Nucleic Acids Research, 2006, Vol. 34, No. 9 ***e70***
doi:10.1093/nar/gkl122

Relationship between gene expression and observed intensities in DNA microarrays—a modeling study

G. A. Held*, G. Grinstein and Y. Tu

IBM TJ Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA

Received October 28, 2005; Revised December 28, 2005; Accepted March 13, 2006

ABSTRACT

A theoretical study of the physical properties which determine the variation in signal strength from probe to probe on a microarray is presented. A model which incorporates probe-target hybridization, as well as the subsequent dissociation which occurs during stringent washing of the microarray, is introduced and shown to reasonably describe publicly available spike-in experiments carried out at Affymetrix. In particular, this model suggests that probe-target dissociation during the stringent wash plays a critical role in determining the observed hybridization intensities. In addition, it is demonstrated that non-specific hybridization introduces uncertainties which significantly limit the ability of any model to accurately quantify absolute gene expression levels while, in contrast, target folding appears to have little effect on these results. Finally, for data from target spike-in experiments, our model is shown to compare favorably with an existing statistical model in determining target concentration levels.

INTRODUCTION

DNA microarrays allow the measurement of the expression levels of thousands of genes simultaneously (1,2). Despite the widespread use of microarrays, however, there remains much uncertainty in the quantitative determination of expression levels obtained using these technologies. The raw data from these assays often exhibit large, unexplained fluctuations, and the methods used to infer expression levels are typically empirical or statistical in nature.

The central component of all DNA microarray technologies is an array of different oligonucleotides (each between 20 and several hundred bases long), called probes, deposited onto a single substrate. A solution containing a labeled nucleic acid sample is brought into contact with the substrate, and each transcript in the sample hybridizes to those probes that are complementary to it. Following hybridization, the substrate is washed, so as to remove unbound and weakly bound (i.e. mismatched) target oligonucleotides from the sample. Bound target molecules are then stained with a fluorophore and the microarray is scanned. Detection of the fluorescent signal allows one to quantify the presence of various sequences, and thus the expression levels of various genes, within the sample.

One class of DNA microarrays is exemplified by Affymetrix Genechips (2,3), wherein each transcript is probed by multiple, short oligomers (typically 25mers). Typically, ~16 perfect match (PM) probes, and an equal number of mismatch (MM) probes correspond to each transcript. Each PM probe exactly complements a short region of the transcript, referred to as the target. The corresponding MM probe is identical to the PM probe except at its centermost base, which is non-complementary to the transcript. Experimentally, it is readily observed that the degree of hybridization varies significantly between those multiple PM probes which hybridize to different regions of a common transcript (3).

The Affymetrix Microarray Suite version 5.0 (MAS v5) algorithm for inferring quantitative transcript expression levels from microarray data begins by subtracting the MM intensity from the corresponding PM intensity (with adjustments to the MM value if MM>PM). The expression level is then taken to be a weighted average of those differences obtained from all of the pairs that probe the given transcript (4).

More recent statistical algorithms have incorporated a probe binding affinity which accounts, in part, for the observed variation in intensity between the different probes which are designed to bind to different regions of a particular transcript (5–7). In addition, there have been efforts to quantify observed probe intensities through physical modeling of the hybridization process (8–13). Thus far, these 'physical' models have yielded expression levels with accuracies comparable with, but not significantly greater than, purely statistical models (14,15). Physical modeling is nonetheless important in that it can provide a means of inferring absolute expression levels more accurately—both by providing guidance in the optimum choice of probe sequences and in suggesting potential modifications to assay protocols and data analysis. Furthermore, a

*To whom correspondence should be addressed. Tel: +1 914 945 2609; Fax: +1 914 945 2141; Email: gaheld@us.ibm.com

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

model which correctly incorporated all of the physical properties which determine the observed hybridization intensities would, in principle, yield gene expression levels with an accuracy limited only by noise intrinsic to the measurements (16,17).

While several physical models have addressed the effects of hybridization energy (11,12) and probe site saturation (8,9), no model thus far has considered the effects of target-probe dissociation during the washing of the microarray which follows hybridization. In this paper, we present a physical model of microarray hybridization which incorporates both the binding of target and probe during the hybridization phase of the assay and the dissociation of target and probe during the washing step. We demonstrate that variations in the saturation intensity of different probes are a central feature of microarray behavior and show that these variations can largely be explained by the varying degrees of dissociation which occur during the washing of the microarray.

We find that much of the uncertainty in determining expression levels is a consequence of the uncertainties in background signal—uncertainties which result from the non-specific binding of probes to oligonucleotides other than their intended targets (18–21). In addition, we present evidence that probe and target folding do not significantly affect the observed hybridization intensities. Finally, we demonstrate that our model compares favorably with MAS v5 in determining target expression levels.

EXPERIMENTS AND DATA

All of the data shown and analyzed in this paper are taken from the publicly available results of experiments carried out at Affymetrix [www.affymetrix.com/analysis/download_center2.affx, (22)]. In particular, the data are the results of a series of controlled, 'spike-in' experiments, in which a transcript group comprised of known concentrations (each between 0 and 1024 pM), of 14 human genes is spiked into a background consisting of a mixture of mRNA from a human pancreatic tissue source. (None of the spike-in genes was expressed in the tissue source.) This mixture is then labeled and hybridized onto Affymetrix U95A Genechips following the manufacturers protocols (23) (T. A. Webster, personal communication). Fourteen different transcript groups, each containing different concentrations of the various spike-in genes (following an experimental design known as a Latin Square), are then each hybridized onto a genechip. The net result is that each of the genes is spiked into one of the transcript groups at each of the following concentrations: 0, 0.25, 0.5, 1, 2, 4, . . . , 1024 pM; i.e. data are collected for all 14 genes, each at 14 different concentrations. Sixteen distinct PM/MM probe pairs interrogate each transcript. Finally, each of the measurements for a given gene at a given concentration was replicated between 2 and 12 times. The raw data provided by Affymetrix contain intensities for each PM and MM probe (the intensities provided corresponding to the 75th percentile of the intensity of the scanned pixels associated with the probe site). We performed no normalization of this data prior to the analysis discussed below. In addition, we consider only the PM probes in our analysis. Note that throughout this paper we use Affymetrix notation to identify genes.

MODELS OF HYBRIDIZATION

During hybridization, the formation of bound target-probe duplexes is governed by the equation:

$$[\text{single-strand probe}] + [\text{single-strand target}] \rightleftharpoons [\text{target-probe duplex}]. \quad \mathbf{1}$$

Assuming $k_f^{(h)}$ and $k_b^{(h)}$ are the forward and backward rate constants for this reaction under hybridization conditions, and the concentrations [single-strand probe], [single-strand target] and [target-probe duplex] are given by $(n_{\text{probe}} - n_B)/V_{\text{probe}}$, $(n_0 - n_B)/V_{\text{total}}$ and n_B/V_{probe}, where n_{probe}, n_0, V_{probe} and V_{total} are equal to the number of probe molecules at the given probe site, the number of transcript molecules in the target solution, the volume of the probe site and the volume of the target solution, respectively, then the number of bound target-probe pairs, n_B, is determined by the rate equation as follows.

$$\frac{\partial n_B}{\partial t} = k_f^{(h)}(n_{\text{probe}} - n_B)\left(\frac{n_0 - n_B}{N_A V_{\text{total}}}\right) - k_b^{(h)} n_B. \quad \mathbf{2}$$

N_A is Avogadro's number and V_{probe} is defined as the area of a probe site multiplied by the average probe height. Note that since V_{probe} does not appear in Equation 2, the precise definition of the average probe height is not critical. Assuming that the system achieves equilibrium (24), and that $n_{\text{probe}} \ll n_0$ [$n_{\text{probe}} \approx 10^7$ (2,24) and $n_0 \approx 2 \times 10^8$ for a 0.25 pM target solution (23)], it follows that

$$n_{B,\text{equilibrium}} = n_{\text{probe}} \frac{c}{(K_d + c)}, \quad \mathbf{3}$$

where $K_d \equiv k_b^{(h)}/k_f^{(h)}$ and $c \equiv (n_0/N_A)/V_{\text{total}}$ is the target concentration in moles/liter. It follows from thermodynamics that the dissociation constant, K_d satisfies the following equation:

$$K_d = e^{\Delta G_{\text{hyb}}/RT}, \quad \mathbf{4}$$

where ΔG_{hyb} is the change in free energy associated with the target-probe hybridization (25) and the units of K_d are moles/liter. Note that in earlier work (9) we referred to K_d as n_e. Note too that when the system approaches equilibrium (24), the concentrations of the target molecules in solution become spatially uniform, whereupon diffusion of the targets ceases to play a role, and the simple form Equation 2 of the rate equation for hybridization is justified (26,27).

Target hybridization is followed by washing of the probe array, a process wherein the array is flushed with a non-stringent wash buffer ($[Na^+] = 1$ M), followed by a stringent wash buffer ($[Na^+] = 0.1$ M), followed in turn by a second non-stringent wash buffer (23). In the course of this washing, unbound and weakly bound (i.e. mismatched) target molecules are removed from the genechip. In addition, some complementary bound target molecules will also dissociate; the number of bound target molecules at a given probe site will be reduced by a factor of $\exp(-k_b^{(w)}t)$, where $k_b^{(w)}$ is the backwards rate constant for Equation 1 under the conditions of the stringent wash, t is the time duration of the stringent wash and we have assumed that the dissociation of complementary bound targets occurs primarily during the

stringent wash. Note that in the course of this wash, the wash buffer is repeatedly replaced (23), whereupon the concentration of dissociated target molecules in the washing fluid is so small that one may assume they are all removed from the system. Therefore, one need not worry about the diffusive process that carries the newly released targets away, but only about the fraction of such targets, which is controlled by $k_b^{(w)}$.

The rate constant $k_b^{(w)}$ may be expressed as $k_b^{(w)} = k_f^{(w)} \exp(\Delta G_{wash}/RT)$, where $k_f^{(w)}$ is the forward rate for Equation 1 under the stringent washing conditions and ΔG_{wash} is the change in free energy associated with duplex formation under stringent wash buffer conditions. The rate constant $k_f^{(w)}$ is observed to be relatively independent of oligonucleotide sequence both in solution (28) and on microspheres (29), whereas $k_b^{(w)}$ can vary by many orders of magnitude (28).

Following washing, the bound target molecules are stained with a fluorophore and the microarray is scanned. Combining the probe-target dissociation resultant from washing with Equation 3, and recasting the result to yield observed intensity, I, as a function of concentration, c, we obtain the following equation:

$$I = \frac{An_{probe} e^{-k_b^{(w)} t} c}{(K_d + c)} + bg_e, \quad \mathbf{5}$$

where we have added a sequence dependent background term bg_e to account for the hybridization of probes to nucleic acids other than their intended targets, and the proportionality constant A in Equation 5 relates n_B to the corresponding fluorescence intensity.

