

Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed

Conrad J. Burden¹

¹Centre for Bioinformation Science

John Curtin School of Medical Research and Mathematical Sciences Institute
Australian National University, Canberra, ACT 0200, Australia

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

Abstract. The Affymetrix U95 and U133 Latin Square spike-in datasets are reanalysed, together with a dataset from a version of the U95 spike-in experiment without a complex non-specific background. The approach uses a physico-chemical model which includes the effects the specific and non-specific hybridisation and probe folding at the microarray surface, target folding and hybridisation in the bulk RNA target solution, and duplex dissociation during the post-hybridisation washing phase. The model predicts a three parameter hyperbolic response function that fits well with fluorescence intensity data from all three datasets. The importance of the various hybridisation and washing effects in determining each of the three parameters is examined, and some guidance is given as to how a practical algorithm for determining specific target concentrations might be developed.

PACS numbers: 87.15.-v, 82.39.Pj

1. Introduction

A number of papers [15, 16, 7, 4, 6, 10, 14, 8, 17] have used chemical adsorption models to analyse data from two well-known Affymetrix Latin-Square spike-in experiments [1], known as the U95 and U133 datasets. The immediate aim of these papers has been to study the physical processes responsible for converting concentrations of specific target RNA in prepared solutions hybridised onto microarrays to measured fluorescence intensities. Their ultimate aim has been to provide biologists with a practical algorithm for estimating absolute specific target concentrations in the presence of a complex non-specific background from fluorescence intensity data. Even though there are a number of existing algorithms (or “expression measures”) of varying degrees of statistical sophistication currently available to and widely used by biologists, such algorithms pay little attention to the detailed physics and chemistry of hybridisation at the microarray surface. As a result they are prone to underestimating fold changes at high target concentrations because saturation effects are not modelled, and at low target concentrations because of a failure to account adequately for non-specific hybridisation [9]. Furthermore, the measures reported are not target concentrations as such, but surrogates derived directly from fluorescence intensities. At best, they could be described as an unknown increasing function of specific target concentrations, modulo statistical noise.

Before a reliable algorithm for inferring target concentrations can be developed, an accurate model of the physics and chemistry of the process is needed. In a recent review of chemical adsorption effects at the microarray surface [2], Binder has described in detail a number of processes influencing fluorescence intensity measurements, including bulk dimerisation of target molecules in solution, non-specific hybridisation and probe folding at the microarray surface, and partial zippering of duplexes. Analyses of the Affymetrix spike-in data have provided strong evidence that these effects cannot be ignored. For instance, Binder and Preibisch [6] have isolated the effects of specific and non-specific binding, Carlon and Heim[10] and Heim et al. [14] have stressed the importance of bulk hybridisation and target folding in solution, and by analysing other data sets, Matveeva et al.[19] have produced evidence for the importance of probe folding.

More recently, analyses of the U95 dataset [8, 17] and subsequent experimental evidence [21] have demonstrated the significant influence of the post-hybridisation washing step in determining fluorescence intensities. Although the washing step has an important scaling effect over the whole range of target concentrations, it is particularly noticeable as a probe sequence dependent scaling of the asymptote of the measured intensity isotherm at saturation concentrations. Because adsorption theories of the hybridisation step which neglect the post-hybridisation washing predict a common asymptote for these isotherms, many adsorption-model-based analyses of the Affymetrix Latin Square data sets have forced the data to fit a common asymptote [16, 10, 14, 6], in spite of strong statistical evidence to the contrary [7, 15].

In this paper we reanalyse both the U95 and U133 datasets, and a third dataset

which is analogous to the U95 dataset, but without the complex background. Our analysis uses a chemical adsorption model which includes a number of chemical reactions occurring during the hybridisation and post-hybridisation washing steps. Details of the datasets are summarised in Section 2, and all three datasets are shown to be consistent with the empirical observations previously observed for the U95 dataset, namely that measured fluorescence intensities follow a hyperbolic isotherm with probe-sequence dependent parameters [7]. Our physico-chemical model is described in Section 3 and demonstrated to be quantitatively consistent with these observations in Section 4. A quantitative analysis of how well the model fits the three datasets is given in Section 5. An analysis in Section 6 gives some indication of how intensity measurements over a whole microarray might be used to infer some of the isotherm parameters for each feature on the microarray, which in turn has the potential to contribute towards development of a practical algorithm for estimating target concentrations. Concluding comments are given in Section 7.

2. Datasets and empirical fits

Three datasets are analysed in this paper, the two publicly available Affymetrix Latin-Square spike-in experiments [1], known as the U95 and U133 datasets, and a version of the U95 dataset without the complex human pancreas background, kindly made available to us by Affymetrix. We will refer to these datasets as numbers I, II and III respectively.

In the manufacture of Affymetrix GeneChip® 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 nanometres apart. A microarray chip surface is divided into some hundreds of thousands of regions called features, commonly 11 to 20 microns square, and with the single strand DNA probes within each feature being synthesized to a specific nucleotide sequence.

The main step in the laboratory process of gene detection with microarrays is the hybridization of cRNA target molecules, fractionated to lengths of typically 50 to 200 bases and with biotin labels attached to their U and C bases, onto the single strand DNA probes. The hybridisation is performed at 45°C for 16 hours. The microarray is then washed to remove excess cRNA target, the biotin labels stained with fluorescent dye, and the density of hybridized probe-target duplexes in each feature detected via intensity measurements of the fluorescent dye. Each gene or EST is represented by a set of pairs of features (16 pairs in the case of the U95 dataset and 11 pairs for U133) 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.

In the Affymetrix spike-in experiments, 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 (dataset I), human adenocarcinoma cell line (dataset II), or no background (dataset III). Each of datasets I and III consist of PM and MM fluorescence intensity values from a set of 14 probe sets corresponding to 14 separate genes, each containing 16 probe pairs. For each probe set a set of fluorescence intensity values was obtained for the 14 spiked-in concentrations 0, 0.25, 0.5, 1, 2, \dots , 1024 pM. In common with previous analyses of dataset I, the following analysis of datasets I and III is restricted to 12 of the 14 genes, omitting data from the two defective probesets, 36889_at and 407_at. Dataset II consists of PM and MM fluorescence intensity values from a set of 38 probe sets corresponding to 38 separate genes, each containing 11 probe pairs. For each probe set a set of fluorescence intensity values was obtained for the 14 spiked-in concentrations 0, 0.125, 0.25, 0.5, 1, \dots , 512 pM. Dataset II also contains data from a further 4 bacterial gene probe sets each containing 20 probe pairs, which we do not include in the current analysis.

In a previous paper [7] it was demonstrated that the dataset I is consistent with the empirical observation that the measured fluorescence intensities $I(x)$ at spike-in concentration x are sampled from a Gamma distribution with constant coefficient of variation about a mean given by a hyperbolic response curves of the form

$$I(x) = A + B \frac{Kx}{1 + Kx}. \quad (1)$$

Importantly, it was further shown using a thorough statistical analysis that each of the parameters A , B and K is feature (i.e. probe-sequence) dependent, and that the asymptote, $\lim_{x \rightarrow \infty} I(x) = A + B$, is almost invariably lower for the MM feature than the neighbouring perfect-match PM feature within any PM/MM pair.

On the assumption that by far the greatest proportion of the intensity signal across a whole microarray in either dataset I or II comes from the complex background, we have preprocessed the data by carrying out a quantile normalisation across each of these two datasets [18]. Thus all microarrays within a particular dataset have a common distribution of fluorescence intensities after preprocessing. Because of the absence of a complex background, we have not carried out this step for dataset III. Instead we have included in the fit the “wafer dependent scaling” described as Model **E** in ref. [7] to allow for the fact that the three replicates of the experiment used microarrays constructed from three separate wafers. That is to say, we fit to the data a model of the form $I(x) = \lambda_w [A_p + B_p K_p x / (1 + K_p x)]$, where the index $w = 1, 2, 3$ labels the three replicates, the index p labels a feature within a probe set and the scaling factors satisfy $\frac{1}{3} \sum_w \lambda_w = 1$.

In Figures 1 and 2 are plotted fits of fluorescence intensity data to the response curve Eq. 1 for one of the spiked-in transcripts for datasets I and III. A complete set of analogous plots for all transcripts from all three datasets can be found at http://dayhoff.anu.edu.au/~conrad/Spike_in_Isotherms/.

For a small minority of features the fit gives negative values to some of the parameters A , B or K , whereas the physical model set out in Section 3 predicts that all

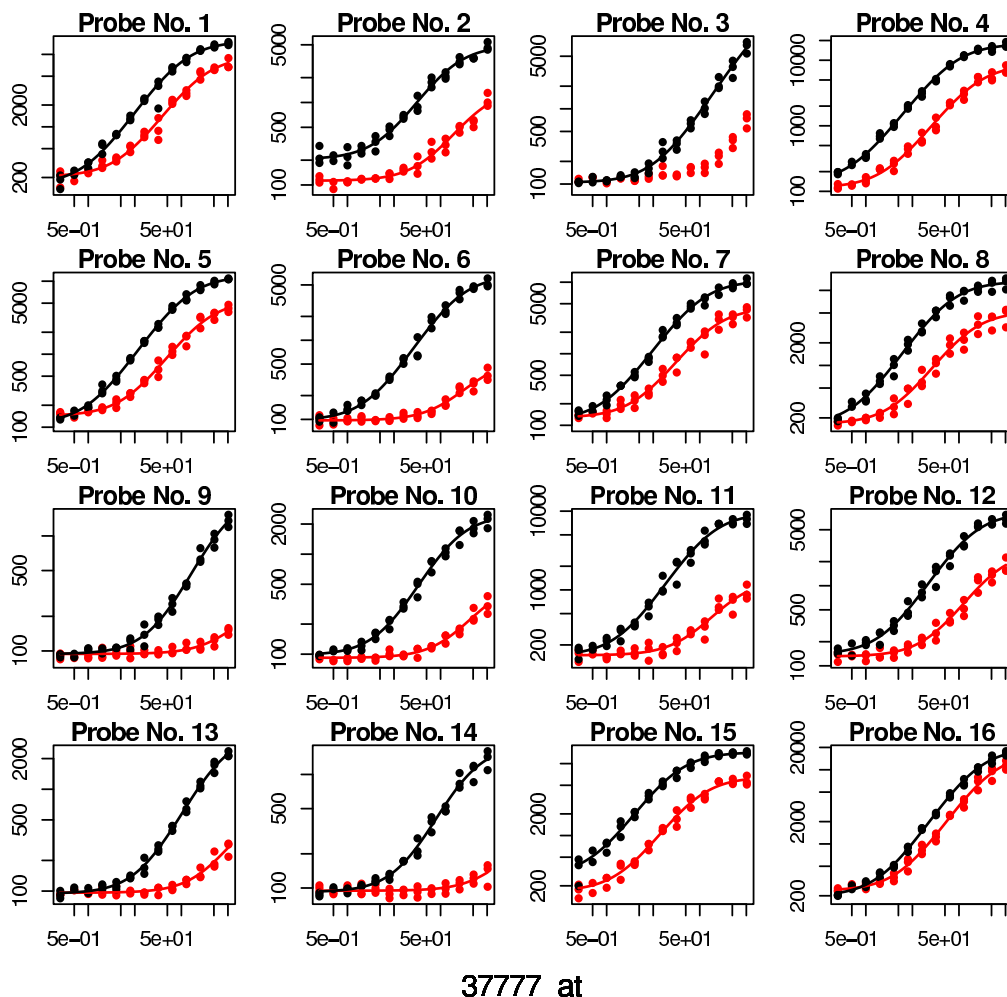


Figure 1. Fits of fluorescence intensity data to the hyperbolic response curve Eq. 1 for the spiked-in transcript `37777_at` for dataset I. PM data are in black and MM data in red. Spike-in concentrations (horizontal axes) in picomolar on a logarithmic scale, and fluorescence intensities after quantile normalisation (vertical axes) are in the arbitrary units used in Affymetrix cel files on a logarithmic scale.

three parameter should be positive. This problem is more slightly more prevalent for MM features than for PM features, and is most acute for dataset II. In some cases, such as probe number 3 in Fig. 1, it appears that the data does not extend far enough into the high concentration, i.e. saturation, limit to allow an acceptable fit. In other cases, such as probe number 9 of Fig. 2, the data may be too noisy. The range of spike-in concentrations used in Dataset II is shifted downwards from that of datasets I and III, and as such may not be adequately probing the saturation region to give acceptable fits in all cases. Furthermore, the spike-in concentrations at the lower end may be probing the regime in which the target concentration is depleted during the hybridisation step, which is beyond the applicability of the model leading to Eq. 1 which we describe below. The analyses in Sections 4 and 5 below are restricted to the subset of fits to Eq. 1 for which all three parameters A , B and K are positive. Table 1 gives the coefficient of

