Published online 18 October 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 20* ***6585–6591***
doi:10.1093/nar/gkn740

A competitive hybridization model predicts probe signal intensity on high density DNA microarrays

Shuzhao Li*, Alex Pozhitkov and Marius Brouwer

Gulf Coast Research Lab, Department of Coastal Sciences, University of Southern Mississippi, Ocean Springs, MS 39564, USA

Received August 20, 2008; Revised October 1, 2008; Accepted October 2, 2008

ABSTRACT

A central, unresolved problem of DNA microarray technology is the interpretation of different signal intensities from multiple probes targeting the same transcript. We propose a competitive hybridization model for DNA microarray hybridization. Our model uses a probe-specific dissociation constant that is computed with current nearest neighbor model and existing parameters, and only four global parameters that are fitted to Affymetrix Latin Square data. This model can successfully predict signal intensities of individual probes, therefore makes it possible to quantify the absolute concentration of targets. Our results offer critical insights into the design and data interpretation of DNA microarrays.

INTRODUCTION

Current DNA microarray technology utilizes multiple oligonucleotide probes to detect the concentration of target molecules. These probes, even though interrogating the same target, often yield very different signal intensities. Without understanding the physicochemistry underlying this problem, the quantification of absolute gene abundance is unattainable and inter-probe comparison is unjustified, leaving DNA microarray technology severely compromised.

A number of physical models have been proposed to address this problem, mostly in the form of Langmuir derivatives (1–10). The Langmuir model is a generic mathematical form that also fits the description of first-order chemical reactions, which is frequently used for probe/target binding on DNA microarrays:

$$\theta = \frac{T}{T+K}, \qquad \mathbf{1}$$

where θ is the fraction of occupied probes, T free target concentration, K dissociation constant.

According to the Langmuir model, all probes should saturate at the same level, which is clearly not the case in microarray hybridizations. Various modifications were proposed to accommodate this difference in saturation levels. A generic version may be written as

$$\theta = \frac{\chi T}{T+K}, \qquad \mathbf{2}$$

where χ is a probe-specific factor. While a physical meaning of χ is difficult to obtain, some (7,8) tried to explain χ through the washing step in microarray experiments. That is, all probes reach the same saturation level in the end of hybridization, but they lose the bound targets to different extents during the washing step. This 'washing model' suggests a significant loss of signals upon each washing cycle. In experimental observations, the first washing cycle usually removes a considerable amount of partially bound targets, but it is clear that signal intensities do not decrease dramatically after extra washing cycles (11). This contradicts the above 'washing model'. Furthermore, the Langmuir derivatives predict that, in response to increasing target concentrations, probes with higher binding affinities saturate first. In experimental observations, on the contrary, low-affinity probes generally saturate first. Although Langmuir models seem to work well on simple surface hybridizations, no Langmuir derivative has adequately predicted probe signals in 'real' experimental settings, such as those in the Affymetrix Latin Square data with complex backgrounds (12).

The best prediction of probe signals to date was reported by Zhang *et al.* (13). They accounted for both specific binding and nonspecific binding in the form of $\hat{T}/(1+K)$, where $\hat{T}$ is total target concentration, while fitting 83 parameters to the data. Mei *et al.* (14) also sought a linear composition of binding energy, where the single base energy contribution alone used 75 parameters. Over-parameterization has been a concern in all these previous studies and invited criticism on their general applicability (15).

After all, a valid physical model of microarray hybridization will have to explain the probe difference through sequence-specific thermodynamics, as its oligonucleotide sequence is the defining property of a microarray probe. The free energy of polynucleotide hybridization in bulk

*To whom correspondence should be addressed. Tel: +1 228 872 4278; Fax: +1 228 872 4204; Email: shuzhao.li@gmail.com

solution has been successfully described by a nearest neighbor (NN) model (16,17). However, this NN model is widely regarded as not applicable to high-density microarray hybridization, as it was either modified and re-parameterized (5,7,9,13,18), or abandoned (1,14,19).

We will first demonstrate that the NN model is applicable to probes free of secondary structures. With the thermodynamic component calculated from the NN model, we then propose a new competitive hybridization model to describe the kinetics. Our model, using only four global parameters that are fitted to Affymetrix Latin Square data, can successfully predict signal intensities of individual probes, and therefore, achieve the absolute quantification of target concentrations.

METHODS

The Affymetrix Latin Square spike-in data U133A were retrieved from (12). They contained 14×3 hybridizations where spike-in targets were added at various concentrations from 0 pM to 512 pM. Probe information was obtained through (20), where only 30 of the 42 probesets were found. A total of 365 probes matched to target sequences. Among them, 10 probes with very low signal intensities (under 900 at highest target concentration) were removed. In total, 355 probes are included in this study. Background was taken as the signal intensity at zero spike-in concentration, and subtracted from data at other concentrations. No normalization was performed on these data. The probe self-folding energy, ΔG_o, was computed by RNAStructure [version 4.5, function OligoWalk (21)]. Duplexing energy, ΔG_d, was computed by the current NN model with the parameters from Ref. (17). Compiled data and computational scripts used in this study are available upon request.

