Non-Linear Analysis of GeneChip Arrays
Diana Abdueva\textsuperscript{a,c}, Dmitriy Skvortsov\textsuperscript{a,c}, and Simon Tavaré\textsuperscript{a,b},

\textsuperscript{a}Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1340, USA
\textsuperscript{b}Department of Oncology, University of Cambridge, Cambridge, England
\textsuperscript{c}These authors contributed equally.

Abstract

The application of microarray hybridization theory to Affymetrix GeneChip data has been a recent focus for data analysts. It has been shown that the hyperbolic Langmuir isotherm captures the shape of the signal response to concentration of Affymetrix GeneChips. We demonstrate that existing linear fit methods for extracting gene expression measures are not well adapted for the effect of saturation resulting from surface adsorption processes. In contrast to the most popular methods, we fit background and concentration parameters within a single global fitting routine instead of estimating the background prior to obtaining gene expression measures. We describe a non-linear multi-chip model of the perfect match signal that effectively allows for the separation of specific and non-specific components of the microarray signal and avoids saturation bias in the high intensity range. Multimodel inference, incorporated within the fitting routine, allows a quantitative selection of the model that best describes the observed data. The performance of this method is evaluated on publicly available data sets, and comparisons to popular algorithms are presented.

Key words: high-density microarrays, non-linear fit, gene expression, Langmuir adsorption

version = 1.0 2006 January 11

* Corresponding author.
Email address: abdueva@usc.edu (Diana Abdueva).
Introduction

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 Affymetrix GeneChip technology, introduced in 1996, has become one of the most widely used platforms for whole-genome expression analysis. Each gene is represented on the GeneChip by 10-20 ‘perfect match’ (PM) probes, consisting of 25 bp complementary to different coding sequence regions of the gene in question. In addition, there is a corresponding set of ‘mismatch’ (MM) probes, where the middle base has been substituted with its complement.

Most software packages available for the calculation of gene expression levels from fluorescence intensities rely on algorithms of a purely statistical or empirical nature; Affymetrix Microarray Suite 5.0 is based on the Tukey biweight algorithm [1], DNA-Chip Analyzer (dChip) implements a linear model for expression analysis of oligonucleotide arrays [2] and Bioconductor’s RMA employs a log-additive robust multi-array analysis [3]. More recently, several algorithms based on physical properties of the hybridization process have been introduced. GCRMA performs a background adjustment using sequence information [4] and Zhang et al. [5] implement a dependent nearest-neighbor model [PDNN].

It has been shown that the hyperbolic Langmuir isotherm describes the shape of GeneChip signal response to concentration [6]. Most of the developed methods rely, however, on a linear fit, while in reality probe intensities often span outside the linear regime region of the isotherm within the working concentration range for commercial arrays. The availability of calibration experiments allows for the development and validation of statistical signal models for microarray data. We use three publicly available spike-in datasets (U95, U133 and wholly
defined control spike-in datasets) to test and validate a novel multi-chip non-linear fitting algorithm [7], [8].

The removal of non-specific background is a key issue in microarray data analysis. Several existing methods use MM probes to estimate the background signal. An underlying assumption of these approaches is that the majority of MM signal is non-specific [1], [4]. However, mismatch signal contains mostly specific signal in addition to a random non-specific component that is different from the non-specific component of PM. Hence direct subtraction of MM is unlikely to be useful. We therefore exclude MM information from our model.

Materials and Methods

Langmuir isotherm

In this paper we propose to use a physical model of the hybridization process based on Langmuir adsorption theory. Previous work in this field was carried out by Halperin et al. [6], Peterson et al. [9], Vainrub and Pettitt [10], and Dai et al. [11]. Attempts to match Langmuir adsorption isotherms to the Affymetrix spike-in experiment data include those of Held et al. [12], Hekstra et al. [13] and Zhang et al. [5]. Langmuir adsorption theory is based on the assumption that there are two competing processes driving hybridization: adsorption, i.e. the binding of target molecules to immobilized probes to form duplexes, and desorption, i.e. the reverse process of duplexes dissociating into separate probe and target molecules. The form of the equation results in a non-linear concentration-intensity response, i.e. decreased probe affinity with increase of concentration. In the absence of non-specific binding, the equation would be of the form:

\[ PM(c) = I \cdot \frac{kc}{1 + kc}, \]

where \( c \) is a concentration, \( I \) is a saturation intensity and \( k \) is an equilibrium constant.
It has been shown that the saturation intensity and equilibrium constants depend on numerous factors and differ greatly from probe to probe. Several attempts have been made to predict these constants from probe sequence content [5], [13]. For example, Hekstra et al. used a Langmuir adsorption model with an additive background from non-specific binding [13]. In this approach parameters for the Langmuir model were first fit into the model using known concentrations values provided in a spike-in experiment. Next, the resulting parameters were fit into linear combinations of the numbers of each nucleotide for each probe-target pair and estimates of concentrations were obtained for each probe by inverting the Langmuir equation. The averages of predicted concentrations across each probeset were then reported as expression measures. Burden et al. [14] demonstrated that such an approach returns poor estimates of concentration; up to 60% of predicted values had to be discarded according to the suggested truncation scheme.

