A probabilistic model for the extraction of expression levels from oligonucleotide arrays

M. Milo*1, A. Fazeli†, M. Niranjan* and N.D. Lawrence*

*Department of Computer Science, University of Sheffield, 211 Portobello Street, Regent Court, Sheffield S1 1DP, U.K. and †Academic Unit of Reproductive and Developmental Medicine, University of Sheffield, Level 4, The Jessop Wing, Tree Root Walk, Sheffield S10 2SF, U.K.

1 To whom correspondence should be addressed

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 mRNA transcripts for thousands of genes. Affymetrix GeneChip array is an oligonucleotide based array technology. In this technology each gene is represented by a set of 11–20 pairs of oligonucleotides that we refer to as probes. Each probe pair is composed of a PM (perfect match) probe, a section of mRNA molecule of interest, and a MM (mismatch) probe that is created by changing the middle base of the PM. To define a measure of the expression level associated to each gene, it is necessary to summarize 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 that we refer to as gamma models for oligonucleotide array signals (gMOS). We use these models to summarize 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 version 5.0) [1] and with RMA (robust multi-array average) expression measure [2]. Our comparison on the spike-in study is expressed in terms of: consistency of fold change, using mean squared differences that may exist among different probe sets used to interrogate the same gene.

detect differential expression, using ROC (receiver operating characteristic) 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 version 5.0 measurements.

Materials and methods

Different methods and models have been proposed to summarize probe level data, all based on an empirical statistical model [1–3]. Here we propose a probabilistic model based on the assumption that the underlying probability distribution for both the PM and MM signals is a γ distribution, Gamma(α, b), with the same inverse scale factor, b, and different shapes, α and a. We can describe the model as follows:

\[ y_{ij} = m_j + s_{ij} \quad \text{with} \quad i = 1, \ldots, N \quad \text{and} \quad j = 1, \ldots, N \]

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 \( n \) is the number of probes in the \( j \)th probe set. Assuming that \( m_j \approx \Gamma(a_j, b_j), s_{ij} \approx \Gamma(a_j + a_j, b_j) \) and \( s_j \approx \Gamma(a_j + b_j, b_j) \) we can derive the following probability expressions:

\[ p(m_{ij} | a_j, b_j) = \frac{b_j^{a_j}}{\Gamma(a_j)} m_j^{a_j-1} \exp(-b_j m_j) \]

\[ p(s_{ij} | a_j, b_j) = \frac{b_j^{a_j}}{\Gamma(a_j)} s_j^{a_j-1} \exp(-b_j s_j) \]

\[ p(y_{ij} | a_j + a_j, b_j) = \frac{b_j^{a_j}}{\Gamma(a_j + a_j)} y_{ij}^{a_j-1} \exp(-b_j y_{ij}) \]

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

The parameters \( a_j, a_j \) and \( b_j \) are estimated by maximizing the joint likelihood:

\[ L(a_j, a_j, b_j) = L(a_j, b_j) + L(a_j + a_j, b_j) \]

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

$$\langle s_j \rangle = \frac{\alpha_j}{b_j} \quad \text{and} \quad \sigma_j^2 = \frac{\alpha_j b_j^2}{\gamma \Gamma(\alpha_j)}.$$ 

In our experiments we used the expected log true probe signal, which can be derived as

$$\langle \ln(s_j) \rangle = \psi(\alpha_j) - \ln(b_j), \quad \text{where} \quad \psi(\alpha_j) = \frac{\partial}{\partial \alpha_j} \ln(\Gamma(\alpha_j)).$$

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 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 four groups and calculated the fold change of all six pairs using the three expression measures. The spike-in dataset is available at http://www.affymetrix.com/analysis/download-center2.affx.

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: gMOS, 0.4%; MAS version 5.0, 0.8%; RMA, 0.7%.

For the sensitivity and specificity study we calculated the number of false positives as the number of non-spiked-in genes with fold change estimate larger than the cut-off value. Conversely, the number of true positives 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 was: gMOS, 0.96; MAS version 5.0, 0.94; RMA, 0.95.

A second comparison was carried out on a dataset obtained from a study investigating the reaction of mouse oviduct to sperm, performed in our laboratory [5]. RNA obtained before and after mating from mouse oviducts was hybridized to MG-U74 Affymetrix array chips. Two genechip arrays were hybridized to RNA samples obtained from oviduct before mating (duplicate) and three to that after mating (triplicate). Table 1 summarizes the results of our analysis in comparison with MAS version 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 version 5.0.

![Figure 1](#) ROC curves for six pairs of arrays chosen at random from Affymetrix spike-in experiments

Table 1 | Summary of the results of our analysis in comparison with MAS version 5.0 and RMA on a selected subset of highly differentially expressed genes before and after mating in the oviduct experiments

<table>
<thead>
<tr>
<th>UniGene</th>
<th>Name</th>
<th>Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Gamma</td>
</tr>
<tr>
<td>Mm.613</td>
<td>Top of form Anp32a acidic (leucine-rich)</td>
<td>7.3</td>
</tr>
<tr>
<td>Mm.3137</td>
<td>Ptgs2, prostaglandin endoperoxide synthase 2</td>
<td>6.3</td>
</tr>
<tr>
<td>Mm.4312</td>
<td>Slc9a1, solute carrier family 9 (Na+/H+ exchanger)</td>
<td>5.7</td>
</tr>
<tr>
<td>Mm.21013</td>
<td>Cxcl1, chemokine (C-X-C motif) ligand 1</td>
<td>6.4</td>
</tr>
<tr>
<td>Mm.108678</td>
<td>Cytip11a1, cytochrome P450, family 11, subfamily a, polypeptide 1</td>
<td>4.1</td>
</tr>
<tr>
<td>Mm.257330</td>
<td>Expressed sequence tags</td>
<td>2.5</td>
</tr>
<tr>
<td>Mm.250422</td>
<td>Serpine 1, serine (or cysteine) proteinase inhibitor, clade E, member 1</td>
<td>2.4</td>
</tr>
<tr>
<td>Mm.245967</td>
<td>Expressed sequence tags</td>
<td>3.9</td>
</tr>
<tr>
<td>Mm.1408</td>
<td>Adm, adrenomedullin</td>
<td>7.2</td>
</tr>
<tr>
<td>Mm.4063</td>
<td>Ndr1, N-myc downstream regulated 1</td>
<td>2.3</td>
</tr>
<tr>
<td>Mm.4639</td>
<td>Cebpδ, CCAAT/enhancer-binding protein, delta</td>
<td>2.1</td>
</tr>
</tbody>
</table>
and prostaglandin endoperoxide synthase 2 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 statistics-based models, outperforming them for consistency of fold change. Moreover in our study we did not perform any background correction or normalization of the data, which are instead carried out by both MAS version 5.0 and RMA. The overlapping of the ROC curves for gMOS and RMA shows that it is possible to combine the two methods [6], in order to achieve better results. This last point is a current matter of research.

**References**


Received 4 July 2003