12

ASSESSING THE POTENTIAL EFFECT OF CROSS-HYBRIDIZATION ON OLIGONUCLEOTIDE MICROARRAYS

Seman Kachalo, Zarema Arbieva and Jie Liang

Dept. of Bioengineering and Core Genomic Facility, University of Illinois at Chicago, Chicago, IL.

Abstract: We introduce a computational method which estimates non-specific binding associated with hybridization signal intensities on the oligonucleotide-based Affymetrix GeneChip arrays. We consider a simplified linear hybridization model that should work well when the target DNA concentration is low or when the probe-target affinity is weak, and use the quadratic programming technique to estimate the parameters of this model (binding coefficients). We show that binding coefficients correlate with the degree of homology between the probe and target sequences. Detectable contribution into DNA binding was found to start from the matches of 7-8 nucleotides. The method suggested here may prove useful for the interpretation of hybridization results and for the assessment of true target concentrations in microarray experiments.

Key words: oligonucleotide microarray, cross-hybridization, linear model

1. INTRODUCTION

At the present time, DNA microarray-based comparative expression analysis has become an important tool in a variety of research areas, including cancer research, pharmacogenomics, population studies, etc. Many current microarray platforms utilize alternative probe formats bound to a solid support. This method was introduced based on the observation that single-stranded DNA binds strongly to a nitrocellulose membrane in a way that prevents strands from re-association with each other, but permits hybridization to complementary strands [Gillespie and Spiegelman, 1965]. Regardless of the probe format, all microarray based applications utilize a fundamental property of nucleic acids to re-associate separate strands in solutions in a fashion dependant on salt concentration, strand composition and sequence, as well as the degree of homology.

Hybridization of nucleic acid targets to tethered DNA probes in a multiplex or heterogeneous fashion is the central event in the detection of nucleic acids on microarrays. An immediate problem is associated with the fact that many target single strands are present in the same reaction. If their sequences are so predisposed, these target sequences can anneal with other (target and probe) strands that are not fully complementary, forming partially duplex states that are reasonably stable at assay temperatures. Obviously such "side reactions", or cross-hybridization, lower accuracy and complicate the interpretation of the microarray data. The ability to estimate the input of the cross-hybridization effect may potentially facilitate both more accurate processing of the registered hybridization intensities and more rational probe design as well.

In relation to spotted arrays, a few attempts have been made to approach the cross-hybridization issue in a more specific and quantitative manner. Riccelli *et al.* developed a new analytical method, which provides evidence of the presence of both perfectly matched and heteromorphic duplex states [Riccelli *et al.,* 2002]. The effect of the subtle sequence composition characteristics (one, two or tandem base pair mismatches and also the context surrounding the mismatch) on duplex stability and cross-hybridization propensity is under discussion. It has also been reported that for a given nucleotide probe any "non-target" transcripts >75% similar over the 50 base target may show cross-hybridization, thus contributing to the overall signal intensity. In addition, if the 50 base pair target region is marginally similar, it must not include a stretch of complementary sequence > 15 contiguous bases [Kane *et al.,* 2000].

To address the problem of cross-hybridization on Affymetrix oligonucleotide microarrays, a PM/MM approach was proposed [Affymetrix, 2002]. Each probe pair consists of a perfect match (PM) probe and a mismatch (MM) probe that is identical to the PM oligonucleotide except for a single base substitution in a central position. It is assumed that PM and MM probes are equally affected by cross-hybridization, while the PM probe has a higher level of specific hybridization. By subtracting the MM signal from the PM signal one expects to cancel the terms related to cross-hybridization and obtain a refined specific signal. However, the general assumption of equal effects of cross-hybridization on PM and MM probes is not always correct. As reported in [Naef *et al.,* 2002] about one-third of all probe pairs detect MM>PM. If the above assumption is true, these probe pairs should indicate negative gene expression.

Additional complexity is introduced by the fact that specific hybridization levels depend on the sequence of the probe. It was shown in [Li and Wong, 2001] that most individual probes are less variable between arrays than different probes within the same probe set on the same array.

This study was designed to investigate the effect of a nucleotide sequence on hybridization and the contribution of low-homologous DNA sequences into cross-hybridization.

