RPA: Probabilistic analysis of probe performance and robust summarization

Leo Lahti$^1$, Laura L. Elo$^2$, Tero Aittokallio$^{2,3}$, and Samuel Kaski$^4$

$^1$University of Helsinki, Department of Veterinary Bioscience $^2$University of Turku, Department of Mathematics
$^3$University of Helsinki, Institute for Molecular Medicine. $^4$Aalto University, Department of Information and Computer Science and University of Helsinki, Department of Computer Science, Helsinki Institute for Information Technology HIIT

14.6.2011

Abstract

Motivation: Probe-level models have led to improved performance in microarray studies but the various sources of probe-level contamination are still poorly understood. Data-driven analysis of probe performance can be used to quantify the uncertainty in individual probes and to highlight the relative contribution of different noise sources. Improved understanding of the probe-level effects can lead to improved preprocessing techniques and microarray design.

Results: We have implemented probabilistic tools for probe performance analysis and summarization on short oligonucleotide arrays. In contrast to standard preprocessing approaches, the methods provide quantitative estimates of probe-specific noise and affinity terms and tools to investigate these parameters. Tools to incorporate prior information of the probes in the analysis are provided as well. Comparisons to known probe-level error sources and spike-in data sets validate the approach.

Availability: Implementation is freely available in R/BioConductor: http://www.bioconductor.org/packages/release/bioc/html/RPA.html.

Contact: leo.lahti@iki.fi

1 Introduction

Probe defects are a major source of noise and uncertainty in microarray studies. Probe performance is affected by RNA degradation, non-specific hybridization, annotation errors and other, potentially unknown factors. While the use of multiple probes and modeling of the probe effects through probe-specific parameters have been shown to yield improved
Figure 1: Demonstration of probe set-level signal estimation for probe set 200046_at in the ALLMLL example data set in BioConductor. Top: the 11 probe-level signals across 20 measurement samples are indicated by gray lines; the black line illustrates the probeset-level summary estimate. Middle: estimated noise level (standard deviation) for each probe. Bottom: probe affinity effects with respect to the probeset-level signal estimate.

estimates of the target signal [6, 8], the various sources of the probe-level noise and their relative contributions remain poorly understood.

We introduce targeted probabilistic tools for investigating probe performance directly based on the expression measurements and independently of external information such as genomic alignments of the probes. The Robust Probabilistic Averaging (RPA) package provides tools to quantify and investigate probe affinity and stochastic noise levels based on the framework introduced in [7], where detailed analysis of probe-level parameters was used to quantify relative contributions from known probe-level error sources such as SNPs, GC-content, genomic mismatches and probe interrogation position along the target sequence. In many cases the source of probe contamination remained unknown, however, highlighting the need for data-driven methods to assess probe performance. The implementation provides tools to assess probe performance and tools to guide microarray preprocessing and probe design in the future studies.
2 Robust Probabilistic Averaging

RPA assumes a Gaussian model for probe effects with a probe-specific mean and variance parameter [7]. These parameters are directly interpretable as a constant affinity term and a stochastic noise level for each probe (Figure 1). The affinity parameter indicates constant shift of the probe-level signal from the probeset-level estimate; low and high binding affinities will lead to reduced and increased signal intensities with respect to the probeset-level signal, respectively. The stochastic noise parameter quantifies the overall accuracy of a probe with respect to the probeset-level signal shape which is common for all probes in a probeset. Probes with smaller variance will follow the probeset-level signal shape more accurately than noisy probes with high variance. Similar affinity parameters are utilized for instance by the widely-used RMA preprocessing algorithm which, in contrast to our model, assumes an equal stochastic noise level for all probes and focuses on probe summarization, while our model is particularly designed for probe performance analysis [7].

Another key difference between RPA and standard probe-level models, such as RMA is the use of probe-level differential expression estimates, which have been shown to improve cross-platform comparability [3, 7]. The model gives tools to quantify probe-level effects and to assess the relative contributions of the many factors that can affect probe performance [7]. Each probe in a probeset is assumed to capture the underlying target signal with probe-specific binding affinity and noise level. Probe performance can be assessed by investigating the observations across multiple arrays. More formally, the probe-level signal for probe $j$ in sample $i$ is modeled in terms of a constant intercept term $\mu$, shape parameter $d_i$, probe affinity $\mu_j$ and stochastic probe-level noise $\varepsilon_{ij} \sim N(0, \tau_j^2)$ as $s_{ij} = \mu + d_i + \mu_j + \varepsilon_{ij}$. The first step of the analysis consists of estimating the signal shape $d = [d_1, \ldots, d_N]$ and probe-specific variances $\tau^2 = [\tau_1^2, \ldots, \tau_P^2]$. Considering the differential expression profile of a given probe $j$, with respect to an arbitrarily selected reference array $r$, the parameters $\mu$ and $\mu_j$ cancel out. This allows efficient estimation of $d$ and $\tau^2$; uncertainty in the reference effect $\varepsilon_{rj}$ is marginalized out [7].