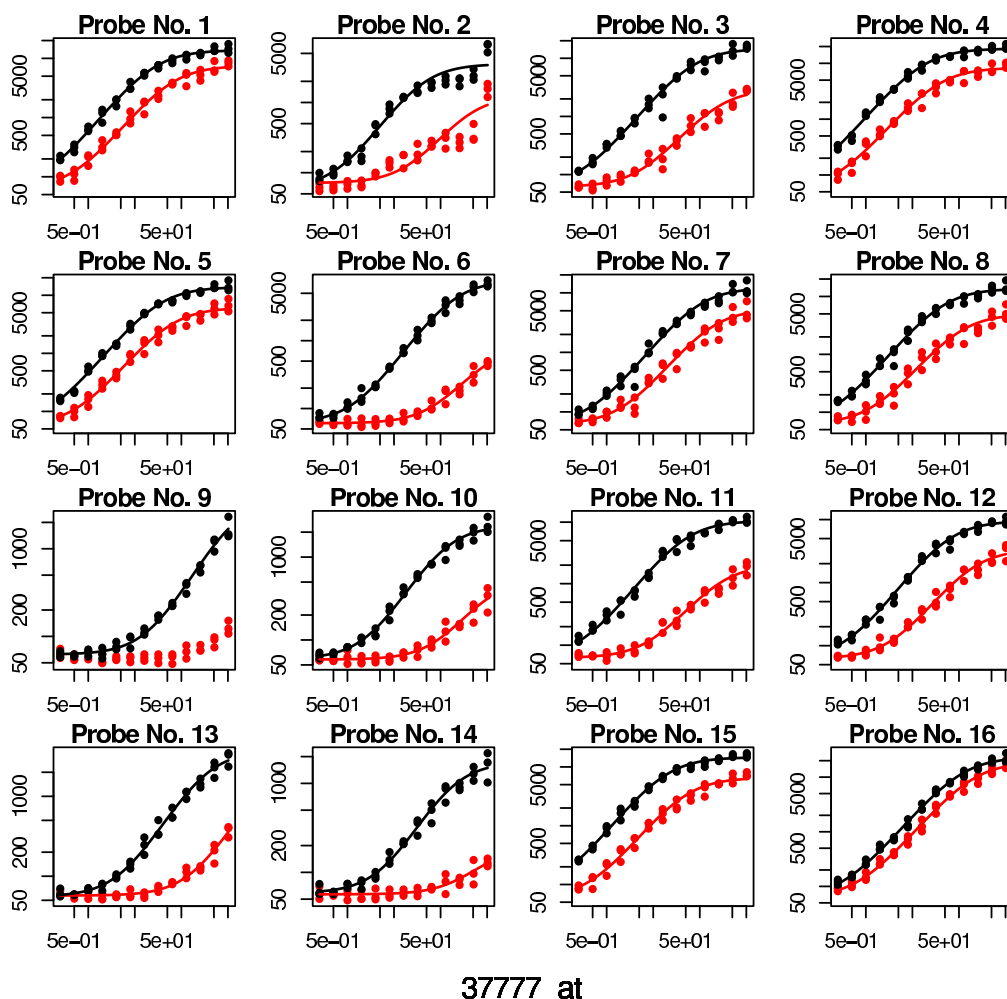


Figure 2. Fits of fluorescence intensity data to the hyperbolic response curve Eq. 1 for the spiked-in transcript *37777_at* for dataset III, without the complex human pancreas background.

variation of the fitted data for each dataset, and the percentage of features for which an acceptable fit with three positive parameters to the hyperbolic response function was obtained. In general, agreement with a hyperbolic response curve with positive parameters is excellent for datasets I and III, and reasonable for dataset II.

3. The model

Consider the response of a given feature to a spike-in concentration x of a particular RNA transcript in the presence of an unknown complex background of non-specific target RNA. We write the measured fluorescence intensity $I(x)$ in the form

$$I(x) = a + b\theta(x), \quad (2)$$

where a is a physical background due to effects unrelated to fluorescent label carrying duplexes, such as reflection from the glass surface of the microarray, and b is the

Table 1. Coefficient of variation of the data, assumed to be distributed from a Gamma distribution with constant coefficient of variation within any one dataset about a mean given by Eq. 1. The two right hand columns give the percentage of probesets which for which the fit gives positive values to all three parameters A , B and K .

Dataset	Coef. of variation	% of accepted fits	
		PM	MM
I	0.12	97.9	91.6
II	0.14	72.5	37.5
III	0.17	100	98.4

Table 2. Chemical species present in the model.

S	Single strand PM-feature-specific RNA target in solution
NS_i	Single strand non-specific RNA target in solution of species i
$S.NS_i$	Bound RNA/RNA duplex in solution unavailable for hybridisation
S'	Folded specific target in solution rendered unavailable for hybridisation
P	Unbound probe at the microarray surface
$P.S$	Duplex formed from PM-feature-specific RNA target binding to probe
$P.NS_i$	Duplex formed from non-specific RNA target binding to probe
P'	Folded probe at the microarray surface rendered unavailable for hybridisation

maximum fluorescence intensity, that is, the contribution from fluorescent dye if all probes on the feature were occupied with labelled probe-target duplexes. It is argued in ref. [2] that the maximum intensity b should vary only weakly due to differing probe sequences. Throughout this paper we assume a and b to be fixed constants for a given microarray. The coverage fraction, $\theta(x)$, is the fraction of probes on the feature carrying probe-target duplexes at the time of scanning. It satisfies $0 \leq \theta(x) \leq 1$.

The coverage fraction is determined by a number of reactions between various chemical species. The species and reactions considered in our model are set out in Tables 2 and 3 respectively. The first five reactions in Table 3 are assumed to reach equilibrium during the hybridisation step. The rate constants K_i^{bulk} , K^{Sfold} , K^{S} , K_i^{NS} and K^{Pfold} are the ratio of the forward to backward rates for each reaction. The washing phase, which is primarily designed to remove unbound targets before scanning, also dissociates some duplexes[8, 17]. Thus the last two reactions are unidirectional as dissociated duplexes are continuously flushed out of the cartridge and replaced with a buffer solution containing no RNA.

The effect of the first two reactions, bulk hybridisation and specific target folding in solution, is to reduce the concentration of specific target available for hybridisation

Table 3. Chemical reactions occurring in the model. Rate constants, defined as the ratio of forward to backward reaction rates, are given in the right hand column.

In bulk solution	Non-specific hybridisation	$S + NS_i \rightleftharpoons S.NS_i$	K_i^{bulk}
	Specific target folding	$S \rightleftharpoons S'$	K^{Sfold}
At the microarray surface	Specific hybridisation	$P + S \rightleftharpoons P.S$	K^{S}
	Non-specific hybridisation	$P + NS_i \rightleftharpoons P.NS_i$	K_i^{NS}
	Probe folding	$P \rightleftharpoons P'$	K^{Pfold}
During the washing phase	Dissociation of specific duplexes	$P.S \rightarrow P (+S)$	
	Dissociation of non-specific duplexes	$P.NS_i \rightarrow P (+NS_i)$	

onto the microarray from its spike-in value of x to a value $[S]$, that is the concentration of single strand RNA target S in solution. (Following the usual convention, square brackets indicate concentrations.) For these reactions we follow the analysis of ref. [14]. The label i in Table 3 ranges over all possible subsequences of the 25-mer part of the specific target RNA sequence complementary to the PM probe. The species NS_i in this reaction includes any target RNA molecule containing a subsequence complementary to the i th subsequence. At equilibrium, we have

$$\frac{[S']}{[S]} = K^{\text{Sfold}}, \quad \frac{[S.NS_i]}{[S][NS_i]} = K_i^{\text{bulk}}. \quad (3)$$

The relation $x = [S] + [S'] + \sum_i [S.NS_i]$ then gives

$$[S] = \frac{x}{1 + K^{\text{Sfold}} + X^{\text{bulk}}}, \quad (4)$$

where

$$X^{\text{bulk}} = \sum_i [NS_i] K_i^{\text{bulk}}. \quad (5)$$

The next three reactions, occurring at the microarray surface, determine the duplex coverage fraction of the feature at the end of the hybridisation step, and before washing. Let the fraction of probes on the feature that have formed duplexes with either specific or non-specific target mRNA molecules and survived to a time t_W after the commencement of the washing process be $\theta(x, t_W)$. We split the fraction of probes which have formed duplexes at the end of the hybridisation step and before washing into two contributions:

$$\theta(x, 0) = \theta^{\text{S}} + \theta^{\text{NS}}. \quad (6)$$

The first contribution, $\theta^{\text{S}} \propto [P.S]$, is that due to duplexes formed with specific mRNA targets, and the second contribution, $\theta^{\text{NS}} \propto \sum_i [P.NS_i]$, is that due to duplexes which have formed with non-specific mRNA targets, the sum being over targets containing a subsequence complimentary to the i th subsequence of the probe.

Either by balancing equilibrium concentrations against chemical rate constants [4] or by considering the Gibbs distribution of states at constant chemical potential [13, 8] one obtains the isotherms

$$\theta^S = \frac{X^S}{1 + K^{\text{Pfold}} + X^S + X^{\text{NS}}} \quad (7)$$

$$\theta^{\text{NS}} = \frac{X^{\text{NS}}}{1 + K^{\text{Pfold}} + X^S + X^{\text{NS}}} \quad (8)$$

where, following the notation of ref.[2], we define

$$X^S = [S]K^S, \quad X^{\text{NS}} = \sum_i [NS_i]K_i^{\text{NS}}. \quad (9)$$

The calculation leading to these isotherms assumes negligible depletion of target molecules in bulk solution during the hybridisation process.

Note that, in the asymptotic limit of high spike-in concentration, namely $x \rightarrow \infty$ while holding $[NS_i]$ constant, Eqs. 4 to 9 imply $\theta^S \rightarrow 1$ and $\theta^{\text{NS}} \rightarrow 0$, implying that the feature becomes saturated with specific duplexes. This is contrary to the differing isotherm asymptotes observed empirically. To explain the differing asymptotes, we include in our model the final two reactions in Table 3, namely the washing step [8]. Define the probability that a given probe-target duplex has survived up to a washing time t_W to be $s^S(t_W)$ for a specific duplex and $s_i^{\text{NS}}(t_W)$ for a non-specific duplex of species i . The survival functions s^S and s_i^{NS} depend on probe and target base sequences, satisfy $s^S(0) = s_i^{\text{NS}}(0) = 1$, are positive and are monotonically decreasing. Defining an average non-specific survival function $s^{\text{NS}}(t_W)$ by

$$X^{\text{NS}}s^{\text{NS}}(t_W) = \sum_i [NS_i]K_i^{\text{NS}}s_i^{\text{NS}}(t_W), \quad (10)$$

the coverage fraction at washing time t_W is then

$$\theta(x, t_W) = \theta^S s^S(t_W) + \theta^{\text{NS}} s^{\text{NS}}(t_W). \quad (11)$$

Substituting Eqs. 7 and 8 and rearranging gives

$$\theta(x, t_W) = \frac{X^{\text{NS}}s^{\text{NS}}(t_W)}{1 + K^{\text{Pfold}} + X^{\text{NS}}} + \left(s^S(t_W) - \frac{X^{\text{NS}}s^{\text{NS}}(t_W)}{1 + K^{\text{Pfold}} + X^{\text{NS}}} \right) \frac{(1 + K^{\text{Pfold}} + X^{\text{NS}})^{-1} X^S}{1 + (1 + K^{\text{Pfold}} + X^{\text{NS}})^{-1} X^S}. \quad (12)$$

Finally, with help from Eqs. (4) and (9), and suppressing the t_W dependence, we get

$$\theta(x) = \alpha + \beta \frac{Kx}{1 + Kx}, \quad (13)$$

where

$$\alpha = \frac{X^{\text{NS}}s^{\text{NS}}}{1 + K^{\text{Pfold}} + X^{\text{NS}}}, \quad (14)$$

$$\beta = s^S - \alpha, \quad (15)$$

$$K = \frac{K^S}{(1 + K^{\text{Sfold}} + X^{\text{bulk}})(1 + K^{\text{Pfold}} + X^{\text{NS}})}. \quad (16)$$

This model, summarised by Eqs. (2) and (13) is consistent with the empirical observation of Eq. (1), with

$$A = a + b\alpha, \quad B = b\beta. \quad (17)$$

Eqs. 14 to 17 (with the help of Eqs. 5, 9 and 10) relate the empirically fitted parameters A , B and K to underlying physical quantities, namely a , b , the concentrations of chemical species in Table 2, reaction rates in Table 3 and washing survival functions s^S and s_i^{NS} .

Physical effects which have not been included in the model include target depletion, which should only manifest at extremely low target concentrations, incomplete probe synthesis during the manufacturing process [12], and probe-probe interactions. Each of these effects will, in theory, cause the response curve to deviate from a hyperbolic form. A discussion of probe-probe interactions, for instance, can be found in ref. [8]. The choice of model in this paper is guided by a desire to balance complexity of the problem with practicality.

