Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays

2006 J. Phys.: Condens. Matter 18 5545

(http://iopscience.iop.org/0953-8984/18/23/024)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **18** (2006) 5545–5565 doi:10.1088/0953-8984/18/23/024

Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays

Conrad J Burden[1,2], **Yvonne Pittelkow**[1] **and Susan R Wilson**[1]

[1] Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia
[2] John Curtin School of Medical Research, Australian National University, Canberra, ACT 0200, Australia

E-mail: Conrad.Burden@anu.edu.au

Received 23 January 2006, in final form 28 April 2006
Published 26 May 2006
Online at stacks.iop.org/JPhysCM/18/5545

Abstract

Analysis of data from an Affymetrix Latin Square spike-in experiment indicates that measured fluorescence intensities of features on an oligonucleotide microarray are related to spike-in RNA target concentrations via a hyperbolic response function, generally identified as a Langmuir adsorption isotherm. Furthermore, the asymptotic signal at high spike-in concentrations is almost invariably lower for a mismatch feature than for its partner perfect match feature. We survey a number of theoretical adsorption models of hybridization at the microarray surface and find that in general they are unable to explain the differing saturation responses of perfect and mismatch features. On the other hand, we find that a simple and consistent explanation can be found in a model in which equilibrium hybridization is followed by partial dissociation of duplexes during the post-hybridization washing phase.

1. Introduction

Oligonucleotide microarrays are designed to enable the evaluation of simultaneous expression of large numbers of genes in prepared messenger RNA samples. Details of the technology and the design and manufacture of Affymetrix GeneChip arrays, the focus of this paper, can be found in the review of Nguyen *et al* [21] or at the Affymetrix website http://www.affymetrix.com/technology/index.affx. The purpose of this paper is to examine physical models of hybridization of RNA at the microarray surface in the light of differing responses of perfect match and mismatch probes.

In the manufacture of Affymetrix arrays, single strand DNA probes, 25 bases in length, are synthesized base by base onto a quartz substrate using a photolithographic process. They are attached to the substrate via short covalently bonded linker molecules roughly 10 nm apart. A

microarray chip surface is divided into some hundreds of thousands of regions called features, commonly 11–20 μm square, and with the single strand DNA probes within each feature being synthesized to a specific nucleotide sequence.

A key step in the laboratory process of gene detection with microarrays is the hybridization of complementary RNA (cRNA) target molecules fractionated to lengths of typically 50–200 bases onto the single strand DNA probes. The density of hybridized probe–target duplexes in each feature is detected via intensity measurements of fluorescent dye attached to the target cRNA molecules. Each gene or expressed sequence tag (EST) is represented by a set of 11–20 (dependent on the chip type) pairs of features using sequences of length 25 selected for their predicted hybridization properties and specificity to the target gene. The first element of the pair, termed the perfect match (PM), is designed to be an exact match to the target sequence, while the second element, the mismatch (MM), is identical except for the middle (13th) base being replaced by its complement.

A number of studies have demonstrated the appropriateness of Langmuir adsorption theory for understanding probe–target hybridization at the surface of microarrays. Experimental work includes that of Nelson *et al* [20], Peterson *et al* [22, 23] and Dai *et al* [8]. Analyses which have sought to match Langmuir adsorption isotherms with data from an Affymetrix spike-in experiment include those of Held *et al* [13], Hekstra *et al* [12], Lemon *et al* [17], Burden *et al* [6] and Binder *et al* [5].

The ultimate aim of such work is to establish a functional relationship between measured fluorescence intensities and underlying target concentration parameterized by known physical properties such as probe base sequences. If such a relationship could be established, it would offer the possibility of an absolute measure of RNA target concentration, as opposed to an arbitrarily defined 'expression measure'. Fundamental to establishing this relationship is a model which accurately describes the physics of the various steps involved in producing a set of intensity measurements from a given messenger RNA (mRNA) target concentration. The two steps we focus on in this paper are hybridization at the microarray surface and the subsequent washing step, designed to removed unbound target molecules.

A little-recognized shortcoming of existing hybridization models based on Langmuir adsorption theory is their inability to explain the differing responses of PM and MM fluorescence intensity signals at saturation concentrations of RNA. That the asymptotic response of a MM feature at high PM-specific spike-in concentration should be less than that of the neighbouring PM feature is hardly news to an experimental biologist, and yet this observation is surprisingly difficult to reconcile with Langmuir adsorption theory (see section 4). This problem was discussed in the early experimental work of Forman *et al* [10], who serendipitously recognized the 'unexpected benefit' of the phenomenon of differential response between PM and MM, but failed to find a satisfactory physical explanation. It is stated on the manufacturer's web page that 'The reason for including a MM probe is to provide a value that comprises most of the background cross hybridization and stray signal affecting the PM probe. It also contains a portion of the true target signal' [1]. Consequently, many researchers have come to view the MM signal as primarily an attempt to measure non-specific hybridization and other background signals, though in practice there are problems with using the MM signals for this purpose [15]. Since the MM signals are more than a measure of non-specific hybridization, we will concentrate in this paper on the view that MM features are primarily less responsive versions of the PM features, and seek to understand their differing responses at saturation. The difference between PM and MM probe signals can then be exploited as the result of a single, well-controlled change in one of the many parameters influencing the complicated process of hybridization. From this perspective one can obtain powerful insights into the physics and chemistry of hybridization at the microarray surface.

In section 2 we review the Langmuir or hyperbolic isotherm and its relationship to a well-known Affymetrix spike-in data set. Section 3 concentrates on an extension of the adsorption based hybridization models of Hekstra *et al* [12] and Halperin *et al* [11] which include the effects of non-specific hybridization, and which we show to be essentially equivalent to each other. This model is consistent with a hyperbolic response function, as observed in data from spike-in experiments. However, as we point out in section 4, it is unable to explain the observed difference between PM and MM signals at saturation concentrations. Section 5 is a survey of a number of possible improvements to our starting model of hybridization at the microarray surface, which seek to overcome this shortcoming. Many of these ideas have been canvassed in the literature, though in general they have not been rigorously examined in the light of the Hekstra/Halperin model. In general, we find no convincing way of explaining the PM/MM difference at saturation by reference only to the hybridization step. In section 6 we consider the post-hybridization washing step, and find this to be the most promising explanation for the PM/MM difference. In section 7 we summarize our findings and draw conclusions. Many of the technical calculations are relegated to appendices.

2. The Langmuir isotherm model

Langmuir adsorption theory is based on an assumption that there are two competing processes driving hybridization: adsorption, i.e. the binding of target molecules to immobilized probes to form duplexes, and desorption, i.e. the reverse process of duplexes dissociating into separate probe and target molecules

$$\text{Probe} + \text{Target} \rightleftharpoons \text{Duplex}. \tag{1}$$

Herein we shall always use the word 'probe' to indicate single strand DNA immobilized on the microarray, 'target' to indicate RNA in solution and 'duplex' to indicate a bound probe–target pair. Both the forward and reverse processes are determined by chemical rate constants which depend on a number of factors including activation energies and temperature. Adsorption models of microarrays often lead to a hyperbolic response function, or equilibrium Langmuir isotherm, relating RNA target concentration x to a measured equilibrium fluorescence intensity y, namely

$$y(x) = y_0 + b\frac{x}{x+K}. \tag{2}$$

The isotherm is defined by three parameters: y_0 is the measured background intensity at zero target concentration, b is the saturation intensity above background at infinite target concentration and K is the target concentration required to reach half saturation. The physical origins of these parameters will be discussed in detail below.

In a previous paper we have carried out an extensive statistical analysis [6] of fits of the hyperbolic and other response functions to the PM probes in the publicly available data from the Affymetrix Human HG-U95A Latin Square spike-in experiment (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). In this experiment genes (or, more precisely, RNA transcripts) were spiked in at cyclic permutations of the set of known concentrations, together with a background of cRNA extracted from human pancreas. The data consist of fluorescence intensity values from a set of 14 probesets corresponding to 14 separate genes, each containing 16 probe pairs. For each probeset a set of fluorescence intensity values was 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. In common with previous analyses of this data set, our study concentrated on data from 12 of the 14 genes, omitting data from two defective genes.

Fits of a number of functions to the fluorescence intensities were compared using a rigorous statistical analysis. The optimum model of those considered for this data set is summarized as follows:

(i) Measured fluorescence values can be approximated by a gamma distribution with a mean given by equation (2) and a constant coefficient of variation, here ≈ 0.17.
(ii) The equilibrium isotherm equation (2) tracks fold changes from both PM and MM probes over the range of spiked-in concentrations from <1 pM to >1000 pM.
(iii) All three parameters y_0, b and K are dependent on the probe sequence (in contrast with the findings of Held *et al* [13]).
(iv) MM features almost invariably saturate at a lower asymptotic intensity $y_0 + b$ than their PM counterparts.

