

Physico-chemical modelling of target depletion during hybridization on oligonucleotide microarrays

Conrad J Burden¹ and Hans Binder²

¹ Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia

² Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Haertelstrasse 16-18, Germany

E-mail: Conrad.Burden@anu.edu.au and binder@rz.uni-leipzig.de

Received 29 July 2009


Accepted for publication 17 November 2009

Published 21 December 2009

Online at stacks.iop.org/PhysBio/7/016004

Abstract

The effect of target molecule depletion from the supernatant solution is incorporated into a physico-chemical model of hybridization on oligonucleotide microarrays. Two possible regimes are identified: local depletion, in which depletion by a given probe feature only affects that particular probe, and global depletion, in which all features responding to a given target species are affected. Examples are given of two existing spike-in data sets experiencing measurable effects of target depletion. The first of these, from an experiment by Suzuki *et al* using custom built arrays with a broad range of probe lengths and mismatch positions, is verified to exhibit local and not global depletion. The second data set, the well-known Affymetrix HGU133a latin square experiment, is shown to be very well explained by a global depletion model. It is shown that microarray calibrations relying on Langmuir isotherm models which ignore depletion effects will significantly underestimate specific target concentrations. It is also shown that a combined analysis of perfect match and mismatch probe signals in terms of a simple graphical summary, namely the hook curve method, can discriminate between cases of local and global depletion.

 This article has associated online supplementary data files

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Physico-chemical models describing the processes involved in converting concentrations of specific RNA or DNA targets hybridized onto oligonucleotide microarrays to observed fluorescence intensities have become commonplace [5–7, 10, 13–17, 19–21, 25, 26]. The ultimate aim of such models has in general been to provide biologists with practical algorithms for estimating absolute specific target concentrations in the presence of a complex non-specific background from fluorescence intensity data. Early models inspired by Langmuir adsorption theory, which applied standard physical chemistry to the hybridization of specific

and non-specific targets to the microarray surface, predicted a hyperbolic response function [19, 20] which has been verified with reasonable accuracy [14] for the Affymetrix U95a latin square spike-in experiment [1]. Refinements of the model to include the effects of probe and target folding and bulk hybridization in the supernatant solution [5, 16] maintain the hyperbolic shape of the response function while decreasing the effective adsorption rate constant. Including the effects of post-hybridization washing [15, 21] also maintains the hyperbolic shape of the response function and is able to explain an asymptotic response in the limit of high target concentrations which is below full saturation of the probe feature and decreases with probe-target binding affinity [13].

The above physico-chemical models generally assume that the concentration of target molecules in the supernatant solution is not appreciably depleted by the hybridization reaction. However, in order to explain their data from spike in experiments which run to very low spike-in concentrations [30], Ono *et al* [27] have recently extended the accepted adsorption model to include such target depletion effects. Their model predicted an interesting saturation effect which was borne out by experiment. As well as the usual saturation effect, in which the number of available probe molecules becomes exhausted in the limit of high target concentration for a fixed probe type, a second saturation effect occurs when the number of target molecules is exhausted in the limit of high binding affinity at fixed target concentration. This limit was realised by including on a custom-built microarray a series of features of increasing probe length.

In the current paper we extend the Ono model by identifying two types of target depletion, which we term ‘local depletion’ and ‘global depletion’. By local depletion we mean that depletion of target molecules in the supernatant solution by a hybridization to a given probe feature only affects that particular probe feature. This is essentially Ono *et al*’s ‘finite hybridization model’. This regime is relevant when diffusion and/or convection of targets is slow compared with the hybridization and probe features responsive to the same target species are spatially separated on the microarray. By global depletion we mean that all probe features responding to a given target species are mutually affected by depletion of that species from the supernatant solution. Global depletion is relevant for spatially separated features undergoing permanent agitation of the hybridization solution, if equilibrium includes rapid diffusion of transcripts through the microarray cartridge, or for neighbouring features such as the perfect match/mismatch (PM/MM) probes on the older designs of Affymetrix GeneChips.

We fit the models to two spike-in data sets. The first, that of Suzuki *et al* [30], which covers a broad range of spike-in concentrations and probe lengths and for which we verify that the local model, not the global model, is relevant, is dealt with in section 2. The second, the U133A Affymetrix latin square data set [1], for which the global model is appropriate, is dealt with in section 3. For this data set we find that the global model of depletion entails a substantial improvement on earlier reported fits by a hyperbolic response function [13]. As well as fitting response functions or so-called ‘isotherms’, we analyse the data sets in terms of the recently developed hook curve formalism [8, 11] designed for the calibration of microarrays whose design includes PM/MM pairs. The hook curve method turns out to be a clear and easily implemented indicator of which depletion regime, local or global, is relevant to a particular data set.

Full details of our local and global depletion models, including specific and non-specific hybridization of target molecules to probes at the microarray surface and of targets within the supernatant solution and the folding of target and probe molecules, are set out in Appendix A. Some technical details of the analysis of the global depletion model are given in Appendix B.

Other than the work of Ono *et al* and a related project [26], we are aware of only one other extensive attempt to incorporate target depletion from the supernatant solution during hybridization into a physico-chemical model of microarrays, namely a recent publication by Li *et al* [23]. In section 4 we give a critical evaluation pointing out a number of errors in the Li *et al* model, with details given in Appendix C.

2. The Suzuki data set: an example of local depletion

Suzuki *et al* [30] have carried out experiments in which a set of 150 cDNA target sequences, with and without a complex background, are hybridized onto custom arrays containing features with probes ranging in length from $\ell = 14$ to 25 DNA bases. The probe designs include perfect matches and mismatches, the mismatches being in each possible position $(1, \dots, \ell)$ and of each possible nucleotide. Spike-in concentrations covered a broad range from 1.4 fM to 1.4 nM. The purpose of the experiment was to determine probe lengths and mismatch positions which optimize the discrimination between PM and MM signals. Because the spike-in concentrations run to very low values, depletion cannot be ignored [27]. Of the two physico-chemical models described in Appendix A, we demonstrate below that this data set is an example of local rather than global target depletion. This result is reasonable. The large set of PM and MM probes addressing any one target species must extend over distances large compared with the nearest neighbour distance on the chip. The remaining question, which we settle below in favour of local depletion, is whether diffusion or convection of target molecules is slow (local depletion) or fast (global depletion) relative to the rate of hybridization.

2.1. Theory

For the case of local depletion, the coverage fraction $0 \leq \theta \leq 1$ of fluorescent dye carrying target molecules bound to a given probe feature at the microarray surface is shown in Appendix A.1 to be

$$\theta = \frac{X_N + K_S(x_S - p\theta_S)}{1 + X_N + K_S(x_S - p\theta_S)}, \quad (1)$$

where x_S is the spiked-in probe-specific target concentration, p is an effective molar concentration of probe molecules immobilized on the microarray surface, X_N , called the non-specific binding strength, is a dimensionless measure of the degree of non-specific binding and K_S is an effective equilibrium constant for the binding of specific targets accounting for several chemical reactions including surface and bulk hybridization and molecular folding. The coverage fraction θ_S of specific targets only is given by equation (A.31). The model of Appendix A.1 also allows for the consideration of post-hybridization washing, which is signalled by differing responses of PM and MM features to saturation target concentrations [15]. There seems to be little evidence that washing is significant for this data set (see figure 3), and for convenience we set the washing survival factors to unity in the current analysis.

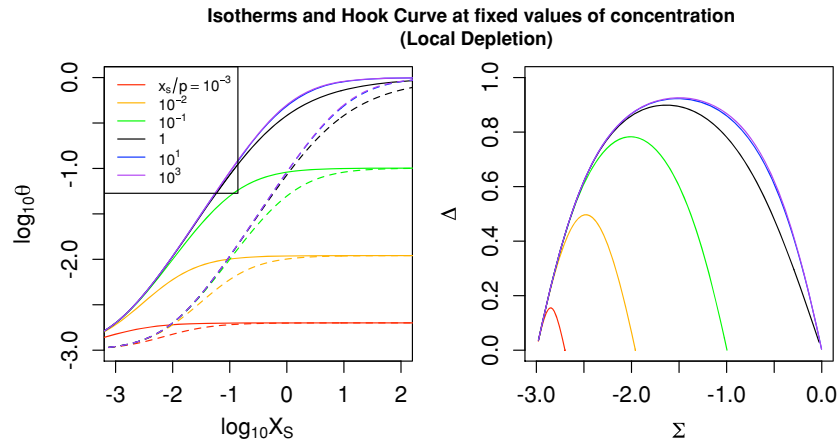


Figure 1. Theoretical isotherms and hook curves derived from the local depletion model of Appendix A.1. Each curve represents the response of the coverage fraction θ to variations of the specific binding strength for the PM probes, $X_S^{\text{PM}} = K_S^{\text{PM}} x_S$ at fixed specific target concentration x_S . Since $\log K_S^{\text{PM}} \propto$ free binding energy of hybridization (see the text), the horizontal scale in the isotherm plot can be thought of as a measure of probe length. Input parameters are: x_S/p as indicated by colour in the legend; $X_N = 10^{-3}$ and $K_S^{\text{MM}} = 0.1 K_S^{\text{PM}}$. Isotherms for PM probes are plotted as solid lines, and for MM probes as dashed lines.

The log of the effective equilibrium constant K_S is expected to be approximately proportional to probe length. This follows from the definition (A.27) and the relationship $K_{\text{PS}} \propto e^{\Delta G/(RT)}$ relating the hybridization constant to free binding energy ΔG , which is well approximated by the SantaLucia nearest neighbour stacking model [28]. Ono *et al* [27] make use of this result to consider isotherms relating coverage fraction to probe length, which we reproduce from the theoretical local depletion model in the left panel of figure 1. In calculating these curves we use an assumption that the ratio $K_S^{\text{PM}}/K_S^{\text{MM}}$ is independent of probe length. This is justified since $K_S^{\text{PM}}/K_S^{\text{MM}} \approx e^{\Delta\Delta G/(RT)}$ where $\Delta\Delta G = \Delta G^{\text{PM}} - \Delta G^{\text{MM}}$ is, on average, independent of probe length by virtue of the nearest neighbour stacking model.

Two saturation behaviours in the limit of large probe length, or high binding strength $X_S = K_S x_S$, are immediately apparent. From equations (A.24) and (A.31) one obtains

$$\lim_{X_S \rightarrow \infty} \theta = \lim_{K_S \rightarrow \infty} (\theta_S + \theta_N) = \begin{cases} 1 & \text{if } x \geq p, \\ x/p & \text{if } x < p. \end{cases} \quad (2)$$

In the case $x > p$, where the concentration of specific target exceeds the effective concentration of probes, the probes become saturated ($\theta = 1$) and any residual unbound targets remain in solution. In the case $x < p$, where the probe concentration exceeds that of the targets, the free targets are completely depleted and the maximum fluorescence intensity decreases with decreasing target concentration ($\theta = x/p$). Note also that for any PM/MM pair, the saturation intensity depends only on specific target concentration and not on the presence of mismatches.