For oligomers in solution, the change in free energy associated with hybridization is well described by a nearest-neighbor (nn) stacking energy model (28,30,31). In nn models, the hybridization free energy of any base pair depends not only on whether that pair is a C–G or an A–T, but also on which base pairs occupy the neighboring positions along the strand. To calculate the hybridization free energy for any sequence, one sums the contributions for each of these stacked pairs along the chain, and then adds a correction for the base pairs terminating the sequence at each end. For DNA/RNA duplexes (such as those present on hybridized Affymetrix microarrays) there are 16 independent stackings of 2 base pairs along an oligonucleotide chain, $\varepsilon(b_1,b_2)$, where $b_{1,2}$ = A, C, G, T. Thus, the free energy of hybridization of a 25mer DNA/RNA duplex in solution may be written as follows.

$$\sum_{i=1}^{24} \varepsilon(b_i, b_{i+1}), \quad \mathbf{6}$$

where we assume that the corrections for the base pairs terminating the sequence at each end may be neglected.

In the microarray geometry, one expects sequence independent factors such as changes in the electrostatic energy (32–36) of the system to contribute to the free energy of hybridization. For this reason, we define $\tilde{\varepsilon}(b_1,b_2) \equiv \varepsilon(b_1,b_2) + (\Delta\gamma_{hybrid}/24)$, where $\Delta\gamma_{hybrid}$ is a sequence independent change in free energy associated with hybridization. As our model does not allow us to independently determine the 16 $\varepsilon(b_1,b_2)$ and $\Delta\gamma_{hybrid}$ (see Discussion), we have chosen to incorporate a fraction $\Delta\gamma_{hybrid}/24$ into each of the nn energy terms.

Hybridization on a microarray differs from hybridization in solution in that the contributions of paired nucleotides to the energy of hybridization are position dependent. In solution, each of the hybridized base pairs contributes equally to the energy of hybridization, as in Equation 6. However, it has been shown that, in a microarray geometry, those bases closest to the center of the oligonucleotide chain contribute most significantly to the energy of hybridization; the contributions of the other pairs to the energy of hybridization falls off roughly parabolically relative to the contribution of the centermost nucleotides (11,12). Incorporating both a sequence independent term and positional weighting into our model of hybridization, we find

$$\Delta G_{hyb} = \sum_{i=1}^{24} w_i [\tilde{\varepsilon}(b_i, b_{i+1})]. \quad \mathbf{7}$$

In the models which follow, we consider both weighting factors w_i which are constant along the oligonucleotide (i.e. $w_i = 1$ for all i) and ones which are parabolically weighted and centered at the middle of the probe oligonucleotide. For this second case, we express the weighting factors as a function of a, the curvature of the parabola:

$$w_i = \frac{a(i - 12.5)^2}{2} + 1 \quad \mathbf{8}$$

As noted earlier, the change in free energy ΔG_{hyb}, associated with the formation of a given duplex under the hybridization conditions differs from the free energy associated with the same reaction under the stringent washing conditions, ΔG_{wash}. However, nn stacking energy models of hybridization in solution predict that the change in free energy of hybridization resulting from a change in salt concentration should be sequence independent (30). In addition, the change in free energy associated with electrostatic interactions at the surface of a microarray are also expected to be sequence independent (32,35). Given these observations, we assume the relationship

$$\Delta G_{wash} = \Delta G_{hyb} + \Gamma, \quad \mathbf{9}$$

where Γ is independent of oligonucleotide sequence. Note that one effect of sequence-specific probe folding would be to render the surface electrostatic interactions probe-sequence dependent. However, given that the average spacing between probe molecules is only of order 25 Å (24,37,38), we assume that the effects of probe folding are minimal.

Given Equations 7–9, we can rewrite Equation 5 as follows:

$$I = \frac{I_p c}{(K_d + c)} + bg_e, \quad \mathbf{10}$$

where the factor I_p, defined as

$$I_p = An_{probe} e^{-k_f^{(w)} t \exp((\Gamma + \Delta G_{hyb})/RT)} \quad \mathbf{11}$$

is a function of probe sequence. The dependence of intensity on concentration in Equation 10 is known as a Langmuir isotherm (25), a function which describes the adsorption of neutral adsorbates at a surface comprising a finite number of sites, each capable of accommodating but a single adsorbate.

Note that in a model which considers neither the disassociation of target molecules during the stringent wash nor other sequence dependent effects such as self-hybridization of probes, one would expect I_p to be a constant independent of probe sequence (8,9). Note also that the intensity I defined by Equation 10 is the total intensity from a given probe spot. If one wanted to apply this model to microarrays with variable spot sizes, it would be necessary to express this equation in terms of average intensity, by dividing I, I_p and bg_e by the area of each probe site. For the case of the Affymetrix Genechips treated here, all of the probe sites are the same size, so such a conversion is not necessary.

MODELING OF PROBE INTENSITY DATA

General considerations

At low concentrations of target molecules, the Langmuir isotherm (Equation 10) is characterized by a linear dependence of hybridization fraction (and, thus, observed intensity) on target concentration, while at higher target concentrations the number of bound target molecules saturates at a constant value. In Figure 1a is plotted the experimentally observed fluorescent intensity from bound target molecules as a function of target concentration for a single probe (gene 37777_at, probe 16), with each data point corresponding to an average of the intensities obtained from between 2 and 12 replicate measurements. (The intensities of the individual replicate measurements are shown in Figure 1b.) The data in this figure are well described by Equation 10, the solid line being a best-fit to this equation, where I_p and K_d are adjustable parameters and bg_e is fixed at the average value of the zero concentration spike-in measurements for this probe.

We begin our analysis of the data by noting that, given accurate values of I_p and K_d for a sufficient number of probe sequences, one can perform a least squares fit to these values of I_p and K_d using Equations 4 and 11 (with the free energies defined by Equations 7–9). In doing these fits, one obtains best-fit values for the 16 nn energy parameters $\varepsilon(b_1,b_2)$, as well as for $I_{probe} \equiv An_{probe}$, Γ and a. From these best-fit values, one may then predict values of I_p and K_d (using Equations 4, 7–9 and 11) for any given probe sequence. Finally, from these results one may, in turn, generate from Equation 10 model-based predictions of intensity as a function of probe sequence and target concentration.

To obtain reliable values of I_p and K_d for a broad range of probe sequences, we begin by plotting intensity as a function of target concentration for each probe in the Latin Square data. This yields 224 (14 genes each with 16 probes) plots equivalent to Figure 1a. Next, we identify those plots from which we can obtain reliable estimates of both I_p and K_d. We set as our first criterion that the relationship between intensity and target concentration approximately follow a Langmuir isotherm. This results in the exclusion of data from defective probe spots [such as those identified by Affymetrix (www.affymetrix.com/analysis/download_center2.affx) because they exhibited weak fluorescence with no discernible correlation between target spike-in concentration and spot intensity]. We fit the remaining data to Equation 10, with the background bg_e for each probe spot fixed at the experimentally observed value (i.e. the average of all of the zero concentration spike-in replicates for that probe). We then limit our consideration to those probes for which the best-fit value of I_p is <1.1 times greater than the experimentally observed intensity at 1024 pM. This criterion is necessary because it is not possible to accurately estimate the saturation value I_p from data which is not close to saturation at the highest measured target concentrations. In total, we find that 95 probes satisfy the above criteria, and it is these probes which we use in the subsequent analysis. Supplementary Figure 1 comprises plots of intensity as a

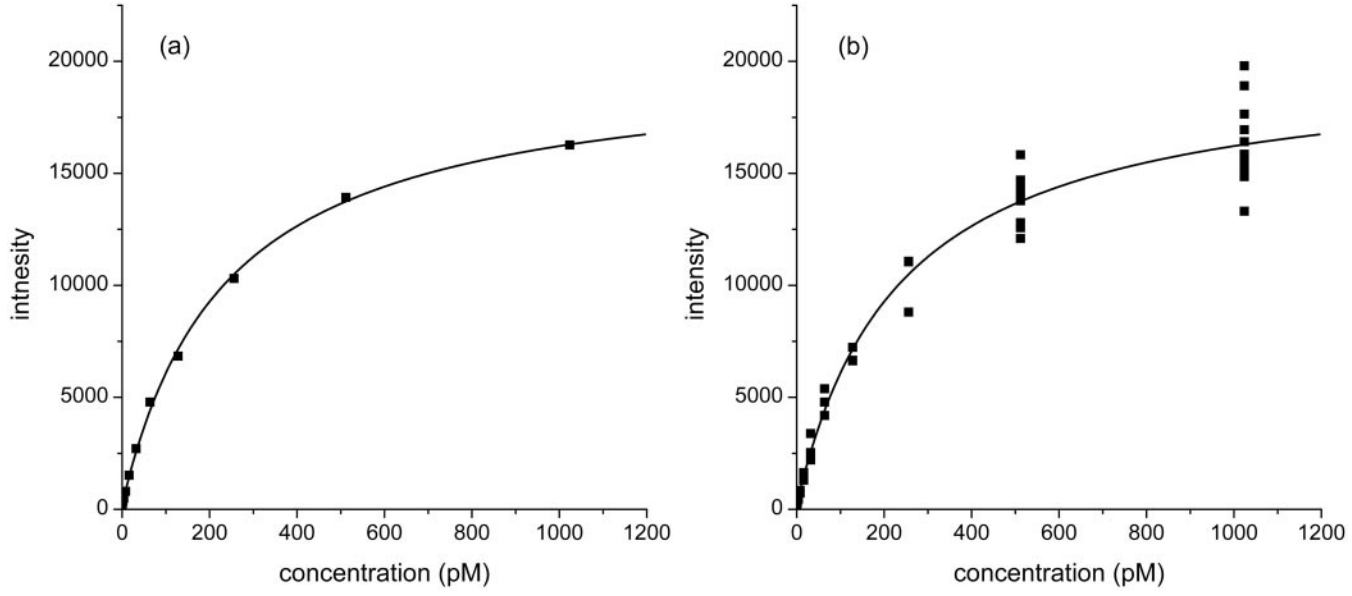

Figure 1. Observed hybridization intensity as a function of spike-in target concentration of PM probe 16 of gene 37777_at. The solid line is a best-fit of the data to Equation 10 with I_p and K_d as adjustable parameters and bg_e set at the experimentally observed average signal for data taken at zero spike-in concentration. In (**a**) each data point shown is the average of all of the replicate measurements taken at a given spike-in concentration. Each of the replicate measurements is shown as a separate data point in (**b**).

