A Probabilistic Model for the Extraction of Expression Levels from Oligonucleotide Arrays

Marta Milo, Alireza Fazeli+, Mahesan Niranjan, and Neil D. Lawrence
Department of Computer Science
and
+Academic Unit of Reproductive and Developmental Medicine,
University of Sheffield, U.K.

Abstract

In this work we present a probabilistic model to estimate summaries of Affymetrix GeneChip probe level data. Comparisons with two different models were made both on a publicly available dataset and on a study performed in our laboratory, showing that our model performs better for consistency of fold change.

Introduction

Oligonucleotide expression array technology has been adopted in many areas of biomedical research to measure simultaneously the level of messenger ribonucleic acid (mRNA) transcripts for thousands of genes. Affymetrix GeneChip array is an oligonucletide based array technology. In this technology each gene is represented by a set of 11 to 20 pairs of oligonucleotides that we refer to as probes. Each probe pair is composed of a perfect match (PM) probe, a section of mRNA molecule of interest, and a mismatch (MM) probe that is created by changing the middle base of the PM. To define a measure of the expression level associated to each gene is necessary to summarise to a single expression level the probe intensity values for each probe set. The analysis of such experiments is not trivial because the probe signals are affected by many levels of variation introduced at different stages of the experiments. A further difficulty is represented by the large differences that may exist among different probe sets used to interrogate the same gene.

In this work we propose a novel approach that makes use of probabilistic models for the PM and MM samples. We use these models to summarise probe expression levels and to extract a level of uncertainty associated with each probe set. We evaluated our approach on a publicly available Affymetrix Spike-in study and a mouse oviduct gene expression study. We compared the results with the expression measures provided by the default Affymetrix microarray suite (MAS v5.0)[1] and with Robust Multi-array Average expression measure (RMA)[2]. Our comparison on the Spike-in study is expressed in terms of: consistency of fold change, using mean squared errors; specificity and sensitivity of the measures' ability to detect differential expression, using receiver operating characteristic (ROC) curves. We also present a table of the most differentially expressed genes in the oviduct gene expression study, arranged in magnitude of fold change with respect to MAS 5.0 measurements.

Material and Methods

Different methods and model have been proposed to summarise probe level data, all based on empirical statistical model[1,2,3], here we propose a probabilistic model based on the assumption that the underling probability distribution for both the PM and MM signals is a gamma distribution $Gamma(\cdot,b)$ with same inverse scale factor, b and different shapes, α and a. We can describe the model as follows:

$$y_{ij} = m_{ij} + s_{ij} \qquad \text{with } i = 1,\ldots,n_j \text{ and } j = 1,\ldots,N$$

where *y* is the PM observed signal, *m* is the MM observed signal, *s* is the true probe signal, *N* is the number of probes on the chip and n_j is the number of probes in the *j*th probe-set. Assuming that $m_{ij} \approx Gamma(a_j, b_j)$, $y_{ij} \approx Gamma(\alpha_j + a_j, b_j)$, $s_j \approx Gamma(\alpha_j, b_j)$ we can derive the following probability expressions:

$$p(m_{ij} \mid a_j, b_j) = \frac{b_j^{a_j}}{\Gamma(a_j)} m_{ij}^{a_j - 1} \exp(-b_j m_{ij})$$

$$p(s_{ij} \mid \alpha_j, b_j) = \frac{b_j^{\alpha_j}}{\Gamma(\alpha_j)} s_{ij}^{\alpha_j - 1} \exp(-b_j m_{ij})$$

$$p(y_{ij} \mid a_j + \alpha_j, b_j) = \frac{b_j^{a_j + \alpha_j}}{\Gamma(a_j + \alpha_j)} y_{ij}^{a_j + \alpha_j - 1} \exp(-b_j y_{ij})$$

where $\Gamma(\cdot)$ is the Gamma function.

The parameters α_j, a_j and b_j are estimated by maximising the joint likelihood:

$$L(a_j, \alpha_j, b_j) = L(a_j, b_j) + L(\alpha_j + a_j, b_j)$$

using the conjugate gradient optimisation algorithm[4]. Thus the expected true probe signal $< s_j >$ and the associated precision $1/\sigma_j^2$ are respectively given by:

$$< s_j >= \frac{\alpha_j}{b_j} \quad \text{and} \quad \sigma_j^2 = \frac{\alpha_j}{b_j^2}.$$

In our experiments we used the expected log true probe signal, which can be derived as $< \ln(s_j) >= \psi(\alpha_j) - \ln(b_j)$, where $\psi(\alpha_j) = \frac{\partial}{\partial \alpha_j} \ln(\Gamma(a_j))$.

There is no golden standard to compare and test summaries of probe level data. For this reason we chose to assess our model on a publicly available Affymetrix spike-in dataset. The Affymetrix experiment[5] consists of 14 groups of human genes spiked-in at known cRNA concentrations, arranged in a cyclic Latin square design with each concentration appearing once in each row and column. Each group has three repetitions. We randomly sampled 4 groups and calculate the fold change of all 6 pairs using the three expression measures.

To assess the consistency of fold change we calculated the mean relative concentration for each spiked-in probe and compared it with the mean relative signal prediction. The accuracy of the prediction on this data is measured by the mean squared error which provided the following results: Gamma: 0.4%, MAS 5.0: 0.8%, RMA: 0.7%.

