Nucleic Acids Research, 2005, Vol. 33, No. 9 ***e84***
doi:10.1093/nar/gni082

Sequence dependence of cross-hybridization on short oligo microarrays

Chunlei Wu, Roberto Carta[1] and Li Zhang*

Department of Biostatistics and Applied Mathematics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd-447, Houston, TX 77030, USA and [1]Department of Statistic and Actuarial Sciences, University of Central Florida, Orlando, FL 32816–2370, USA

Received October 23, 2004; Revised February 12, 2005; Accepted May 2, 2005

ABSTRACT

One of the critical problems in the short oligo microarray technology is how to deal with cross-hybridization that produces spurious data. Little is known about the details of cross-hybridization effect at molecular level. Here, we report a free energy analysis of cross-hybridization on short oligo microarrays using data from a spike-in study. Our analysis revealed that cross-hybridization on the arrays is mostly caused by oligo fragments with a run of 10–16 nt complementary to the probes. Mismatches were estimated to be energetically much more costly in cross-hybridization than that in gene-specific hybridization, implying that the sources of cross-hybridization must be very different between a PM–MM probe pair. Consequently, it is unreliable to use MM probe signal to track cross-hybridizing signal on a corresponding PM probe. Our results also showed that the oligo fragments tend to bind to the 5′ ends of the probes, and are rarely seen at the 3′ ends. These results are useful for microarray design and data analysis.

INTRODUCTION

Microarray technology has become a powerful tool for genomic-scale studies of gene expression (1). One of the popular platforms of this technology, as exemplified by Affymetrix GeneChip, uses 25mer short DNA oligonucleotides as probes to hybridize to biotinylated RNA molecules to measure gene expression (2). Because hybridization to probe oligonucleotides at such length is known to have limited specificity, a key issue is how to avoid getting spurious signals from cross-hybridization. Current approach (2) to this problem is to use multiple probe pairs, which is referred as a 'probeset', to target a single gene; one of each pair exactly matches a fragment of the gene (PM probe) and the other contains a single mismatching nucleotide in the center (MM probe). The contrast between the signals from the probe pairs is used to reduce the effect of cross-hybridization. The rationale behind this design is that an MM probe, relative to the PM probe, should have much less gene-specific signal due to the mismatch, but the same amount of cross-hybridization signal. However, a number of studies (3–8) have noted that the rationale might be flawed. For genes expressed at high levels, the ratio between signals in a probe pair is close to one, indicating that MM signal contains gene-specific signals very much like the PM signal; for genes expressed at low levels, PM and MM probe signals do not approach to the same level, which results in negative PM–MM signals.

We recently developed a simple physical model of hybridization interaction on short oligonucleotide arrays (9) that partially explains the mechanism of the observed behavior of probe signals. The model assumes that the observed probe signals come from two idealized sources: gene-specific binding (GSB) and nonspecific binding (NSB). GSB refers to the formation of DNA/RNA duplexes with exact complementary sequences at the length of 25 bp. NSB refers to the formation of duplexes with many mismatches between the probe and the attached RNA molecule. The number of duplexes with few mismatches should be rare because the probes are pre-selected to avoid this type of binding (2,10). The free energy of GSB is formulated as a weighted sum of stacking energies of nearest-neighbors of base pairs. The weights depend on the position of the pairs along the DNA/RNA duplex. We called our model the positional dependent nearest-neighbor model (PDNN). The free energy of NSB is formulated similarly, under the assumption that NSB on a probe depends on the probe sequence only, and it is independent of gene expression. The source of NSB is approximated by a random mixture of all possible short oligonucleotides. It was shown that the model is able to explain most of the variations of probe signals in a probe-set and an algorithm based on this model was designed for estimation of gene expression from the probe signals (9).

However, this model of NSB is clearly too simplistic to account for all of the cross-hybridization signals.

*To whom correspondence should be addressed. Tel: +1 713 563 4298; Fax: +1 713 563 4243; Email: lzhangli@mdanderson.org

The model is unable to identify changes in NSB signals from different RNA samples hybridized on different chips, because NSB signals are assumed to be independent of any gene's expression. To advance the microarray technology, a better understanding of the cross-hybridization effect is necessary. Here, we report a detailed analysis of cross-hybridization effects on the microarrays. We used a spike-in dataset provided by Affymetrix, which was generated using a Latin Square design that allowed an easy identification of probes with cross-hybridization signals. Our aim was to identify potential sources of cross-hybridization, where they bind on the probes, and how much they bind depending on their sequences.

MATERIALS AND METHODS

Microarray data

The dataset used in this analysis was downloaded from Affymetrix website (http://www.affymetrix.com/support/datasets.affx). The dataset included 42 HG-U133A array images. The experiments were designed to follow a Latin Square with 14 spike-in gene groups and 14 concentrations. All spike-ins were made by T7 RNA polymerase to generate anti-sense transcripts and subsequently mixed into a cRNA sample isolated from total RNA in HeLa cell line (ATCC CLL-13). Each spike-in group had three different RNA transcripts and at each concentration, the experiment was replicated three times. The 14 concentrations were 0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256 and 512 in pM.