Plots of fits of the hyperbolic response function to intensity data from the 16 PM and MM features corresponding to a typical one of the 12 genes considered is reproduced from [6] in figure 1. A measure of the closeness of the fit is the unscaled deviation, defined by equation (8) of [6]. This quantity is the analogue for a generalized linear model of the mean square error in a standard linear regression. For each of the 12 genes in question, the unscaled deviation per degree of freedom is much the same, the gene shown in figure 1 being somewhere near the middle of the range. Since the complete set of models considered in [6] was a set of nested models, we were able to use standard statistical tests based on accepted principles of balancing accuracy and parsimony to reject alternative functional forms for the fluorescence intensity response function, in favour of the hyperbolic form of equation (2). The rejected response functions included a Sips isotherm [25] and a function modelling non-equilibrium adsorption (see equation (15)). While details of the analysis were only reported for PM features in our earlier paper, we have subsequently also confirmed points (i) to (iii) for MM features (see appendix A for comparison of hyperbolic and Sips isotherms). Point (iv) was confirmed by fits of the hyperbolic response function by Hekstra *et al* (see figure 2(A) of [12] and accompanying text) and our own calculations, and is apparent from figure 1.

3. Physical models leading to the hyperbolic isotherm

In what follows we define 'specific' to mean PM specific. All other hybridization will be referred to as 'non-specific'. Hekstra *et al* [12] have modelled hybridization at the microarray surface in the combined presence of a specific cRNA target species and a single, non-specific target species using classical chemical adsorption kinetics. The model gives a hyperbolic response function of the form equation (2) and predicts values for the parameters y_0, b and K in terms of chemical rate constants and physical properties of the microarray. It is straightforward to extend their results to any number of non-specific species [5].

The hyperbolic isotherm is equivalently derivable from statistical mechanics by considering the Gibbs distribution at constant chemical potential [14]. Halperin *et al* [11] have used this approach to study adsorption in microarray chips in the presence of non-specific hybridization. In order to establish a notation for subsequent sections, we rederive here the hyperbolic isotherm using the Halperin approach. We shall further augment the approach to include partial zippering of duplexes, that is, the idea is that a particular probe–target duplex can exist in a number of possible partially zipped-up configurations $\alpha = 1, 2, \ldots$ (see, for example, [9]).

For a given feature on the microarray surface, whether PM or MM, let the concentration of target molecules specific to the PM feature of the matched pair be x, and the concentration of the non-specific species i be z_i. Further, let θ_α be the fraction of a given feature covered

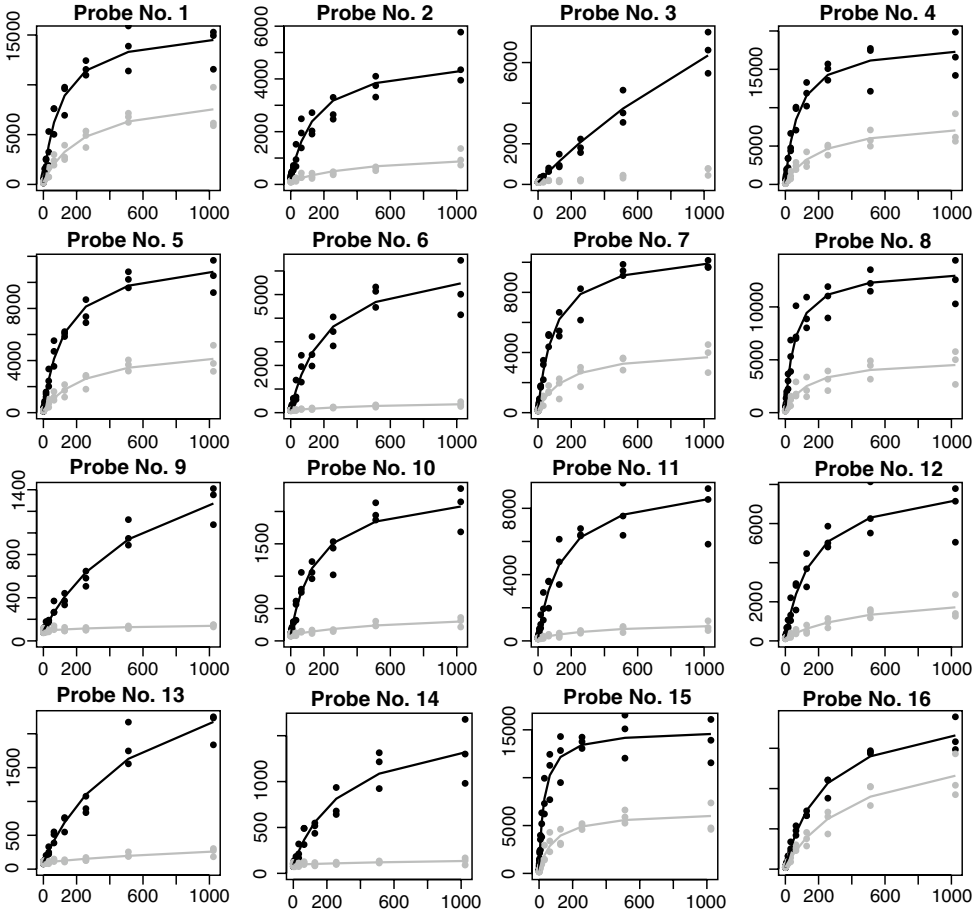

Figure 1. Fits of equation (2) to fluorescence intensity data for the 16 PM (black) and 16 MM (grey) features of the gene 37 777_at probeset of the Affymetrix spike-in experiment. Concentrations (horizontal axes) are picomolar and fluorescence intensities (vertical axes) are in the arbitrary units used 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.

by specific duplexes in the partially zippered configuration α, and likewise $\phi_{i\alpha}$ be the fraction covered by duplexes formed with non-specific target species i in configuration α. The fraction covered with unmatched single strand probes is therefore $1-\theta-\sum_i \phi_i$, where the total fraction of sites holding, respectively, specific and non-specific duplexes of the ith species is $\theta = \sum_\alpha \theta_\alpha$ and $\phi_i = \sum_\alpha \phi_{i\alpha}$. The free energy per mole of probe sites at the microarray surface is

$$\gamma = RT\left[\sum_\alpha \theta_\alpha \ln\theta_\alpha + \sum_{i,\alpha} \phi_{i\alpha}\ln\phi_{i\alpha} + \left(1-\theta-\sum_i \phi_i\right)\ln\left(1-\theta-\sum_i \phi_i\right)\right] + \sum_\alpha \theta_\alpha \mu^0_{\mathrm{pt}\alpha} + \sum_{i,\alpha}\phi_{i\alpha}\mu^0_{\mathrm{pti}\alpha} + \left(1-\theta-\sum_i\phi_i\right)\mu^0_{\mathrm{p}}, \tag{3}$$

where $\mu^0_{\mathrm{pt}\alpha}$, $\mu^0_{\mathrm{pti}\alpha}$ and μ^0_{p} are respectively reference state chemical potentials per mole of

specific and non-specific probe–target duplexes in configuration α, and of unmatched probes[3]. R is the gas constant and T the absolute temperature. The exchange chemical potentials of the various species of probe–target duplexes are

$$\frac{\partial\gamma}{\partial\theta_\alpha} = RT\left[\ln\theta_\alpha - \ln\left(1-\theta-\sum_i\phi\right)\right] + \mu^0_{\mathrm{pt}\alpha} - \mu^0_{\mathrm{p}},$$
$$\frac{\partial\gamma}{\partial\phi_{i\alpha}} = RT\left[\ln\phi_{i\alpha} - \ln\left(1-\theta-\sum_i\phi\right)\right] + \mu^0_{\mathrm{pt}i\alpha} - \mu^0_{\mathrm{p}}.$$

At equilibrium these exchange chemical potentials balance the chemical potentials of the corresponding target molecule species in solution. Assuming that the bulk concentrations of target molecules are not appreciably affected by hybridization, these are given in terms of reference values μ^0_{t} and μ^0_{ti} at reference concentrations x_0 and z_{0i} of specific and non-specific target molecules by

$$\begin{aligned}\mu_{\mathrm{t}} &= \mu^0_{\mathrm{t}} + RT\ln\frac{x}{x_0},\\ \mu_{ti} &= \mu^0_{ti} + RT\ln\frac{z_i}{z_{0i}}.\end{aligned} \tag{4}$$

Matching exchange chemical potentials with target chemical potentials gives

$$RT\ln\frac{x}{x_0} = RT\left[\ln\theta_\alpha - \ln\left(1-\theta-\sum_i\phi\right)\right] + \Delta G_\alpha,$$
$$RT\ln\frac{z_i}{z_{0i}} = RT\left[\ln\phi_{i\alpha} - \ln\left(1-\theta-\sum_i\phi\right)\right] + \Delta G_{i\alpha},$$

where we have defined the duplex binding free energies

$$\Delta G_\alpha = \mu^0_{\mathrm{pt}\alpha} - \mu^0_{\mathrm{p}} - \mu^0_{\mathrm{t}}, \qquad \Delta G_{i\alpha} = \mu^0_{\mathrm{pt}i\alpha} - \mu^0_{\mathrm{p}} - \mu^0_{ti}. \tag{5}$$

