Sensitivity and specificity of five abundance estimators for high-density oligonucleotide microarrays

Andrew C. James, Jim G. Veitch, Ali R. Zareh and Timothy Triche

1Research School of Biological Sciences, Australian National University, PO Box 475, Canberra ACT 2601, Australia, 2Corimbia Inc., Berkeley, CA 94710, USA and 3Department of Pathology, Children’s Hospital Los Angeles, CA 90027, USA

Received on August 24, 2003; accepted on November 3, 2003
Advance Access publication February 5, 2004

ABSTRACT

Motivation: A number of algorithms have been proposed for the processing of feature-level data from high-density oligonucleotide microarrays to give estimates of transcript abundance. Performance in the common task of detecting differential expression between samples can be quantified by the statistical concepts of sensitivity and specificity, and represented by the use of receiver operating characteristic curves. These have been previously presented for small numbers of genes known to be differentially present in spiked-in samples. We present here a study of performance over a large number (thousands) of transcripts for which there is strong evidence of differential expression, with corresponding false positive rates controlled by comparisons between replicates.

Results: The straight-line regression analysis of a mixture series with replicates by five estimation algorithms produces a consensus set of 4462 transcripts with differential expression of agreed direction and high significance \( p < 0.01 \) according to all algorithms. The more difficult task of two-sample tests between adjacent mixture levels produces performance curves of fraction true positive detected against significance level. Performance varies significantly between algorithms: at \( p < 0.01 \) level, the detection rate varies between 41 and 66%. A control using comparisons between replicates at the same levels indicates that the tests produce empirical false positive rates closely matching the nominal \( p \)-values.

Contact: andrew.james@anu.edu.au

INTRODUCTION

Affymetrix GeneChip technology is widely used for the parallel assessment of the level of expression of large numbers of genes. Each measured gene is assayed by a set of features or cells comprising a probeset, consisting generally of 11–20 probe pairs. The perfect match probes are 25-base oligonucleotides exactly matching a segment of the gene of interest. Each is paired with a mismatch probe, which has the complementary base substituted in the 13th position. After hybridizing a fluorescently labeled sample of RNA to the chip, washing, and optionally amplifying with an antibody stain, the chip is scanned and the resulting image processed to give an intensity value for each feature.

The processing of these probe level intensities to give levels of gene expression can be termed as abundance estimation. A number of algorithms have been proposed and implemented in software packages. Each consists essentially of two major steps. The first step transforms data for each chip, attempting to remove non-meaningful chip-to-chip variation, particularly in sample concentration, scanner gain and background offset. This typically includes two operations, background correction and normalization. The second step is probeset summarization, which combines the intensities of a probeset to give a single abundance measure for each assayed transcript. An optional further step might normalize again, transforming the abundance estimates of each chip. Although these steps can be considered separately, for the researcher deciding which system to use, the relative performance of a complete analysis system is of interest, and will be studied here.

A common goal of a gene expression study is to detect the transcripts with differential expression levels between samples representing differing conditions. This paper assesses the relative performance of five abundance estimation systems by measuring the sensitivity and specificity of detection of differential expression over a large population of transcripts. Sensitivity (also called power) is defined as the probability of detecting a difference, when the difference is really there, i.e. the true positive rate. Specificity is the probability of correctly calling no difference, when there really is none; hence, it is one minus the false positive rate. Each of these rates is generally a function of a parameter, which controls the stringency of the test. At low stringency, most differences are detected (high sensitivity), but there will be more false positives as
Sensitivity and specificity of abundance estimators

well (low specificity). At high stringency, the converse will occur. The overall performance of the detection process can be summarized by a curve of true positive rate on false positive rate, termed the receiver operating characteristic (ROC) curve, developed in the context of radar signal detection.