Also shown in the right panel of figure 1 are the predicted hook curves for varying probe length in the case of local depletion. The hook curve method [8, 11] was originally developed to analyse data from microarrays whose design includes PM/MM pairs, but can be applied to any pair of probe features addressing the same specific target. The method

processes the PM/MM intensities I^{PM} and I^{MM} using the transformation

$$\begin{aligned} \Delta &= \log_{10} I^{\text{PM}} - \log_{10} I^{\text{MM}}, \\ \Sigma &= \frac{1}{2} (\log_{10} I^{\text{PM}} + \log_{10} I^{\text{MM}}), \end{aligned} \quad (3)$$

where, for Affymetrix GeneChips, the angular brackets denote averaging over probes within a probeset. Smoothing the Δ versus Σ plot provides a hook curve, whose characteristic shape typically assumes the concave downwards curve shown in figure 1.

In previous implementations [8, 11] the hook curve has been considered as a trajectory in the Σ - Δ plane as the specific binding strength $X_S = K_S x_S$ varies due to changes in specific target concentration x_S while the binding affinity K_S is held fixed. For the Suzuki data set we use a *different* and more appropriate implementation which specifically exploits the broad range of binding affinities arising from probe lengths which vary from 14 to 25 mer. That is, figure 1 plots the hook curve as a trajectory traced out by varying binding affinity K_S at fixed values of target concentration x_S . The left-hand end of the hook curve ($X_S = 0$) is determined by non-specific hybridization and will not vary significantly with probe length. The right-hand end of the hook curve ($X_S \rightarrow \infty$) is determined by the saturation intensity, and is expected to shift leftwards for subcritical specific target concentrations $x < p$.

By contrast, the theoretical isotherms and hook curves for the case of global depletion are shown in figure 2. For a given probe feature P , the theoretical isotherm, derived in Appendix A.2, is now (see equation (A.48))

$$\theta^P = \frac{X_N^P + K_S^P (x_S - p\theta_{\text{sum}})}{1 + X_N^P + K_S^P (x_S - p\theta_{\text{sum}})}, \quad (4)$$

where $\theta_{\text{sum}} = \sum_P \theta^P$ is the sum total of specific target coverage fractions over all probe features addressing the relevant target species, and is determined by an equation analogous to equation (A.47). For illustrative purposes the curves in figure 2 are calculated for the case of a single

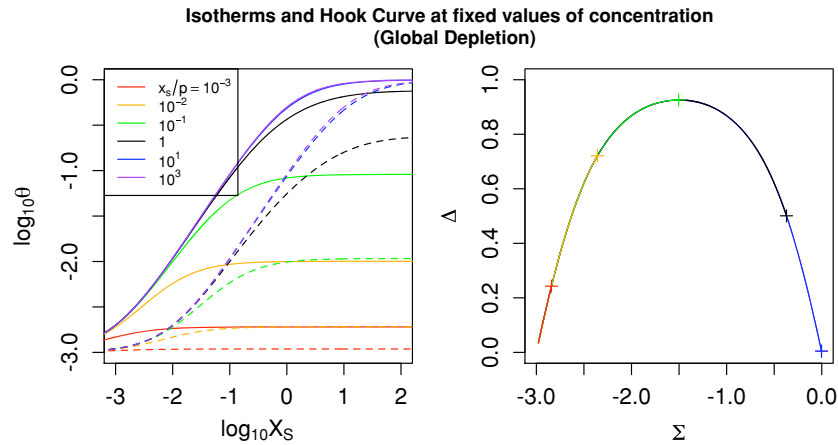


Figure 2. Theoretical isotherms and hook curves derived from the global depletion model of Appendix A.2. The input parameters and curve conventions are the same as for figure 1. Also shown in the right panel are the right-hand end points of individual hook curves indicated by a + sign. As explained in the text, the shape of the hook curve remains unchanged as x_S varies, except that the curve terminates at different right-hand points at subcritical concentrations. Hook curves for all values of x_S/p start at the same left-hand point. Note that the critical value of specific concentration, below which the isotherm saturates at $\theta < 1$ now occurs at a value $x_{crit}/p > 1$ as the depleted targets are shared among more than one probe feature.

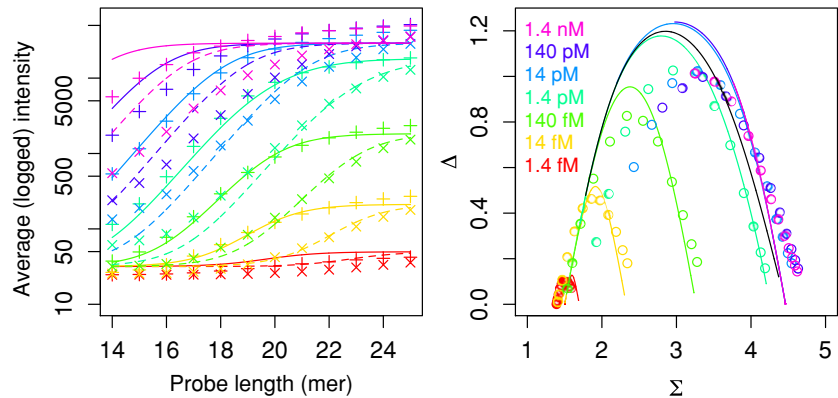


Figure 3. The Suzuki *et al* data set without complex background. Left panel: fluorescence intensities from PM (\times) and MM in the central nucleotide position (+) probes obtained by taking average logged intensities of 150 spiked cDNA sequences. Fits to the model equation 5 are shown as a solid curve (PM) and dashed curve (MM). Right panel: the corresponding hook curves and fits. The black hook curve corresponds to the fitted critical concentration $x_{crit} = p = 2.26$ pM.

PM/MM pair of probe features addressing the target in question. Two differences with the local depletion case are immediately apparent. Firstly, since the depleted targets are shared among more than one probe feature, the asymptotic behaviour of the isotherms at subcritical concentrations differs between different probes addressing the the same target species (i.e. $\lim_{x_S \rightarrow \infty} \theta_{PM} > \lim_{x_S \rightarrow \infty} \theta_{MM}$). Secondly, the shape of the hook curve remains unchanged as the probe length varies by the following argument. Since we use an assumption that K_S^{PM}/K_S^{MM} is independent of probe length, one can think of the hook curve as being parameterized by the variable $K_S^{PM}(x_S - p\theta_{sum})$ where θ_{sum} has some functional dependence on x_S , p , K_S^{PM} and the fixed ratio K_S^{PM}/K_S^{MM} . Changing the value of x_S then simply effects an identical reparameterization in this variable of both θ^{PM} and θ^{MM} . Individual points will migrate along the path of the hook curve, and at subcritical concentrations the curve will be truncated at the right-hand end at different points, but otherwise the shape of the hook curve remains unchanged.

2.2. Experiment

The Suzuki spike-in experiment [30] includes spike-in runs of 150 cDNA target sequences both with and without a complex background. To keep the analysis simple we analyse only the data set without a complex background. The data set with complex background provides very similar results with respect to target depletion.

In figure 3 are plotted average logged intensities ($I_{av.log} = 10^{\log_{10}(I)}$) over three replicates of the 150 sequences for PM and MM probes of varying lengths, the mismatches being in the central position of the probe. Thus each plotted data point is an average over 450 raw intensities. Some sort of averaging over probe sequences to account for the dependence of binding affinity on individual probe sequences was necessary in order to separate out the dependence on probe length. This is handled in the implementation of the hook curve for Affymetrix chips by correcting intensities with position- and nucleotide-dependent sensitivity profiles

Table 1. Parameters fitting the local depletion model equation (5) to the Suzuki data set.

Optical background intensity	A	31.7
Saturation intensity above background	B	2.90×10^4
Equilibrium constant of 20 mer PM probe	κ_{PM}	0.500 pM^{-1}
Equilibrium constant of 20 mer MM probe	κ_{MM}	0.022 pM^{-1}
Logarithmic length increment of K_S per nucleotide	λ	1.02
Bulk equivalent concentration of probes	p	2.26 pM

determined from intensity distributions over the whole array [11]. However, this method cannot be used for the Suzuki data set as each array contains a range of probe lengths, making it difficult to define meaningful sensitivity profiles. The use of average logged intensities rather than averaged intensities was an appropriate and simple solution accounting for the fact that microarray data is generally observed to have multiplicative errors. Comparison of the resulting hook curves in figure 3 with the theoretical hook curves in figures 1 and 2 shows clear evidence for local rather than global depletion.

Also shown in figure 3 are fits of a six-parameter model to the 168 data points (12 probe lengths \times 7 concentrations \times PM and MM). The model, based on the theoretical solution equation (1) for local depletion with $X_N = 0$, is defined by

$$I_{\text{av.log}}^P = A + B\theta^P, \quad P = \text{PM, MM} \quad (5)$$

where θ^P is the solution to

$$\theta^P = \frac{\kappa_P e^{\lambda(\ell-20)}(x - p\theta^P)}{1 + \kappa_P e^{\lambda(\ell-20)}(x - p\theta^P)}, \quad (6)$$

ℓ is the probe length and x the spike-in concentration. The parameters A and B account for the optical background intensity and saturation intensity, respectively, and the effective equilibrium constant in equation (1) is modelled by $K_S^P = \kappa_P e^{\lambda(\ell-20)}$. The fitted parameter values are listed in table 1. The fitted value of the effective probe concentration $p = 2.26 \text{ pM}$ is consistent with the observations of Ono *et al* [27].

3. The affymetrix latin square data set: an example of global depletion

Affymetrix have produced two well-known data sets [1] from experiments in which RNA transcripts were spiked in at cyclic permutations of a set of known concentrations together with a complex background of cRNA extracted from human pancreas or human adenocarcinoma cell line and hybridized onto U95a or U133 GeneChips, respectively. In a previous analysis [13] the U95a data set was shown to fit very well, and the U133 data set moderately well, to a physico-chemical model in which the target concentration was assumed not to be significantly depleted from the supernatant solution by hybridization to the microarray surface. This model was the $p = 0$ limit of the models in Appendix A. In this section we reanalyse the U133 data set and demonstrate that the global model of target depletion provides a significantly improved fit to this data.

3.1. Theory

The global model of target depletion is relevant to U133 Affymetrix GeneChips as the elements of a PM/MM pair of features are located in neighbouring locations on the microarray surface. Although each targeted gene is represented by 11 such probe pairs, we ignore depletion from other features within the same probeset as the design of the chip is such that those features are located elsewhere on the chip, and in general will target parts of the gene sequence further removed than the typical target fragment size of about 200 bases.

