A POWERFUL DIFFERENTIAL EXPRESSION TEST FOR PROBE-LEVEL OLIGONUCLEOTIDE MICROARRAY DATA

Roland Nilsson[1,3,5], *Johan Björkegren*[1,2,4], *Jesper Tegnér*[1,2,3]

[1] Clinical Gene Networks AB, Stockholm, Sweden.
[2] Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden.
[3] Computational Biology, IFM, Linköping University, Linköping, Sweden.
[4] King Gustaf V Research Institute, Karoliska Institutet, Stockholm, Sweden.
[5] To whom correspondence should be addressed: rolle@ifm.liu.se

ABSTRACT

The most common objective of microarray studies is to look for changes in the expression levels of genes, typically comparing data from two biologically distinct sources such as healthy vs. diseased tissue samples. Since microarray experiments are expensive, results from these studies commonly suffer from low numbers of replicates, rendering low statistical power. Here, we propose an alternative analysis based on a paired test of oligonucleotide probe intensities. Using one separate array for characterizing hybridization noise, we develop a Bayesian statistical test for differential expression. We demonstrate that our approach provides low error rates while requiring only a single array hybridization for each biological sample, making gene expression experiments more affordable. We also show that our statistical test predicts error rates with good precision, allowing the researcher to select gene sets in a more rational way.

BACKGROUND

Genome-wide expression analysis has become an increasingly important tool for identifying gene function, disease-related genes and transcriptional patterns related to drug treatments. The most common measure in expression analysis is the estimate of differential expression between two distinct samples, such as experiments *vs.* controls or diseased *vs.* healthy tissue samples. Although there are typically many sources of variation in this measure unrelated to the biological question at hand, the concept remains central.

The Affymetrix GeneChip technology [1], introduced in the mid-90's, has become the most widely used platform for whole-genome expression analysis. With this technology, each gene is represented by 10-20 "perfect match" (PM) 25-mer oligo probes, complementary to different coding sequence regions. There is also a corresponding set of "mismatch" (MM) probes, where the middle base has been substituted for its complement. These were originally intended to be estimators of non-specific hybridization [1], although later on it became evident that they too bind specifically [2].

The typical goal of GeneChip data analysis is to combine all probe signals into an estimate of *transcript abundance*, a measure of the amount of transcripts present. The method first proposed by Affymetrix was to form the average difference $\langle \mathrm{PM} - \mathrm{MM} \rangle$ across all probes within a probe set. This method is based on two assumptions, both of which has subsequently been shown to be false: (1) by using the $\mathrm{PM} - \mathrm{MM}$ difference, one assumes that the MM probes only measure background noise; and (2) by averaging differences, one assumes that the probes involved have identical binding characteristics. Assumption (1) has been thoroughly discussed elsewhere[3], and we will avoid the details by simply not using MM probe intensities. Assumption (2) has also been addressed by several authors, and several estimators of transcript abundance have recently been proposed, including the Model-based Expression Index (MBEI)[2] and the Robust Multi-array Analysis (RMA) [4].

There is also considerable problems associated with determining the statistical significance of differential expression. The very amount of hypotheses tested (usually on the order of 20,000) means that classic statistical tests designed for a few hypotheses cannot be applied directly. Moreover, all statistical tests rely on some estimate of variation (noise) in the measurements, and it is not clear how to obtain reliable noise estimates given the complexity of probe hybridization.

In the present study, we propose the following analysis method to tackle the problems outlined above. We avoid estimating abundances altogether by using a paired statistic for probe-level data. For this statistic, we estimate the hybridization noise using a separate replicate array. Then, we develop a statistical test that includes this noise estimate as "prior knowledge", and produces a direct estimate of the false positives rates for a given set of selected genes. We demonstrate that our approach provides a reliable measure of differential expression using a single GeneChip per condition, and also predicts error rates with good precision.