**Statistical model**

Although duplex formation in solution has been extensively studied using a nearest-neighbour model, it has been difficult to apply those results to microarrays [5]. Binding interactions on the microarray surface are complicated by many factors including steric hindrance on the surface, probe-probe interactions and RNA secondary structure formation. Hence, we avoid predicting physical parameters of the model. We use the notion of background as a probe-specific part of the signal that is introduced by all the genes in the sample pool other than the target gene. We propose the following statistical model:

\[
\log(PM_{pjl}) = \log \left( I_p \frac{k_p c_j}{1 + k_p c_j} + bg_p \right) + \varepsilon_{pjl},
\]

where \( p \) is a probe index \( (p = 1, \ldots, P) \), \( j \) is a condition (concentration) index \( (j = 1, \ldots, J) \), \( c \) is a concentration, \( l \) is replicate \( (l = 1, \ldots, L) \), \( I \) is a saturation intensity, \( k \) is an equilibrium constant and \( bg \) is the background component of the signal. The \( \varepsilon_{pjl} \) are independent error terms, with mean 0 and constant variance. Thus, each probe is parameterized by three
parameters and each experimental condition is characterized by one concentration.

Algorithm

We used an iterative non-linear procedure using a Newton-type optimization method to fit the model to the data. Prior to fitting, the data were normalized. For Affymetrix spike-in datasets we used a probe level quantile normalization [15], implemented in Bioconductor, an open source project for the analysis of genomic data [16]. For wholly defined control spike-in dataset, we adopted constantsubset invariant set normalization, proposed in the original paper [8].

The first step of this method is to obtain estimates for initial concentrations by fitting log intensities to a linear model using a robust additive model that employs Tukey’s median polish procedure:

\[ \log(PM_{pji}) = \text{concentration}_{pj} + \text{probe affinity}_{p} + \epsilon_{pji} \]

The \texttt{medpolish()} routine is available in R, a widely used open source language and environment for statistical computing and graphics.

Next, probe parameters \( I, k, bg \) and concentration terms \( c \) are iteratively refined according to the full non-linear model through \texttt{nlm()}, a non-linear least square optimization routine in the R statistical system. The search is carried out iteratively by minimizing the sum of square of the residual matrix obtained in the previous fit. First, the estimators of probe parameters are updated using current concentration values. Then concentration estimates are optimized based on these new probe parameters. The iterative scheme continues until convergence is obtained, that is, until fitted parameters from a current step are sufficiently close to the parameters of the previous step. Convergence was observed after 5-10 iterations.
Model selection

In the presence of a specific target gene, the model can be rewritten as

$$\log(PM_{pjl}) = \log(SP_{pj} + bg_p) + \varepsilon_{pjl}$$

where $p$ is a probe index, $j$ is a condition (concentration) index, $SP_{pj}$ is the specific component of the signal and $bg_p$ is the non-specific fraction of the signal. If gene expression does not change from condition to condition, then the signal model should be simplified to

$$\log(PM_{pjl}) = \log(SP_p + bg_p) + \varepsilon_{pjl}$$

Hence, a simplified model with fewer parameters is more appropriate when the target gene is not differentially expressed. In order to address this question, for each gene we perform a goodness-of-fit test and select the appropriate model based on formal model selection criteria using the statistic:

$$\frac{(RSS_1 - RSS_0)/(3P + J - P)}{RSS_0/(PJL - 3P - J)}$$

where $RSS_1$ denotes the residual sum of squares from fitting the reduced model and $RSS_0$ denotes the residual sum of squares from fitting the full model. The nested model is constructed so that the simpler one-concentration model can be obtained from the multi-concentration signal model as described above. Thus choosing among models reduces to determining the appropriateness of the additional concentration parameters. The significance of F-values is assessed by permutation analysis; randomization of the probe data for the genes that were not spiked in the test datasets was performed to establish cutoff level for p-values.