2. DATA

We used the Human portion of the Affymetrix Latin Square dataset [Affymetrix, 2001], which can be found on Affymetrix corporate website at http://www.affymetrix.com/analysis/download or on the CAMDA website at http://www.camda.duke.edu/camda02. This dataset contains signal intensities for a total of 409,600 probes on Affymetrix HG-U95A microarray chips in 59 experiments. Experiments are divided into two groups of twenty and one group of nineteen experiments.

In each experiment fourteen labeled DNA targets with known concentrations were spiked into labeled complex targets and hybridized to the arrays. Two of fourteen targets (transcripts corresponding to the probe sets 37777_at and 407_at) are at equal concentrations in each experiment; therefore, there are only 13 distinct targets of varying concentrations in the dataset. The composition of the complex target is not specified, however, it was identical within each of the three groups of experiments. We introduced three additional variables to represent these complex targets. As the actual concentrations of complex targets had no special meaning in our study, each of these variables was assigned the value of one in one group of experiments and zero in the other two groups.

Oligonucleotide probe sequences and target definitions for HG-U95A microarray chip can be found at Affymetrix corporate website. Complete cDNA sequences for the spiked targets can be retrieved from GenBank database (http://www.ncbi.nlm.nih.gov).

3. MODELS

DNA binding to oligonucleotide probes on a microarray is a dynamic process [Tibanyenda *et al.* , 1984; Ikuta *et al.* , 1987; Wang *et al.* , 1995; Vernier *et al.* , 1996; Persson *et al.* , 1997]. The rate R_+ of DNA molecules associating with the spot is proportional to the concentration of DNA x and to the number N_{unocc} of unoccupied oligonucleotides on the microarray spot:

$$R_{+} = k_{+}xN_{unocc}. \tag{1}$$

The rate R_{-} of DNA dissociating is proportional to the amount of DNA bound to the spot or to the number N_{occ} of occupied oligonucleotides:

$$R_{-} = k_{-}N_{occ}. \tag{2}$$

Here, k_{+} and k_{-} are the coefficients of proportionality that can depend on DNA structure, oligonucleotide sequence and many other factors. The total number of oligonucleotides per spot $N = N_{unocc} + N_{occ}$ does not change.

When equilibrium is achieved, the rates of DNA associating and dissociating become equal, i.e.:

$$N_{occ} = kxN_{unocc}, \tag{3}$$

where $k = k_{+} / k_{-}$, or, after making all substitutions,

$$N_{occ} = \frac{kxN}{1+kx} \tag{4}$$

Because the probe signal intensity is proportional to the amount of DNA molecules bound to the probe, the same relation can be applied for the probe signal intensity y:

$$y = \frac{kxy_{sat}}{1+kx}, \tag{5}$$

where y_{sat} is the probe intensity in saturated state when all probe oligonucleotide molecules are bound to DNA. The dependency of signal intensity on DNA concentration is hyperbolic. However, when $kx << 1$ (i.e. when the probe signal intensity is low), it can be approximated by the linear function:

$$y = bx, \tag{6}$$

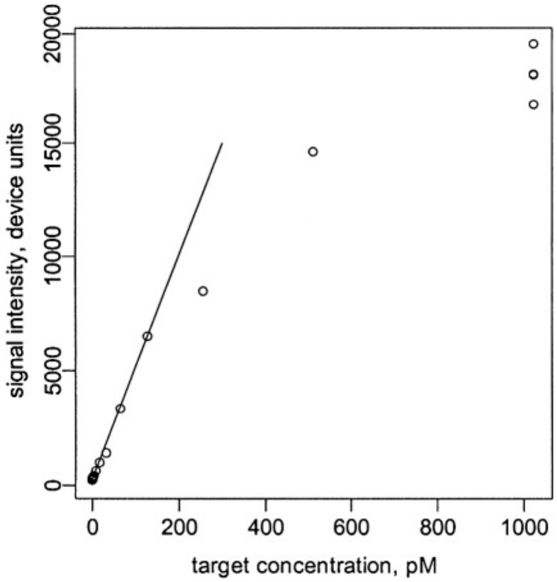

Figure 1. **Dependency of probe signal intensity from target concentration. DNA transcript 684_at; probe [517:489]; first group of experiments. Probe [517:489] is specific to the transcript 684_at. The dependency can be approximated by linear function for low target concentrations.**

where we define $b = ky_{sat}$ as the binding coefficient. The binding coefficient is closely related with the probe affinity effect discussed in [Li and Wong, 2001] and is equal to the logarithm of affinity effect defined in [Irizarry *et al.,* 2003].