For the sensitivity and specificity study we calculated the number of false positive as the number of non-spiked-in genes with fold change estimate larger than the cut-off value. Conversely, the number of true positive was calculated as the number of spiked-in genes with fold change estimate larger than the cut-off value. We used a large range of fold change cut-off values. The ROC curve derived is illustrated by Figure 1 where the area under the ROC curve (AUROC) was: Gamma: 0.96, MAS 5.0: 0.94, RMA: 0.95.

A second comparison was carried on a dataset obtained from a study investigating the reaction of mouse oviduct to sperm, performed in our laboratory[6]. RNA obtained before and after mating from mouse oviducts were hybridised to MG-U74 Affymetrix array chips. Two genechip arrays were hybridised to RNA samples obtained from oviduct before mating (duplicate) and three to that after mating (triplicate). The following table summarises the results of our analysis in comparison with MAS 5.0 and RMA on a selected subset of highly differentially expressed genes before and after mating. It shows the top log2 fold change differentiation for Oviduct experiments, in descending order with respect to MAS 5.0.

UniGene	Name	gamma FC	MAS FC	RMA FC
Mm.613	Anp32a acidic (leucine-rich)	7.3	4	0.9
Mm.3137	Ptgs2 prostaglandin-endoperoxide synthase 2	6.3	3	1.2
Mm.4312	Slc9a1 solute carrier family 9 (sodium/hydrogen exchanger)	5.7	2.2	0.5
Mm.21013	Cxcl1 chemokine (C-X-C motif) ligand 1	6.4	2.2	1.2
Mm.108678	Cyp11a1 cytochrome P450, family 11, subfamily a, polypeptide 1	4.1	2.1	1.3
Mm.257330	ESTs	2.5	1.8	2.1
Mm.250422	Serpine1 serine (or cysteine) proteinase inhibitor, c E, member 1	2.4	1.8	0.7
Mm.245967	ESTs	3.9	1.7	2.2
Mm.1408	Adm adrenomedullin	7.2	1.7	1
Mm.4063	Ndr1 N-myc downstream regulated 1	2.3	1.7	1.5
Mm.4639	Cebpd CCAAT/enhancer binding protein (C/EBP),	2.1	1.7	1.6

Our method associated higher fold change values to genes that are highly differentiating in this study. Real time RT-PCR analysis was used to check expression values of two of the above genes: adrenomedullin (ADM) and prostaglandin endoperoxide synthase-2 (PTGS2) in mice oviducts before and after mating (Figure 2).

Discussion

The probabilistic model described in this paper has proved to be consistently comparable with robust statistical based models, outperforming them for consistency of fold change. More over in our study we did not perform any background correction or normalisation of the data, which are instead carried out by both MAS 5.0 and RMA. The overlapping of the ROC curves for gamma and RMA shows that it is possible to combine the two methods[7], in order to achieve better results. This last point is current matter of research.

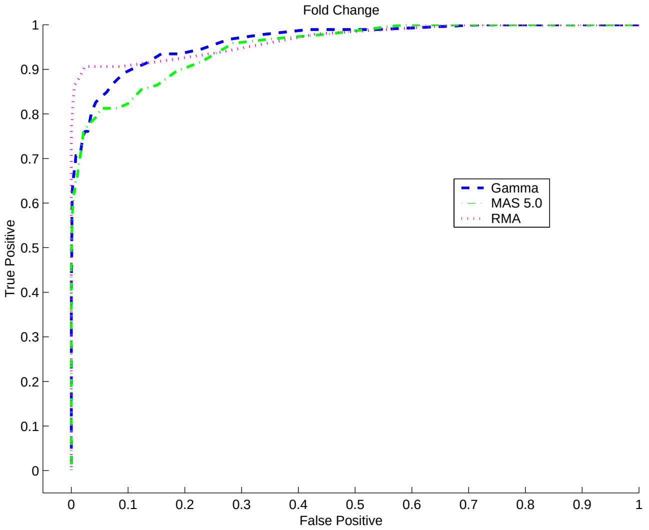

Figure1. ROC curves for 6 pairs of arrays chosen at random from Affymetrix spike-in experiments

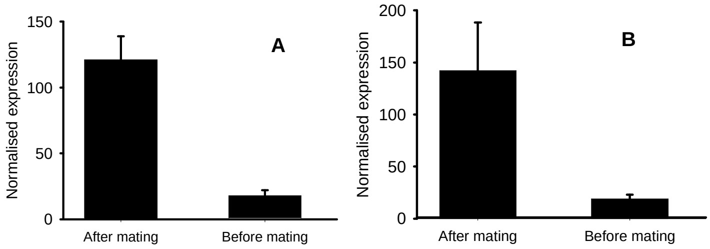

Figure 2. Adrenomedullin (A) and prostaglandin endoperoxide synthase-2 (B) expression values (normalised based on ß-actin expression values) in mice oviduct before and after mating as determined by real time PCR analysis.

[1] Affymetrix (2001) *Microarray Suite User Guide version 5.0.* Santa Clara CA:Affymetrix,Inc.

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

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

[4] Nabney,I.T.(2001). *NETLAB:Algorithms for Pattern Recognition.*Springer.

[5] Spike-In dataset available at http://www.affymetrix.com/analysis/download_center2.affx

[6] Fazeli A(2002). *Transcriptome alteration in mouse oviduct 6 hours after mating*. In: Gordon Research Conferences, Mammalian Gametogenesis & Embryogenesis; 2002; Connecticut College.

[7] Scott, M.J.J, M. Niranjan, R.W. Prager (1998). *Realisable Classifiers: Improving Operating Performance on Variable Cost Problems.* Tech report CUED/F-INFENG/Tr. 323. http://www.spc.eeng.liv.ac.uk/~andyp/papers/scott98parcel.pdf