4. Qualitative behaviour of the fits

Before considering a detailed analysis of the ability of the model to explain the parameters of the empirical fits, one can carry out a number of simple qualitative checks. The first three panels of Figure 3 compare the fitted saturation asymptotes $A + B$ for PM/MM pairs of features for each of the three datasets. Consistent with the hypothesis that a portion the bound probe-target duplexes are removed during the washing step, the asymptotes cover a broad range of values. The MM asymptote is almost always less than its PM partner, consistent with the scenario that a saturated feature of PM duplexes will lose less duplexes to washing than the partner saturated feature of less tightly bound MM duplexes [8]. The observed pattern breaks down at higher values of $A + B$, as the fits must be extrapolated further past the highest spike-in concentration to estimate the asymptote, and numerical accuracy is lost. This is most evident for dataset II, for which spike-in concentrations only extend to 512 pM, compared with 1024 pM for datasets I and III.

From Eqs. 15 and 17, the saturation asymptote is given by $I(\infty) = A + B = a + bs^S$. This depends only on the rate s^S at which specific duplexes are dissociated by the washing process, and not on the properties of non-specific duplexes. Thus the model predicts that the asymptote of the response function is unaffected by non-specific hybridisation. The fourth panel of Figure 3, which compares the asymptote for dataset I, (U95 with a complex non-specific background), with that for dataset III, (U95 without a non-specific background), confirms this. That is, the asymptote $I(\infty)$ for any feature is the same for dataset I as for dataset III to within the standard errors of the isotherm fits.

The parameter A in Eq. 1 is the baseline intensity estimate at zero spike-in concentration. From Eq. 17, it consists of a physical background component a , and

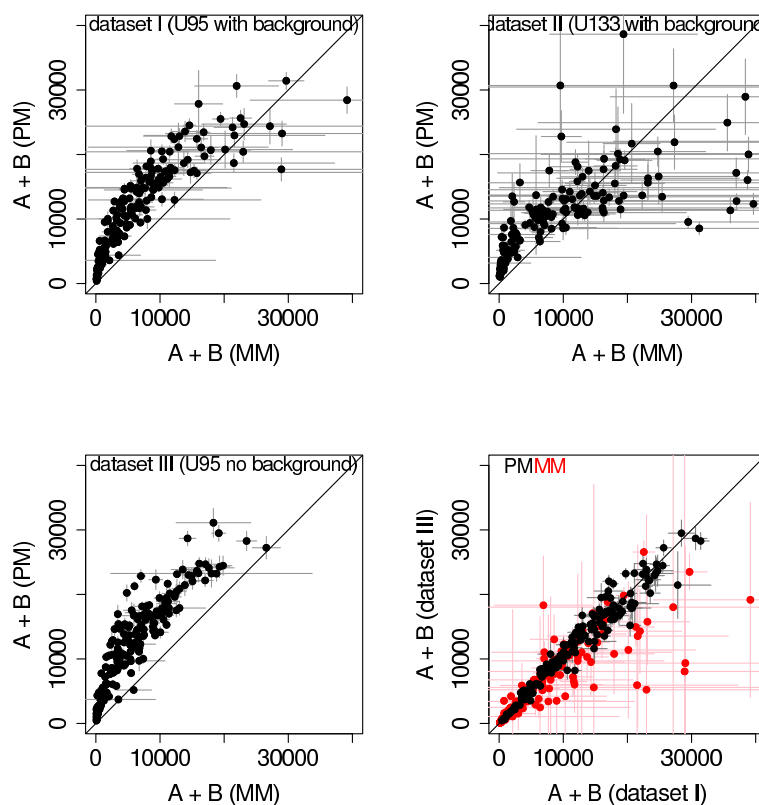


Figure 3. The first three panels show fitted asymptotes $I(\infty) = A + B$, defined in Eq. 1 for PM/MM pairs for each of the three datasets. The fourth panel compares the asymptotes for dataset I (with non-specific background) with those for dataset III (without non-specific background). Standard errors arising from the fits to Eq. 1 are also shown.

a component due to non-specific hybridisation, $b\alpha$. Consistent with this, the A values, shown in Fig. 4, are spread over a much broader range for datasets I and II in which a complex mRNA background was present than for dataset III with no background and therefore little non-specific hybridisation.

An obvious pattern, which emerges from comparing A from PM/MM pairs of features in datasets I and II, is that non-specific hybridisation is stronger for a probe whose middle base is a pyrimidine (C or T) than for its partner probe whose middle base is a purine (A or G). This effect has been noted previously for microarray intensity data generally, and there is some debate about its physical origins [20, 5, 11]. The effect is better understood at the level of individual letter frequencies. Binder et al.[4] have noted that probe sensitivities increase with C-content, and decrease with A-content, while the G- or T-content of the probe has little effect. There are probably two effects contributing to this pattern. Firstly, the averaged contributions to DNA/RNA binding energies calculated from nearest neighbour stacking models [22] are ordered as [3, 11]

$$|\Delta G_C^{\text{av}}| > |\Delta G_G^{\text{av}}| \approx |\Delta G_T^{\text{av}}| > |\Delta G_A^{\text{av}}|, \quad (18)$$

causing the substitution of a pyrimidine by a purine to decrease probe sensitivity and vice versa. Secondly, there is the simple geometric effect that pyrimidines, having a small

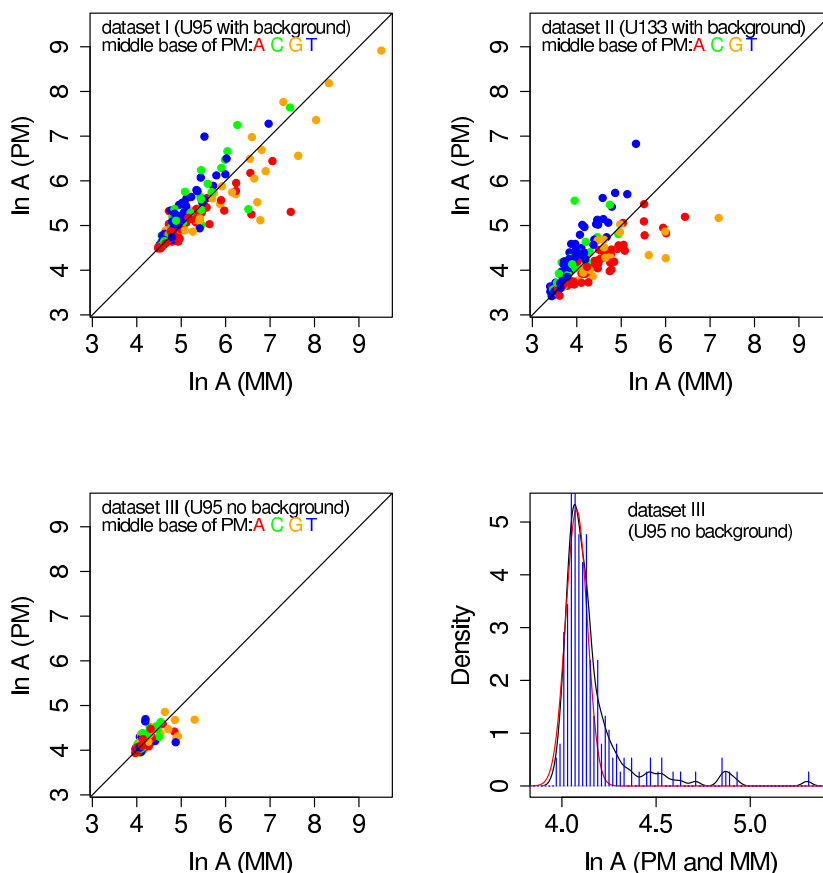


Figure 4. The log of the fitted baseline fluorescence intensities $I(0) = A$, defined in Eq. 1, for PM/MM pairs for datasets I and III. The middle (13th) base of the PM probe sequence is colour coded as indicated. The fourth panel shows a histogram of $\ln A$ (in blue), a kernel density estimate of the histogram using a gaussian kernel with a bandwidth of 0.025 (in black), and a fit of the left part of the histogram to a Gamma distribution in A with a mean of 59 and a coefficient of variation of 0.057 (in red).

single ring structure, will tolerate mismatches more easily than purines, which have a double ring structure, and would therefore need to deform the molecular backbone to bind with a target closely matching the probe sequence either side of the central base. No obvious pyrimidine/purine asymmetry is observed in the A values from dataset III. This is to be expected as the parameter A in this case is essentially the physical background without hybridisation contributions.

The fourth panel of Fig. 4 shows a histogram of A values from dataset III, for which there is no complex background present. There is very little non-specific hybridisation, and most of this data represents statistical noise in the physical background parameter a (defined in Eq. 2). Ignoring the tail, which we assume to be due to a small amount of non-specific hybridisation from the other spiked-in transcripts in the latin square protocol, we estimate a by fitting the left hand part of the histogram to a gamma distribution in the unlogged data. The fitted distribution has a mean of 59 and a coefficient of variation of 0.057.

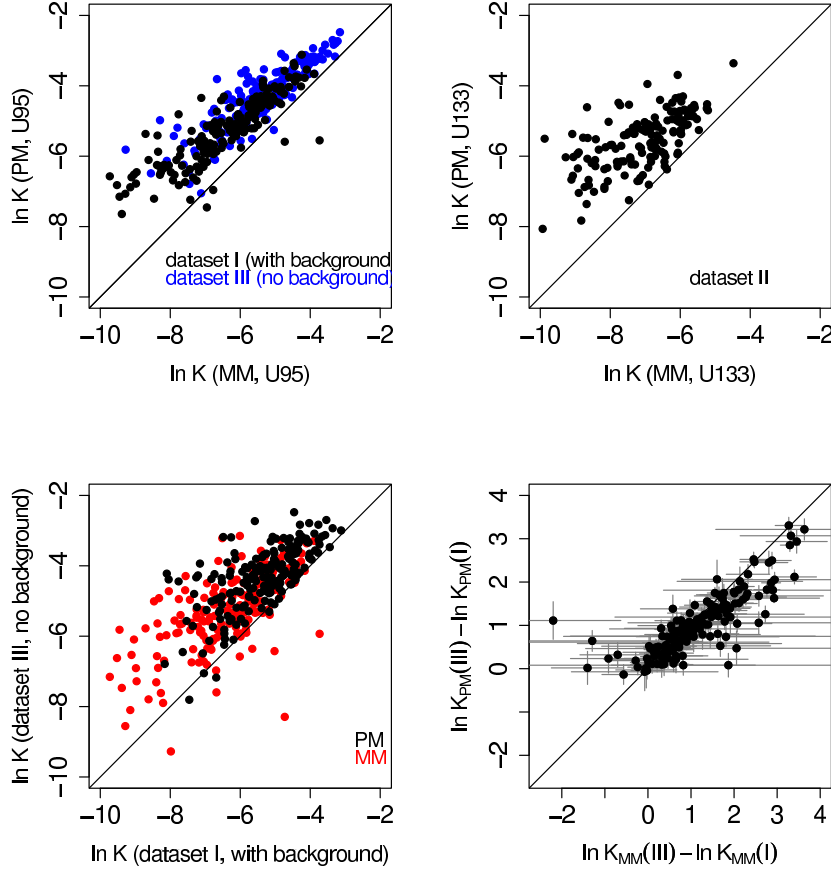


Figure 5. The log of the fitted parameters K , defined in Eq. 1. The first and second panels compare PM/MM pairs for each of the three datasets. The third panel compares datasets I and III, with and without the complex human pancreas background respectively. The fourth panel compares the increase in $\ln K$ for MM with that for PM when the complex human pancreas background is removed.

Various comparisons of the effective adsorption rate constant K are shown in Fig. 5. This parameter is modelled in terms of more fundamental rate constants via Eq. 16. One can reasonably assume the differences between PM and MM for the indirect effective rate constants X^{bulk} , K^{Pfold} and X^{NS} to be small compared with that for the specific rate constant K^{S} , because of averaging over large numbers of non-specific species. Thus, to a reasonable approximation, $\ln K_{\text{PM}}/K_{\text{MM}} \approx \ln K_{\text{PM}}^{\text{S}}/K_{\text{MM}}^{\text{S}} \approx -\Delta\Delta G/RT$, where $\Delta\Delta G$ is the difference in specific binding energies between a specific mRNA target and a PM and MM probe respectively. This empirical result has been noted previously as a shift away from the diagonal in a plot of K_{PM} versus K_{MM} for dataset I [15] and for dataset II [6], and is confirmed here for all three datasets.

Estimates of $-\Delta\Delta G$ were obtained by taking a weighted average of $-\Delta\Delta G_p = (\ln K_{\text{PM}} - \ln K_{\text{MM}})_p$ over fitted isotherms p for each of the central base letters. The weighting $\sum_p (-\Delta\Delta G_p/s_p^2) / \sum_p (1/s_p^2)$, where s_p is the standard error in $\Delta\Delta G_p$ estimated from the standard error in K_p arising from the isotherm fit is chosen to minimise the error in the average. The results, given in Table 4, are consistent with

Table 4. $-\Delta\Delta G$ estimated from a weighted average of $(\ln K_{\text{PM}} - \ln K_{\text{MM}})$ for each of the four central bases.