METHODS

Probe hybridization model

For a given gene g, we model the normalized log-intensity of probe i in the corresponding probe set as a stochastic variable X_{gi},

$$X_{gi} = p_{gi} + a_g + \epsilon \qquad (1)$$

where p_{gi} is referred to as the *probe effect*, a_g is the log abundance of the transcript of for gene g, and ϵ is the noise contribution, which is assumed to have mean 0. This is the same model used previously by among others Irizarry et al.[4]. Now, consider a microarray experiment assessing differential expression between

two distinct samples. For two arrays, containing different RNA populations, we will obtain two relationships,

$$X_{gi} = p_{gi} + a_g + \epsilon$$
$$X'_{gi} = p_{gi} + a'_g + \epsilon$$

Previous techniques would at this point attempt to estimate a_g and a'_g, and then calculate the log fold change of the gene, which we denote by $\delta_g = a'_g - a_g$. This procedure requires estimation of all probe effects p_{gi}. However, if we are only interested in the log fold change δ_g (which is often the case) we can avoid this step and simply remove the problematic probe bias by forming the difference

$$D_{gi} = X'_{gi} - X_{gi} = \delta_g + 2\epsilon$$

We now assume that ϵ is identical for all probes. Then, the stochastic variable D_{gi} is also identical, and we may estimate δ_g by the mean $D_g = \sum_i D_{gi}/n$. For hypothesis testing, we must also estimate the standard deviation σ of the statistic D_g. This can be obtained with good precision using a single replicate chip, since for replicates, $\delta_g = 0$ for all genes, so $\sigma^2 = \sum_g D_g^2/(n-1)$. We will refer to σ as *technical variation*. Note that none of these estimates require independence between the probes in a given probe set.

Empirical-Bayes analysis

The question of significance of each selected gene requires a careful treatment due to the amount of hypotheses tested. We chose an empirical-Bayes approach [5] to this problem because it makes assumptions (priors) explicit and testable.

Denote by a_0 the fraction of genes that do not differ in abundance, that is, genes for which the null hypothesis $H_0 : \delta = 0$ is true. Let $\pi_1(\delta)$ be the distribution of δ for the remaining fraction $1 - a_0$ of genes. A reasonable choice for a prior $\pi(\delta)$ is then

$$\pi(\delta) = \begin{cases} a_0, & \delta = 0 \\ (1-a_0)\pi_1(\delta), & \delta \neq 0 \end{cases} \quad (2)$$

For all genes, we now assume that the mean D_g is $\sim N(\delta_g, \sigma)$ (the central limit theorem provides some justification for this). The marginal distribution is then (dropping the subscript g for clarity)

$$m(d) = a_0 N(d|0,\sigma) + (1-a_0)m_1(d) \quad (3)$$

where m_1 is the marginal density of D with respect to π_1,

$$m_1(d) = \int N(d|\delta,\sigma)\pi_1(\delta)\,\mathrm{d}\delta$$

In other words, the marginal distribution, which is what one would observe in a histogram of d (figure 1), is a mixture of $N(d|0,\sigma)$ from the genes that have not changed, and another distribution $m_1(d)$ from those that have.

We then obtain the posterior probability of H_0 from Bayes theorem:

$$\pi(\delta = 0|d) = \frac{a_0 N(d|0,\sigma)}{m(d)}$$

To compute this probability, we must determine the shape of π_1 and the value of a_0. Assuming $\pi_1 \sim N(\mu,\tau)$, the marginal (3) becomes a gaussian mixture, and we can easily estimate the parameters using a standard expectation-maximization (EM) algorithm [6]. It is certainly not clear that π_1 should be normal; however, it can be shown that the exact functional form has little bearing on the resulting posterior probability, as long as we have a reasonable estimate of the variance of π_1 [5](pp.151).

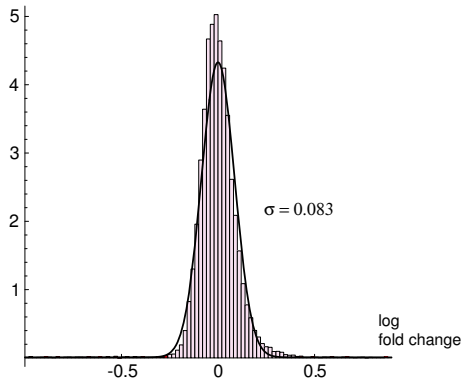

Fig. 1. Plot of the marginal $m(d)$ (solid line) as estimated by the EM algorithm, together with the corresponding data histogram, for configuration 1.

