130. Simulation Model for Gene Expression Data

Mikko Katajamaa and Jaakko Hollmén[1]

Keywords: data analysis, gene expression data, simulation model, test bench

1 Introduction

Microarray technology is widely used to yield thousands of simultaneous measurements. The measurement data must be analyzed in order to make inferences about the phenomenon under interest. However, the experiments are often designed with the needs of biological and medical sciences in view and only a few sample points are available. This is also caused by the high cost of conducting additional measurements. Therefore, it is utterly important to understand the limitations of the measurement process, specifically with regard to the measurement noise. For these reasons, we have developed a simulation model of the cDNA microarray measurement process [1]. The simulation model replaces the subsequent early stages of the measurement process preceding data analysis and uses numerical user-defined protein production levels instead of the source tissue sample. With the simulation model it is possible to produce an artificial data set of desired size under measurement conditions controlled with a set of adjustable parameters. Currently, systematic and random noise components can be controlled.

2 Simulation model for creating artificial microarray data

The input to the simulation model comprises of two source profiles, which are vectors of user-defined protein levels. The model itself consists of three successive layers, each of which corresponds to a central stage in the microarray measurement process. The profiles enter the layers in order; each layer disturbs the profile independently of the other layers. The last layer produces an intensity image as an output.

Table 1: The simulation model contains three layers. Types of disturbances and the parameters for controlling the disturbances are listed in the middle and right column of the table, respectively.

Layer name	Type of disturbance	Model Parameter
Layer for biochemical noise	Multiplicative noise. Picked up independently for every value in both profiles.	Multiplicative noise variance σ_1^2
Layer for hybridization distortion	Non-linear mapping. Same type of mapping for both profiles, but different slope.	Maximum departure from linear mapping a_{max}
Layer for scanning noise	Additive noise. Picked up independently for every pixel on both color channels of intensity image.	Additive noise variance σ_3^2

[1] Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FIN-02015 HUT, Finland. e-mail: Jaakko.Hollmen@hut.fi

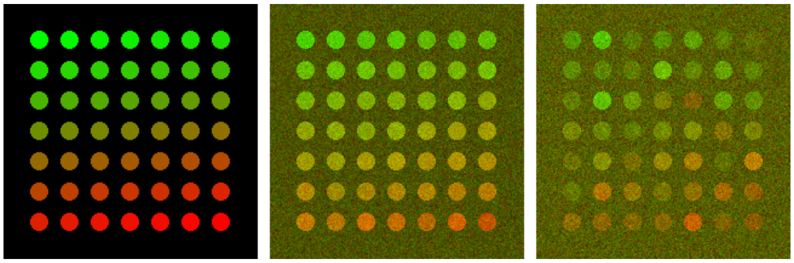

Figure 1: Three intensity images created by the simulation model with different parameter settings. The left image contains no noise, and the center and right images contain intermediate and strong levels of noise, respectively.

3 Test bench for data analysis methods

The simulation model is a central part of a test bench framework (see Figure 2), which can be used for testing the data preprocessing and data analysis methods for the cDNA microarray data. As already described, the simulation model distorts the ideal protein levels defined by the user, and the results from the model are fed to the data analysis method. It is possible to verify the performance of a data analysis method by comparing the results of analyzing the original, ideal data and the data distorted by the simulator.

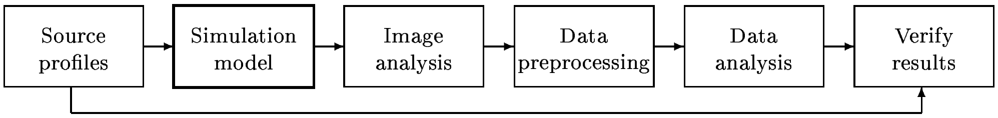

Figure 2: User-defined source profiles enter the simulation model, which distorts the profiles and produces an intensity image. The rest of the process follows the usual steps in the preprocessing and data analysis process.

The components following the simulation model should be provided by the user. They are not included in the simulation model itself, to enable testing different implementations for those components. However, if one is interested in testing only a new data analysis algorithm, for example, it is not necessary to implement all other components. For instance, it is possible to apply a third-party image analysis software, or to experiment with advanced algorithms for image analysis [2]. We have carried out preliminary experiments with the test bench in gene ranking and patient profile clustering tasks with varying noise levels [1].

References

[1] Mikko Katajamaa. *Simulation Model for Exploring Variations in the gene Expression Data and its Analysis.* Manuscript for a Master's thesis, 2003

[2] Salla Ruosaari and Jaakko Hollmén. Image analysis for detecting faulty spots from microarray images. *Proceedings of the 5th International Conference on Discovery Science (DS 2002)*, volume 2534 of *Lecture Notes in Computer Science*, pages 259–266. Springer-Verlag, 2002.