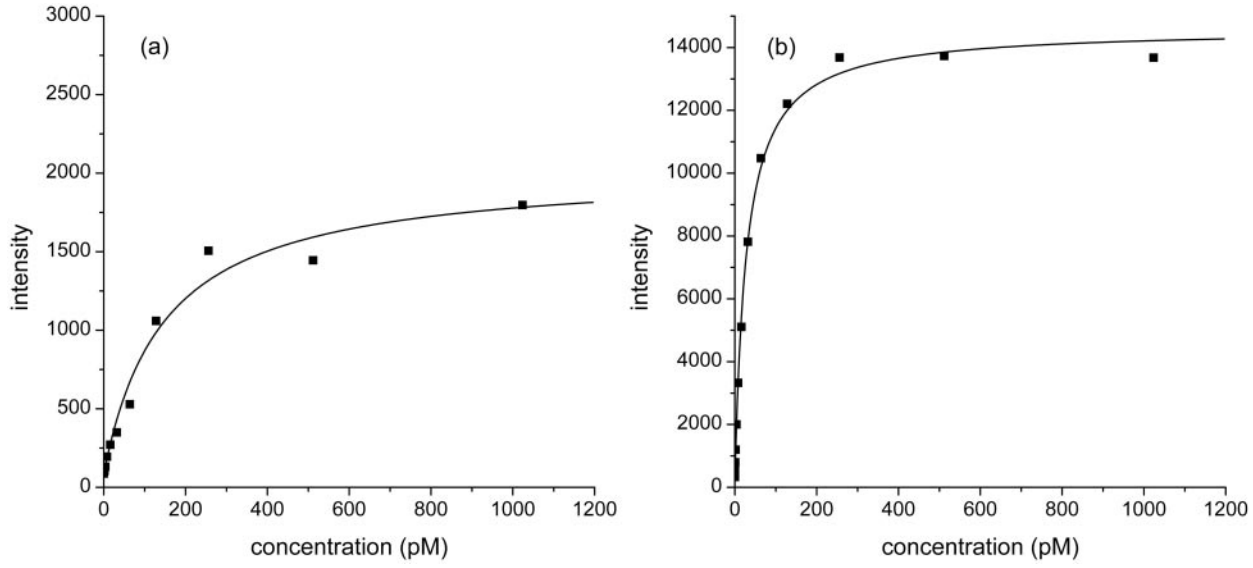

Figure 2. Observed hybridization intensity as a function of spike-in target concentration of (**a**) PM probe 9 of gene 1597_at and (**b**) PM probe 15 of gene 37777_at. The solid lines are best-fits of the data to Equation 10 with I_p and K_d as adjustable parameters and bg_e set at the experimentally observed average signal for data taken at zero spike-in concentration. Each data point shown is the average of all of the replicate measurements taken at a given spike-in concentration.

function of target concentration for these 95 probes. The black line in each plot is a best-fit of the data to Equation 10.

In Figure 2 are shown plots of intensity as a function of target concentration for two probes (gene 1597_at, probe 3 and gene 37777_at, probe 15). The best-fit value of the saturation intensity I_p for the data in Figure 2a is 1942, a factor of 10.2 smaller than that for the data shown for a different probe in Figure 1a. Likewise, the best-fit value of 28.5×10^{-12} mol/l obtained for K_d from the data in Figure 2b is 8.3 times smaller than that observed in Figure 1a. In Figure 3a and b we plot histograms of the values of I_p and K_d obtained by fitting the data for the 95 probes in Supplementary Figure 1 to Equation 10. The data in Figures 1–3 clearly demonstrate that any physical model which purports to explain the observed gene expression data must be able to account for significant probe to probe variations in both K_d and the saturation intensity I_p.

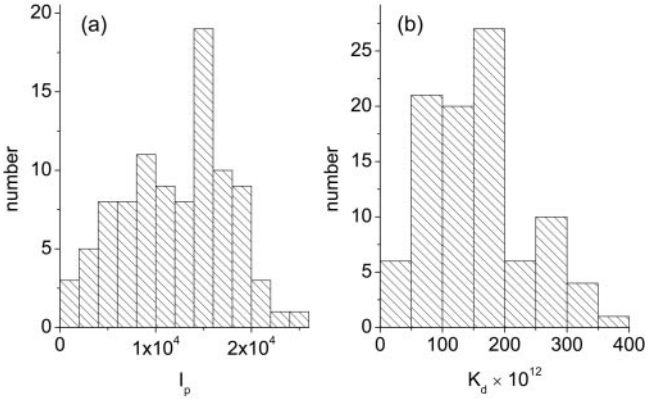

Figure 3. Histograms of the values of (**a**) I_p and (**b**) K_d obtained by fitting the hybridization intensity as a function of spike-in concentration for 95 selected PM probes (see text) to Equation 10 with I_p and K_d as adjustable parameters and bg_e set at the experimentally observed average signal for data taken at zero spike-in concentration. Units of K_d are mol/l.

Physical model

We now introduce and study a physical model for fitting the gene expression data. In addition, we explore the importance of the various elements of this model by considering variants of it wherein a particular element is modified. Comparison of the results of the original model with those of the variants provides insight into the significance of the modified elements.

Model I. As our main model, hereafter referred to as Model I, we simultaneously fit the values of K_d and I_p obtained for the 95 good probes to Equations 4 and 11, respectively, using the sixteen $\tilde{\varepsilon}(b_1, b_2)$, $I_{probe} \equiv An_{probe}$, Γ and a as adjustable parameters. The best-fit values for the energies $\tilde{\varepsilon}(b_1, b_2)$ are plotted in Figure 4, while the best-fit values for I_{probe}, Γ and a are 79 410, 1.95 kcal/mol and -0.0116, respectively. It may be seen in Figure 4 that the best-fit values of complementary energy terms [e.g. $\tilde{\varepsilon}(A, G)$ and $\tilde{\varepsilon}(C, T)$] are not equal. Such differences are found for DNA/RNA hybridization in solution (31), and so are not unexpected for hybridization of the DNA probe to the RNA target. These differences may also be due in part to the fact that on the target RNA only pyrimidine nucleotides are biotinylated (and subsequently fluorescently labeled); similar differences between the hybridization energies associated with complementary nucleotides have been attributed to this fact in previous studies (11,39).

In performing these fits, we have assumed a value of 10^6 $(\text{mol/l})^{-1}$ s^{-1} for $k_f^{(w)}$ (28) and set t to the experimental value of 600 s (23). The precise value assumed for $k_f^{(w)}$ does not alter the quality of the fit, the effect of a different assumed value for $k_f^{(w)}$ being to change the best-fit value of Γ accordingly. We have chosen not to allow $k_f^{(w)}$ to vary with probe sequence because experimental evidence suggests that it varies far less than $k_b^{(w)}$ (28,29). In addition, it is unclear precisely how one would model the dependence of $k_f^{(w)}$ on probe sequence without rendering the fitting process intractable. Note that in simultaneously fitting K_d and I_p it is necessary

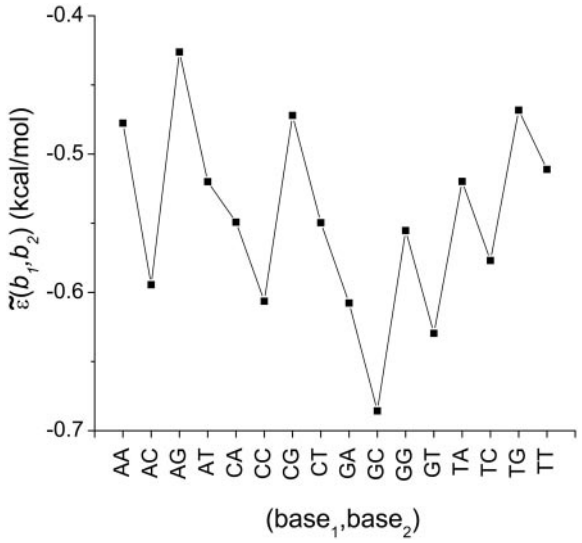

Figure 4. Best-fit values of the energies $\tilde{\varepsilon}(b_1,b_2)$ for $b_{1,2}$ = A,C,G,T obtained by simultaneously fitting the values of I_p and K_d for 95 selected PM probes to Equations 4, 7–9 and 11 using the 16 $\tilde{\varepsilon}(b_1,b_2)$, I_{probe}, a and Γ as adjustable parameters (see text).

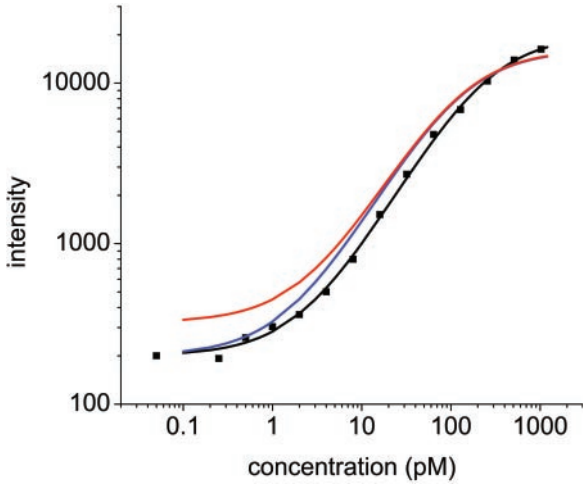

Figure 5. Observed hybridization intensity, averaged over all replicate measurements, as a function of spike-in target concentration of PM probe 16 of gene 37777_at, plotted on a log–log scale. The zero concentration (leftmost) data point is plotted at 0.05 pM. The black line is a best-fit of the data to Equation 10 with I_p and K_d as adjustable parameters and bg_e set at the experimentally observed intensity at zero spike-in concentration. The blue and red lines are plots of Equation 10 with I_p and K_d calculated from the probe sequence using Model I, discussed in the text, and bg_e set at the experimentally observed intensity at zero spike-in concentration (blue line), or determined by Equation 13 (red line).

to choose an arbitrary proportionality constant to weight the importance of a unit of least square error in K_d relative to a unit of least square error in I_p. We have chosen this factor to be 60, as for this value the quality of the fits to K_d and I_p appear qualitatively comparable.

To determine the accuracy of Model I, we use our best-fit values of $\tilde{\varepsilon}(b_1,b_2)$, I_{probe}, Γ and a to calculate ΔG_{hyb} and ΔG_{wash} from Equations 7–9 for each of the 95 good probe sequences. From these, we then calculate predicted values of K_d and I_p from Equations 4 and 11, respectively. Finally, we incorporate these predicted values of I_p and K_d with the experimentally observed values for bg_e into Equation 10 and obtain predicted intensities for each probe at each experimentally observed concentration between 0.25 and 1024 pM. These predicted intensities are shown as solid blue lines in Figure 5 for one probe (gene 37777_at probe 16) and in Supplementary Figure 1 for all of the 95 probes used in the analysis.

