Dynamic detection of unexpressed genes in microarray mRNA

Kajsa Møllersen
Department of community medicine, UiT The Arctic University of Norway

Introduction

In a cell nucleus, the DNA strings with all their genes is stored. Any cell activity - good or bad - is dependent on the transcription of DNA to a multiple of shorter RNA strings. This is referred to as gene expression, measured as the quantity of mRNA strings corresponding to a specific DNA sequence. A major obstacle for statistical analysis on gene expression data is the presence of noise, which makes it difficult to draw valid conclusions. In microarrays, so called *negative control probes* measure the noise, whereas *regular probes* measure the actual gene expression. If the distribution of a regular probe is sufficiently different from the noise distribution, its corresponding gene is considered expressed.
The standard method for estimating the noise distribution relies on normality assumption and a cut-off dependent on user-set parameters.
The normality assumption does not reflect the actual noise distribution, and the result is sensitive to the parameters.

Aims

A new method, quantifying the *dissimilarity* between a gene and the noise, not relying on normality assumption, is presented.

Method

Let $p_{noise}(x)$ and $p_{probe,j}(x)$ be the probability density functions (pdfs) of the noise and probe number j, respectively. Then the Kullback-Leibler information D is defined as

$$D(j, noise) = \int p_{probe,j}(x) \log \frac{p_{probe,j}(x)}{p_{noise}(x)} dx.$$

Results

Each probe is no longer classified as expressed or not expressed, but its dissimilarity to the noise distribution is measured.
The method is demonstrated on a data set from the NOWAC study with 256 samples, analysed on microarrays with about $47,000$ probes. A random sample of negative controls is set aside for validation.
The dissimilarity between regular probes and noise will be presented as a distribution, and compared with the standard method using the validation set.

Conclusions

The strength of the new method is its dynamic approach to noise estimation, and possibility of a data-driven threshold instead of user-set parameters.

Keywords: Gene expression; Noise distribution; Microarray; Non-parametric; Kullback-Leibler
Preference: Oral