RESULTS AND DISCUSSION

Thermodynamic predictability in DNA microarrays

Microarray probes and targets may form secondary structures by intramolecular self-folding. These structural effects are not accounted for in the NN model, posing a problem to the thermodynamic calculation. As a first step, we investigated the structural effects through the self-folding energy of probes, ΔG_o. In the Affymetrix Latin Square data (see Methods Section for details), about 45% of the probes can be selected by the criterion $\Delta G_o > -1$. For these probes, a clear correlation appears between log signal intensities (*SI*) at the highest target concentration and the duplexing energy ΔG_d that are computed by the current NN model with existing parameters (Figure 1, $R^2 = 0.58$). If the selection criterion is relaxed to $\Delta G_o > -2.5$, 75% probes are included and the $\log SI - \Delta G_d$ correlation has $R^2 = 0.45$ (data not shown). However, the $\log SI - \Delta G_d$ correlation diminishes at lower target concentrations (data not shown). These observations suggest that the current NN model offers a certain degree of predictability, but they cannot be accommodated by previous, Langmuir-like models. A new kinetic model is needed.

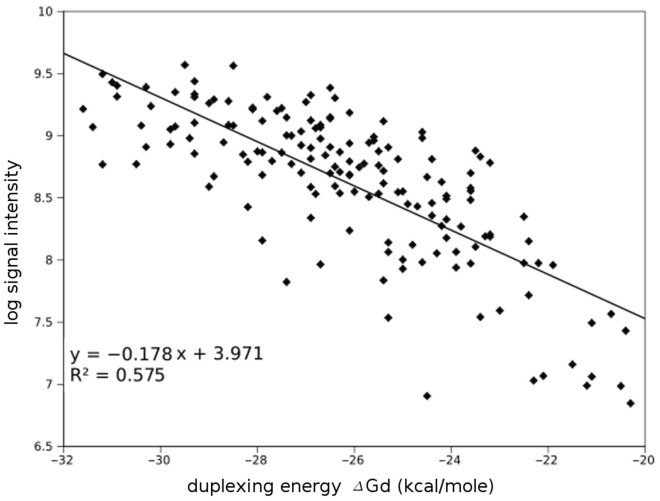

Figure 1. Duplexing energy calculated by NN model correlates with signal intensity at 512 pM, the highest spike-in target concentration, for probes free of secondary structures ($\Delta G_0 > -1$, 160 probes from Affymetrix U133A data). Each dot represents a probe.

A competitive hybridization model

We treat DNA microarray hybridization as two subprocesses, the binding of targets to probes and the dissociation of target/probe duplexes. Assuming that equilibrium is reached at the end of hybridization and the binding rate is the same for all target molecules (see below), the dissociation rate is governed by the duplexing energy between paired target/probe. A kinetic equilibrium between binding and dissociation should be observed.

Two types of targets are explicitly modeled: 'specific targets' (perfect match) with probe-specific dissociation rate k_d, and 'cross-hybridizing targets' with dissociation rate k_n. These cross-hybridizing targets are present in large quantities because partially matching sequences are abundant in a transcriptome. For the moment, we simplify them as a uniform mixture with a probe-nonspecific k_n.

The target/probe duplex formation is commonly believed to start with an initiation step, the base-pairing between a small number of nucleotide bases, and then extend to the rest of complementary regions (22,23). If the initiation step sets the rate limit, the binding rate should be hardly specific to probe sequences. We therefore assume a single binding rate, k_b, for all target molecules. How the specific factors (24,25), including adsorption and electrostatics (26), steric and brush effects (27) and labeling (19,28), come into play is not yet entirely clear. In this study, we postulate that the available area of probe spots is the limiting factor in adsorption, so that the binding is described as

$$\frac{\dot{n}_{in}}{N_A V} = (1 - \alpha - \beta) \cdot p \cdot k_b, \qquad \mathbf{3}$$

where $\dot{n}_{in}$ is the number of target molecules going into the exposed probes over a unit of time, N_A the Avogadro

constant, V the volume of hybridization solution. On the right side, α is the fraction of probes bound to specific targets, β the fraction of probes bound to cross-hybridizing targets. p is the total number of probes in unit of molar concentration (for simplicity, as if they were dissolved in the hybridization solution).

On the other hand, the dissociation is described by

$$\frac{\dot{n}_{out}}{N_A V} = \alpha \cdot p \cdot k_d + \beta \cdot p \cdot k_n, \quad \mathbf{4}$$

where $\dot{n}_{out}$ is the number of target molecules leaving target/probe duplexes over a unit of time; k_d and k_n are dissociation rates for specific targets and cross-hybridizing targets, respectively.