The RNA sequences of the spike-ins were also obtained from Affymetrix website. We noted that four of the spike-ins (for AFFX-DapX-3_at, AFFX-LysX-3_at, AFFX-PheX-3_at and AFFX-ThrX-3_at) were sense sequences instead of supposed anti-sense sequences. Because the signals of these probe-sets behaved as expected, we concluded that the experiments were conducted correctly but the sequences were provided in the wrong strand. We hence corrected these sequences.

We found that there may have been some contamination in the spike-ins. For example, all of the probe signals in probesets 204890_s_at and 204891_s_at displayed large signal variations consistent with those of the group 1 spike-ins. Probesets 204890_s_at and 204891_s_at both are targeted to a gene called lymphocyte-specific protein tyrosine kinase (LCK). However, we found no significant alignments between LCK and any of the spike-ins. Thus, the sources of these probe signals could not be identified. We thought that it was not likely that these were caused by cross-hybridization, because cross-hybridization typically affects individual probes rather than the whole probesets. We postulated that these probesets were affected by contamination, probably due to PCR artifacts during preparation of the spike-ins. Similar contamination problems were uncovered (11) in an older spike-in experiment (12). Since the probes in probesets 204890_s_at and 204891_s_at might be affected by contamination, we did not count them as probes affected by cross-hybridization to the spike-ins. There were two additional probesets, 213060_s_at and 203173_s_at, which also seemed to react to group 1 spike-ins but bear no sequence similarity to the spike-ins. These problematic probesets were excluded from further analysis.

Normalization

We used the quantile normalization method (13) to normalize probe intensities in order to ensure that the distribution of probe intensity was the same for each array. The normalization procedure involves a non-linear transformation. But because the probe intensity distributions were nearly identical for all of the arrays, the transformations were essentially linear in every case.

Sequence alignment by BLAST

BLAST (14) algorithm was used to align the probe sequences to the spike-in transcripts. To identify the probes that are expected to respond to the spike-ins due to GSB, we collected all perfectly matched alignments with an alignment length equal to 25. To identify the probes that may respond to the spike-ins due to cross-hybridization, we collected all the alignments with an alignment length >7 (allowing no mismatches and no gaps).

Free energy modeling of cross-hybridization

Linear regression was used to quantify the binding affinity of cross-hybridizing probes. A cross-hybridization signal caused by a fragment on a certain spike-in transcript is assumed to be linearly dependent on the nominal concentration of the spike-in. The regression slope represents the binding affinity of the spike-in fragment binding on the probe. We assumed that the logarithm of the binding affinity could be taken as the binding free energy according to Boltzmann distribution at thermal equilibrium, and modeled the free energy as a weighted sum of nearest-neighbor stacking energies:

$$G = \sum \omega(k) * \varepsilon(b_k, b_{k+1}), \quad \mathbf{1}$$

where $\omega(k)$ is a weight factor that depends on the position on the probe sequence, $\varepsilon(b_k, b_{k+1})$ is the stacking energy of a pair of base pairs adjacent to each other along the probe at positions k and $k + 1$ with b_k and b_{k+1} as their nucleotide types, respectively. The weight factor is expressed as a fifth-order polynomial function of k. The summation covers the range of the aligned fragment between the probe sequence and the spike-in sequence. All of the parameters involved in Equation 1 were treated as unknown.

Because a probe may bind multiple fragments excised from a single spike-in transcript, contributions from multiple alignments were summed up as follows:

$$\hat{A} = \sum \exp(-G_i/k_B T), \quad \mathbf{2}$$

where $\hat{A}$ is the model-expected binding affinity of a probe, and G_is are the values of binding free energy for each alignment, T is the temperature and k_B, the Boltzmann constant.

A least-square-fit method was used to optimize the fit between $\hat{A}$ values and the binding affinities computed from linear regression using microarray data.

RESULTS

Identification of probes affected by spike-ins

To identify the probes that were affected by the spike-ins, we searched for probes that displayed reproducible variation

