Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 35

Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays

Conrad J. Burden* Yvonne E. Pittelkow†

Susan R. Wilson‡

*Australian National University, conrad.burden@anu.edu.au

†Australian National University, yvonne.pittelkow@anu.edu.au

‡Australian National University, sue.wilson@anu.edu.au

Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays*

Conrad J. Burden, Yvonne E. Pittelkow, and Susan R. Wilson

Abstract

Recent analyses have shown that the relationship between intensity measurements from high density oligonucleotide microarrays and known concentration is non linear. Thus many measurements of so-called gene expression are neither measures of transcript nor mRNA concentration as might be expected.

Intensity as measured in such microarrays is a measurement of fluorescent dye attached to probe-target duplexes formed during hybridization of a sample to the probes on the microarray. We develop several dynamic adsorption models relating fluorescent dye intensity to target RNA concentration, the simplest of which is the equilibrium Langmuir isotherm, or hyperbolic response function. Using data from the Affymerix HG-U95A Latin Square experiment, we evaluate various physical models, including equilibrium and non-equilibrium models, by applying maximum likelihood methods. We show that for these data, equilibrium Langmuir isotherms with probe dependent parameters are appropriate. We describe how probe sequence information may then be used to estimate the parameters of the Langmuir isotherm in order to provide an improved measure of absolute target concentration.

KEYWORDS: Gene expression, microarrays, Langmuir adsorption

---

*We thank Peter Hall for helpful discussion on the validity of the bootstrap approach to obtaining confidence intervals for the median for the situation described here. This research was partly supported by ARC Grant DP 0343727

1. Introduction

Oligonucleotide microarrays are a technology which enable the simultaneous testing for the presence and quantification of large numbers of genes in prepared target RNA samples. Affymetrix GeneChip arrays, the focus of this paper, consist of a substrate onto which short single strand DNA oligonucleotide probes have been synthesized using a photolithographic process. A chip surface is divided into some hundreds of thousands of regions typically tens of microns in size, the probes within each region being synthesized to a specific nucleotide sequence. Throughout this paper we use the word 'probe' to refer either to a single strand of synthesised DNA, or a region of identically synthesised strands. Dependent on the organism, each gene is represented by a set of between 11 to 16 probe pairs termed a probeset. One element of each pair is synthesised as a 'perfect match' (PM) sequence of length 25 bases, and the other a 'mismatch' (MM) sequence in which the middle (13th) base has been replaced by its complement. Each PM probe is chosen to be a non-overlapping subsequence of the full gene sequence, chosen for its predicted hybridisation properties and specificity to the potential target gene. The target RNA sample is hybridized onto the chip to form probe-target duplexes, and the chip scanned to obtain fluorescence intensity readings from dyes incorporated during the laboratory procedures. For further details, see Nguyen et al. (2002).

In principle, with suitable calibration, intensity readings are intended to be in some sense a measure of concentration of matching target RNA in the sample. A number of expression indices exist which seek to extract a measure of 'gene expression' (see for example Affymetrix Inc. (2002) or Irizarry et al. (2003)). Such indices are generally calculated by subtracting an estimate of background, often estimated from MM readings, and summarizing over readings from probes within a probeset. The approach is purely empirical: little or no attempt is made to understand the physical processes driving hybridization, and consequently neither the effects of saturation at high target concentration nor the effects of probe sequence specificity are accounted for. Also the true meaning of the MM intensity reading is not properly understood. Furthermore, expression indices are given in arbitrary units and not units of concentration. While this may have application to comparisons between different treatments of a given transcript (gene or EST) within the same experiment, no comparison can be made between the concentrations of expressed RNA from different transcripts, or the same transcript in different experiments. As we shall see below, fluorescence intensity measurements are very strongly sequence dependent, and consequently probe dependent.

Recently studies have begun to address these issues by appealing to models based on well established principles of physical chemistry (Hekstra et al., 2003; Held et al., 2003). Such models, known generically as chemical adsorption models, offer the possibility of predicting absolute target concentrations as opposed to relative expression measures, and hence have the potential to enable comparisons between expression levels of different genes. It is also hoped that an accurate understanding of the physical processes driving hybridization will lead to improvements in microarray design. While this line of research shows great promise, we are unaware of any thorough and rigorous comparative statistical analysis assessing the various existing versions of adsorption models for Affymetrix microarrays. The purpose of this paper is to carry out such an analysis, with the eventual aim of establishing a practical relationship between measured fluorescence intensities and the underlying concentration of mRNA in a biological sample.

In Section 2 we develop several dynamic adsorption models, including a simple equilibrium Langmuir isotherm, or hyperbolic response function, a non-equilibrium Langmuir model, and the Sips isotherm. In Section 3 we apply our approach to the publicly available data from the Affymetrix Human HG-U95A Latin Square spike-in experiment (available off `http://www.affymetrix.com/support/technical/sample_data`). Our comparative statistical analysis is able to resolve the differing conclusions of Hekstra et al. (2003) and Held et al. (2003), for example, both of whom have analysed this data set using approaches based on the equilibrium isotherm model. Our main findings are that measured fluorescence intensity readings for both PM and MM probes are described by a three parameter hyperbolic response function over a range from $< 1\,\mathrm{pM}$ to $> 1000\,\mathrm{pM}$, and that the response function parameters are strongly probe sequence dependent. Furthermore, fold changes in target concentration are clearly not linearly related to fold changes in fluorescence intensity readings, as is generally assumed.

The practical efficacy of these findings lies in the fact that the three parameters defining the response function can only depend on the probe sequence, and are therefore universal to all experimental applications of that particular sequence. The remaining challenge, to establish an algorithm for extracting isotherm parameters from any given probe sequence, is not dealt with in this paper, but is the subject of ongoing work. Such an algorithm, when established, can then be impemented as part of an expression measure of absolute concentration. To gauge the feasibility of such an approach we analyse in Section 4 a simple linear model for predicting isotherm parameters suggested by Hekstra et al. (2003), and thereby develop a method that reduces potential bias inherent in, and provides confidence intervals for, any estimate of

mRNA concentration based on inverting adsorption isotherms. Finally we plot for comparison MAS5 and RMA expression measures for the spike-in data to illustrate that, even using this simple linear estimation of isotherm parameters, a Langmuir isotherm model outperforms existing expression measures in recovering target concentration fold changes over the full range of spiked in concentrations.

2. Dynamic adsorption models

We consider a number of models based on a process of competing adsorption and desorption of target RNA to form probe-target duplexes at the chip surface (Forman et al., 1998). Let $\theta(t)$ be the fraction of sites within a probe region occupied by probe-target duplexes at time t after the commencement of hybridization, and k_{f} and k_{b} be the forward adsorption and backward desorption rate constants respectively. The forward adsorption reaction is assumed to occur at a rate $k_{\mathrm{f}}x(1-\theta(t))$, proportional to target concentration x and fraction $(1-\theta(t))$ of unoccupied probe sites. The backward desorption reaction is assumed to occur at a rate $k_{\mathrm{b}}\theta(t)$, proportional to the fraction of occupied probe sites. The fraction of probe sites occupied by probe-target duplexes is then given by the differential equation

$$\frac{d\theta(t)}{dt} = k_{\mathrm{f}}x(1-\theta(t)) - k_{\mathrm{b}}\theta(t), \tag{1}$$

with initial condition $\theta(0) = 0$. This solves to give

$$\theta(t) = \frac{x}{x+K}\left[1 - e^{-(x+K)k_{\mathrm{f}}t}\right], \tag{2}$$

where $K = k_{\mathrm{b}}/k_{\mathrm{f}}$. Setting y to be the measured fluorescence intensity and assuming the intensity above the background value y_0 at zero concentration to be proportional to θ, we arrive at the relationship

$$y = y_0 + b\frac{x}{x+K}\left[1 - e^{-(x+K)k_{\mathrm{f}}t}\right]. \tag{3}$$

In Figure 1 we graph the dimensionless quantities $(y-y_0)/b$ against x/K for various values of the dimensionless inverse time $\tau = (k_{\mathrm{b}}t)^{-1}$. For times much shorter than the inverse backward rate constant, $t << {k_{\mathrm{b}}}^{-1}$ or $\tau >> 1$, we find

$$\frac{y-y_0}{b} = \frac{x}{K\tau} + O\left(\frac{1}{\tau^2}\right). \tag{4}$$

This linear response is evident in Figure 1.