The experimental dependency of probe signal intensity from DNA concentration is illustrated on Figure 1.

The assumption of linearity allows us to develop a linear binding model for simultaneous binding of many different DNA targets to many different probes in a series of experiments:

$$y_{ik} = \sum_j b_{ij} x_{jk} + \varepsilon_{ik} , \tag{7}$$

where $y_{ik} \geq 0$ is the signal intensity for the i -th probe in the k -th experiment, $x_{jk} \geq 0$ is molar concentration of the j -th target in the k -th experiment, $b_{ij} \geq 0$ is the binding coefficient for the j -th target and the i -th probe, and ε_{ik} is random noise.

For further comparison, we also used a random binding model that assumes that the probe signal intensities are random and independent of target molar concentrations:

$$y_{ik} = \overline{y_i} + \hat{\varepsilon}_{ik}, \qquad (8)$$

where $\overline{y_i}$ is the mean signal intensity of the i -th probe in the whole set of experiments.

4. EXPERIMENTAL BINDING COEFFICIENTS

By dropping index i from (7), for each probe we can write:

$$y_k = \sum_j b_j x_{jk} + \varepsilon_k . \qquad (9)$$

Provided that target concentrations x and probe signal intensities y are known for the set of experiments, binding coefficients b_j can be found as the solutions of the classical quadratic programming problem [Boot, 1964]:

minimize $\sum \varepsilon_k{}^2$ in (9),

subject to: $b_j \geq 0$. (10)

The program for solving the problem (10) was implemented as a combination of C++ and Matlab code. For each of 409,600 probes the program was used to calculate 16 binding coefficients (for thirteen known targets and three complex targets) from 59 data points.

The obtained binding coefficients were substituted in (9) to calculate the minimized error $\sum \varepsilon_k{}^2$, which was compared with the minimized error $\sum \hat{\varepsilon}_k{}^2$ of the random binding model (8).

As seen on Figure 2, the minimized error of the linear model is smaller than the minimized error of the random model; however, the difference is less than one order of magnitude. This can be explained by the high level of noise as well as by the nonlinearity of signal from many probes due to high probe signal intensity.

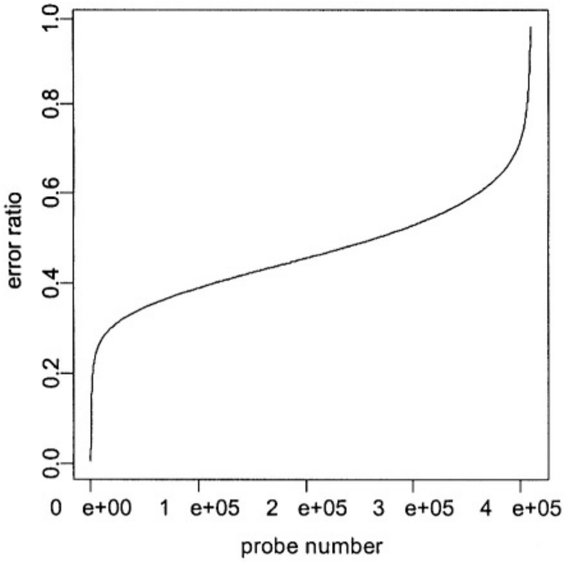

Figure 2. Sorted error ratios $\sum \varepsilon_k^{\,2} / \sum \hat{\varepsilon}_k^{\,2}$ calculated for 409,600 probes.

For further study, a subset of 304 probes was selected for which we expected binding coefficients to be found with best accuracy. First, from the complete set, there were a few-hundred probes chosen for which the quadratic programming problem (10) solution gave the best optimization:

$$\sum \varepsilon_k^{\,2} / \sum \hat{\varepsilon}_k^{\,2} \le 1/10 .$$

Next, the probes specific to, or having high similarities to the thirteen known targets were excluded from the analysis. Because of high target concentrations in the experiments, these probes were expected to demonstrate nonlinear concentration-intensity dependency.