The coverage fraction θ^P , $P \in \{\text{PM, MM}\}$, of fluorescent dye carrying target molecules bound to the PM or MM feature at completion of the hybridization step is given by equation (4) where $\theta_{\text{sum}} = \theta^{\text{PM}} + \theta^{\text{MM}}$ is found by solving equation (A.47), and X_N^P and K_S^P are the non-specific binding strength and effective equilibrium constant for specific binding respectively. The loss of fluorescence intensity due to the post-hybridization washing step cannot be ignored for Affymetrix GeneChips [15, 21, 29], and we introduce into our model specific and non-specific washing factors w_S^P and w_N^P , respectively, where $1 > w_S^P > w_N^P > 0$. The post-washing coverage fraction is then given by equation (A.49). Finally, the observed fluorescence intensity is

$$\begin{aligned} I^P &= a + b\theta_{\text{after.wash}}^P \\ &= a + b \frac{w_N^P X_N^P + w_S^P K_S^P (x_S - p\theta_{\text{sum}})}{1 + X_N^P + K_S^P (x_S - p\theta_{\text{sum}})}, \quad P = \text{PM, MM}, \end{aligned} \quad (7)$$

where a and b are the physical background and absolute saturation intensities, assumed to be constant across the entire microarray.

3.2. Experiment

For the purposes of comparing fits of the spike-in data to a null-hypothesis model without depletion ($p = 0$) and the one-sided alternate hypothesis with depletion ($p > 0$), we rewrite the model defined by equations (7) and (A.47) in the form

$$I^P(x) = A^P + B^P \frac{K^P (x - p\theta_{\text{sum}})}{1 + K^P (x - p\theta_{\text{sum}})}, \quad P = \text{PM, MM}, \quad (8)$$

where $\theta_{\text{sum}}(x; K^{\text{PM}}, K^{\text{MM}}, p)$ is the solution in the physically relevant interval $0 \leq \theta_{\text{sum}} \leq 2$ to

$$\theta_{\text{sum}} = \sum_{P=\text{PM,MM}} \frac{K^P (x - p\theta_{\text{sum}})}{1 + K^P (x - p\theta_{\text{sum}})}. \quad (9)$$

Here we have suppressed the subscript S on the PM-specific spike-in concentration x_S and introduced the parameterization

$$A^P = a + bw_N^P \frac{X_N^P}{1 + X_N^P}, \quad (10)$$

$$B^P = b \left(w_S^P - w_N^P \frac{X_N^P}{1 + X_N^P} \right), \quad (11)$$

$$K^P = \frac{K_S^P}{1 + X_N^P}, \quad P = \text{PM, MM}. \quad (12)$$

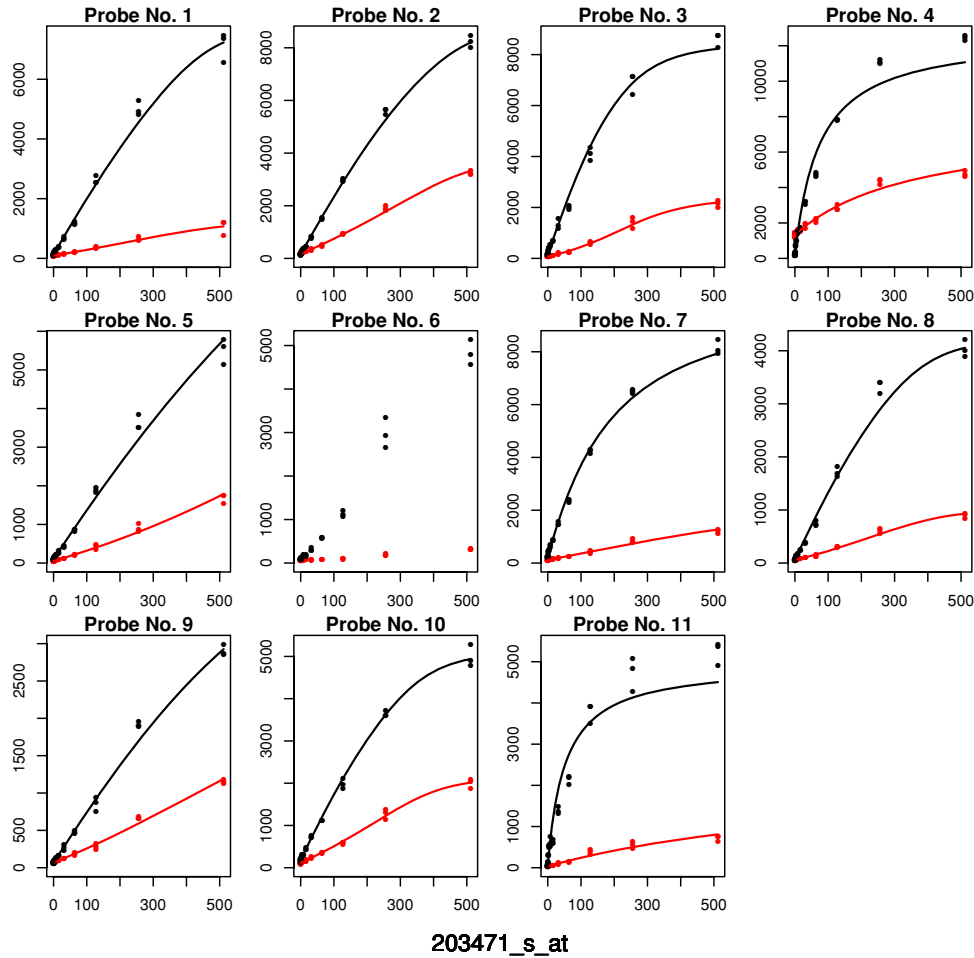


Figure 4. Fits of measured fluorescence intensities in .cel file units against spike-in concentrations in pM from a selected probeset of the spiked transcripts in the Affymetrix latin square U133 experiment to the seven-parameter model defined by equation (8). Note that a flattening of the PM isotherm and an inflection point in the MM isotherm, predicted in section 3.3 to be a characteristics of target depletion in certain parameter regimes, is clearly visible for most of these probes.

Equations (8) and (9) define a seven-parameter model (A^P , B^P , K^P , p) to which intensity data from a PM/MM pair of features for a range of spike-in concentrations x can be fitted. The $p = 0$ case, corresponding to no significant target depletion, defines a six-parameter model which was previously fitted to the U95a data in [14] and to both the U95a and U133 data in [15]. Below we use standard statistical methods to distinguish between a null hypothesis, $p = 0$, and alternate hypothesis $p > 0$.

Fluorescence intensities for each of 11 probe pairs from each of 38 spike-in transcripts of the U133 latin square experiment were fitted assuming the data to be Gamma distributed with mean given by the model of equation (8). The assumption of Gamma-distributed data was used in previous analyses [14] to accommodate a constant coefficient of variation as expected for data with multiplicative errors, and is easily implemented using the function `glm()` from the statistical computing environment R [2]. Fits of the model to the data of one of the spiked transcripts are plotted in figure 4, and analogous plots for all 38 spiked transcripts are available in the supplementary material (available at stacks.iop.org/PhysBio/7/016004/mmedia) or at the website of one of the authors [3].

Of the 418 probe pairs in the data set, 276 (or 66.0%) were successfully fitted to physically relevant values of the effective probe concentration restricted to the range of concentrations $p \geq 0$ with physical values for the remaining parameters, i.e. A_{PM} , B_{PM} , K_{PM} , A_{MM} , B_{MM} and K_{MM} all > 0 . This should be compared with fits to the $p = 0$ model without depletion in [13], for which only 37.5% of probes were successfully fitted to PM/MM probe pairs. A histogram of the fitted values of the effective probe concentration parameter p is shown in figure 5.

Immediately noticeable is that the distribution is bimodal: a number of fits are simply the ‘no depletion’ solution at $p = 0$, while most of the the remaining cases cluster around $p = 200$ pM. To understand this, note that equation (8) makes clear that the effect of depletion is to reduce the true target concentration x to an effective concentration

$$x_{\text{eff}} = x - p\theta_{\text{sum}}, \quad (13)$$

where θ_{sum} is the sum of the PM and MM hybridization fractions due to specific binding only, and is obtained by solving equation (9). In figure 6 is plotted the corresponding effective binding strength $K^{\text{PM}}_{x_{\text{eff}}}$ against the true binding strength K^{PM}_x . One sees that, below a certain binding strength

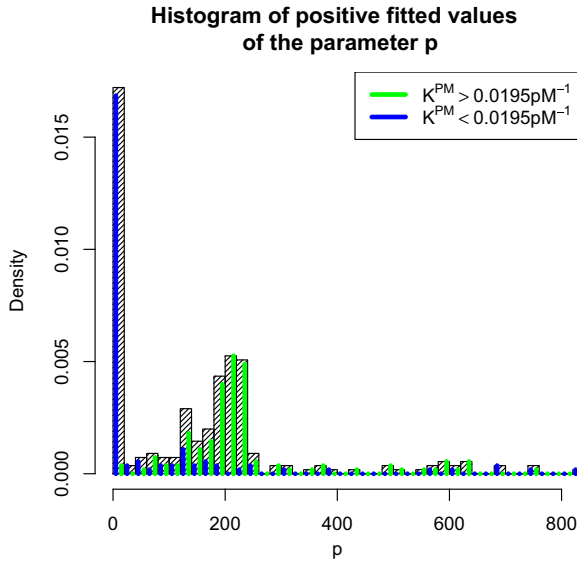


Figure 5. Histogram of the fitted value of the parameter p (pM) for all of the 276 fits with physically meaningful parameter values (hatched bars). Also shown are histograms of the subsets corresponding to high and low values of the parameter K^{PM} . The low- K^{PM} probes correspond to data lying to the left of the vertical dotted line in figure 6 and are an approximation to the set of probes for which depletion data can also be fitted to a no-depletion model with an effective equilibrium constant given by equation (14).

indicated by the horizontal dotted line $K^{\text{PM}}x_{\text{eff}} = 1$, the true concentration is reduced by a factor which is approximately constant over a range of x . In fact, from equation (9) we have that, for $K^{\text{PM}}x_{\text{eff}} \ll 1$, i.e. the linear, low-concentration part of the isotherm, $\theta_{\text{sum}} \approx (K^{\text{PM}} + K^{\text{MM}})x_{\text{eff}}$, from which it follows using equation (13) that $x_{\text{eff}} \approx x/[1 + (K^{\text{PM}} + K^{\text{MM}})p]$. It follows that any probe whose data points lie within this range will be fitted equally well by a hyperbolic, no-depletion, isotherm $I^P = A^P + B^P K_{\text{eff},x}^P / (1 + K_{\text{eff},x}^P)$, with an underestimated equilibrium constant:

$$K_{\text{eff}}^P = \frac{K^P}{1 + (K^{\text{PM}} + K^{\text{MM}})p}, \quad P = \text{PM}, \text{MM}. \quad (14)$$