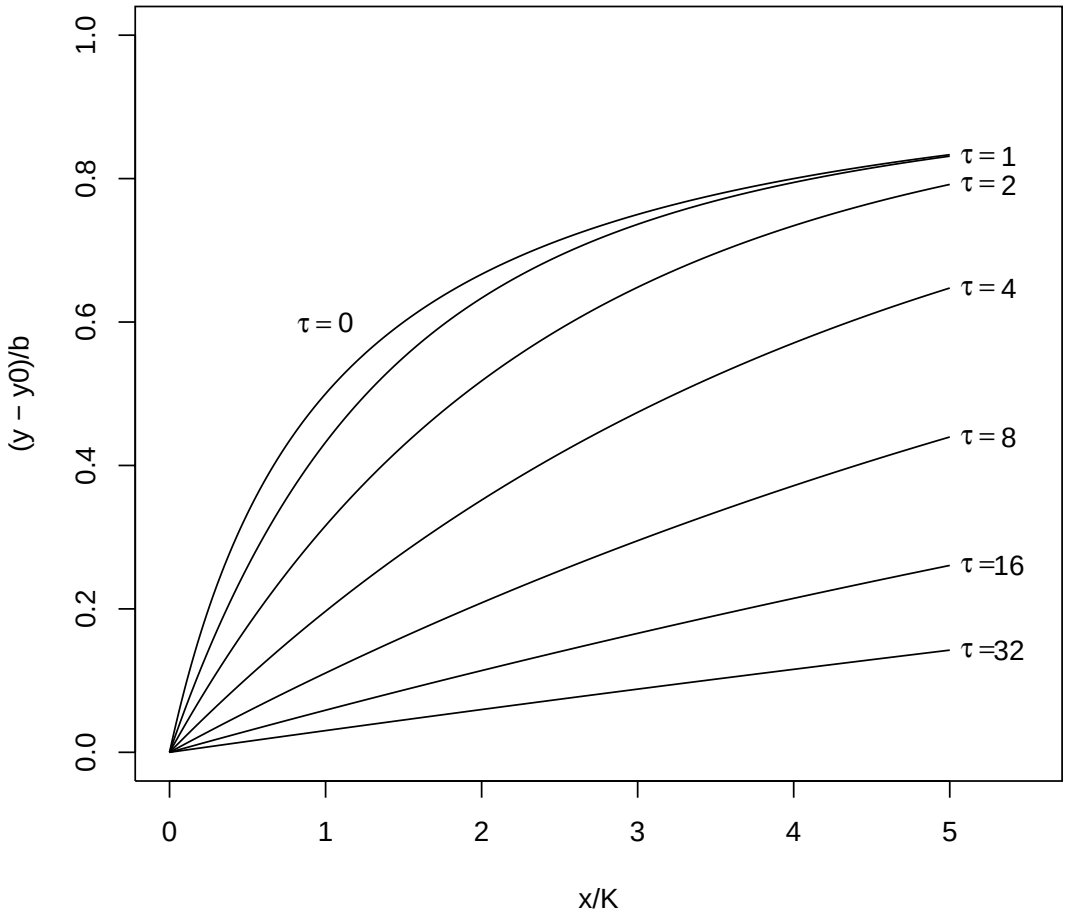

Figure 1. Plot of the dimensionless fluorescence intensity $(y - y_0)/b$ against dimensionless concentration x/K for various values of the dimensionless inverse time $\tau = (k_{\mathrm{b}}t)^{-1}$. In the equilibrium limit $\tau \to 0$ the Langmuir isotherm is recovered.

In the equilibrium limit $t \to \infty$ we recover the well known Langmuir isotherm, or hyperbolic response function,

$$y = y_0 + b\frac{x}{x+K}. \tag{5}$$

The quantity $y_0 + b$ is the saturation intensity in the limit of high target concentration. In this limit, the simple Langmuir model predicts $\theta = 1$, i.e. all probe sites are occupied by probe-target duplexes. In practice however, probe efficiency may be affected by a number of factors, particularly on high density chips (Peterson et al., 2001). The background y_0 is generally considered to consist of a physical component from sources such as reflection and photomultiplier dark current, and a biological component from non-specific hybridization (Hekstra et al., 2003). Experimentally one typically finds that $y_0 << b$. The parameter K is the target concentration at half-saturation and is expected on thermodynamic grounds (Atkins, 1994) to be proportional to $\exp(-\Delta G/k_{\mathrm{B}}T)$, where ΔG is the probe-target binding free energy, k_{B} is Boltzmann's constant and T the absolute temperature. Hekstra et al. (2003) have argued that the parameters y_0, b and K may be augmented in the presence of non-specific hybridization while leaving the form of Eq. (5) unchanged.

We shall also consider a variant of the Langmuir isotherm, known as the Sips isotherm (Sips, 1948)

$$y = y_0 + b\frac{x^\alpha}{x^\alpha + K^\alpha}, \tag{6}$$

which arises by assuming the binding free energy to have an approximately Gaussian distribution about its mean value. The parameter α takes values between 0 and 1, the variance of the distribution decreasing as α increases. In the limit $\alpha \to 1$ the distribution becomes a delta function and the Langmuir isotherm is recovered. Experiments at high target concentrations up to 1μM indicate that the Sips model may be appropriate in some circumstances (Peterson et al., 2002).

3. Models and analyses for the Affymetrix data

For the Affymetrix Human HG-U95A Latin Square experiment genes were spiked in at cyclic permutations of the set of known concentrations, together with a background of RNA extracted from human pancreas. The data consists of fluorescence intensity values from a set of 14 probesets corresponding to 14 separate genes, each containing n_{probe} probe pairs where $n_{\mathrm{probe}} = 16$. For each probeset a set of intensity values are obtained for the 14 spiked-in concentrations (0, 0.25, 0.5, 1, 2, 4, ..., 1024) pM. The experiment was replicated

three times using microarray chips from different wafers. We decided to (i) omit from our analysis two genes which suffered from defective probes (407_at and 36889_at), leaving only the remaining 12, and an extreme probe outlier in gene 37777_at, (ii) performed no normalization on the data before the analysis described below, and (iii) consider only the perfect match (PM) probes, as did Held et al. (2003).

As described in the appendix, the stochastic noise in this data set has the property of having an approximately constant coefficient of variation across a broad range of fluorescence intensity values. A gamma distribution with constant shape parameter conveniently models this property and takes as its argument only physically meaningful positive real values. Furthermore it is consistent with the fact that fluorescence intensity, being a photon count or, at a more fundamental level, a count of probe-target duplexes at the chip surface, is an extensive variable in the sense that it is a physically additive quantity (Cox and Snell, 1981). As pointed out by McCullagh and Nelder (1989), page 286, a gamma distribution is preferable to, say, a lognormal distribution in such situations since the mean of an extensive property maintains the property of being extensive, whereas the mean of its log, or of any nonlinear function for that matter, does not.

Direct physical justification for the gamma assumption is found in the work of Dennis and Patil (1984) who demonstrate that the gamma distribution serves as a general model of a population fluctuating about a steady state equilibrium. The class of dynamic models considered by Dennis and Patil includes the adsorption-desorption model of Eq. (1). In the appendix we argue that a physically reasonable model leading to an exact gamma distribution can be constructed to describe stochastic nature of the adsorption process. We also note work by Chen et al. (1997) who argue that intensity measurements are linked by having a constant coefficient of variation across a broad range of target concentrations and of Newton et al. (2001) on spotted microarrays who also consider fluorescence data to be drawn from a gamma distribution.

For the remainder of this paper we therefore assume that the stochastic component of the fluorescence intensity y is drawn from a gamma distribution with mean μ and (constant) shape parameter ν having density

$$\frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu}\exp\left(-\frac{\nu y}{\mu}\right)d(\ln y); \tag{7}$$

the unscaled deviance (McCullagh and Nelder, 1989) given by

$$D(\mathbf{y};\mu) = -2\sum_{i}\left[\ln(y_i/\mu_i) - (y_i-\mu_i)/\mu_i\right]. \tag{8}$$

Table 1

Six generalized linear models fitted to the Affymetrix Latin Square experiment. The indices g, p and w indicate dependence of fitting parameters on gene, probe (within each gene's probeset) and wafer respectively.