The obtained results reveal the existence of a relationship between the binding coefficient and the degree of homology of the probe with the target nucleotide sequences. As shown on Figure 3, the correlation between the binding coefficient and the length of the longest common substring is over 60%. An almost identical relationship is observed when using the Smith-Waterman [Smith and Waterman, 1981] alignment score with various parameters instead of a common substring length.

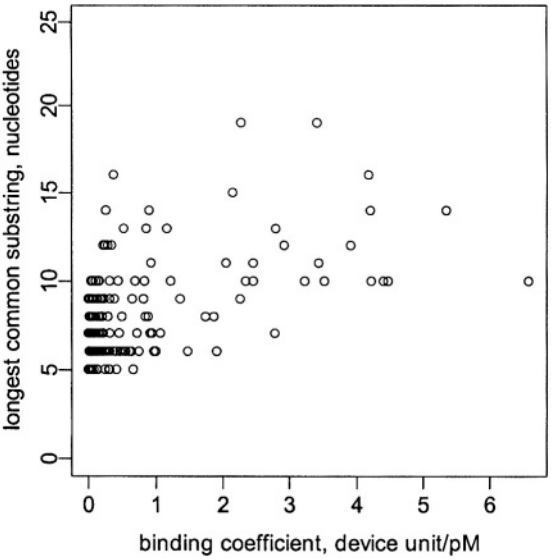

Figure 3. **Binding coefficients depend on the degree of sequence similarity. Binding coefficients and longest common substring lengths for the 304 top probes and transcript 684_at are 61% correlated.**

5. ESTIMATED BINDING COEFFICIENTS

As suggested by the above results, even a modest similarity may result in cross-hybridization. It is natural to think that DNA binds to the probe not only at the site of the best match, but also at the sites of weaker matches. To model this situation, many kinds of binding patterns can be introduced as multiple non-overlapping areas of similarity between the probe and target sequences that together contribute to the binding coefficient:

$$b = \sum_a n_a c_a + \varepsilon , \tag{11}$$

where b is the binding coefficient between any fixed probe and target, n_a - number of matches of type a found between these probe and target sequences, c_a - contribution of each pattern of type a into the binding coefficient and ε - error (not to be confused with errors in equations 8 and 9).

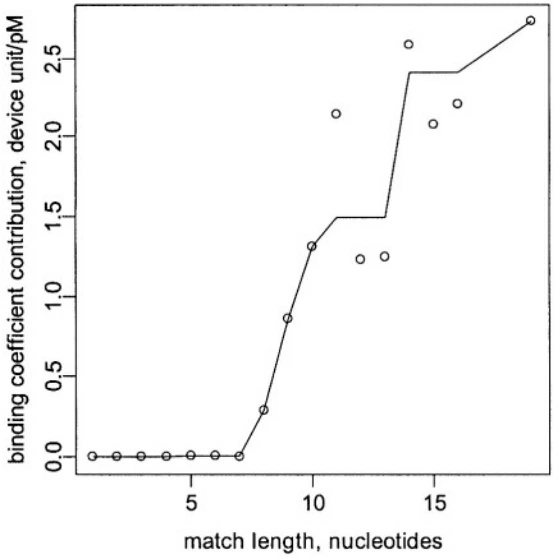

Figure 4. **Contributions of perfect matches of different length into the binding coefficient, calculated for transcript 684_at and the top 304 probes. Dots are the solution of problem (13) with no additional condition; the solid line is the solution of the same problem with the additional condition (14).**

Once the set of binding patterns is defined, it's easy to calculate the number of each pattern occurrence within the sequences of probe and target. If the binding coefficients are known for a number of probe-target pairs, the contribution of each binding pattern can be found by methods of quadratic programming similar to those applied for solving problem (10).

