I/NI-calls: a novel unsupervised feature selection criterion

Djork-Arné Clevert[1], Willem Talloen[2], Hinrich W.H. Göhlmann[2] and Sepp Hochreiter[1]

[1] *Institute of Bioinformatics, Johannes Kepler Universität Linz 4040 Linz, Austria*
[2] *Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica n.v., Beerse, Belgium*

Motivation:

High-density oligonucleotide microarrays, and in particular Affymetrix GeneChip arrays, are successfully applied in many areas of biomedical research. However, the large number of gene expression values, small sample sizes, and high noise levels lead to high false positive rates in extracting genes which are differentially expressed in different conditions. The false positive rate due to random correlations is a serious problem for biologists and medical researchers because if the significance level of their results is low or, even worse, they are misguided.

If the conditions are withhold in the first preprocessing step then random correlations between condition and expression values are considerably reduced in the second step. The first preprocessing step should exclude all genes, which do not contain a signal or are non-informative. As supervised feature selection approaches often suffer from overfitting, unsupervised feature filtering techniques are scarce and only look at the information content of the final expression value (signal intensity or non-Gaussianity). However for Affymetrix array data more information is available at probe level because a whole probe set records the expression value of a single gene. If probes of a probe set are governed by a common latent variable, then we associate this variable with the mRNA concentration and its variation with the mRNA variation, i.e. with the signal. Intuitively speaking, if probes of a probe set change synchronously across the arrays then this effect is very unlikely produced by noise and one should assume they are driven by a signal. Therefore we propose a novel unsupervised criterion that is based on a probabilistic latent variable model at probe level. Only such probe sets are filtered out where a variation of the latent variable can reliably be detected by a maximum a posterior optimization which combines and trades-off noise and signal likelihood.

Results:

We have applied this technique on 30 different data sets including recent publications from top journals (like Nature, Science and PNAS), covering six of the most commonly used gene chips. We found that our method excluded the non-informative probe sets without loss of sensitivity and specificity. The exclusion rates are about 98% while never losing a spiked-in gene in spiked-in data sets. As this objective technique results in uninformative feature reduction, it offers a critical solution to the curse of high-dimensionality in the analysis of microarray data.