	Model	Parameters
A	$\mu = y_0 + bx/(x+K)$	y_{0pg}, b_g, K_{pg}
B	$\mu = y_0 + bx/(x+K)$	y_{0pg}, b_{pg}, K_{pg}
C	$\mu = y_0 + bx/(x+K)\left[1 - e^{-(1+x/K)/\tau}\right]$	y_{0pg}, b_g, K_{pg}, τ_{pg}
D	$\mu = y_0 + bx/(x+K)\left[1 - e^{-(1+x/K)/\tau}\right]$	y_{0pg}, b_{pg}, K_{pg}, τ_{pg}
E	$\mu = \lambda\left[y_0 + bx/(x+K)\right]$	λ_w, y_{0pg}, b_{pg}, K_{pg}
F	$\mu = y_0 + bx^\alpha/(x^\alpha + K^\alpha)$	y_{0pg}, b_{pg}, K_{pg}, α_{pg}

Our aim is to select a model that is (i) parsimonious (i.e. without unnecessary parameters), and (ii) accurate over the full set of data. To compare models 1 and 2, with r_1 and r_2 residual degrees of freedom and deviances D_1 and D_2 respectively, where model 1 is nested within model 2 (i.e. $r_1 > r_2 >> 1$), we use

$$\Delta D_{\text{scaled}} = (D_1 - D_2)\frac{r_2}{D_2}. \tag{9}$$

If the extra parameters present in model 2 are not statistically significant the scaled change in deviance is distributed approximately as a chi-squared distribution with $r_2 - r_1$ degrees of freedom.

3.1 *Models*

The six models we have fitted to the data are summarized in Table 1. Models **A** and **B** are the Langmuir isotherms of Eq. (5). They differ from each other in that the asymptotic saturation intensity above background at high concentration b is taken to be common to all probes within a probeset in model **A** and allowed to vary across probes in model **B**. Model **B** is that used by Hekstra et al. (2003) and Webster et al. (2003). The parameters y_0 and K can reasonably be expected to vary across probes within a probeset given expected variations in cross hybridization levels and free binding energies respectively.

Models **C** and **D**, corresponding to the non-equilibrium model of Eq. (3), are extended versions of models **A** and **B** including a new set of parameters τ representing a dimensionless inverse time. We have included these models to investigate the hypothesis of a common asymptotic intensity b across probes given the possibility that the hybridization had not reached equilibrium. We

feel that this possibility should be pursued given that experimental studies such as that quoted by Held et al. (2003) as evidence of equilibrium, namely Forman et al. (1998), or the more recent work of Peterson et al. (2001) and Peterson et al. (2002), are carried out at significantly higher RNA concentrations (> 1 nM) than the pM concentrations of the Affymetrix data set.

While our main concern is a comparison of the four models **A** to **D**, we include two extra models, **E** and **F**, for completeness. We shall see below that, of the four models described so far, model **B** was selected. For this reason, the remaining two models are based on model **B**. Model **E** is an extension of model **B** with a wafer dependent factor λ_w included to account for the possibility of systematic variation across the three replicate experiments. The λs are scaled so that $\frac{1}{3}\sum_w \lambda_w = 1$, entailing that models **B** and **E** differ by only 2 degrees of freedom. Finally, model **F** is an extension of model **B** to the Sips isotherm Eq. (6). The parameters α_{gp} are taken to be gene and probe dependent. The relationship between models **A** to **E** is shown in Figure 2.

3.2 *Statistical analysis*

Table 2 shows calculated changes in scaled deviance for four pairwise comparisons between models **A**, **B**, **C** and **D**, which address the question of whether the hypothesis of an asymptotic saturation intensity b above background common to all probes can be supported, given that hybridization may or may not have reached equilibrium. A small number of probes have been omitted from this analysis, either because fits gave unphysical negative values to the parameters b and K, or, in the case of probes omitted from genes 38734_at and 1091_at, to remove outlying values of intensity at high concentration so as to enable fits of the common asymptote models **A** and **C**.

Minimization of the deviance over the non-linear parameters K and α was carried out using an algorithm described in McCullagh and Nelder (1989), while minimization over the parameters λ and τ was carried out using a steepest descent algorithm. For some probes we found the minimum over τ occurred in the limit $\tau \to \infty$, $K \to 0$ with $K\tau$ finite, consistent with the linear response limit Eq. (4). We found that limiting the steepest descent algorithm to a maximum value of $\tau_{\max} = 60$ to handle these cases made no noticeable difference to the calculated deviance.

Examining values of ΔD_{scaled} for comparisons $\mathbf{A} \to \mathbf{B}$ and $\mathbf{C} \to \mathbf{D}$, it is clear that the extra parameters introduced to allow for a probe dependent asymptote are significant. That is, the hypothesis of a common asymptotic saturation intensity cannot be supported, irrespective of whether the equilibrium or non-equilibrium model is assumed. The comparison $\mathbf{B} \to \mathbf{D}$ then shows that the extra parameters introduced by assuming a non-equilibrium model

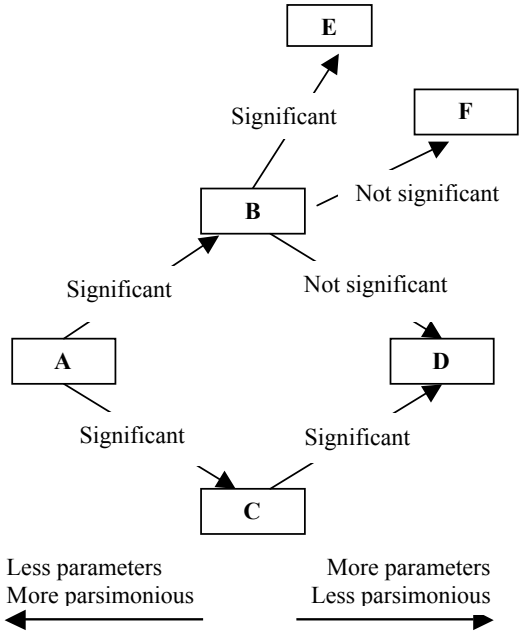

Figure 2. The relationship between the six models listed in Table 1. An arrow running from model 1 to model 2 indicates that model 1 is nested in model 2 by the addition of extra parameters in model 2. We also indicate the conclusions from our analysis.

Table 2

Pairwise comparisons of models **A**, **B**, **C** *and* **D**. Δr *is the decrease in residual degrees of freedom for each gene and remains unchanged for these four comparisons.* ΔD_{scaled} *is the corresponding scaled decrease in deviance calculated from Eq. (9).*

gene	Δr	ΔD_{scaled}				omitted probes
		A → **B**	**A** → **C**	**B** → **D**	**C** → **D**	
37777_at	15	428.5	39.4	4.15	363.0	
684_at	14	662.9	55.1	2.03	544.5	7
1597_at	14	32.5	34.0	5.03	3.7	14
38734_at	11	266.5	33.9	0.57	212.1	7,8,14,15
39058_at	14	514.8	47.4	0.34	421.5	1
36311_at	15	722.6	23.1	1.13	659.6	
1024_at	15	178.1	41.5	1.85	126.9	
36202_at	15	369.9	25.7	7.71	333.9	
36085_at	15	106.1	20.3	2.58	83.9	
40322_at	15	420.6	20.2	0.00	378.0	
1091_at	11	196.0	40.0	0.00	140.0	8,10,11,12
1708_at	14	661.7	37.2	2.99	577.8	13
All genes	168	4560.2	418.0	28.38	3844.9	

Table 3
Comparisons of models $\mathbf{B} \rightarrow \mathbf{E}$ *and* $\mathbf{B} \rightarrow \mathbf{F}$*. Column headings have the same meanings as in Table 2.*

model comparison and gene	Δr	ΔD_{scaled}	omitted probes
$\mathbf{B} \rightarrow \mathbf{E}$			as in
All genes	2	2894.8	$\mathbf{B} \rightarrow \mathbf{F}$ below
$\mathbf{B} \rightarrow \mathbf{F}$			
37777_at	16	4.48	
684_at	15	18.74	7
1597_at	15	8.75	14
38734_at	16	29.25	
39058_at	15	24.74	1
36311_at	16	7.27	
1024_at	16	3.48	
36202_at	16	8.17	
36085_at	16	11.90	
40322_at	16	49.53	
1091_at	16	4.58	
1708_at	15	5.86	13
All genes	188	187.46	

are not significant. We therefore accept model **B**, the equilibrium Langmuir isotherm augmented with probe dependent asymptotic saturation intensities, as the best supported of models **A** to **D** for these data.

Table 3 shows calculated values of ΔD_{scaled} for the extensions of model **B** to model **E** with the introduction of an overall wafer dependent scaling, and to model **F** with the introduction of a probe dependent Sips parameter. Once again we have omitted from the analysis probes corresponding to fits giving unphysical negative values to the parameters b and K. The decrease in scaled deviance for the comparison $\mathbf{B} \rightarrow \mathbf{E}$ is well beyond the 0.001 significance level, indicating a scaling effect arising from the three replicate wafers. The fitted values obtained are

$$\lambda_1 = 0.883, \quad \lambda_2 = 1.104, \quad \lambda_3 = 1.014. \tag{10}$$

Considering the comparison $\mathbf{B} \rightarrow \mathbf{F}$, we see that the introduction of Sips parameters α_{pg} is significant at the 95% level for two out of the twelve probes,

and that for all probesets taken collectively, ΔD_{scaled} is close to the 50th χ^2 percentile. We conclude that the introduction of a Sips parameter to account for a distribution of probe-target binding energies overall is not necessary for PM probes. A histogram of the fitted values to α_{pg} from model **F** is given in Figure 3.