In figure 5 we have partitioned the fitted values of p into those matching with high and low fitted values of the equilibrium constant, $K^{\text{PM}} \gtrless (\frac{10}{512} \approx 0.0195) \text{pM}^{-1}$, respectively. The cutoff is chosen as a simple way to separate out an approximate set of probe pairs satisfying the conditions leading to the result of equation (14). Recall that the spike-in concentrations in the U133 experiment are bounded above by 512 pM, so for the low- K^{PM} probes $\log_{10} K^{\text{PM}}x < 1$. That is, the fitted isotherms of these probes are determined solely from data lying to the left of the vertical dotted line in figure 6 for which the curves relating $\log x$ to $\log x_{\text{eff}}$ are approximately straight. Returning to figure 5, one observes that the high-equilibrium-constant isotherms, $K^{\text{PM}} > 0.195 \text{pM}^{-1}$, fit predominantly to the depletion model with p consistently around $p = 200 \text{pM}$, and the low-equilibrium-constant isotherms fit predominantly to the no-depletion model with, we infer, the fitted parameter K^P underestimated according to equation (14). A rough estimate

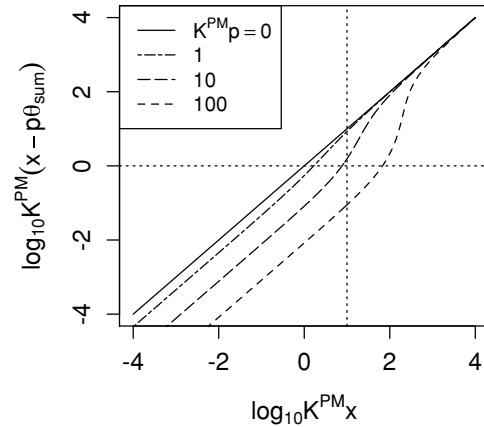


Figure 6. The relationship between the effective binding strength $K^{\text{PM}}x_{\text{eff}}$ and true binding strength $K^{\text{PM}}x$ for a range of effective probe concentrations. The curves are calculated with the help of equation (9), assuming $K^{\text{PM}}/K^{\text{MM}} = 5$, though in practice the shape of the curves is not very sensitive to this ratio. The horizontal dotted line is the upper limit of binding strengths for which depletion data can also be fitted to a no-depletion model with an effective equilibrium constant given by equation (14). The vertical dotted line is the right-hand limit of binding strengths determining the set of low- K^{PM} probes in figure 5.

Histogram of P-values indicating significance of the extra parameter p

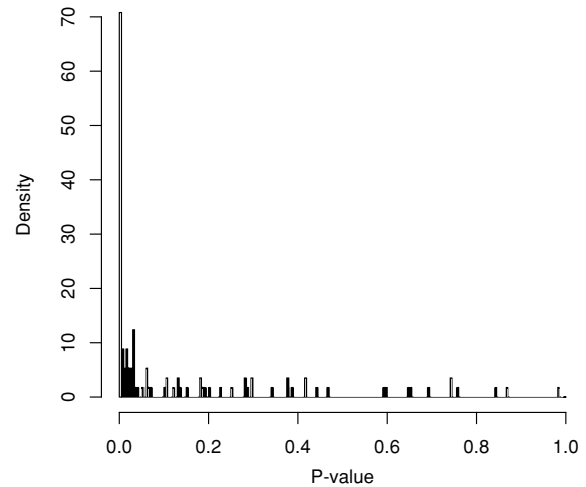


Figure 7. Histogram of the fitted P-values under the null hypothesis of no target depletion ($p = 0$) tested against the alternate hypothesis of global target depletion ($p > 0$) for each of those probe pairs which admit physically meaningful fits to both models. Just over 60% of cases fall within the 5% confidence interval (P-values < 0.05), favouring the alternative hypothesis.

of the lower limit of the underestimation factor, assuming $K^{\text{MM}} \ll K^{\text{PM}}$, is $(1 + 0.0195 \times 200)^{-1} \approx 0.2$.

For the subset of probe pairs which admit physically meaningful fits to both the alternate hypothesis model with depletion, and to the null hypothesis model without depletion, and for which the fitted value of p is strictly positive, we calculated one-sided P-values under the null hypothesis assumption using the analysis appropriate to generalized models [24] described in detail in [14]. The histogram of these P-values, figure 7, shows that they are heavily bunched

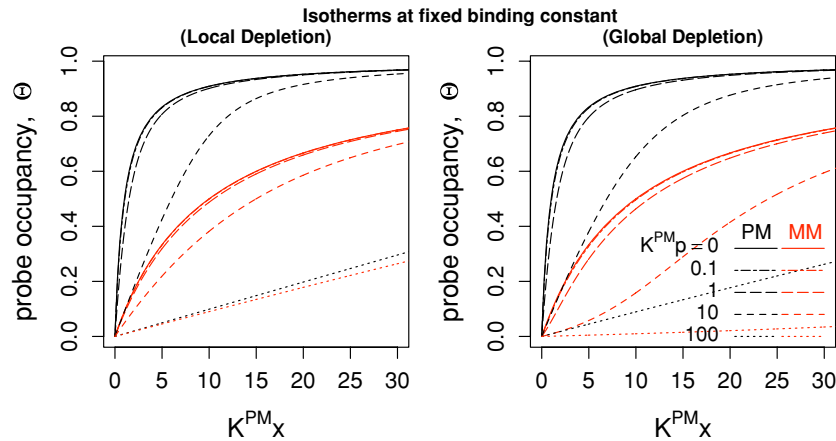


Figure 8. Theoretical isotherms for PM and MM probes derived from local (left) and global (right) depletion models. The isotherms are scaled to dimensionless units $K^{\text{PM}}x$, $\Theta^P = (I^P - A^P)/B^P$ for $s = 10$ and various values of the dimensionless depletion factor $K^{\text{PM}}p$. As explained in the text, in the global PM/MM model, the MM isotherms have an inflection point for $K^{\text{PM}}p > (s - 1)^{-1}$, whereas the PM isotherms do not have an inflection point for any value of this parameter. Isotherms from the local model have no inflection point for any parameter values. Note that these isotherms are plotted at fixed values of binding constant K^{PM} , whereas the isotherms in figures 1 and 2 are plotted at fixed values of specific target concentration x , and consequently have different asymptotic properties as $K^{\text{PM}}x \rightarrow \infty$.

to the left: depletion is confirmed at the 5% confidence level for just over 60% of those cases for which the comparison could be made.

3.3. Shape of the isotherms

It is interesting to examine the shape of the isotherm fits in the global PM/MM depletion model to see how they differ from the well-known hyperbolic Langmuir form of the model without depletion, and from the isotherms of the local depletion model. It is convenient to define dimensionless quantities

$$\Theta^P = \frac{I^P(x) - A^P}{B^P}, \quad s = \frac{K^{\text{PM}}}{K^{\text{MM}}}. \quad (15)$$

On physical grounds we expect $s > 1$, which is observed in general in fits of spike-in data to models with and without depletion. Equations (8) and (9) become

$$\begin{aligned} \Theta^{\text{PM}} &= \frac{K^{\text{PM}}(x - p\theta_{\text{sum}})}{1 + K^{\text{PM}}(x - p\theta_{\text{sum}})}, \\ \Theta^{\text{MM}} &= \frac{K^{\text{PM}}(x - p\theta_{\text{sum}})}{s + K^{\text{PM}}(x - p\theta_{\text{sum}})}, \end{aligned} \quad (16)$$

with θ_{sum} the solution to

$$\theta_{\text{sum}} = \frac{K^{\text{PM}}(x - p\theta_{\text{sum}})}{1 + K^{\text{PM}}(x - p\theta_{\text{sum}})} + \frac{K^{\text{PM}}(x - p\theta_{\text{sum}})}{s + K^{\text{PM}}(x - p\theta_{\text{sum}})}. \quad (17)$$

Plots of Θ^P as a function of the dimensionless $K^{\text{PM}}x$ for the realistic value $s = 10$ and a range of values of the dimensionless depletion parameter $K^{\text{PM}}p$ are shown in the right panel of figure 8. Also shown for comparison (left panel) are the equivalent isotherms from the local depletion model, for which θ_{sum} in equation (16) is replaced by Θ^{PM} or Θ^{MM} , respectively. The effect of depletion is to depress the response function at small specific target concentrations, as the available effective specific target concentration is

effectively decreased. For the case of the PM/MM global depletion model, we show in Appendix B that for $K^{\text{PM}}p > (s - 1)^{-1}$, and provided $s > 1$, the MM response curve acquires an inflection point, while the PM curve flattens without forming an inflection point. Physically, the effect of depletion on the MM response is more pronounced as the PM probes more strongly deplete the available target in solution. This behaviour is clearly evident in fits to the U133 spike-in data (see figure 4 and the supplementary material available at stacks.iop.org/PhysBio/7/016004/mmedia). A straightforward calculation shows that isotherms from the local depletion model, on the other hand, do not have an inflection point for either PM or MM probes for any parameter values.

3.4. Shape of the hook curve

Theoretical hook curves assuming either a local or global depletion model and parameter values typical of fits to the U133 latin square data set and a range of the probe density parameter p are shown in figure 9. For these curves the trajectory is that of the pair (Σ, Δ) defined by equation (3) traced out as the specific binding strength $X_S^{\text{PM}} = K_S^{\text{PM}}x_S$ varies over a range of specific spike-in concentrations x_S at fixed values of all other parameters in the model. For the case of global depletion the hook coordinates are calculated from the post-washing coverage fractions equation (A.49) with θ_{sum} given by equation (A.47). For the case of local depletion θ_{sum} is replaced by θ_S^{PM} or θ_S^{MM} , respectively, defined by equation (A.31).