Solving for the duplex coverage fractions θ_α and $\phi_{i\alpha}$, and summing over configurations α, we obtain the isotherms

$$\theta = \frac{x/K_{\mathrm{S}}}{1 + x/K_{\mathrm{S}} + \sum_i z_i/K_i} \tag{6}$$

$$\phi_i = \frac{z_i/K_i}{1 + x/K_{\mathrm{S}} + \sum_j z_j/K_j}, \tag{7}$$

where K_{S}^{-1} and K_i^{-1} are effective equilibrium constants for specific and non-specific hybridizations given by

$$K_{\mathrm{S}}^{-1} = x_0^{-1}\sum_\alpha \mathrm{e}^{-\Delta G_\alpha/RT}, \qquad K_i^{-1} = z_{0i}^{-1}\sum_\alpha \mathrm{e}^{-\Delta G_{i\alpha}/RT}. \tag{8}$$

Introducing proportionality constants b_{S} and b_i for the specific and non-specific hybridizations and a physical optical background a, the measured fluorescence intensity is given by

$$y(x) = a + b_{\mathrm{S}}\theta + \sum_i b_i\phi_i \tag{9}$$

$$= y_0 + b\frac{x}{x+K}, \tag{10}$$

[3] Halperin *et al* [11] also include a term for the charge density dependent electrostatic free energy, which we discuss briefly in section 5.3.

where

$$y_0 = a + A, \qquad b = b_S - A, \qquad K = K_S B, \tag{11}$$

and

$$A = \frac{1}{B}\sum_i \frac{b_i z_i}{K_i}, \qquad B = 1 + \sum_i \frac{z_i}{K_i}. \tag{12}$$

The presence of non-specific hybridization does not spoil the hyperbolic form of the Langmuir isotherm equation (2), but does influence the parameters y_0, b and K. The purpose of equations (8), (11) and (12) is to relate the estimated isotherm parameters to the underlying physical parameters: a (the physical background value in the absence of any hybridization), b_S and b_i (proportionality constants relating the incremental change in measured intensity to an incremental change in duplex fraction for the specific and non-specific hybridizations, respectively), duplex binding energies ΔG_α and $\Delta G_{i\alpha}$, and a set of non-specific background target concentrations z_i. The parameters b_S and b_i are a measure of the amount of fluorescent light emitted per hybridized target molecule. Fluorescent dye is bound only to the target molecules (in fact only to U and C bases), so b_S and b_i can only be functions of specific and non-specific target sequences, and not probe sequences. Equations (10)–(12) are a generalization of equation (2) of Hekstra *et al* [12].

4. Inconsistency of adsorption models with observed PM/MM saturation intensities

The model given by equations (8)–(12) inescapably leads to a conclusion that the PM and MM intensity measurements for a given probe pair must saturate at the same asymptotic intensity value, contradicting the observed fits to experimental data. This point has been inferred previously in regard to adsorption models [10], but does not appear to be generally appreciated in the literature, with the exception of work by Peterson *et al* [23].

Consider two neighbouring features on a microarray, one PM and one MM, their probe sequences differing only by the middle base. Recall that, in this paper, we define the word 'specific' to mean those target cRNAs which are exact complements to the PM sequence, even when dealing with the MM feature. For our purposes, this definition will prove useful given that, for most probe pairs, the dominant part of the MM signal at high spike-in concentrations in the Affymetrix experiment appears to come from hybridization of spiked-in target RNAs complementary to the PM sequence. Parameters relating to the PM and MM features will be indicated by superscripts PM and MM, respectively.

Although the sums occurring in equation (12) will be over the same set of non-specific targets for PM as for MM, one can expect $A^{\mathrm{MM}} \neq A^{\mathrm{PM}}$ since in general $K_i^{\mathrm{MM}} \neq K_i^{\mathrm{PM}}$. Considering the asymptotic intensities at high concentration, however, equations (10) and (11) imply that, under the Hekstra model, the non-specific hybridization effects cancel out:

$$\begin{aligned} y^{\mathrm{MM}}(\infty) &= y_0^{\mathrm{MM}} + b^{\mathrm{MM}} = a + b_S, \\ y^{\mathrm{PM}}(\infty) &= y_0^{\mathrm{PM}} + b^{\mathrm{PM}} = a + b_S. \end{aligned} \tag{13}$$

An essential step in this argument is the claim that the parameters a and b_S do not differ between intensity measurements from a neighbouring PM/MM pair of features. For the physical background a this is clearly a reasonable assumption: physical properties of the chip in the absence of any hybridization, such as reflectance, are unlikely to vary significantly over a distance of a few micrometres. For the parameter b_S the argument is more subtle. From equation (9), b_S is, up to a multiplicative constant, the expected number of biotin labels per hybridized specific target molecule. Importantly, b_S confers on equation (9) no information about probe–target binding affinities, this information being contained in the coverage fraction

θ. By our current definition of 'specific', target molecules contributing to the specific part of the signals of a given PM/MM pair of features are drawn from the same subset of molecules in the RNA solution, namely those containing a contiguous PM-specific subsequence of 25 bases. Hence b_S is the same for both members of a neighbouring PM/MM pair. The Hekstra or Halperin model formulated above then necessarily entails that $y_0^{MM} + b^{MM} = y_0^{PM} + b^{PM}$, in obvious contradiction with the values of y_0, and b obtained by fitting the spike-in data.

The source of the problem is that any model leading to the coverage fraction given by equation (6) entails that, at sufficiently high specific target concentration, all probes form duplexes: as $x \to \infty, \theta \to 1$. That is, all probes in the feature are predicted to form duplexes if saturated with enough specific target, even in the case of the MM feature. A subtle point to note is that this is true irrespective of the bulk solution melting temperature of duplexes, which is defined as the temperature at which half the total number of single strand targets are free and half are bound as duplexes in bulk solution. This temperature can be calculated [4] in terms of enthalpy and entropy by balancing forward and backward reaction rates under the constraints of stoichiometry, namely: $2[T] + [T.T] =$ constant, where [T] and [T.T] are bulk concentrations of single strand and duplex targets respectively[4]. However, this stoichiometric constraint does not apply for the adsorption reaction at the microarray surface: because the solution target volume is effectively infinite, the target concentration is unchanged as hybridization proceeds and the coverage fraction θ increases towards its finite equilibrium value $\theta \leqslant 1$. Even above the bulk solution melting temperature, the forward reaction can be forced by setting the target concentration sufficiently high. The upshot is that, irrespective of temperature, Langmuir adsorption theory tells us that a feature will saturate at infinite target concentration.

5. Other hybridization effects

Clearly the above model does not account for all possible effects during the complex process of hybridization. In this section we consider a number of other possible hybridization effects, some of which have been proposed in the literature as putative explanations for the differing measured PM/MM saturation intensities. In general, we find none of these effects to be a strong candidate, and believe that the explanation of the PM/MM saturation difference is unlikely to lie with the hybridization step.

5.1. Sips isotherm

The problem of differential PM/MM saturation was recognized in the context of a simple Langmuir model without non-specific hybridization by Peterson *et al* [23], who explain their experimental data by invoking a Sips isotherm to explain a lower MM response curve at high target concentrations. The Sips isotherm [25] is an empirical response curve believed to correspond to an adsorption model in which chemical reaction rates are drawn from a pseudo-Gaussian distribution. Peterson *et al*'s experimental results are indeed a good fit to the Sips isotherm; however, their experiment differs from the conditions of the hybridization of Affymetrix chips in one important aspect, namely the hybridization temperature. The Peterson experiment was carried out at a hybridization temperature of 20 °C, while Affymetrix microarrays are hybridized at 45 °C. Furthermore, Peterson *et al* found that heating the hybridization buffer to 37 °C and then cooling back to 20 °C almost completely removed any difference in equilibrium saturation intensities between PM and MM probes. This appears to be the effect of a first order phase transition which sets in at a temperature well below the

[4] In section 5.5 we argue that, for the Affymetrix spike-in data set $[T] \approx x$, the spike-in concentration.

Affymetrix hybridization temperature. We comment on the problem of determining the phase structure in section 5.4.

To determine whether the hyperbolic or Sips isotherm is more appropriate for the Affymetrix spike-in data we have carried out a statistical analysis comparing the fits of the MM data to both isotherms. Our results, summarized in appendix A, show that for the Affymetrix spike-in data the extra parameters involved in invoking the Sips isotherm are not significant, and that a hyperbolic response function adequately describes the data. We conclude that, at a hybridization temperature of 45 °C, the more appropriate empirical fit to the spike-in data is equation (2), with $y^{\mathrm{MM}}(\infty) < y^{\mathrm{PM}}(\infty)$.