In summary, we find model **E**, the Langmuir isotherm with probe dependent parameters y_0, b and K and overall scaling λ_w to account for a systematic wafer dependence, to be the best supported model of those listed in Table 1. Plots of the fits of the fluorescence intensity data of both PM and MM probes for gene 37777_at to model **E** are shown in Figure 4. The strong dependence of the parameters of the Langmuir isotherm on individual probe sequences is manifest; note the different scales on the vertical axes. The relationship between the respective isotherm parameters of the PM and MM probes is studied in a companion paper (Burden et al., 2004).

4. Estimation of mRNA concentration

Ultimately one would like to predict absolute target concentrations from measured fluorescence intensities using an adsorption model with parameters solely determined from specific probe sequences. As a first step in this direction, Hekstra et al. (2003) have fitted their estimated parameters from the Langmuir isotherm model of Eq. (5) (our model **B**) to a simple linear model

$$\begin{pmatrix} \ln b \\ \ln K \\ \ln y_0 \end{pmatrix} = \begin{pmatrix} \gamma_{bA} & \gamma_{bC} & \gamma_{bG} \\ \gamma_{KA} & \gamma_{KC} & \gamma_{KG} \\ \gamma_{y_0A} & \gamma_{y_0C} & \gamma_{y_0G} \end{pmatrix} \begin{pmatrix} n_A \\ n_C \\ n_G \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}, \quad (11)$$

where n_A, n_C and n_G are the number of A, C and G nucleotides occurring in each probe. Note that the constraint $n_A + n_C + n_G + n_T = 25$ obviates the need for an explicit n_T dependence. They then used the fitted parameters γ_{ij} and C_i to test the ability of the model to recover known spiked in RNA concentrations given individual probe sequences and fluorescence intensities. Here we redo this analysis, extending the methodology to accommodate more accurate estimation of errors and reduction of bias.

Our results of fitting the parameters y_0, b and K obtained from the probe and wafer dependent Langmuir isotherm model **E** to the regression parameters γ_{ij} and C_i are shown in Table 4. Only those probes which give unphysical negative values to the parameters b or K, namely those listed in Table 3, have been omitted from our analysis. By contrast, Hekstra et al. only work with two of the three wafers and omit almost 30% of the remaining probes which they consider to be unsuitable.

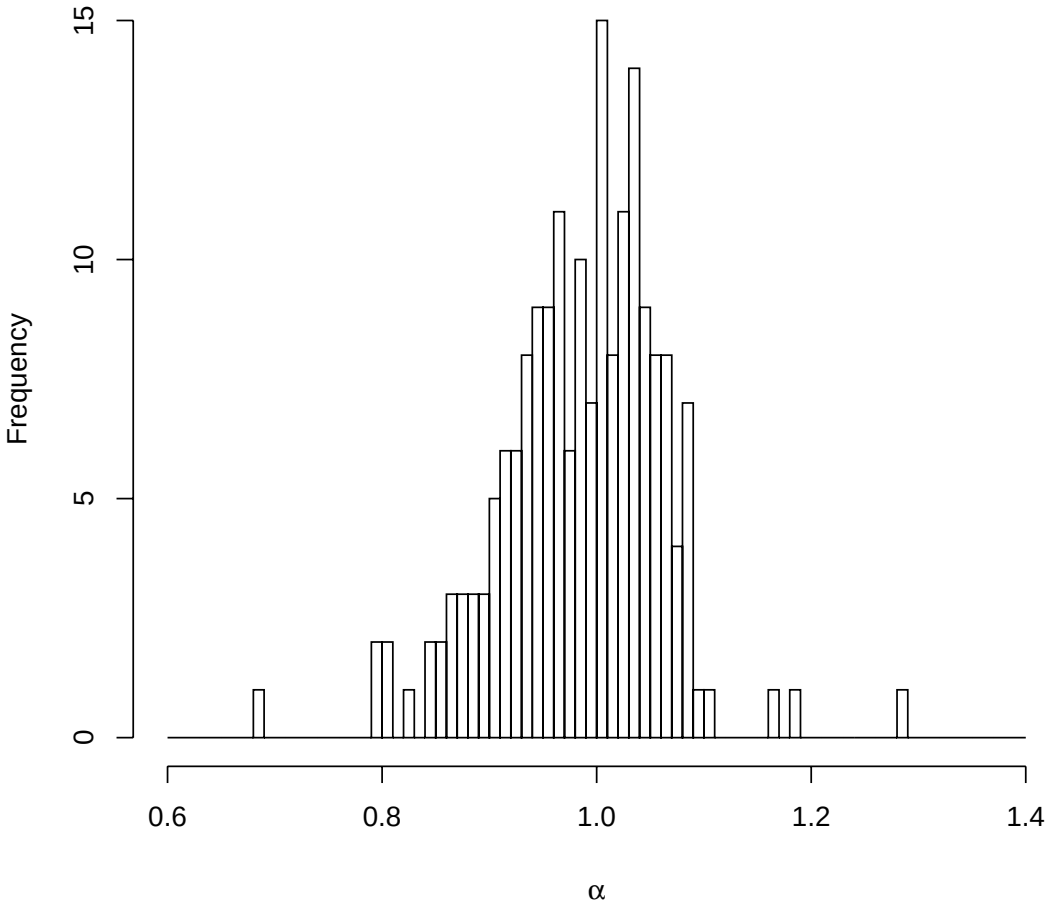

Figure 3. Histogram of fitted values of Sips parameters α_{pg} of model **F**.