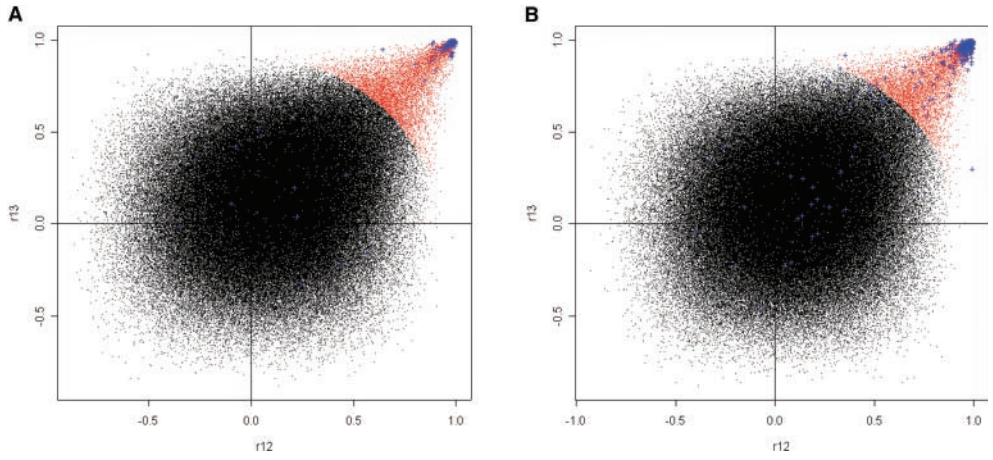

Figure 1. Scatter plot of signal correlation for each probe between the three replicated experiments. (**A**) PM probes; (**B**) MM probes. *x*-axis is the Pearson's correlation of PM probe signals between experiment batches 1 and 2 (r12); *y*-axis is the Pearson's correlation of PM probe signals between experiment batches 1 and 3 (r13). Red-dot spots are marked as the probes satisfying these criteria: $r12 > 0$, $r13 > 0$ and $r12^2 + r13^2 > 0.81$. These probes are identified as the probes with reproducible signals across three replicated experiments. Blue-cross spots are marked as the probes targeted to spike-in transcripts.

amongst the batches of replicated experiments. Figure 1 shows correlation of probe signals between the three batches of replicated experiments. Each r12 value in the figure represents a Pearson's correlation for a probe calculated between observed PM signals on the probe in batch 1 and those in batch 2. Similarly, an r13 value is calculated from PM probe signals in batches 1 and 3. The data points in Figure 1 form a circular pattern except those at the upper right corner. Since the circular pattern is expected if probe signal variations are purely random, we assumed that the probes in the upper right corner were affected by the spike-ins, due to either cross-hybridization or gene-specific hybridization. We found 8020 PM probes and 7145 MM probes, located outside the circle with a radius of 0.9 (marked as red dots in Figure 1). These probes constitute 3% of all the probes in the array.

To reduce the false positives in these affected probes, we excluded the probes that displayed little signal variation across the 42 samples. Figure 2 shows the distribution of SD of log-transformed probe signal for PM probes. The major peak in the distribution represents the probes with random noise only, and the minor peak with larger SD represents the probes affected by the spike-ins. Among the affected probes identified in Figure 1, there are 1281 PM probes (1307 MM probes) with SD > 0.3. It is these probes that are hereafter assumed to be the probes affected by the spike-ins.

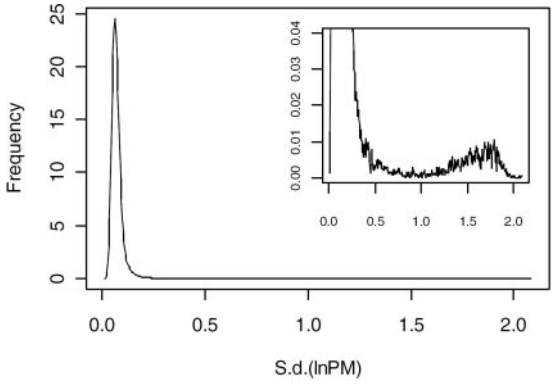

Figure 2. Distribution of SD of log-transformed probe signals for PM probes. The distribution shows a major peak and a minor peak, which represent contributions from random noise and the effects of the spike-ins, respectively. Inset shows the minor peak that is only visible after zoom-in.

Identifying the sources of cross-hybridization

To study the sequence dependence of cross-hybridization, we needed a collection of probes for which the sources of cross-hybridization signals can be clearly identified. We used multiple criteria to screen for such probes. First, we searched for spike-ins that may be the responsible sources for these 1281 probes. For each probe and a spike-in transcript, we computed the Pearson's correlation between the observed probe signals on the 42 array images and the known concentrations of the spike-in. The highest correlation among all of the spike-ins for a specific probe (r_{max}) was taken to be the responsible source for the probe. Among the 1281 affected PM probes, 97% of the probes showed $r_{max} > 0.9$, which was expected if one of the spike-ins was indeed the responsible source. Note that if a probe was affected by multiple spike-ins, then r_{max} may be low, making it difficult to trace the sources of cross-hybridization. To avoid such complications, we consequently selected 1248 PM probes (1285 MM probes) that had $r_{max} > 0.9$ for further analysis.

Then, we used BLAST algorithm to align sequences of these probes with the corresponding spike-ins to identify the cross-hybridizing fragments. A PM probe sequence that could be perfectly matched to a 25 nt fragment of a spike-in transcript was assumed to be affected by the spike-in transcript due to gene-specific hybridization: 807 PM probes were identified as such probes. The remaining 441 PM were assumed to be produced by cross-hybridization.