5.2. Non-equilibrium hybridization

In an earlier paper [6] we examined the possibility that hybridization had not reached equilibrium in the Affymetrix spike-in experiment. We considered the simple non-equilibrium one-step model without non-specific hybridization, namely,

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = k_{\mathrm{f}} x(1-\theta) - k_{\mathrm{b}}\theta, \tag{14}$$

where k_{f} and k_{b} are forward and backward chemical reaction rates. The solution corresponding to the initial condition $\theta(x, 0) = 0$ is

$$\theta(x,t) = \frac{x}{x+K_{\mathrm{S}}}\left[1 - \mathrm{e}^{-(x+K_{\mathrm{S}})k_{\mathrm{f}}t}\right], \tag{15}$$

where $K_{\mathrm{S}} = k_{\mathrm{b}}/k_{\mathrm{f}}$. A statistical analysis of the data showed that the extra degree of freedom distinguishing the non-equilibrium from the equilibrium solution (equation (2)) is not significant. That is, our finding was that the equilibrium solution is the more appropriate model.

However, textbook descriptions of duplex formation (see, for instance, [7], pp 1215–1219) imply that hybridization is more accurately described as a two-step process: a slow rate determining step in which an initial two or three base pairs form, followed by a fast 'zipping-up' step involving the remaining base pairs. Measured forward reaction rates for duplex formation may typically be of the order of 10^6 mol^{-1} s^{-1} [28], potentially translating to timescales of several hours at picomolar concentrations. In order to establish more rigorously that the hybridization had reached equilibrium in the spike-in experiment, we have considered in appendix B a quasi-equilibrium hybridization model with two timescales. Chemical reaction rates leading to the initiation configuration with two or three base pairs formed are taken to be slow, while other reaction rates are assumed to equilibrate on short timescales. Again this model leads to non-equilibrium solutions taking the form of equation (15), which differs from the hyperbolic form observed in the data. This confirms that our previous statistical analysis is appropriate even when a two-step hybridization process is taken into account. We therefore believe that equilibrium thermodynamics to be the correct framework for studying hybridization for this data set.

5.3. Electrostatic surface potential

Halperin *et al* [11] include in the free energy equation (3) a term γ_{el} for the charge density dependent electrostatic free energy. The effect of this term is to change the effective equilibrium constants K_{S} and K_i by a finite amount via the replacements $\Delta G_\alpha \to \Delta G_\alpha + \partial\gamma_{\mathrm{el}}/\partial\theta_\alpha$ and $\Delta G_{i\alpha} \to \Delta G_{i\alpha} + \partial\gamma_{\mathrm{el}}/\partial\phi_{i\alpha}$ in equation (8). This introduces a θ dependence to K_{S} and has the potential to change the shape of the isotherm from a hyperbolic form [27]. However, it cannot be the explanation for differing PM/MM saturation intensities, as the adjusted form of equation (6) still satisfies $\theta \to 1$ as $x \to \infty$.

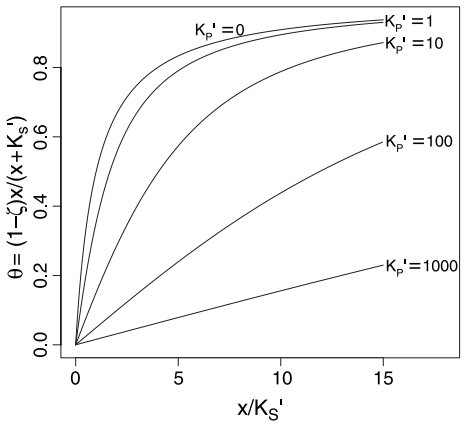

Figure 2. Plots of the coverage fraction θ of probe–target duplexes against the dimensionless target concentration x/K'_{S} given by the solution equations (C.11) and (C.12) to the one-dimensional model described in appendix C, for various values of the effective probe–probe duplex equilibrium constant $K'_{\mathrm{P}} = K_{\mathrm{P}}/(1 + K_{\mathrm{Q}})^2$. Probe duplex or probe self-interaction free energies of ΔG_{P} or $\Delta G_{\mathrm{Q}} = 0$, -1 and -2 kcal mol^{-1} at 45 °C correspond to K_{P} or K_{Q} values of 1, 4.9 and 23.7, respectively.

5.4. Competitive hybridization with probe–probe pairs and probe self-interactions

Forman *et al* [10] advanced the hypothesis that the observed divergence of saturation intensities between PM and MM features is caused by hybridization of neighbouring probe–probe pairs, rendering a certain fraction of each feature unavailable for target binding to form probe–target duplexes. Probe–probe interactions are possible if we assume probes of approximate length 8 nm and bound by flexible linker molecules to a glass substrate to have an average interprobe distance of the order of 10 nm [24], especially given that some clustering of probes is to be expected. It has also been recognized [5] that self-interaction of probes via probe folding may render a fraction of probes unavailable for hybridization and affect adsorption isotherms.

In appendix C we discuss how hybridization in the presence of probe–probe and probe self-interactions may be modelled. In agreement with reference [5] we find that probe self-interactions have the effect of scaling the equilibrium constant K_{S} for the adsorption process. However, this cannot be the explanation for differing PM/MM saturation intensities as it does not change the saturation asymptote. The probe–probe interactions, on the other hand, are more complex, and we show in appendix C that one is naturally led to the random lattice version of a two-dimensional statistical mechanics model known as the monomer–dimer model. No solution to this model exists, even for the more tractable cases of regular lattices, though some numerical work has been done for the regular square lattice monomer–dimer model [2].

In appendix C we tackle the unphysical but analytically tractable one-dimensional model of competitive hybridization with probe–target and probe–probe duplexes. We see that a probe–probe binding energy of 1 or 2 kcal mol^{-1} is enough to make a noticeable difference to the adsorption isotherm in this approximation (see figure 2). The one-dimensional model saturates at 100% coverage of probe–target duplexes at high target concentration and so is unable to explain the divergence of PM and MM saturation intensities. However, it is well known that the behaviour of statistical mechanics models in one and two dimensions can be very different. It is known, for instance, that a one-dimensional model with local interactions cannot lead to a phase transition, whereas a number of two-dimensional models are known to exhibit phase transitions at critical temperatures or densities [3].

The evidence from numerical calculations of the monomer–dimer model on a regular square lattice is that it does not have a phase transition for non-zero monomer density [2], but we are unaware of any numerical simulations for the random lattice case more relevant to our problem. It is therefore still possible that the microarray surface configuration could undergo a phase transition from a disordered phase with low concentration of probe–probe duplexes to an ordered phase in which a high concentration of probe–probe duplexes line up along a particular direction. This could explain the differing intensity measurement curves of MM features observed before and after quenching in the experiments of Peterson *et al* [23]. Whether the Forman hypothesis can explain the observed difference in PM/MM saturation intensities, however, remains an open question, though any such function is unlikely to be consistent with the observed hyperbolic response function.

5.5. Competitive bulk hybridization

By competitive bulk hybridization we mean the hybridization of specific target molecules T in solution either with (i) other specific target molecules T′ which might happen to be, at least in part, self-complementary ($\mathrm{T} + \mathrm{T} \rightleftharpoons \mathrm{T.T}$), (ii) non-specific target molecules which happen to have approximately complementary nucleotide sequences ($\mathrm{T} + \mathrm{T}' \rightleftharpoons \mathrm{T.T}'$), or (iii) target self-interactions ($\mathrm{T} \rightleftharpoons \mathrm{T}_{\mathrm{folded}}$). Halperin *et al* [11] have considered the effect on equilibrium isotherms of the first two types of bulk hybridization, and type (iii) can be dealt with in a similar way. Assuming that probe–target hybridization has a negligible effect on bulk target concentrations, they argue that equilibrium isotherms can be obtained from isotherms such as equation (6) by replacing the spike-in target concentration x with the single strand concentration [T] obtained by applying the law of mass action to the bulk hybridization reaction in solution.

For all three types of hybridization, we argue here that competitive bulk hybridization cannot explain differential PM/MM saturation. In each case, application of the law of mass action entails that $[\mathrm{T}] \to \infty$ as $x \to \infty$, so equation (6) with x replaced by [T] still implies 100% saturation of features in the high spike-in concentration limit for both PM and MM features.