The simplest example of binding patterns can be a set of non-overlapping substrings of different lengths that are common in the probe and target sequences. Since the length of all probes on the HG-U95A microarray is 25 nucleotides, there are only 25 types of binding patterns in the set. If the binding coefficients are known for some set of probes and targets, equation (11) can now be rewritten as:

$$b_{ij} = \sum_{l} n_{ijl} c_l + \varepsilon_{ij}, \qquad (12)$$

where b_{ij} is the binding coefficient for the i -th probe and the j -th target, n_{ijl} - number of matches of length l found between these probe and target sequences, c_l - contribution of each match of length l into the binding coefficient and ε_{ij} - random noise. The optimization problem to find the match contribution in this case will be:

$$\text{minimize } \sum \varepsilon_{ij}^{2} \text{ in (12),}$$

$$\text{subject to: } c_l \geq 0\,, \tag{13}$$

$$\text{additional condition: } c_{l+1} \geq c_l\,. \tag{14}$$

We used experimental values of binding coefficients for 304 probes, selected above to calculate the contributions of matches of various lengths to DNA binding. For each probe-target pair, a histogram was built for the number of non-overlapping common substrings of one to twenty five nucleotides in length. Following that, the optimization problem (13) was solved with and without additional conditions (14) using Matlab code. The problem was solved for the complete set of thirteen targets and for each target separately, revealing very similar results. Figure 4 shows the perfect match contributions obtained for one of the targets with and without additional conditions (14). A slight disagreement between these two solutions for matches longer than 10 nucleotides can be explained by the relative rarity of long matches and high level of noise, caused by that fact.

As seen from the figure, matches of length eight or greater contribute significantly to cross-hybridization. Though it's not easily apparent on the plot, contributions to cross-hybridization from matches of length seven are also detectable.

One could expect faster growth of the match contribution function with an increase in match length. Slow growth of this function for longer matches is due to the fact that probes with high similarities to targets have high signal intensities through the experiments. Because of possible non-linearity their binding coefficients may be underestimated.

Calculated match contributions were substituted back into (12) to obtain estimated binding coefficients that were then compared with experimental binding coefficients obtained in the previous section. Figure 5 illustrates the results of this comparison. The method based on the use of binding patterns performs better than the method based on match scores. We expect that this method can be further improved by using a more diverse set of binding patterns rather than the set of matches of different length. This will require, however, a larger set of experimental data.

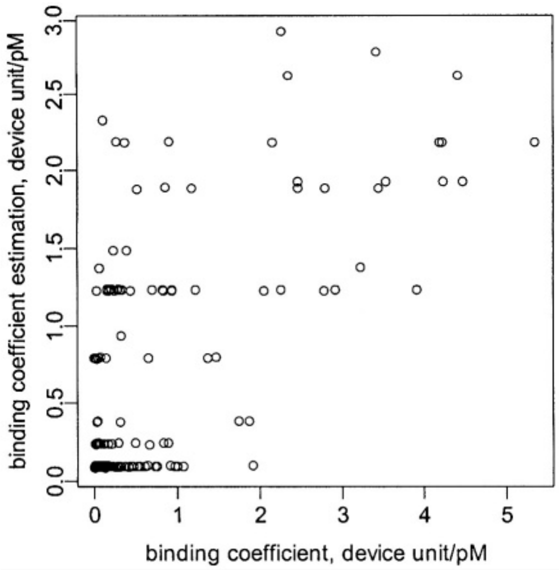

Figure 5. **Experimental and estimated binding coefficients. Estimated binding coefficients for the top 304 probes and transcript 684_at are 71% correlated with experimental binding coefficients.**

6. DISCUSSION

Our results demonstrate that cross-hybridization can contribute significantly to the hybridization signal, potentially introducing substantial error. By rough estimation, in the case of randomly uniformly distributed nucleotides, for any DNA transcript of 500 nucleotides in length there is about a 50% chance of a 7-nucleotide match with any 25-nucleotide probe. This suggests that any transcript, which is present in high abundance in the hybridization mixture, can affect the signal intensity for half of the probes on the microarray. In seven of nineteen possible cases, a 7-nucleotite match will cover the central nucleotide of 25-nucleotide probe. Thus, cross-hybridization differentially affects PM signal and its corresponding MM signal. This ratio is even worse for longer matches that are not as frequent as 7-nucleotide matches, but produce much stronger contribution into the signal. As reported in [Naef *et al.,* 2002], MM>PM for about one-third of all probe pairs. The only explanation of this fact is strong cross-hybridization. Though most PM/MM-based algorithms [Affymetrix, 2002;

Li and Wong, 2001] ignore such pairs, as well as other outliers, there is no guarantee that the remaining probe pairs are free from significant cross-hybridization.