Data set

The data used in this paper is publicly available from Affymetrix [7]. It contains 42 transcripts added at known concentrations, referred to as *spike-in* genes. The concentrations are exponentially spaced as $(1/8, 1/2, \ldots, 256, 512)$ pM. In all experiments, an identical background from human cells is present in addition to the spike-ins. We refer to this data set as the Latin Square data set.

RESULTS

By comparing spike-in genes from the Latin Square data set against each other in various pairwise arrangements, different fold changes can be obtained. We arranged the data in two configurations, obtaining two "virtual" experiments with (1) 1638 spike-in genes, with fold changes exponentially spaced as $2, 4, 8 \ldots, 4096$, and (2) 1512 spike-in genes, all with fold change 2. In addition to the spike-ins we added the background from a random pair of replicates chips. We refer to these constructed data sets as configuration 1 and 2, respectively. For both configurations, data was normalized at the probe level by the quantile-normalization method [8] and log-transformed (base 2). No background correction was used in our analysis; we tried the procedure proposed by Irizarry et al.[10], but found that it did not improve our results. We then computed the statistic d for each gene. Figure 2 shows plots of d versus average signal for the two configurations, demonstrating a low amount of false positives (thin black region). Note the underestimation of fold changes in fig. 2A, where the largest $\log_2$ nominal fold change is 12, and also the inconsistency of fold change estimates in fig. 2B, especially for low intensity transcripts.

We estimated the technical variation σ to 0.083 and 0.078 for configuration 1 and 2, respectively (differences reflect chip reproducibility). We then estimated a_0, μ and τ using the EM algorithm (Table 1). The algorithm has a slight tendency to overestimate the fraction of H_0 close to $\delta = 0$, which is understandable. These small discrepancies do not seem to have any discernible impact on the subsequent analysis. The fit of the estimated marginal distribution (3) is shown for configuration 1 in figure 1; the H_1 fraction is too small to be visible here.

Using the above estimates we calculated the posterior proba-

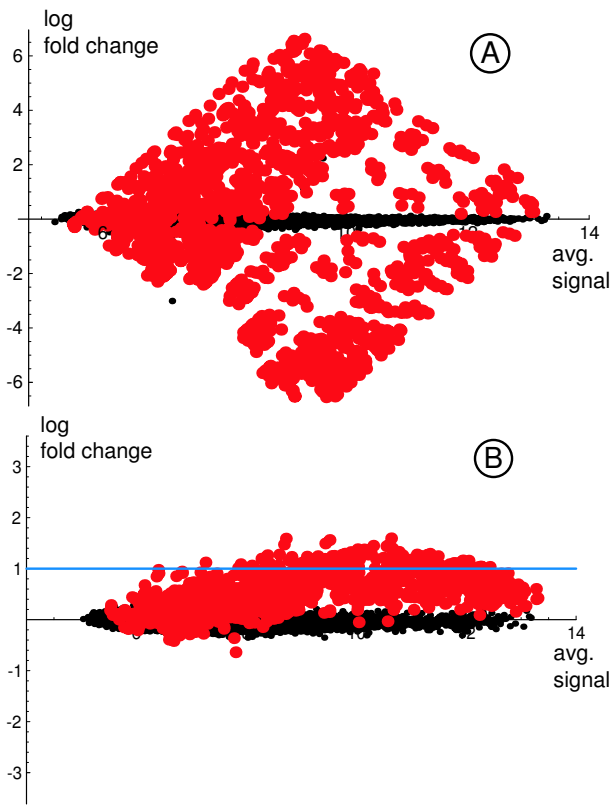

Fig. 2. (A) For configuration 1, a plot of estimated $\log_2$ fold change d versus average $\log_2$ signal $(x' + x)/2$. Spike-in genes are highlighted in red. (B) Same as (A) for configuration 2. Here, the blue line marks the expected 2-fold difference.

bility of no differential expression $\pi(\delta = 0|d)$. From this probability we determined gene ranks and estimated the expected number of false positives ($\widehat{\text{FP}}$) and true positives ($\widehat{\text{TP}}$) in a selected set of genes G as

$$\widehat{\text{FP}} = \sum_{g \in G} \pi(\delta_g = 0|d_g)$$

$$\widehat{\text{TP}} = |G| - \widehat{\text{FP}}$$

These estimates are plotted along with true FP and TP values in Receiver Operator Characteristic (ROC) curves (figure 3). For reference, we also compute classical p-values with the Bonferroni correction, and make the corresponding error estimates using these. Our method is slightly overoptimistic for configuration 1, possibly because of the extreme fold changes in this set (up to 4096-fold). Overall though, the agreement is good enough to warrant the use of this error estimate for gene selection. In contrast, the classical p-values are clearly not suitable: even with the conservative correction (Bonferroni), the error rates are severely underestimated in both configurations. The Area Under Curve (AUC) measure found with our method (figure 3) is on par with the best scoring methods in the Affycomp benchmark [11], although direct comparisons are not possible since our data comes from a more recent GeneChip version.

