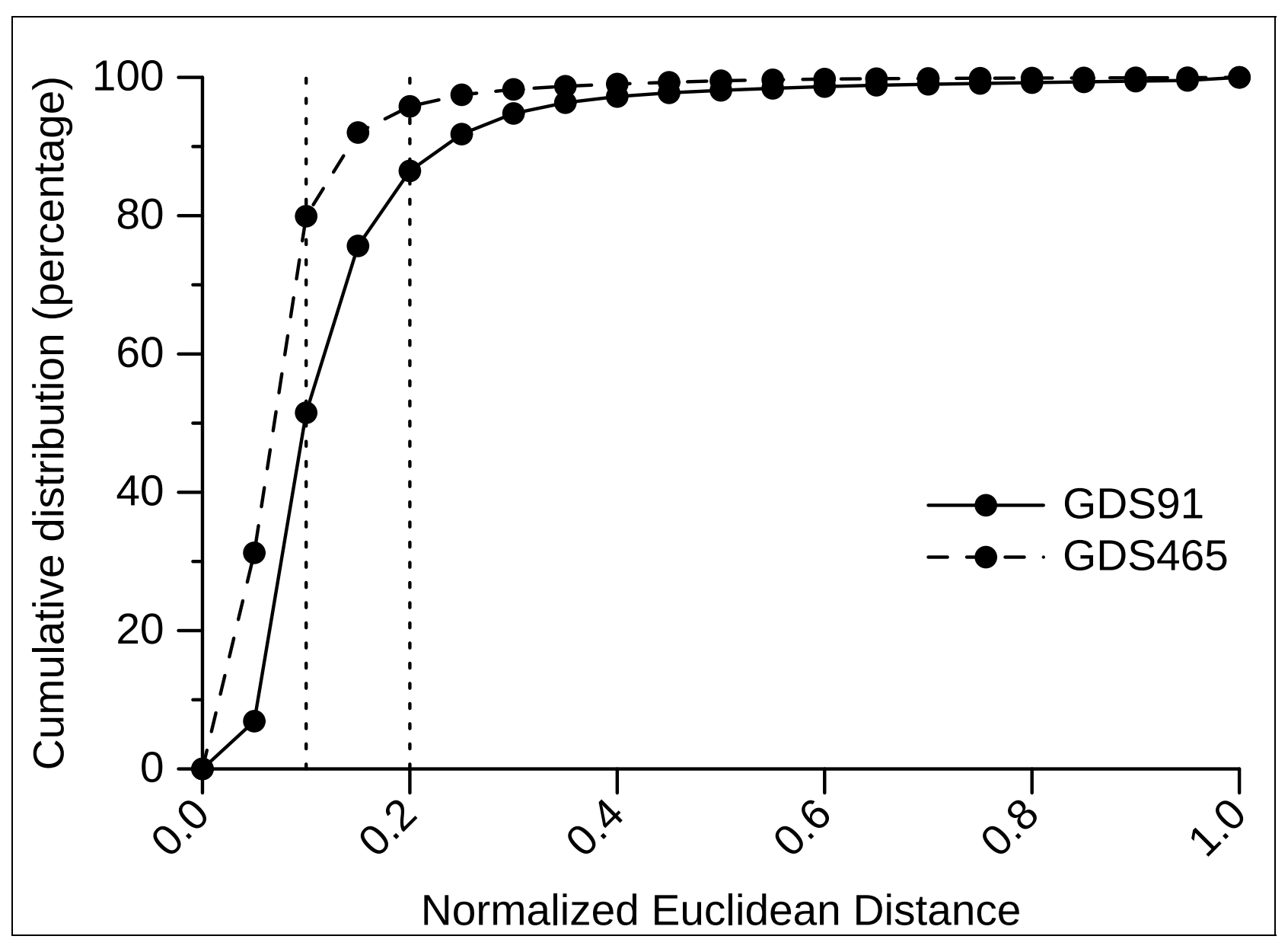

Figure 2: *Distribution of edges.*

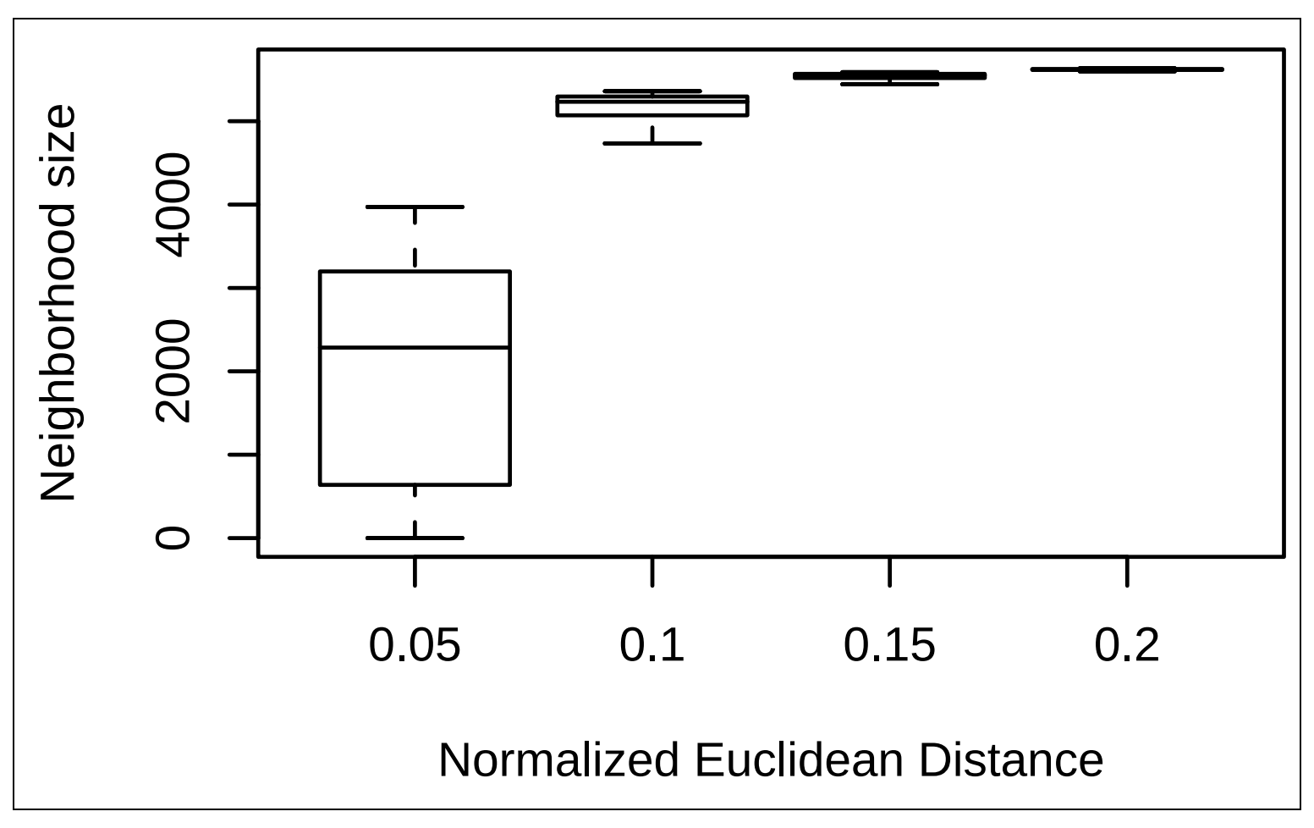

Figure 3: *Neighborhood size for `GDS465`.*

A sample graph of four genes and the similarity between gene profiles is shown in Figure 4.