To identify fragments of the spike-ins that bound probes due to cross-hybridization, we collected the BLAST alignments with alignment-length >7 to be candidates of cross-hybridizing fragments. For simplicity, we also left out the probes whose optimum BLAST alignment scores to the spike-ins are not unique. This should help to reduce the cases in which single probes were affected by multiple spike-ins. Through the

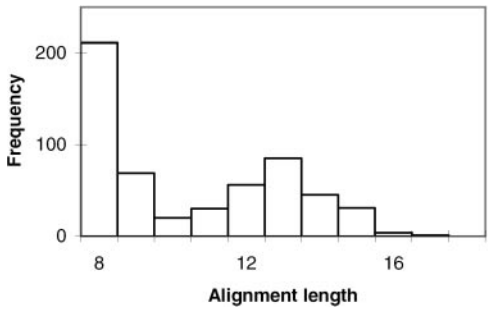

Figure 3. Histogram of sizes of cross-hybridizing fragments. The fragments were collected from BLAST alignments between spike-ins and the PM probes that were affected by cross-hybridization to the spike-ins.

analysis above, we obtained 287 PM probes that have identifiable cross-hybridization sources and the sequences of the cross-hybridizing fragments.

Figure 3 shows the histogram of the fragment sizes collected from the 287 PM probes. The histogram shows a peak ~13, a minimum at 10, and it rapidly rises for smaller sizes. We found that aligned fragments with <8 are so prevalent that they can be found between any probe and any spike-in transcript. Because shorter alignments are in general less likely to cause binding, it seems reasonable to expect that cross-hybridization generally would happen to alignment sizes between 10 and 16. Longer alignment than 16 are rare simply because the probes had been pre-selected in array design to avoid cross-hybridization.

Free energy modeling of cross-hybridization

Using the binding affinities along with the sequences of cross-hybridizing fragments, we fitted a free energy model (see Materials and Methods). Our model assumes that the cross-hybridization signal respond linearly to the spike-in concentrations. This may seem to be a crude approximation since it was noted that the response curve is non-linear and it can be better characterized by Langmuir isotherm (15). However, we noted that although Langmuir saturation is apparent for GSB, it does not happen to cross-hybridization, presumably due to the fact that the magnitude of cross-hybridization is much smaller. As shown in Figure 4, cross-hybridizing probes reacted linearly to the spike-ins.

The optimized energy parameters are shown in Figure 5. The optimized ln ($\hat{A}$) values correlated well with the observed ln(A) values for the data collected from the 287 PM probes (Pearson's correlation between the values of ln($\hat{A}$) and ln(A) in Figure 5C is 0.7). The weight factor displays a distinct pattern: there is a sharp drop on the right side, which is at the 3′ end of the probes. This pattern indicates that there is a 5′ end bias for cross-hybridization effects.

To further examine this bias, we mapped the locations of the cross-hybridizing fragments on the probes. We collected short oligo fragments that displayed cross-hybridization effects on multiple probes (Figure 6). As expected, we found that the fragments closer to the 5′ ends tend to incur stronger cross-hybridization affinities.

The 5′ end bias was also observed when our PDNN model was used to fit the PM probe signals on an entire array. Figure 7B shows that the weight factors for NSB are higher on the left-hand side of the figure. We also observed such

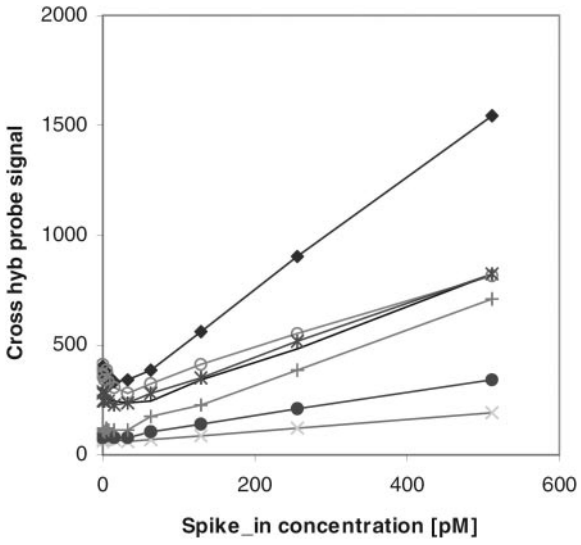

Figure 4. Linearity of cross-hybridization signals. Seven probes that were found to cross-hybridize to group 1 spike-ins are included here. Each line represents a probe. The *x*-axis shows the spike-in concentration. The *y*-axis shows the averaged probe signals over three replicated experiments. This figure shows that the cross-hybridization signals respond linearly to source transcripts.