One sees that the effect of local depletion is to flatten the peak and introduce an asymmetry in the hook curve. The flattening is caused by a decrease in the difference between the PM and MM responses as more specific target is extracted from solution in the vicinity of the PM probe feature. Global depletion, on the other hand, has no effect on the shape or end points of the hook curve as it effects an identical reparameterization $x_S \rightarrow x_S - p\theta_{\text{sum}}$ in the formulae for

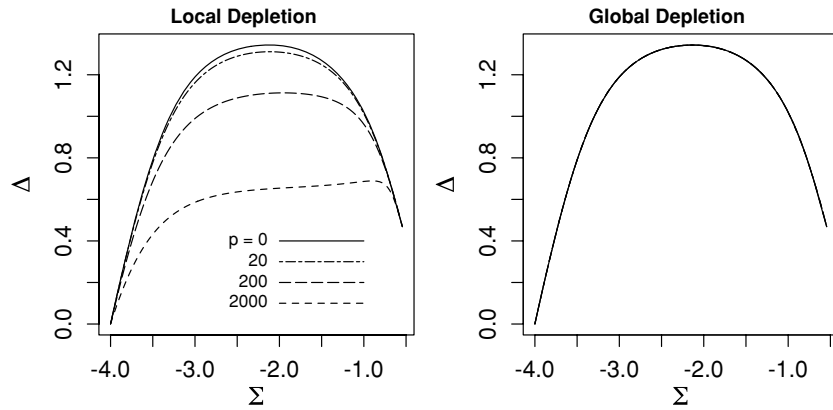


Figure 9. Theoretical hook curves determined from isotherms of a PM/MM pair assuming local depletion (left) and global depletion (right) for a range of the effective probe concentration parameter p (pM). The following parameter values, typical of fits to the U133 latin square spike-in data set, were used: $X_N^{\text{PM}} = X_N^{\text{MM}} = 10^{-3}$, $K_S^{\text{PM}} = 5 \times 10^{-3} \text{pM}^{-1}$, $K_S^{\text{MM}} = 5 \times 10^{-4} \text{pM}^{-1}$, and washing survival fractions $w_N^{\text{PM}} = w_N^{\text{MM}} = 0.1$, $w_S^{\text{PM}} = 0.5$, $w_S^{\text{MM}} = 0.2$. Note that for global depletion the shape of the hook curve is independent of p .

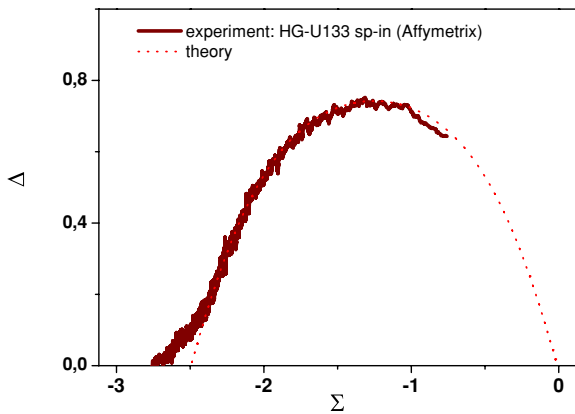


Figure 10. Experimental hook curve of one array of the U133 latin square data set and a fit using assuming the hyperbolic Langmuir response function without depletion. Note the symmetric shape of the experimental hook curve, compatible with global depletion. The deviation between the experimental and theoretic curve at small Σ is caused by non-specific hybridization not discussed here (see [8, 11]).

both θ^{PM} and θ^{MM} . However, as p is increased, internal points corresponding to a given probe-pair value of the binding strength migrate progressively to the left along the curve, reflecting a decrease in both the PM and MM fluorescence intensities.

A typical hook curve from one of the arrays of the U133 latin square data set using the algorithm in [4] is shown in figure 10. This algorithm includes a moving average over ~ 100 probesets and correction of raw intensities for probe binding affinities using position- and nucleotide-dependent sensitivity profiles [11]. Hook curves have been similarly evaluated for a number of experimental data sets relating to Affymetrix GeneChips in [8], including the latin square spike-in experiments, with the result that no evidence for an asymmetric hook curve has yet been observed. We conclude, within this particular set of data sets corresponding to chip designs with neighbouring PM/MM pairs, that target depletion, if significant, fits the global model rather than the local model.

3.5. Correction of expression estimates

Estimates of expression levels using algorithms such as the hook method [8] and the inverse Langmuir method [25] have to date ignored target depletion and therefore been based on the assumption of a hyperbolic Langmuir isotherm. In the previous section we have seen that this is equivalent to underestimating the true specific target concentration x_S by a shift $x_S \rightarrow x_{\text{eff}} = x_S - p\theta_{\text{sum}}$, where θ_{sum} is the sum of the PM and MM hybridization fractions due to specific binding only. θ_{sum} can be calculated from the observed total coverage fractions θ^{PM} and θ^{MM} , which include both specific and non-specific binding, as follows. From equation (A.47),

$$\theta_{\text{sum}} = \frac{K_S^{\text{PM}} x_{\text{eff}}}{1 + X_N + K_S^{\text{PM}} x_{\text{eff}}} + \frac{K_S^{\text{MM}} x_{\text{eff}}}{1 + X_N + K_S^{\text{MM}} x_{\text{eff}}}, \quad (18)$$

where the nonspecific strength $X_N = X_N^{\text{PM}} \approx X_N^{\text{MM}}$ is assumed to be common for all probes on the microarray after correction for binding affinities via sensitivity profiles. X_N can be measured from the width of the hook curve and is typically of order 10^{-3} . From equation (A.48),

$$\theta^P = \frac{X_N + K_S^P x_{\text{eff}}}{1 + X_N + K_S^P x_{\text{eff}}}, \quad P = \text{PM, MM}, \quad (19)$$

which rearranges to give $K_S^P x_{\text{eff}} = \theta^P / (1 - \theta^P) - X_N$. Substituting back into equation (18) then gives $\theta_{\text{sum}} = (1 + X_N)(\theta^{\text{PM}} + \theta^{\text{MM}}) - 2X_N$. Thus, the true specific target concentration is given in terms of the effective, depleted target concentration by

$$x_S = x_{\text{eff}} + p[(1 + X_N)(\theta^{\text{PM}} + \theta^{\text{MM}}) - 2X_N]. \quad (20)$$

In principle, this formula gives the correction for target depletion over the entire range of target concentrations, including an interpolation between the two regimes illustrated in figure 6. Note that x_{eff} , X_N and the coverages θ^P can be estimated by established methods such as the hook curve or inverse Langmuir method. Equation (20) then requires knowledge of the probe concentration p , which, for example, is expected to depend on the chip type. Its estimation requires further efforts which will be the subject of future investigations.

4. Critical evaluation of an alternate hybridization model

Recently an alternate competitive hybridization model incorporating target depletion by Li *et al* [23] has appeared in the literature. This model is applied to the Affymetrix U133 data set and is purported to be capable of predicting signal intensities of individual probes and of achieving quantification of absolute target concentrations from microarray fluorescence intensity data. Here we point out a number of errors in the basic assumptions of Li *et al*'s model and argue that it does not represent any advance over previously existing hybridization models.

Of particular interest to Li *et al* is the asymptotic behaviour of fluorescence intensities for individual probes in the limit of saturation concentrations of specific target. Standard reaction kinetic models applied to the hybridization step of the Affymetrix protocol implies that in the high specific target concentration limit, all probes should saturate at the same observed fluorescence intensity, regardless of the nucleotide probe sequence or resulting probe-target binding free energy. For either of the models in Appendix A, for instance, we have $\lim_{x_s \rightarrow \infty} \theta = 1$, where the limit is taken with other variables being held constant. This is at variance with observations from spike-in experiments, for which the PM element of a PM/MM almost invariably saturates at a higher intensity than its MM partner.

An acceptable explanation, which has been demonstrated to fit well the saturation behaviour to both the U95a and U133 Affymetrix spike-in experiments [15, 21], is to explain the differing asymptotes via the post-hybridization washing step, which not only removes unbound targets, but also dissociates both specific and non-specific bound targets (see equation (A.49)). For reasons which are not clear, but which appear to be based on a misinterpretation of Skvortsov *et al*'s experimental results [29], Li *et al* reject the washing hypothesis. Instead, they proceed to develop their own thermodynamic model, which is not consistent with accepted principles of physical chemistry, but which nevertheless predicts response functions with binding free energy-dependent asymptotes resulting from the hybridization step alone. In their model, the washing step is assumed to have little effect on specific targets bound to probes.

In Appendix C we explain in detail a fundamental error in their application of the law of mass action to hybridization at the microarray surface, and show that when the error is corrected, their model essentially agrees with existing treatments inspired by Langmuir adsorption theory, together with the depletion extension of Ono *et al* [27]. We also note that their derived formula for the coverage fraction of specific targets is demonstrably wrong in that it disagrees with the results of the Affymetrix latin square spike-in experiments without complex background. Lastly, the algorithm proposed by Li *et al* for inferring absolute specific target concentrations requires subtraction of the intensity at zero spike-in concentration as a way of dealing with non-specific hybridization (see equations (22) and (27) of [23]). This value is of course unknown in any biomedical application

of microarrays, and it is the problem of calibrating a correction for non-specific hybridization which is the subject of much current activity in physico-chemical modelling of microarrays (see [6, 22] for instance). That Li *et al* are able to produce estimates of spike-in concentrations at the higher end of the scale (>1 pM) by cross validation from a crude four-parameter formula based on incorrect physical assumptions is not surprising and is not an improvement on any existing expression measure.

5. Conclusions and outlook

The physico-chemical models of equilibrium microarray hybridization described here involve microarray probes, specific and non-specific targets and their interactions on the chip surface. As well as probe-target hybridization, bulk hybridization and probe and target folding, the important innovation is a careful consideration of depletion of target molecules from the supernatant solution by hybridization of specific targets. Consideration of target depletion is important when the target concentration is comparable with or less than the effective probe molecule concentration, which we determine to be of the order of 200 pM for HG133 generation Affymetrix GeneChips. If the sensitivity of microarrays is to be pushed to lower specific target concentrations, a proper understanding of and appropriate correction for this phenomenon is important.

Two possible scenarios are considered, local and global depletion. In the first scenario, studied by Ono *et al* [27], depletion by hybridization to a given probe feature only affects that particular feature. This scenario is relevant when probe features addressing the same target species are physically separated on the microarray, and the rate of diffusion or convection over the distance between features is small compared with the rate of hybridization. The second scenario, global depletion, has not been considered previously. In this scenario some or all of the features addressing a given target species are effected. This is relevant, for instance, for chip designs which include mismatch features located in close proximity on the microarray surface to their perfect match partners.

We analysed data obtained in two experimental situations: firstly, the intensity response of PM and MM probes of varying probe length at fixed target concentration (the Suzuki *et al* data set), and secondly, the intensity response of PM and MM probes of fixed length at varying target concentration (the Affymetrix latin square data set). The PM/MM design of the chips allows for a combined analysis of both probe types via the 'hook plot', the shape of which gives a clear discrimination between local and global depletion.

We have confirmed conclusively using the hook curve analysis that the spike-in data set of Suzuki *et al* [30] is an example of local and not global depletion. A six-parameter fit of the local depletion model verifies the earlier analysis of Ono *et al* [27]. The hook curve analysis has proved particularly useful for this type of analysis because of the marked qualitative difference in the behaviour of these plots between the two possible scenarios.