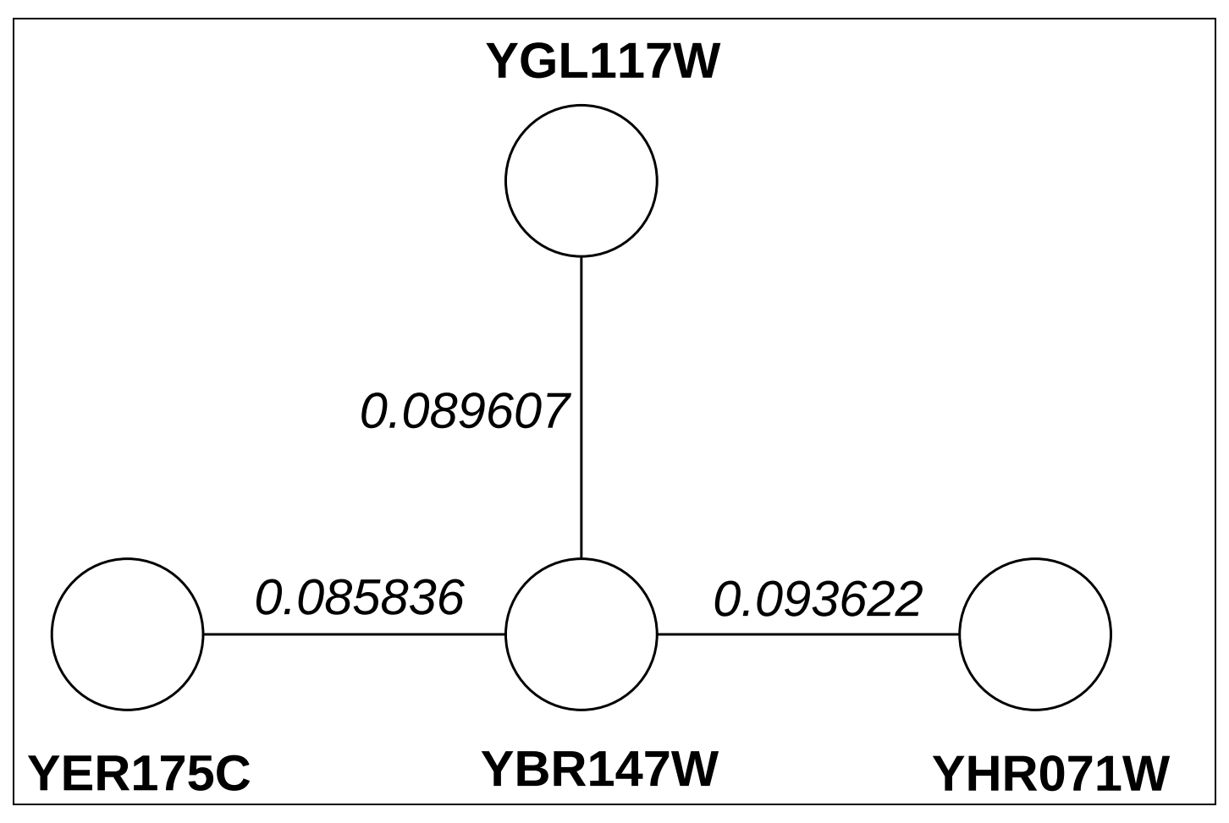

Figure 4: *Example of an MRF for a small microarray data set of four genes.*

Cleaning a data set is done an experiment at a time by MEP-Clean. First, an energy is applied to the field according to Equation 1. This equation consists of two terms whose strengths are controlled by the parameters α and β. The first term minimizes change to each expression level by taking the squared difference between its current value $\tilde{v}_i$ and its initial value $\tilde{v}_i^*$. The second term minimizes the difference between each expression level and its neighborhood. This formula is similar to the one found in image restoration [Li, 2001]. Vertices in the graph are cleaned in order of *decreasing average similarity*.