Furthermore, we can rule out any significant effect on the isotherm from T.T hybridization for the probe sequences studied in the Affymetrix spike-in experiment by the following argument. The law of mass action implies that the behaviour of the single strand concentration goes from $[\mathrm{T}] \approx x$ at spike-in concentrations $x \ll K_{\mathrm{bulk}}^{-1}$ (where K_{bulk} is the equilibrium constant for the reaction $\mathrm{T} + \mathrm{T}' \rightleftharpoons \mathrm{T.T}'$) to $[\mathrm{T}] \approx (x/2K_{\mathrm{bulk}})^{1/2}$ at spike-in concentrations $x \gg K_{\mathrm{bulk}}^{-1}$. A significant effect from T.T hybridization would therefore lead to a Sips isotherm with parameter $\gamma = \frac{1}{2}$ at high spike-in concentration, which, by the analysis of appendix A, is not observed over the range of concentrations in the Affymetrix spike-in experiment.

6. The washing step

The hybridization step is followed by a washing step designed to remove unbound target molecules before scanning the microarray. During the washing step the target solution is flushed out of the cartridge containing the microarray and replaced by a washing buffer containing no RNA. Thus the ambient concentration of target molecules is set to zero, switching off the forward adsorption reaction. We argue here that the washing step is responsible for the measured differences between PM/MM intensity measurements at saturation concentrations. This idea has been proposed briefly by Zhang [29], but requires further analysis.

Let us assume that, immediately prior to washing, duplex coverage fractions on a given feature are given by the equilibrium model set out in section 3. That is, the fraction θ of sites

on a feature occupied by specific probe–target duplexes and the fraction ϕ_i covered by non-specific duplexes of species i are given by equations (6) and (7). During the washing process some of the duplexes will be dissociated. Suppose that the probability that a given probe–target duplex has survived up to a washing time t_W is $s(t_W)$ for a specific duplex and $s_i(t_W)$ for a non-specific duplex of species i. The survival functions s and s_i depend only on probe and target base sequences and not the ambient target concentrations x and z_i present during the prior hybridization step. They satisfy $s(0) = 1$ and are monotonically decreasing. The specific and non-specific duplex coverage fractions at time t_W are then

$$\theta(x, t_W) = \frac{s(t_W)x/K_S}{1 + x/K_S + \sum_i z_i/K_i} \tag{16}$$

$$\phi_i(x, t_W) = \frac{s_i(t_W)z_i/K_i}{1 + x/K_S + \sum_j z_j/K_j}. \tag{17}$$

Repeating the assumption used in section 2, that the measured fluorescence intensity is a linear function of the duplex coverage fractions, that is $y(x, t_W) = a + b_S\theta(x, t_W) + \sum_i b_i\phi(x, t_W)$, we find that at fixed t_W the hyperbolic form required by points (ii) and (iii) in section 2, namely

$$y(x, t_W) = y_0(t_W) + b(t_W)\frac{x}{x + K}, \tag{18}$$

is maintained, and that the 'observed' parameters y_0, b and K are now given by

$$y_0(t_W) = a + A(t_W), \qquad b(t_W) = s(t_W)b_S - A(t_W), \qquad K = K_S B, \tag{19}$$

where

$$A(t_W) = \frac{1}{B}\sum_i \frac{s_i(t_W)b_i z_i}{K_i}, \qquad B = 1 + \sum_i \frac{z_i}{K_i}. \tag{20}$$

Note that the parameter K is unaffected by the length of the washing process, and depends only on duplex binding free energies via the hybridization step. The asymptotic fluorescence intensity at high target concentration,

$$y(\infty, t_W) = y_0(t_W) + b(t_W) = a + s(t_W)b_S, \tag{21}$$

is depressed by the presence of the survival fraction $s(t_W)$.

To model the survival function $s(t_W)$, one expects the rate of dissociation of specific probe–target duplexes to be the product of the fraction $\theta(t_W)$ of probes forming specific duplexes and a washing rate κ which depends only on the probe and target nucleotide sequences. Assuming then that κ is independent of t_W, the survival function is

$$s(t_W) = \mathrm{e}^{-\kappa t_W}. \tag{22}$$

Since the binding affinity of a PM-specific target to a MM probe is less than to a PM probe, we expect in general that $\kappa^{MM} > \kappa^{PM}$, or equivalently, $s^{MM}(t_W) < s^{PM}(t_W)$ and hence $y^{MM}(\infty) < y^{PM}(\infty)$ as required.

Ideally one would like to test directly the veracity of the survival function equation (22) using data from a range of washing times. While this is not possible with spike-in data corresponding to a single value of t_W, we can at least check for qualitative agreement of the above scenario with probe sequence information.

From equations (19) and (22) one obtains

$$\kappa t_W = \log b_S - \log\left[y_0(t_W) + b(t_W) - a\right]. \tag{23}$$

For fixed t_W, the left-hand side is a measure of the rate at which probe–target duplexes dissociate due to washing, and should increase with decreasing binding affinity. The right-hand side

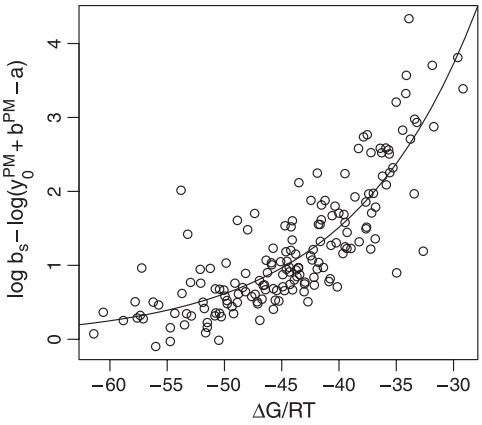

Figure 3. Plots of the estimate of the washing rate κ (times the washing time t_W, which is constant across all probes) given by equation (23) using Langmuir isotherm parameter fits for the PM probesets of the spike-in experiment described in section 2. ΔG is the RNA/DNA duplex free binding energy in bulk solution calculated using the nearest-neighbour stacking model of [26]. The solid curve is the exponential fit equation (24) with parameter values given in the text.

depends on the fitted isotherm parameters, $y_0(t_W)$ and $b(t_W)$, and two unknown parameters: a, the physical background, and b_S, the fluorescence intensity above background of a feature fully saturated with PM-specific probe–target duplexes. In order to make comparisons across the fitted spike-in data, we will take the two unknown parameters to be constant across all features of the microarray. While this may seem to be a radical assumption for b_S, we argue that, because the target mRNA is fractionated randomly to lengths of between 50 and 200 bases, the distribution of the number of U and C bases carrying biotin labels on PM-specific targets will not be strongly influenced by the relatively short 25-base subsequence of the probe. The total number of bases carrying labels on a saturated feature could therefore be replaced by a typical representative value independent of the feature.

In figure 3, we plot the right-hand side of equation (23) against RNA/DNA duplex free binding energy in bulk solution calculated using the nearest-neighbour stacking model and parameters of [26]. Values of y_0 and b are from fits of the PM data to hyperbolic isotherms as described in section 2. The value $a = 50$ was chosen to be slightly less than the lowest intensity value from the entire data set, though in practice any positive value up to 100 gives an almost identical plot. The choice of parameter $b_S = 31\,100$ only affects the vertical offset, and has been determined by setting $\log b_S = -\alpha$, where α is the intercept in a linear regression to an empirical function (also plotted in figure 3)

$$-\log\left[y_0(t_W) + b(t_W) - a\right] = \alpha + \beta\, e^{-\Delta G/(\lambda RT)}, \tag{24}$$

in which $\lambda = 11.1$ has been chosen to minimize the residual standard error. The linear regression gives $\alpha = 10.3$ and $\beta = 55.4$. We see that the data are consistent with a rate of duplex removal during washing that decreases exponentially to zero with increasing binding energy $-\Delta G$. The factor λ reflects the fact that effective duplex binding energies at the microarray surface are considerably less than bulk solution binding energies due to effects such as electrostatic blocking [18] (see also section 5.3) and a consequent enhancement of partial zippering (see equation (8)).