Previous attempts to fit a hyperbolic Langmuir isotherm model to the second data set, the Affymetrix U133 latin square spike-in, had only met with partial success [13]. In our current reanalysis of this data set we have had markedly improved success using the global model of target depletion, which we believe is relevant because of close proximity of partner PM and MM probe features. The global depletion model provides a significantly improved fit to a large portion of these data, namely that portion for which the effective equilibrium constant of the hybridization reaction is above a certain threshold value. Importantly, we have demonstrated that if the effective equilibrium constant K is below the inverse of the range of concentrations of a spike-in experiment, the ability to detect target depletion through response curve fits is masked and the data may mistakenly be fitted to the linear part of a non-depleted hyperbolic Langmuir isotherm with an underestimated equilibrium constant given by equation (14).

For the Affymetrix spike-in data our depletion model is also able to explain certain qualitatively observed phenomena related to the shape of the isotherms. The MM response function typically has an inflection point at low concentrations which may serve as a signal for global depletion in spike-in experiments, whereas the PM response function is typically flattened but does not form any such inflection point. Another characteristic of global depletion is the shape of the hook curve which continues to be symmetric as the effective concentration of free specific targets is reduced by hybridization. Local depletion, on the other hand, is predicted to entail an antisymmetric hook curve.

In the final section we have given a critique pointing out a number of serious errors in a competing physico-chemical model dealing with target depletion in microarray hybridization experiments by Li *et al* [23]. After correction of these errors one gets a solution which, in the limit of no depletion, is the well-established and accepted Langmuir model. With depletion included it is a simplified version of our local depletion model or the model of Ono *et al* [27].

The observations made herein, particularly those for the Affymetrix U133 data set, have consequences for existing physico-chemistry-based algorithms and methods for microarray calibration. By calibration we mean obtaining estimates of transcript abundance, ideally as an absolute concentration or, at the very least, relative measures which are related linearly to transcript abundance. It must include not only systematic correction for the effects such as non-specific background, saturation and sequence-specific binding affinities of probes [12], but also, as we have shown, depletion of targets from the supernatant solution.

Physico-chemical calibration algorithms rely directly or indirectly on obtaining estimates of the effective equilibrium constant K from probe sequences via position-dependent affinities [9, 10] or via free binding energies ΔG calculated from nearest neighbour stacking models [17, 18]. To date they have assumed a hyperbolic Langmuir isotherm and involve fits to spike-in data sets including the Affymetrix HGU133 data set. We have shown here that estimates of K from this data set are compromised in a predictable way by target depletion if a hyperbolic isotherm is assumed. It is consequently not

surprising that attempts to find a clear and unambiguous relationship between K obtained in this way and ΔG have met with limited success (see section 5.2 and 5.3 of [13]). Clearly more work has to be done in correcting this aspect of calibration algorithms to take into account target depletion. Finally, irrespective of whether calibration algorithms rely on inverting a theoretical isotherm [25] or first extracting an effective binding strength $X_{\text{eff}} = K x_{\text{eff}}$ from, say, the hook curve [12], a solution must be found to the problem of extracting the true target concentration x from the microarray-depleted concentration x_{eff} . In section 3.5 we show that the information required to do this is, in principle, inherent in the measured fluorescence intensities via equation (20). A practical implementation of this will be the subject of future work.

Acknowledgments

This work was supported by an Australian Research Council Discovery Project Grant (DP0987298), an Australian Academy of Science Scientific Visits to Europe Grant and by Deutsche Forschungsgemeinschaft (BIZ6-06).

Appendix A. Physico-chemical model

In the physico-chemical model presented below the equilibrium coverage fraction θ ($0 \leq \theta \leq 1$) of fluorescent dye carrying target molecules bound to a given probe feature at the microarray surface at the end of the hybridization step is calculated assuming standard equilibrium physical chemistry. The model differs from previous models considered by the current authors [5, 13] in that the Ono model [27] of target depletion from the supernatant solution by hybridization to the array is included. Two regimes are considered: the first of these, local depletion, in which depletion by a given probe feature only affects that particular probe, is a slight variant of the ‘finite hybridization model’ including competitive specific and non-specific hybridization presented by Ono *et al* [27]. It differs from the Ono model in that all chemical reactions, namely folding, bulk hybridization and surface hybridization, are integrated *ab initio*, leading to slightly different formulae for the final coverage fraction. A detailed derivation of local depletion is included here for completeness and to establish the notation and a framework for the second regime, global depletion, in which all features responding to a given target species are affected.

A.1. Local depletion

In this case there is assumed to be no interaction between different probe features. The set of chemical species considered is set out in table A1. For a given probe feature, the input parameters to the model are (1) the total specific target concentration,

$$x_S = [S] + [S'] + [P \cdot S] + [S \cdot N] + [S \cdot S], \quad (\text{A.1})$$

(2) an effective total non-specific target concentration, assumed to be common to all probe features:

$$x_N = [N] + [N'] + [P \cdot N] + [S \cdot N] + [N \cdot N], \quad (\text{A.2})$$

Table A1. Chemical species present in the model.

	Unfolded	Folded
Specific target in the solution	S	S'
Non-specific effective target in solution	N	N'
Probe at the surface (not bound to the target)	P	P'
Duplexes in the solution	$S \cdot S, S \cdot N, N \cdot N$	
Duplexes at the microarray surface	$P \cdot S, P \cdot N$	

(3) an effective probe concentration for the feature:

$$p = [P] + [P'] + [P \cdot S] + [P \cdot N] \quad (\text{A.3})$$

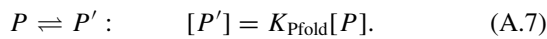
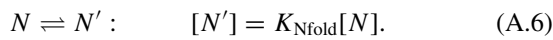
and (4) a set of equilibrium constants K_r , where $r \in \{\text{Sfold}, \text{Nfold}, \dots\}$, for the reactions (A.5)–(A.12) set out below. Following the usual convention square brackets indicate the molar concentration of a chemical species.

Our aim is to determine the total coverage fraction

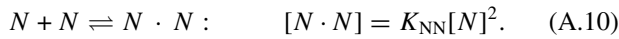
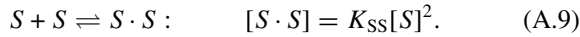
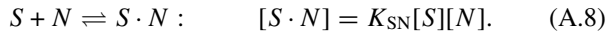
$$\theta = \theta_S + \theta_N = \frac{[P \cdot S]}{p} + \frac{[P \cdot N]}{p} \quad (\text{A.4})$$

of both specific and non-specific duplexes resulting from the following chemical reactions:

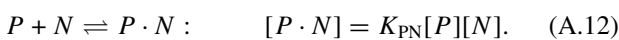
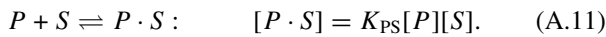
Folding



Bulk hybridization



Surface hybridization



We begin by using equations (A.5)–(A.12) to eliminate concentrations of folded species and of most duplex species from equations (A.1) to (A.4). From equations (A.3) and (A.4) we obtain

$$\theta_S = \frac{K_{\text{PS}}[S]}{(1 + K_{\text{Pfold}}) + K_{\text{PS}}[S] + K_{\text{PN}}[N]}, \quad (\text{A.13})$$

$$\theta_N = \frac{K_{\text{PN}}[N]}{(1 + K_{\text{Pfold}}) + K_{\text{PS}}[S] + K_{\text{PN}}[N]}, \quad (\text{A.14})$$

and from equations (A.1) and (A.2)

$$x_S = (1 + K_{\text{Sfold}} + K_{\text{SN}}[N])[S] + K_{\text{SS}}[S]^2 + [P \cdot S], \quad (\text{A.15})$$

$$x_N = (1 + K_{\text{Nfold}} + K_{\text{SN}}[S])[N] + K_{\text{NN}}[N]^2 + [P \cdot N]. \quad (\text{A.16})$$

Following [6] we make the reasonable assumptions.

- (i) $K_{\text{SS}}[S] \ll 1$: specific targets will not easily encounter each other in bulk solution;

- (ii) $K_{\text{SN}}[S] \ll 1$: very little of the depletion of nonspecific targets by bulk hybridization is due to encounters with the specific targets in question;

- (iii) $[P \cdot N] \ll x_N$: the proportion of nonspecific background depleted by hybridization to the microarray is negligible.

With these assumptions, the above equations reduce to

$$x_S = (1 + K_{\text{Sfold}} + K_{\text{SN}}[N])[S] + [P \cdot S], \quad (\text{A.17})$$

$$x_N = (1 + K_{\text{Nfold}})[N] + K_{\text{NN}}[N]^2. \quad (\text{A.18})$$

Equation (A.18) is quadratic in $[N]$ whose solution we will write as

$$[N] = f_N(x_N, K_{\text{NN}}, K_{\text{Nfold}}). \quad (\text{A.19})$$

Previously (equation (2.8) of [5] and equation (1) of [6]) the following approximation

$$[N] \approx \frac{x_N}{1 + K_{\text{Nfold}} + K_{\text{NN}}x_N} \quad (\text{A.20})$$

has been used, though this approximation is not necessary in the current context and is only included for comparison with previous work. We also have

$$[S] = \frac{x_S - [P \cdot S]}{1 + K_{\text{Sfold}} + K_{\text{SN}}f_N(x_N, K_{\text{NN}}, K_{\text{Nfold}})} \quad (\text{A.21})$$

$$\approx \frac{x_S - [P \cdot S]}{1 + K_{\text{Sfold}} + K_{\text{SN}}x_N}, \quad (\text{A.22})$$

once again employing the same approximation.

Substituting back into equations (A.13) and (A.14) gives

$$\theta_S = \frac{K_S(x_S - [P \cdot S])}{1 + X_N + K_S(x_S - [P \cdot S])}, \quad (\text{A.23})$$

$$\theta_N = \frac{X_N}{1 + X_N + K_S(x_S - [P \cdot S])}, \quad (\text{A.24})$$

where

$$X_N = \frac{K_{\text{PN}}}{1 + K_{\text{Pfold}}} f_N(x_N, K_{\text{NN}}, K_{\text{Nfold}}) \quad (\text{A.25})$$

$$\approx \frac{K_{\text{PN}}x_N}{(1 + K_{\text{Pfold}})(1 + K_{\text{Nfold}} + K_{\text{NN}}x_N)}, \quad (\text{A.26})$$

and

$$K_S = \frac{K_{\text{PS}}}{(1 + K_{\text{Pfold}})(1 + K_{\text{Sfold}} + K_{\text{SN}}f_N(x_N, K_{\text{NN}}, K_{\text{Nfold}}))} \quad (\text{A.27})$$

$$\approx \frac{K_{\text{PS}}}{(1 + K_{\text{Pfold}})(1 + K_{\text{Sfold}} + K_{\text{SN}}x_N)}. \quad (\text{A.28})$$

Finally, using equation (A.4) gives

$$\theta_S = \frac{K_S(x_S - p\theta_S)}{1 + X_N + K_S(x_S - p\theta_S)}, \quad (\text{A.29})$$

$$\theta_N = \frac{X_N}{1 + X_N + K_S(x_S - p\theta_S)}. \quad (\text{A.30})$$

The quantity X_N is known as the non-specific binding strength, and in the approximation of equation (A.26) is often written

in the form $X_N = K_N x_N$ where K_N is an effective equilibrium constant for non-specific binding. It is also common to define a specific binding strength $X_S = K_S x_S$ in terms of the effective specific equilibrium constant K_S and specific target concentration.

