38. Accounting for Probe-level Noise in Principal Component Analysis of Microarray Data

Guido Sanguinetti

sanguinetti@dcs.shef.ac.uk

PCA is widely used in microarray data analysis, however its outcome can be greatly influenced by the presence of noise on the data. Recently developed techniques for the extraction of expression levels from microarray data allow for an estimate of the error associated with each gene expression in each experiment. We develop a probabilistic PCA model that takes thes probe-level uncertainties into account and we demonstrate how the model can be used to denoise a microarray data set, leading to more robust data analysis and more coherent clusterings.

Principal Component Analysis (PCA) is one of the most popular dimensionality reduction techniques for the analysis of high dimensional datasets. However, in its standard form, it does not take into account any error measures associated with the data points beyond a standard spherical noise. This indiscriminate nature provides one of its main weaknesses when applied to biological data with inherently large variability, such as expression levels measured with microarrays. Methods now exist for extracting credibility intervals from the probe-level analysis of cDNA and oligonucleotide microarray experiments. These credibility intervals are gene and experiment specific, and can be propagated through an appropriate probabilistic downstream analysis.

We propose a new model-based approach to PCA that takes into account the variances associated with each gene in each experiment. We develop an efficient EM-algorithm to estimate the parameters of our new model. The model provides significantly better results than standard PCA, while remaining computationally reasonable. We show how the model can be used to `denoise' a microarray data set leading to improved expression profiles and tighter clustering across profiles. The probabilistic nature of the model means that the correct number of principal components is automatically obtained.

Normalization of cDNA microarrays using external control spikes Kristof Engelen, Bart Naudts, Koen Van Leemput, Bart De Moor and Kathleen Marchal

Normalization of microarray measurements is the first step in a microarray analysis flow. It aims at removing consistent sources of variations to make measurements mutually comparable. Reliable normalization is essential since the results of all subsequent analyses (such as e.g. clustering) might largely be influenced by the normalization procedure. For normalization of cDNA different methods have been described. Although some approaches inherently work with absolute intensities (e.g. ANOVA[1,2]), in general, preprocessing, of cDNA microarrays largely depends on the calculation of the log-ratios of the measured intensities. A common normalization step consists of the linearization of the Cy3 vs. Cy5 intensities (e.g. loess[3]). It assumes the distribution of gene expression is balanced and shows little change between the biological samples tested (Global Normalization Assumption). Global mRNA changes that result in an uneven distribution of expression changes however, have been shown to occur more frequently than what is currently believed[4,5], and could have a significant

impact on the interpretation of data normalized according to the Global Normalization Assumption. Therefore, in this study we describe a different way of normalizing cDNA microarray data. In contrast to previous approaches, our methodology is based on a physically motivated model, consisting of two major components. We explicitly model the hybridization of mRNA transcripts to their corresponding cDNA probes and the relation between the measured fluorescence and the amount of hybridized, labeled mRNA. The parameters of this model and the incorporated error distributions are estimated from external control spikes: mRNA transcripts that are added to the hybridization solution in known concentrations. Using a publicly available data set, we show that our procedure, due to the inherent nonlinearity of the model, is capable of adequately linearizing the data, without making any assumptions on the distribution of gene expression (as opposed to the Global Normalization Assumption). More importantly, since our model links mRNA concentration to measured intensity, we are able to estimate the absolute concentrations of mRNA transcripts in the hybridization solution with fair accuracy.

Kerr MK, Martin M, en Churchill GA 2000. Analysis of variance for gene expression microarray data. J. Comput. Biol. 7: 819-837.

Jin W, Riley RM, Wolfinger RD, White KP, Passador Gurgel G en Gibson G 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 29: 389-395.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J en Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30: e15.

van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC. 2003 Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep. 4(4):387-393.

van Bakel H, Holstege FC. 2004 In control: systematic assessment of microarray performance. EMBO Rep. 5(10):964-969.