In figure 4 we examine the dependence of the estimated washing rate equation (23) on the nucleotide composition of probe sequences. The upper four bar charts show estimated PM (MM) washing rates averaged over sequences with a particular base at the ith position

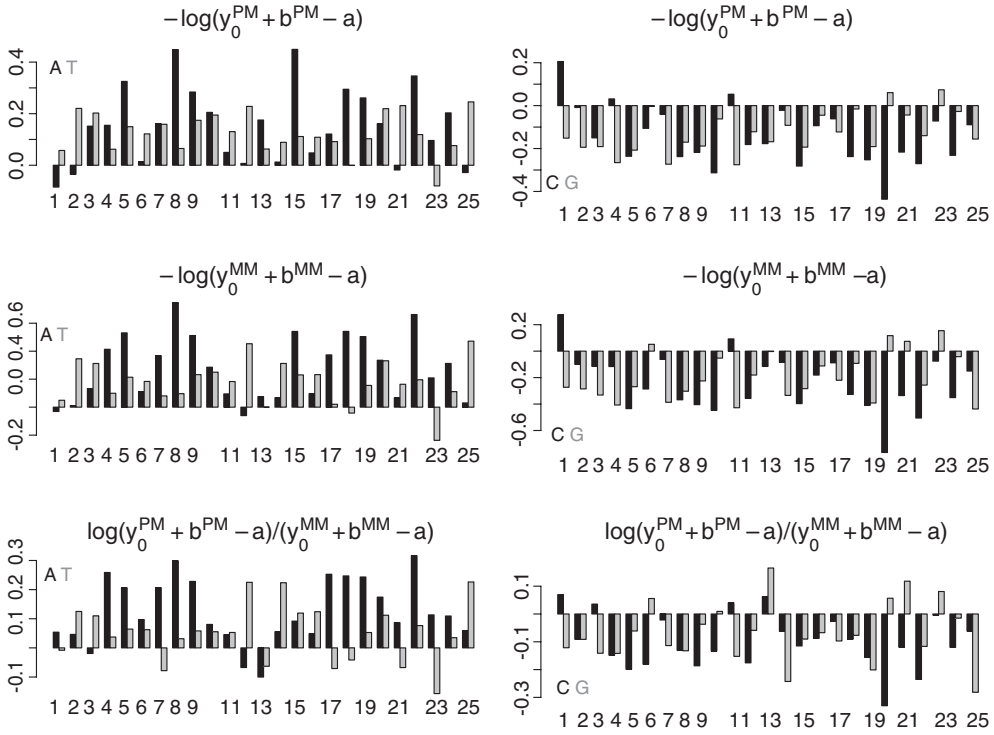

Figure 4. Bar charts of estimates of $\kappa^{PM}t_W$ (top), $\kappa^{MM}t_W$ (middle) and $(\kappa^{MM}-\kappa^{PM})t_W$ (bottom) from equation (23) averaged over DNA probe sequences with base A, T, (black, grey respectively, left-hand plots), C or G (black, grey respectively, right-hand plots) at each position along the PM probe sequence minus the corresponding averages over all probe sequences.

($i = 1, \ldots, 25$) minus estimated washing rates averaged over all PM (MM) probes. As expected, washing rates are generally lower than average for strong hydrogen bonded bases C and G occurring in the DNA probe sequences and higher than average for A and T. This is the case for both PM and MM probes. Interestingly, with the exception of the mismatched central MM base, there seems to be no obvious relationship between the strength of the effect and position along the probe.

The remaining two bar charts show the analogous contrasts for the difference $(\kappa^{MM} - \kappa^{PM})t_W$. The estimate of this quantity, determined from equation (23), is independent of b_S, and so conclusions drawn from from this bar chart do not rely on the assumption that b_S is uniform from one feature to another. Here the effect of the mismatched base at position 13 is quite noticeable: removing a triple hydrogen bond (C≡G) raises the washing rate more than removing a double hydrogen bond (A=U) or (T=A). Conversely, the effect of a central mismatch on the washing rate is almost always greater when any of the remaining 24 bases is a weakly bound A or T than a strongly bound C or G. This is entirely in keeping with the washing scenario.

7. Summary and conclusions

An understanding of the physical processes driving hybridization is essential if the design of expression measures is to advance to a point where target concentration can be measured in

absolute terms. The aim of this paper has been to gain an improved understanding of the physics of oligonucleotide microarrays by exploiting the observed differences in the responses of PM and MM features to known cRNA target concentrations. The starting point of this paper is an adsorption model of hybridization at the surface of oligonucleotide microarrays based on models proposed independently by Hekstra *et al* [12] and Halperin *et al* [11]. Though arrived at from different approaches the Hekstra and Halperin models are essentially equivalent, and are an improvement on their predecessors in that they allow for the presence of cross-hybridization from non-specific targets.

We have mainly concentrated on seeking to explain the commonly observed difference between fluorescence intensity measurements from a neighbouring PM/MM pair of features at high specific target concentration. That is, if a sufficiently high concentration of PM-specific RNA target is spiked in to the target solution, both the PM and MM fluorescence intensity signals will reach an asymptote, but the MM asymptote is almost invariably observed to be lower than the PM asymptote. Our starting Hekstra/Halperin model incorrectly predicts 100% coverage with PM-specific duplexes of both PM and MM features under these conditions, which in turn incorrectly implies that the asymptotic PM and MM fluorescence signals will be equal.

We have sought to resolve this discrepancy, first by taking a more detailed look at the hybridization step, and second by examining the subsequent washing step. In general, we find that more detailed variants of our starting model of the hybridization step, many of which have been independently suggested or alluded to previously, are unable to resolve the problem. Given our previous analysis of data from the Affymetrix Latin Square spike-in experiment [6], we are able to dismiss the Sips isotherm and non-equilibrium models of hybridization including multi-step models which take into account a slow initiation step followed by a rapid zipping up. We are also able to dismiss the effects of electrostatic screening at the microarray surface and bulk target–target hybridization as a possible explanation of differential PM/MM intensity measurements at saturation.

We are as yet unable to dismiss entirely the possibility that competitive hybridization from probe–probe duplexes at the microarray surface renders a fraction of DNA probes unavailable to target molecules, as suggested by Forman *et al* [10]. To make progress with this problem, one needs to carry out a numerical simulation of a dimer-like statistical mechanics model on a two-dimensional random lattice, probably by Monte Carlo methods. Analysis of the equivalent one-dimensional model suggests that this form of competitive hybridization could well have a measurable quantitative effect on the equilibrium adsorption isotherm, though the two-dimensional case is unlikely to lead to the observed hyperbolic response curve.

By comparison, we find that the post-hybridization washing step is able to provide a promising and straightforward explanation for the PM/MM difference at saturation. We have considered a scenario in which the equilibrium state predicted by our starting Hekstra/Halperin model is attained by the end of the hybridization step, following which the washing phase dissociates a fraction of bound duplexes. The portion of both the PM and MM signals above background decays exponentially during the washing phase, but since the MM binding affinity is less than that for PM features, the decay rate is faster for MM features. The results of our analysis of the dependence of inferred washing rates on probe base sequences support this scenario. The advantages of this model are that it preserves the observed hyperbolic shape of the Langmuir isotherm and that it explains both the partial (i.e. $<100\%$) coverage of each feature by duplexes at saturation spike-in concentrations and the fact that the MM feature almost invariably asymptotes to a lower measured fluorescence intensity than its PM partner.

The analysis presented in this paper argues that the solution to providing a practical method of estimating absolute concentration of target mRNA from microarray data lies in

understanding the physics of hybridization and washing at the microarray surface. Ideally one would like to be able to estimate isotherm parameters from probe sequence information and physical parameters including microarray design parameters, hybridization temperatures and washing times. It is hoped that theoretical analysis can serve as a guide to the design of experimental work. In particular, the results set out in this paper illustrate a strong need for further spike-in experiments carried out with varying washing times or continuous monitoring of fluorescence intensities during the washing step.

Acknowledgments

We thank Rodney Baxter and Andrew James for helpful discussions and anonymous referees for helpful input. This research was partially supported by Australian Research Council Discovery Grant DP0343727.

Appendix A. Statistical comparison of Langmuir and Sips isotherms

In this appendix we carry out a statistical analysis of fits to the Langmuir isotherm, equation (2), and the Sips isotherm

$$y = y_0 + b\frac{x^\gamma}{x^\gamma + K^\gamma}, \tag{A.1}$$

to determine which model is the better fit to the MM data of the Affymetrix spike-in experiment. The method used is described in detail in an earlier paper which compares fits of the PM data to a number of isotherm models [6].

The stochastic component of the fluorescence intensity y is assumed to be drawn from a gamma distribution. The data are fitted using the generalized linear model formalism as defined in [19], in which the negative log likelihood of the fit, or deviance, is minimized over the parameters y_0, b, K and, in the case of the Sips isotherm, also γ. To compare fits to the Langmuir and Sips models with r_L and r_S residual degrees of freedom and deviances D_L and D_S, respectively, we use the scaled deviance

$$\Delta D_{\text{scaled}} = (D_L - D_S)\frac{r_S}{D_S}. \tag{A.2}$$

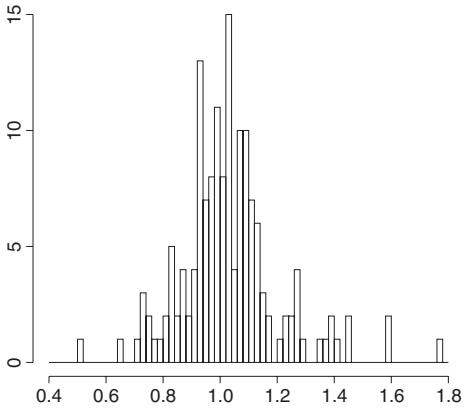

Figure A.1. Histogram of fitted values of the Sips parameter γ for the MM data.