For given x_S, x_N, p and equilibrium constants K_r , equation (A.29) is quadratic in θ_S with a unique solution in $[0, 1]$, namely

$$\theta_S = \frac{1}{2} \left[\frac{1 + X_N}{K_S p} + 1 + \frac{x_S}{p} - \sqrt{\left(\frac{1 + X_N}{K_S p} + 1 + \frac{x_S}{p} \right)^2 - 4 \frac{x_S}{p}} \right]. \quad (\text{A.31})$$

The required result is then

$$\theta = \theta_S + \theta_N = \frac{X_N + K_S (x_S - p\theta_S)}{1 + X_N + K_S (x_S - p\theta_S)}. \quad (\text{A.32})$$

If post-hybridization washing is significant, it is introduced into the model via specific and non-specific survival factors w_S and w_N , where $1 > w_S > w_N > 0$, giving

$$\theta_{\text{after.wash}} = w_S \theta_S + w_N \theta_N = \frac{w_S X_N + w_N K_S (x_S - p\theta_S)}{1 + X_N + K_S (x_S - p\theta_S)}. \quad (\text{A.33})$$

A.2. Global depletion

In the case of global depletion the target concentration specific to a given feature is assumed to be depleted by the hybridization to all features which target the same chemical species. Below we consider the case of a PM/MM pair of probe features, though the analysis readily generalizes to any number of features addressing the same specific species. We use superscripts PM and MM to indicate probe molecules on respective elements of a PM/MM pair, and denote by S the target species complementary to the PM probe. With these changes the set of input parameters become (1) the total specific target concentration

$$x_S = [S] + [S'] + [P^{\text{PM}} \cdot S] + [P^{\text{MM}} \cdot S] + [S \cdot N] + [S \cdot S], \quad (\text{A.34})$$

(2) an effective total non-specific target concentration

$$x_N = [N] + [N'] + [P^{\text{PM}} \cdot N] + [P^{\text{MM}} \cdot N] + [S \cdot N] + [N \cdot N], \quad (\text{A.35})$$

(3) an effective probe concentration, assumed to be the same for PM and MM,

$$p = [P^{\text{PM}}] + [P^{\text{PM}'}] + [P^{\text{PM}} \cdot S] + [P^{\text{PM}} \cdot N] = [P^{\text{MM}}] + [P^{\text{MM}'}] + [P^{\text{MM}} \cdot S] + [P^{\text{MM}} \cdot N] \quad (\text{A.36})$$

and a set of equilibrium constants K_r^P , which may or may not depend on $P = \text{PM, MM}$, depending on the reaction r . Our aim is now to determine a coverage fraction

$$\theta^P = \theta_S^P + \theta_N^P = \frac{[P^P \cdot S]}{p} + \frac{[P^P \cdot N]}{p}, \quad P = \text{PM, MM} \quad (\text{A.37})$$

for both elements of a probe pair.

Analogous to equations (A.13) and (A.14) we have

$$\theta_S^P = \frac{K_{\text{PS}}^P [S]}{(1 + K_{\text{Pfold}}^P) + K_{\text{PS}}^P [S] + K_{\text{PN}}^P [N]}, \quad P = \text{PM, MM}, \quad (\text{A.38})$$

$$\theta_N^P = \frac{K_{\text{PN}}^P [N]}{(1 + K_{\text{Pfold}}^P) + K_{\text{PS}}^P [S] + K_{\text{PN}}^P [N]}, \quad P = \text{PM, MM}. \quad (\text{A.39})$$

After making the ‘reasonable assumptions’ of the previous section, equation (A.17) becomes

$$x_S = (1 + K_{\text{Sfold}} + K_{\text{SN}}[N])[S] + [P^{\text{PM}} \cdot S] + [P^{\text{MM}} \cdot S], \quad (\text{A.40})$$

and equations (A.18) and (A.19) remain unchanged. Then equations (A.23) and (A.24) become

$$\theta_S^P = \frac{K_S^P (x_S - [P^{\text{PM}} \cdot S] - [P^{\text{MM}} \cdot S])}{1 + X_N^P + K_S^P (x_S - [P^{\text{PM}} \cdot S] - [P^{\text{MM}} \cdot S])}, \quad (\text{A.41})$$

$$\theta_N^P = \frac{X_N^P}{1 + X_N^P + K_S^P (x_S - [P^{\text{PM}} \cdot S] - [P^{\text{MM}} \cdot S])}, \quad (\text{A.42})$$

where

$$X_N^P = \frac{K_{\text{PN}}^P}{1 + K_{\text{Pfold}}^P} f_N(x_N, K_{\text{NN}}, K_{\text{Nfold}}), \quad (\text{A.43})$$

and

$$K_S^P = \frac{K_{\text{PS}}^P}{(1 + K_{\text{Pfold}}^P)(1 + K_{\text{Sfold}} + K_{\text{SN}} f_N(x_N, K_{\text{NN}}, K_{\text{Nfold}}))}. \quad (\text{A.44})$$

Using equation (A.37) then gives

$$\theta_S^P = \frac{K_S^P [x_S - p(\theta_S^{\text{PM}} + \theta_S^{\text{MM}})]}{1 + X_N^P + K_S^P [x_S - p(\theta_S^{\text{PM}} + \theta_S^{\text{MM}})]}, \quad (\text{A.45})$$

$$\theta_N^P = \frac{X_N^P}{1 + X_N^P + K_S^P [x_S - p(\theta_S^{\text{PM}} + \theta_S^{\text{MM}})]}. \quad (\text{A.46})$$

Summing equation (A.45) over P and defining $\theta_{\text{sum}} = \theta_S^{\text{PM}} + \theta_S^{\text{MM}}$ gives

$$\theta_{\text{sum}} = \sum_{P=\text{PM,MM}} \frac{K_S^P (x_S - p\theta_{\text{sum}})}{1 + X_N^P + K_S^P (x_S - p\theta_{\text{sum}})}. \quad (\text{A.47})$$

This equation is cubic in θ_{sum} , and can easily be solved numerically as a function of x_S, K_S^P, X_N^P and p using a Newton–Raphson algorithm. The required coverage function is then

$$\theta^P = \frac{X_N^P + K_S^P (x_S - p\theta_{\text{sum}})}{1 + X_N^P + K_S^P (x_S - p\theta_{\text{sum}})}, \quad P = \text{PM, MM}. \quad (\text{A.48})$$

Again post-hybridization can be introduced into the model via specific and non-specific survival factors, giving

$$\theta_{\text{after.wash}}^P = \frac{w_S^P X_N^P + w_N^P K_S^P (x_S - p\theta_{\text{sum}})}{1 + X_N^P + K_S^P (x_S - p\theta_{\text{sum}})}, \quad (\text{A.49})$$

$P = \text{PM, MM}$.

Appendix B. Analysis of the shape of the isotherms

We demonstrate that in the global PM/MM depletion model with $s = K^{\text{PM}}/K^{\text{MM}} > 1$ considered in section 3.3, the MM response curve acquires an inflection point for sufficiently high values of K^{PM} , while the PM response curve flattens without forming an inflection point as K^{PM} increases.

Defining $\phi(x) = K^{\text{PM}}(x - p\theta_{\text{sum}})$, equation (16) gives $\Theta^{\text{PM}} = \phi/(1 + \phi)$ and $\Theta^{\text{MM}} = \phi/(s + \phi)$, and thus

$$\begin{aligned} \frac{d^2\Theta^{\text{PM}}}{dx^2} &= \frac{\phi''}{(1 + \phi)^2} - 2\frac{(\phi')^2}{(1 + \phi)^3}, \\ \frac{d^2\Theta^{\text{MM}}}{dx^2} &= \frac{s\phi''}{(s + \phi)^2} - 2s\frac{(\phi')^2}{(s + \phi)^3}, \end{aligned} \quad (\text{B.1})$$

while differentiating equation (17) twice gives

$$-\frac{\phi''}{pK^{\text{PM}}} = \frac{\phi''}{(1 + \phi)^2} - 2\frac{(\phi')^2}{(1 + \phi)^3} + \frac{s\phi''}{(s + \phi)^2} - 2s\frac{(\phi')^2}{(s + \phi)^3}. \quad (\text{B.2})$$

One easily checks that $\phi(0) = 0$, and thus equation (B.2) implies

$$\phi''(0) = \frac{2(1 + s^2)}{s(1 + s + s/(pK^{\text{PM}}))} \phi'(0)^2. \quad (\text{B.3})$$

Substituting back into equation (B.1) at $x = 0$ gives

$$\begin{aligned} \left. \frac{d^2\Theta^{\text{PM}}}{dx^2} \right|_{x=0} &= \phi''(0) - 2\phi'(0)^2 \\ &= 2 \left(\frac{1 + s^2}{s(1 + s + s/(pK^{\text{PM}}))} - 1 \right) \phi'(0)^2 \\ &< 2 \left(\frac{1 + s^2}{s(1 + s)} - 1 \right) \phi'(0)^2 \\ &= 2 \frac{1 - s}{s(1 + s)} \phi'(0)^2 < 0, \end{aligned} \quad (\text{B.4})$$

for $s > 1$. That is, the PM response curve is concave downwards at the origin for all physically relevant values of s . In fact $d^2\Theta^{\text{PM}}/dx^2|_{x=0}$ increases from $-2(K^{\text{PM}})^2$ at $p = 0$ to 0 as $p \rightarrow \infty$ and hence the response curve flattens to an almost straight line.

Similarly we have

$$\begin{aligned} \left. \frac{d^2\Theta^{\text{MM}}}{dx^2} \right|_{x=0} &= \frac{1}{s}\phi''(0) - \frac{2}{s^2}\phi'(0)^2 \\ &= \frac{2}{s^2} \left(\frac{1 + s^2}{1 + s + s/(pK^{\text{PM}})} - 1 \right) \phi'(0)^2, \end{aligned} \quad (\text{B.5})$$

from which it follows that

$$\left. \frac{d^2\Theta^{\text{MM}}}{dx^2} \right|_{x=0} \leq 0 \quad \text{according as} \quad pK^{\text{PM}} \leq \frac{1}{s-1}. \quad (\text{B.6})$$

Thus, the MM response curve has an inflection point for $pK^{\text{PM}} > 1/(s-1)$.