It was observed that even genes with no concentration changes exhibit a probe fluorescence intensity variation range of $\sim 10\%$. Hence, in practice, a reduced model is fitted first, as it was noticed that full model fit is at times unnecessary since fitting a reduced model and evaluating RSS leads to clear evidence of no change in concentration across all conditions. This step allows us to reduce computational time significantly without affecting the accuracy.
of concentration prediction.

The test data sets

*Affymetrix Human Genome U95 Dataset*

In the course of developing and validating the Affymetrix Microarray Suite (MAS) 5.0 algorithm, Affymetrix produced data from a set of 59 arrays (HGU95) organized in a Latin square design. This dataset 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. The concentrations of the 14 gene groups in the first experiment are 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024 pM. Each experiment contains three replicates. The Affymetrix Human Genome U95 spike-in dataset is available at [http://www.affymetrix.com/support/technical/sample_data/datasets.affx](http://www.affymetrix.com/support/technical/sample_data/datasets.affx).

*Affymetrix Human Genome U133 Dataset*

This data set consists of 3 technical replicates of 14 separate hybridizations of 42 spiked transcripts in a complex human background at concentrations ranging from 0.125 pM to 512 pM. Thirty of the spikes are isolated from a human cell line, four spikes are bacterial controls, and eight spikes are artificially engineered sequences believed to be unique in the human genome. The Affymetrix Human Genome U133 spike-in dataset is available at [http://www.affymetrix.com/support/technical/sample_data/datasets.affx](http://www.affymetrix.com/support/technical/sample_data/datasets.affx).

*Wholly Defined Control Spike-In Dataset*

Choe et al. [8] generated a new control dataset for the purpose of evaluating methods for identifying differentially expressed genes between two sets of replicated hybridizations to Affymetrix GeneChips. This dataset contains 1309 individual cRNAs that differ by known
relative concentrations between the spike-in and control samples. This large number of defined RNAs provides accurate estimates of false-negative and false-positive rates at each fold-change level. The dataset includes low fold changes, beginning at only a 1.2-fold concentration difference, up to 4 fold. The dataset uses a defined background sample of 2,551 RNA species present at identical concentrations in both sets of microarrays, rather than a biological RNA sample of unknown composition.

**Results**

**Langmuir adsorption model**

As noted earlier, the response curve is non-linear and it can be better characterized by a Langmuir isotherm \[13\]. The resulting fluorescence signal versus target concentration data are shown in Fig. 1, providing a demonstration that the Langmuir model captures the physical chemistry of GeneChip hybridization.

**Extraction of gene expression and estimates of differential expression**

We fit our nonlinear model to the U133 spike-in data without prior knowledge of the concentrations used in this experiment. Fig. 2 shows the reconstructed concentration range plotted against the true mRNA concentration.

The sensitivity of the algorithm can be assessed by examining local slopes, i.e. the observed log fold-change for genes with true fold-change of 2. Since concentration groups in both datasets are arranged in a Latin Square design and thus differ by a multiple of 2, the ideal local slope would be 1. Fig. 3 shows the predicted concentration increments for the full concentration ranges.
Comparison with existing algorithms

We compared the performance of our non-linear multi-chip fitting procedure to other popular algorithms that compute gene expression measures using the two Affymetrix spike-in datasets mentioned earlier. Algorithms selected for comparison included RMA [3], RMA’s successor GCRMA [4], Affymetrix MAS 5.0 [1] and the latest Affymetrix algorithm, PLIER (Probe Logarithmic Intensity Error) [1]. The results for MAS and PLIER methods were obtained using BioConductor implementation [16]. Surprisingly, the performance of these algorithms varies greatly depending on the dataset used, e.g. GCRMA shows improved performance on the HG-U95 spike-in control dataset while performing significantly worse on the HG-U133 dataset [Fig. 4 and Fig. 5]. In contrast, MAS 5.0 shows improved reconstructed concentration curves on the U133 spike-in dataset and performs poorly on the U95 spike-in dataset. The non-linear multi-chip fit performs consistently on both U133 and U95 since it does not rely on pre-fitted parameters optimized for a particular dataset.

We used the Affycomp benchmark [17] to evaluate, compare, and display the performance of expression level estimators for proposed global multi-chip non-linear fit. Using controlled spike-in experiments and dilution series, Affycomp systematically assesses the performance of the methods at different biologically relevant spike-in concentrations. The results of the assessment have been submitted to the Affycomp website for comparison other probe-level analysis algorithms. Due to the structure of NLFIT results, several assessments scores should not be considered for evaluation and comparison. For example, the proposed global non-linear multi-chip algorithm accounts for experiment design, and fits replicated data into a single measure. Thus, assessments that account for between-replication variation in expression level estimators should not be considered. For the spike-in experiment using the HG-U133A and HG-U95 chip, our proposed global non-linear multi-chip algorithms achieved first scores in three relevant assessments (from a total of 13 scores). The complete assessment report is
provided in Table 1.