We quantify the difference between these predicted intensities and the experimentally observed intensities by calculating a normalized sum log difference square *lds* defined as follows.

$$lds \equiv \frac{1}{N}\sum_{i=1}^{95}\sum_{c=0.25}^{1024pM}\left[\log\left(I_{predicted}(i,c)\right) - \log\left(I_{observed}(i,c)\right)\right]^2, \quad \textbf{12}$$

where $I_{predicted}(i,c)$ and $I_{observed}(i,c)$ are the predicted and experimentally observed intensities for probe i and spike-in target concentration c, and $N \equiv 95 \times 13 = 1235$ normalizes *lds* with respect to the number of predicted intensities included in the double sum. We calculate the difference between the natural logarithms of the intensities, as opposed to the linear difference, so as to give equal weight to errors of equal magnitude at all of the measured target concentrations. As a second measure of the accuracy of our model, we calculate the Pearson correlation coefficient between all of the values of $\log(I_{observed})$ and $\log(I_{predicted})$. The values of the *lds* and correlation coefficient for this model are given in the first row of Table 1. We use these measures below to compare the accuracy of this model with several variants of the model which are based on different physical assumptions. In doing so, we are able to identify those attributes of the model which contribute most significantly to agreement between model and experiment.

The normalized sum *lds* represents the typical value of $[\log(I_{predicted}/I_{observed})]^2$ for a given probe at a given concentration. The quantities $\exp(\sqrt{lds})$ and $\exp(-\sqrt{lds})$ therefore provide a rough measure of the range over which the ratio of predicted to observed intensities, $I_{predicted}/I_{observed}$, can be expected to vary. For Model I, where $lds = 0.17$, e.g. this ratio should range between 0.67 and 1.5.

The extent to which our model captures the experimentally observed variations in I_p and K_d may be seen in Figure 6a and b, in which the values of I_p and K_d (as obtained by fitting the experimental data) are plotted as a function of ΔG_{hyb}, where this free energy of hybridization is determined from Equation 7 and the best-fit values of $\tilde{\varepsilon}(b_1,b_2)$ and a obtained from Model I. The predicted dependencies of I_p and K_d on ΔG_{hyb} (from Equations 11 and 4, respectively), are shown as solid red lines in these figures. Note that in Figure 6 we have assigned error bars to the experimentally observed values of I_p and K_d. These error bars are defined to be the greater of two estimates of uncertainty. The first is simply the statistical uncertainty (1 SD) of the best-fit values of I_p and K_d obtained by fitting the replicate averaged data of Figure 1a and Supplementary Figure 1 to Equation 10. The second estimate is the standard deviation of the set of values for I_p and K_d obtained by dividing the set of replicate data for each probe into three groups—each one corresponding to the replicate measurements taken from the genechips grown on a specific wafer

Table 1. Sum log square difference (*lds*) (as defined in text) and Pearson correlation coefficient between the logarithms of the observed and predicted probe intensities are shown for Model I and several variants

Model Number	Model parameters				Sum log difference square (*lds*)	Pearson correlation coefficient	Figure
	I_p	K_d	Background bg_e	Positional weighting w_i			
I	Fit by model	Fit by model	Observed values	Parabolic	0.170	0.963	6a and b
I: Variant 1	Fit by model	Fit by model	Observed values	None	0.204	0.955	6c and d
I: Variant 2	Best-fit to constant	Fit by model	Observed values	Parabolic	0.321	0.928	6e and f
I: Variant 3	Fit by model	Best-fit to constant	Observed values	Parabolic	0.148	0.968	6g and h
I: Variant 4	Fit by model	Fit by model, also spike-in concentrations between genes allowed to vary	Observed values	Parabolic	0.128	0.972	6i and j
I/fit_bg	Fit by model	Fit by model	Equation13	Parabolic	0.224	0.952	6a and b

Models and variants are distinguished through the methods by which I_p, K_d, bg_e and w_i are determined; details are described in the text. Panels of Figure 6 which utilize the values of ΔG_{hyb} derived from each model or variant are indicated in the last column. Parabolic positional weighting (column 5) indicates that values of w_i were determined by a best-fit to Equation 8. No positional weighting indicates values of unity for all w_i. Observed values of background (column 4) are the average of zero spike-in concentration replicates.

(www.affymetrix.com/analysis/download_center2.affx)—and fitting each group separately to Equation 10. These error bars are intended to provide a qualitative sense of the degree of uncertainly associated with the fitted values for I_p and K_d; they are not defined in a statistically rigorous fashion. For this reason, we do not utilize them in determining our best-fit values of the various adjustable parameters. Nonetheless, the magnitude of these error bars does capture the extent to which variability between replicate measurements (17) (such as that shown in Figure 1b) limits the ability of any model to precisely describe the data.

We now calculate the Pearson correlation function and *lds* between the predicted and observed probe intensities for several variants of Model I. By observing the effect that modifying the model has on the reliability with which it predicts the observed data, we are able to identify those physical processes which are most important to incorporate into a model of microarray hybridization. We emphasize that there is no a priori reason for making these modifications, which are studied simply to provide insight into the practical significance of the various elements of Model I.

Variant 1 of Model I. In this first variant of Model I, we take the contribution of nucleotides to the energy of hybridization to be independent of position along the probe, i.e. we fix the values of all the weights w_i in Equation 7 to unity. This change increases the *lds* from 0.17 to 0.204—indicating that, as reported previously (11,12), weighting the energetic contributions of base pairs in the center of the probe more highly than those at the ends of the probe does indeed improve the accuracy of the model. In Figure 6c and d we plot I_p and K_d as a function of the values of hybridization energy ΔG_{hyb} calculated from the best-fit parameters obtained for Variant 1.

Variant 2 of Model I. This second variant of Model I addresses the importance of allowing the dissociation constant $k_b^{(w)}$, and, thus, I_p, to vary with probe sequence. We do this by allowing the parameters $\tilde{\varepsilon}(b_1, b_2)$ and the curvature a to be determined by a least squares fit of only the values of K_d of the 95 probes to Equations 4, 7 and 8. The probe intensities are then determined using Equation 10, where I_p is fixed at a constant value independent of probe sequence. The value of I_p is chosen so as to minimize the value of *lds*. This model results in an *lds* value of 0.321—approximately twice that observed when one allows I_p to vary with probe sequence. This result demonstrates the importance of incorporating into any model a physical mechanism through which I_p can vary with probe sequence. In Figure 6e and f we plot I_p and K_d as functions of the values of hybridization energy ΔG_{hyb} calculated from the best-fit parameters obtained using Variant 2. As may be seen in Figure 6e, the values of I_p are not correlated with ΔG_{hyb}, as expected given that they were not utilized in obtaining these values of the free energy. The blue line in Figure 6e is the best-fit value of I_p (10 700) while the green line is the average of the 95 I_p values obtained by fitting the data for each probe individually to Equation 10 (12 065).

In earlier work (9), we modeled the same set of spike-in data using a fixed value of I_p (as in Variant 2). However, rather than fitting the nn stacked base pair energies, we used experimental values obtained from measurements of hybridization in solution (30). We found that in order to obtain a reasonable fit of the data to Equation 10, it was necessary to introduce a phenomenological constant multiplying ΔG_{hyb} in the expression for K_d (Equation 4) as an extra fitting parameter. This underscores the difficulty of accounting for the data without any *ad hoc* assumptions unless one considers processes, such as washing, which allow for variations in I_p from probe to probe.

Variant 3 of Model I. Next we investigate the importance of allowing K_d to vary with probe sequence, by considering a variant—Variant 3—wherein the parameters $\tilde{\varepsilon}(b_1, b_2)$, a, I_{probe} and Γ are derived from a best-fit of only the values of I_p of the 95 probes to Equations 7–9 and 11, while a probe sequence independent K_d is chosen so as to minimize the value of *lds*. This amounts to ignoring the probe sequence-dependence of the hybridization process, but including the sequence dependence of the dissociation during microarray washing. In this case, the value of *lds* is 0.148, smaller than that for a model which uses the same probe dependent energy for duplex formation during both the hybridization and washing processes. While taking K_d to be independent of probe sequence seems a rather drastic approximation, the low value of *lds*

