BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 10 2008, pages 1278–1285
doi:10.1093/bioinformatics/btn109

Gene expression

An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays

Naoaki Ono[1], Shingo Suzuki[1], Chikara Furusawa[1,2,*], Tomoharu Agata[1], Akiko Kashiwagi[1], Hiroshi Shimizu[1] and Tetsuya Yomo[1,2,3]

[1]Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, 2-1 Yamadaoka, [2]Complex Systems Biology Project, ERATO, 2-1 Yamadaoka and [3]Graduate School of Frontier Biosciences, Osaka University, 1-3 Yamadaoka, Suita, Osaka 565-0871, Japan

Received on December 10, 2007; revised on March 23, 2008; accepted on March 24, 2008

Advance Access publication March 31, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: High-density DNA microarrays provide useful tools to analyze gene expression comprehensively. However, it is still difficult to obtain accurate expression levels from the observed microarray data because the signal intensity is affected by complicated factors involving probe–target hybridization, such as non-linear behavior of hybridization, non-specific hybridization, and folding of probe and target oligonucleotides. Various methods for microarray data analysis have been proposed to address this problem. In our previous report, we presented a benchmark analysis of probe–target hybridization using artificially synthesized oligonucleotides as targets, in which the effect of non-specific hybridization was negligible. The results showed that the preceding models explained the behavior of probe–target hybridization only within a narrow range of target concentrations. More accurate models are required for quantitative expression analysis.

Results: The experiments showed that finiteness of both probe and target molecules should be considered to explain the hybridization behavior. In this article, we present an extension of the Langmuir model that reproduces the experimental results consistently. In this model, we introduced the effects of secondary structure formation, and dissociation of the probe–target duplex during washing after hybridization. The results will provide useful methods for the understanding and analysis of microarray experiments.

Availability: The method was implemented for the R software and can be downloaded from our website (http://www-shimizu.ist.osaka-u.ac.jp/shimizu_lab/FHarray/).

Contact: furusawa@ist.osaka-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

DNA microarrays have been used for a wide range of studies on expression analysis. High-density oligonucleotide microarrays use a set of short oligonucleotide probes to measure gene expression and they allow us to analyze the expression of thousands of genes quantitatively in a single experiment (Lipshutz *et al.*, 1999; Selinger *et al.*, 2000). Various methods have been studied to improve the quality of analysis measured by the microarrays (Cope *et al.*, 2004; Irizarry *et al.*, 2006). In the standard protocol provided by Affymetrix's tool [Microarray Analysis Suite ver. 5.0 (MAS)], a number of probes are designed for a single gene and an expression level is estimated by the weighted mean of their signal intensities. To improve the accuracy and robustness of the expression analysis, various statistical models (Irizarry *et al.*, 2003; Li and Wong, 2001; Wu and Irizarry, 2004) have been proposed, in which affinities between each probe and target are estimated using multiple array data. These methods relied mainly on the linearity between the concentration of any target molecule and the amount of hybridization measured by the fluorescent intensity of its probe. However, experimental results showed that the linearity is maintained within a rather narrow range of concentration, about 2–3 orders of magnitude (Chudin *et al.*, 2002), which depended on both the lower limit of fluorescence measurement and the saturation level of probe–target hybridization.

To expand the dynamic range of the measurement, a promising approach is to model the non-linear behavior of hybridization in detail. It has been accepted that the Langmuir adsorption model explains the behavior of hybridization (Burden *et al.*, 2006; Hekstra *et al.*, 2003; Held *et al.*, 2003). Analysis based on that model showed that the signal intensity significantly depended on the hybridization free energy between probe and target (Mei *et al.*, 2003) and the free energy can be estimated from the probe sequence (SantaLucia, 1998; Zhang *et al.*, 2003). However, it has been problematic in that the intensity level observed in usual spike-in experiments includes the effect of non-specific target, namely, ensembles of oligonucleotide fragments that do not complement the probes perfectly (Wu *et al.*, 2005). Although intense studies on the calibration of this model have been performed (Shippy *et al.*, 2004; Yuen *et al.*, 2002), experiments under more ideal conditions where such non-specific hybridization can be negligible are needed for the more accurate analysis of microarray data.

In Suzuki *et al.* (2007a) we presented spike-in experiments without background, namely, in which only artificially synthesized oligonucleotides were hybridized onto a custom designed

*To whom correspondence should be addressed.

microarray as a dilution series. The results provided us clear and accurate information of hybridization behavior because we could neglect the signal intensities of non-specific targets. Analyzing these experimental results, we found that the intensity showed two types of saturation, depending on the target concentration. When the target concentration was high, the probe intensity saturated to the same level as Langmuir-type models predict. This indicated that all probe molecules hybridized with the target. On the other hand, when the target concentration was low, the intensity saturated to different lower levels. Since the levels were correlated with the target concentration, these results suggested that the target molecules were depleted in the hybridization process [see Suzuki *et al.* (2007a) for details].