At equilibrium between binding and dissociation,

$$(1-\alpha-\beta)\cdot p \cdot k_b = \alpha \cdot p \cdot k_d + \beta \cdot p \cdot k_n \quad \mathbf{5}$$

Equilibrium is established for both specific and cross-hybridizing targets. The proportions of specific targets and cross-hybridizing targets are determined by their concentrations:

$$\alpha \cdot p \cdot k_d = \frac{\dot{n}_{in}}{N_A V} \cdot \frac{[T]}{[T]+[N]}, \quad \mathbf{6}$$

$$\beta \cdot p \cdot k_n = \frac{\dot{n}_{in}}{N_A V} \cdot \frac{[N]}{[T]+[N]}, \quad \mathbf{7}$$

where $[T]$ is the concentration of free specific targets, $[N]$ the concentration of free cross-hybridizing targets.

Equations (6) and (7) can be combined to express β as:

$$\beta = \frac{k_d[N]}{k_n[T]} \cdot \alpha \quad \mathbf{8}$$

Then, Equations (5) and (8) give the fraction of specific binding

$$\alpha = \frac{1}{1 + k_d\left(\frac{1}{k_b} + \left(\frac{1}{k_n} + \frac{1}{k_b}\right)\frac{[N]}{[T]}\right)} \quad \mathbf{9}$$

Here $[T]$, the concentration of free specific target molecules, is less than nominal spike-in concentration by the amount of probe binding:

$$[T] = \hat{T} - \alpha \cdot p \quad \mathbf{10}$$

with $\hat{T}$ as the nominal spike-in concentration (total amount).

We assume the concentration of cross-hybridizing targets, $[N]$, is large and can be treated as constant in this model. Let

$$\gamma = \left(\frac{1}{k_n} + \frac{1}{k_b}\right)[N], \quad \mathbf{11}$$

then Equation (9) becomes

$$\alpha = \frac{1}{1 + k_d(1/k_b + \gamma/(\hat{T} - \alpha \cdot p))} \quad \mathbf{12}$$

An analytical solution of Equation (12) is

$$\alpha = \frac{1}{p}\left(\Gamma - \sqrt{\Gamma^2 - \frac{p\hat{T}}{1 + k_d/k_b}}\right), \quad \mathbf{13}$$

where

$$\Gamma = \frac{\hat{T}}{2} + \frac{p + \gamma k_d}{2(1 + k_d/k_b)} \quad \mathbf{14}$$

It can be shown that the other analytical solution of Equation (12), which bears a plus sign before the square root, has no valid physical meaning and merits no further discussion.

So α, the fraction of probes bound to specific targets, is described by three global parameters: p, k_b and γ, one probe-specific parameter k_d and one variable $\hat{T}$. k_d can be expressed as:

$$k_d = e^{\frac{\xi \Delta G_d}{RT}}, \quad \mathbf{15}$$

where R is the molar gas constant, T absolute temperature (318 K in the Affymetrix hybridization experiments), ΔG_d the energy computed from NN model, ξ as a scaling factor to account for binding to immobilized probes.

The physical meaning of our model is clear. Both specific binding α and cross-hybridization β compete for the same probe sites. As a result, high affinity probes (small k_d) can achieve a higher fraction of specific binding, while low-affinity probes (large k_d) saturate at a lower fraction. γ serves as a cross-hybridization factor. We made assumptions that are important to real experimental settings: a large quantity of cross-hybridizing targets are present; k_b is uniform for all targets and the adsorption is limited by the available area of probe spots. These assumptions make our model fundamentally different from previous competitive kinetic models (29,30).

Experimentally, signal intensity is what is observed after washing, where most of cross-hybridized targets have been washed off:

$$SI = A \cdot \alpha \cdot p + \tau + \iota, \quad \mathbf{16}$$

where SI is the observed signal intensity, τ the residual intensity from cross-hybridized targets, ι scanner bias, A the detection coefficient of fluorescence. As the unit of signal intensities is arbitrarily digitized, it only comes to a physical meaning through A.

Explanation to the log $SI - \Delta G_d$ correlation

First of all, we shall demonstrate that our model is capable of explaining the $\log SI - \Delta G_d$ correlation at high target concentration in Figure 1.

