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.

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 $d$ 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.

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 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.

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.

References


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