	C	G	T	A
Dataset I	1.09 ± 0.05	0.94 ± 0.06	0.98 ± 0.04	0.74 ± 0.04
Dataset II	1.74 ± 0.14	1.43 ± 0.10	1.26 ± 0.05	0.84 ± 0.06
Dataset III	1.10 ± 0.04	0.90 ± 0.04	0.97 ± 0.03	0.66 ± 0.03

the ordering of binding energies calculated from nearest neighbour stacking models in Eq. 18.

The third panel of Fig. 5 compares the effective rate constant K for the U95 experiments with and without the complex human pancreas background. From Eq. 16, one sees that removing the background, which has the effect of setting X^{bulk} and X^{NS} to zero, should increase K . This is confirmed in the plot, and is also evident as a shift in the inflection points to the left between Figs. 1 and 2. The fourth panel compares the amount by which $\ln K$ increases as the complex background is removed for PM and MM. We observe that removing the effects of X^{bulk} and X^{NS} affects K for a PM probe and its MM partner by a similar factor. The small number of points away from the diagonal line to the left of the plot are caused by the difficulty in fitting the MM isotherm when K_{MM}^{-1} is beyond the upper limit of the range of spike-in concentrations.

5. Quantitative behaviour of the fits

In this section we explore the ability of the model to explain the quantitative relationship between the fitting parameters of the hyperbolic response curve Eq. 1 and known physical properties of microarrays. We divide the analysis into two parts: The parameters A and B which set the vertical scale of the hyperbolic response curve, and the parameter K which sets its horizontal scale.

5.1. Vertical scale parameters

The parameters A and B are explained in the model in terms of the more fundamental quantities a , the physical background and b , saturation intensity above background (which together set the intensity scale in terms of the dimensionless duplex coverage fraction) and α and β (which are driven by the chemical reactions in Table 3).

We begin with a and b , which are assumed to be fixed for an entire microarray. In Fig. 6 are plotted histograms of measured fluorescence intensities across microarrays from datasets I and II. Because these data have been quantile normalised, a common histogram will apply to all microarrays within a given dataset. In the absence of statistical noise, the parameters a and $a + b$ should provide bounding limits for these histograms. However, given the coefficients of variation reported in Table 1, the raw intensity measurements could well extend beyond these limits. According to Eq. 17 the

Table 5. Fitted parameters, to 3 significant figures, occurring in the analysis of Section 5.

Defining equation	Parameter	Dataset I	Dataset II	Dataset III
(2)	a	88.4	30.0	
	$\log_{10} a$	1.95	1.48	
	b	32300	48800	
	$\log_{10} b$	4.51	4.69	
(20)	c_0	62.2	37.9	
	λ_0	0.0920	0.0841	
(24) (α Model 2)	c_1	-16.1	-14.8	
	c_2	-0.186	-0.148	
	c_3	0.124	0.102	
(28) (α Model 5)	c_1	-14.8	-14.8	
	c_2	-0.200	-0.149	
	c_3	0.0776	0.101	
	λ_α	0.176	0.909	
	μ_α	4.57	-6.14	
(43) (K Model 7)	λ_S	0.145	0.0824	0.0944
	μ_S	-62.9	69.4	-73.4
	$\lambda_{S\text{fold}}$	0.204	0.131	0.202
	$\mu_{S\text{fold}}$	-59.3	-51.5	-73.7
	$\lambda_{P\text{fold}}$	0.268	0.100	0.385
	$\mu_{P\text{fold}}$	-0.757	136	0.917

parameter a should also be a lower cutoff on the fitted parameter A (corresponding to the case of negligible non-specific hybridisation), while the analysis of A in Section 4 for dataset III (see the fourth panel of Fig. 4) suggests a much smaller coefficient of variation for the fitted value of A than for the intensity data generally. For these reasons we use as an estimate of a the minimum over all fits within a dataset of the parameter A . These estimates are shown in Fig. 6, together with bars extending two standard deviations either side using the coefficients of variation in Table 1.

To estimate the saturation parameter b from fits to the hyperbolic response function, and to explain the observed values of the combination $\alpha + \beta$, we make use of the model's prediction that the asymptotic intensity at high spike-in concentration, $I(\infty)$, is determined by the washing-step survival function of PM-specific duplexes, $s^S(t_W)$ [8]. Thus from Eqs. 15 and 17 we have

$$I(\infty) = A + B = a + b(\alpha + \beta) = a + bs^S(t_W) = a + be^{-\kappa t_W}, \quad (19)$$

where we assume a uniform washing rate κ that depends only on the probe and target nucleotide sequences via their binding affinity. Following Ref. [8], we model κ in terms

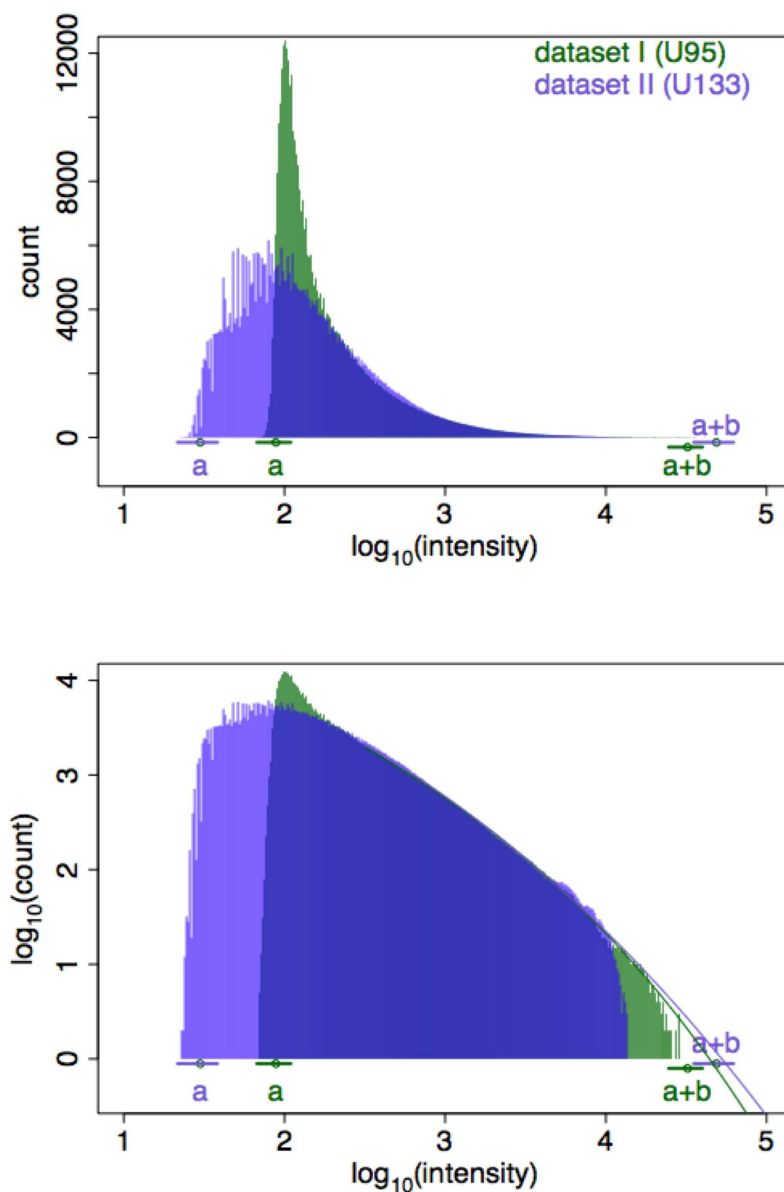


Figure 6. Histograms of quantile normalised fluorescence intensities across microarrays in datasets I and II on a linear (upper) and logarithmic (lower) scale. Both PM and MM intensities are included. Counts are from bins of size 0.01 on the log intensity axis. Also indicated are estimates of the parameters a and b for each dataset, with bars indicating two standard deviations of the spread in the intensity data either side. The curves fitted to the histograms in the lower panel are explained in Section 6.

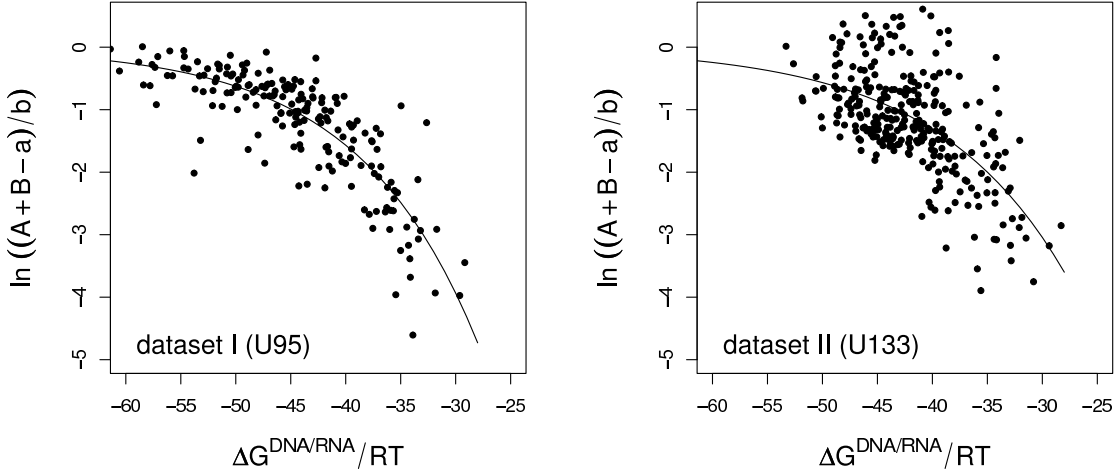


Figure 7. Fits of Eq. 21 to the parameter combination $A + B$ of the hyperbolic response function fits to the PM data for datasets I and II. The parameter b has been absorbed into a shift in the ordinate.

of the RNA/DNA duplex free binding energy in bulk solution:

$$\kappa t_W = c_0 e^{\lambda_0 \Delta G^{\text{DNA/RNA}}/(RT)}. \quad (20)$$

Here $\Delta G^{\text{DNA/RNA}}$ is calculated using the nearest neighbour stacking model and parameters of Ref. [22], R is the gas constant, T the absolute temperature and c_0 and λ_0 are undetermined constants. We use the convention that $\Delta G^{\text{DNA/RNA}}$ is negative for a bound state. Rearranging gives

$$\ln(\alpha + \beta) = \ln \frac{A + B - a}{b} = -c_0 e^{\lambda_0 \Delta G^{\text{DNA/RNA}}/(RT)}, \quad (21)$$

where the A and B are determined for each feature from fitted hyperbolic response functions, a has been estimated above, and $\ln b$, c_0 and λ_0 are fitting parameters. Fits to datasets I and II are shown in Fig. 7, and fitting parameters listed in Table 5. The fits were done by minimising the sum of the squares of the residuals with respect to the three fitting parameters.

The parameter α is in principle predicted by the model in terms of fundamental physical constants via Eqs. 10 and 14. It is determined mainly by non-specific hybridisation, including a strong influence from the probe sequence pyrimidine content as suggested by Fig. 4, and by probe folding. Without a detailed knowledge of the composition of the non-specific target solution, a direct evaluation of α is of course impossible. However, one can aim for an ansatz in terms of those quantities which are known. Following the reasoning used above, the washing-step survival function of the i th non-specific species $s_i^{\text{NS}}(t_W)$ can be assumed to take the form of a double exponential function of the corresponding free binding energy $\Delta G_i^{\text{DNA/RNA}}$. The value of the double exponential function ($f(x) = e^{-e^x}$) undergoes a changeover from 1 for $x \ll 0$ to 0 for $x \gg 0$ over a narrow range of its argument. Thus the factor $s_i^{\text{NS}}(t_W)$ in Eq. 10 can be thought of as a switch which removes from the sum any binding configuration i less tightly bound than some threshold.

The numerator of Eq. 14 is then a sum of reaction rate constants K_i^{NS} , weighted by the number of target molecules in solution with nucleotide sequences complementary to the i th subsequence of the probe. Numerical estimates carried out in the context of bulk hybridisation in solution[14] show that such a sum can be well approximated as an exponential function of free binding energy of the entire probe sequence to its complement. Assuming then that the behaviour of α is dominantly exponential in $\Delta G^{\text{DNA/RNA}}$, and taking into account the added effect of the probe's pyrimidine count, we have tested the following nested models:

$$\text{Model 0: } \ln \alpha = c_1 + \epsilon, \quad (22)$$

$$\text{Model 1: } \ln \alpha = c_1 + c_2 \Delta g^{\text{DNA/RNA}} + \epsilon, \quad (23)$$

$$\text{Model 2: } \ln \alpha = c_1 + c_2 \Delta g^{\text{DNA/RNA}} + c_3 n_{\text{pyr}} + \epsilon, \quad (24)$$

$$\text{Model 3: } \ln \alpha = c_1 + c_2 \Delta g^{\text{DNA/RNA}} + c_3 n_{\text{pyr}} + c_4 n_{\text{pyr}} \Delta g^{\text{DNA/RNA}} + \epsilon, \quad (25)$$

where c_1, \dots, c_4 are fitting parameters to be determined, $\Delta g^{\text{DNA/RNA}} = \Delta G^{\text{DNA/RNA}}/(RT)$, n_{pyr} is a count of the number of pyrimidines in the 25-mer probe sequence and ϵ is the residual error, which is assumed Gaussian.