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

3. Results

The experimental framework employed is shown in Figure 5. MEP-Clean is applied twice so that the first application takes the data set to a baseline state. In between applications, constant noise of $+0.20$ is added at percentages ranging from 5% to 95%.

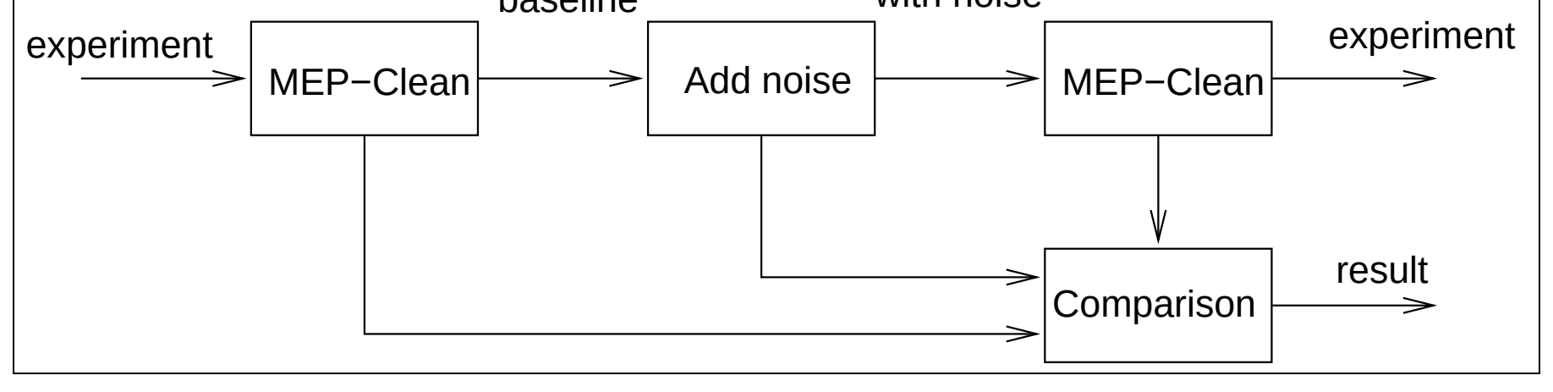

Figure 5: *The experimental framework.*

Thus, each gene's expression level can be categorized according to Table 1. Two metrics can be derived. Precision measures the ratio of cleaned values which were originally noisy; accuracy is the ratio of values cleaned.

	Noise	Noiseless
cleaned ($\leq \lvert\Delta\rvert$)	$\tilde{C}^+$	C^+
not cleaned ($> \lvert\Delta\rvert$)	$\tilde{C}-$	C^-

Table 1: *Classifications of expression levels, where $\Delta = 0.10$.*

Using this experimental framework and these two metrics, the results for `GDS91` and `GDS465` are shown in Figure 6.