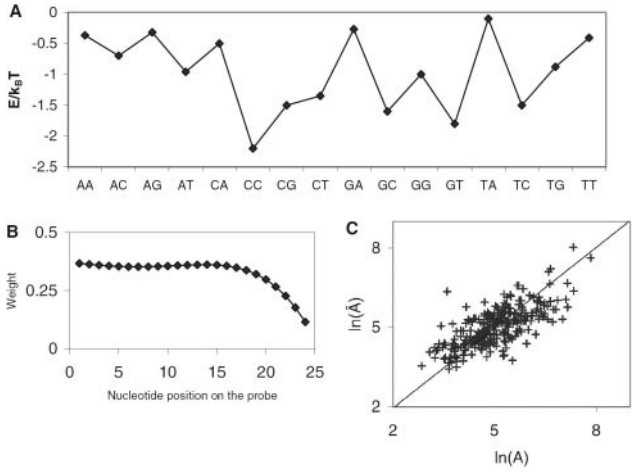

Figure 5. Free-energy model fitting. (**A**) Nearest-neighbor stacking energy. (**B**) Weight factors for cross hybridization. (**C**) Model fitting. The model fitted binding affinities ($\hat{A}$) are plotted against the observed affinities (A) on a logarithmic scale.

pattern on other samples from our own experiments using the HG-U133A chips. Thus, the 5′ end bias seemed to be general on such type of arrays.

Discordant behavior of PM and MM probes in probe pairs

PM–MM probe pairs are designed to track the effects of cross-hybridization. Thus, it seems important to find out how the probe pairs react to the spike-ins when there is cross-hybridization. Figure 8 shows the SD of log-transformed probe signals for all the probe pairs in the dataset. Other than the probes targeted to the spike-ins (shown in red), most of other probe pairs can be found near either the *x*-axis or the *y*-axis. This result showed that either a PM or an MM probe

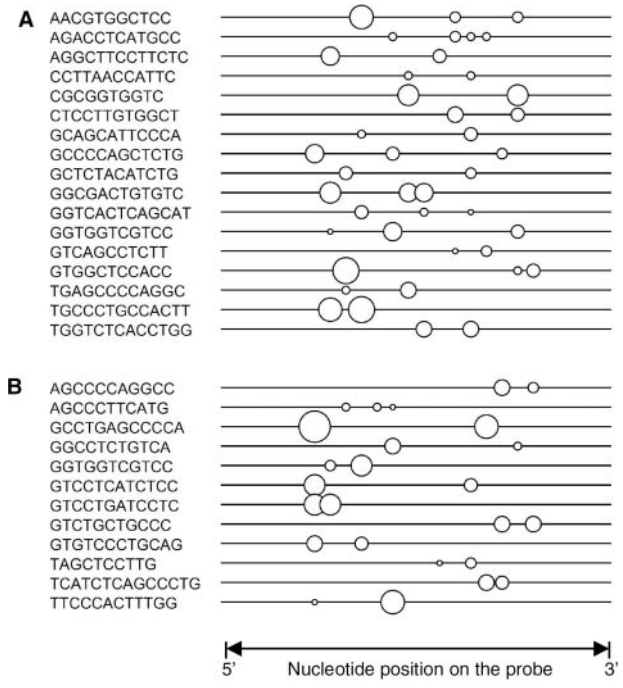

Figure 6. Location of cross-hybridizing fragments on the probes. Each row represents a cross-hybridizing fragment from the spike-ins. For each oligo-fragment, the line represents probe (5′ end at the left); the center of a circle indicates the mid-point of oligo-fragment on the probe; and the area of the circle represents the magnitude of cross hybridization binding affinity. With a few exceptions, the oligo-fragments appear to bind stronger at the 5′ ends of the probes. A cross-hybridization oligo-fragment is shown here only if it was found to cross-hybridize to multiple probes. (**A**) Fragments collected from PM probes. (**B**) Fragments collected from MM probes.

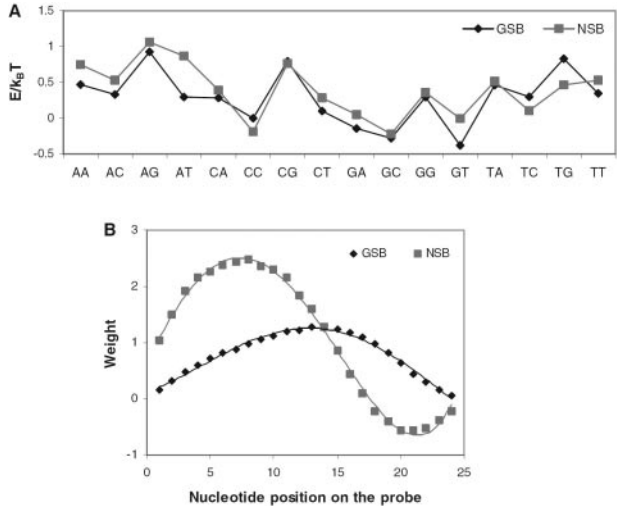

Figure 7. Stacking energies and weight factors from PDNN model. (**A**) Stacking energies of NSB and GSB; (**B**) Weight factors of GSB and NSB. The nucleotide positions were counted from the 5′ end of the probes. The array type used here was human genome HG-U133A manufactured by Affymetrix Inc. The weight factors for NSB are higher on the left-hand side of the figure, showing a bias to the 5′ end.

could be affected by cross-hybridization, but not simultaneously.