To include the effect of probe folding, we approximate K^{Pfold} in Eq. 14 by a single exponential term

$$K^{\text{Pfold}} = \exp[\lambda_\alpha(\mu_\alpha - \Delta g^{\text{DNA-fold}})], \quad (26)$$

where λ_α and μ_α are fitting parameters and $\Delta g^{\text{DNA-fold}} = \Delta G^{\text{DNA-fold}}/(RT)$, where $\Delta G^{\text{DNA-fold}}$ is calculated for each 25-mer probe sequence from Zuker's Mfold web server [24] with the temperature set to 45°C and other parameters set to their default values. The Mfold web server calculates the free energy of the most energetic folding configuration of a given single strand DNA sequence, though ideally one should include a Boltzman weighted sum over all possible folding configurations. Models 1 and 2 are then nested within Models 4 and 5 respectively:

$$\begin{aligned} \text{Model 4: } \ln \alpha &= c_1 + c_2 \Delta g^{\text{DNA/RNA}} \\ &\quad - \ln\{1 + \exp[\lambda_\alpha(\mu_\alpha - \Delta g^{\text{DNA-fold}})]\} + \epsilon, \end{aligned} \quad (27)$$

$$\begin{aligned} \text{Model 5: } \ln \alpha &= c_1 + c_2 \Delta g^{\text{DNA/RNA}} + c_3 n_{\text{pyr}} \\ &\quad - \ln\{1 + \exp[\lambda_\alpha(\mu_\alpha - \Delta g^{\text{DNA-fold}})]\} + \epsilon. \end{aligned} \quad (28)$$

The above nested models can be tested for the significance of the extra parameters introduced in going from a less to a more complicated model. For instance, to test the significance of the extra parameter distinguishing model m_2 from the simpler m_1 , consider for each model the residual sums of squares $D = \sum \epsilon^2$, where the sum is taken over fitted data points. Under the null hypothesis that the extra complexity is not significant, and assuming the data points to be independent, the test statistic defined by

$$F = \frac{(D_1 - D_2)/(d_1 - d_2)}{D_2/d_2}, \quad (29)$$

Table 6. P-values testing significance of the extra parameter related to nested pairs of models in Eqs. 22 to 28. Smaller p-values indicate that the extra parameters in the more complicated model are significant. The second column gives the extra parameters included in the more complicated of the two models.

	Parameter	Dataset I	Dataset II
model 0 to model 1:	c_2	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$
model 1 to model 2:	c_3	1.2×10^{-9}	1.9×10^{-6}
model 2 to model 3:	c_4	0.0098	0.648
model 1 to model 4:	$\lambda_\alpha, \mu_\alpha$	1.1×10^{-10}	0.60
model 2 to model 5:	$\lambda_\alpha, \mu_\alpha$	3.4×10^{-5}	0.81
model 4 to model 5:	c_3	0.00064	2.7×10^{-6}

(where d_1 and d_2 count the residual degrees of freedom of m_1 and m_2 respectively) has an F distribution with degrees of freedom equal to $d_1 - d_2$ and d_2 . This allows us to assign a p-value to the significance of model m_2 over m_1 .

We have fitted each of the five models to the combination $\alpha = (A - a)/b$ using a and b from Table 5 and A from the hyperbolic response function fits of both PM and MM data for datasets I and II separately. The number of data points fitted, that is, the number of fitted hyperbolic response functions for which all three parameters A , B and K are positive, was 364 for dataset I and 460 for dataset II. Table 6 gives the calculated p-values.

The parameters c_2 and c_3 modelling linear effects in $\Delta G^{\text{DNA/RNA}}$ and n_{pyr} respectively are highly significant in both datasets. The parameter c_4 defining a mixed effect is barely significant at the 1% level in dataset I and not significant in dataset II, and we shall ignore it from here on.

The probe folding effect is highly significant for dataset I, but not significant for dataset II. Note that Models 4 and 5 contain the functional form

$$-\ln\{1 + \exp[\lambda(\mu - \Delta g)]\} \approx \begin{cases} \lambda(\Delta g - \mu), & \Delta g \ll \mu, \\ 0, & \Delta g \gg \mu, \end{cases} \quad (30)$$

for $\lambda > 0$. Thus the probe folding effect “switches on” once the energy of a folded probe is below some threshold μ_α , at which point the effect becomes linear in $\Delta g^{\text{DNA-fold}}$. From Table 5, the fitted value of μ_α in Model 5 for dataset I is 4.57, whereas the range of probe folding energies calculated by Mfold for the probe sequences of the U95 microarray is $-8.18 < \Delta g^{\text{DNA-fold}} < 2.98$. Thus μ_α is well above the folding energy of any of the probes, implying that the probe folding effect is effectively linear for dataset I. On the other hand, the fitted value of μ_α in Model 5 for dataset II is -6.14 , which is below the range $-5.49 < \Delta g^{\text{DNA-fold}} < 3.11$ of calculated probe folding energies for the U133 microarray, implying that the probe folding is switched off for dataset II. There is no obvious reason why the probe folding parameter μ_α should shift markedly from one spike-in experiment to another, given that, although

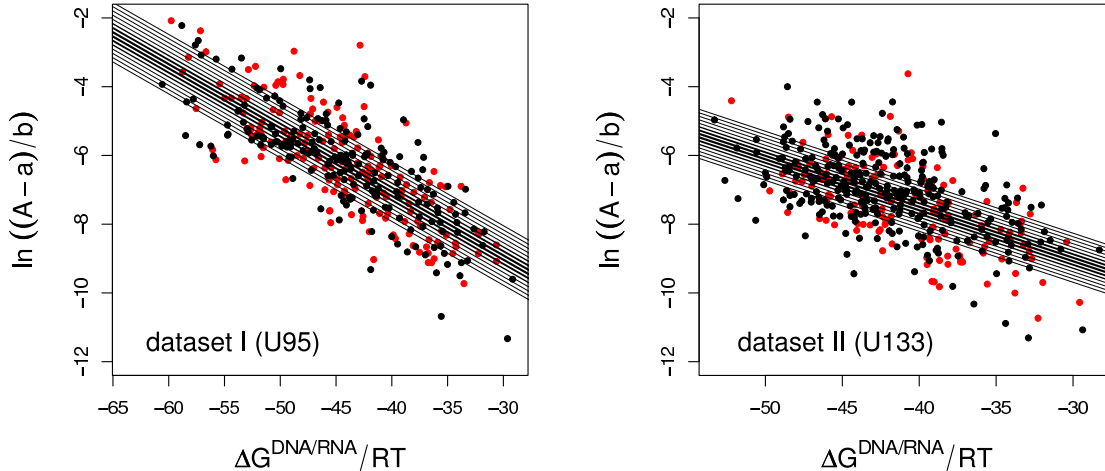


Figure 8. Fits of Eq. 24 (Model 2) to the parameter A of the hyperbolic response function fits of both PM (black) and MM (red) data for datasets I and II. The parameters a and b are from Table 5. The plotted lines correspond to the range $6 \leq n_{\text{pyr}} \leq 20$ of pyrimidine counts, with n_{pyr} increasing from bottom to top.

the U133 probeset nucleotide sequences are completely redesigned from those of the U95 microarray, the experimental protocol and geometry of the microarray should be similar. One, albeit prosaic, explanation may simply be that dataset II is noisier (see Table 1) and the probe folding effect has been lost in the noise.

Fitted parameter values of Models 2 and 5 are given in Table 5. As expected from the above argument, the fitted parameters c_1 , c_2 and c_3 for dataset II differ very little between Models 2 and 5. Fits of Model 2 to the data are shown in Fig. 8.

5.2. Horizontal scale parameter

The parameter K sets the horizontal scale of the hyperbolic response function Eq. 1. K^{-1} is an estimate of the spike-in concentration required to give a fluorescence intensity half way between the background, zero concentration level and the asymptotic, infinite concentration level. Our model in Section 3 explains K as an effective rate reaction constant which is determined by the reactions occurring during the hybridisation step, and which is unaffected by the washing step. As was the case for the vertical scale parameters, much of the information required to evaluate K from first principles is unknown, and so we try for an ansatz based on probe sequences and free binding energies.

Eqs. 16, 5 and 9 give K in terms of reaction rate constants and concentrations of reactants. In general, each term K^r or X^r occurring in Eq. 16, where r labels one of the reactions in Table 3, is a sum of terms of the form $\text{const.} \times e^{-\Delta G_i^r/RT}$, weighted by the concentration of reactant i . Once again we will be guided by Heim et al.'s numerical estimate of X^{bulk} [14], and approximate each sum as a single exponential term. Thus we write

$$K^r \text{ or } X^r = \exp[\lambda_r(\mu_r - \Delta g^r)], \quad (31)$$

where the μ_r and λ_r are fitting parameters, and $\Delta g^r = \Delta G^r/RT$, with the effective binding energy ΔG^r for each reaction is estimated from some external physical model. With the sign convention that ΔG^r is defined negative for a bound state, each λ_r is expected to be positive.

Consider first dataset III, for which the complex non-specific background is absent. In Eq. 16 we can set the non-specific binding terms X^{bulk} and X^{NS} to zero, giving

$$\ln K = \ln K^{\text{S}} - \ln(1 + K^{\text{Sfold}}) - \ln(1 + K^{\text{Pfold}}). \quad (32)$$

The rate constants are modelled as

$$K^{\text{S}} = \exp[\lambda_{\text{S}}(\mu_{\text{S}} - \Delta g^{\text{DNA/RNA}})], \quad (33)$$

$$K^{\text{Sfold}} = \exp[\lambda_{\text{Sfold}}(\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})], \quad (34)$$

$$K^{\text{Pfold}} = \exp[\lambda_{\text{Pfold}}(\mu_{\text{Pfold}} - \Delta g^{\text{DNA-fold}})]. \quad (35)$$

For the free binding energy $\Delta G^{\text{DNA/RNA}}$ we use the nearest neighbour stacking model parameters of Sugimoto et al.[22], for $\Delta G^{\text{RNA/RNA}}$ we use Xia et al.'s nearest neighbour stacking parameters for RNA binding to RNA [23], and for $\Delta G^{\text{DNA-fold}}$ we use Zuker's Mfold web server [24]. The Mfold web server also has the facility to calculate folding energies of RNA targets. However, since for RNA target folding we are interested in the propensity for the 25-mer stretch of target complimentary to a given probe to bind with any segment of the much longer target RNA (or possibly another RNA molecule), we believe it is more appropriate to model target folding in solution using an RNA-to-RNA binding energy.