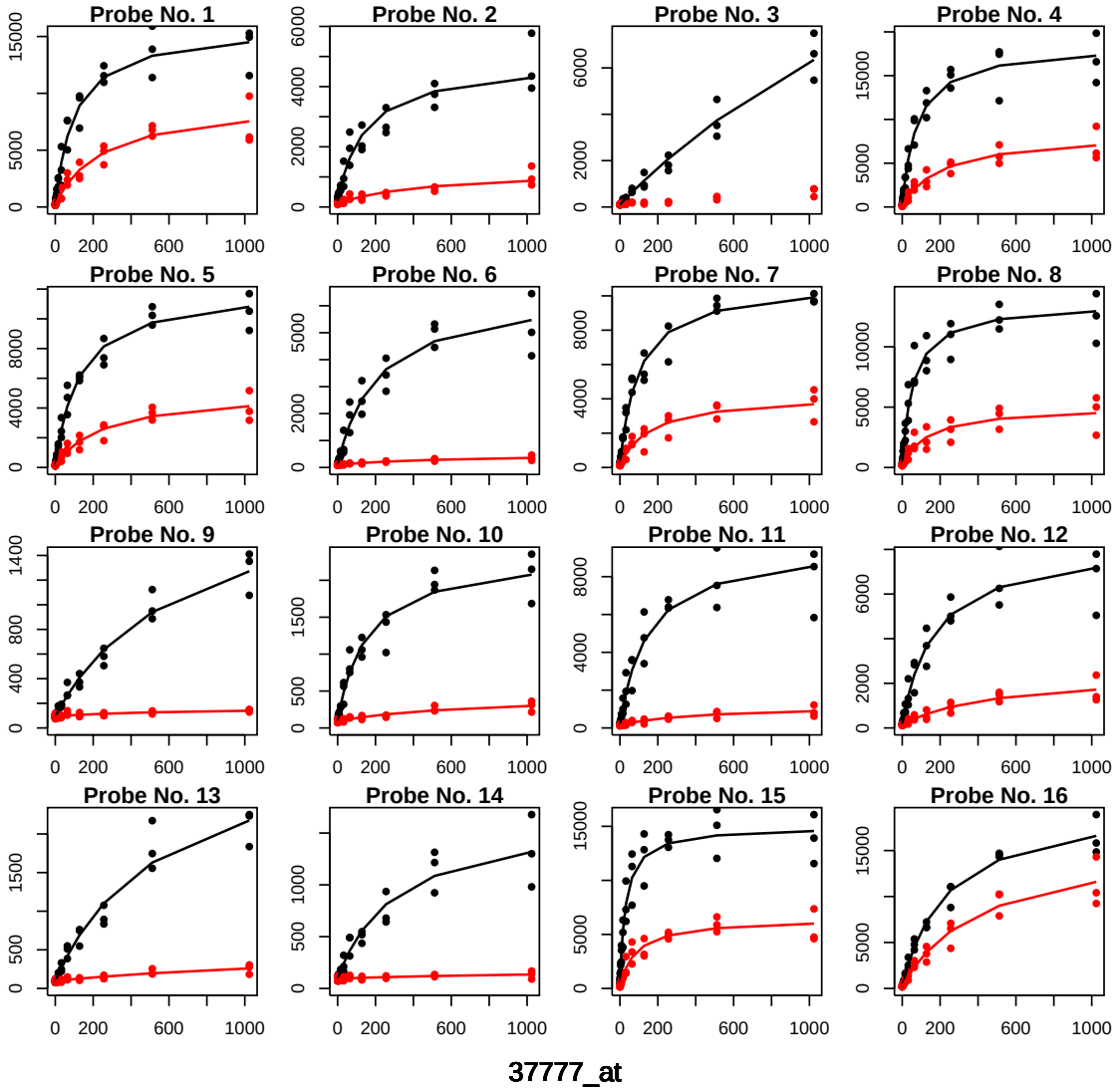

Figure 4. Fits to model **E** of fluorscence intensity data for the 16 PM (black) and 16 MM (red) features of the gene 37777_at probeset of the Affymetrix Latin Square experiment. Concentrations (horizontal axes) are in picomolar and fluorescence intensties (vertical axes) are in the arbitrary units used in Affymetrix .cel files. The fit to MM probe No. 3 gave unphysical negative values to the parameters K and b and is not shown.

Table 4

Fitted parameters (with standard errors in brackets) for the model of Eq. (11) using the parameters y_0, b and K obtained from the Langmuir isotherm model **E**.

	C_i	γ_{iA}	γ_{iC}	γ_{iG}
$\ln b$	5.957 (0.294)	-0.030 (0.019)	0.304 (0.025)	0.223 (0.024)
$\ln K$	0.917 (0.417)	0.210 (0.027)	0.175 (0.036)	0.341 (0.034)
$\ln y_0$	4.437 (0.263)	-0.133 (0.017)	0.185 (0.023)	0.067 (0.021)

Having constructed a model for obtaining estimates $\hat{b}_p$, $\hat{K}_p$ and $\hat{y}_{0p}$ of the Langmuir parameters from probe sequences, Eq. (5) can be inverted to give an estimate of concentration from fluorescence intensity y for each probe,

$$\hat{x}_p = \frac{\hat{K}_p(y - \hat{y}_{0p})}{\hat{b}_p + \hat{y}_{0p} - y}, \tag{12}$$

where the wafer dependence λ_w in model **E** of Table 1 has been omitted to reflect the fact that, in an experimental situation, the overall scaling effect will not be known.

Two questions immediately arise. Firstly, what is the best way to extract a single concentration estimate from the n_{probe} values obtained for a complete probeset, and secondly, how can unphysical estimates obtained from Eq. (12) be dealt with? Note that unphysical estimates of $\hat{x}_p$ can arise in two ways: (i) if $y < \hat{y}_{0p}$, the measured fluorescence intensity falls below the estimated background level $\hat{y}_0$, yielding a negative concentration, and (ii) if $y > \hat{y}_{0p} + \hat{b}_p$ the measured intensity falls above the estimated saturation intensity and the wrong branch of the hyperbola is read from Eq. (12). Both situations can be expected to occur, given the underlying statistical nature of the observed intensity.

Hekstra et al. (2003) propose a target gene concentration estimate given by averaging logged probe estimates. They deal with the problem of unphysical concentration estimates by removing the offending probes from the probeset, giving

$$\ln \hat{x}_{\text{gene}} = \frac{1}{n_S} \sum_{p \in S} \ln \hat{x}_p, \tag{13}$$

where S is the subset of probes within the probeset for which $\hat{y}_0 < y < \hat{y}_{0p} + \hat{b}_p$ and n_S is the number of elements in S. This introduces downward bias at

high target concentrations and upward bias at low target concentrations because valid data contributing to the random distribution about the expected intensity value has been removed. The extent of the problem is apparent in Figure 5 which shows the percentage of $\hat{x}_p$ values discarded at each concentration. The situation can be improved at low intensities by replacing the logarithm in Eq. (13) by arcsinh, which approximates a linear function near zero and so accepts negative arguments. However, this does nothing to solve the saturation problem at high concentrations.

Our alternative proposal is to replace Eq. (12) by

$$\ln \hat{x}_p = \begin{cases} X & \text{if } y > \hat{y}_{0p} + \hat{b}_p \\ \ln\left[\hat{K}_p(y - \hat{y}_{0p})/(\hat{b}_p + \hat{y}_{0p} - y)\right] & \text{if } \hat{y}_{0p} < y < \hat{y}_{0p} + \hat{b}_p \\ -X & \text{if } y < \hat{y}_0 \end{cases} \tag{14}$$

where X is some arbitrarily chosen large value acknowledging the existence of data beyond the capacity of the estimated inverse Langmuir isotherm. We then replace the mean in Eq. (13) by a median,

$$\ln \hat{x}_{\text{gene}} = \text{median}(\ln \hat{x}_p), \tag{15}$$

which will in general be unbiased by the magnitude of our artificially introduced outliers $\pm X$.

Confidence intervals can be placed on the concentration estimates using bootstrapping with uniform resampling (Hall, 1992). A set of B estimates of concentration is obtained, where at each iteration a probeset is constructed by randomly resampling with replacement n_{probe} probes from the original probeset and treating each resampled probeset analogously to the original data. Approximate equal tail 95% confidence interval limits are then given by the 2.5 and 97.5 percentiles of this set of B estimates.

Figures 6 and 7 show median log and mean log estimates of concentrations obtained from Eqs. (15) and (13) respectively, plotted against known spiked-in concentrations for each of the 12 genes considered here. Error bars in each case are equal tailed 95% confidence intervals obtained from bootstrap resampling with $B = 100$. Comparing the two methods we see that the expected downward bias at high concentration has been corrected in several of the genes (namely 684_at, 38734_at, 36311_at, 36202_at and, to some extent, 1708_at) by using the unbiased median estimator instead of the mean log. The downward bias in genes 37777_at and 40322_at appears to have been overcorrected by the median estimator, though this could be the effect of attempting to estimate concentrations at saturation levels for these probesets.

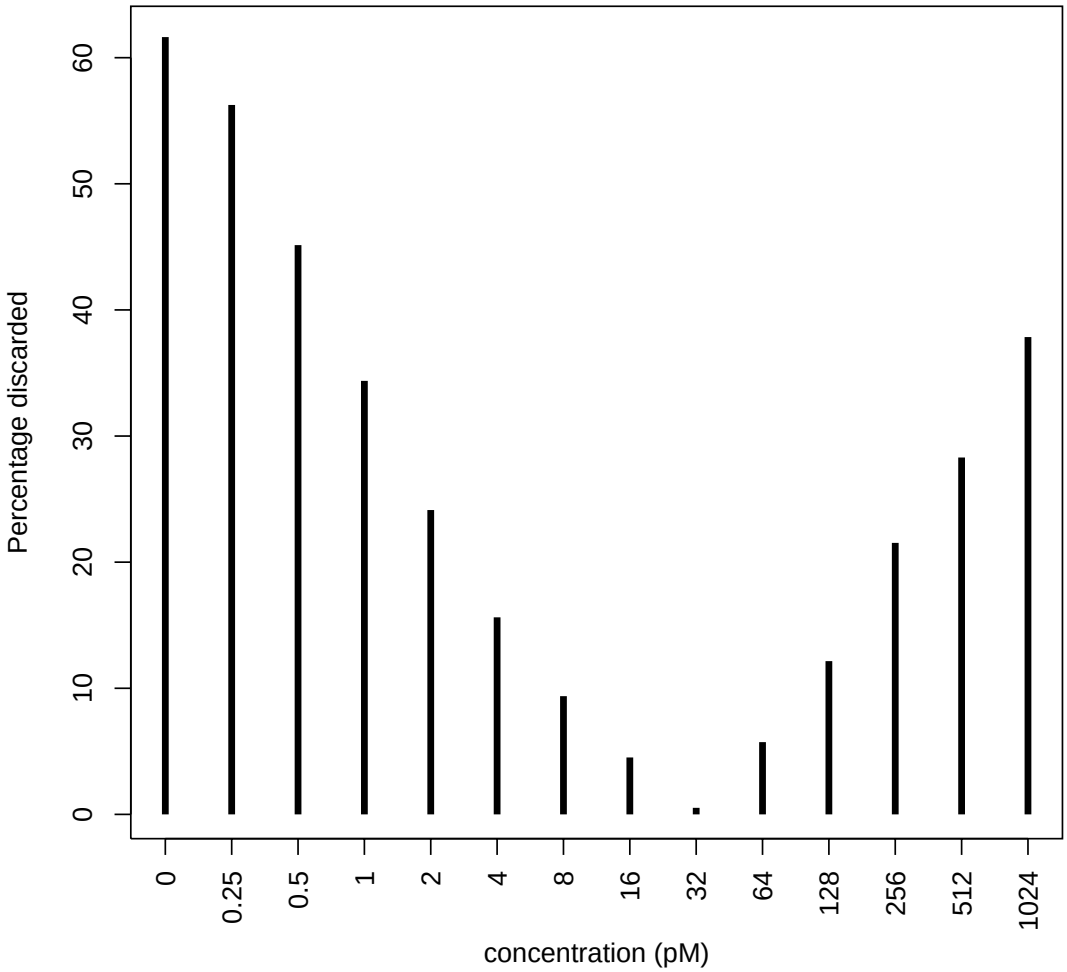

Figure 5. Percentage of concentration estimates $\hat{x}_p$ discarded by Eq. (13) at each concentration averaged over the 12 genes considered in our analysis.

