# Quantitative expression analysis using oligonucleotide microarrays based on a physico-chemical model

Naoaki Ono
Graduate school of Information Science and Technology, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
nono@ist.osaka-u.ac.jp

Shingo Suzuki
Graduate school of Information Science and Technology, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
shingo@ist.osaka-u.ac.jp

Chikara Furusawa
Graduate school of Information Science and Technology, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
ERATO, JST
furusawa@ist.osaka-u.ac.jp

Hiroshi Shimizu
Graduate school of Information Science and Technology, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
shimizu@ist.osaka-u.ac.jp

Tetsuya Yomo
Graduate school of Information Science and Technology, Graduate School of Frontier Biosciences, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
ERATO, JST
yomo@ist.osaka-u.ac.jp

## ABSTRACT

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 nonlinear behavior of hybridization, nonspecific 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 [7], we presented a benchmark analysis of probe–target hybridization using artificially synthesized oligonucleotides as targets, in which effect of nonspecific 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. The experiments showed that finiteness of both probe and target molecules should be considered to understand detail behavior of hybridization.

In this paper, we present an extension of the Langmuir-model that reproduces the experimental results consistently and the 3-base nearest neighbor model to improve prediction accuracy. We also introduced 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.


*Bionetics* '08, November 25-28, 2008, Hyogo, Japan


## Keywords

microarray, expression analysis

## 1. INTRODUCTION

Recently, analysis of transcriptome, i.e. comprehensive analysis of the all genes that expressed in a cell, has been intensely studied to investigate the behavior of an organism. Detail analysis of expression changes in various environments has been provided us important information to understand how cells response to the environment. These studies revealed complex interaction networks of gene regulatory system. In order to analyze more detail dynamics of the interaction networks and understand the underlying principles that controls the behavior of organisms, more quantitative and high-throughput measurement of gene expression levels are required [4].

DNA microarrays have become one of the most popular tool for transcriptome 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 in a single experiment. However, it is known that linearity of the measurement is maintained within a rather narrow range of concentration, about 2–3 orders of magnitude [3], because of a lower limit of fluorescence measurement and saturation of probe–target hybridization.

In [11] we presented spike-in experiments without back-

ground, namely, in which only artificially synthesized oligonucleotides were hybridized onto a custom designed microarray as a dilution series. In this paper, we introduce an extended thermodynamic model of hybridization in order to improve accuracy and sensitivity of the prediction and the dynamic range of measurements.

# 2. METHODS

## 2.1 Finite Hybridization Model

In this study we use the Finite Hybridization (FH) model which was introduced in [7].

In this model, we consider equilibrium of probe-target duplex formation,

$$\mathrm{P}^{\mathrm{free}} + \mathrm{T}^{\mathrm{free}} \quad \stackrel{K^{\mathrm{sp}}}{\Longleftrightarrow} \quad \mathrm{PT}^{\mathrm{sp}}, \tag{1}$$

where $\mathrm{P}^{\mathrm{free}}$, $\mathrm{T}^{\mathrm{free}}$ are free probe and target molecules and $\mathrm{PT}^{\mathrm{sp}}$ is their duplex, $K^{\mathrm{sp}}$ gives the equilibrium constant of gene-specific hybridization between them.

We considered the effect of nonspecific target, that is, random DNA segments included in the solution as chemical background. To explain the effect of nonspecific target, we introduced two effects: nonspecific hybridization and bulk hybridization. We also took into account other physical effects such as folding of probes [7].