To try to understand the relative importance of each of the effects contributing to the effective rate constant K we have analysed a set of models containing nested pairs, the relationship between which is illustrated in Fig. 9:

$$\text{Model 0: } \ln K = k_0 + \epsilon, \quad (36)$$

$$\text{Model 1: } \ln K = \lambda_{\text{S}}(\mu_{\text{S}} - \Delta g^{\text{DNA/RNA}}) + \epsilon, \quad (37)$$

$$\text{Model 2: } \ln K = k_0 - \ln\{1 + \exp[\lambda_{\text{Sfold}}(\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})]\} + \epsilon, \quad (38)$$

$$\text{Model 3: } \ln K = k_0 - \ln\{1 + \exp[\lambda_{\text{Pfold}}(\mu_{\text{Pfold}} - \Delta g^{\text{DNA-fold}})]\} + \epsilon, \quad (39)$$

$$\begin{aligned} \text{Model 4: } \ln K &= \lambda_{\text{S}}(\mu_{\text{S}} - \Delta g^{\text{DNA/RNA}}) \\ &\quad - \ln\{1 + \exp[\lambda_{\text{Sfold}}(\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})]\} + \epsilon, \end{aligned} \quad (40)$$

$$\begin{aligned} \text{Model 5: } \ln K &= \lambda_{\text{S}}(\mu_{\text{S}} - \Delta g^{\text{DNA/RNA}}) \\ &\quad - \ln\{1 + \exp[\lambda_{\text{Pfold}}(\mu_{\text{Pfold}} - \Delta g^{\text{DNA-fold}})]\} + \epsilon, \end{aligned} \quad (41)$$

$$\begin{aligned} \text{Model 6: } \ln K &= k_0 - \ln\{1 + \exp[\lambda_{\text{Sfold}}(\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})]\} \\ &\quad - \ln\{1 + \exp[\lambda_{\text{Pfold}}(\mu_{\text{Pfold}} - \Delta g^{\text{DNA-fold}})]\} + \epsilon, \end{aligned} \quad (42)$$

$$\begin{aligned} \text{Model 7: } \ln K &= \lambda_{\text{S}}(\mu_{\text{S}} - \Delta g^{\text{DNA/RNA}}) \\ &\quad - \ln\{1 + \exp[\lambda_{\text{Sfold}}(\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})]\} \\ &\quad - \ln\{1 + \exp[\lambda_{\text{Pfold}}(\mu_{\text{Pfold}} - \Delta g^{\text{DNA-fold}})]\} + \epsilon, \end{aligned} \quad (43)$$

where ϵ is the residual error, which is assumed Gaussian. The first term in each of Models 1, 4, 5 and 7 could equally well be written as $k_0 + k_1 \Delta g^{\text{DNA/RNA}}$ to make

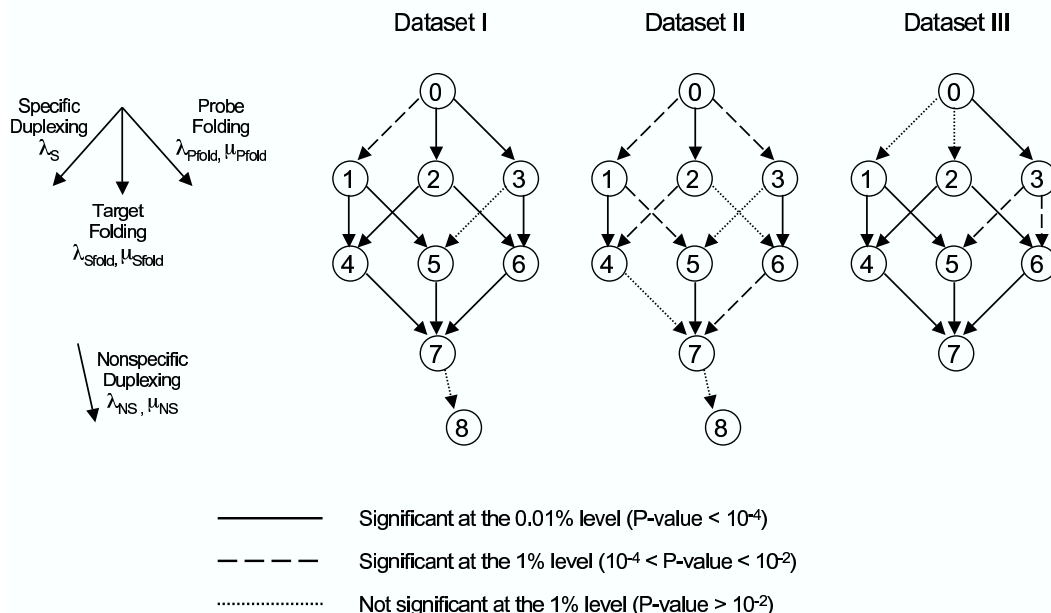


Figure 9. The relationship between the nine models fitting the effective rate constant K , namely Eqs. 36 to 43 and 46. The directions of the arrows indicate the extra effects included in going from a simpler, nested model to a more complicated model, and the style of the arrows indicate P-values from Table 7. Smaller P-values indicate that the extra parameters in the more complicated model are significant. The fitted parameters corresponding to the favoured model, Model 7, are given in Table 5

the nesting explicit, but for convenience we use a parameterisation based on Eq. 33. Models 1 to 3 include only the effects of specific hybridisation, target folding in bulk solution and probe folding respectively. Models 4 to 6 include pairwise effects, and Model 7 includes all three effects. A recurring theme in these models is the functional form of Eq. 30. Thus the target and probe folding effects “switch on” once the binding energy is below (i.e. more tightly binding than) thresholds $\mu_{S\text{fold}}$ and $\mu_{P\text{fold}}$ respectively, and have the effect of reducing the effective rate constant K .

P-values calculated from the F-statistic, Eq. 29, testing the pairwise comparative significance of these models are given for Dataset III in the right hand column of Table 7 (see also Fig. 9). Results are shown for PM data only as no complete set of stacking model parameters for $\Delta G^{\text{DNA/RNA}}$ with mismatches is available. Comparisons of Model 0 with Models 1 to 3 indicate that, taken in isolation, the specific hybridisation and bulk target folding effects appear not to be significant, whereas the probe-folding effect appears to be highly significant. This is also illustrated in Fig. 10. However, when taken in combination with probe folding, the analysis shows specific binding and target folding to be significant at the 1% level (see the comparisons Model 3 to 5 and 3 to 6 in Table 7). Thus we accept Model 7 for Dataset III. The fitted parameters are given in the right hand column of Table 5. Note that each λ_r is positive as required of the model.

For Dataset III, the apparent non-significance of the specific hybridisation and

Table 7. P-values testing significance of the extra parameters related to nested pairs of Models 0 to 8 fitting the effective rate constant K . Smaller p-values indicate that the extra parameters in the more complicated model are significant. The number of fitted PM hyperbolic response functions for which all three parameters A , B and K positive, and hence the number of points fitted to the models, is 188 for dataset I, 303 for dataset II and 192 for dataset III.

	Parameter	Dataset I	Dataset II	Dataset III
model 0 to model 1:	λ_S	0.00028	0.00021	0.51
model 0 to model 2:	$\lambda_{S\text{fold}}, \mu_{S\text{fold}}$	3.5×10^{-11}	1.1×10^{-7}	0.32
model 0 to model 3:	$\lambda_{P\text{fold}}, \mu_{P\text{fold}}$	1.6×10^{-12}	0.00011	4.8×10^{-15}
model 1 to model 4:	$\lambda_{S\text{fold}}, \mu_{S\text{fold}}$	$< 2 \times 10^{-16}$	2.3×10^{-6}	1.8×10^{-9}
model 1 to model 5:	$\lambda_{P\text{fold}}, \mu_{P\text{fold}}$	3.0×10^{-10}	0.0077	$< 2 \times 10^{-16}$
model 2 to model 4:	λ_S	4.4×10^{-12}	0.0055	5.9×10^{-10}
model 2 to model 6:	$\lambda_{P\text{fold}}, \mu_{P\text{fold}}$	6.8×10^{-7}	0.089	$< 2 \times 10^{-16}$
model 3 to model 5:	λ_S	0.087	0.023	0.00016
model 3 to model 6:	$\lambda_{S\text{fold}}, \mu_{S\text{fold}}$	1.4×10^{-5}	9.6×10^{-5}	0.00012
model 4 to model 7:	$\lambda_{P\text{fold}}, \mu_{P\text{fold}}$	9.1×10^{-5}	0.056	7.7×10^{-14}
model 5 to model 7:	$\lambda_{S\text{fold}}, \mu_{S\text{fold}}$	3.8×10^{-13}	1.7×10^{-5}	2.0×10^{-5}
model 6 to model 7:	λ_S	7.9×10^{-10}	0.0034	2.4×10^{-5}
model 7 to model 8:	λ_{NS}, μ_{NS}	0.016	0.23	

bulk target folding effects in Models 1 and 2 can be explained as follows. In Model 7, the bulk folding is a stronger effect than specific hybridisation by a factor of 2 ($\lambda_{S\text{fold}} \approx 2\lambda_S$ in Table 5). Furthermore, from Eq. 30, the bulk folding effect is opposite in sign to the specific hybridisation effect, and only comes into effect for $\Delta g^{\text{RNA/RNA}} < \mu_{S\text{fold}} = -73.7$. Also, it turns out that $\Delta g^{\text{DNA/RNA}}$ and $\Delta g^{\text{RNA/RNA}}$ are very highly correlated, with a Pearson correlation coefficient of 0.89. Examination of the first two plots in Fig. 10 shows a tendency of the data to increase with Δg to start with, while the bulk target folding dominates, and then to decrease once the bulk folding effect switches off and the specific hybridisation effect takes over. Attempting to fit a straight line through data which first increases and then decreases has resulted in the conclusion that the term linear in ΔG in Model I is not significant. A related statement has been made by Carlon and Heim [10], namely that the effective target concentration needs to be appropriately “rescaled” for those targets with a high binding affinity in bulk solution in order to see the expected relationship between K and $\Delta G^{\text{DNA/RNA}}$.

We now turn to Datasets I and II. In the presence of a complex non-specific background, X^{bulk} and X^{NS} are reinstated in Eq. 16. The bulk hybridisation effect will be a sum of exponentials of $\Delta G_i^{\text{RNA/RNA}}$, and its modelling can be absorbed into that for bulk target folding, while the non-specific effect will be a sum of exponentials

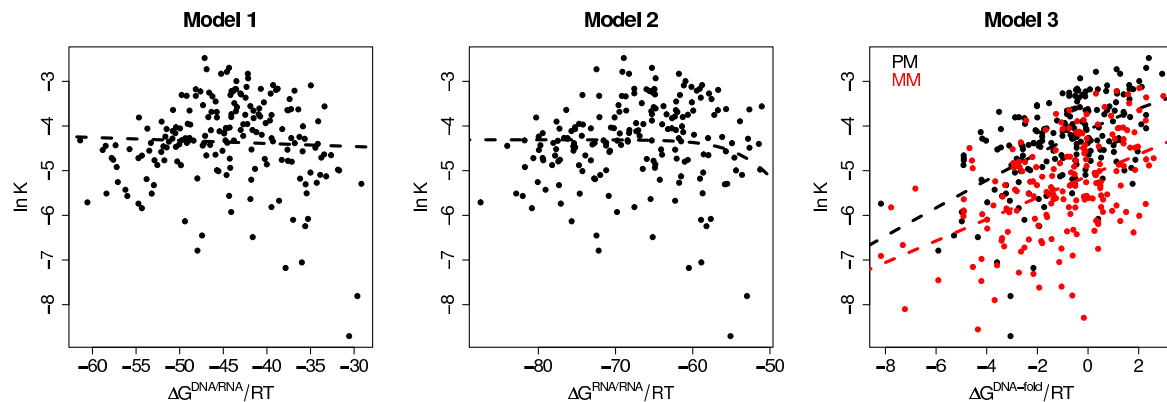


Figure 10. Fits of $\ln K$ estimated from Dataset III to Models 1, 2 and 3. Mismatch data is also shown for Model 3 since ΔG^{Pfold} can be obtained from the Mfold web site for all probe sequences.

of $\Delta G_i^{\text{DNA/RNA}}$. Thus we set

$$K^{\text{Sfold}} + X^{\text{bulk}} = \exp[\lambda_{\text{Sfold}}(\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})], \quad (44)$$

$$X^{\text{NS}} = \exp[\lambda_{\text{NS}}(\mu_{\text{NS}} - \Delta g^{\text{DNA/RNA}})], \quad (45)$$

which suggests one further model:

$$\begin{aligned} \text{Model 8: } \ln K = & \lambda_{\text{S}}(\mu_{\text{S}} - \Delta g^{\text{DNA/RNA}}) \\ & - \ln \left\{ 1 + \exp[\lambda_{\text{Sfold}}(\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})] \right\} \\ & - \ln \left\{ 1 + \exp[\lambda_{\text{Pfold}}(\mu_{\text{Pfold}} - \Delta g^{\text{DNA-fold}})] \right. \\ & \left. + \exp[\lambda_{\text{NS}}(\mu_{\text{NS}} - \Delta g^{\text{DNA/RNA}})] \right\} + \epsilon. \end{aligned} \quad (46)$$

Turning to Table 7, columns I and II, we discover that the extra parameters introduced to account for non-specific probe-target binding are not significant at the 1% level (see also Fig. 9). This surprising result can be explained by the fact that the fitted values of μ_{NS} are in both cases close to the maximum value of $\Delta g^{\text{DNA/RNA}}$ within the dataset, so most of the fitted points fall into the $\Delta g^{\text{DNA/RNA}} \ll \mu_{\text{NS}}$ regime of Eq. 30, and the effect is adequately covered by the λ_{S} term of Model 7. To further illustrate the point, fits to Models 1, 2 and 3 are plotted In Fig. 11. If Model 1 is taken in isolation, λ_{S} appears to have the “wrong” sign, as the non-specific probe-target binding and target folding and binding in solution all combine to dominate the specific binding effect. A similar result is observed for dataset II.

The generally small p-values in the first column of Table 7 indicate that Model 7 is an appropriate description of the parameter K for dataset I. For dataset II the picture is less clear. In agreement with the analysis of the parameter α , the probe folding is in general less significant. Nevertheless, for consistency we list the fitting parameters of Model 7 to both datasets in Table 5, while acknowledging there is redundancy in the Dataset II parameters.