In this article, we introduce an extension of Langmuir-type thermodynamic model of hybridization to reproduce these behavior of hybridization and improve the accuracy and dynamic range of measurements. This model considers basic duplex formation and depletion of both probe and target molecules so that it explains the experimental results reported in Suzuki *et al.* (2007a) consistently. Furthermore, based on this hybridization model, we took other physical effects of probe–target interaction into account in order to improve the accuracy of the model. First, though it has been pointed out that the probes undergo folding (Binder *et al.*, 2004), the contribution of this to microarray hybridization has not been estimated quantitatively. In this study, we evaluated the effect of secondary structure and integrated it into our model. We also considered the effects of dissociation of the probe–target duplex during the washing process after hybridization (Wick *et al.*, 2006), to explain the difference of the saturation level between observed intensity and that expected by the equilibrium model.

Next, we present the results of another spike-in experiment with background, namely, the oligonucleotide controls were mixed with cDNA sample obtained from the transcriptome of *Escherichia coli*. The comparison between the experiments with and without transcriptome background made clear the effects of the non-specific targets for microarray analysis. We found two major effects. First, when the target concentration was low, some probes in the case with cDNA sample showed much greater intensity than those in the case without cDNA sample. This can be attributed to non-specific hybridization which has long been discussed as a cause of spurious signals (Kane *et al.*, 2000; Naef and Magnasco, 2003; Wu *et al.*, 2005). On the other hand, we also found that the probes tended to show lower intensity in experiments with cDNA sample than those without them, when the target concentration was high. The result suggested that target molecules that hybridize with non-specific targets in the bulk solution decrease the effective target concentration (Binder, 2006; Halperin *et al.*, 2004). For an accurate estimate of target concentrations, these effects of non-specific hybridization and bulk hybridization should be taken into account. In this study, we introduced the terms for both hybridization effects and showed that the model reproduced the behavior of the observed intensity. Finally, using this model, we estimated the nominal target concentration from observed intensity in the experiments with non-specific targets. This showed that the dynamic range of the measurement achievable with our improved physico-chemical model was over 5 orders of magnitude.

2 MODELS

2.1 Langmuir model

The Langmuir adsorption model has been used widely to model microarray hybridization (Burden *et al.*, 2006; Hekstra *et al.*, 2003; Held *et al.*, 2003). In the Langmuir model, the probe intensity is given as follows:

$$I^{\mathrm{Langmuir}} = \alpha \frac{Kx}{1+Kx} + I^{\mathrm{bg}}, \tag{1}$$

where α gives the scale of intensity, K gives the equilibrium constant of probe–target duplex formation, x gives the concentration of target molecules and I^{bg} denotes the optical background intensity. The equilibrium constant is defined by $K = \exp(-\Delta G/RT)$, where ΔG denotes the free energy of the hybridization, R denotes the gas constant and T denotes the temperature. In this model, when $Kx \gg 1$, namely, the affinity of the probe is very strong or the target concentration is high enough, the first term saturates to the constant α, which implies that all probes bind to their target molecules.

2.2 Zhang's gene-specific hybridization model

Zhang and others used slightly different functions to estimate their position-dependent nearest neighbor model (Zhang *et al.*, 2003), which is given by:

$$I^{\mathrm{Zhang}} = \alpha' \left(\frac{x}{1+K'} + \frac{N}{1+K''} \right) + I^{\mathrm{bg}}, \tag{2}$$

where x and N denote the population of the target molecules for gene-specific hybridization and that of RNA molecules that contributes to non-specific hybridization, respectively. $K' = \exp(E)$ and $K'' = \exp(E^*)$, where the E and E^* are the free energy for gene-specific hybridization and the average free energy for non-specific hybridization, both scaled by RT, respectively. This model also assumes that the intensity saturates as the affinity of the probe increases, but that the saturation level is proportional to the target concentration. It represents the state where all available target molecules are bound to their probes.

2.3 Finite hybridization model

In this article, we introduce a Finite Hybridization (FH) model. Held and others proposed a simple equilibrium model of the binding between probe and target molecules (Held *et al.*, 2003) based on the equation:

$$\mathrm{P}^{\mathrm{free}} + \mathrm{T}^{\mathrm{free}} \overset{K^{\mathrm{sp}}}{\Longleftrightarrow} \mathrm{PT}^{\mathrm{sp}}, \tag{3}$$

where $\mathrm{P}^{\mathrm{free}}$, $\mathrm{T}^{\mathrm{free}}$ are free probe and target molecules and $\mathrm{PT}^{\mathrm{sp}}$ is their duplex, K^{sp} gives the equilibrium constant of gene-specific hybridization between them. If one assumes that the system reaches equilibrium, and takes mass conservation of probe and target molecules into account, the amount of probe and target molecules can be described by the following equations:

$$\left[\mathrm{PT}^{\mathrm{sp}}\right] = K^{\mathrm{sp}} \left[\mathrm{P}^{\mathrm{free}}\right] \left[\mathrm{T}^{\mathrm{free}}\right] \tag{4}$$

$$\left[\mathrm{P}^{\mathrm{total}}\right] = \left[\mathrm{P}^{\mathrm{free}}\right] + \left[\mathrm{PT}^{\mathrm{sp}}\right] \tag{5}$$

