Improving temporal gene expression profiles with probabilistic models

Marta Milo[1], M.C. Holley[1], M. Rattray[2], M. Niranjan[3] and N. D. Lawrence[3]

Keywords: temporal gene expression profiles, high-density short oligonucleotide microarrays, probabilistic models, gamma distribution.

1 Introduction.

High-density short oligonucleotide microarrays are a widely used tool for measuring gene expression on large scale [3]. A key issue for short oligonucleotide probes, such as the one used by Affymetrix, Inc., is the way of selecting probe sequences with high sensitivity and specificity. The use of multiple probe pairs (11-20 pairs), referred as *probe set*, to target a single gene is the current approach to this problem. One of each pair exactly matched the DNA fragment of the gene (*PM probe*) and the other contains a single mismatch base in the middle (*MM probe*). The sensitivity of the gene expression signal is given by the PM values and the MM values, offering a measure of non-specific binding, improve the specificity. However, around 30% of the pairs consistently have a negative signal, indicating that the MM values are unreliable as pure measure of non-specific binding [5, 7]. Moreover the variation of the order of magnitude of the probe signals within a probe set suggests that not all the probes have optimal sensitivity [2]. Given the above limitations, it becomes crucial to design a model for the extraction of gene expression signals from oligonucleotide arrays. Many recent studies have focused on statistical methods [1, 2] to summarise these expression levels, choosing not to use the MM probes for the unreliability mentioned above. In this work we chose to use a probabilistic model to describe both the sensitivity and the specificity of the probe set including in the model the MM observed signals. The model, gamma Model for Oligonucleotide Signal (gMOS) [4] proved to perform well both on publicly available data set and on a real case study. To improve the specificity of the model we take in account the correlation between MM probes and PM probes, using a different approach that allows us to learn the parameter of the joint distribution of PM and MM from the observed data. The *modified gMOS* (mgMOS) is tested on a temporal profile of a cell line, UB/OC-1, [6] derived from epithelial cells in the cochlear duct at embryonic day 13.5 (E13.5). The extracted profiles are compared with a real time RT-PCR profile and with profiles obtained with both the Affymetrix MAS v.5 and with different statistical methods.

2 Material and Methods.

For each probe set we model the PM signal (y) and MM (m) using a gamma probability functions. The gene expression signal (s) is then derived from the joint probability of y and m. In the basic gMOS we assume independency in the probe set. The joint probability of y and m is defined as:

$$p(y_i, m_i) = p(y_i \mid a, \alpha, b)\, p(m_i \mid a, b)$$

and $i=1,...N$. N is the number of probe pairs in a probe set. The parameters a, α, b are estimated using Maximun Likelihood. The parameter b does not vary within the probe set, modeling the genetic affinity of corresponding short oligonucleotide with the gene sequence as constant within

[1] Institute of Molecular Physiology, Department of Biomedical Science, Addison Building, Western Bank, Sheffield, S10 2TN, UK E-mail: M.Milo@sheffield.ac.uk

[2] Department of Computer Science, University of Manchester, UK. E-mail: magnus@cs.man.ac.uk

[3] Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield S1 4DP,UK. E-mail: niranjan@dcs.shef.ac.uk, neil@dcs.shef.ac.uk

the probe set. The mgMOS is a latent variable model that defines this genetic affinity as varying within the probe set. This difference is defined by modeling the correlation between *y* and *m* that is also detectable from the observed *y* and *m* signals. In the modified model the parameter *b* now varies within the probe set. The joint probability distribution of *y* and *m* becomes:

$$p(y_i, m_i) = \int p(y_i \mid a, \alpha, b_i)\, p(m_i, a, b_i)\, p(b_i)\, db_i$$

The equation is tractable for Gamma distributions. Given $p(b_i) = \frac{d^c}{\Gamma(c)} b_i^{c-1} \exp(-db_i)$ it is possible to calculate the joint distribution of *y* and *m* and therefore obtain the estimate of the signal *s* with the above equation. The parameters are again estimated using Maximum Likelihood.

3 Results and conclusions.

The two methods are tested on benchmark data and on a temporal profile of the transcription factor *gata3* from an inner ear cell line. The temporal profile consists in 12 time points sample in 14 days of development after differentiation. The method gives very promising results both for the application to real dataset (Figure 1) and for further theoretical exploitation.

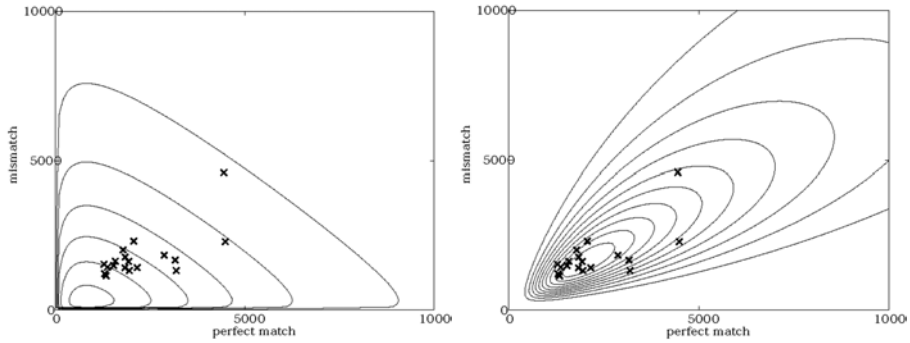

Figure1: Correlation of PM and MM for one time point of the gata3 profile. The plot the left shows the contours of the distributions obtained with gMOS; the plot on the right shows contours obtained with mgMOS. The plot clearly shows that mgMOS better fits the observed data.

References

[1] Irizarry R., B. Bolstad, F. Collin, L. Cope, B. Hobbs and T. Speed. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acid Research*, **31**, No. 4 e15.

[2] Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: expression index, computation and outlier detection. *Proc. Natl Acad. Sci.*USA, **98**, 31-36

[3] Lockhart, D.J. and Winzeler E.A. 2000. Gene expression and DNA arrays. *Nature*, 405, 827-836.

[4] Milo, M., Fazeli, A., Niranjan M. and N. D. Lawrence 2003. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Transations*, **31**, 6.

[5] Naef, F. Hacker, C.R., Patil N., and Magnasco M. 2002. Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biology*, **3**, research0018

[6] Rivolta, M. N.,Hasall,A., Johnson, C.M., Tones, M.A., Holley, M.C. 2002. Transcript profiling of functionally related Groups of Genes During Conditional Differentiation of Mammalian Cochlear Hair Cell Line. *Genome Research,* **12**, 7, pp. 1091-1099.

[7] Zhou, Y. and Abagyan, R. 2003. Match-Only Integral Distribution (MOID) Algorithm for high density oligonucleotide array analysis. *BMC Bionformatics*, **3**, 3.