How could a single mismatch have such a drastic effect? To evaluate the free energy cost of a single nucleotide

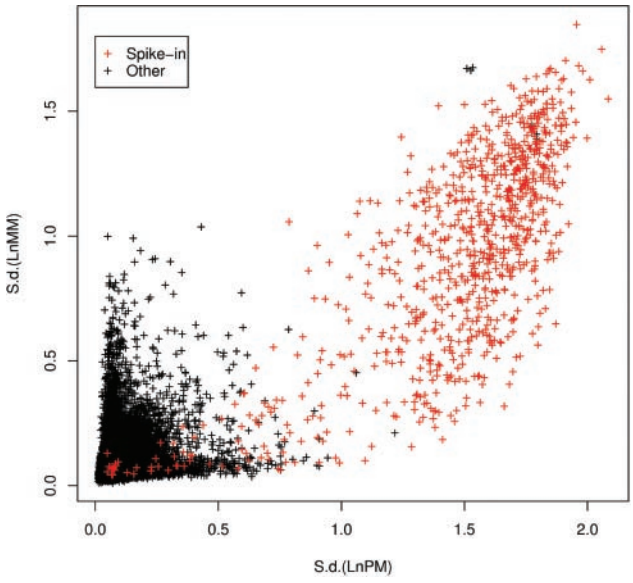

Figure 8. Scatter plot of SD of log-transformed probe signal of probe pairs. *x*-axis is for PM probes and *y*-axis for MM probes. Each point represents a probe pair. Red cross represents the probes targeting spike-ins. Except for the red crosses, most of the others are close to either *x*-axis or *y*-axis, showing very discordant behavior.

Table 1. Free energy cost of a single nucleotide mismatch (in k_BT units) in cross-hybridization

Mismatch	Match	ΔG(Mismatch) − ΔG(Match) (k_BT)
dT•rU	dA•rU	2.2 ± 0.9
dG•rG	dC•rG	3.4 ± 0.8
dC•rC	dG•rC	2.7 ± 0.9
dA•rA	dT•rA	2.9 ± 1.1

The values were obtained from averaging the binding free energies between PM and MM probe pairs that were affected by cross-hybridization to the spike-ins. We only included the cases in which the mismatch position (base 13 on the probe) is not at the ends of the aligned segment between the probe and the spike-in transcript.

mismatch, we first selected from the probes affected by cross-hybridization of which the PM–MM probe pairs reacted to the same spike-ins: 167 such probe pairs were selected. Among them, 97 probe pairs have the mismatch positions located within the aligned fragment. Table 1 shows the averaged differences of binding energies ($\Delta G = \Delta G_{mm} - \Delta G_{pm}$) between probe pairs. When a mismatch is located outside the region bound to the cross-hybridizing fragment, no mismatch is expected. Indeed, we found that the magnitude of ΔG ($0.8 \pm 1.5\ k_BT$) is much smaller in such cases. Similarly, the effect of a mismatch at the either end of a cross-hybridizing fragment should be small, and we found $\Delta G = 1.3 \pm 1.0\ k_BT$.

Using the free energy model to identify probes highly prone to cross-hybridization

The free energy model developed in this work can be used to identify probes highly prone to cross-hybridization. To demonstrate how this can be achieved, we used the free energy parameters derived from the PM signals shown in Figure 5C and applied them to the MM probes. Figure 9A shows the