Since its original publication the model has been extended to estimate the remaining terms in the probe-level model. We require that the expected affinity effect of the probes tends to be close to zero by assuming a Gaussian prior $\mu_j \sim N(0, \sigma_j^2)$ for the affinities. Then the expected sum of probe-specific affinities $\mu_j$ is zero. This gives a probabilistic interpretation for RMA, which would obtained by setting identical prior $\sigma^2$ for all probes. However, instead of giving equal weight for all probes in affinity estimation, we consider an alternative approach, where the probes are weighted according to their noisiness by setting $\sigma_j^2 = \tau_j^2$. This yields a more flexible but robust model that takes into account the varying degrees of reliability of individual probes, where the expectation of probe affinities will remain at zero but more noisy probes will have less effect on the estimated signal level $\mu$. In the limit of large sample size, the solution will converge to the mean of the probe-level observations weighted by probe-specific variances that quantify the noise level on each probe; the probabilistic formulation is robust to uncertainties in the data when the sample size is limited. The probe-level estimates quantify probe performance; the probeset-level
signal estimate is useful for preprocessing purposes. Moreover, hyperpriors of the model parameters allow incorporation of prior information on the probes in the analysis.

RPA assumes background-corrected, normalized, and log-transformed probe-level data. The BioConductor package provides standard tools for preprocessing, including support for alternative CDF environments and allows compatible downstream analysis with other BioConductor tools. By default, RPA uses the standard RMA background correction [5] and quantile normalization [1]. See the package documentation for further options.

3 Experimental validation

Comparisons to known probe-level error sources, such as SNPs, GC-content and genomic mismatches, have been used to validate the estimates of probe-specific noise, and RPA has previously been shown to enhance cross-platform comparability in differential gene expression studies [7]. To further validate the model we calculated here the average ranking of various preprocessing methods across the 14 tests on AffyCompII [2] on two spike-in data sets (Supplementary Tables 1-2). Notably, while RPA is primarily targeted at probe performance analysis it also outperformed many widely-used preprocessing algorithms, such as RMA [6], which supports the validity of the probe-level model. Also certain other algorithms, including variants of GC-RMA (Wu and Irizarry, 2001) and FARMS (Hochreiter, 2006), outperformed RMA. The differences in preprocessing performance reflect the complexity of probe-level models: FARMS and GCRMA have more detailed models for probe effects, and RMA is obtained as a special case of the RPA algorithm when the stochastic noise is assumed equal between all probes.

4 Conclusion

Robust Probabilistic Averaging provides tools for probe performance analysis and preprocessing. In contrast to standard preprocessing packages, RPA provides explicit data-driven estimates of the affinity and noise level for individual probes and tools to interpret this information. The information can be used to assess the relative contributions from different probe-level noise sources, to guide preprocessing and to verify the end results of a microarray study. Better understanding of the probe-level effects can ultimately lead to improved probe and microarray design.

References

[1] Bolstad, B. M. et al. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19, 185–193.
[2] Cope, L. M. et al. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20, 323–331. http://affycomp.jhsph.edu/AFFY2/TABLES.hgu/0.html (13 March 2011, date last accessed).

[3] Elo, L. L. et al. (2005). Integrating probe-level expression changes across generations of Affymetrix arrays. *Nucl. Acids Res.*, 33, e193.

[4] S. Hochreiter et al. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics*, 22, 943–949.

[5] Irizarry, R. A. et al. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 249–264.

[6] Irizarry, R. A. et al. (2003b). Summaries of Affymetrix GeneChip probe level data. *Nucl. Acids Res.*, 31, e15.

[7] Lahti, L. et al. (2011). Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Tr. Comp. Biol. Bioinf.*, 8, 217–225.

[8] Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Nat’l Acad. Sci., USA*, 98, 31–36.

[9] Wu, Z and Irizarry, R. (2004). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. In Bourne, P.E. and Gusfield, D. (eds.). *Proc. 8th Conf. Research in Comp. Mol. Biol. (RECOMB’04)*. ACM Press, New York, pp. 98–106.