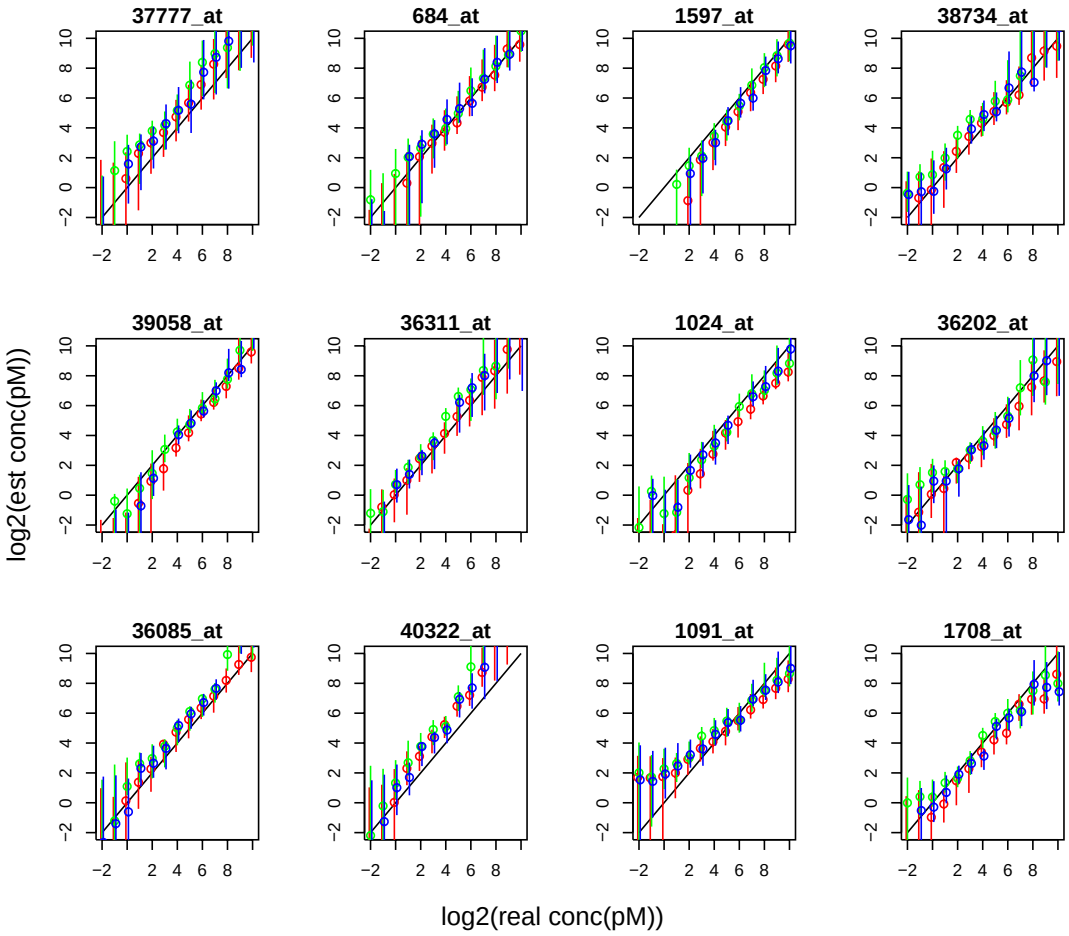

Figure 6. Estimates of RNA concentration from medians of estimates of probe intensities, Eq. (15), plotted against actual spiked-in concentrations. Error bars are approximate 95% confidence intervals obtained by bootstrap resampling. The line indicates the perfect relationship between predicted and actual concentration.

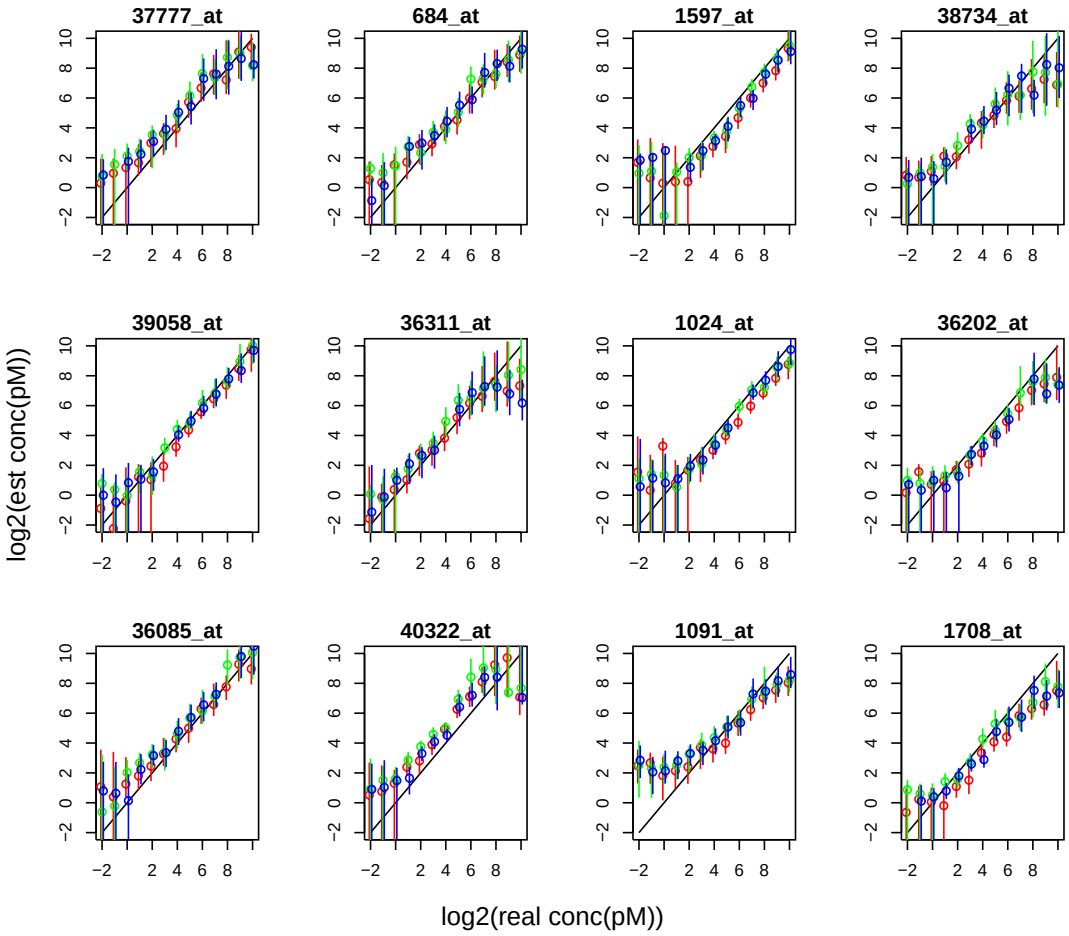

Figure 7. Estimates of RNA concentration from averaged logged estimates of probe intensities, Eq. (13), plotted against actual spiked-in concentrations. Error bars are approximate 95% confidence intervals obtained from bootstrap resampling. The line indicates the perfect relationship between predicted and actual concentration.

At low concentrations the median estimator has corrected the upward bias in genes 38734_at, 1024_at, 36202_at, 36085_at, 40322_at and 1708_at, and overcorrected somewhat for genes 684_at, 1597_at and 39058_at. While there has been a slight improvement for gene 1091_at, for which both methods perform poorly, we see that the bootstrap error bars more effectively include the correct value when applied to the full probeset used for the median method rather than the truncated probeset used in the mean log method.