Table A.1. Comparisons of fits to Langmuir and Sips isotherms. Δr is the decrease in residual degrees of freedom for each gene and ΔD_{scaled} is the corresponding scaled decrease in deviance from equation (A.2).

Gene	Δr	ΔD_{scaled}	Omitted probes
37777_at	14	6.43	3, 9
684_at	12	3.62	3, 5, 7, 8
1597_at	12	14.56	9, 11, 14, 15
38734_at	9	10.11	1, 3, 4, 9, 11, 12, 6
39058_at	5	11.34	1, 2, 3, 5, 6, 7, 9, 10, 12, 14, 16
36311_at	13	3.46	7, 8, 14
1024_at	16	15.19	
36202_at	15	6.18	6
36085_at	15	7.29	13
40322_at	16	39.58	
1091_at	14	3.83	1, 2
1708_at	12	2.36	11, 12, 13, 14
All genes	153	133.80	

Note that $r_{\text{L}} > r_{\text{S}} \gg 1$. To evaluate the null hypothesis, $\gamma = 1$, ΔD_{scaled} can be compared with a chi-squared distribution with $\Delta r = r_{\text{L}} - r_{\text{S}}$ degrees of freedom [19].

We were able to obtain fits with positive parameter values to both the Langmuir and Sips isotherms for about 80% of the probes. For most of the remaining cases the MM response was too small to provide a useful fit. Results for the scaled deviance are shown in table A.1. The total deviance of 133.8 lies at the 13th percentile of a chi-squared distribution with 153 degrees of freedom, showing no reason to consider a more complex model that the Langmuir isotherm. Finally, a histogram of the fitted values of the Sips parameter, figure A.1, shows that the Sips parameter is symmetrically distributed about $\gamma = 1$, as expected if the Langmuir isotherm is the more accurate model.

Appendix B. Quasi-equilibrium model with nucleation

We consider the hybridization model illustrated in figure B.1 in which the forward, duplex-forming, reaction involves two steps: a slow rate determining step in which the first two or three base pairs form, following a fast zipping-up step in which the remaining base pairs form. The probe and target molecules are denoted by P and T, respectively, the partially formed duplex after the rate determining step by P.T*, and the completed target–probe duplex by P.T. For simplicity we consider the case without cross-hybridization.

Let the target concentration be x, the fraction of probes in a feature which have formed a fully zipped up duplex P.T be θ and the fraction which have formed an initiated duplex P.T * be ζ. The remaining fraction of free single strand probes is defined as $\chi = 1 - \theta - \zeta$. The chemical rate equations are

$$\frac{\mathrm{d}\chi}{\mathrm{d}t} = -k_1 x \chi + k_{-1}\zeta, \tag{B.1}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = k_2\zeta - k_{-2}\theta. \tag{B.2}$$

The reaction rates $k_1 x$ and k_{-2} are assumed to be slow (on the order of hours) and the rates k_{-1} and k_2 fast. Accordingly we define

$$k_1 = \epsilon\kappa_1, \qquad k_{-2} = \epsilon\kappa_{-2}, \qquad \zeta = \epsilon\hat{\zeta}, \tag{B.3}$$

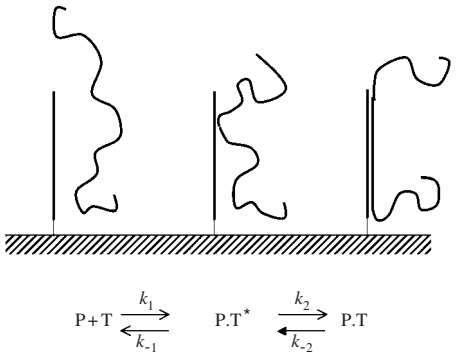

Figure B.1. Hybridization proceeding from probe plus target (P + T) to partially formed duplex in which two or three bases pair (P.T*) to a zipped-up duplex (P.T). k_1, k_{-1}, k_2 and k_{-2} are chemical reaction rates.

where $\epsilon \ll 1$. This gives

$$\frac{\mathrm{d}\chi}{\mathrm{d}\tau} = -\kappa_1 x\chi + k_{-1}\hat{\zeta}, \tag{B.4}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = k_2\hat{\zeta} - \kappa_{-2}\theta, \tag{B.5}$$

where $\tau = \epsilon t$ is O(1) on timescales of the slow nucleation reactions. We solve these equations to zeroth order in ϵ, subject to the constraints $\theta + \chi = 1 + \mathrm{O}(\epsilon)$ and $\mathrm{d}\theta/\mathrm{d}\tau = -\mathrm{d}\chi/\mathrm{d}\tau + \mathrm{O}(\epsilon)$. Eliminating $\hat{\zeta}$ and χ with the help of the constraints gives

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = \frac{\kappa_1 k_2 x}{k_{-1} + k_2} - \frac{\kappa_1 k_2 x + \kappa_{-2} k_{-1}}{k_{-1} + k_2}\theta + \mathrm{O}(\epsilon). \tag{B.6}$$

The solution to zeroth order, with initial condition $\theta(0) = 0$, is

$$\theta(t) = \frac{k_1 k_2 x}{k_1 k_2 x + k_{-1} k_{-2}} \left[1 - \mathrm{e}^{-(k_1 k_2 x + k_{-1} k_{-2})t/(k_{-1} + k_2)}\right], \tag{B.7}$$

after reinstating the original variables. This is of the form of equation (15) where $K_\mathrm{S} = k_{-1}k_{-2}/(k_1 k_2)$ and $k_\mathrm{f} = k_1 k_2/(k_{-1} + k_2)$.

Appendix C. Equilibrium model with competition between probe–target and probe–probe duplexes

We consider here the equilibrium thermodynamics of the microarray surface when pairwise interactions between neighbouring probes and self-interaction of individual probes are taken into account. The formation of probe–probe duplexes or folded probes will render a fraction of the probes unavailable for RNA target hybridization.

For a given feature, we define M to be the total number of probe sites on that feature, N to be the number of probe–target duplexes, P to be the number of probe–probe duplexes and Q to be the number of self-interacting (i.e. folded) probes. In this appendix we will for simplicity ignore hybridization of non-specific targets and partial zippering. The number of configurations consistent with the above partitioning is

$$g(M, N, P, Q) = \frac{\nu(P, M)(M - 2P)!}{N!Q!(M - N - 2P - Q)!}, \tag{C.1}$$

where $\nu(P, M)$ is the number of ways of forming P neighbouring pair duplexes on an array of M sites, where $0 \leqslant 2P \leqslant M$. The contribution to the canonical partition function from the entire feature is

$$\mathrm{e}^{-M\hat{\gamma}/k_\mathrm{B}T} = g(M, N, P, Q) \times \exp\left(\frac{1}{k_\mathrm{B}T}\left[\hat{\mu}^0_{\mathrm{pt}}N + \hat{\mu}^0_{\mathrm{pp}}P + \hat{\mu}^0_{\mathrm{q}}Q + \hat{\mu}^0_{\mathrm{p}}(M - N - 2P - Q)\right]\right), \tag{C.2}$$

where $\hat{\gamma}$ is the free energy per site, $\hat{\mu}^0_{\mathrm{pt}}$, $\hat{\mu}^0_{\mathrm{pp}}$, $\hat{\mu}^0_{\mathrm{q}}$ and $\hat{\mu}^0_{\mathrm{p}}$ are reference state chemical potentials per site of a probe–target duplex, probe–probe duplex, self-interacting probe and unmatched probe, respectively, and k_B is Boltzmann's constant.

For illustrative purposes we begin with an analysis of the relatively easily solved one-dimensional model. For a one-dimensional lattice in which nearest-neighbour sites may form duplexes, one easily obtains $\nu(P, M) = (M - P)!/[P!(M - 2P)!]$, and hence

$$g(M, N, P, Q) = \frac{(M - P)!}{N!P!Q!(M - N - 2P - Q)!}. \tag{C.3}$$

Applying the Stirling approximation $\log N! = N \ln N - N + \mathrm{O}(\ln N)$ and setting

$$\theta = \frac{N}{M} = \text{fraction of feature covered by P–T duplexes}$$
$$\zeta = \frac{2P}{M} = \text{fraction of feature covered by P–P duplexes}$$
$$\xi = \frac{Q}{M} = \text{fraction of feature covered self interacting probes}$$
$$\gamma = \frac{R}{k_\mathrm{B}}\hat{\gamma} = \text{surface free energy per mole of probe sites}$$

gives, in the bulk limit $M \to \infty$,

$$\gamma = RT\Big[-(1 - \tfrac{1}{2}\zeta)\ln(1 - \tfrac{1}{2}\zeta) + \tfrac{1}{2}\zeta \ln \tfrac{1}{2}\zeta + \xi \ln \xi + \theta \ln \theta + (1 - \theta - \zeta - \xi)\ln(1 - \theta - \zeta - \xi)\Big] + \theta\mu^0_{\mathrm{pt}} + \tfrac{1}{2}\zeta\mu^0_{\mathrm{pp}} + \xi\mu^0_{\mathrm{q}} + (1 - \theta - \zeta - \xi)\mu^0_{\mathrm{p}}, \tag{C.4}$$

where μ^0_{pt}, μ^0_{pp}, μ^0_{q} and μ^0_{p} are reference state chemical potentials per mole and R is the gas constant.