$$\left[\mathrm{T}^{\mathrm{total}}\right] = \left[\mathrm{T}^{\mathrm{free}}\right] + \left[\mathrm{PT}^{\mathrm{sp}}\right] \tag{6}$$

where $[P^{total}]$ and $[T^{total}]$ represent their total concentration. From Equations (4–6), we obtain the amount of hybridized molecules. Then the intensity expected by the FH model is given as follows:

$$\begin{aligned} I^{FH} &= C\left[PT^{sp}\right] + I^{bg} \\ &= \frac{C}{2}\left\{\frac{1}{K^{sp}} + A + x \right. \\ &\quad \left. - \sqrt{\left(\frac{1}{K^{sp}} + A + x\right)^2 - 4Ax}\right\} + I^{bg}, \end{aligned} \tag{7}$$

where C is the scale of intensity, $A = \left[P^{total}\right]$, $x = \left[T^{total}\right]$ and I^{bg} is the optical background intensity.

It is worth noting that the both Langmuir and Zhang models are limiting cases of the FH model. Namely, when $x \gg A$, Equation (7) can be approximated by the following equation:

$$I^{FH} \simeq AC\frac{K^{sp}x}{1 + K^{sp}x} + I^{bg}. \tag{8}$$

It is clear that Equation (8) is identical to the Langmuir equation [Equation (1)], given that $\alpha = AC$. On the other hand, when $x \ll A$, Equation (7) is approximated by:

$$I^{FH} \simeq AC\frac{K^{sp}x}{1 + AK^{sp}} + I^{bg}. \tag{9}$$

Given that $\alpha' = C$ and $K' = 1/AK^{sp}$, it corresponds to the first term of Equation (2).

2.4 Nearest neighbor model

The free energy of specific hybridization is calculated using the nearest neighbor (NN) model (SantaLucia, 1998). Given the base sequence of the probe provided by $\mathbf{b} = (b_1, \ldots, b_l)$, the hybridization free energy is given as follows:

$$\Delta G^{sp}(\mathbf{b}) = \sum_{k=1}^{l-1} \epsilon^{sp}(b_k, b_{k+1}), \tag{10}$$

where ϵ^{sp} (b_1, b_2) denotes the binding and stacking energy of two given base pairs and l indicate the probe length.

2.5 Effect of secondary structure

Folding of the probes affects the efficiency of hybridization (Binder *et al.*, 2004). To estimate this effect, we consider the equilibrium of the probes between the folded (P^{fold}) and free (P^{free}) states:

$$P^{free} \overset{K^{fold}}{\Longleftrightarrow} P^{fold}. \tag{11}$$

Taking mass conservation into account, the amount of probe–target duplex at equilibrium is described by the same equation as Equations (4–6) except that the equilibrium constant K^{sp} is replaced by an effective equilibrium constant:

$$K^{eff} = \frac{K^{sp}}{1 + K^{fold}}. \tag{12}$$

An algorithm named UNAfold, based on the thermodynamics of DNA folding, has been proposed (Markham and Zuker, 2005) to calculate the free energy of hybridization. Although the folding of microarray probes is affected by the interaction with surface of microarray, we assume that the free energy of the folding is proportional to that in the bulk solution calculated by UNAfold. Therefore, the equilibrium constant of the folding is as follows:

$$K^{fold} = \exp\left(-\frac{w^{fold}\Delta G^{fold}}{RT}\right), \tag{13}$$

where w^{fold} is an adjustable weight factor and ΔG^{fold} denotes the free energy of the folding calculated by UNAfold.

2.6 Effect of dissociation

Dissociation of the probe–target duplex during the washing process has been considered as a non-equilibrium process that decrease signals (Burden *et al.*, 2006; Held *et al.*, 2006; Wick *et al.*, 2006). We assume that the dissociation rate constant k^{dis} depends on the hybridization energy, namely, it is proportional to $\exp(-w^{dis}\Delta G^{sp}/RT)$, where w^{dis} is an adjustable weight parameter. According to the dissociation rate, the amount of duplex after the wash decreases exponentially, thus, Equation (7) is changed as follows:

$$d = \exp\left\{-B\exp\left(-\frac{w^{dis}G}{RT}\right)\right\} \tag{14}$$

$$I^{FH} = Cd\left[PT^{sp}\right] + I^{bg}, \tag{15}$$

where B is another adjustable constant related to the duration of the wash.

2.7 Competitive hybridization of specific and non-specific targets

Next, we consider the competitive hybridization of specific and non-specific targets. To explain the effect of non-specific targets observed in experiments with cDNA samples, we introduced two effects: non-specific hybridization and bulk hybridization. By addition of Equations (3) and (11), we consider the following reactions:

$$P^{free} + T^{ns} \overset{K^{ns}}{\Longleftrightarrow} PT^{ns} \tag{16}$$

where T^{ns} represents the non-specific target, that is, the ensemble of DNA fragments that hybridize randomly with the probe or target, K^{ns} and K^{bulk} represent the average affinity of non-specific hybridization and bulk hybridization, respectively. Zhang and others described the average free energy of non-specific hybridization using a similar model to that of specific hybridization (Zhang *et al.*, 2003). In our model, the average free energy of non-specific hybridization is calculated using another set of parameters as follows:

$$\Delta G^{ns}(\mathbf{b}) = \sum_{k=1}^{l-1} \epsilon^{ns}(b_k, b_{k+1}), \tag{17}$$

where ϵ^{ns} represents the binding energy of non-specific hybridization.