In Figure 8 we show for comparison plots of MAS5 (Microarray Suite version 5) (Affymetrix Inc., 2002) and RMA (Robust Multiarray Average) (Irizarry et al., 2003) expression measures for the Latin Square data set computed using functions supplied by the Bioconductor software package (Gautier et al., 2003) with default settings. Also plotted are lines indicating the expected slope of these plots assuming the log(MAS5) and RMA expression measures are intended to be indicators of logged mRNA target concentration. In both cases concentration fold changes are underestimated, particularly at higher concentrations. It is clear that the use of inverse Langmuir isotherms are a considerable improvement over existing measures in estimating fold changes over a broad range of concentrations, including the approach to saturation.

We mention in passing a claim by Yung et al. (2004) that target concentation can be recovered from the MAS5 expression measure using a linear fit of $\log_2$ of the MAS5 index to $\log_2$ of the spike-in concentration, namely by a universal linear fit with a slope of 0.71 to all of the upper curves in Figure 8. We question the accuracy of this claim, which relies on the linearity of a log plot of expression index fold changes against spike-in concentration fold changes over the full range of possible fold changes (see Figure 3B of Yung et al., 2004). We claim that this linearity is an artifact of having averaged over all possible pairs of log ratios with a given spike-in fold change. This procedure masks the systematic downward bias in expression index fold changes at saturation concentrations evident in Figure 8. Furthermore, unlike the approach from Langmuir adsorption theory, the use of a universal fit fails to take any account of probe sequence dependence, also evident in Figure 8.

Finally we note that a method for estimating target concentration from fluorescence intensities was also given by Held et al. (2003). In this case the parameters y_0 and K were estimated from probe binding free energies and the parameter b was assumed to be universal to all genes, an assumption inconsistent with our findings in Section 3.

5. Discussion and Conclusions

We have carried out a comparative statistical analysis of various forms of adsorption-desorption models of oligonucleotide microarray chips using PM

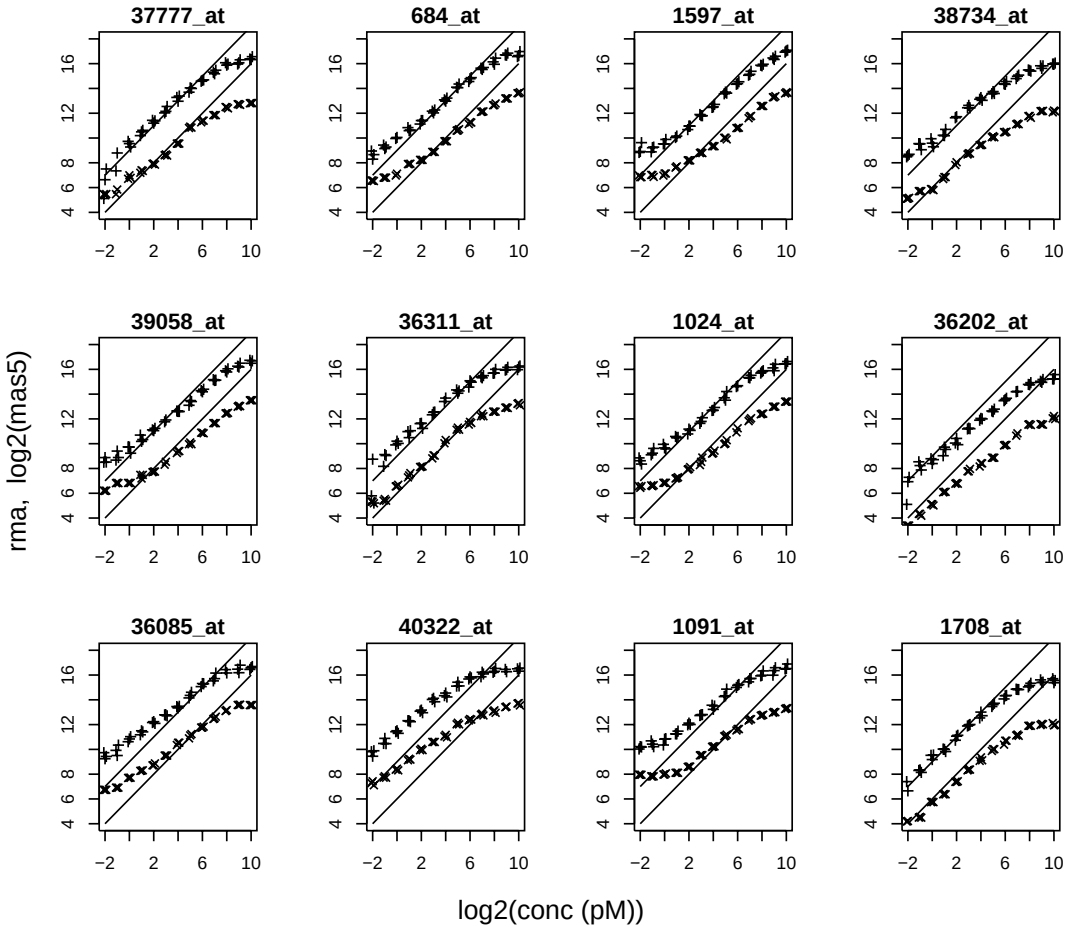

Figure 8. Log(MAS5) (+) and RMA (×) expression measures calculated from the Affymetrix Latin Square data. The slope of the lines indicate the behavior required of an expression measure which accurately tracks concentration fold changes

probe data from the Affymetrix Latin Square spike-in experiment. This analysis is a first step towards an accurate understanding of the physical and chemical processes driving the hybridization of labelled mRNA targets at the surface of Affymetrix Genechips. This paper has concentrated on statistical data analyses and we have been guided by broad physical principles in our choice of models.

Before summarizing our results, it is important to place the work in the context of broader goals. The first of these is to provide a practical method of estimating the absolute concentration of mRNA in biological samples taken in a real experimental situation. We have determined the appropriate chemical adsorption model for oligonucleotide microarrays to be a three parameter, probe sequence dependent Langmuir isotherm. The next step is to find physical and chemical explanations for the probe dependence of the parameters. Once an algorithm for determining isotherm parameters from probe sequences is established, then for each microarray there would be no practical impediment to supplying a data file of all parameters necessary for determining adsorption isotherms, and appropriate software for extracting a measure of mRNA concentration, complete with bootstrap confidence intervals. A second broad goal is to improve the design of microarray chips. If the probe sequence dependence of any Langmuir isotherm can be determined, one might ask whether it is possible to reduce the number of probes required per transcript, or whether the probe sequences could be chosen more effectively. Given that the response of the MM probes is also described by a Langmuir isotherm (see Figure 4), it is reasonable to ask whether better might use be made of the MM data, or whether the MM could probes be dispensed with altogether. Construction of a detailed model based on established principles of physical chemistry and consistent with the findings of this statistical analysis is the subject of ongoing work (Burden et al., 2004).

Returning to the statistical analyses in this paper, we based our comparison of models on an analysis of deviance within the framework of generalised linear models (McCullagh and Nelder, 1989). We chose to assume a gamma distribution with constant shape parameter as this treats the observed fluorescence intensity as an extensive variable. The coefficient of variation was observed to be significantly less than one for the data analysed. In this case the gamma distribution is similar in shape to a log normal distribution, and indicates that in an exploratory analysis, a log transformation may be satisfactory. Also, the gamma distribution could be generalized to allow for the observed slight upward tendency in the coefficient of variation (see appendix).

The most appropriate of the models considered was determined to be an equilibrium Langmuir isotherm specified by three parameters: background

intensity at zero concentration, saturation intensity at high concentration, and target concentration at half-saturation intensity. Importantly, all three parameters are necessarily probe dependent. This is in stark contrast with the claims of Held et al. (2003) who fit all probe responses across all genes to a common asymptote after binning the data by each probe's binding free energy. Their choice of model is more parsimonious than our model **A**, which we reject as being too parsimonious, and this choice implicitly assumes that all probes are 100% efficient. We contend that their binning procedure has no statistical justification, and that a physical explanation for a probe dependent saturation intensity must be sought elsewhere in the gene chip technology. In the current work we have also been able to dismiss the possibility that a common asymptotic intensity is masked by non-equilibrium dynamics.

To gauge the effectiveness of using the Langmuir isotherm to recover absolute target concentrations from fluorescence intensity data and probe sequence properties we have followed an approach by Hekstra et al. (2003), but have introduced improvements to reduce bias in the estimators and to enable estimation of confidence intervals. We find that a substantial downward (upward) bias at high (low) concentrations can be corrected by using an appropriately defined median over probes within a probeset rather than the mean.

We also find the Langmuir isotherm approach to be an improvement on existing expression measures such as MAS5 and RMA, both in its ability to predict absolute rather than relative target concentrations and in its ability to allow for saturation effects. There may well be useful aspects of existing expression measures which could be combined with adsorption models to produce a universal expression measure whose input parameters include probe sequence properties, and this could be explored.

