A Probabilistic model for Affymetrix probe-level data analysis

Xuejun Liu[1], Marta Milo[2], Neil D. Lawrence[2] and Magnus Rattray[1]
[1]School of Computer Science, University of Manchester [2]Department of Computer Science, University of Sheffield
http://www.bioinf.man.ac.uk/resources/puma/

1. Introduction

Microarrays are an essential tool to simultaneously measure gene expression on a large scale. However, they are associated with high levels of experimental uncertainty. Probe-level analysis is required to determine an accurate summary of the expression level for a particular gene in the face of this uncertainty. It is also useful to associate this measurement with a level of confidence that can then be propagated and incorporated into further analyses using probabilistic models or Bayesian methods. Affymetrix GeneChip® arrays are currently the most widely used microarray technology. We devised an Affymetrix probe-level probabilistic model to obtain the uncertainty of gene expression values. Results show that this uncertainty is useful in the analysis of microarray data.

2. multi-mgMOS

multi-mgMOS[2] is the latest version of our gMOS family of models[1,2]. It uses a latent variable to model the probe affinity and allows the binding of specific signal to MM probes. multi-mgMOS is defined by

$$y_{gjc} \sim Ga\left(a_{gc} + \alpha_{gc}, b_{gj}\right), \qquad g - \text{gene}$$
$$m_{gjc} \sim Ga\left(a_{gc} + \phi\alpha_{gc}, b_{gj}\right), \qquad j - \text{probe}$$
$$b_{gj} \sim Ga\left(c_g, d_g\right), \qquad c - \text{chip}$$

Where y_{gjc} and m_{gjc} represent respectively the PM and MM intensities. By integrating out the latent variable b_{gj}, the MAP solution can be found by fast gradient-based optimisation. We can find the posterior distribution of the expected log signal by using truncated Gaussian to approximate $P(\alpha_{gc} \mid D)$.

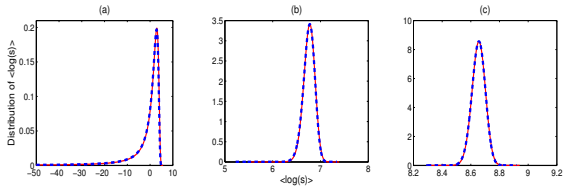


Fig. 1. The posterior distribution of expected log expression level for a spike-in gene at concentration (a) 0, (b) 8 and (c) 512 pM. The thin solid lines are numerically calculated while the thick dashed lines are from the truncated Gaussian approximation.

3. Making use of biological replicates

multi-mgMOS obtains the measurement error for each gene in a single chip. Replicates are commonly used to estimate the variance between chips. The following Bayesian hierarchical model defines the combination of signal from replicates

$$s_{cr} \sim N\left(\theta_c, \sigma^2 + \beta_{cr}^{-1}\right), \qquad r - \text{replicate}$$
$$\theta_c \sim N\left(\mu, \tau^2\right), \qquad c - \text{condition}$$

Where β_{cr} is the measurement precision obtained from multi-mgMOS. In the case where the number of chips and the number of conditions are large, we prefer to use a non-informative prior,

$$P\left(\mu, \log\sigma, \log\tau\right) = \tau.$$

When there are few chips or conditions in an experiment, a conjugate prior may be better,

$$1/\sigma^2 \sim Ga\left(\alpha, \beta\right).$$

In these two cases, we obtain the posterior distribution of the signal θ_c for each condition.

4. Results

We use several data sets to show the usefulness of the measurement error obtained from multi-mgMOS.

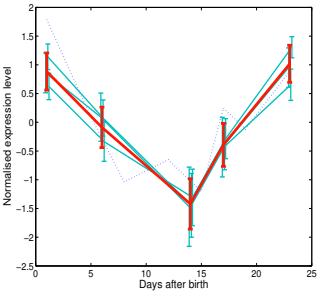


Fig. 2. Temporal profile of gene Fbln1 in the dataset from [3] (25 chips). Error bars are the 2.5-97.5% credibility intervals. Blue lines are from the 3 biological replicates processed by multi-mgMOS and the red line is the combined signal from the 3 replicates using a non-informative prior. The dotted line is the result from quantitative real-time PCR data.

After combination of replicates, the final signal becomes more confident in the lower end, while getting less confident in the higher end.

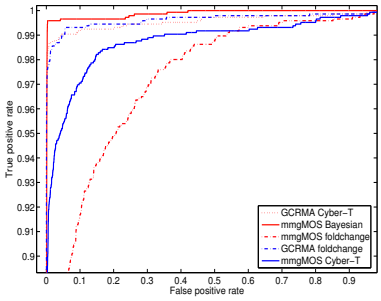


Fig. 3. ROC curves for all pairs of chips for the whole Affymetrix HG-U95a Latin Square data set (59 chips). mmgMOS Bayesian is our Bayesian hierarchical model using a non-informative prior.

With shared between-replicate variance, The Bayesian hierarchical model can easily find the differentially expressed genes, and obtain the dependence between variance and signal for differentially expressed genes across groups.

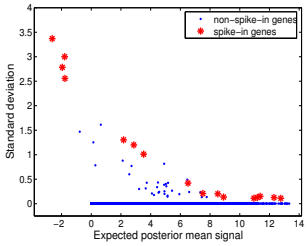
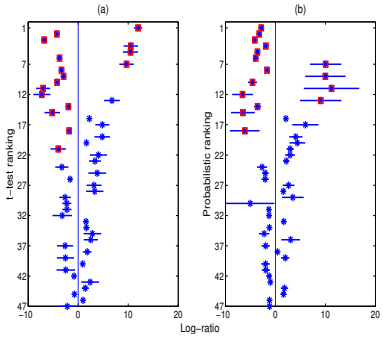


Fig. 4. Standard deviation vs. Log signal



By incorporation of measurement error, the Bayesian hierarchical model reduces the number of false positives.

Fig. 5. The ranking and 5-95% credibility intervals of the posterior distribution of log-ratio of expression level between two randomly selected groups from a spike-in data set. The 16 spike-in genes are indicated with a box. (a) t-test ranking on Cyber-T results, and (b) Bayesian hypothesis test ranking on the Bayesian hierarchical combination with conjugate prior.

5. Conclusion

There are many sources of variability in microarray experiments. It is a waste to discard this uncertainty. We described a probabilistic way to obtain this uncertainty and use it for detecting differential gene expression. Other downstream analyses of microarray data can also be modified to include probe-level uncertainty and this will improve the performance of these downstream analyses in a similar way as we described above. Recent work on PCA[4] demonstrates how probe-level uncertainties can also be incorporated into other probabilistic models, leading to improved methods.

References

[1] Milo M, Fazeli A, Niranjan M, Lawrence ND. A probabilistic model for extracting of expression levels from Oligonucleotide arrays. *Biochemical Society Transactions* 2003; 31:1510-1512.
[2] Liu X, Milo M, Lawrence ND, Rattray M. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics* 2005, doi: 10.1093/bioinformatics/bti583.
[3] Lin KK, Chudova D, Hatfield GW, Smyth P, Andersen B. Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. *PNAS* 2004; 101(45):15955-15960.
[4] Sanguinetti G, Milo M, Rattray M and Lawrence ND. Accounting for probe-level noise in principal component analysis of microarray data, *Bioinformatics* 2005, doi:10.1093/bioinformatics/bti617.