The curve produced depends not only on the analysis algorithms, but also on the two sets of test cases, the real positives and real negatives, that are considered. For example, if the real differences are large, sensitivity will be high for each given specificity. Hence what is of interest is relative performance on a common set of cases representative of real-world data. To calculate empirically the true and false positive rates for a detection system requires having test cases known to have real difference and no real difference, respectively. Cases with no real difference are readily produced by considering replicate hybridizations of the same sample. For test cases of real differences, some studies have used the transcripts in spike-in datasets which were known to be at different levels (Irizarry et al., 2003a; Lemon et al., 2002), however, these studies typically only spike-in a relatively small number of transcripts, the other transcripts being a constant background sample. This paper takes an alternative approach: considering replicated data along a mixture series between two distinct tissue samples, and performing stringent tests to select a consensus subset of transcripts with strong evidence of differential expression. This procedure from Bioconductor release 1.1 was run with default background correction being used.

ALGORITHMS

MAS5 algorithm

This algorithm (Affymetrix, 2001, http://www.affymetrix.com/support/technical/manuals.affx) is often the default choice of users. It has default background correction being subtraction of spatially interpolated values of the mean of lowest 2% of intensities calculated in 16 sectors of the chip. The normalization is then to divide by the trimmed mean with the extreme 2% of values at each end excluded. Difference values are taken by subtracting perfect match minus corresponding mismatch values. For each transcript probeset, let PM$_{ij}$ and MM$_{ij}$ denote the perfect-match and mis-match intensities for probe pair $j$ of sample $i$. In the case of MM$_{ij}$ > PM$_{ij}$, an ideal mismatch value IM$_{ij}$ is imputed from the relationship

\[ IM_{ij} = \frac{2}{\text{IQR}} \text{MM}_{ij} - \mu \]

where symbols are as for the Li–Wong model, but an offset $v_j$ is included for each probe pair, allowing that the PM and MM values can be classed as outliers according to criteria based on the standard errors or residuals, and excluded from fitting, with iteration to convergence. A variation that is set as the default in dChip version 1.2 is to fit the bilinear model to only the PM values, ignoring the MM values. The abundance estimates from these two algorithms have been termed model-based expression indices (MBEIs). We label these two variants as dChipDiff and dChipPM, respectively.

Robust multichip analysis (RMA)

RMA (Irizarry et al., 2003a,b) (available from http://www.bioconductor.org/) is a multichip log-scale algorithm. It subtracts a background intensity estimated pointwise by decomposing observed intensities into signal plus background, with assumed parametric distributions for signal and background fitted chipwise (Irizarry et al., 2003a). Quantile normalisation is then done over a batch of chips (Bolstad et al., 2003). For each probeset, the logs of the resulting positive intensity signals are fitted with an additive model having a chipwise abundance term and a probewise affinity term. Fitting is by a robust median polish procedure (Tukey, 1977). Estimates are reported on the log2 scale. In this study, the RMA procedure from Bioconductor release 1.1 was run with default parameters, and the log2-scale estimates were then transformed to linear scale by antilog base 2 to compare with other algorithms.

The dChip algorithm

Li and Wong (2001a,b) introduced a model addressing the variation in probe affinity within a probeset, implemented in the software package dChip (http://www.dchip.org/). It processes batches of chips, first performing a nonlinear normalization (Li and Wong, 2001b). The original algorithm then fitted for each probeset a bilinear model, expressing the differences PM$_{ij} - \text{MM}_{ij}$ for sample $i$ and probe pair $j$ as the product of an abundance value $\theta_i$ and a probe pair affinity value $\phi_j$:

\[ \text{PM}_{ij} - \text{MM}_{ij} = \theta_i \phi_j + \epsilon_{ij}. \]

Estimates of abundances and affinities are fitted by unweighted least squares. Probe pairs, chips or individual values can be classed as outliers according to criteria based on the standard errors or residuals, and excluded from fitting, with iteration to convergence. A variation that is set as the default in dChip version 1.2 is to fit the bilinear model to only the PM values, ignoring the MM values. The abundance estimates from these two algorithms have been termed model-based expression indices (MBEIs). We label these two variants as dChipDiff and dChipPM, respectively.