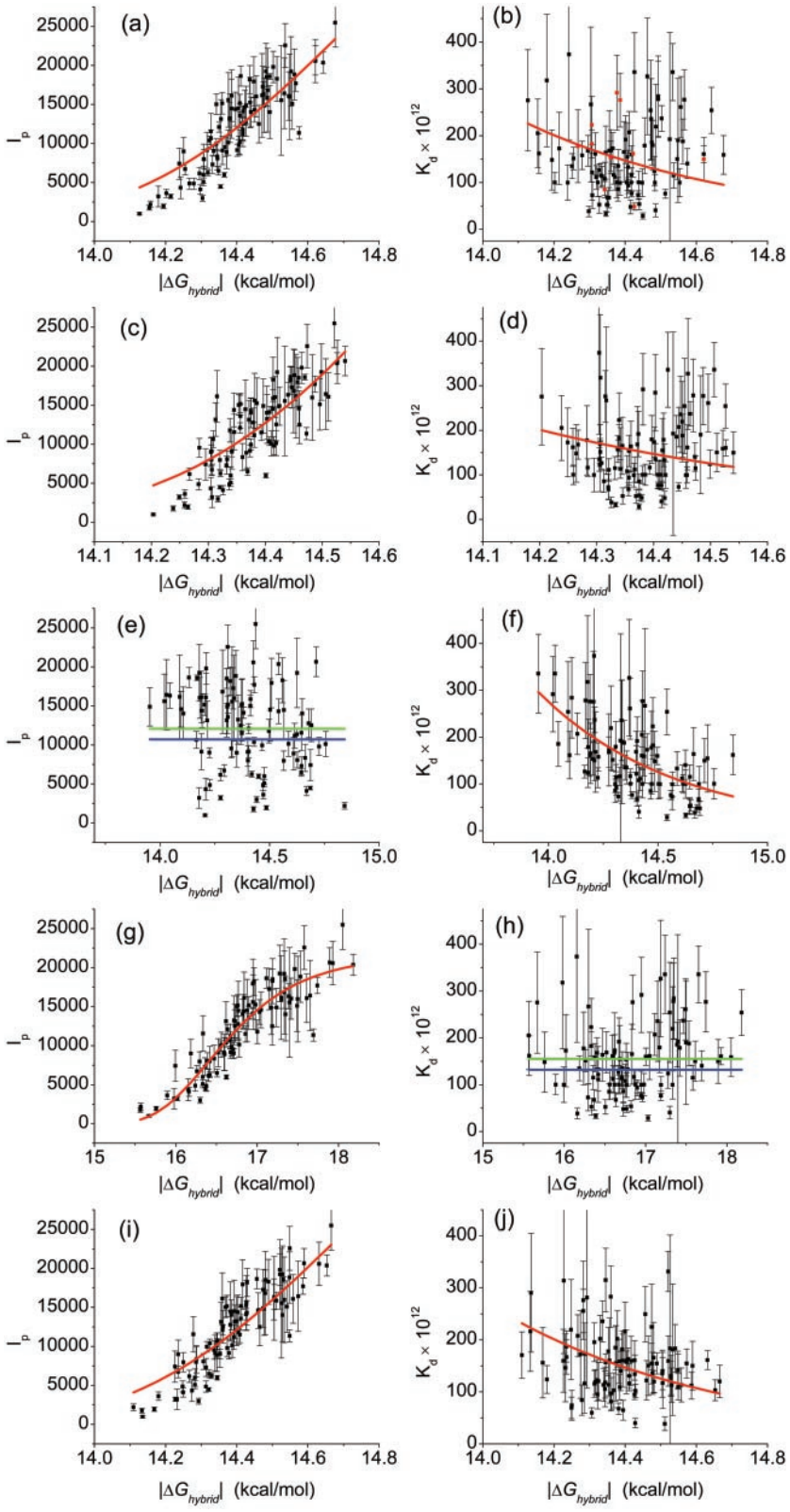

Figure 6. Values of I_p and K_d (obtained by fitting the hybridization intensity as a function of spike-in concentration to Equation 10 for 95 selected PM probes—see text) plotted as functions of $|\Delta G_{hyb}|$, for $|\Delta G_{hyb}|$ calculated using several different models. Method of determining error bars for I_p and K_d is discussed in the text. In all plots the sign of ΔG_{hyb} is negative, and the red lines show values of I_p and K_d predicted by the model. Units of K_d are moles/liter. (**a**) I_p and (**b**) K_d from Model I; (**c**) I_p and (**d**) K_d from Variant 1 of Model I; (**e**) I_p and (**f**) K_d from Variant 2 of Model I. Blue line in (e) shows value of I_p which minimizes *lds* (see text). Green line in (e) shows value of I_p averaged over the 95 plotted values. (**g**) I_p and (**h**) K_d from Variant 3 of Model I. Blue line in (h) shows value of K_d which minimizes *lds*. Green line in (h) shows value of K_d averaged over the 95 plotted values. (**i**) I_p and (**j**) K_d from Variant 4 of Model I.

obtained with this model shows that the experimentally observed variations in fluorescent intensity with probe sequence owe more to the varying degrees of target dissociation during the washing process than to hybridization. This is not unexpected, as the values of I_p for the various probes have a double exponential dependence on ΔG_{hyb} (Equation 11), whereas the values of K_d are related to ΔG_{hyb} by a simple exponential dependence (Equation 4). In Figure 6g and h we plot I_p and K_d as functions of the values of hybridization energy ΔG_{hyb} calculated from the best-fit parameters obtained using Variant 3. In this case, the values of K_d (Figure 6h) are not correlated with ΔG_{hyb}, as they were not utilized in obtaining the values of ΔG_{hyb}. The blue solid line in Figure 6h is the best-fit constant value of K_d (1.32×10^{-10} mol/l) while the green solid line is the average of the 95 K_d values obtained by fitting the data for each probe individually to Equation 10 (1.55×10^{-10} mol/l); the red solid line in Figure 6g shows the dependence of I_p on ΔG_{hyb} predicted by the model.

Although it is tempting to use the comparative results for *lds* and the Pearson correlation coefficient in Table 1 to rank the efficacy of Model I and its variants, our goal here is to construct a model motivated entirely by hybridization, washing and other physical processes. From this viewpoint, there is no justification for incorporating the hybridization energy dependence of I_p in Equation 10 while ignoring the energy dependence of K_d (Equation 4), as in Variant 3. We do not, therefore, use the results in Table 1 to promote Variant 3 as superior to Model I, e.g. but rather to understand how strongly the various physical processes affect measured intensities. For example, the results for Variants 2 and 3 imply that sequence dependent differences in I_p due to washing play a much more significant role in determining intensities than do the sequence dependent differences in K_d.

Variant 4 of Model I. Thus far, we have assumed that the actual concentrations of all of the spiked-in genes are equal to their nominal (i.e. reported) values. In fact, it is difficult to determine the exact concentration of one gene relative to another; the absolute concentration of each gene is quantified by measuring optical density using a spectrophotomer (T. A. Webster, personal communication). On the other hand, all of the spiked-in concentrations of a single gene are determined by serial dilution of a common sample and thus, relative to one another, are well controlled (T. A. Webster, personal communication). With this in mind, we construct a variant—Variant 4—in which the nominal relative concentrations (i.e. 1/4:1/2:1: . . . :1024), of each spiked-in gene are assumed correct but the actual concentrations of two genes, both with the same nominal spike-in concentration, may differ. This model therefore requires 13 adjustable scale factors to reflect the relative differences in actual spike-in concentration for the 14 genes used in the analysis. The best-fit values of these scale factors range from 0.5 to 2.5. Plots of I_p and K_d as a function of the calculated free energy ΔG_{hyb} for Variant 4 are shown in Figure 6i and j. The values of *lds* and the Pearson correlation coefficient are given in Table 1. While this variant describes the data more accurately than any of the others, it is impossible to attach significance to this improvement given the large number of adjustable parameters relative to Model I and the other variants.

Model I/fit_bg. Neither Model I nor its variants are truly predictive, because we have used experimentally determined values for the background; i.e. we have used the average (across all replicate measurements) of the zero concentration spike-in data for each probe as the value for bg_e in Equation 10. In practice, one would expect that those probes which hybridize most strongly with their complementary targets would also hybridize most strongly with background targets that are not fully complementary. Hence, assuming that bg_e results from probe molecules binding to oligonucleotides other than their specific targets and, further, assuming that all possible sequences are represented roughly equally within these non-specific target oligonucleotides, one could reasonably expect those probes with the largest values of $|\Delta G_{hyb}|$ to exhibit the strongest background signals.

In Figure 7a, the experimentally observed background for each of the 95 probes studied is plotted as a function of ΔG_{hyb}, as calculated from Model I. The same data are plotted in Figure 7b, after first being binned into nine energy bins. While the unbinned data do not show a well defined trend as a function of hybridization free energy, the binned data are well described by the sum of a constant and a term which scales exponentially with ΔG_{hyb}. Specifically,

$$bg_e = 10^{15} e^{6\Delta G_{hyb}/RT} + 58.5. \quad \textbf{13}$$

This best fit is shown as a solid line in Figure 7a and b. The deviations of the individual probes from this simple model suggest that, in fact, all sequences are not equally represented within the non-specific (i.e. background) target oligonucleotides and, thus, the observed values of bg_e are strongly affected by cross-hybridization (19) to certain background target molecules which are present at anomalously high concentrations. In other words, the unknown, non-uniform composition of the background solution makes it impossible to predict accurate values of bg_e from a physical model.

The background signal produced by cross-hybridization presents a fundamental limit for any predictive model of absolute gene expression. No model can predict hybridization intensities more accurately than it can predict background intensities due to cross-hybridization. In Figure 5 and Supplementary Figure 1 we plot, as solid red lines, results from Model I/fit_bg, i.e. Equation 10 with values of I_p and K_d derived from Model I and values of bg_e derived from Equation 13. As expected, these lines fit the data significantly worse at low concentrations than do plots of Equation 10 which use the same values for I_p and K_d but use the experimentally observed signal at zero spike-in concentration for bg_e (i.e. the blue solid lines). This is seen quantitatively in the value of *lds*; using Equation 13 to model bg_e results in an *lds* of 0.224–33% higher than the equivalent model (Model I) which incorporates the 'true' background. Model I/fit_bg, which uses three parameters to fit the background data, has three more fitting parameters than Model I, but uses 95 fewer experimental numbers, since Model I directly incorporates the 95 measured background intensities. Of course, in realistic microarray applications, separate background data will not be available and, thus, in the absence of information concerning the presence of particular cross-hybridizing target molecules (19) one must, of necessity, use phenomenological expressions such as Equation 13 to approximate the sequence dependence of bg_e.

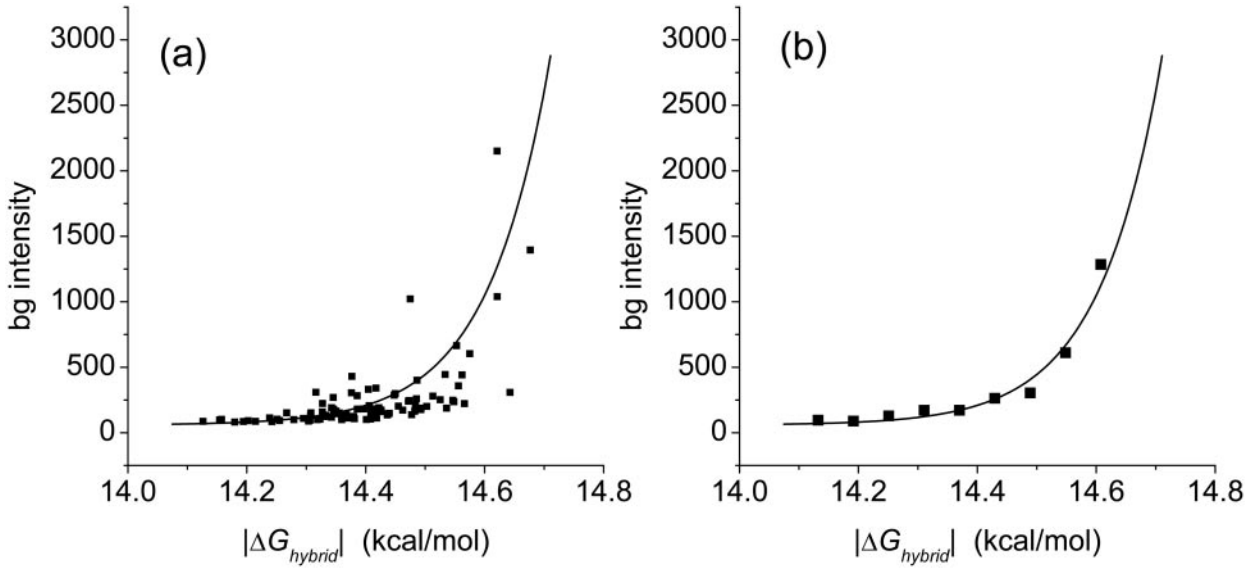

Figure 7. (**a**) Experimentally observed background (i.e. average zero concentration spike-in signal) for selected 95 PM probes plotted as a function of $|\Delta G_{\mathrm{hyb}}|$, where ΔG_{hyb} is determined by simultaneously fitting I_{p} and K_{d} to Equations 4, 7–9 and 11. Values of ΔG_{hyb} for data plotted are negative. (**b**) Same data as (a), binned into nine energy bins. Solid lines in (a) and (b) follow Equation 13.