Equation (12) can be rearranged to a logarithmic form:

$$\log\frac{\alpha}{1-\alpha} = -\log k_d - \log(1/k_b + \gamma/(\hat{T} - \alpha \cdot p)) \quad \mathbf{17}$$

Note that the second item on the right side still contains the probe-dependent variable α. However, at high target

concentration, the bound targets are minor comparing to free targets. This means, $\hat{T} \gg \alpha \cdot p$, and $\hat{T} \approx \hat{T} - \alpha \cdot p$. Hence, Equation (17) at high target concentration is approximated as:

$$\log\frac{\alpha}{1-\alpha} = -\log k_d - \log(1/k_b + \gamma/\hat{T}) \quad \textbf{18}$$

In a long range for $0 < \alpha < 1$, a linear approximation can be drawn between $\log\frac{\alpha}{1-\alpha}$ and $\log\alpha$. As in Figure 2,

$$\log\frac{\alpha}{1-\alpha} = 1.248 \cdot \log\alpha + 0.702 \quad \textbf{19}$$

Combining Equations (15), (18) and (19), we get

$$\log\alpha = -0.801 \cdot \frac{\xi \Delta G_d}{RT} - 0.801 \cdot \log(1/k_b + \gamma/\hat{T}) - 0.563 \quad \textbf{20}$$

At high target concentration, both cross-hybridization and scanner bias can be neglected. Therefore Equation (16) can be simplified to $SI = A \cdot \alpha \cdot p$. We substitute the α in Equation (20) with $SI/(A \cdot p)$:

$$\log SI = -\frac{0.801\xi}{RT}\Delta G_d + C, \quad \textbf{21}$$

where $C = \log(A \cdot p) - 0.801 \cdot \log(1/k_b + \gamma/\hat{T}) - 0.563$, a constant for fixed $\hat{T}$. Thus, $\log SI$ is inversely correlated to ΔG_d. The observed $\log SI - \Delta G_d$ correlation is explained by our competitive hybridization model. At low $\hat{T}$, the premise $\hat{T} \approx \hat{T} - \alpha \cdot p$ is less valid; as a result, $\log SI$ is less correlated to ΔG_d. A similar effect may be created by a very low ΔG_d, where a large fraction of targets is bound to probes and taken out of solution.

A bonus here is the determination of ξ. Since the coefficient for ΔG_d in Equation (21) should equate the slope in Figure 1, we get $\xi = 0.140$.

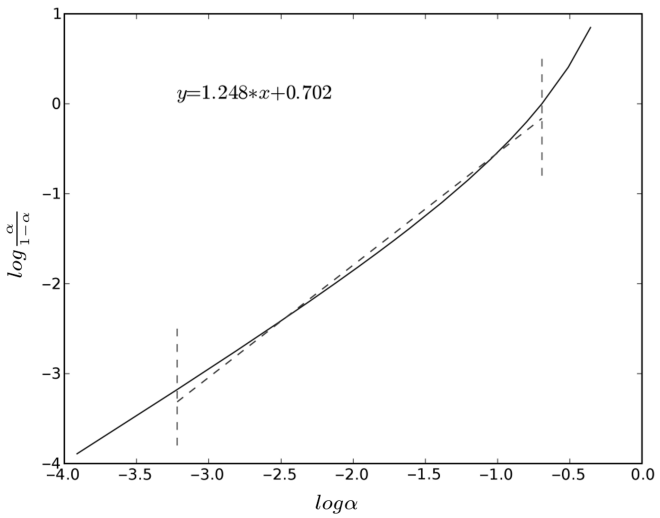

Figure 2. Linear relationship between $\log(\alpha/(1-\alpha))$ and $\log\alpha$.

Procedure of fitting model to Latin Square data

In DNA microarray experiments, signal intensities are measured in place of fluorescent densities of bound targets. However, common photomultiplier tube scanners usually carry a significant nonlinearity for low signal intensities (31). This means, the lower end of these Affymetrix data may deviate from the true fluorescent densities, a problem difficult to correct without knowledge of the specific instrument calibration data. And the signals from targets below 1 pM are hardly distinguishable from backgrounds, therefore, data from spike-in concentration 1 pM and above are used for our modeling.

The model fitting is to match the theoretical calculation of signal intensity, $\widehat{S}$, to the experimentally observed counterpart $\bar{S}$. $\bar{S}$ is defined from Equation (16):

$$\bar{S} = SI - \tau - \iota \quad \textbf{22}$$

Here, the background levels τ are observed values in these Affymetrix data (signal intensities at zero spike-in concentration). With background τ subtracted, the signal intensity should approach zero when the target concentration approaches zero. However, there is usually a deviation from zero that is known as scanner bias, ι, which can be therefore estimated by extrapolating the signal intensities at low target concentrations. For the data in this study, $\iota = -20$ is taken. This value of ι is relatively small and has no significant effect on our model parameters. Though it is useful for stabilizing the small numbers in the fitting process.

With the theoretical value

$$\widehat{S} = A \cdot \alpha \cdot p, \quad \textbf{23}$$

Equation (13) can be written as

$$\widehat{S} = A \cdot \left(\Gamma - \sqrt{\Gamma^2 - \frac{p\hat{T}}{1 + k_d/k_b}}\right), \quad \textbf{24}$$

where Γ is defined in Equation (14). In this equation, $\hat{T}$ is known, $\bar{S}$ the observed value for $\widehat{S}$, and k_d can be calculated from Equation (15). So we only need to fit four global parameters: A, p, k_b and γ.