We used a wholly defined control dataset to assess the performance of our non-linear multi-chip fitting procedure. This dataset contains a very large set of 1309 individual cRNAs that differ by known relative concentrations between the spike-in and control samples as well as defined background sample of 2,551 RNA species present at identical concentrations throughout the experiment, rather than a biological RNA sample of unknown composition. To perform a direct comparison of the popular methods to the proposed non-linear routine, we reproduced the results of RMA, GCRMA, Affymetrix MAS 5.0 and PLIER. We compared the observed fold changes with known fold changes and the resulting plot is shown in Fig. 6 and Fig. 7.

These figures demonstrate the advantage of NLFIT over others popular procedures i.e. RMA, GCRMA, PLIER, MAS5. The effect is more pronounced due to increased number of spiked-in genes available in the wholly defined control dataset. The results, reproduced for RMA, GCRMA, Affymetrix MAS 5.0 and PLIER agree with the ones obtained in the original paper [8]. It can be observed that these algorithms consistently underestimate fold changes by a factor of two. In contrast, NLFIT reconstructs fold changes without significant distortion.

**Discussion**

Examining fluorescence intensity versus nominal concentration plots shows that the assumption of linearity between measured intensity and concentration is inaccurate in the case of Affymetrix GeneChips and results in curves saturating according to a hyperbolic Langmuir isotherm. By employing a physical model that accounts for chemical saturation we improve the accuracy in differential gene expression estimates, especially in high concentrations ranges.

Additionally, by implementing background and gene expressions estimation within the same
fitting procedure, we are able to provide estimates of differential expression with a significant reduction in bias without a concomitant decrease in the signal-to-noise ratio. This allows us to estimate background on each probe in the context of all experimental conditions. Other algorithms [RMA, GCRMA] subtract background estimates based on theoretical predictions prior to fitting expression measures.

Comparisons of the various algorithms are limited by the small number of available control datasets. The two Affymetrix spike-in data sets used in this study are among the few available for benchmarking Affymetrix GeneChip expression measures. Certain algorithms were trained on these datasets and hence a direct comparison of such results must be excluded.

Examination of Figures 4 and 5 reveals that both PLIER and GCRMA are fairly insensitive to low concentrations of less than 1 pm, whereas NL-FIT and MAS 5.0 are able to detect concentration changes at a level of 0.1 pm. Methods that use MM as a background predictor, i.e. MAS 5.0 and GCRMA, tend to overestimate concentration changes (Figure 5). It is also apparent that methods that rely on linear models (MAS 5.0, PLIER, and GCRMA) become less sensitive in high target concentration range, while NLFIT correctly estimates concentration changes across the entire range of values.

Test of the algorithm’s performance on the wholly defined spike in dataset [8] demonstrates the reduced sensitivity of the selected popular algorithms and the advantage of NLFIT in accurate fold change reconstruction. While the cause of this compressive fold change effect is not clear, we attribute this effect to the high inhomogeneity of the genomic background that results in cross hybridization, unaccounted for by these methods.

References

[1] Affymetrix Statistical Algorithms Description Document. (2001).
[2] Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98, 31–6.

[3] Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004) affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3), 307–15.

[4] Wu, Z. and Irizarry, R. A. (2004) Preprocessing of oligonucleotide array data. *Nat Biotechnol*, 22, 656–8; author.

[5] Zhang, L., Miles, M. F., and Aldape, K. D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol*, 21, 818–21.

[6] Halperin, A., Buhot, A., and Zhulina, E. B. (2004) Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophys J*, 86, 718–30.

[7] Affymetrix (2001) Latin Square data for expression algorithm assessment.

[8] Choe, S., Boutros, M., Michelson, A., Church, G., and Halfon, M. (2005) Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol*, 6(2), R16.

[9] Peterson, A. W., Heaton, R. J., and Georgiadis, R. M. (2001) The effect of surface probe density on DNA hybridization. *Nucleic Acids Res*, 29, 5163–8.

[10] Vainrub, A. and Pettitt, B. M. (2002) Coulomb blockage of hybridization in two-dimensional DNA arrays. *Phys Rev E Stat Nonlin Soft Matter Phys*, 66, 041905.

[11] Dai, H., Meyer, M., Stepaniants, S., Ziman, M., and Stoughton, R. (2002) Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays. *Nucleic Acids Res*, 30, 86.