Following the model of bulk hybridization proposed in (Binder, 2006) we regard the bulk hybridization as hybridization with non-specific targets in the solution.

$$T^{free} + T^{ns} \overset{K^{bulk}}{\Longleftrightarrow} TT^{ns}. \tag{18}$$

As the analogy of non-specific hybridization with the probes we estimate K^{bulk}, assuming that the free energy of bulk

hybridization is proportional to that of non-specific hybridization $\Delta G^{\mathrm{bulk}} = w^{\mathrm{bulk}} \Delta G^{\mathrm{ns}}$, where w^{bulk} is an adjustable parameter which represents the difference of hybridization condition between solution and the array surface.

Because we consider the average effect of non-specific hybridization with various fragments of DNA, we assume that the total amount of non-specific target N does not depend on the probes, and it is much larger than either the amount of the specific target or that of the probes. Then, the intensity is given by the sum of specific and non-specific hybridization, thus:

$$I^{\mathrm{FH}} = C(d^{\mathrm{sp}}[\mathrm{PT}^{\mathrm{sp}}] + d^{\mathrm{ns}}[\mathrm{PT}^{\mathrm{ns}}]) + I^{\mathrm{bg}}, \tag{19}$$

where d^{sp} and d^{ns} denote the dissociation coefficients for specific and non-specific targets given by Equation (14) and

$$[\mathrm{PT}^{\mathrm{sp}}] = \frac{1}{2}\left\{ \frac{1}{K^{\mathrm{eff}}} + A + x - \sqrt{\left(\frac{1}{K^{\mathrm{eff}}} + A + x\right)^2 - 4Ax} \right\} \tag{20}$$

$$[\mathrm{PT}^{\mathrm{ns}}] = \frac{(A - [\mathrm{PT}^{\mathrm{sp}}])K^{\mathrm{ns}}N}{1 + K^{\mathrm{fold}} + K^{\mathrm{ns}}N} \tag{21}$$

$$K^{\mathrm{eff}} = \frac{K^{\mathrm{sp}}}{(1 + K^{\mathrm{fold}} + K^{\mathrm{ns}}N)(1 + K^{\mathrm{bulk}}N)}. \tag{22}$$

2.8 Parameter optimization

In the FH model, there are 27 parameters that are adjusted to fit the model to the observed data: 10 parameters of the NN model to estimate the free energy of hybridization; 10 parameters for non-specific hybridization; three parameters to scale the system, namely, the total amount of the probes, the coefficient for intensity, and optical background constant and four weighting factors for the estimation, one for folding, two for dissociation and one for bulk hybridization, respectively. We optimized these model parameters by minimizing the mean residual error (R) between the observed and expected probe intensity:

$$R = \sum_{i,j} (\log_{10} I_{ij}^{\mathrm{obs}} - \log_{10} I_{ij}^{\mathrm{pre}})^2 / M, \tag{23}$$

where I_{ij}^{obs} and I_{ij}^{pre} are the observed and predicted probe intensities of the ith probe in jth experiments, respectively, and M is the number of data points. In this study, $M = 37\,800$ data points—5400 probes in seven experiments—were used for the analysis. The optimization of the parameters was performed using a greedy method based on Monte Carlo simulation the detailed algorithm is described in Supplementary Material.

3 RESULTS

3.1 Design of oligonucleotide probes

We synthesized 150 species of 25 mer oligonucleotides using artificial random sequences as control targets, and designed a custom microarray whose probes were complementary to the control targets. The oligonucleotide microarray were synthesized on the Maskless Array Synthesizer platform (Nuwaysir *et al.*, 2002; Singh-Gasson *et al.*, 1999). We arranged 25 mer probes, which were perfectly complementary to the targets, but also placed shorter probes to observe the effect of any difference in hybridization affinity. The original 25 mer probes were shortened from one end by one base, so that 12 different probe lengths ranging from 14 mer to 25 mer were designed for each of the 150 targets. Because we arranged three copies for each probe, 5400 probes could be used in total for the analysis (see Suzuki *et al.*, 2007a for detail). The extracted microarray data were analyzed using custom-designed scripts in R software (R Development Core Team, 2006). In each experiment, replicates correlated well ($r > 0.94$), indicating a high level of reproducibility. To obtain a single absolute signal intensity for each probe, we average logged values of the replicated measurements.

