Significance analysis of microarray transcript levels in time series experiments

Di Camillo B, Toffolo G, Cobelli C

Information Engineering Department, University of Padova, 35131 Padova, Italy

Motivation

Microarray time series studies are essential to understand the dynamics of biological molecular events. In order to limit the analysis to those genes that change expression over time, a first necessary step is to select differentially expressed transcripts. This is often accomplished using an empirical or statistically based fold change threshold and comparing samples time by time. This approach is far-from-ideal since it does not account for the dynamic nature of the data and is particularly sensitive to random fluctuations due to the noise. To overcome these limitations a variety of methods were proposed; among them ANOVA based procedures, run test, autocorrelation and approaches based on regression modeling. However, these methods are seldom applicable in practice since they require either a large number of replicates or a relatively high number of time samples, while microarray time series experiments usually consist of a limited number of samples and replicates are only available for a limited number of them. Here we present a novel algorithm to select differentially expressed genes, which accounts for the entire dynamic profile and explicitly handles the experimental error. The method requires a relatively small number of replicates, which makes it applicable for time series microarray analysis.

Methods

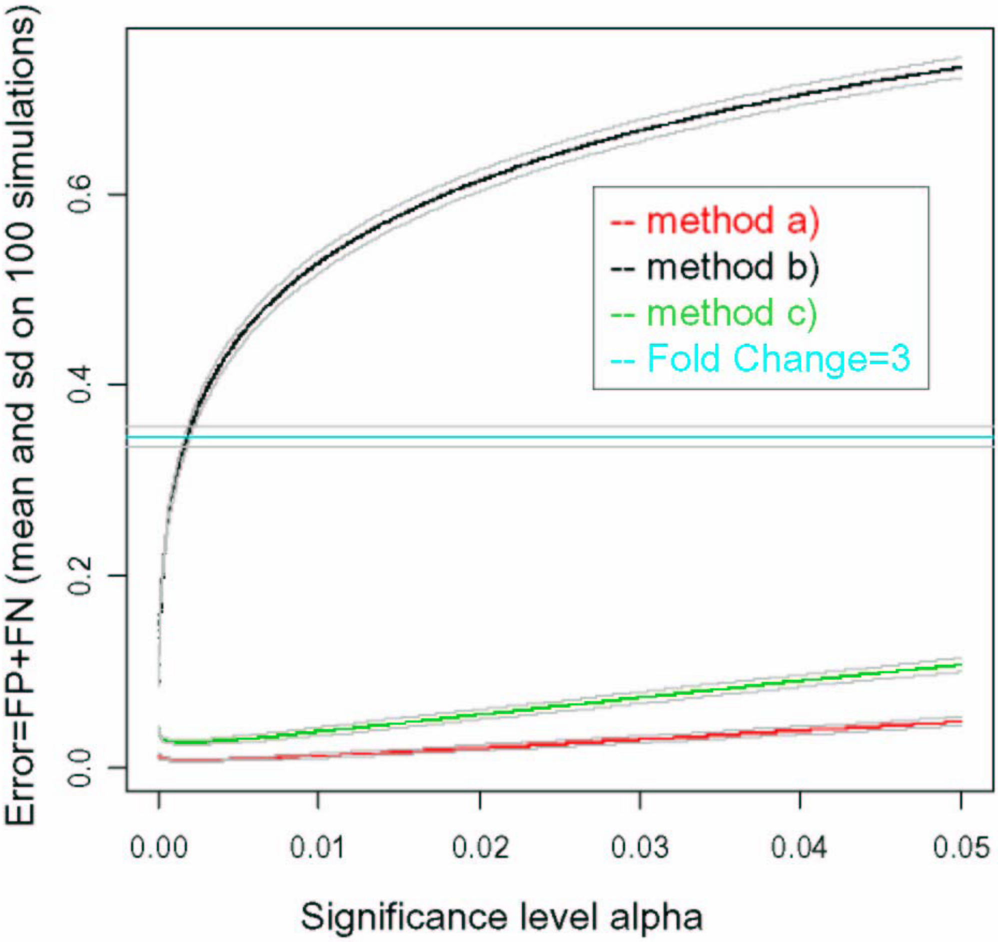
Three approaches to select differentially expressed genes are tested and their performances are compared using synthetic data: a novel method (a) and two methods proposed in the literature (b and c). Approach a) Let's call xT(tk) and xC(tk) the log-expression measurements in treated (T) and control (C) experiments, available for a generic gene x at time sample tk (k=1, ..., M). The rationale adopted to decide whether a gene x is differentially expressed in condition T and C is to calculate the area Ax of the region bounded by the two profiles and to consider it significantly different from 0 if it exceeds a threshold level D. For a given significance level alpha, the threshold D is derived from the distribution of the variable Ax when xT(tk) and xC(tk) are experimental replicates, i.e. the difference in the measurements represents only experimental variability. To this purpose Ax is calculated using a Monte Carlo approach for 10000 random time series, obtained by sampling the distribution of the following variable: dk=xT(tk)-xC(tk) derived when xT(tk) and xC(tk) are experimental replicates. Approach b) A threshold is derived based on the error distribution to be applied to dk for each time sample; genes are selected as differentially expressed if dk exceeds the threshold in at least one comparison. Approach c) Time series are fitted using a quadratic model of time. Genes are selected as differentially expressed if the differences between parameters estimated in T and C are significantly different from 0. Synthetic data 100 synthetic data-sets are generated, each consisting of 2000 genes and 10 time samples. In each data set 100 time series are generated using a Markov model in which dk+1 depends on dk according to a given probability model based on real data observations (multiple sets of differentially expressed genes), while 1900 time series are generated as random noise, modeled as a mixture of Gaussian as observed from real data.

Results

Method a outperforms methods b and c both in sensitivity and specificity, over a wide range of significance levels alpha. In Figure 1 the error (sum of false positives and false negative classifications) is reported as average (+/- standard deviation) on 100 simulations, for alpha ranging from 0 to 0.05. Results obtained using a constant fold change method (not depending on alpha) for each time sample, and selecting a profile as differentially expressed if it is differentially expressed in at least one time sample are also reported.

Availability: http://www.dei.unipd.it/~dicamill/Abstract

Contact email: dicamill@dei.unipd.it


Error=FP+FN (mean and sd on 100 simulations)
Significance level alpha
-- method a)
-- method b)
-- method c)
-- Fold Change=3