		EM Estimates	True values
configuration (1)	a_0	0.931	0.931
	μ	-0.18	- 0.15
	τ	3.03	3.04
configuration (2)	a_0	0.945	0.936
	μ	0.57	0.50
	τ	0.35	0.37

Table 1. Estimates of the prior parameters from the EM algorithm. True values for τ and μ refer to mean and standard deviation of the set of spike-in genes, calculated directly.

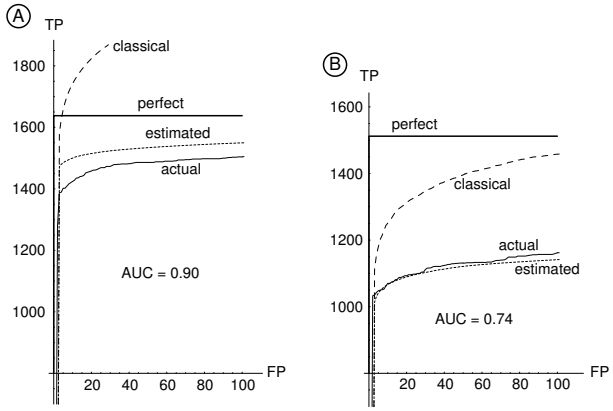

Fig. 3. (A) ROC curve for configuration 1, with genes selected by $\pi(\delta|d)$. Curves displayed are the perfect ROC (thick, solid); actual ROC (thin, solid); ROC predicted by our method (dotted); ROC predicted by classical Bonferroni-corrected p-values (dashed). Area under ROC curve (AUC) is calculated for the actual ROC. (B) Same as A, for configuration 2.

We next determined confidence intervals for δ from the posterior, given the estimated fold change d. Confidence intervals are quite narrow (about $d \pm 0.12$ at the 95% level), but unfortunately there is substantial bias in the estimates d when the true fold change is large. Figure 4 shows true versus estimated log fold changes for the spike-in probes in configuration 1. There is a considerable underestimation of fold changes, especially at higher δ. The Pearson correlation coefficient is only 0.81. We believe that this low agreement originates in nonlinear, sequence-dependent characteristics of Affymetrix probes [12, 13] which is not adequately captured by the linear model (1). Colors indicate the average expression $(a + a')/2$ in both arrays (using the known concentrations). It is clear that a lower overall expression level compresses fold changes more than a higher level. This has previously been attributed to lack of background correction [10], but we were not able improve the performance more than marginally using the background correction proposed in [10].

DISCUSSION

The Affymetrix GeneChip technology is currently the most popular platform for global gene expression analysis, an approach generally expected to be a key tool for discovering gene functions.

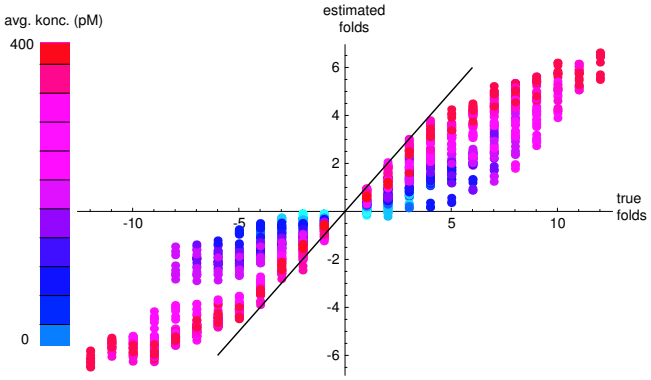

Fig. 4. True versus estimated fold changes for configuration 1 (see text). Colors indicate the average actual concentration in the two arrays. Solid line marks $d = \delta$.

