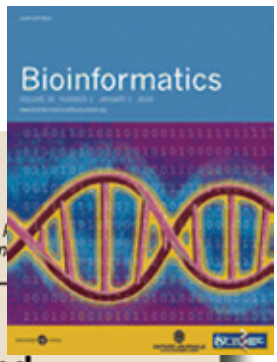
Model-based Analysis of Non-specific Binding for Background Correction of High-density Oligonucleotide Microarrays

Paper in journals : this is the first page of a paper published in ***Bioinformatics***.

[***Bioinformatics***] **25**, 36-41 (2009)



ORIGINAL PAPER

Vol. 25 no. 1 2009,
doi:10.1093/bioinform

Gene expression

Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays

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

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

Received on June 17, 2008; revised on September 16, 2008; accepted on October 30, 2008

Advance Access publication October 31, 2008

Associate Editor: Trey Ideker

ABSTRACT

Motivation: High-density DNA microarrays provide us with useful tools for analyzing DNA and RNA comprehensively. However, the background signal caused by the non-specific binding (NSB) between probe and target makes it difficult to obtain accurate measurements. To remove the background signal, there is a set of background probes on Affymetrix Exon arrays to represent the amount of non-specific signals, and an accurate estimation of non-specific signals using these background probes is desirable for improvement of microarray analyses.

Results: We developed a thermodynamic model of NSB on short nucleotide microarrays in which the NSBs are modeled by duplex formation of probes and multiple hypothetical targets. We fitted the observed signal intensities of the background probes with those expected by the model to obtain the model parameters. As a result, we found that the presented model can improve the accuracy of prediction of non-specific signals in comparison with previously proposed methods. This result will provide a useful method to correct for the background signal in oligonucleotide microarray analysis.

Availability: The software is implemented in the R language and can be downloaded from our website (http://www-shimizu.ist.osaka-u.ac.jp/shimizu_lab/MSNS/).

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

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

1 INTRODUCTION

High-density oligonucleotide microarrays such as those provided by Affymetrix allow the genome-wide quantitative analysis of gene expression, genetic variations and regulatory factor binding sites using the ChIP-chip method (Buck and Lieb, 2003; Lipshutz *et al.*, 1999; Selinger *et al.*, 2000). To improve the quality of data analysis measured by the microarrays, various methods have been studied (Cope *et al.*, 2004; Held *et al.*, 2006; Irizarry *et al.*, 2006; Li and Wong, 2001; Ono *et al.*, 2008; Wu and Irizarry, 2004). A key issue in such microarray technology using short oligonucleotide probes is how to remove the effect of the false signal due to non-specific binding (NSB) between probe and target, which is inevitable when a complex mixture of DNA/RNA fragments are hybridized to millions of probes simultaneously (Shippy *et al.*, 2004; Wu *et al.*, 2005; Yuen *et al.*, 2002; Zhang *et al.*, 2003). The current approach to solving this problem adopted in Affymetrix's platform [Microarray Analysis suite ver. 5.0 (MAS5)] is to use a set of probe pairs, i.e. a perfect match (PM) probe which matches a fragment of the corresponding gene exactly and a mismatch (MM) probe containing a single nucleotide MM in the center (Affymetrix, 2001). It is assumed that the signal intensities of the MM probes provide a measure of NSB to the corresponding PM probes, and thus the use of the signal intensities of MM probes allows one to remove the effect of NSB. However, this method has the following two problems. First, it has been pointed out that around 30% of raw MM intensities are larger than corresponding raw PM intensities, and to extract information of target hybridization from such probe pairs is difficult. This fact indicates that the MM intensities are not always available for the compensation of NSB in PM probes (Naef *et al.*, 2002). Second, this method requires a huge number of MM probes, as many as there are PM probes, for the background correction. Since the number of available PM probes is an important factor in the accuracy of microarray analysis, especially in the tiling-array analysis (Bertone *et al.*, 2004), a method of signal correction using a smaller number of MM probes is desirable.

Recently, Affymetrix released a new platform, namely Exon arrays, designed for high resolution analysis of exon-level expression. One significant feature of the Exon arrays is that they have no MM probes. Instead, they include a set of probes, called background probes, for which it is expected that there is no significant gene-specific signal caused by exactly matched targets, so the observed signal intensities of the background probes mostly originate from NSB (Affymetrix, 2005). Thus, by analyzing how the signal intensities of the background probes depend on their probe sequences, it is expected that we can estimate how the NSB signal contributes to signal intensities of all probes. In Affymetrix's algorithm for the estimation of NSB, called the GC composition-based background correction (PM-GCBG), the non-specific signal intensity of a given PM probe is estimated as the median of

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

▲From ***Bioinformatics,*** 25, Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays, 36-41, 2009. Reprinted with permission from Oxford University Press.

The following is a comment on the published paper shown on the preceding page.

