Selection of Insulin Regulated Gene Expression Profiles Based on Intensity-Dependent Noise Distribution of Microarray Data

Barbara Di Camillo[(1)], GiannaToffolo[(1)], Claudio Cobelli[(1)], K. Sreekumaran Nair [(2)]

[(1)] Dept of Information Engineering, University of Padova, Via Gradenigo 6B, 35131 Padova
dei@unipd.it

[(2)] Mayo Clinic, 200 First Street SW, 5-194 Joseph Rochester, MN, USA 55905
mayo@edu

Keywords. Insulin, microarrays, gene expression profiles

Introduction

Insulin resistance in skeletal muscle plays a key role in the development of Type 2 diabetes. To define the molecular mechanisms underlying insulin-induced changes in gene expression, recent studies, performed using microarrays techniques, identified genes involved in insulin resistance in control vs diabetic subjects, before vs after insulin treatment, i.e. exploiting only steady state information. Although extremely useful in order to identify candidate genes involved in analyzed processes and to develop new physiological hypothesis, these data can tell little about the interactions among genes. To infer genes regulation, it is of paramount importance to monitor dynamic expression profiles, i.e. time-series of expression data collected during the transition from one physiological state to another. A first necessary step, in order to limit the analysis to those genes that actually change expression over time, is to select differentially expressed genes.

Methods proposed in the literature usually deal with comparison of static conditions rather than time-course experiment data, and are based on application of modified t-test and ANOVA test which assume Gaussian distribution of analyzed variables. These methods test the significance of the differential expression gene by gene, and their application requires at least two replicated experiments per each condition.

In time course experiments, a number of samples is monitored across time and complete replicates of the experiment are seldom available, mainly for cost reasons. Therefore, differentially expressed genes are often selected using an empirical fold change (FC) threshold. This is a far-from-ideal situation, since it is based on an arbitrary choice (e.g. FC=2). In the case of Affymetrix chips, this choice is even more questionable since a constant threshold does not take in account the intensity dependence of the measurement errors, which is a well-known feature of this technology..

Here, we propose a novel method for gene selection, to be applied on dynamic gene expression profiles, which explicitly accounts for the properties of the measurement errors and addresses the situation where a relative small number of replicates is available.

Materials and Methods

The method was developed to select insulin-regulated genes in L6 skeletal muscle cells under 8 hours insulin stimulation. Twenty Affymetrix chips RG_U34A (monitoring 8.799 transcripts) were hybridized using eight samples collected every 1 h, from a control culture and from an insulin treated culture plus two replicates of the basal sample for each culture.

Error analysis was based on replicates of the basal sample, denoted for a generic gene as x_0 and x_{0bis}. Assuming that the error is log additive, the average error:

$$\delta = \frac{\ln(x_0) - \ln(x_{0bis})}{2} \qquad (1)$$

showed a standard deviation SD dependent in a non-linear fashion on the gene expression intensity, expressed as:

$$\bar{x}_0 = \frac{\ln(x_0) + \ln(x_{0bis})}{2} \qquad (2)$$

Moreover, while the error showed on intensity dependent distribution, the standardized error:

$$\delta' = \frac{\ln(x_0) - \ln(x_{0bis})}{2 \cdot SD(\delta)} \qquad (3)$$

did not.

The distribution of δ was fitted using a sum of N Gaussians and a 95% confidence threshold Θ was calculated under the null hypothesis. Genes that showed a value of expression in insulin treated culture at time j, x^I_j, significantly different from the corresponding sample in control culture, x^C_j:

$$\left|\ln(x_j^I) - \ln(x_j^C)\right| > 2 \cdot SD(\delta) \cdot \Theta \qquad (4)$$

were considered differentially expressed with respect to the baseline. Since SD(δ) depends on the gene expression level, also the threshold Θ'=[2SD·Θ] is intensity dependent (figure 1).

Only genes that were differentially expressed in at least one of the eight comparisons between chips 1…8 and the baseline were selected. The threshold was calculated by using Bonferroni correction of factor 8, since 8 comparisons were performed for each gene.

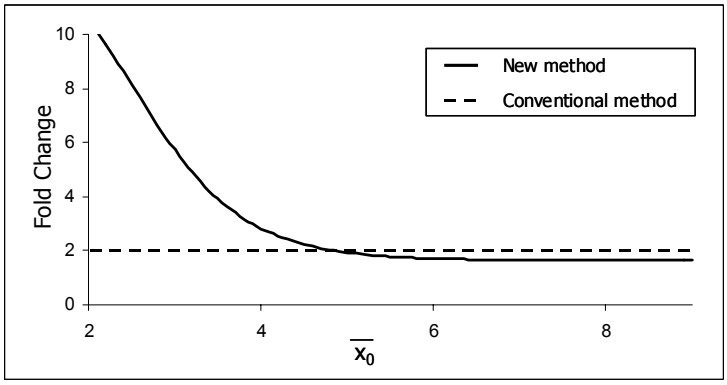

Fig. 1 Comparison between the thresholds Θ'=[2SD·Θ] obtained using a 95% confidence interval and expressed as fold change threshold, and a conventional "two-fold threshold".

Results and Conclusions

379 genes were selected using this method, which explicitly accounts for the intensity dependent properties of the Affymetrix measurement noise. The novelty of the method is that the intensity dependence of the noise behaviour is considered, rather than using an arbitrary fold change threshold. The measurement error depends strongly on the expression level and the use of a more conventional constant fold change would not consider this dependency thus penalizing highly expressed genes with respect to the lowly expressed ones.