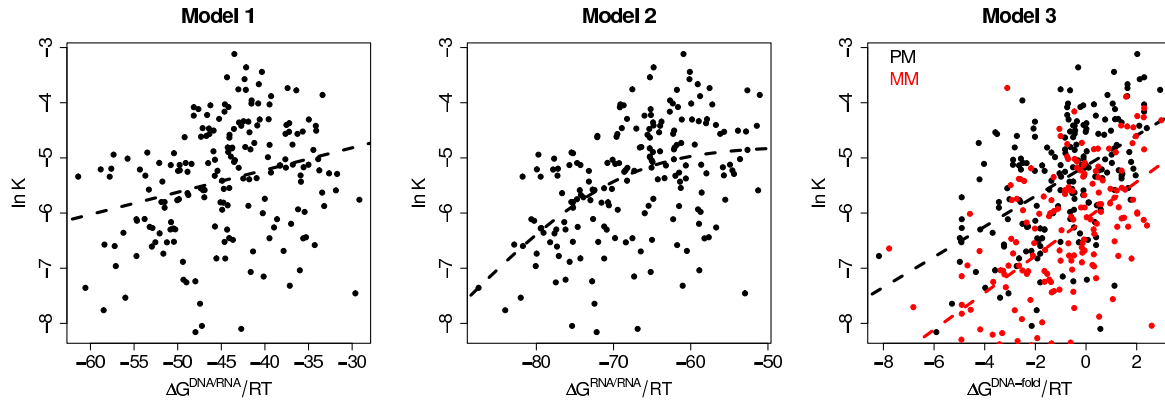


Figure 11. The same as Fig. 10 for Dataset I.

5.3. How close are the fits?

Fig. 12 gives some idea of how much information has been lost in the above fits. The plot compares estimated parameters A , B and K of the fitted hyperbolic response curves, such as those in Fig. 1, with those which would be predicted by the fitting constants and parameters listed in Table 5, namely

$$A = a + be^{c_1 + c_2 \Delta g^{\text{DNA/RNA}} + c_3 n_{\text{pyr}} - \ln\{1 + \exp[\lambda_\alpha (\mu_\alpha - \Delta g^{\text{DNA-fold}})]\}}, \quad (47)$$

$$B = a + be^{-c_0 \exp(\lambda_0 \Delta g^{\text{RNA/RNA}})} - A, \quad (48)$$

$$K = \frac{e^{\lambda_S (\mu_S - \Delta g^{\text{DNA/RNA}})}}{\left[1 + e^{\lambda_{\text{Sfold}} (\mu_{\text{Sfold}} - \Delta g^{\text{RNA/RNA}})}\right] \left[1 + e^{\lambda_{\text{Pfold}} (\mu_{\text{Pfold}} - \Delta g^{\text{DNA-fold}})}\right]}. \quad (49)$$

Dotted lines either side of the diagonal are the boundary of the region within which predicted values do not differ from the original fitted parameters by more than a factor of 2. A clear majority of estimates of A and B fall within this range. Clearly the most difficult parameter to explain adequately is the horizontal scale K , owing to the large number of contributing chemical reactions. In general, dataset II has proved to be more problematic than dataset I, probably because the concentration range tested does not extend far enough into the saturation regime to demonstrate a clear hyperbolic isotherm.

6. Parameter prediction

For the above model to be of value in constructing a practical algorithm for inferring target concentrations, some or all of its parameters should ideally be predictable using only information available to experimental biologists. That available information consists of fluorescence intensities for the complete set of features on each microarray used in an experiment, the probe sequences of each feature, and any parameters associated with the experimental protocol. By contrast, the fitted parameters of Table 5 were obtained from spike-in experiments. Comparing datasets I and III in Figs. 4 and 5, one sees for instance that the unknown nature of the complex background has a

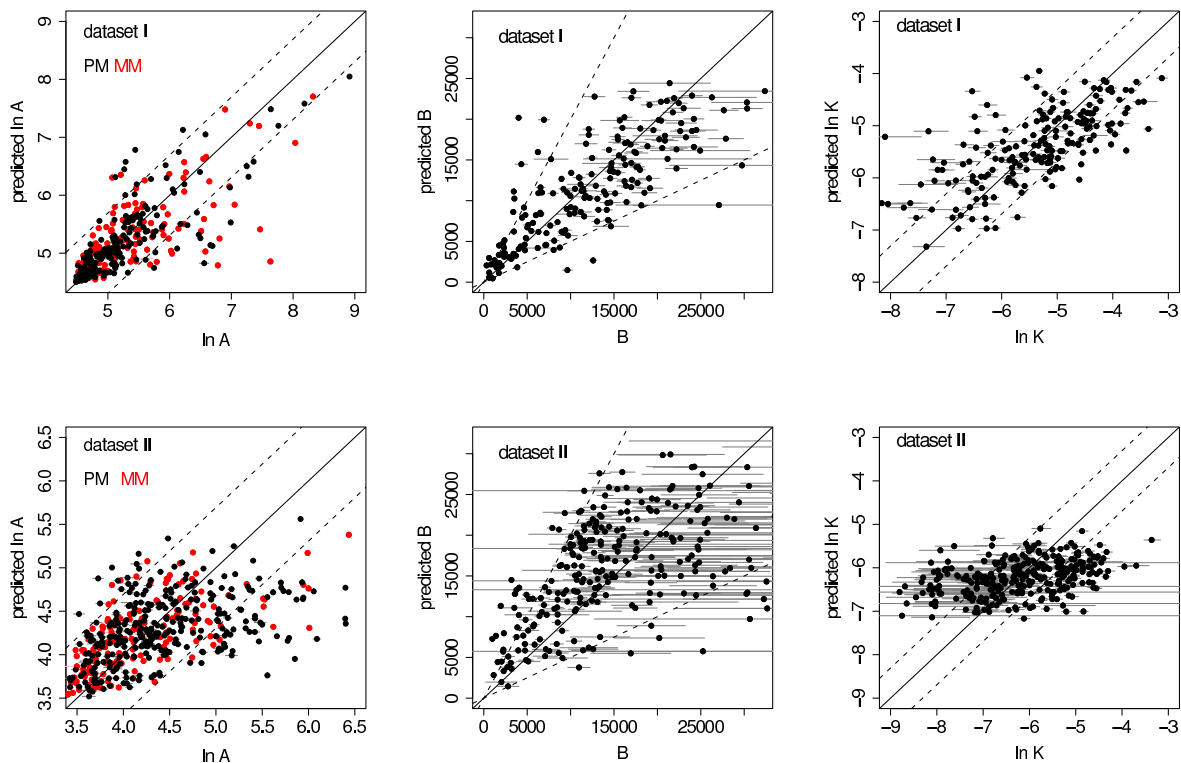


Figure 12. Estimated parameters obtained by fitting the the hyperbolic response function Eq. 1 to datasets I and II (horizontal axes, together with error bars showing standard errors) plotted against with the values that would be predicted by the quantitative fits of Section 5 (vertical axes). The dotted lines indicate a factor of 2 either side of the diagonal.

profound effect on the parameters A and K of the hyperbolic response function. At first sight it appears one may need a new set of spike-in data for each experiment, which is clearly not a practical consideration. However, we argue here that if one exploits the distribution of fluorescence intensities from the entire microarray, an estimation of vertical scale parameters at least may be possible.

In the following qualitative description we propose a two step process for the vertical scale parameters, in which the physical background a and maximum intensity b for a microarray are first determined from the entire distribution of intensities over the microarray. The intensities $I(x)$ can then be scaled to the dimensionless coverage fraction $\theta(x)$ via Eq. 2, and one is left with the remaining problem of estimating the parameters α and β , which are driven by the chemical reactions of Table 3.

To estimate a and b , consider the histograms in Fig. 6. For both datasets I and II, our estimate of the physical background a , based on hyperbolic response curves derived from spike-in data, is close to two standard deviations above the minimum measured fluorescence intensity. Assuming an experiment consisting of a number of technical replicates of each hybridisation setup, the data can be quantile normalised across replicates. A representative minimum intensity a_{\min} can be obtained by fitting a suitable smooth curve to the logged histogram (i.e. the lower panel of Fig. 6), and the

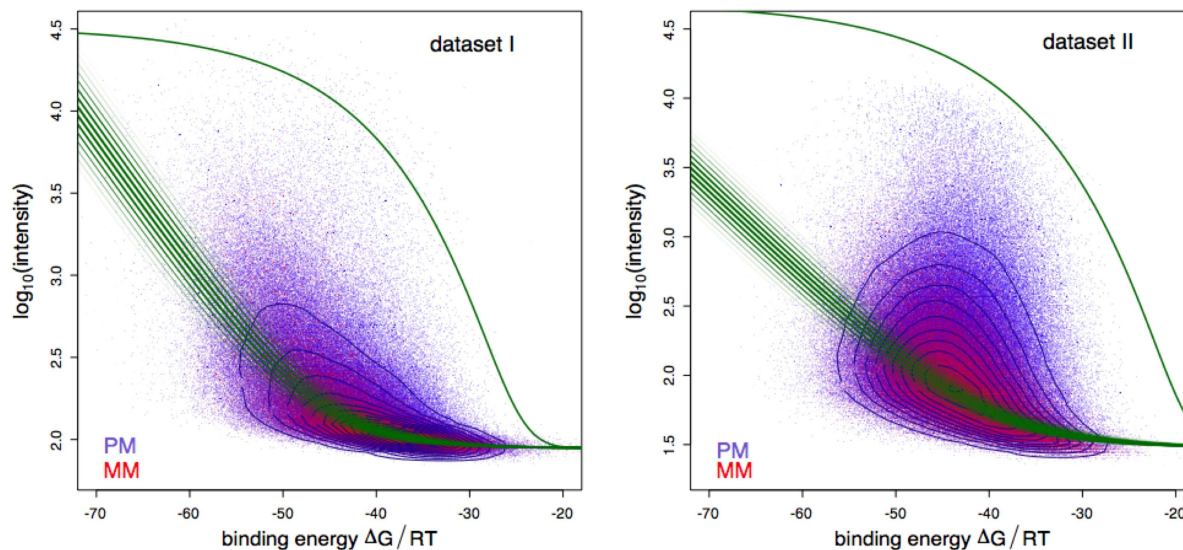


Figure 13. Scatter plots of measured fluorescence intensities against the theoretical DNA/RNA free binding energies. The upper curve is the fit to $A + B$ of Eq. 48, and the lower set of curves are the fits to A of Eq. 47 for pyrimidine counts $6 < n_{\text{pyr}} < 20$, with n_{pyr} increasing from bottom to top. Also shown are contour lines of the density of points.

coefficient of variation in the data η easily estimated from the replicate intensity values over the whole microarray. $(1 + 2\eta)a_{\min}$ then gives an estimate of a .

Estimating b from the histogram proves to be quite difficult because of the gradual tail at its the right hand end. With some experimentation we find that a cubic fit to the logged histogram over the range $[l + 0.25(u - l), l + 0.875(u - l)]$, where l and u are the lower and upper extremities of the histogram, crosses the $\log_{10}(\text{count}) = 0$ line close to two standard deviations above the previously obtained estimate of $a + b$. Calling this point $(a + b)_{\max}$, an estimate of $a + b$ is then $(1 - 2\eta)(a + b)_{\max}$. However, we find that such a method is highly sensitive to the range over which the cubic is fitted.

To gain some insight into how α and β may be estimated, consider the scatter plot, Fig. 13, of fluorescence intensities against the theoretical free binding energy $\Delta G^{\text{DNA/RNA}}$ obtained from the probe letter sequences using the nearest neighbour stacking model [22]. Superimposed on these plots are the fits from Eqs. 47 and 48 to the asymptotic saturation intensity $A + B$ and background intensity at zero spike-in concentration A , using the parameters of Table 5. According to our model, the asymptote curve should form an upper envelope to the data, with some slight leakage across the envelope due to the finite coefficient of variation in the data. Because the vast majority of the genes are not expressed in RNA samples taken from a typical cell, most of the data is expected to lie along, or close to, the lower set of background intensity curves. Indeed this is precisely what is seen. Conversely, in the absence of spike-in data, there is potential to estimate the upper, asymptote intensity curve by fitting an envelope

to the data and the lower, background intensity curve by fitting a curve through the ridge of the scatter plot’s contour lines. In principle, these fitted curves, together with a and b then determine estimates of α and β for each feature on the microarray.

7. Conclusions and outlook

This paper concentrates mainly on the immediate aim of understanding the physical processes at work in the operation of microarrays, but in doing so highlights some of the challenges and, we hope, gives some guidance, for meeting the ultimate aim of providing an algorithm for converting the set raw fluorescence intensities from microarrays to absolute target concentrations.

The model we have examined includes the effects of specific and non-specific hybridisation, folding and hybridisation in bulk solution of target RNA, the folding of probes at the microarray surface, and the removal of signal during the post-hybridisation step. It leads to the hyperbolic response curve (or Langmuir isotherm) of Eq. 1 with three fitting parameters A , B and K , which depend on a set underlying physical parameters including chemical reaction rate constants, washing survival functions and RNA target concentrations. In more practical terms, all three parameters will depend on the probe letter sequence, whether the probe is PM or MM, the nature of the complex non-specific background, and experimental protocols such as hybridisation temperature and washing times. Determining the parameters only from information likely to be known to biologists in a practical situation, as opposed to a highly controlled spike-in experiment, remains a formidable task.