One method for estimating bg_e which would be less phenomenological and based more directly on the physics of hybridization makes use of the 16 fitted nn energy parameters, $\tilde{\varepsilon}(b_1, b_2)$, to calculate the energy of hybridization for any fragment of a background molecule with a string of bases complementary to some subsequence of a given probe. Assuming that the background is ideally uniform, one could, in principle, average the hybridization probability over all such fragments, producing an estimate for the overall probability that the probe cross-hybridizes to the background and, hence, an estimate for the intensity. An approach along these lines was taken by Zhang *et al*. (12), although these authors introduced a second set of nn-binding energies as adjustable parameters to model the observed non-specific binding. It is difficult to physically justify the need for a separate set of binding energies to describe the binding of non-specifically bound oligonucleotides and, thus, we have chosen to model the non-specific binding contributions in terms of ΔG_{hyb} through Equation 13. The consequences of this decision are not great because, in practice, the non-uniformity of the background would, as we saw above, limit the accuracy of any such estimate (19). In addition, one might consider modeling the MM probe intensities as a means of reducing the uncertainty introduced by cross-hybridization. We have not pursued this approach because it would require another full set of nn energy parameters to describe the mismatch hybridization energy. Further, previous work (19) has shown that non-specific binding to the MM probes is sufficiently different from equivalent binding to PM probes so as to make this method of estimating cross-hybridization ineffective.

CALCULATION OF GENE EXPRESSION LEVELS

Our analysis thus far has dealt with the ability of various physical models to accurately predict probe intensities as a function of target concentration. However, the true utility of any model lies in the accuracy with which it can utilize measured probe intensities to predict gene expression levels. To determine the concentration, c, of a given transcript from the data obtained from a series of replicate measurements, we begin by averaging the PM probe intensities across all of the replicates. We next plot these averaged intensities for all probes corresponding to the transcript in question as a function of calculated probe hybridization free energy ΔG_{hyb}. We then perform a least squares fit of this data to Equation 10, where the values of I_{p}, K_{d} and bg_e are held constant and c, the only adjustable parameter, is constrained to lie between 0 and 2048 pM. The means by which the values of I_{p}, K_{d} and bg_e are determined is discussed below. For a fit of this type to be meaningful, one needs intensity measurements from multiple probes for a given gene. Only 8 of the 14 spiked-in genes have 8 or more probes included among the 95 good probes studied in Modeling of Probe Intensity Data. In the analysis which follows, we limit our consideration to seven of these eight spiked-in genes [we discard one of these eight genes, 407_at, which was identified by Affymetrix as having defective probes (www.affymetrix.com/analysis/download_center2.affx)].

To insure unbiased results, it is important that the probes associated with the gene being assayed are not included in the 'training set' from which the parameters I_{p}, K_{d} and bg_e used to assay that gene are derived. Thus, for each of the seven genes which we assay, we calculate a unique set of values for these parameters. That is, for each of the seven genes, we perform a least squares analysis to calculate best-fit values of $\bar{\boldsymbol{\epsilon}}(b_1, b_2)$, I_{probe}, Γ and a using Model I and only those probes (from the set of 95 good probes) which are not associated with that particular gene. We then use that particular set of parameters to calculate values of I_{p} and K_{d} for each probe from Equations 11 and 4, respectively. In addition, for each gene, we calculate a function (of the form of Equation 13) for bg_e derived using only zero concentration spike-in data for

probes not associated with that gene. The seven sets of values for the parameters $\tilde{\varepsilon}(b_1, b_2)$ are plotted in Figure 8. Ideally, all of these sets should be identical. While we observe some variation from set to set, it is clear from this figure that all of the sets do indeed follow a common overall trend. In carrying out this procedure, we have effectively constructed seven sets of parameters for Model I/fit_bg from seven different 'training sets.' We do not follow a similar procedure for any of the variants of Model I because Variants 1–3 do not represent physical models, while the large number of additional parameters included in Variant 4 prevent a meaningful comparison to other models. That is, Model I/fit_bg represents our basic physical model and, as such, it is the only model whose predictive power we wish to compare with Affymetrix MAS v5 software.

The above procedure for calculating target concentrations from probe intensities is illustrated in Figure 9 with data taken for gene 36085_at at a nominal spike-in concentration of 1024 pM. The observed intensities of nine PM probes for this transcript are plotted as a function of the calculated ΔG_{hyb}. The black line is a best-fit of this data to Equation 10. The best-fit value of c is 759 pM. Similar fits for all of the seven genes analyzed in this fashion, carried out at all concentrations between 0.25 and 1024 pM, are shown in Supplementary Figure 2.

A comparison between nominal spike-in concentration, concentration as determined using the above procedure, and concentration determined using MAS v5 is shown in Figure 10 for all seven genes and all concentrations between 0.25 and 1024 pM. The results from the MAS v5 analysis have been normalized by a multiplicative factor chosen so as to minimize the sum of the squares of the differences between the calculated and nominal concentrations. In general, we find that our method yields results closer to the nominal concentrations than does MAS v5. Specifically, the sum of the squares of the differences between the logarithms of the calculated and nominal concentrations, normalized by the number of genes and concentrations calculated, for our algorithm is 1.09 while for MAS v5 it is 2.1. Note that in calculating these values, it was necessary to exclude those data points for which either our model or MAS v5 predicted a concentration of zero.

To improve our analysis, it is desirable to identify those probe intensities which are either anomalously high or low (9). Such points can be identified by their large distance from the best-fit curve of the model just described and illustrated in Figure 9. In particular, we calculate the mean and the standard deviation of the distances from each of the probe data points for a given transcript to the best-fit curve. Discarding all data points whose distance from the best-fit exceeds the mean distance by at least 1 SD we then refit the remaining data. These refit curves are shown as red lines in Figure 9 and Supplementary Figure 2. The data points which were excluded from these fits are shown in red. When the data in Figure 9 are refit without these points, the best-fit value for c is 881 pM, closer to the known spike-in value of 1024 pM. However, when we calculate the normalized sum (over all seven genes considered) of the squares of the differences between the logarithms of the nominal concentrations and the concentrations calculated in this manner, the result is 1.25—not an improvement from the value, 1.09, resultant from fitting all of the probes. Note that the probes which are identified as outliers for

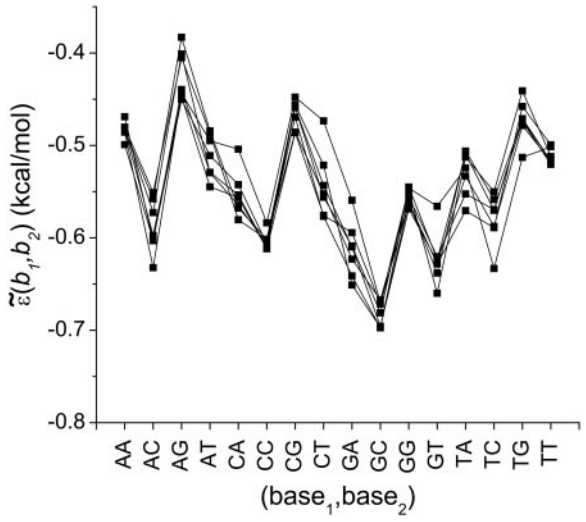

Figure 8. Seven sets of best-fit values of the energies $\tilde{\varepsilon}(b_1, b_2)$ for $b_{1,2}$ = A,C,G,T obtained by simultaneously fitting values of I_{p} and K_{d} to Equations 4, 7–9 and 11. These sets were derived using the same data and method as in Figure 4, except that each set was derived after excluding data from one of the following genes: 37777_at, 36311_at, 1024_at, 36202_at, 36085_at, 40322_at and 1708_at. The similarity between sets illustrates the extent to which the energy values are independent of the datasets used to derive them.

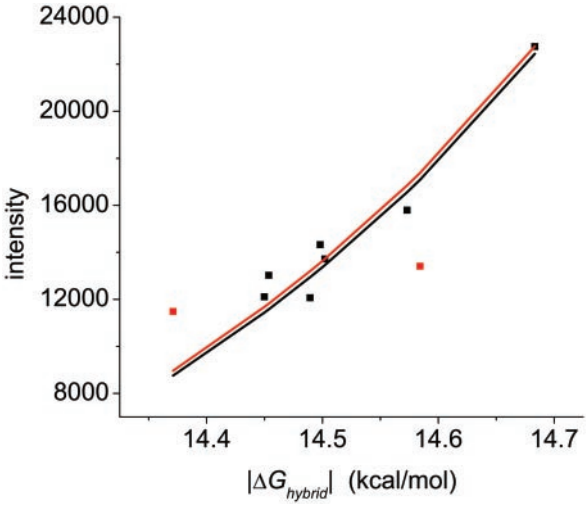

Figure 9. Observed hybridization intensity as a function of calculated $|\Delta G_{\mathrm{hyb}}|$ for those PM probes of gene 36085_at which are included in the 95 selected probes (see text). Each data point is the average of all replicate measurements taken for a given probe at target spike-in concentration 1024 pM. The black line is a best-fit of the data to Equation 10, with concentration c the only adjustable parameter; the best-fit c is 759 pM. The red line is a best-fit to only the black data points, the red ones having been identified as statistical outliers (see text); the best-fit c in his case is 881 pM, closer to the known spike-in value of 1024 pM.

measurements at one concentration tend consistently to be outliers at other concentrations. For example, note in Supplementary Figure 2 that the same two data points are identified as outliers for gene 36085_at data at all spike-in concentrations between 128 and 1024 pM. This suggests that the deviations of individual probe intensities from fits to our model are not simply the result of measurement errors and, therefore, that the model could potentially be improved by incorporating additional physical features.