We use a fitness function of weighted squares [similar to (1)]. For a probe i, the fitting error is calculated as

$$E_i = \sum_t \frac{(\widehat{S}_{i,t} - \bar{S}_{i,t})^2}{\bar{S}_{i,t}}, \quad \textbf{25}$$

where $\bar{S}$ is observed signal intensity, $\widehat{S}$ the calculated value by Equation (24), t one of the nominal target concentrations $\hat{T}$ (1–512 pM). Our model in Equation (24) is fitted to the training data by minimizing the sum of E_i through brute-force searches as heuristic ranges of the four parameters can be obtained based on their physical meanings.

A useful constraint to the fitting is the value of $P_0 = A \cdot p$. This is the signal intensity in Equation (23) when $\alpha = 1$, often referred as the saturation level of hybridization. It is obvious that P_0 should be larger but not infinitely larger than the highest signal intensity

observed in the experiment. When varying P_0 is used, as shown in Figure 3, the overall fitting results are not very sensitive to P_0 beyond a certain value. So we choose $P_0 = 30\,000$ here.

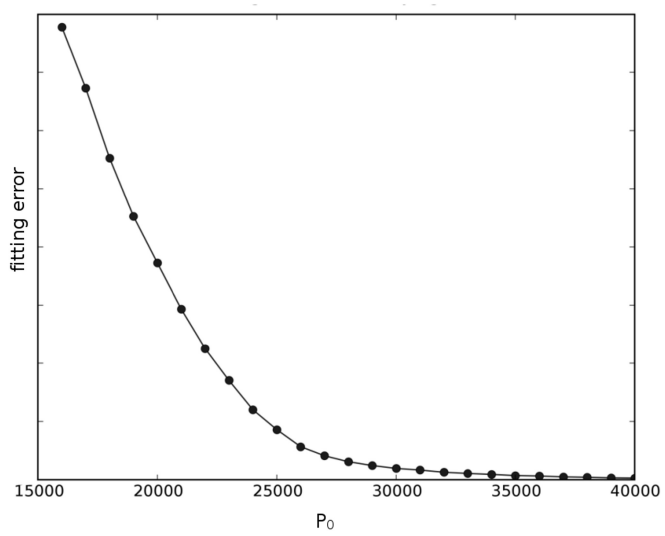

Figure 3. The fitting result is not sensitive to P_0 beyond a certain value. All four parameters, A, p, k_b and γ, are fitted simultaneously. $P_0 = A \cdot p$; the fitting error is computed as in Equation (25).

Probe signal intensities can be successfully modeled

Figure 1 shows that ΔG_d (hence k_d) can be reasonably approximated by the current NN model for the probes free of secondary structures. We use half of these probes to fit our competitive hybridization model, and determine the four global parameters, A, p, k_b and γ. The evaluation is then performed on the rest of probes.

The results indicate that our model captures the probe properties well. Figure 4A shows the modeling of individual probe signals on both the training data and evaluation data. Overall, the prediction on training data has $R^2 = 0.866$ (Figure 4B), and $R^2 = 0.880$ on evaluation data (Figure 4C). If we relax the probe selection criterion to $\Delta G_o > -2.5$, about 75% of total probes are included, with prediction $R^2 = 0.844$ (Figure 4D). The rest 25% of probes, which are presumably under stronger influence of secondary structures, can still be modeled with the same parameters but less accuracy at $R^2 = 0.735$.

In the previous, heavily parameterized models, the best prediction on $log\widehat{S}$ was correlation coefficient $r = 0.85$ in Ref. (14) and $r > 0.9$ in Ref. (13). In comparison, our model of four parameters produces $r = 0.889$ for all probes, and $r = 0.919$ for 75% probes after a preliminary selection by secondary structures (i.e. Figure 4D). In conclusion, our competitive hybridization model can not only predict probe signals successfully, but also opens up paths to future improvements.