The ProbeProfiler algorithm

ProbeProfiler is a software package implementing an alternative model-based algorithm, along with a variety of quality control features (http://www.corimbia.com/). Normalisation is done by division by the interquartile range (IQR) (75th minus 25th percentiles), in each of 16 sectors of the chip. Probeset summarization is done by a principal components analysis (PCA) of PM – MM differences. This is equivalent to a bilinear model with offset:

\[ \text{PM}_{ij} - \text{MM}_{ij} = \theta_i \phi_j + v_j + \epsilon_{ij}, \]

where symbols are as for the Li–Wong model, but an offset $v_j$ is included for each probe pair, allowing that the PM and
MM probes may not have the same background intensity at zero abundance. In PCA terms, the affinity coefficients are the weights and the abundances are the scores. Considering the data as vectors in the probe-space indexed by \( j \), the model fits the line of least perpendicular squared distances through the cloud of sample points (Fig. 1). The background vector \( \nu \) offsets from the origin, the affinity vector \( \phi \) defines the direction of the line of fit and the abundance estimate \( \theta_i \) for each sample parameterizes distance along the line of fit. There is indeterminacy in the scale of affinities, resolved by constraining their sum of squares to be one. The sign of the direction vector is hence taking the vector \( \phi \) as being perpendicular to the affinity direction (scaled by five times). In the full dimensional space these vectors are orthogonal. Note the significant offset defines the affinity direction (scaled by five times). In the full dimensional space these vectors are orthogonal. Note the significant offset defines the affinity direction (scaled by five times). In the full dimensional space these vectors are orthogonal. Note the significant offset.

Since, the sum of squares of affinities is unity, and the sum of products of affinities and background offsets is zero. Probes for which the apparent difference affinity is negative have MM values increasing with abundance more rapidly than PM, and are excluded from analysis. Point outliers are flagged when beyond a fixed multiple of the spread expected at a given probe as measured by median absolute deviations. These outliers as well as points affected by photomultiplier tube saturation are replaced by imputed values. An initial imputed value is taken as the mean value of unaffected data in the probeset of the point, estimating a probe-effect profile across the samples, then taking as final imputed value the prediction from fitting the probe-effect profile to the unaffected points of the probeset. Where saturation was detected, the larger of the raw value and the imputed value is taken, hence using the information that the probe was at least as bright as the saturating value. A second normalization of abundance estimates on each chip is finally done, scaling to constant interquartile range.

**DATA AND ANALYSIS**

The dataset used is a mixture series generated by GeneLogic (http://quolotus02.genelogic.com/datasets.nsf), consisting of five replicates at each of five levels of mixture (0, 25, 50, 75, 100%) of a central nervous system (CNS) sample with the complementary amount of a liver sample. Samples were hybridized to HG_U95Av2 chips (however denoted as HG_U95A chips in some distributed data). These data were part of a larger study also including dilution series of the liver and CNS samples. We use the mixture component of the study here because it involves a contrast between two tissues, with all chips run at the recommended RNA concentration levels, providing a reasonable correspondence with routine laboratory usage. The 25 CEL files of the series were processed with each of the analysis systems, using the standard background correction, normalization and abundance estimation settings. Five scanners had been used in scanning the replicate chips, and we rely on the normalization algorithms to remove scanner effect. This provides a test for a common problem in microarray usage, the variation in scanner gain between chip scans.

The true level of a given transcript should lie on a straight line when plotted against proportion of CNS in the mixture, either rising, flat or falling depending on relative amount in the liver sample. To determine a subset of transcripts that had strong evidence of differential expression between the liver and CNS samples, for each algorithm and transcript, a regression line was fitted by least squares, for abundance on CNS proportion. The regression line uses all available information in the experimental design. For each fit we note the sign of the slope, and the significance level of a \( t \)-test for the slope being non-zero. The subset of transcripts was then taken for which all algorithms indicated the same sign, and had high significance level (\( p < 0.01 \)) for non-zero slope. This consensus
subset is taken as having real differential expression, by a stringent criterion which is fair with regard to each algorithm. It contained 4462 transcripts, equal to 35.3% of those assayed on the chip. We then assess the performance of the algorithms on the more difficult task of detecting differential expression by two-sample $t$-tests between adjacent mixture levels.