Model-based Analysis of Non-specific Binding for Background Correction of High-density Oligonucleotide Microarrays

FURUSAWA Chikara and ONO Naoaki

(Graduate School of Information Science and Technology)

Introduction

High-density oligonucleotide microarrays such as those provided by Affymetrix allow genome-wide quantitative analysis of gene expression, genetic variations, and regulatory factor binding sites using the ChIP-chip method [1,2]. A key issue in such microarray technology using short oligonucleotide probes is how to remove the effect of the false signal due to non-specific binding (NSB) between probe on the arrays and fluorescent labeled targets, which is inevitable when a complex mixture of DNA/RNA fragments is hybridized to millions of probes simultaneously. Thus, the development and the evaluation of algorithms to predict the amount of NSB are important topics for the improvement of microarray analysis.

In this study, we sought to develop a thermodynamic model of NSB on short nucleotide microarrays. In the model, we assumed that NSB caused by a complex mixture of DNA/RNA fragments can be approximated by hybridization of the probes and multiple hypothetical targets, which reflect major components of the mixture. We also assumed the binding affinities of these hypothetical targets to the probes can be estimated by the nearest neighbor model [3] from the sequence of each probe. In the original nearest neighbor model, the model parameters depends on permutation of every adjacent base pairs, we extended it to consider permutation of every adjacent n-base pairs. We fitted the observed signal intensities of NSB with those expected by the model to obtain the model parameters as representing the concentrations of the hypothetical targets and the binding affinities. As a result, we found that our model improved the accuracy of prediction of NSB intensities compared with previously proposed approaches.

A thermodynamic model for the non-specific hybridization signals

To estimate the NSB signal we used the data of "background probes" on Affymetrix's Exon Arrays, for which it is expected that there is no significant gene-specific signal caused by exactly matched targets, so the observed signal intensities of the background probes mostly originate from NSB [4]. Thus, by analyzing how the signal intensities of the background probes depend on their probe sequences, it is expected that we can estimate how the NSB signal contributes to signal intensities of all probes.

In this study, we introduce a multi-source non-specific hybridization (MSNS) model to estimate the NSB signal intensity of background probes. The schematic representation of this model is shown in Fig.1. In this approach, we assume that the non-specific signals of background probes can be represented by a thermodynamic equilibrium model of the bindings between probes and multiple hypothetical targets, based on the following reactions:

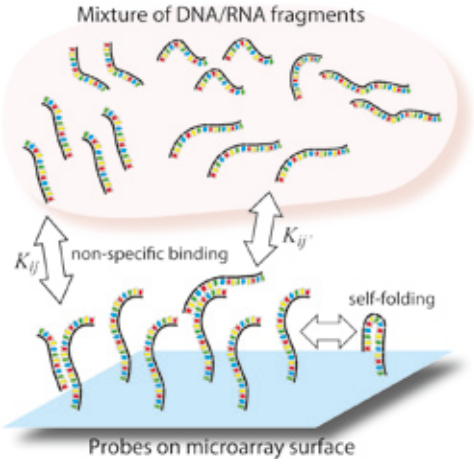


Fig. 1 Schematic representation of the multi-source non-specific hybridization (MSNS) model

$$P_i^{free} + T_j^{free} \xleftrightarrow{K_{ij}} P_i T_j \qquad j = 1,2,\cdots,m$$

where P_i^{free} and T_j^{free} are free i-th probes and j-th hypothetical targets, and P_iT_j is their duplex, K_{ij} is equilibrium constant between them and m is the number of hypothetical targets. Here we assume the following points; i) the equilibrium the system, ii) mass conservation of probe and target molecules, and iii) the number of target molecules is enough larger than the number of probe molecules. From these assumptions, we obtain that the intensity of non-specific hybridization of the i-th probe is represented as follows:

$$I_i^{NS} = C\sum_{j=1}^{m}\left[P_i T_j\right] + I^{bg} = C\frac{\left[P_i^{total}\right]\sum_j K_{ij}\left[T_j^{free}\right]}{1+\sum_j K_{ij}\left[T_j^{free}\right]} + I^{bg}$$

where I_i^{NS} is the non-specific signal intensity of the i-th probe, C is the scale of intensity and I^{bg} is the optical background intensity. For the equilibrium constant K_{ij}, we assume that the free energy of hybridization between probe and hypothetical non-specific targets is calculated using the n-nearest neighbor model, which is an expansion of the nearest neighbor model [3] to include the effect of n-neighboring nucleotides for the calculation. We also consider the equilibrium of the probes between the folded and unfolded states, where the equilibrium constant of folding/unfolding of probes are calculated by an algorithm named UNAfold [5]. In the MSNS model, there are

$m\times4^n+m-n+28$ parameters that are adjusted to fit the model of the observed background intensities: $m\times4^n$ parameters for the n-nearest neighbor parameter for each hypothetical target; m parameters for the concentration of m hypothetical targets; $26-n$ for the position dependence of the weight factors on the probes; one parameter for the optical background constant; and one for the weight factor representing the coefficient for probe folding. We optimized these model parameters by minimizing the mean residual error between observed and expected probe intensity in the background probes.

