Probabilistic networks for cDNA microarray experiments: from raw data to gene expression effects

F.M. Stefanini

Dipartimento di Statistica "G. Parenti" - Firenze

Raw microarray data often contain substantial noise due to structural and random components of variation. It is widely recognized that raw data must be processed before studying the differential gene expression, and several transformations have been proposed.

Despite the lack of a formal model, each transformation makes implicit assumptions. A transformation may result in the sub-optimal processing of information if any of the implicit assumptions fails. Similarly, a waste of resources and biased inferences are expected from the inefficient analysis of replicated experiments.

Statistical quality control procedures require a parametric statistical model to perform optimal on-line monitoring and intervention. The effort required to develop a parametric statistical model of the measurement process is rewarded by the possibility of comparing different models according to the observed fit and of suggesting the range of application. Outliers and artefacts are identified by inspecting model residuals.

In this work a probabilistic network is proposed to normalize gene expression data using the information provided by control spots. The parametric model accounts for random components due to array manufacturing, print tips effects and differences due to dyes. All the point estimates of model parameters are provided with the indication of their precision. The final distribution of model parameters may be used to calculate gene expression effects to be further investigated by means of appropriate statistical techniques.

Results from the analysis of an actual case study show that estimated variance components due to array and subarray random effects overwhelm the variance of the experimental error. The analysis of residuals suggests that a source of artefacts may be present in the measurement process as regards one dye within one array. Moreover, one subarray has shown outlying features with respect to the other seven subarrays.

The probabilistic network described in this work may be extended in several ways, for example to study the differential gene expression. This and other issues are currently under investigation.