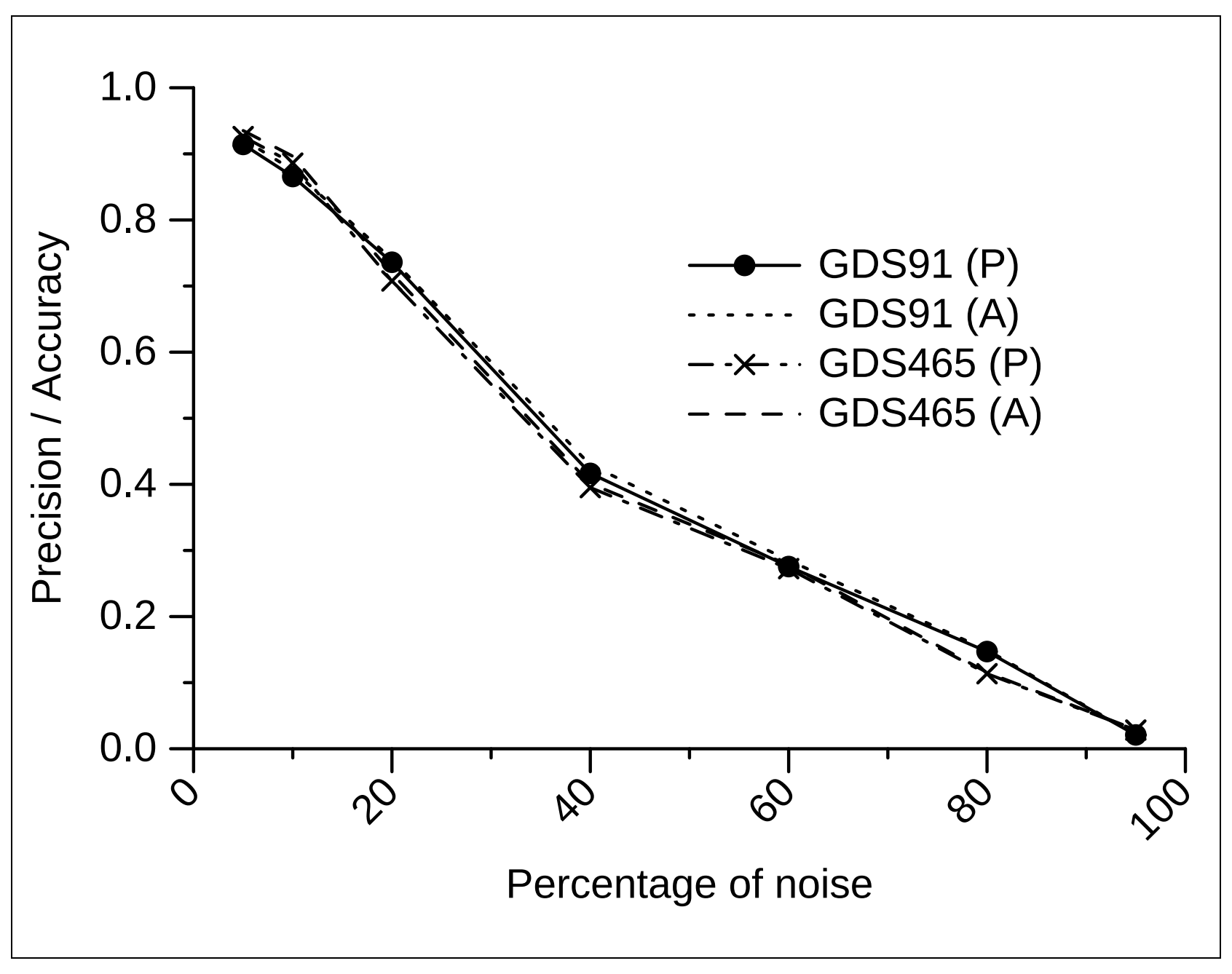

Figure 6: *Precision and accuracy of noise cleaning with* MEP-Clean *for `GDS91` and `GDS465`. The distance threshold was set at 0.10.*

4. Discussion

Preliminary results have shown that a microarray data set contaminated with artificial constant noise can be cleaned with 95% precision and accuracy when there is a chance of 5% noise. In order to avoid cleaning useful information, our method should be broken down into two steps: noise identification followed by user-directed noise cleaning.

However, the effectiveness of our system cannot be properly assessed unless real noise is cleaned. In essence, a real microarray data set is required whose "cleaned values" are accurate enough to compare to.

5. Recent Work

Recently, we are in the progress of improving our work as follows:

- We focus on outliers (noise that is rare and stronger in intensity) since small fluctuations in expression levels are typical across microarray data sets.
- We are comparing with existing methods such as statistical and normalization.

References

R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002.

J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. Unpublished, 1971.

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.

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.

```
Any suggestions on our on-going work
would be greatly appreciated.
```

SAC 2005, 20th ACM Symposium on Applied Computing, 14-17 March, 2005, Santa Fe, New Mexico