3.2 Evaluation of the three hybridization models

First, we evaluated the three hybridization models, i.e. the Langmuir, Zhang and FH models, by the experiments without background. We optimized the three models using intensity data of the probes that were complementary to the control targets in the seven experiments. Then, we compared how the models reproduced the behavior of the observed intensity at 1.4 fM to 1.4 nM. Although the microarray had some other probes whose sequences were irrelevant to these targets, the intensities of these probes were very low compared with that of the specific hybridization (data not shown). Thus, the effects of non-specific hybridization were negligible in this series of experiments.

Remember that we have arranged different lengths of probes for each target. As Equation (10) implies that ΔG^{sp} is roughly proportional to the length of the probe, we first focused on the dependency of probe intensity on probe length. Figure 1a shows the results of experiments at seven target concentration levels. Each line represents the average intensity of 150 probes observed in the experiments as a function of the probe length. The intensity saturated as the probes become longer, i.e. as the affinity of each probe increases. However, the behavior depended on the target concentration. When the target concentration was lower than 1.4 pM, the saturation level was proportional to the concentration. On the other hand, when the target concentration was higher than 14 pM, the intensity saturated to the same level.

Next, Figures 1b–d illustrate the averages of the predicted intensity as a function of the probe length using each model, after parameter optimization. The Zhang model (Fig. 1b) reproduced the actual behavior in that the saturation level was proportional to the target concentration when it was lower than 1.4 pM. The Langmuir model (Fig. 1c) explained saturation to its maximum intensity, when the target concentration was higher than 14 pM. The FH model reproduced both types of saturation, so that it fit to the observed data better than the other models over the whole range of target concentrations (Fig. 1d). Still, there was a difference between the experimental data and those expected by the model, that is, the signal intensities of experimental data gradually increased with the probe length in the range of longer probes (e.g. longer than 22 mer), where the prediction of the model converged to constant levels. This difference will be explained in the next section by the effect of dissociation during the washing process.

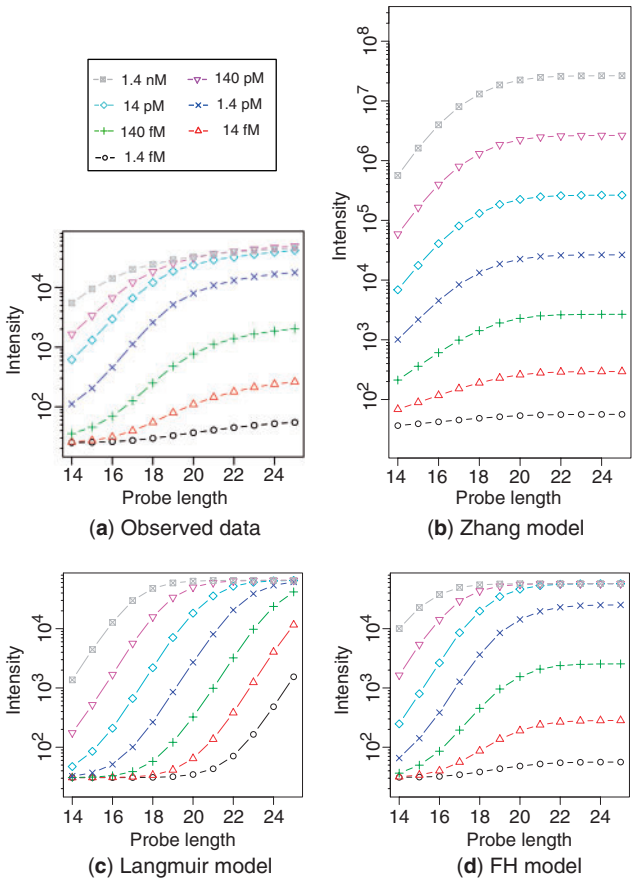

Fig. 1. Behavior of the observed probe intensity and comparison with theoretical models. The average intensities of all 150 species are plotted as functions of the probe length. (**a**) Observed intensity. When the target concentration x was lower than 1.4 pM, the saturation level depended on the target concentration, whereas it saturated to the same level when the target concentration is higher than 14 pM. Average intensity is shown as predicted by the Zhang (**b**), Langmuir (**c**) and FH (**d**) models. Because the Zhang model ignores the saturation of probe molecules and the Langmuir model ignores the depletion of target molecules, they fit only partially to the observed data. The FH model reproduced the behavior of observed intensity well over the whole concentration range.

3.3 Effect of secondary structure and dissociation

Based on the FH model evaluated in the previous section, we then attempted to improve the accuracy of the model's predictions. First, we took the effect of secondary structure formation into account. Although it has been pointed out that probes with stable secondary structures tend to show lower intensity (Matveeva *et al.*, 2003), it has been difficult to quantify this effect. To evaluate this effect on probe intensity, we compared the stability of the expected secondary structure of the probes against the residual errors between the observed intensity and that calculated by the FH model. The stability of the secondary structure was evaluated using the UNAfold model, as proposed by Zuker and others (Markham and Zuker, 2005). We found that the residual errors between observed and predicted intensity correlated negatively ($r = -0.36$) with the free energy of the secondary structure of probes calculated using UNAfold (Supplementary Fig. 1). This suggested that incorporating the effect of secondary structure formation into the FH model can help decrease the residual errors. To incorporate this effect, we took the equilibrium between the folded and unfolded state of the probes Equation (11) into account so that the equilibrium constant K^{sp} was replaced by K^{eff} given by Equation (12). Using this model, we re-optimized all parameters to reduce the residual errors.