The concept of differential expression plays a central role in this field. In the present paper, we have proposed an alternate method for measuring differential expression. We have showed that our method can provide an accurate measure of differential expression using a single GeneChip per condition. We have also demonstrated that our method provides accurate predictions of the error rates for a given set of selected genes.

Hitherto, most studies have estimated technical variation separately for each gene, using several replicate GeneChips per condition. The main advantage to our approach stems from the treatment of technical variation as constant across genes, which allows us to assess statistical significance using a single array per condition. This can obviously save substantial time and resources for microarray researchers. Our statistical treatment does not depend on the exact method of fold change estimation; in fact, it may even be applicable to other types of microarrays, as long as the underlying assumption of constant σ remains valid. For our data, we have verified that the GeneChips used are well reproducible concerning σ (estimates ranging from about 0.07 to 0.09 $\log_2$ fold changes), and also that σ is indeed uniform for all genes, independent of signal intensity, motivating the use of a single estimate.

In fact, the only case where this assumption does not hold is when there actually is a substantial difference in RNA levels. In this case, deviations from σ may well be evidence of nonlinearity in probe hybridization, which cannot be captured by the linear model (1). If so, gene-specific estimates would include model bias, misleading the statistical analysis. This line of reasoning is also consistent with the fact that, although our statistical model seems promising, we have not been able to achieve accurate quantitative estimates of fold changes with our method (figure 4).

For microarray researchers, the most useful output of our method is probably the estimated error rates for a given set of selected genes. The accuracy of these predictions rely on our estimates of the parameters in the prior. Hence, we carried out a sensitivity analysis for these parameters and found that σ is the most sensitive parameter. Fortunately, this parameter is also the easiest to estimate, given the large amounts of data used. For the remaining parameters, the method is fairly robust.

In closing, we must emphasize that we have not considered biological variation in this paper: having determined a statistically significant change of RNA levels between two samples says little of the biological interpretation of that change. On the other hand, a reliable technical procedure opens up possibilities of investigating biological variation. For example, a set of 10 arrays, each measuring the transcriptional profile of a human tissue for different individuals or conditions, would allow for $10 \cdot 9/2 = 45$ comparisons in an all-to-all scheme, providing more information for higher-level data analysis.

1. REFERENCES

[1] Robert J. Lipschutz, Stephen P.A. Fodor, Thomas R. Gingeras, and David J. Lockhart, *High density synthetic oligonucleotide arrays*, Nature Genet. **21** (1999), 20–24, Supplement.

[2] Cheng Li and Wing Hung Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection*, Proc. Natl. Acad. Sci. USA **98** (2001), no. 1, 31–36.

[3] Felix Naef, Daniel A. Lim, Nila Patil, and Marcelo Magnasco, *DNA hybridization to mismatched templates: a chip study*, Phys. Rev. E **65** (2002), no. 040902.

[4] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed, *Summaries of Affymetrix GeneChip probe level data*, Nucleic Acids Res. **31** (2003), no. 4, e15–.

[5] James O. Berger, *Statistical decision theory and Bayesian analysis*, 2 ed., Springer series in statistics, Springer-Verlag New York, Inc., 1985.

[6] A.P. Dempster, N.M. Laird, and D.B.Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. (Ser. B) **39** (1977), 1–38.

[7] `www.affymetrix.com/support/datasets.affx`.

[8] Douglas C. Montgomery, *Design and analysis of experiments*, 5 ed., Wiley, 2001.

[8] B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*, Bioinformatics **19** (2003), no. 2, 185–193.

[10] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence C. Speed, *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*, Biostatistics **4** (2003), no. 2, 249–264.

[11] Leslie M. Cope, Rafael A. Irizarry, Harris A. Jaffee, Zhijn Wu, and Terence P. Speed, *A benchmark for Affymetrix GeneChip expression measures*, Bioinformatics, In press, 2004.

[12] Doeke Hekstra, Alexander R. Taussig, Marcelo Magnasco, and Felix Naef, *Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays*, Nucleic Acids Res. **31** (2003), no. 7, 1962–68.

[13] G. A. Held, G. Grinstein, and Y. Tu, *Modeling of DNA microarray data by using physical properties of hybridization*, Proc. Natl. Acad. Sci. USA **100** (2003), no. 13, 7575–7580.