If one assumes that the system reaches equilibrium, and takes mass conservation of probe and target molecules into account, the intensity expected by the FH model is given as follows:

$$I = C(d^{\mathrm{sp}}\left[\mathrm{PT}^{\mathrm{sp}}\right] + d^{\mathrm{ns}}\left[\mathrm{PT}^{\mathrm{ns}}\right]) + I^{\mathrm{bg}}, \tag{2}$$

$$\left[\mathrm{PT}^{\mathrm{sp}}\right] = \frac{1}{2}\Big\{\frac{1}{K^{\mathrm{eff}}} + A + x - \sqrt{(\frac{1}{K^{\mathrm{eff}}} + A + x)^2 - 4Ax}\Big\} \tag{3}$$

$$\left[\mathrm{PT}^{\mathrm{ns}}\right] = \frac{(A - \left[\mathrm{PT}^{\mathrm{sp}}\right])K^{\mathrm{ns}}N}{1 + K^{\mathrm{fold}} + K^{\mathrm{ns}}N} \tag{4}$$

$$K^{\mathrm{eff}} = \frac{K^{\mathrm{sp}}}{(1 + K^{\mathrm{fold}} + K^{\mathrm{ns}}N)(1 + K^{\mathrm{bulk}}N)}. \tag{5}$$

where $C$ is the scale of intensity, $d^{\mathrm{sp}}$ and $d^{\mathrm{ns}}$ denote the dissociation coefficients for specific and nonspecific targets, $I^{\mathrm{bg}}$ is the optical background intensity. $A$, $x$ and $N$ give the total concentration of probes, target, and nonspecific target molecules. $K^{\mathrm{sp}}$,$K^{\mathrm{ns}}$,$K^{\mathrm{fold}}$ and $K^{\mathrm{bulk}}$ are reaction coefficients estimated from free energy change of specific hybridization, nonspecific hybridization, secondary structure and bulk-hybridization, respectively. $d^{\mathrm{sp}}$ and $d^{\mathrm{ns}}$ represent dissociation of probe-target duplex during wash process after hybridization.

## 2.2 The 3-base Nearest Neighbor Model

The Nearest Neighbor (NN) model [9] has been used for estimate hybridization energy from nucleotide sequence. Given the base sequence of the probe provided by $\mathbf{b} = (b_1, \ldots, b_l)$, the hybridization free energy is given as follows:

$$\Delta G^{\mathrm{sp}}(\mathbf{b}) = \sum_{k=1}^{l-1} \epsilon^{\mathrm{sp}}(b_k, b_{k+1}), \tag{6}$$

where $l$ denotes the probe length, $\epsilon^{\mathrm{sp}}(b_k, b_{k+1})$ denote the binding and stacking energy of the given base pairs.

In this study, we introduced the model considering triplets of the bases instead of combination of nearest two bases. Namely, hybridization energy of a probe is estimated as follows:

$$\Delta G^{\mathrm{sp}}(\mathbf{b}) = \sum_{k=1}^{l-2} \epsilon^{\mathrm{sp}}(b_k, b_{k+1}, b_{k+2}). \tag{7}$$

The free energy change of nonspecific hybridization is also estimated by the same formula, using another set of parameters $\epsilon^{\mathrm{ns}}$

## 2.3 Parameter Optimization

This model includes 71 adjustable parameters to fit observed data. We optimized these model parameters by minimizing the mean residual error $R$ between the predicted and observed probe intensity:

$$R = \sum_{i,j} (\log_{10} I_{ij}^{\mathrm{obs}} - \log_{10} I_{ij}^{\mathrm{pre}})^2 / M, \tag{8}$$

where $I_{ij}^{\mathrm{obs}}$ and $I_{ij}^{\mathrm{pre}}$ are the observed and predicted probe intensities of the $i$-th probe in $j$-th experiments, respectively, and $M$ is the number of data points. In this study, $M$ = 37800 data points—5400 probes in seven experiments—were used for the analysis.

Due to the complex interaction among probe and target molecules, it is difficult to calculate the each physical parameters from observed data directly. Instead, we used random sampling and greedy optimization method based on the Monte-Carlo simulation to estimate the parameter values. In short, the initial model parameters are randomly modified in each step, then the values which prediction error is the smallest are selected. Repeat this process until the prediction error no longer decreases.

We designed artificial random sequences as control target oligonucleotides and designed a custom microarray whose probes were complementary to the control targets. Using this microarray, we evaluated our model using experimental data from a spike-in experiment. First we observed the probe intensity in without chemical background. Namely, only artificially synthesized oligonucleotides were hybridized onto a custom designed microarray as a dilution series. Next, in order to evaluate the model in a more realistic condition, cDNA sample obtained from the transcriptome of *Escherichia coli* were mixed with the control oligonucleotide as chemical background.