As we pointed out in Figures 1a and d, the signal intensities of shorter probes are smaller than the expected saturation levels. To explain this difference, we focused on the relationship between the saturation level and the hybridization free energy of the probes. We compared the saturation level of the probes with different hybridization energies and confirmed that the saturation levels of probes with lower hybridization free energies are significantly lower than those with higher free energies (Supplementary Fig. 2). However, the model expects that their intensities reach the same saturation level when all probes are hybridized with target molecules. Possible causes of this difference were that the hybridization had not yet reached equilibrium, or that the probe–target duplexes dissociated after hybridization. Therefore, even if the saturation level at equilibrium is the same, the dissociation rate in the washing process might depend on probe–target affinity. It has been pointed out that the washing process after hybridization of the labeled targets affects the intensity of the less stable probe–target duplexes, for example, duplexes containing mismatched base pairs are washed out more easily (Suzuki *et al.*, 2007b; Wick *et al.*, 2006). Following the previous studies, we introduced terms of dissociation Equations (14) and (15) into the model to estimate their effect.

To confirm validity of the introduced parameters, we evaluated the three models: 1) estimation of duplex formation based on only the NN model; 2) the effect of introducing a secondary structure and 3) the effect of dissociation by estimating their prediction error using 5-fold cross-validation method. As we added three adjustable parameters (w^{fold} for the secondary structure model, and B and w^{dis} for the dissociation model) to the normal NN model, the estimated prediction errors of these models reduced to 6.7×10^{-2} and 6.1×10^{-2}, from that of the original model (7.0×10^{-2}). The difference of the prediction errors between these models were significant (by Mann–Whitney U-test, $P < 10^{-2}$).

3.4 Effects of non-specific hybridization

In this section, we introduce the effects of non-specific hybridization to the FH model and evaluate it using experimental data from a spike-in experiment with background. In this experiment, the spike-in control oligonucleotides were mixed with cDNA generated from the total RNA of *E.coli*. The concentration levels of the spike-in controls were the same as in previous experiments: i.e. 1.4 fM to 1.4 nM (Suzuki *et al.*, 2007a).

First, we compared the intensity of spike-in controls observed under the condition without background (I^{without}) and that mixed with the background (I^{with}). It is worth noting that two different effects can be observed in the distribution of the intensity ratio ($I^{\mathrm{with}}/I^{\mathrm{without}}$) (Fig. 2), depending on the

target concentration. When the target concentration was lower than 14 fM, I^{with} tended to show a higher intensity than I^{without} because of non-specific hybridization. On the other hand, as the target concentration increased, I^{with} became slightly smaller than I^{without}. This result can be attributed to bulk hybridization. It has been suggested that the target molecules might hybridize with other target molecules non-specifically in bulk solution (Binder, 2006; Burden *et al.*, 2006; Halperin *et al.*, 2004). This effect would decrease the amount of free target oligonucleotides available to hybridize with the probes.

In the FH model, we introduced 12 more parameters to estimate the effect of non-specific and bulk hybridization, namely, 10 parameters for the estimation of non-specific hybridization energy (ϵ^{ns}), the weight parameter w^{bulk} for bulk hybridization, and the total amount of molecules that contribute to the non-specific hybridization (N). We optimized all parameters again using the observed data of the spike-in experiments with background. We confirmed that the NN model for non-specific hybridization provided a reliable estimation of signals caused by the background addition. The analysis for non-specific hybridization is shown in Supplementary Material.

Figure 3 shows the distribution of the residual error of the prediction using the FH model. Scatter plots of observed against expected intensity are shown in Supplementary Figure 3. The mean residual error was 6.1×10^{-2}, and 93% of the observed points were within a 3-fold range of the predicted intensity. Since the model was heavily parameterized, we evaluated the prediction error of the model using 5-fold cross-validation method. When the model of specific hybridization presented in the previous section was applied for the spike-in experiments with background, the estimated prediction error was $R = 8.1 \times 10^{-2}$, while that of the model with the effects of non-specific hybridization reduced to $R = 6.1 \times 10^{-2}$ and the difference was significant (by Mann–Whitney U-test, $P < 10^{-2}$). Since the test datasets were separated from training data for parameter optimization in the scheme of cross-validation, it is clear that the results were not an artifact due to over parameterization.

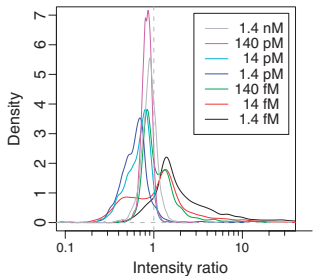

Fig. 2. Effects of a non-specific target. Distribution of the intensity ratio between spike-in experiment with and without background ($I^{\text{with}}/I^{\text{without}}$). The right-hand peak was observed when the target concentration was lower than 14 fM, suggesting the effect of non-specific hybridization. On the other hand, when the target concentration was higher than 140 fM, the average ratio was smaller than 1, suggesting the effects of bulk hybridization.