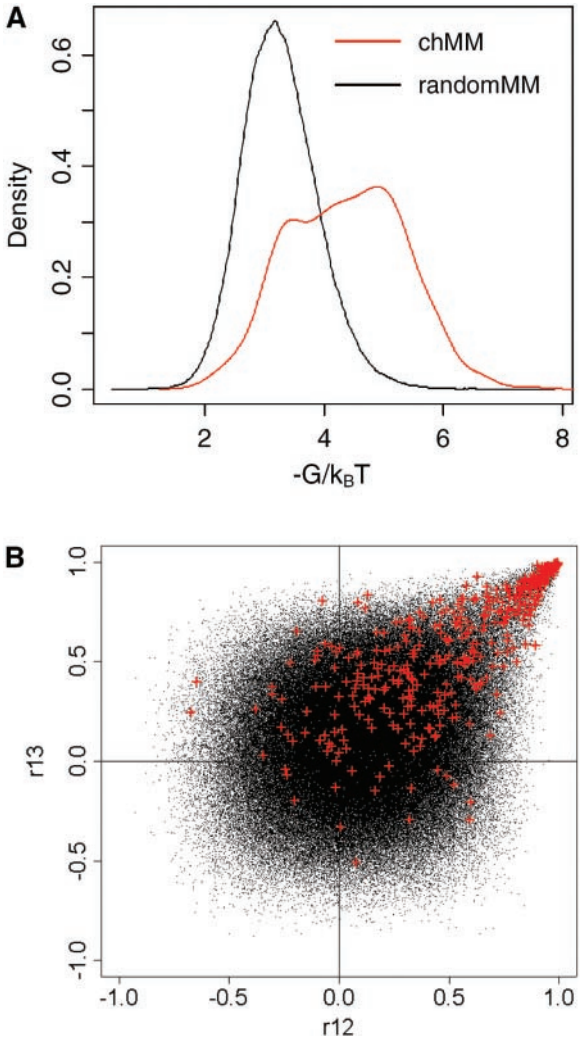

Figure 9. Predicting cross-hybridization effects. (**A**) Distribution of free energy of binding of chMM probes and randomMM probes. The MM probes targeted to the spike-ins were excluded in this figure. chMM represents the MM probes affected by cross-hybridization. They were selected according to the following criteria: r12 > 0, r13 > 0, $r12^2 + r13^2 > 0.81$ and SD of log-transformed probe signal >0.3. randomMM represents the rest of MM probes. The free energy values ($-G/k_BT$ on *x*-axis) were calculated according to the parameters shown in Figure 5 and using BLAST alignments of MM probe sequences with the spike-ins. (**B**) Scatter-plot of r12 versus r13. The data are the same as shown in Figure 1B. The probes shown in red were selected if $-G/k_BT$ >5.5 k_BT.

distribution of free energies for all MM probes that may bind to the spike-ins. We included all MM probes that were aligned to at least one of the spike-ins, but excluded the probes whose corresponding PM probes matched perfectly to the spike-ins. The MM probes showing cross-hybridization effects were identified using the same criteria as used for PM probes: r12 > 0, r13 > 0, $r12^2 + r13^2 > 0.81$, SD of log-transformed probe signal >0.3. The minor peak in Figure 9A represents probes that are prone to cross-hybridization to the spike-ins. Using binding free energy >5.5 (in k_BT units) as a threshold, we predicted 421 MM probes with high risk of cross-hybridization. To check if these probes were affected by cross-hybridization, we looked at the correlations of the probes between batches of replicated experiments (Figure 9B). Most of the predicted probes turned out to be in the upper right corner (shown in red), which means that the variations of the probe signals are reproducible, suggesting that these probes were indeed affected by cross-hybridization as predicted.

DISCUSSION

In this paper, we have developed a free energy model to assess the sequence dependence of cross-hybridization effect on short oligonucleotide arrays. Our analysis revealed that cross-hybridization on the arrays are mostly caused by oligo fragments with a run of 10–16 nt complementary to the probes. Our analysis also revealed that cross-hybridization tends to be biased towards the 5′ ends of probes, which means that cross-hybridizing molecules tend to cling to the tips of the probes, as the 3′ ends of the probes are attached to the microarray surface. This finding may help to refine the algorithms used in probe design (10,16,17). The number of probes highly prone to cross-hybridization may be reduced by using our free energy model. The physical mechanism of this 5′ end bias is unclear, but a possible cause is the interaction with the microarray surface, which has been shown to affect binding on microarray surface (18–20). We also expect that more insights will be obtained using new technologies (21) that are able to directly detect cross-hybridization on microarrays.

We noted that some of the weights in Figure 7B are slightly negative. It is not clear how to interpret these negative values, but one possible explanation is that the partitioning of NSB and GSB in PDNN model is biased (L. Zhang, unpublished data). To avoid confusion, it may be important here to distinguish the cross-hybridization signals evaluated in our linear regression modeling from the NSB signals estimated in PDNN model. The former only concerns contribution from the spike-ins, while the latter concerns contribution from any random transcripts. This distinction may also explain the differences between the stacking energies seen in Figures 5A and 7A.

It is important to point out that our free energy model has not taken into account all the technical factors such as RNA secondary structure that may affect the observed probe signals (22,23). The cross-hybridizing fragments identified from sequence alignments were relatively short, but the actual molecules bound on the array due to cross-hybridization may be much longer. It is not known how the unmatched regions on a cross-hybridizing molecule might influence the binding.

Our model also omitted cross-hybridizations that involve mismatches, but this omission seems justifiable. We observed that a PM–MM probe pair behaves very differently when it comes to cross-hybridization. The high free energy cost of a single mismatch ($\sim 3\ k_BT$) implies that mismatches are generally avoided in cross-hybridization interaction. It is interesting to compare our results of the free energy cost of single mismatch with that identified in aqueous solution. In Table 1, the trend of stabilities of the mismatches is dT•rU > dC•rC ≈ dA•rA > dG•rG. However, for DNA/RNA duplex in solution, the order is dG•rG > dA•rA ≈ dT•rU (24) (dC•rC is very unstable, but quantitative data for dC•rC is not available). For DNA duplex formation in solution, the order is dG•dG > dT•dT ≈ dA•dA > dC•dC (25). For RNA duplex formation in solution, the order is rG•rG > rU•rU > rA•rA

(26). These studies noted that the mismatch cost could vary substantially depending on the neighboring sequences. The striking discrepancy is that G•G is a relatively more stable mismatch than C•C in solution but the reverse is true on microarrays. We think the discrepancy may be related to the fact that C and U nucleotides in the target RNA molecules are biotin labeled. It is also notable that C•C as a stable mismatch has been observed before, and it was explained in terms of the stacking energies (9). Based on our observations, on all GeneChip expression arrays, the contrast between PM and MM probe pairs is the least when the mismatch type is C•C, suggesting that C•C is generally a stable mismatch on microarrays.