A new successful algorithm not based on the PM/MM principle was recently suggested in [Irizarry *et al.,* 2003]. The model used in this algorithm can be written as

$$T(PM_{ij}) = e_i + a_j + \varepsilon_{ij}, \tag{15}$$

where T represents the transformation that corrects background, normalizes and logs the PM intensities; e_i represents the $\log_2$ scale expression value found on array i; a_j represents the log scale affinity effects for probe j; and ε_{ij} corresponds to relative error of j -th probe on the i -th array. There is an obvious tight relation between the models (15) and (7). Both models assume a linear dependency of probe signal from target concentration and imply that probe affinity to the target may be different for different probes. However, the possible effects of cross-hybridization are ignored in the model (15).

The main benefit of using the linear binding model suggested here is the opportunity to eliminate the impact of cross-hybridization. Once the binding coefficients are determined either by experiment or theoretically, finding target concentrations in (7) from the known probe signal intensities becomes a trivial linear algebra problem that can be effectively solved computationally.

The main limitation of the linear model is the fact that hybridizations should be performed at lower target concentrations than those commonly used in microarray experiments, which may result in higher relative noise level. However, the linear model (15) was shown to outperform PM/MM-based methods, probably because the concentration of spike DNA in the datasets used was about 10 times lower than in the Affymetrix Latin Square dataset.

To adopt an experiment with high target concentration, a non-linear model with more than one parameter for each probe-target pair should be applied. Its disadvantage compared to a linear model is that calculation of target concentrations from the signal intensities can be a difficult mathematical problem requiring substantially longer computational time.

ACKNOWLEDGMENTS

We would like to thank Seby Edassery and Peter Larsen from the Core Genomics Facility at UIC for their generous help during all stages of our work. We are also grateful to the members of CAMDA'02 Organizing Committee for giving us the inspiration to begin this study and for the very interesting datasets provided. This work is supported in part by a grant from The Whitaker Foundation.

REFERENCES

Affymetrix, Inc. (2001) New statistical algorithms for monitoring gene expression on GeneChip probe arrays. Technical report.

Affymetrix, Inc. (2002) Statistical algorithms description document. Technical paper.

Boot J. C. G. (1964) Quadratic Programming. North-Holland

Gillespie D., Spiegelman S. (1965) A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. J. Mol. Biol. 12(3):829-42

Ikuta S., Takagi K., Wallace R.B., Itakura K. (1987) Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs. Nucleic Acids Res. 26;15(2):797-811

Irizarry R., Bolstad B., Collin F., Cope L., Hobbs B., Speed T. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31:e15

Kane M., Jatkoe TA., Stumpf C., Lu J., Thomas J., Madore S. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucl. Acids Res. 2000 Nov 15;28(22):4552-7.

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

Naef F., Lim D., Patil N., Magnasco M. (2002) DNA hybridization to mismatched templates: a chip study. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 65, 040902

Persson B., Stenhag K., Nilsson P., Larsson A., Uhlen M., Nygren P. (1997) Analysis of oligonucleotide probe affinities using surface plasmon resonance: a means for mutational scanning. Anal. Biochem. 246(1):34-44

Riccelli P., Hall T., Pancoska P, Mandell K., Benight A. (2003) DNA sequence context and multiplex hybridization reactions: melting studies of heteromorphic duplex DNA complexes. J. Am. Chem. Soc. Jan 8;125(1):141-50

Smith T.F. and Waterman M.S (1981) Identification of common molecular subsequences. J. Mol. Biol. 147:195-197.

Tibanyenda N., De Bruin S.H., Haasnoot C.A., van der Marel G.A., van Boom J.H., Hilbers C.W. (1984) The effect of single base-pair mismatches on the duplex stability of d(T-A-T-T-A-A-T-A-T-C-A-A-G-T-T-G). d(C-A-A-C-T-T-G-A-T-A-T-T-A-A-T-A). Eur. J. Biochem. 15;139(1):19-27

Vernier P., Mastrippolito R., Helin C., Bendali M., Mallet J., Tricoire H. (1996) Radioimager quantification of oligonucleotide hybridization with DNA immobilized on transfer membrane: application to the identification of related sequences. Anal. Biochem. 235(1):11-9

Wang S., Friedman A.E., Kool E.T. (1995) Origins of high sequence selectivity: a stopped-flow kinetics study of DNA/RNA hybridization by duplex- and triplex-forming oligonucleotides. Biochemistry 34(30):9774-84