Appendix: Microarray intensity data and the gamma distribution

Throughout our analysis we have assumed fluorescence intensity data from olignucleotide data to be drawn from a gamma distribution with constant shape parameter, or equivalently, with constant coefficient of variation. Here we give details of the arguments leading to this assumption.

In the Affymetrix HG-U95A Latin square experiment, each intensity measurement at a given target concentration for a given gene g and probe p is replicated with chips from $n_{\text{wafer}} = 3$ separate wafers w. An estimate of the coefficient of variation can be obtained by considering the set of quantities

$$\eta_{pg} = \frac{\sqrt{\text{Var}(y_{pg})}}{\bar{y}_{pg}} \tag{16}$$

where

$$\bar{y}_{pg} = \frac{\sum_{w=1}^{n_{\text{wafer}}} y_{pgw}}{n_{\text{wafer}}}, \quad \text{Var}(y_{pg}) = \frac{\sum_{w=1}^{n_{\text{wafer}}} (y_{pgw} - \bar{y}_{pg})^2}{n_{\text{wafer}} - 1}, \tag{17}$$

and y_{pgw} are individual fluorescence intensity measurements. The average over the 2688 values of η_{pg} obtained from the PM data for the 12 genes analyzed in this paper is $\hat{\eta}_{pg} = 0.17$. To estimate the variation in η over the range of intensity measurements we fitted the linear model

$$\log_{10} \eta = \alpha + \beta \log_{10} \bar{y} + \epsilon, \tag{18}$$

obtaining $\beta = 0.080$. Over the three orders of magnitude of intensity values represented this corresponds to a slight upward tendency in the coefficient of variation from about 0.12 at the lowest intensities to 0.18 at the highest intensities. We have repeated this analysis for each gene separately and find that the estimate $\hat{\eta}_{pg}$ of the coefficient of variation changes little from gene to gene, staying within the range 0.14 to 0.19 for the 12 genes studied.

Evidence for the choice of a gamma distribution is found in the work of Dennis and Patil (1984) on stochastic fluctuations of populations about a stable equilibrium. This work is concerned with dynamic population models governed by a differential equation of the form

$$\frac{d\theta}{dt} = \theta g(\theta), \tag{19}$$

which exhibit a stable equilibrium solution $\bar{\theta}$ satisfying

$$g(\bar{\theta}) = 0, \quad g'(\bar{\theta}) < 0. \tag{20}$$

The adsorption model of Eq. (1) is the specific case of these models with

$$g(\theta) = k_{\text{f}} \left\{ \frac{x}{\theta} - (x + K) \right\}, \tag{21}$$

and equilbrium solution given by the Langmuir isotherm $\bar{\theta} = x/(x + K)$. Dennis and Patil consider in detail a stochastic version of Eq. (19),

$$\frac{d\theta}{dt} = \theta \left[g(\theta) + h(\theta) z(t) \right], \tag{22}$$

in which a Gaussian white noise $z(t)$, with variance σ^2, times a density dependence $h(\theta)$ has been added to $g(\theta)$. They further find the probability density function of the variable θ to be given by (up to normalization)

$$f(\theta) = \exp \left\{ \frac{2}{\sigma^2} \int_{\bar{\theta}}^{\theta} \frac{g(n) dn}{n h^2(n)} - \frac{2\omega}{\sigma^2} \log \theta - \frac{2\omega}{\sigma^2} \log h(\theta) \right\}, \tag{23}$$

where ω is the "Ito-Stratonovich" parameter determining the stochastic integral by which the model of Eq. (22) is interpreted. It takes the values

$$\omega = \begin{cases} \sigma^2 & \text{if Ito calculus is used;} \\ \sigma^2/2 & \text{if Stratonovich calculus is used.} \end{cases} \tag{24}$$

Expanding the terms $g(n)/h^2(n)$ and $\log h(\theta)$ in the integrand of Eq. (23) to first order in a Taylor series about the stable equilibrium solution $\bar{\theta}$, Dennis and Patil show that $f(\theta)$ is approximately a gamma distribution.

Suppose now that we consider a simple expedient but realistic model $h(\theta) = c(k_\mathrm{f} x/\theta)^{1/2}$, where c is constant, so that the stochastic version of Eq. (1) becomes

$$\frac{d\theta}{dt} = k_\mathrm{f} x(1-\theta) - k_\mathrm{b}\theta + c(k_\mathrm{f} x\theta)^{1/2} z(t), \tag{25}$$

and the scale of the stochastic noise increases montonically with both the target concentration and fraction of occupied probes. Then it is straightforward to show that the distribution in Eq. (23) becomes an exact gamma distribution

$$f(\theta) = \psi \theta^{\nu-1} e^{-\alpha\theta} \tag{26}$$

where

$$\alpha = \frac{2(x+K)}{c^2\sigma^2 x}, \quad \nu = \frac{2}{c^2\sigma^2} + 1 - \frac{\omega}{\sigma^2}, \tag{27}$$

and ψ is a normalization constant ensuring $\int_0^\infty f(\theta) = 1$. Note that the shape parameter ν is constant with respect to x, ensuring constant coefficient of variation, and that, if the Ito scheme is used, the mean of the distribution is $\nu/\alpha = x/(x+K) = \bar{\theta}$.

The fluorescence intensity is taken in Section 2 to be the sum of a term proportional to the fraction θ of occupied probes and a background component assumed mainly to be due to cross hybridization, which will itself be driven by a stochastic process similar to that described above.

References

Affymetrix Inc. (2002). Statistical algorithms description document. Available at http://www.affymetrix.com/support/technical/whitepapers.affx.

Atkins, P. W. (1994). *Physical Chemistry*. Oxford University Press, Oxford, UK, 5th edition.

Burden, C. J., Pittelkow, Y. and Wilson, S. R. (2004). An adsorption model of hybridization behaviour on oligonucleotide microarrays. ArXiv q-bio.BM/0411005.

Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarrays. *Journal of Biomedical Optics* **2**, 364–374.

Cox, D. R. and Snell, E. J. (1981). *Applied Statisitics*. Chapman and Hall, New York, NY, USA, 1st edition.

Dennis, B. and Patil, G. P. (1984). The gamma distribution and weighted multimodal gamma distributions as models of population abundance. *Mathematical Biosciences* **68**, 187–212.

Forman, J. E., Walton, I. D., Stern, D., Rava, R. P. and Trulson, M. O. (1998). Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesised oligonucleotide arrays. In Leontis, N. B. and SantaLucia, J., editors, *Molecular Modeling of Nucleic Acids, ACS Symposium Series*, volume 682, pages 206–228. Am. Chem. Soc., Washington, DC, USA.

Gautier, L., Irizarry, R. A., Cope, L. and Bolstad, B. (2003). Textual description of affy. Vignette available at the Bioconductor wesite http://www.bioconductor.org.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York, NY, USA, 1st edition.

Hekstra, D., Taussig, A. R., Magnasco, M. and Naef, F. (2003). Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research* **31**, 1962–1968.

Held, G. A., Grinstein, G. and Tu, Y. (2003). Modeling of DNA microarray data by using physical properties of hybridization. *Proceedings of the Natlional Academy of Science* **100**, 7575–7580.

Irizarry, R. A., Hobbs, B., Collin, F. and Beazer-Barclay, Y. D. (2003). Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, UK, 2nd edition.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Nguyen, D. V., Arpat, A. B., Wang, N. and Carroll, R. J. (2002). Microarray experiments: Biological and technological aspects. *Biometrics* **58**, 701–717.

Peterson, A. W., Heaton, R. J. and Georgiadis, R. M. (2001). The effect of surface probe density on DNA hybridization. *Nucleic Acids Research* **29**, 5163–5168.

Peterson, A. W., Wolf, L. K. and Georgiadis, R. M. (2002). Hybridization of

mismatched or partially matched DNA at surfaces. *Journal of the American Chemical Society* **124**, 14601–14607.

Sips, R. (1948). On the structure of a catalyst surface. *Journal of Chemical Physics* **16**, 490–495.

Webster, T., Mei, R., Hubbell, E., Shen, M. and Bekiranov, S. (2003). Use of langmuir adsorption isotherm models to predict probe response. Talk at the Affymetrix Genechip Microarray low level workshop, August 7-8 (2003), Berkeley, Ca.

Yung, C. K., Halperin, V. L., Tomaselli, G. F. and Winslow, R. L. (2004). Gene expression profiles in end-stage human idiopathic dilated cardiomyopathy: altered expression of apoptotic and cytoskeletal genes. *Genomics* **83**, 281–297.