The equilibrium isotherm is obtained by balancing exchange chemical potentials for P–T duplexes with the chemical potential of the target species in solution and setting the chemical potentials for P–P duplexes and self-interacting probes to zero, that is

$$\frac{\partial\gamma}{\partial\theta} = \mu_\mathrm{t}, \qquad \frac{\partial\gamma}{\partial\zeta} = 0, \qquad \frac{\partial\gamma}{\partial\xi} = 0, \tag{C.5}$$

where μ_t is given by equation (4). This leads to

$$\theta = (1 - \zeta - \xi)\frac{x}{x + K_\mathrm{S}}, \tag{C.6}$$

$$\tfrac{1}{2}\zeta(1 - \tfrac{1}{2}\zeta) = K_\mathrm{P}(1 - \theta - \zeta - \xi)^2, \tag{C.7}$$

and

$$\xi = \frac{1 - \theta - \zeta}{1 + K_\mathrm{Q}^{-1}}, \tag{C.8}$$

where the equilibrium constants K_S for the P–T duplex forming reaction, K_P for the P–P duplex forming reaction and K_Q for the self-interaction are

$$K_S = x_0\, e^{\Delta G/RT}, \qquad K_P = e^{-\Delta G_P/RT}, \qquad K_Q = e^{-\Delta G_Q/RT}, \tag{C.9}$$

where

$$\Delta G = \mu^0_{pt} - \mu^0_p - \mu^0_t, \qquad \Delta G_P = \mu^0_{pp} - 2\mu^0_p, \qquad \Delta G_Q = \mu^0_q - \mu^0_p. \tag{C.10}$$

Eliminating ξ gives

$$\theta = (1-\zeta)\frac{x}{x + K'_S}, \tag{C.11}$$

$$\tfrac{1}{2}\zeta(1 - \tfrac{1}{2}\zeta) = K'_P(1-\theta-\zeta)^2, \tag{C.12}$$

where $K'_S = (1+K_Q)K_S$ and $K'_P = K_P/(1+K_Q)^2$. That is, the effect of probe self-interaction is to rescale the remaining equilibrium constants.

From equation (C.12) one finds that the P–P coverage fraction ζ decreases smoothly from a maximum value $\zeta_{max} = 1 - \frac{1}{2}(K'_P + \frac{1}{4})^{-1/2}$ at $\theta = 0$ to zero at $\theta = 1$. This has two consequences. First, there is no phase transition, as expected for a one-dimensional model with local interactions. Second, we see from equation (C.11) that θ asymptotes to 1 in the limit of high target concentration $x \to \infty$. Thus the simple one-dimensional model of P–P duplexes is unable to explain partial saturation of the feature at high concentration. A plot of θ against target concentration for a range of values of K'_P is given in figure 2.

Ideally we need to solve the model defined by equation (C.2) for a random two-dimensional lattice. The presence of self-interactions of individual probes involves no interaction between sites and consequently cannot complicate the phase structure. In fact, by comparing equations (3) and (C.4) we see that, for the purposes of determining phase structure, probe self-interactions and non-specific hybridization are mathematically identical problems. Nearest-neighbour probe–probe interactions, on the other hand, are less tractable. In this case one needs to calculate $\nu(P, M)$ for a random two-dimensional lattice with some reasonable definition of 'neighbouring'. To analyse the bulk limit, it can be shown that one only needs $(1/M)\log\nu(P, M)$ in the limit $M,\ P \to \infty$ for given fixed $2P/M$. This is the random lattice analogue of the monomer–dimer model which is usually defined on a regular two-dimensional lattice, and for which no exact solution has been found. For a square lattice, though, numerical calculations strongly suggest the model has no phase transition at non-zero monomer density [2]. (At zero monomer density, that is $2P = M$, the square lattice monomer–dimer model is critical, corresponding to the critical point of the Ising model [16].)

A review of most of the two-dimensional statistical models which have been solved exactly can be found in [3]. These include the close packed dimer model on a square lattice, which is equivalent to calculating $\nu(\frac{1}{2}M, M)$, and the hard hexagon model, in which sites of a triangular lattice are occupied subject to the constraint that no two neighbouring sites may be occupied simultaneously. The model we are interested in is similar in some ways to the hard hexagon model, except that in our case links of a lattice are occupied subject to the constraint that no two adjoining links may be simultaneously occupied. The hard hexagon model does undergo a phase transition between a liquid phase (uncorrelated positioning of hexagons at low density) and a solid phase (close packing of hexagons centred on one of three possible sublattices). Whether the random lattice duplex model relevant to the case in hand undergoes a phase transition from a disordered phase at low duplex density or high temperature to an ordered phase at high duplex density or low temperature is unknown.

References

[1] Affymetrix Inc. 2002 *Statistical Algorithms Description Document* Available at http://www.affymetrix.com/support/technical/whitepapers.affx

[2] Baxter R J 1968 Dimers on a rectangular lattice *J. Math. Phys.* **9** 650–4

[3] Baxter R J 1982 *Exactly Solved Models in Statistical Mechanics* (London: Academic)

[4] Bhanot G, Louzoun Y, Zhu J and DeLisi C 2003 The importance of thermodynamic equilibrium for high throughput gene expression arrays *Biophys. J.* **84** 124–35

[5] 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

[6] 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

[7] Cantor C R and Schimmel P R 1980 *Biophysical Chemistry, Part 3: The Behaviour of Biological Macromolecules* 1st edn (San Francisco, CA: Freeman)

[8] Dai H, Meyer M, Stepaniants S, Ziman M and Stoughton R 2002 Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays *Nucleic Acids Res.* **30** e86

[9] Deutsch J M, Liang S and Narayan O 2004 Modeling of microarray data with zippering *Preprint* q–bio.BM/0406039

[10] 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 synthesized oligonucleotide arrays *Molecular Modeling of Nucleic Acids* (*ACS Symp. Series* vol 682) ed N B Leontis and J SantaLucia (Washington, DC: American Chemical Society) pp 206–28

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

[12] 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

[13] 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

[14] Hill T L 1960 *An Introduction to Statistical Thermodynamics* (Reading, MA: Addison-Wesley)

[15] Irizarry R A *et al* 2003 Exploration, normalization and summaries of high density oligonucleotide array probe level data *Biostatistics* **4** 249–64

[16] Kasteleyn P W 1963 Dimer statistics and phase transitions *J. Math. Phys.* **4** 287–93

[17] Lemon W J, Liyanarachchi S and You M 2003 A high performance test of differential gene expression for oligonucleotide arrays *Genome Biol.* **4** R67.1–R67.11

[18] Levicky R and Horgan A 2005 Physicochemical perspectives on DNA microarray and biosensor technologies *Trends Biotechnol.* **23** 143–9

[19] McCullagh P and Nelder J A 1989 *Generalized Linear Models* 2nd edn (London: Chapman and Hall)

[20] Nelson B P, Grimsrud T E, Liles M R, Goodman R M and Corn R M 2001 Surface plasmon resonance imaging measurements of DNA and RNA hybridization adsorption onto DNA microarrays *Anal. Chem.* **73** 1–7

[21] Nguyen D V, Arpat A B, Wang N and Carroll R J 2002 DNA microarray experiments: Biological and technological aspects *Biometrics* **58** 701–17

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

[23] Peterson A W, Wolf L K and Georgiadis R M 2002 Hybridization of mismatched or partially matched DNA at surfaces *J. Am. Chem. Soc.* **124** 14601–7

[24] Pirrung M C 2002 How to make a DNA chip *Angew. Chem. Int. Edn* **41** 1276–89

[25] Sips R 1948 On the structure of a catalyst surface *J. Chem. Phys.* **16** 490–5

[26] Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H and Ohmichi T 1995 Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes *Biochemistry* **34** 11211–6

[27] Vainrub A and Pettitt M 2002 Coulomb blockage of hybridisation in two-dimensional DNA arrays *Phys. Rev.* E **66** 041905

[28] Wang S, Friedman A E and Kool E T 1995 Origins of high-sequence selectivity: a stopped-flow kinetics study of DNA/RNA hybridization by duplex- and triplex-forming oligonucleotides *Biochemistry* **34** 9774–84

[29] Zhang L 2003 Modeling sequence dependence of probe signals on oligonucleotide microarrays *Talk at the Affymetrix Genechip Microarray Low Level Workshop (Berkely, CA, August 2003)* available at http://affymetrix.com/corporate/events/scientific-archive-2003.affx