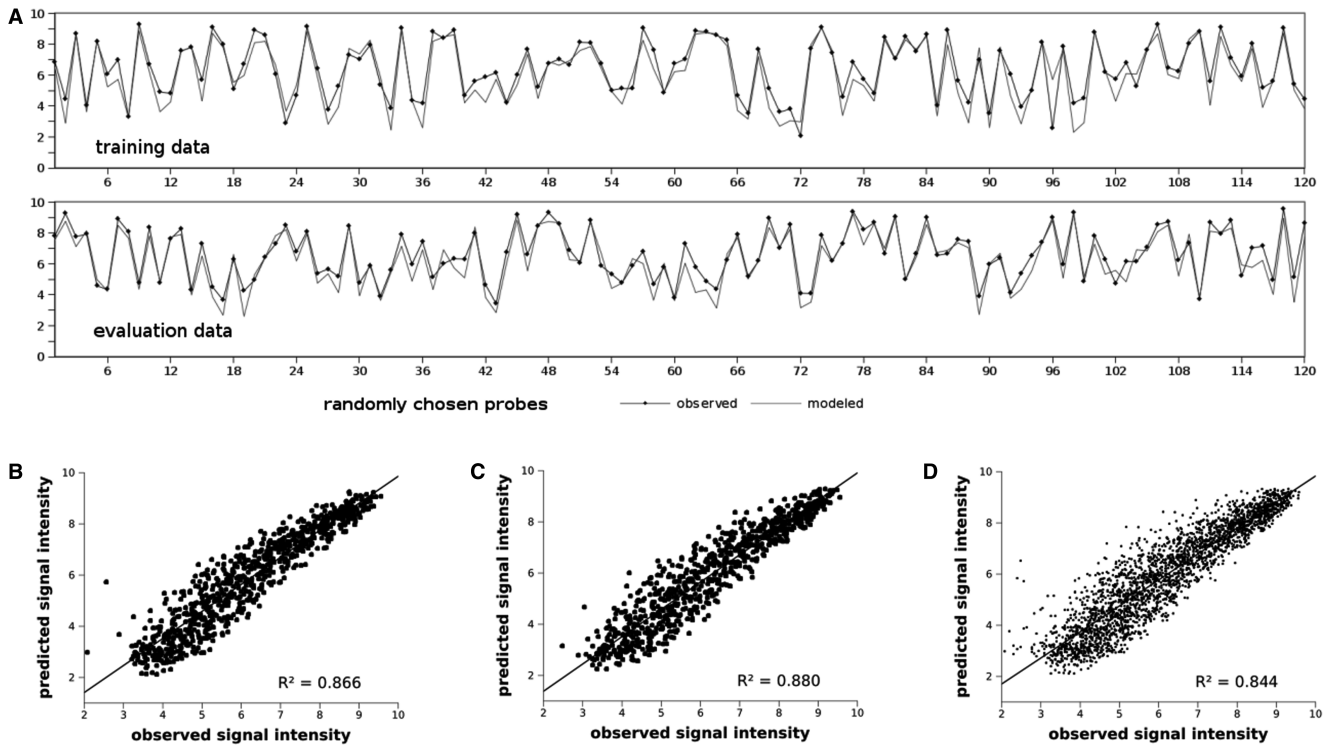

Figure 4. Our model can successfully predict probe signal intensities. (**A**) The prediction on randomly chosen probes at random target concentrations. Top: the training data; Bottom: the evaluation data. (**B**) Scatter plot of all training data. (**C**) Scatter plot of the evaluation data. The training data are consisted of half of the probes from Figure 1, and evaluation data from the other half. (**D**) Extended evaluation on 266 (75% of total) probes that satisfy $\Delta G_o > -2.5$. All signal intensities are in log scale. The parameters in this figure are $A = 33.408$ (pM)$^{-1}$, $p = 898$ pM, $k_b = 1.348E\text{-}3\ s^{-1}$ and $\gamma = 245\,500$ pM s.

Prediction of target concentrations

With the four global parameters, target concentration can be calculated from Equation (12):

$$\hat{T} = \alpha p + \frac{k_d \gamma}{1/\alpha - k_d/k_b - 1} \quad \textbf{26}$$

If we substitute $\alpha = \bar{S}/Ap$,

$$\hat{T} = \frac{\bar{S}}{A} + \frac{k_d \gamma}{Ap/\bar{S} - k_d/k_b - 1} \quad \textbf{27}$$

Since k_d calculation is more accurate for probes free of secondary structures, we focus on 19 out of the 30 probesets (transcripts) in this study that have five or more probes with $\Delta G_o > -1$. For these transcripts, Equation (27) is applied to calculate a target concentration from each probe. And the final concentration of a transcript is taken as the median of the data from its probes (Figure 5A). Figure 5B shows the prediction at gene level for all 19 transcripts. In fact, comparable results can be obtained by using the few probes with $\Delta G_o > -1$ alone. At low concentrations, the predicted values in Figure 5B tend to be higher than the nominal concentrations. We think this is likely to be a reflection of scanner nonlinearity in the low signal range, which can be corrected by an instrument calibration.

Discussion

In DNA microarray experiments, systematic variations stem from sample preparation and instrument operations. They are likely reflected in the global parameters of our model, A, k_b and γ. Therefore, batch variations can be expected in these parameters. The highest signal intensity in the Latin Square data was about 16 000. Comparing with the saturation level $P_0 = 30\,000$, this means about half of those probe sites are bound to specific targets at $\hat{T} = 512$ pM. Thus, the fitted value of $p = 898$ pM (as in Figure 4) seems to be reasonable. Since our model has only four global parameters, they can be easily calibrated if control probes are built into array design. For instance, a set of targets complementary to the control probes can be spiked into the hybridization at various concentrations. Signal intensities of the control probes along with the known target concentrations can then be used to calibrate our model every time a hybridization experiment is performed.