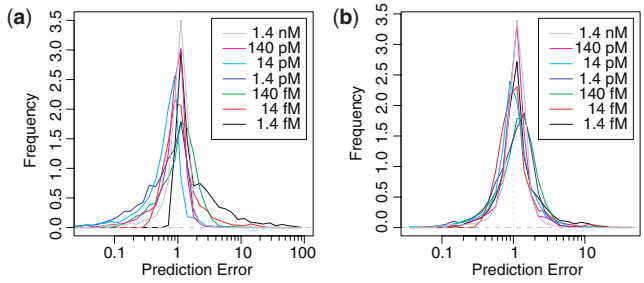

Fig. 3. Distribution of the prediction error. (**a**) The ratio of the observed intensity to the prediction of the model without the effects of a non-specific target. (**b**) The ratio of the observed intensity to the prediction with the effects of a non-specific target.

3.5 Accuracy test

Next, based on the FH model, we propose a method to estimate the target concentration from the observed intensity. Given the probe sequences and the model parameters, the residual error R in Equation (23) is computed as a function of target concentration. Therefore, the target concentration can be estimated by minimizing the residual error between the observed and predicted intensity. In this section, we evaluated this method using the data of the spike-in experiments under the condition with the background. To evaluate this method. We used 100 sets of 25 probes to estimate the target concentration, which were randomly chosen from all the probes on the array. Using the same set of probes, we also estimated target concentrations using a) Affymetrix's MAS (Affymetrix, 2001) and b) Robust Multiarray Average (RMA) (Irizarry *et al.*, 2003) (Figs 4a and b). We also compared the Langmuir and Zhang models, in that we estimated the optimal target concentration using Equations (8) or (9) instead of the first term of 20, and other

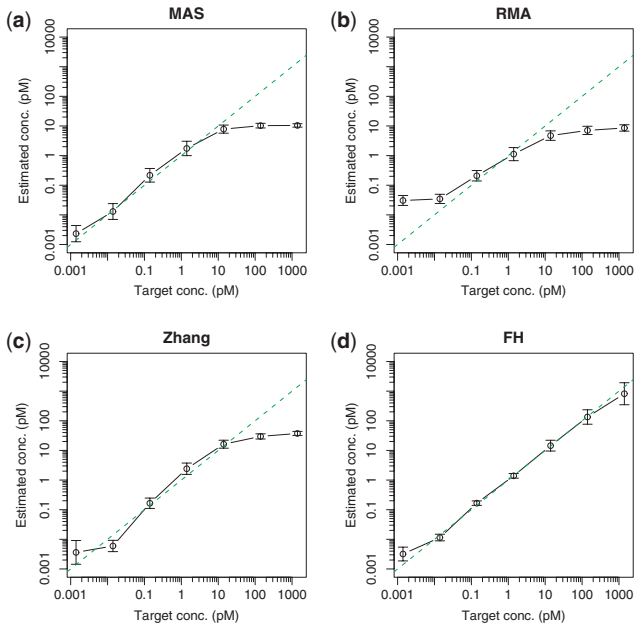

Fig. 4. Estimation of target concentrations. We evaluated 100 probe sets each contains randomly chosen 25 probes. The average estimated concentrations are plotted against nominal concentration. The error bars represent the SD of the 100 estimations.

secondary effects were calculated as in the FH model. The estimations produced by MAS and RMA tended to be lower than nominal values at higher concentrations because these methods are based on linear models that do not consider any saturation, as shown in Figures 1a and b. Furthermore, the Langmuir model failed to fit the parameters because of the difference in saturation behavior between experimental results and those assumed by the model. For the Langmuir model, the estimated target concentration did not correlate with the nominal concentration (data not shown) and the residual error of each probe set was much larger than that of the other methods. The results of the Zhang and the FH models are shown in Figures 4c and d. Estimation using the Zhang model was also affected by saturation. However if all the probes in the given probe set do not completely saturate, the FH model could estimate the nominal target concentration quantitatively by comparing the prediction and observed intensity.

Finally, we evaluated our model using Affymetrix's HUG133a Latin square spike-in experiments data (http://www.affymetrix.com/analysis/download-center2.affx). Since the physical features of the microarray and the conditions of hybridization were different, the model parameters were changed to fit the given data. The results showed that the prediction errors of the model were as small as that in our experiments (see Supplementary Material, Section 3, in detail), and the estimations of target concentration by our model reproduced the nominal target concentration quantitatively in all concentration range as shown in Supplementary Figure 4.