The model is tested against the Affymetrix U95 spike-in datasets with and without a complex non-specific background and the Affymetrix U133 spike-in dataset. In general, agreement with a hyperbolic response curve is excellent for the U95 spike-in datasets and reasonable for the U133 spike-in dataset (see Table 1).

The response function parameters A and B set the vertical scale of the isotherm, that is, the scale of the measured fluorescence intensities. The parameter A is a combination of a relatively straightforward physical background, and a non-trivial contribution from the complex non-specific background. It is important to understand the nature of the non-specific background component as it is responsible for the “bright mismatches” problem which complicates the naive PM – MM subtraction scheme used in the MAS5 algorithm, for instance, for dealing with non-specific hybridisation. Our analysis shows that the DNA/RNA binding energy, pyrimidine content, and (in the case of U95 dataset) the folding of probes, all contribute significantly to the value of this parameter. The dependence of A on binding energies and pyrimidine count is illustrated in Fig. 8.

The parameter B (or, more precisely, the combination $A + B$, where $B \gg A$ in general) is mainly concerned with the asymptote at high spike-in concentration, which, according to our model, is driven by the washing step. The qualitative prediction that it should be less for a mismatch feature than for a perfect match feature in a PM/MM

pair is verified for all three datasets in Fig. 3. Its quantitative behaviour as a function of specific probe-target binding energies, using bulk solution free binding energies as a guide, is verified in Fig. 7.

The parameter K can be thought of as an effective overall reaction rate. It sets the horizontal scale of the isotherm, that is, the scale of the specific target concentration. If an algorithm to determine absolute target concentrations is to be constructed, it is necessary to understand this parameter. Because of the large number of hybridisation reactions which have the potential to contribute it is by far the hardest of the three parameters to explain effectively. The model predicts that it is affected by all of the reactions listed in Table 3 occurring in the bulk hybridisation solution and at the microarray surface, but is unaffected by the dissociation during the washing phase. Analysis to determine which reactions are significant is complicated by the fact that the effect on K of non-specific hybridisation and probe and target folding act in the opposite direction to that of specific binding, so obscuring the effects. Nevertheless, our analysis indicates that all of the hybridisation reactions considered are potentially significant contributors.

A common practice in previous studies [15, 16, 7, 10, 14, 17] has been to invert fits of hyperbolic response curves to recover spike-in target concentrations in order to test the predictive ability of models. We have deliberately refrained from doing so here, as one of the results of this study has been to demonstrate the strong dependence of the parameters of the isotherm on the (in practice unknown) complex non-specific background. Recovering spike-in concentrations using fitting parameters which implicitly contain information about the background belonging to a particular dataset is an inherently circular argument and is guaranteed to give unrealistically good results.

Instead, in Section 6, we address the problem of determining the hyperbolic response function parameters from information likely to be available to biologists in a typical microarray experiment, that is, fluorescence intensities for the complete set of features on each microarray, the probe sequences, and parameters associated with the experimental protocol. We argue that information for the vertical scale parameters is in principle implicitly contained in the distribution of intensities across the microarray by partitioning the intensities by quantities which can be estimated from probe sequences such as probe-target binding energies, probe folding energies and probe pyrimidine content. Determination of the horizontal scale parameter is a more formidable and open problem.

Ultimately, our aim is find practical algorithms for biological analysis through an improved understanding of the physics of microarrays. A problem encountered in this paper, particularly with analysis of the horizontal scale parameter of the isotherm K , has been the difficulty encountered unravelling mutual correlations and compensations between competing effects. In such cases one can never be totally certain, given the available data, that the physical interpretation is correct. However, one could argue that is not necessary, nor possibly even helpful, to know all of the contributing physical

effects in order to meet our ultimate aim. If, for instance, our interpretation of the significance analysis of various models of the parameter K in Section 5.2 is correct, we discover that we often do just as well by modelling a number of physical effects by a single effective term. In general, the lesson is that it is important to strike a balance between attempting to understand and incorporate all physical aspects on the one hand, and relying on empirically determined effective models on the other.

The above analysis has concentrated on Affymetrix gene expression microarrays. Genomic microarrays come in a variety of types for a variety of tasks. As well as gene expression arrays there are tiling arrays, which interrogate large contiguous tracts of genome rather than specific genes, resequencing arrays designed to detect the location of mutations and single nucleotide polymorphisms (SNPs), SNP arrays which are used for genotyping, and non-DNA arrays such as protein arrays designed to identify protein-protein interactions and antibody arrays for detecting antigens. In addition to this there exist a number of technologies including photolithographic deposition, inkjet printing and fabrication of probes onto micro-beads.

In each case they share the common property that they detect large biological molecules of specific known letter sequences via their binding to complementary sequences attached to a solid surface. In each case they will almost invariably share the common problems of non-specific hybridisation, saturation and bulk solution target hybridisation and folding dealt with in this work. The physico-chemical model we have explored is consistent with spike-in data over a broad range of target concentrations and should serve as a starting point for a variety of microarray types and platforms.

Acknowledgement

Thanks to Hans Binder, Susan Wilson and Yvonne Pittelkow for useful discussions and advice.

Notation

- a*: Physical background intensity measurement from factors such as reflection off the microarray surface and photomultiplier dark current. Assumed to be constant for all features on a microarray.
- A*: One of three parameters in the hyperbolic response curve Eq. 1 fitted to the measured fluorescence intensity data. A estimates the (background) fluorescence intensity at zero PM-specific spike-in concentration.
- b*: Saturation fluorescence intensity above the physical background before washing, in a hypothetical situation in which all probes on a feature have formed biotin label carrying duplexes. Assumed to be constant for all features on a microarray.
- B*: One of three parameters in the hyperbolic response curve Eq. 1 fitted to the measured fluorescence intensity data. $A + B$ estimates the asymptotic saturation fluorescence intensity at infinite PM-specific spike-in concentration.

- $I(x)$: Measured fluorescence intensity signal at PM-specific spike-in concentration x .
- K : One of three parameters in the hyperbolic response curve Eq. 1 fitted to the measured fluorescence intensity data. K^{-1} estimates the PM-specific spike-in concentration required to give a fluorescence intensity half way between the background level A and asymptotic level $A + B$.
- $s^S(t_W)$: The specific washing survival function, i.e. the probability that a duplex formed with a PM-feature-specific mRNA target existing at the beginning of the washing step will survive to a washing time t_W .
- $s^{NS}(t_W)$: The non-specific washing survival function, i.e. the probability that a duplex formed with a PM-feature-non-specific mRNA target of species i existing at the beginning of the washing step will survive to a washing time t_W .
- t_W : The washing time.
- x : ($= [S] + [S'] + \sum_i [S.NS_i]$) Spike-in concentration of mRNA PM-specific target.
- $\Delta G^{\text{DNA/RNA}}, \Delta G^{\text{RNA/RNA}}, \Delta G^{\text{Pfold}}$: Binding free energies of a DNA/RNA duplex, a RNA/RNA duplex and of DNA probe self-folding. We use the convention that ΔG is negative for a bound state. Δg^r are the corresponding dimensionless binding free energies $\Delta G^r/(RT)$, where R is the gas constant and T the absolute temperature.
- α : The fraction of probes on the feature carrying probe-target duplexes after a washing time of t_W at zero spike-in concentration x . See Eq. 13.
- β : $\alpha + \beta$ is the fraction of probes on the feature carrying probe-target duplexes after a washing time of t_W at infinite spike-in concentration x . See Eq. 13.
- $\theta(x, t_W)$: The fraction of probes on the feature carrying probe-target duplexes after a washing time of t_W , as a result of a spike-in concentration x of mRNA specific to the PM feature. At $t_W = 0$ is the end of the washing time, that is, at the time of scanning, we write simply $\theta(x)$.
- θ_S, θ_{NS} : the fraction of probes on the feature carrying PM-specific and PM-nonspecific duplexes respectively at $t_W = 0$, i.e., after the hybridisation step and before the washing step.

Glossary

Hybridisation. 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 hybridisation. The hybridisation of target molecules with sequences other than those of the intended sequence. Sometimes ‘non-specific’ is used to mean ‘non-PM-specific’, that is, hybridisation 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. PM and MM are defined in Section 2.

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 hybridised onto microarrays.

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

- [1] http://www.affymetrix.com/support/technical/sample_data/datasets.affx.
- [2] H. Binder. Thermodynamics of competitive surface adsorption on DNA microarrays. *J. Phys. (Condens. Matter)*, 18:S491–S523, 2006.
- [3] H. Binder, T. Kirsten, I. L. Hofacker, P. F. Stadler, and M. Loeffler. Interactions in oligonucleotide hybrid duplexes on microarrays. *Journal of Physical Chemistry*, 108:18015–18025, 2004.
- [4] H. Binder, T. Kirsten, M. Loeffler, and P. F. Stadler. Sensitivity of microarray oligonucleotide probes: variability and effect of base composition. *Journal of Physical Chemistry*, 108:18003–18014, 2004.
- [5] H. Binder and S. Preibisch. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophys. J.*, 89:337–352, 2005.
- [6] H. Binder and S. Preibisch. Genechip microarrays – signal intensities, RNA concentrations and probe intensities. *J. Phys. (Condens. Matter)*, 18:S537–S566, 2006.
- [7] C. J. Burden, Y. E. Pittelkow, and S. R. Wilson. Statistical analysis of adsorption models for oligonucleotide microarrays. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 35, 2004.
- [8] C. J. Burden, Y. E. Pittelkow, and S. R. Wilson. Adsorption models of hybridisation and post-hybridisation behaviour on oligonucleotide microarrays. *J. Phys. (Condens. Matter)*, 18:5545–5565, 2006.
- [9] C. J. Burden, Y. E. Pittelkow, and S. R. Wilson. Statistical analysis and physical modelling of oligonucleotide microarrays. In I. A. Deutsch, L. Bruschi, H. Byrne, G. de Vries, and H.-P. Herzel, editors, *Mathematical Modeling of Biological Systems*, volume I, pages 333–346. Birkhäuser, Boston, MA, USA, 2007.
- [10] E. Carlon and T. Heim. Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays. *Physica A*, 362:433–449, 2006.
- [11] E. Carlon, T. Heim, J. Klein Wolterink, and G.T. Barkema. Comment on “Solving the riddle of the bright mismatches: Labelling and effective binding in oligonucleotide arrays”. *Phys. Rev. E*, 73:063901, 2006.
- [12] J. E. Forman, I. D. Walton, D. Stern, R. P. Rava, and M. O. Trulson. Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesised oligonucleotide arrays. In N. B. Leontis and J. SantaLucia, editors, *Molecular Modeling of Nucleic Acids, ACS Symposium Series*, volume 682, pages 206–228. Am. Chem. Soc., Washington, DC, USA, 1998.
- [13] A. Halperin, A. Buhot, and E. B. Zhulina. Sensitivity, specificity and the hybridization isotherms of DNA chips. *Biophysical Journal*, 86:718–730, 2004.
- [14] T. Heim, L.-C. Tranchevent, E. Carlon, and G.T. Barkema. Physical-chemistry-based analysis of Affymetrix microarray data. *J. Phys. Chem. B*, 110:22786–22795, 2006.
- [15] D. Hekstra, A. R. Taussig, M. Magnasco, and F. Naef. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research*, 31:1962–1968, 2003.
- [16] G. A. Held, G. Grinstein, and Y. Tu. Modeling of DNA microarray data by using physical properties of hybridization. *Proceedings of the National Academy of Science*, 100:7575–7580, 2003.
- [17] G. A. Held, G. Grinstein, and Y. Tu. Relationship between gene expression and observed intensities in DNA microarrays - a model study. *Nucleic Acids Research*, 34:e70, 2006.
- [18] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, and et al. Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.
- [19] O.V. Matveeva, S.A. Shabalina, V.A. Nemtsov, A.D. Tsodikov, R.F. Gesteland, and J.F. Atkins. Thermodynamic calculations and statistical correlations for oligo-probe design. *Nucl. Acids Res.*, 31:4211–4217, 2003.
- [20] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: Labelling and effective

- binding in oligonucleotide arrays. *Phys. Rev. E*, 68:011906, 2003.
- [21] D. Skvortsov, D. Abdueva, C. Curtis, B. Schaub, and S. Tavaré. Explaining differences in saturation levels for Affymetrix Genechip® arrays. *Nucleic Acids Research*, 35:4154–4163, 2007.
- [22] N. Sugimoto, S. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, and T. Ohmichi. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, 34:11211–11216, 1995.
- [23] T. Xia, J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochem.*, 37:14719–14735, 1998.
- [24] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.*, 31:3406–3415, 2003. <http://frontend.bioinfo.rpi.edu/applications/mfold/>.