[12] Held, G. A., Grinstein, G., and Tu, Y. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc Natl Acad Sci U S A*, 100, 7575–80.
[13] Hekstra, D., Taussig, A. R., Magnasco, M., and Naef, F. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res*, 31, 1962–8.

[14] Burden, C., Pittelkow, Y. E., and Wilson, S. (2004) Statistical analysis of adsorption models for oligonucleotide microarrays. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

[15] Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19, 185–93.

[16] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5, 80.

[17] Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20, 323–31.
Fig. 1. Fluorescence intensity vs nominal concentration for three genes in the Affymetrix U133 data set. Points represent experimental conditions and solid lines represent Langmuir isotherm fit.

Fig. 2. Gene expression measures obtained using non-linear multi-chip fit for the 42 spike-in genes in the 14 experiments in AFFY HG-U133. Each curve represents extracted expression measures for a gene. Red line denotes the median, dashed line represents the identity line.
Fig. 3. Extracted local slopes vs. nominal concentration. The average over 42 probesets is shown in red. The identity line is also shown by the black dashed line.

Fig. 4. Reconstructed concentration range vs. nominal concentrations of U133 and U95 spike-in dataset for MAS 5.0 [green], PLIER [light blue], GCRMA [dark blue], RMA [fuchsia] and NLFIT [red], shifted to match 1 pM concentration. Averages taken over 42 probesets are shown. Reconstructed curves were shifted vertically to meet zero at 1pM spike-in. The identity line is dashed.
Fig. 5. Reconstructed concentration increments vs. nominal concentrations of U133 and U95 spike-in dataset for MAS 5.0 [green], PLIER [light blue], GCRMA [dark blue], RMA [fuchsia] and NLFIT [red]. Averages taken over 42 probesets are shown. An ideal, $y = 1$, increment line is shown in dashed black.

Fig. 6. Reconstructed log$_2$ fold change vs. nominal log$_2$ fold changes of wholly defined control spike-in dataset for MAS 5.0 [light blue closed circle], PLIER [red +], GCRMA [green open triangle], RMA [blue open circle] and NLFIT [black +]. The $x = y$ line is shown in solid black.
Fig. 7. Reconstructed log$_2$ fold change vs. nominal log$_2$ fold changes of wholly defined control spike-in dataset for MAS 5.0 [light blue closed circle], PLIER [red +], GCRMA [green open triangle], RMA [blue open circle] and NLFIT [black +]. Medium, overeach fold change are shown. The $x = y$ line is shown in solid black.
| Statistics                                                                                                      | HG_U95 | HG_U133A |
|-----------------------------------------------------------------------------------------------------------------|--------|----------|
| 1. [26] Median SD - median SD across replicates                                                                  | 0      | 0.02     |
| 2. [18] null log-fc IQR - Inter-quartile range of the log-fold-changes from genes that should not change.       | 0      | 0.01     |
| 3. [20] null log-fc 99.9% - 99.9% percentile of the log-fold-changes if from the genes that should not change. | 0      | 0.15     |
| 4. [2] Signal detect R2 - R-squared obtained from regressing expression values on nominal concentrations in the spike-in data | 0.89   | 0.85     |
| 5. [1] Signal detect slope - slope obtained from regressing expression values on nominal concentrations in the spike-in data | 0.65   | 0.7      |
| 6. [29] low.slope - Slope from regression of observed log concentration versus nominal log concentration for genes with low intensities. | 0.25   | 0.22     |
| 7. [30] med.slope - As above but for genes with medium intensities.                                            | 0.68   | 0.69     |
| 8. [31] high.slope - As above but for genes with high intensities.                                              | 0.68   | 1.07     |
| 9. [10] Obs-intended-fc slope - slope obtained from regressing observed log-fold-changes against nominal log-fold-changes | 0.64   | 0.7      |
| 10. [11] Obs-(low)int-fc slope - slope obtained from regressing observed log-fold-changes against nominal log-fold-changes for genes with nominal concentrations less than or equal to 2 | 0.34   | 0.29     |
| 11. [21] low AUC - Area under the ROC curve (up to 100 false positives) for genes with low intensity standardized so that optimum is 1. | 0.74   | 0.76     |
| 12. [22] med AUC - As above but for genes with medium intensities.                                               | 0.92   | 0.95     |
| 13. [23] high AUC - As above but for genes with high intensities.                                                | 0.95   | 0.97     |
| 14. [24] weighted avg AUC - A weighted average of the previous 3 ROC curves with weights related to amount of data in each class (low,medium,high). | 0.79   | 0.81     |

**Table 1**

The results from the affycomp assessment of HG-U133A and HG-U95 spike in data.