4 DISCUSSION AND CONCLUSION

These experiments using artificially synthesized oligonucleotides as targets have revealed details of probe–target hybridization. Based on the results of the experiments, we have identified the source of the errors in previous hybridization models and have introduced an improved thermodynamic model. First, the non-linearity between probe intensity and target concentration was attributed to the depletion of probe and target molecules. Second, we took the effect of secondary structures and dissociation during the washing process into account to improve the accuracy of the prediction. Though in this study, we roughly approximated the activation energy of dissociation in wash process to explain the relationship between estimated hybridization energy and the decrease of the intensity observed in the saturated hybridization condition. However, detailed dynamics of dissociation would be more complicated. For example, in Pozhitkov *et al*. (2007), it was pointed out that no significant difference was found between Perfect Match (PM) and the corresponding MisMatch (MM) probes whose hybridization energy is expected to be lower than that of PM probes. Detailed understanding of non-equilibrium dynamics in wash process will be required for more accurate analysis.

Because our model is based on a physico-chemical model of hybridization, it would be easy to add other physical effects, for example, the effect of base position (Zhang *et al*., 2003), mismatch (Binder *et al*., 2005; Naef *et al*., 2002), and others into this framework.

Using this model, we proposed a method for the estimation of target concentration. We confirmed the model using a spike-in experiment and showed that the concentration range over which the estimation was valid over 5 orders of magnitude, which was much wider than preceding methods. This algorithm will allow us to analyze gene expression in more detail. For example, when there are 10^8 cells in a sample, our method makes it possible to measure from 0.01 to 1000 mRNA molecules per cell. Development of analysis based on this method will greatly improve quantitative analyzes of gene-expression levels using microarrays.

ACKNOWLEDGEMENTS

Funding: This work was supported by 'Special Coordination Funds for Promoting Science and Technology: Yuragi Project', and 'the Global Centers of Excellence Program' of the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

Conflicts of Interest: none declared.

REFERENCES

Affymetrix (2001) *New statistical algorithms for monitoring gene expression on GeneChip Probe Arrays*. Technical Note. Affymetrix, http://www.affymetrix.com/support/technical/technotes/statistical_algorithms_technote.pdf

Binder,H. (2006) Thermodynamics of competitive surface adsorption on DNA microarrays. *J. Phys. Condens. Matter*, **18**, S491–S523.

Binder,H. *et al*. (2004) Sensitivity of microarray oligonucleotide probes: variability and effect of base composition. *J. Phys. Chem. B*., **108**, 18003–18014.

Binder,H. *et al*. (2005) Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *Langmuir*, **21**, 9287–9302.

Burden,C. *et al*. (2006) Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays. *J. Phys. Condens. Matter*, **18**, 5545.

Chudin,E. *et al*. (2002) Assessment of the relationship between signal intensities and transcript concentration for affymetrix genechip arrays. *Genome. Biol*., **3**, RESEARCH0005.

Cope,L. *et al*. (2004) A benchmark for affymetrix genechip expression measures. *Bioinformatics*, **20**, 323–331.

Halperin,A. *et al*. (2004) Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophys. J*., **86**, 718–730.

Hekstra,D. *et al*. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res*., **31**, 1962–1968.

Held,G. *et al*. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl Acad. Sci. USA*, **100**, 7575–7580.

Held,G. *et al*. (2006) Relationship between gene expression and observed intensities in DNA microarrays–a modeling study. *Nucleic Acids Res*., **34**, e70.

Irizarry,R. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Irizarry,R. *et al*. (2006) Comparison of affymetrix genechip expression measures. *Bioinformatics*, **22**, 789–794.

Kane,M. *et al*. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*., **28**, 4552–4557.

Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Lipshutz,R. *et al*. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet*., **21**, 20–24.

Markham,N.R. and Zuker,M. (2005) Dinamelt web server for nucleic acid melting prediction. *Nucreic Acids Res*., **33**, W577–W581.

Matveeva,O. *et al*. (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. *Nucleic Acids Res*., **31**, 4211–4217.

Mei,R. *et al*. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **100**, 11237–11242.

Naef,F. and Magnasco,M. (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.

Naef,F. *et al.* (2002) DNA hybridization to mismatched templates: a chip study. *Phys. Rev. E.*, **65**, 040902.

Nuwaysir,E. *et al.* (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.*, **12**, 1749–1755.

Pozhitkov,A. *et al.* (2007) Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. *Nucleic Acids Res.*, **35**, e70.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SantaLucia,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.

Selinger,D. *et al.* (2000) RNA expression analysis using a 30 base pair resolution escherichia coli genome array. *Nat. Biotechnol.*, **18**, 1262–1268.

Shippy,R. *et al.* (2004) Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics*, **5**, 61.

Singh-Gasson,S. *et al.* (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.*, **17**, 974–978.

Suzuki,S. *et al.* (2007a) Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. *BMC Genomics*, **8**, 373.

Suzuki,S. *et al.* (2007b) Insight into the sequence specificity of a probe on an affymetrix genechip by titration experiments using only one oligonucleotide. *Biophysics*, **3**, 47–56.

Wick,L. *et al.* (2006) On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes. *Nucleic Acids Res.*, **34**, e26.

Wu,C. *et al.* (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res.*, **33**, e84.

Wu,Z. and Irizarry,R. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol.*, **22**, 656–658.

Yuen,T. *et al.* (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.*, **30**, e48.

Zhang,L. *et al.* (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.