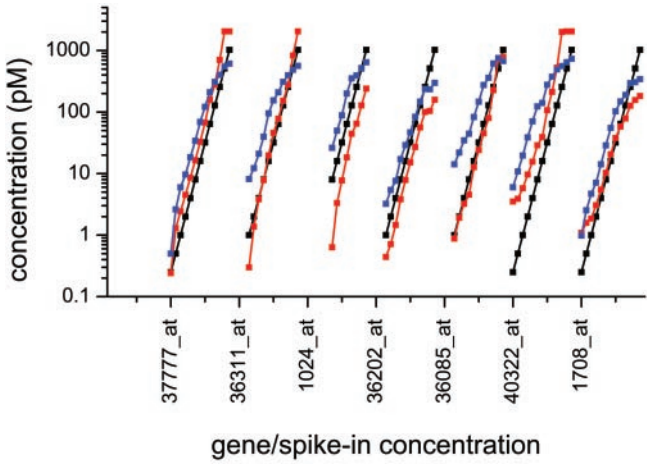

Figure 10. Comparison of best-fit values of concentration of spike-in probes as determined using Model I/fit_bg, as discussed in text (red squares), and MAS v5 (blue squares). Black squares are at nominal spike-in concentrations; deviations from the nominal concentrations appear as deviations from these points. The abscissa indicates the gene fitted as well as the spike-in concentration; for each gene, the nominal spike-in concentration for each data point is twice that of the preceding point. The lowest concentrations shown for genes 37777_at, 36311_at, 1024_at, 36202_at, 36085_at, 40322_at and 1708_at are 0.25, 1, 8, 1, 1, 0.25 and 0.25 pM, respectively. Lower spike-in concentrations are not shown because either our model or MAS v5 predicted a value of zero concentration. Note that the results from the MAS v5 analysis have been normalized by a multiplicative factor chosen so as to minimize the sum of the squares of the differences between the calculated and nominal concentrations.

TARGET FOLDING

One factor not included in our analysis is the effect of target folding in solution (40). Typically, hybridization of the target and probe is a thermally activated process. Helix initiation requires the formation of a 'nucleus' of several base pairs (5 ± 1 bp for oligomers containing only AU pairs and 2 ± 1 bp for oligomers containing at least two CG pairs), following which complete hybridization becomes energetically favorable (28). The secondary structure of most targets in solution results in many of the bases of the target being bound to other bases on the target. Before a target can hybridize to a probe, we hypothesize that there must be at least four bases on the target (in the region complementary to the probe) which are not bound to other target bases.

Prior to the hybridization process, each transcript in a sample is digested such that the typical length of a target is 50 bases (3). We make the approximation that 26 different sequenced target molecules are available to bind to each probe—these targets each being 50-mer regions of the transcript, each beginning at a different location relative to the probe. The assumed target molecules have starting points such that the region complementary to the 25mer probe ranges from the extreme 3′ end of the 50mer target to the extreme 5′ end of the target. We assume that all 26 targets which complement a given probe are present in equal concentration.

For each of these 26 targets associated with a given probe, we use the program 'm-fold' to calculate the most stable secondary structure (41); we then ask whether this structure includes four contiguous unpaired nucleotides in the 25mer region of the target complementary to the probe. We then define the 'target concentration scale factor,' S_F, as the fraction of the 50mer targets which possess such a region of contiguous unbound bases. The product of S_F and the actual concentration of the target represents the effective concentration, c_{eff}, of target molecules present in the solution which are readily able to hybridize to bound probe molecules. In cases where a given 50mer target has multiple stable secondary structures (within 5% of the folding energy of the most stable configuration), we assume all of these configurations to be present in equal fractions and weight the contribution of that 50mer to S_F accordingly. Given these definitions, we find that 86 of the 95 probes used in our analysis have $S_F = 1$ (i.e. all of the stable secondary structures for all 26 of the 50mer target molecules associated with that probe molecule have regions of at least four unbound bases complementary to the probe). Of the remaining nine probes, five have $S_F > 0.9$, while the remaining four have $S_F > 0.74$. Thus, using four contiguous unpaired nucleotides as the criterion for availability of a folded target, we conclude that folding does not significantly limit the availability of targets for hybridization with corresponding probes.

If we instead assume that the secondary structure of the target molecules must have five unbound nucleotides in the region complementary to the appropriate probe molecule, we find that 10 of the 95 probes used in our analysis have $S_F < 0.5$. In Figure 6b the values of K_d obtained for these probes are plotted in red. If the values of c_{eff} (as calculated by requiring five contiguous unbound nucleotides on each target molecule) did indeed reflect the concentration of target molecules available to bind to a given probe, then one would expect intensity as a function of concentration to follow Equation 10, with K_d replaced by K_d/S_F. If this were the case, one would expect the best-fit values of K_d for the 10 probes with $S_F < 0.5$ to all be higher than the corresponding predicted values by factors of at least 2. As can be seen from the positions of the red points in Figure 6b, the best-fit values of K_d for these 10 probes are distributed both above and below the solid line corresponding to the predicted values for K_d. This suggests that the requirement that a target molecule have five contiguous unbound bases which can bind to the probe molecule is too stringent. Assuming, then, that four contiguous bases are sufficient, we conclude that the evidence for folding playing a significant role in determining hybridization intensities is weak. Of course, a more complete analysis would include a detailed picture of how partially folded targets hybridize with their complementary probes. However, given that such a detailed picture is not readily available, it is difficult to formulate a more rigorous assessment of the role of target folding on hybridization intensity.

DISCUSSION AND CONCLUSIONS

In the course of modeling probe intensities, we have found best-fit values for the energy terms $\tilde{\varepsilon}(b_1, b_2)$ and Γ. The average of the 16 energy terms $\tilde{\varepsilon}(b_1, b_2)$ in Figure 4 is -0.55 kcal/mol, whereas, based on measurements of DNA/RNA duplex dissociation in solution, one expects that at 318 K [the hybridization temperature in the Affymetrix protocol (23)] and a salt concentration of 1 M the average such energy term would be -1.46 kcal/mol (in solution under these conditions, $\Delta_{\gamma hybrid} \approx 0$) (31). This indicates that the energy of hybridization is significantly smaller for substrate bound duplexes than it is for duplexes in solution. If we assume that this difference results

largely from sequence independent destabilizing factors, then it follows that $\Delta_{\gamma\text{hybrid}}/24 \approx 0.91$ kcal/mol, i.e. $\Delta_{\gamma\text{hybrid}} \approx 22$ kcal/mol.

One factor which is expected to destabilize hybridization confined to a substrate is the increase in the electrostatic energy which results from localization of the charged oligonucleotides at the substrate surface (32–36). In the limit in which salt cations are the principle source of charge screening (as is the case for a 1 M salt concentration such as that used in the hybridization process) this term may be approximated by

$$\frac{\Delta\gamma_{\text{el, hybrid}}}{RT} = 8\pi N\sigma l_{\text{B}}\frac{r_{\text{D}}^2}{H}, \qquad \textbf{14}$$

where σ is the number charge density per unit area, H is the average height of the probe molecules, N is the average charge number of the probe molecules, $l_{\text{B}} = e^2/\varepsilon k_{\text{B}}T$ is the Bjerrum length ($\approx$7 Å for water at room temperature) and the Debye length $r_{\text{D}} = (8\pi l_{\text{B}}\varphi_{\text{s}})^{-1/2}$ is $\sim$3 Å for a 1 M salt solution (24,32). Assuming an areal probe density of 27 pM/cm^2 and an average probe length of nine bases (24,37,38) (i.e. assuming a 10% termination probability at each base along the probes during probe fabrication), one finds from Equation 14 that, for a fully hybridized probe, $\Delta\gamma_{\text{el,hybrid}}$, which we define as the electrostatic contribution to $\Delta\gamma_{\text{hybrid}}$, is equal to 8.6 kcal/mol, <40% of the experimentally observed value. This suggests that additional factors may be significantly weakening the binding between target and probe in the bound probe microarray geometry. Denoting the sum of all these additional factors by $\Delta\gamma_{a,\text{hybrid}}$, we conclude that $\Delta\gamma_{a,\text{hybrid}} \approx 13.4$ kcal/mol. In addition, we note that the SD of the 16 nn stacking energies $\varepsilon(b_1,b_2)$ measured for bound duplexes in solution is 0.74 kcal/mol (31), whereas the corresponding result for the $\tilde{\varepsilon}(b_1,b_2)$ plotted in Figure 4 is 0.07 kcal/mol, suggesting that the effect of differences in nn pairs along hybridized oligonucleotides is reduced in the anchored probe geometry of microarrays relative to equivalent oligonucleotides in solution.

We note that $\Delta\gamma_{\text{el,hybrid}}$ depends on hybridization fraction (through the resultant change in σ) and, thus, correctly incorporating it into Equations 4 and 7 would result in Equation 10 no longer describing a simple Langmuir isotherm. Rather, Equation 10 would become a transcendental equation, wherein the intensity cannot be expressed as a function of parameters not including intensity itself (32). To avoid this difficulty, we assume an average surface charge which we use to estimate $\Delta\gamma_{\text{el,hybrid}}$ independent of the fraction of bound probe molecules. This approximation is justified by the observation that even in the limit of complete hybridization [itself an overestimate, as probes less than six bases long would probably not be hybridized at 318 K (24)], $\Delta\gamma_{\text{el,hybrid}}$ is <40% of the total value of $\Delta\gamma_{\text{hybrid}}$.

During the stringent wash (for which the salt concentration is 0.1 M and r_{D} is $\sim$10 Å) the nucleotides themselves provide the principle source of screening and $\Delta\gamma_{\text{el,wash}}$ is approximated as follows:

$$\frac{\Delta\gamma_{\text{el, wash}}}{RT} = N\left[\ln\left(\frac{8\pi\sigma_0 l_{\text{B}} r_{\text{D}}^2}{H}\right) + 1\right] + N\ln(1+x), \qquad \textbf{15}$$

where x is the fraction of probes which are hybridized, σ_0 is the number of charges per unit area when x is zero (32), and $\Delta\gamma_{el,\text{wash}}$ and $\Delta\gamma_{a,\text{wash}}$ are the electrostatic and additional contributions to the total $\Delta\gamma_{\text{wash}}$, respectively: $\Delta\gamma_{\text{wash}} = \Delta\gamma_{\text{el,wash}} + \Delta\gamma_{a,\text{wash}}$. From Equation 15 it follows that $\Delta\gamma_{\text{el,wash}}$ ranges from 17.9 to 21.8 kcal/mol as x varies from zero to unity. Considering the limit of complete probe hybridization, the difference between $\Delta\gamma_{\text{el,wash}}$ and $\Delta\gamma_{\text{el,hybrid}}$ is 13.2 kcal/mol, significantly greater than our best-fit value of 1.95 kcal/mol for Γ, which is defined as the difference in free energies of duplex formation during the hybridization and stringent wash stages of sample processing, i.e. $\Gamma = \Delta\gamma_{\text{wash}} - \Delta\gamma_{\text{hybrid}}$. Since $\Delta\gamma_{\text{hybrid}} \approx 22$ kcal/mol, this yields $\Delta\gamma_{\text{wash}} \approx 24$ kcal/mol, which, given $\Delta\gamma_{\text{el,wash}} \approx 21.8$ kcal/mol, implies $\Delta\gamma_{a,\text{wash}} \approx 2.2$ kcal/mol. Thus, under stringent wash conditions, $\Delta\gamma_{\text{wash}}$, which quantifies the reduction in stability of hybridized probes on a DNA microarray relative to hybridized probes in solution, can be accounted for almost entirely by the electrostatic contribution, $\Delta\gamma_{\text{el,wash}}$. The electrostatic contribution, $\Delta\gamma_{\text{el,hybrid}}$, to the reduction in stability during the hybridization process itself, i.e. to $\Delta\gamma_{\text{hybrid}}$, is more modest.