We would like to emphasize that k_d is the only probe-specific factor in our model, and therefore plays a pivotal role in model accuracy. The accuracy of k_d or ΔG_d in this article is limited by the NN model, which is only a coarse approximation and affected by probe/target secondary structures. This can be improved but beyond the scope of this current study.

We assumed a constant cross-hybridization factor γ for all probes, which may not be the case. Further research on γ may improve the accuracy of our model. We did not deal with the background levels in this study, which are not important to signals at high target concentration but will affect signals at low concentrations. Background levels have a clear dependency on ΔG_d, and are well addressed in other studies (32,33).

Conclusion

Our study presents the first model of DNA microarray hybridization that explains probe signal intensities through sequence-based thermodynamic properties without excessive parameter fitting. This fills in the long standing knowledge gap in DNA microarray hybridization. Our model provides a mechanism of absolute quantification,

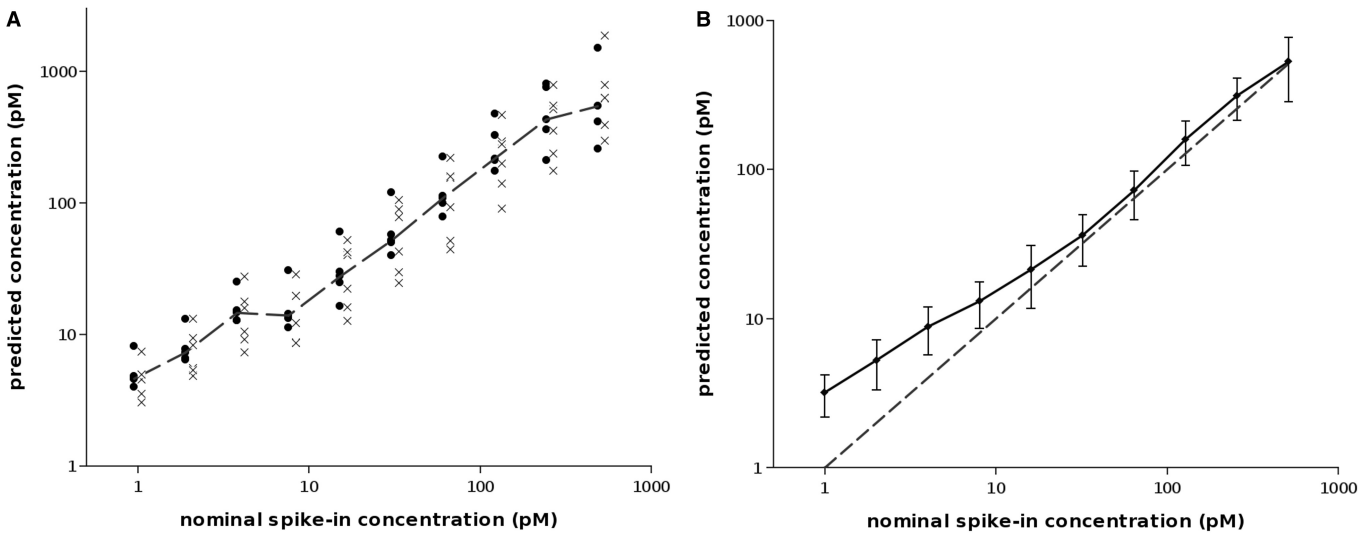

Figure 5. Prediction of transcript concentration. (**A**) Example of the 11 probes for transcript 205267_at. Dots are probes with $\Delta G_o > -1$, other probes in crosses (slightly shifted horizontally for clarity). The transcript concentration (dashed line) is taken as the median value of all probes. (**B**) Prediction of 19 transcripts that have five or more probes with $\Delta G_o > -1$. Correlation coefficient between nominal concentrations and the predictions is $r = 0.89$. Error bars are standard deviations of the 19 transcripts. The predicted values bend away from the ideal line (dashed) at low concentrations probably because of scanner nonlinearity.

and shall improve the quality control and reproducibility of the technology. With only four global parameters, this model can be easily calibrated through control features that are built into microarrays, and adopted in practice. We expect new design and quantification algorithms to take advantage of our results.

FUNDING

National Oceanic and Atmospheric Administration (NA05NOS4261163 and NA06NOS42600117).

Conflict of interest statement: None declared.