Expression estimates on the linear scale were transformed by the generalized log transform (glog) (Durbin et al., 2002; Huber et al., 2002; Munson, 2001, http://stat-www.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/Munson.ppt). This transformation approximately stabilizes variance for data having additive plus multiplicative noise characteristics. It converges toward the log2 transform for large data values, where multiplicative noise dominates, but becomes closer to a linear scaling with offset at low levels, where additive noise dominates. The transformation has one parameter, $\lambda$, which is set to the data value at which the additive and multiplicative noise components have equal variance. In principle, this means the optimal parameter might be different for each transcript as well as for each algorithm. However, for parsimony a single value was estimated for each algorithm, taking the maximum-likelihood estimate of $\lambda$, for the model that the replicate data at each level should be normally distributed about replicate means with uniform variance after transformation (Durbin and Rocke, 2003). For comparison, the analysis that follows was also done on untransformed expression estimates, and also on estimates transformed by a simple log, with negative values treated as missing. The results obtained are essentially the same, and are included in the supplementary material.

For each algorithm, a series of two-sample $t$-tests were performed to detect differential expression. True positive rates were estimated by testing for differential expression between the sets of five replicates at adjacent mixture levels, i.e. 0–25 CNS, 25–50, 50–75 and 75–100%. For those transcripts assumed to have real differential expression according to the consensus of regression tests, a one-sided $t$-test was done for change in the appropriate direction, resulting in four $p$-values for each of 4462 transcripts, totaling 17 848 $p$-values.

To confirm empirically the false positive rate, $t$-tests were done between groups artificially formed from replicates at the same mixture level, for which there should be no difference besides noise. At a given level, the five replicates were divided into a group of two and a group of three. Two-sided $t$-tests were done for each of the 10 possible ways that the division can be made, and this was repeated for each of the five levels, resulting in a total of 50 $p$-values for each of 12 625 transcripts.

RESULTS AND DISCUSSION

The tests for differences on transcripts in the consensus true-difference set produce 17 848 $p$-values for each algorithm. Figure 2 plots the fraction of tests detecting significant difference as a function of $p$-value, for each of the five algorithms. As the $p$-values of interest are small values, indicating significance, such as 0.05, 0.01 or better, we have found it more useful to plot the curve on a reversed log10 $p$-value axis, shown in Figure 3, for which fraction of differences detected declines as significance increases towards the right.
As a control on the assumptions of the $t$-tests, the fraction of false positives for the tests between artificial groups defined at equal mixture levels are also plotted for each algorithm. These are the dashed curves of Figure 3, lying almost on top of the line $X = Y$. On the $-\log_{10}$ axis, the line $X = Y$ becomes the negative exponential curve. The agreement between these empirical false positive rates and the nominal $p$-values is very close and, hence, it appears that all algorithms generate $t$-statistics that follow the true $t$-distribution closely.

While we cannot derive empirical CDFs for the statistics for a $t$-test of two groups of 5, the assumptions of the $t$-test (underlying normality in the distributions of the means) should be even better met than in a $t$-test of 2 versus 3, and hence the observed empirical CDFs for the case of two groups of 5 can be expected to match the theoretical $t$-distribution even more closely.

An ROC curve is a plot of true positive rate on empirical false positive rate. If empirical false positive rate differed from the nominal $p$-value, a ‘realized’ ROC curve would be the trajectory of true positive rate against empirical false positive rate, with nominal $p$-value as an implicit parameter. However, here the concordance of empirical false positives with nominal $p$-value is so high that the curves are essentially unchanged by such an adjustment (see supplementary material).

The detection rate curves show a significant spread of performance between the algorithms, plotted in increasing order as MAS5, dChip Difference, dChip PM only, RMA and ProbeProfiler. The percentages of differences found at significance level $p < 0.01$ are 41, 52, 56, 60 and 66, respectively.

Figure 4 shows percentage detection within stratified bands of average expression level. The 4462 transcripts were stratified into 10th percentile bands of 446 transcripts each, according to a consensus value indicative of abundance level. For each transcript, this value was taken as the mean over algorithms of the log of the abundance at the middle of the straight-line fits done previously. For each band, the percentage of the 446×4 pairwise tests that were significant at $p < 0.01$ were calculated. The curves show generally increasing performance as abundance increases, with the exception of the highest 10% of transcripts, for which four algorithms decline in performance. The order of curves is the same as for the detection curves. The ProbeProfiler curve is uniformly highest, climbing from 46 to 79%. MAS5 is lowest except for the highest band, climbing from 15 to 59%, with the gap particularly wide at lower abundances. The RMA curve runs second highest, but with a relative fall for the lowest 10% band.