As noted earlier, the incorporation of additional physical processes could result in more accurate predictions of probe intensity as a function of target concentration. One factor which we have not addressed here is the distribution of the lengths of the probe molecules at a given probe site. The presence of such a distribution on Affymetrix microarrays has been reported and the effects of this distribution on target-probe duplex melting temperatures (24), as well as on the reliability of single mismatch detection (42) has been reported. For the experiments analyzed here, however, a simple estimate shows that the hybridization is dominated by the contributions from the full-length 25mers. This dominance is a consequence of the longest probes having the highest affinity for binding to their complementary targets, and occurring more abundantly than all but the shortest truncated probes, whose binding affinity is very low.

To see this, recall from Equation 3 that the contribution to the intensity due to the hybridization of a probe n bases in length is proportional to $n_{\text{probe}}^{(n)} c/(K_{\text{d}}^{(n)} + c)$, where $n_{\text{probe}}^{(n)}$ and $K_{\text{d}}^{(n)}$ are, respectively, the number of probes of length n, and the equilibrium constant for those probes. The 10% truncation probability (24) as each base is added to the probe during fabrication implies that $n_{\text{probe}}^{(n)} = f_n n_{\text{probe}}$, where $f_n \equiv 0.9^{n-1}/10$ is the fraction of probes that are truncated at length n (for $1 \le n \le 24$), and n_{probe} is of course the total number of probes on a spot. That leaves $f_{25} = 1 - \sum_{m=1}^{24} f_m$, i.e. $f_{25} \approx 0.08$, as the fraction of probes with the full 25-base length.

As for $K_{\text{d}}^{(n)}$: Since $K_{\text{d}}^{(n)} = \exp(\Delta G_n/RT)$, where ΔG_n is the hybridization energy of an n-base probe, and since ΔG_n can be approximated by $(n-1)\Delta g$, with Δg a typical stacking energy for nearest-neighbor base pairs, then $K_{\text{d}}^{(n)} \approx b^{n-1}$, with $b \equiv \exp(\Delta g/RT)$. For the experiments studied here, the quantity b can be estimated from the average of our fitted values of K_{d} (which equals $K_{\text{d}}^{(25)}$) through the relation $K_{\text{d}} \approx b^{24}$. Since the average value of K_{d} is roughly 150×10^{-12} mol/l (Figure 6), one finds that $b \approx 0.39$, or $b^{-1} \approx 2.57$.

In the limit $c \ll K_{\text{d}}^{(n)}$, which holds for all but the few longest probes and highest concentrations in the experiments studied here, $c/(K_{\text{d}}^{(n)} + c) \approx c(K_{\text{d}}^{(n)})^{-1}$, whereupon the total contribution to the intensity from all hybridized probes with lengths n from 1 to 24 inclusive is proportional

to $X_{1-24} \approx \sum_{n=1}^{24} (0.9b^{-1})^{n-1}/10$ or, $\sim(0.9\mathrm{b}^{-1})^{24}/13.1$. The contribution, X_{25}, from hybridized probes of full length 25 is $X_{25} \approx 0.08b^{-24}$, whereupon the ratio is $X_{25}/X_{1-24} \approx 13.1$. Thus the full-length probes account for $\sim$93% of the observed intensity. For the highest concentration targets, where $c = 1024 \times 10^{-12}$ mol/l and the approximation $c \ll K_{\mathrm{d}}^{(n)}$ does not hold for the largest values of n, this percentage drops to just below 80%, so the full length probes still account for most of the observed signal. Thus, while a systematic treatment of the truncated probes is obviously desirable, such a treatment seems unlikely to produce significant changes in our results. This conclusion is consistent with the results of our earlier work on fitting the Affymetrix data, Ref. (9), Supplementary Data.

In conclusion, we have demonstrated that it is possible to model gene expression data using a model-based entirely on the physical processes underlying hybridization and washing in DNA microarrays, with the resulting best-fit parameters in good agreement with physically reasonable values. We find that dissociation of probe and target during the washing phase of microarray processing provides a reasonable explanation for the large observed variations in intensity among the probes for a given gene and target concentration. Finally, we have shown that our physical model can be used to quantify gene expression levels fairly accurately, and that fluctuations in the background intensity due to cross-hybridization limit the accuracy attainable. We note that by varying the time and stringency of the wash, it should be possible to experimentally verify the role which washing plays in determining probe intensity. It is hoped that experimental work along these lines will be forthcoming.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by IBM.

Conflict of interest statement. None declared.

REFERENCES

1. Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genet*., **21**, 33–37.
2. Lipshutz,R.J., Fodor,S.P.A., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nature Genet*., **21**, 20–24.
3. Lockhart,D.J., Dong,H.L., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C.W., Kobayashi,M., Horton,H. *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol*., **14**, 1675–1680.
4. Statistical Algorithms Reference Guide. Affymetrix Technical Note.
5. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, e15.
6. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
7. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
8. 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**, 1962–1968.
9. Held,G.A., Grinstein,G. and Tu,Y. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl Acad. Sci. USA*, **100**, 7575–7580.
10. Levicky,R. and Horgan,A. (2005) Physicochemical perspectives on DNA microarray and biosensor technologies. *Trends Biotechnol*., **23**, 143–149.
11. Mei,R., Hubbell,E., Bekiranov,S., Mittmann,M., Christians,F.C., Shen,M.M., Lu,G., Fang,J., Liu,W.M., Ryder,T. *et al*. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **100**, 11237–11242.
12. Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol*., **21**, 818–821.
13. Binder,H., Kirsten,T., Loeffler,M. and Stadller,P.F. (2004) Sensitivity of microarray oligonucleotide probes: variability and effect of base composition. *J. Phys. Chem. B*, **108**, 18003–18014.
14. Wu,Z.J. and Irizarry,R.A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol*., **22**, 656–658.
15. Zhang,L., Wu,C.L., Carta,R., Baggerly,K. and Coombes,K.R. (2004) Preprocessing of oligonucleotide array data—response. *Nat. Biotechnol*., **22**, 658–658.
16. Naef,F., Hacker,C.R., Patil,N. and Magnasco,M. (2002) Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol*., **3**, RESEARCH0018.
17. Tu,Y., Stolovitzky,G. and Klein,U. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl Acad. Sci. USA*, **99**, 14031–14036.
18. Dai,H.Y., 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.
19. Wu,C.L., Carta,R. and Zhang,L. (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res*., **33**, e84.
20. Wu,Z.J., Irizarry,R.A., Gentleman,R., Martinez-Murillo,F. and Spencer,F. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc*., **99**, 909–917.
21. Binder,H. and Preibisch,S. (2005) Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophys. J*., **89**, 337–352.
22. (2001) New Statistical Algorithms for Monitoring Gene Expression on GeneChip Probe Arrays. *Affymetrix Technical Note*.
23. (2004) GeneChip Expression Analysis Technical Manual. *Affymetrix Technical Manual*.
24. Forman,J.E., Walton,I.D., Stern,D., Rava,R.P. and Trulson,M.O. (1998) Thermodynamics of duplex formation and mismatch diserimination on photolithographically synthesized oligonucleotide arrays. *ACS Symp. Ser*., **682**, 206–228.
25. Atkins,P. (1994) *Physical Chemistry, 5th edn*. W. H. Freeman, NY.
26. Erickson,D., Li,D.Q. and Krull,U.J. (2003) Modeling of DNA hybridization kinetics for spatially resolved biochips. *Anal. Biochem*., **317**, 186–200.
27. Gadgil,C., Yeckel,A., Derby,J.J. and Hu,W.S. (2004) A diffusion-reaction model for DNA microarray assays. *J. Biotechnol*., **114**, 31–45.
28. Turner,D.H. (2000) Conformational changes. In Bloomfield,V.A., Crothers,D.M. and Tinoco,J.I. (eds), *Nucleic Acids: Structures, Properties, and Functions*. University Science Books, Sausalito, pp. 259–334.
29. Sekar,M.M.A., Bloch,W. and St John,P.M. (2005) Comparative study of sequence-dependent hybridization kinetics in solution and on microspheres. *Nucleic Acids Res*., **33**, 366–375.
30. SantaLucia,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
31. Sugimoto,N., Nakano,S., Katoh,M., Matsumura,A., Nakamuta,H., Ohmichi,T., Yoneyama,M. and Sasaki,M. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, **34**, 11211–11216.
32. Halperin,A., Buhot,A. and Zhulina,E.B. (2004) Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophy. J*., **86**, 718–730.

33. Peterson,A.W., Heaton,R.J. and Georgiadis,R.M. (2001) The effect of surface probe density on DNA hybridization. *Nucleic Acids Res*., **29**, 5163–5168.
34. 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**, 14601–14607.
35. Vainrub,A. and Pettitt,B.M. (2003) Surface electrostatic effects in oligonucleotide microarrays: control and optimization of binding thermodynamics. *Biopolymers*, **68**, 265–270.
36. Vainrub,A. and Pettitt,B.M. (2000) Thermodynamics of association to a molecule immobilized in an electric double layer. *Chem. Phys. Lett*., **323**, 160–166.
37. Fodor,S.P.A., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.
38. Pease,A.C., Solas,D., Sullivan,E.J., Cronin,M.T., Holmes,C.P. and Fodor,S.P.A. (1994) Light-generated oligonucleotide arrays for rapid DNA-sequence analysis. *Proc. Natl Acad. Sci. USA*, **91**, 5022–5026.
39. Naef,F. and Magnasco,M.O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys*., **68**, 011906.
40. Mir,K.U. and Southern,E.M. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol*., **17**, 788–792.
41. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.
42. Jobs,M., Fredriksson,S., Brookes,A.J. and Landegren,U. (2002) Effect of oligonucleotide truncation on single-nucleotide distinction by solid-phase hybridization. *Anal. Chem*., **74**, 199–202.