Appendix C. Critique of Li *et al*

We point out errors in the thermodynamic model proposed in a recent paper by Li *et al* [23]. The primary source of error in this paper is an incorrect use of the law of mass action in equation (3) of their paper describing the rate \dot{n}_{in} of binding of specific and non-specific targets to probes. In the notation of Li *et al*, the corrected form of the equation is

$$\frac{\dot{n}_{\text{in}}}{N_A V} = (1 - \alpha - \beta)pk_b([T] + [N]), \quad (\text{C.1})$$

where α and β are specific and non-specific coverage fractions (equivalent to our θ_S and θ_N), p is the effective probe concentration, $[T]$ and $[N]$ are the free specific and non-specific target concentrations, k_b is the reaction rate for binding (assumed to be determined by a rate-determining initiation step and therefore the same for specific and non-specific targets), N_A is Avogadro's number and V volume of the hybridization solution. The factor $([T] + [N])$ is missing from Li *et al*'s paper, either intentionally or through an oversight, but must be present if the reaction proceeds at a rate proportional to the product of the concentrations of each of the reactants.

With this correction, equation (5) of [23] balancing the forward and backward reaction rates becomes

$$(1 - \alpha - \beta)pk_b([T] + [N]) = \alpha pk_d + \beta pk_n, \quad (\text{C.2})$$

where k_d and k_n are dissociation rate constants for specific and non-specific duplexes, respectively. Equation (8) of [23] is best derived by balancing the forward and backward rates for specific and nonspecific targets separately:

$$\begin{aligned} (1 - \alpha - \beta)pk_b[T] &= \frac{\dot{n}_{\text{in}}^{(T)}}{N_A V} = \frac{\dot{n}_{\text{out}}^{(T)}}{N_A V} = \alpha pk_d, \\ (1 - \alpha - \beta)pk_b[N] &= \frac{\dot{n}_{\text{in}}^{(N)}}{N_A V} = \frac{\dot{n}_{\text{out}}^{(N)}}{N_A V} = \alpha pk_n, \end{aligned}$$

giving

$$\beta = \frac{k_d[N]}{k_n[T]}\alpha,$$

in agreement with equation (8) of [23]. In fact this equation cannot be derived without the assumption that the forward reactions are driven at rates proportional to the target concentrations, as used in equation (C.1) above, but not in equation (3) of [23]. Substituting this back into equation (C.1) gives the corrected form of equation (9) of [23]:

$$\begin{aligned} \alpha &= \frac{1}{1 + (k_d/k_n)([N]/[T]) + (k_d/k_b)(1/[T])} \\ &= \frac{K_T[T]}{1 + K_T[T] + K_N[N]}, \end{aligned} \quad (\text{C.3})$$

where we define specific and non-specific equilibrium constants $K_T = k_b/k_d$ and $K_N = k_b/k_n$, respectively. A similar calculation gives the non-specific coverage fraction as

$$\beta = \frac{K_N[N]}{1 + K_T[T] + K_N[N]}. \quad (\text{C.4})$$

Li *et al* incorporate target depletion by hybridization from the supernatant solution by making the substitution $[T] = \hat{T} - \alpha p$, where \hat{T} is the nominal spike-in concentration.

With appropriate changes of notation, the corrected equations (C.3) and (C.4) with this substitution are essentially nothing more than simplified versions of the Ono model [27], or of our local depletion model equations (A.29) and (A.30), without inclusion of probe or target folding or bulk hybridization in the supernatant solution. Li *et al* then proceed to fit their model to the U133 Affymetrix latin square data set. However, the above substitution corresponds to local, not global, depletion, which we have demonstrated in section 3 is not appropriate for this data set.

Finally we note that Li *et al*'s equation (12) for the specific coverage fraction (the corrected form of which is equation (C.3)), namely

$$\alpha = \frac{1}{1 + k_d[1/k_b + \gamma/(\hat{T} - \alpha p)]} \quad [\text{sic}], \quad (\text{C.5})$$

where $\gamma = (1/k_n + 1/k_b)[N]$, cannot be correct by the following reasoning. In the absence of a non-specific complex background ($[N] \rightarrow 0$, and thus $\gamma \rightarrow 0$), this equation predicts that the coverage fraction should be independent of spike-in concentration ($\alpha \rightarrow 1/(1 + k_d/k_b)$), and indeed equal to their predicted binding affinity-dependent saturation coverage over the whole range of spike-in concentrations \hat{T} . This is obviously wrong, as evidenced by a version of Affymetrix's U95a latin square spike-in experiment without complex background [13] in which the experimentally obtained coverage fraction clearly responds to target concentration.

Glossary

Hybridization. The reversible chemical reaction by which target molecules in solution bind to probes attached to the microarray surface to form duplexes.

Microarray. A high-throughput device for detecting the presence of large biological molecules (DNA, RNA or proteins) of specific known letter sequences via their binding to molecules of complementary sequences attached to a solid surface. They are high-throughput in the sense that large numbers of sequences are tested for in a single device. The microarrays discussed here are oligonucleotide gene expression microarrays, that is, they have short DNA probes and are intended for the detection of expressed genes through their messenger RNA.

Non-specific hybridization. The hybridization of target molecules with sequences other than those of the intended sequence. When dealing with microarrays with a PM/MM (perfect match/mismatch) design, 'non-specific' is used to mean 'non-PM-specific', that is, hybridization of target molecules which are not complementary to the PM sequence, irrespective of whether they are binding to the PM or MM member of a probe pair.

Perfect match/mismatch probes. (Conventionally abbreviated as PM and MM.) A common design in Affymetrix GeneChip microarrays is to represent each targeted nucleotide sequence by two neighbouring probe

features: the PM, whose DNA probe sequence is exactly complementary to the target sequence, and the MM, whose DNA sequence is identical to the PM sequence except that the base in the central position of the probe sequence is replaced by a base complementary to that in the PM sequence. The idea behind the MMs is that they should respond to non-specific targets in a way similar to their PM partner, and can be used as a way of controlling biases due to non-specific hybridization.

Probe. A biological molecule attached to the microarray surface during fabrication.

Spike-in experiment. An experiment in which known concentrations of a specific set of target molecules are artificially added to a solution not otherwise containing those specific targets, and the solution hybridized onto microarrays.

Target. A biological molecule in the solution hybridized onto the microarray during a laboratory experiment.

References

- [1] http://www.affymetrix.com/support/technical/sample_data/datasets.affx
- [2] <http://cran.r-project.org/>
- [3] http://www.maths.anu.edu.au/cbis/~conrad/Spike_in_Isotherms
- [4] http://www.izbi.uni-leipzig.de/englisch/downloads_links/programs/hook.php?group=links
- [5] Binder H 2006 Thermodynamics of competitive surface adsorption on DNA microarrays *J. Phys.: Condens. Matter* **18** S491–523
- [6] Binder H, Brückner J and Burden C J 2009 Nonspecific hybridization scaling of microarray expression estimates: a physicochemical approach for chip-to-chip normalization *J. Phys. Chem. B* **113** 2874–95
- [7] Binder H, Kirsten T, Loeffler M and Stadler P F 2004 Sensitivity of microarray oligonucleotide probes: variability and effect of base composition *J. Phys. Chem.* **108** 18003–14
- [8] Binder H, Krohn K and Preibisch S 2008 'Hook'-calibration of genechip microarrays: chip characteristics and expression measures *Algorithms Mol. Biol.* **3** 11
- [9] Binder H and Preibisch S 2005 Specific and nonspecific hybridization of oligonucleotide probes on microarrays *Biophys. J.* **89** 337–52
- [10] Binder H and Preibisch S 2006 Genechip microarrays—signal intensities, RNA concentrations and probe intensities *J. Phys.: Condens. Matter* **18** S537–66
- [11] Binder H and Preibisch S 2008 'Hook'-calibration of genechip microarrays: theory and algorithm *Algorithms Mol. Biol.* **3** 12
- [12] Binder H, Preibisch S and Berger H 2009 Calibration of microarray gene-expression data *Methods Mol. Biol.* **576** 375–407
- [13] Burden C J 2009 Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed *Phys. Biol.* **5** 016004
- [14] Burden C J, Pittelkow Y E and Wilson S R 2004 Statistical analysis of adsorption models for oligonucleotide microarrays *Stat. Appl. Genet. Mol. Biol.* **3** Article 35

- [15] Burden C J, Pittelkow Y E and Wilson S R 2006 Adsorption models of hybridisation and post-hybridisation behaviour on oligonucleotide microarrays *J. Phys.: Condens. Matter* **18** 5545–65
- [16] Carlon E and Heim T 2006 Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays *Physica A* **362** 433–49
- [17] Heim T, Tranchevent L-C, Carlon E and Barkema G T 2006 Physical-chemistry-based analysis of Affymetrix microarray data *J. Phys. Chem. B* **110** 22786–95
- [18] Heim T, Wolterink J Klein, Carlon E and Barkema G T 2006 Effective affinities in microarray data *J. Phys.: Condens. Matter* **18** S525–36
- [19] 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
- [20] Held G A, Grinstein G and Tu Y 2003 Modeling of DNA microarray data by using physical properties of hybridization *Proc. Natl Acad. Sci.* **100** 7575–80
- [21] Held G A, Grinstein G and Tu Y 2006 Relationship between gene expression and observed intensities in DNA microarrays—a model study *Nucleic Acids Res.* **34** e70
- [22] Kroll K M, Barkema G T and Carlon E 2008 Modeling background intensity in DNA microarrays *Phys. Rev. E* **77** 061915
- [23] Li S, Pozhitkov A and Brouwer M 2008 A competitive hybridization model predicts probe signal intensity on high density DNA micorarrays *Nucleic Acids Res.* **36** 6585–91
- [24] McCullagh P and Nelder J A 1989 *Generalized Linear Models* 2nd edn (London: Chapman and Hall)
- [25] Mulders G C W M, Barkema G T and Carlon E 2009 Inverse langmuir method for oligonucleotide microarray analysis *BMC Bioinformatics* **10** 64
- [26] Nguyen K 2009 Extended investigations in the physics of oligonucleotide microarrays *ANU Undergrad. Res. J.* **1** 29–38
- [27] Ono N, Suzuki S, Furusawa C, Agata T, Kashiwagi A, Shimizu H and Yomo T 2008 An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays *Bioinformatics* **24** 1278–85
- [28] SantaLucia J 1998 A unified view of polymer, dumbbell and oligonucleotide DNA nearest neighbour thermodynamics *Proc. Natl. Acad. Sci.* **95** 1460–5
- [29] Skvortsov D, Abdueva D, Curtis C, Schaub B and Tavaré S 2007 Explaining differences in saturation levels for Affymetrix Genechip@arrays *Nucleic Acids Res.* **35** 4154–63
- [30] Suzuki S, Ono N, Furusawa C, Kashiwagi A and Yomo T 2007 Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays *BMC Genomics* **8** 373