A major factor influencing the intensities observed in a probeset is the differing response gain, or affinity, of the probes (Li and Wong, 2001a). Four of the five algorithms that are considered address this, the two dChip and the ProbeProfiler algorithms by fitting bilinear models of the form abundance times affinity, and RMA by fitting an additive probe effect on the log scale. MAS5 does not allow for probe effect, and ignoring this large effect might explain the overall lower performance of this algorithm on the detection task. Although ignoring an additive factor should only lead to an offset in result, equivalent to a scaling in the linear domain, the biweighting procedure risks treating a probepair with unusually great affinity as being an outlier and down-weighting useful signal.

The noise present in microchip data appears to have both multiplicative and additive components (Rocke and Durbin, 2001). The log transform stabilizes multiplicative variance. However at low levels, where additive noise becomes significant, log transformation causes an increase in variance, to arbitrarily large values as zero is approached. The two log domain algorithms MAS5 and RMA each try to solve this problem by using different methods of thresholding: they prevent signal estimates getting too close to zero by imputing a larger value when the observed signal is ‘low’. This trades the increase in variance for an increase in bias, but either way real variations in low signal will be lost. This is likely to be a problem with these two log-domain algorithms, which do indeed have relatively deteriorating performance at the low abundance end. From the biologists’ perspective, these lower abundance genes are often genes associated with regulatory mechanisms such as signal transduction. The overall lower detection rates of the RMA algorithm compared with the essentially least-squares fitting of ProbeProfiler could indicate that the robust fitting of RMA might be more robust than is necessary, and give too much sacrifice in efficiency.

The ProbeProfiler algorithm fits a bilinear model similar to the dChip difference algorithm, but with different outlier...
detection, and with the model allowing an offset from the origin for difference value at zero abundance. Figure 1 illustrates an example of the line of best fit, with the offset from zero, projected into the coordinates of two of the probe pairs. As in this example, this offset is frequently significant, and forcing the line of fit to pass through the origin, as the dChip Difference algorithm does, can lead to a set of affinity values with lower discrimination. In extreme cases an affinity value in dChip Difference might appear to be negative, leading to the exclusion of that probe pair, whereas with an offset allowed, the affinity is still positive and provides useful discrimination. The dChip PM only algorithm outperforms the difference version in this analysis, lending support to the choice of settings in dChip version 1.2, whereby PM only is the default model.

A relative drop in ProbeProfiler performance for the lowest 10% band indicates some room for improvement at the lowest abundances. However, at the \( p < 0.01 \) level, over all abundances, the true positive fraction translates to an increase in the number of detections by 10, 17, 27 and 63% relative to the other four algorithms.

An apparently paradoxical effect is the decline in performance of four algorithms at the highest abundance levels (Fig. 4). This might be because at these abundance levels the probe features are becoming chemically saturated, whereby the limitation in available probe molecules causes a relatively small gain in intensity for a given difference in concentration.

This study indicates that quite widely varying detection rates are produced by different algorithms in the task of detecting differential expression over a large number of transcripts. It suggests that significant additional value can be extracted from high-density oligonucleotide array data by statistically based processing of the data, utilizing probe-specific properties observed over sets of samples, with the interesting result that the relatively classical methodology of principal components analysis in this case has produced the best results.

ACKNOWLEDGEMENTS

We thank Gene Logic Inc. for permission to use the mixture series data. This work was supported in part by a grant from the National Cancer Institute to T.T., entitled ‘Director’s Challenge: Towards a Molecular Classification of Tumors’, grant number CA 88199-01. Additional support was provided by the Beckman Research Foundation, to the Beckman Macular Research Center at Doheny Eye Institute, Keck School of Medicine, also to the same author.

SUPPLEMENTARY DATA

For Supplementary data please refer to Bioinformatics online.

REFERENCES