## 2.4 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 [10],[6]. 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 [11] for detail). The extracted microarray data were analyzed using custom-designed scripts in R software [8]. 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. RESULTS

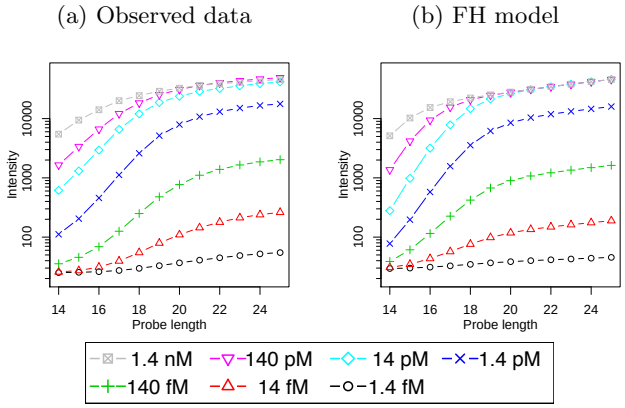


**Figure 1: Average signal intensity as a function of probe length. The model reproduced the saturation behavior in each target concentration level.**

First, we evaluated the FH model by the experiments without background where true signal intensity can be observed. We optimized the model parameters 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 nonspecific hybridization were negligible in this series of experiments.

Remember that we have arranged different lengths of probes for each target. As Eq. 7 implies that $\Delta G^{\mathrm{sp}}$ is roughly proportional to the length of the probe, we first focused on the dependency of probe intensity on probe length. Figure 1(a) shows the results of experiments at seven target concentration levels. Each line represents the average intensity of 450 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. The FH model reproduced the behavior of the observed data over the whole range of target concentrations (Fig. 1(b)).

Next, we evaluated the model under more realistic conditions. In this experiment, the spike-in control oligonucleotides were mixed with cDNA transfered 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 [11].

The scatter plot of observed and predicted intensity is shown in Fig. 2. Even though the background RNA much

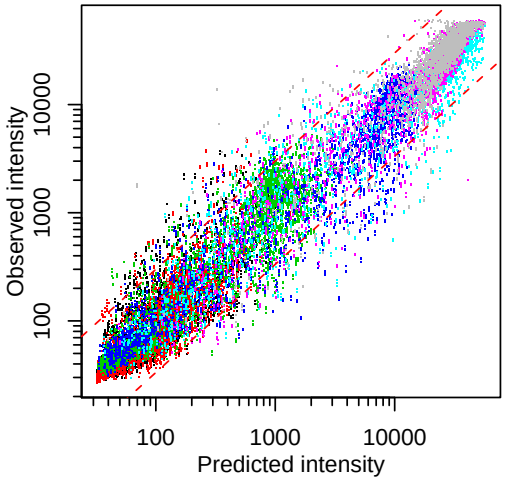


**Figure 2: Comparison between observed and predicted probe intensity. Data points observed in the all experiments with 7 different target concentrations were plotted. Colors of the dots corresponds to concentration levels shown in the Fig. 1. Dashed lines represent $y = 3x$ and $y = 1/3x$. 95% of the data were within this range.**

affected on signal intensity as chemical noise, the model reproduced the behavior of hybridization in all concentration levels.

To confirm that the result was not due to over parameterization, we compared the prediction of the 3-base NN model with the 2-base NN model by 10-fold cross validation method. Namely, we evaluated the prediction for the test data sets using the parameters optimized by the training data from which the test data were excluded. The average prediction error of the 3-base NN model was $5.4 \times 10^2$ while that of the 2-base NN model was $6.6 \times 10^2$. Significance of the difference was confirmed by t-test ($p < 10^{-4}$).