REFERENCES

1. 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.
2. Held,G.A. Grinstein,G. and Tu,Y. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc Nat Acad Sci. USA*, **100**, 7575.
3. Halperin,A. Buhot,A. and Zhulina.E.B. (2004) Specificity, sensitivity and the hybridization Isotherms of DNA chips. *Biophys. J.*, **86**, 718–730.
4. Abdueva,D. Skvortsov,D. and Tavare,S. (2006) Non-linear analysis of GeneChip arrays. *Nucleic Acids Res.*, **34**, e105.
5. Binder,H. and Preibisch,S. (2006) GeneChip microarrays-signal intensities, RNA concentrations and probe sequences. *J.Phys. Condens. Matter*, **18**, S537–S566.
6. Heim,T. Tranchevent L.C., Carlon,E. and Barkema G.T. (2006) Physical-chemistry-based analysis of affymetrix microarray data. *J. phys. chem. B*, **110**, 22786–22795.
7. Held,G.A. Grinstein,G. and Tu,Y. (2006) Relationship between gene expression and observed intensities in DNA microarrays–a modeling study. *Nucleic Acids Res.*, **34**, e70.
8. Burden,C.J. Pittelkow,Y. and Wilson,S.R. (2006) Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays. *J. Phys. Condensed Matter*, **18**, 5545–5565.
9. Bruun,G.M. Wernersson,R. Juncker,A.S. Willenbrock,H. and Nielsen,H.B. (2007) Improving comparability between microarray probe signals by thermodynamic intensity correction. *Nucleic Acids Res.*, **35**, e48.
10. Burden,C.J. (2008) Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed. *Phys. Biol.*, **5**, 16004.
11. Skvortsov,D. Abdueva,D. Curtis,C. Schaub,B. and Tavare,S. (2007) Explaining differences in saturation levels for Affymetrix GeneChip (R) arrays. *Nucleic Acids Res.*, **35**, 4154–4163.
12. Affymetrix Latin Square data. http://www.affymetrix.com/support/datasets.affx (6 May 2008, date last accessed).
13. Zhang,L. Miles M.F., and Aldape KD. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.
14. 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. Nat. Acad. Sci.*, **100**, 11237.
15. Wu,Z. and Irizarry,R.A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol.*, **22**, 656–658.
16. SantaLucia. Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Nat. Acad. Sci. USA*, **95**, 1460.
17. Wu,P. Nakano,S. and Sugimoto,N. (2002) Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. *FEBS J.*, **269**, 2821–2830.
18. Ono,N. Suzuki,S. Furusawa,C. Agata,T. Kashiwagi,A. Shimizu,H. and Yomo,T. (2008) An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. *Bioinformatics*, **24**, 1278–1285.
19. Naef,F. and Magnasco,M.O. (2003) Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E*, **68**, 11906.
20. NetAffx Analysis Center. http://www.affymetrix.com/analysis/index.affx (6 May 2008, date last accessed).
21. Mathews,D.H. Disney,M.D. Childs,J.L. Schroeder,S.J. Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, *Proc. Nat. Acad. Sci. USA*, **101**, 7287–7292.
22. Turner,D.H. (2000) In Bloomfield,V.A. Crothers,D.M., and Tinoco, Jr I. (eds), *Nucleic Acids: Structures, Properties, and Functions*, Sausalito University Science Books, Sausalito, CA, pp. 259–334.
23. Christensen,U. Jacobsen,N. Rajwanshi,V.K. Wengel,J. and Koch,T. (2001) Stopped-flow kinetics of locked nucleic acid (LNA)-oligonucleotide duplex formation: studies of LNA-DNA and DNA-DNA interactions. *Biochem. J.*, **354**, 481–484.
24. Levicky,R. and Horgan,A. (2005) Physicochemical perspectives on DNA microarray and biosensor technologies. *Trends Biotechnol.*, **23**, 143–149.
25. Halperin,A. Buhot,A. and Zhulina,E.B. (2006) On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions. *J. Phys. Condens. Matter*, **18**, S463–S490.
26. 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.
27. Halperin,A. Buhot,A. and Zhulina,E.B. (2005) Brush Effects on DNA Chips: Thermodynamics, Kinetics, and Design Guidelines. *Biophys. J.*, **89**, 796–811.
28. Zhang,L. Hurek,T. and Reinhold-Hurek,B. (2005) Position of the fluorescent label is a crucial factor determining signal intensity in microarray hybridizations. *Nucleic Acids Res.*, **33**, e166.
29. Zhang,Y. Hammer,D.A. and Graves,D.J. (2005) Competitive hybridization kinetics reveals unexpected behavior patterns. *Biophys. J.*, **89**, 2950–2959.
30. Bishop,J. Blair,S. Chagovetz.A.M. (2006) A competitive kinetic model of nucleic acid surface hybridization in the presence of point mutants. *Biophys. J.*, **90**, 831–840.
31. Shi,L. Tong,W. Su,Z. Han,T. Han,J. Puri,R.K. Fang,H. Frueh,F.W. Goodsaid,F.M. Guo,L. *et al.*, (2005) Microarray scanner calibration curves: characteristics and implications. *BMC Bioinformatics*, **6**, S11.
32. Wu,Z. 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.
33. Schuster,E.F. Blanc,E. Partridge,L. and Thornton,J.M. (2007) Estimation and correction of non-specific binding in a large-scale spike-in experiment. *Genome Biol.*, **8**, R126.