The discordant behavior of probe pairs identified in our current work also explains why MM probe have a limited value for assessing cross-hybridization on a PM probe. Because the free energy cost of a single mismatch is so high in cross-hybridization, and given that the mismatch is located at the center of the probes, it would be difficult for an RNA molecule to bind both probes in a probe pair. Consequently, the sources of cross-hybridization signals are different for the probes in probe pair. This means that we cannot expect an MM probe signal to track changes in cross-hybridization in the corresponding PM probe signal, and using PM–MM may not reduce the effect of cross-hybridization.

Furthermore, from the probe pairs targeted to bind the spike-ins, we observed that PM and MM signals behave in nearly perfect concordance as the spike-in concentration changes. It means that in this case PM and MM probes closely track each other when it comes to GSB. Thus, using PM–MM can reduce the gene-specific signal. It is not surprising that a number of studies have found that it is better to ignore the MM signals altogether in gene expression estimation (3–9).

ACKNOWLEDGEMENTS

We thank Kevin R. Coombes, Gary Rosner and Margaret Newell for helpful suggestions on the manuscript. Funding to pay the Open Access publication charges for this article was provided by MD Anderson Cancer Center start-up fund.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S. (1999) Array of hope. *Nature Genet*., **21**, 3–4.
2. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol*., **14**, 1675–1680.
3. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
4. Zhou,Y. and Abagyan,R. (2002) Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics*, **3**, 3.
5. Lemon,W.J., Palatini,J.J., Krahe,R. and Wright,F.A. (2002) Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, **18**, 1470–1476.
6. Chu,T.M., Weir,B. and Wolfinger,R. (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math. Biosci*., **176**, 35–51.
7. Cope,L.M., Irizarry,R.A., Jaffee,H.A., Wu,Z. and Speed,T.P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.
8. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scheerf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
9. Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol*., **21**, 818–821.
10. Mei,R., Hubbell,E., Bekiranov,S., Mittmann,M., Christians,F.C., Shen,M.M., Lu,G., Fang,J., Liu,W.M. and Ryder,T. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **100**, 11237–11242.
11. Hsieh,W.P., Chu,T.M. and Wolfinger,R.D. (2003) *Who are These Strangers in the Latin Square? Methods of Microarray Data Analysis III*. Kluwer Academic Publishers, Boston, MA.
12. Hubbell,E., Liu,W.M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.
13. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
14. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
15. 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.
16. Flikka,K., Yadetie,F., Laegreid,A. and Jonassen,I. (2004) XHM: a system for detection of potential cross hybridizations in DNA microarrays. *BMC Bioinformatics*, **5**, 117.
17. Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
18. Heaton,R.J., Peterson,A.W. and Georgiadis,R.M. (2001) Electrostatic surface plasmon resonance: direct electric field-induced hybridization and denaturation in monolayer nucleic acid films and label-free discrimination of base mismatches. *Proc. Natl Acad. Sci. USA*, **98**, 3701–3704.
19. Jin,R., Wu,G., Li,Z., Mirkin,C.A. and Schatz,G.C. (2003) What controls the melting properties of DNA-linked gold nanoparticle assemblies? *J. Am. Chem. Soc*., **125**, 1643–1654.
20. Vainrub,A. and Pettitt,B.M. (2003) Surface electrostatic effects in oligonucleotide microarrays: control and optimization of binding thermodynamics. *Biopolymers*, **68**, 265–270.
21. Plutowski,U. and Richert,C. (2005) A direct glimpse of cross-hybridization: background-passified microarrays that allow mass-spectrometric detection of captured oligonucleotides. *Angew. Chem. Int. Ed. Engl*., **44**, 621–625.
22. Southern,E., Mir,K. and Shchepinov,M. (1999) Molecular interactions on microarrays. *Nature Genet*., **21,**, 5–9.
23. Mir,K.U. and Southern,E.M. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol*., **17**, 788–792.
24. Sugimoto,N., Nakano,M. and Nakano,S. (2000) Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes. *Biochemistry*, **39**, 11270–11281.
25. Peyret,N., Seneviratne,P.A., Allawi,H.T. and SantaLucia,J.,Jr (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry*, **38**, 3468–3477.
26. Kierzek,R., Burkard,M.E. and Turner,D.H. (1999) Thermodynamics of single mismatches in RNA duplexes. *Biochemistry*, **38**, 14214–14223.