Finally, 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 Eq. 8 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. We evaluated accuracy of this method using 10 sets of randomly chosen 15 probes and the observed data of the spike-in experiments under the condition with the background. We compared the accuracy with the estimation based on the same scheme of 2-base NN model using the same data sets. The results are shown in Fig. 3. The mean squared error of the estimation was significantly decreased. And the results shows that estimation using the 3-base NN model is valid over 5 orders of magnitude.

## 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

(a) 2-base NN model (b) 3-base NN model

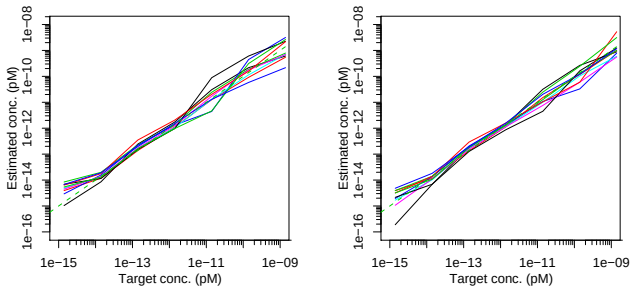


**Figure 3: Estimation of target concentrations. Estimated target concentrations were plotted as a function of nominal concentrations. Each lines represents a probe set which contains randomly chosen 15 probes. Dashed line represents $y = x$. The mean squared error of estimation base on 2-base and 3-base NN model were $10.9 \times 10^{-2}$ and $6.6 \times 10^{-2}$, respectively.**

identified the source of the errors in previous hybridization models and have introduced an improved thermodynamic model.

We presented that the 3-base NN model that estimates hybridization energy for each base considering the bases on the both side of the sequence. The results showed that the model predicts the behavior of hybridization intensity better than the 2-base NN model. 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 [12], mismatch [2, 5], 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, which dynamic ranges are around 2-3 orders [1, 12]. 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 1000 to 0.01 mRNA molecules per cell quantitatively. It implies that the method can measure a wide range of genes from an enzyme that are abundantly expressed to a regulatory gene that controls upstream of signal transduction. Development of analysis based on this method will greatly improve quantitative analyzes of gene expression levels using microarrays.

## 5. ACKNOWLEDGMENTS

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.

## 6. REFERENCES

[1] Affymetrix. *New Statistical Algorithms for Monitoring Gene Expression on GeneChip Probe Arrays*, 2001. Technical Note.

[2] H. Binder, S. Preibisch, and T. Kirsten. Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *Langmuir*, 21(20):9287–302, 2005.

[3] E. Chudin, R. Walker, A. Kosaka, S. Wu, D. Rabert, T. K. Chang, and D. E. Kreder. Assessment of the relationship between signal intensities and transcript concentration for affymetrix genechip arrays. *Genome. Biol.*, 3(1):RESEARCH0005, 2002.

[4] E. Gelenbe. Steady-state solution of probabilistic gene regulatory networks. *Physical Review E*, 76:031903, 2007.

[5] F. Naef, D. Lim, N. Patil, and M. Magnasco. Dna hybridization to mismatched templates: a chip study. *Phys. Rev. E*, 65(4 Pt 1):040902, 2002.

[6] E. Nuwaysir et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.*, 12(11):1749–55, 2002.

[7] N. Ono, S. Suzuki, C. Furusawa, T. Agata, A. Kashiwagi, H. Shimizu, and T. Yomo. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. *Bioinformatics*, 24:1278–1285, 2008.

[8] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[9] A. SantaLucia, S. HT, and P.A. Improved nearest-neighbor parameters for predicting dna duplex stability. *Biochemistry*, 35(11):3555–62, 1996.

[10] S. Singh-Gasson, R. Green, Y. Yue, C. Nelson, F. Blattner, M. Sussman, and F. Cerrina. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.*, 17(10):974–8, 1999.

[11] S. Suzuki, N. Ono, C. Furusawa, A. Kashiwagi, and T. Yomo. Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. *BMC Genomics*, 8(1):373, 2007.

[12] L. Zhang, M. Miles, and K. Aldape. A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, 21(7):818–21, 2003.