Evaluating estimates of the non-specific hybridization signals

To evaluate the MSNS model for the estimation of the NSB signal, we fitted the observed non-specific signal intensities in the training probe set with those expected by the models by tuning the parameters in the models. Then, we evaluated the accuracy of the estimate. In Fig.2, we show the scatter plots of estimated and observed signal intensities of the testing probe set which were not used for the parameter fitting. In the figure, the estimations of two previous methods are also presented. One is Affymetrix's GC composition based background correction (PM-GCBG) method [4], in which the NSB signal intensity of a given PM probe is estimated as the median of signal intensities of background probes having the same GC-content. Another is the model-based analysis of tiling arrays (MAT) method [6], in which a simple linear model is used for the NSB estimation. As clearly shown in Fig.2, the PM-GCBG and MAT methods showed inconsistency in the region of large background signals (i.e. observed intensities larger than 500) and relatively low R^2 value, while the MSNS model succeeded in estimating such large background signals and resulted high R^2 value (R^2~0.8). One reason for the difference in the estimation of large background signals is the use of the n-nearest neighbor model to calculate the hybridization free energy. The parameters in the n-nearest neighbor model indicated that contiguous sequences of cytosine(C) in a probe, such as 'CCCC' and 'CGCCC', are more effective in increasing the amount of nonspecific hybridization than non-contiguous sequences of C, even among probes having the same GC content. Since such effects of contiguous sequences in a probe cannot be represented in the previous models, these models failed to predict the large signal intensity caused by NSB. Also, the results indicate that the inclusion of multiple hypothetical targets is effective for the accurate prediction of the amount of NSB. Since the MSNS model used a large number of fitting parameters for the estimation (e.g., 1052 fitting parameters are used in the case of , m=4 , n=4), we checked the possibility of over-parameterization by using Akaike information criterion (AIC) and Bayesian information criterion (BIC). As results, the analysis using AIC and BIC suggested that the models with $m\approx4$ and $n\approx4$ are appropriate for the non-specific estimation under this condition.

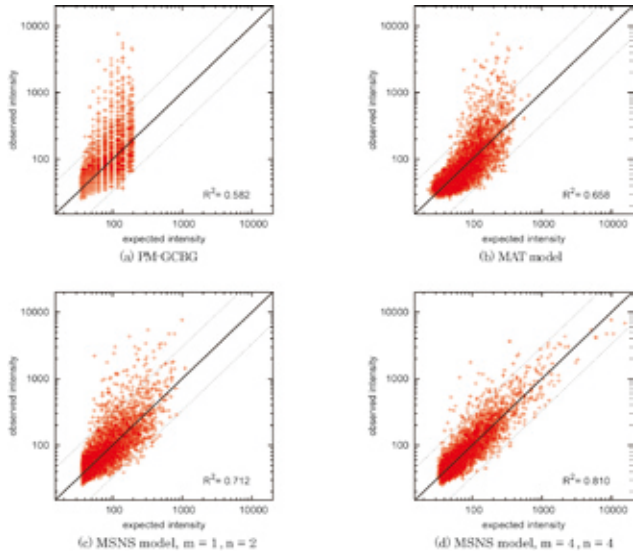


Fig. 2 The relationship between the expected and observed signal intensities of the testing probes. The R^2 values for each estimation are also presented. The solid line indicates y=x, while the dashed lines show y=3×x and y=1/3×x, respectively. The numbers of fitting parameters are 25 for (a), 80 for (b), 43 for (c), and 1052 for (d), respectively. In the case of the MSNS model with m=4, n=4, 99% of probes in the testing dataset were within 3-fold differences, and 96% were within 2-fold differences.

Conclusion

One significant feature of microarray technology is that we can prepare a huge number of probes, and thus can use huge amounts of signal intensity data for analyses. Such a huge amount of data makes it possible to evaluate hybridization models with large numbers of fitting parameters. In this study, we evaluated the model-based approach for the estimation of non-specific hybridization. We expanded the hybridization model to represent the NSB between probes and a complex mixture of oligonucleotide fragments by introducing multiple hypothetical targets and the n-nearest neighbor model for the estimation of binding affinities. Even though the number of fitting parameters in the model becomes larger, we found that the accuracy of predicting the NSB signal intensities increases significantly. Our studies showed that the inclusion of 10^4~10^5 background probes, for which no signal caused by specific binding is expected, provides accurate prediction of the NSB signal, with R^2~0.8. We believe that the accurate background correction with such a small number of background probes can be a key algorithm for future microarray analysis.

References

[1] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ, "High density synthetic oligonucleotide arrays", *Nat. Genet.*, **21**, 20–24 (1999)

[2] Selinger D. *et al.* "RNA expression analysis using a 30 base pair resolution Escherichia Coli genome array", *Nat. Biotechnol.*, **18**, 1262–1268 (2000)

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

[4] Affymetrix, "White Paper: Exon Array Background Correction", available at http://www.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf (2005)

[5] Markham NR. and Zuker M, "Dinamelt web server for nucleic acid melting prediction", *Nucleic Acids Res.*, **33**, W577–W581. (2005)

[6] Johson WE *et al.* "Model-based analysis of tiling-arrays for ChIP-chip", *Proc. Natl Acad. Sci.*, **103**, 12457–12462 (2006)