

ASSESSING 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

Abstract: We introduce computational method, which allows estimating the input of non-specific binding into hybridization signal intensities on the oligonucleotide-based Affymetrix GeneChip arrays.

We consider a simplified linear model of hybridization that should work well for microarray experiments with low DNA concentrations during hybridization, and use the quadratic programming technique to estimate the parameters of this model (binding coefficients).

We show that binding coefficients estimated based on our model depend on the degree of homology between the target and the probe. Detectable contribution into DNA binding starts from 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 [Liang and Kachalo, 2002; Bertucci et al, 2003; Lockhart et al.,1996; Wodicka et al. 1997] and analysis of DNA variation on a genome-wide scale [Liang and Kachalo, 2002; Chakravarti, 1999; Pollack et al. 1999;

Mei et al, 2000] have become an important tool in the variety of research areas, including cancer research, pharmacogenomics, populational studies, etc.

Although different in some aspects, these applications have many common requirements and utilize 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 length and degree of homology.

The starting point for the introduction of the solid support was observation that single-stranded DNA binds strongly to nitrocellulose membrane in a way that prevents strands from re-association with each other, but permits hybridization to complementary strands [Gillespie and Spiegelman, 1965]. The process of recognition or hybridization can be highly parallel; every sequence in a complex solution mixture can, in principle, be interrogated simultaneously.

Based on these principles, a powerful new experimental technology has been developed, which allows fabrication of hundreds of thousands of polynucleotides at high spatial resolution on a solid surface, allowing for parallel detection and analysis of multiple molecular interactions.

Affymetrix arrays represent one of the major versions of the microarray platform and are based on light-directed synthesis with the use of photolithography and solid-phase DNA synthesis. In brief, synthetic linkers modified with photochemically removable protecting groups are attached to a glass substrate; light directed through a photolithographic mask to specific areas on the surface produces localized photodeprotection. The first of a series of chemical building blocks, hydroxyl-protected deoxynucleosides, are incubated with the surface, and chemical coupling occurs at those sites that have been illuminated in the preceding step. Next, light is directed to different regions of the substrate by a new mask, and the chemical cycle is repeated [McGall and Fidanza, 2001]. This highly efficient strategy allows synthesis of the arbitrary polynucleotides at specific locations; given a reference sequence a DNA probe array can be designed and fabricated that consists of a highly dense collection of complementary probes with virtually no constraints on design parameters. The amount of nucleic acid information encoded on the array is limited only by physical size of the array and the achievable lithographic resolution.

A nucleic acid sample is being used for the synthesis of cDNA, fluorescently tagged throughout entire length of the molecules and hybridized to an array. Subsequent washes remove the majority of the non-specifically bound material. Array scanning involves laser excitation of the incorporated fluorophores; emitted fluorescence is then collected by a lens and passes through a series of optical filters to a sensitive detector. By

scanning with laser beam or translating the array, a quantitative two-dimensional fluorescent image is obtained.

Oligonucleotide arrays are designed and synthesized based on sequence information alone. With the use of the 200-300 nucleotides of the most 3' end of a cDNA sequence, independent 25-mers are selected to serve as sensitive, unique sequence-specific detectors. Probes are chosen based on the set of empirically derived, composition-dependant design rules [Lockhart et al., 1996; Wodicka et al. 1997]. These rules are designed with the intension of improving the odds of choosing oligonucleotides with high specificity and to substantially diminish cross-hybridization effects.

The design of the array implies some level of redundancy, including the use of multiple probes derived from different region of the same gene and the use of mismatch (MM) probes, which are identical to their perfect match (PM) partners except for a single base substitution in a central position. The MM probes are thought out to provide a measure of non-specific hybridization and to discriminate between "real signal" and that due to non-specific hybridization. In theory, hybridization of the intended nucleic acid molecule produces higher signal from the PM probes than from MM probes, resulting in the consistent and recognizable patterns that are unlikely to occur by chance. This approach implies that a major component of the signal derived from any given probe set will be due to duplexes formed predominantly with the involvement of the entire length of the 25-nucleotide probe.

However, a question remains whether or not much shorter stretches of nucleotide homologies (or shorter duplexes) may impact total signal intensity and substantially reduce the discriminatory power of the designed pair PM-MM.

This study was designed to investigate the contribution of low-homologous DNA sequences into cross-hybridization.

2. DATA

We use the Human portion of Affymetrix Latin Square dataset [Affymetrix, 2001], which can be found on Affymetrix corporate website at http://www.affymetrix.com/analysis/download_center.affx or on 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 chip in 59 experiments. Experiments are divided into three groups of twenty, twenty and nineteen experiments.

In each experiment fourteen labeled DNA targets with known concentrations were spiked into labeled complex target and hybridized to the array. Two of fourteen targets (transcripts corresponding to probe sets 37777_at and 407_at) have equal concentrations in each experiment; therefore, there are only 13 distinct targets of varying concentrations in the dataset. The composition of complex target is not specified, however, it was identical within each of the three groups of experiments. In this study we can treat unknown complex targets as three additional targets, each with a concentration of one in one group of experiments and zero in two others.

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 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:

$$R_+ = k_+ x N_{unocc} . \quad (1)$$

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:

$$R_- = k_- N_{occ} . \quad (2)$$

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.:

$$N_{occ} = kx N_{unocc} , \quad (3)$$

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

$$N_{occ} = \frac{kxN}{1+kx} \quad (4)$$

As the probe signal intensity is proportional to the amount of DNA molecules bound to the probe, the same formula can be used for the probe signal intensity y :

$$y = \frac{kxy_{sat}}{1+kx}, \quad (5)$$

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, for $kx \ll 1$ (i.e. for low probe signal intensities), it can be approximated by the linear function:

$$y = bx, \quad (6)$$

where $b = ky_{sat}$ will be called binding coefficient. 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 on many different probes in a series of experiments:

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

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

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

$$y_{ik} = \overline{y}_i + \hat{\varepsilon}_{ik}, \quad (8)$$

where \bar{y}_i is the mean signal intensity of the i -th probe in the whole set of experiments.

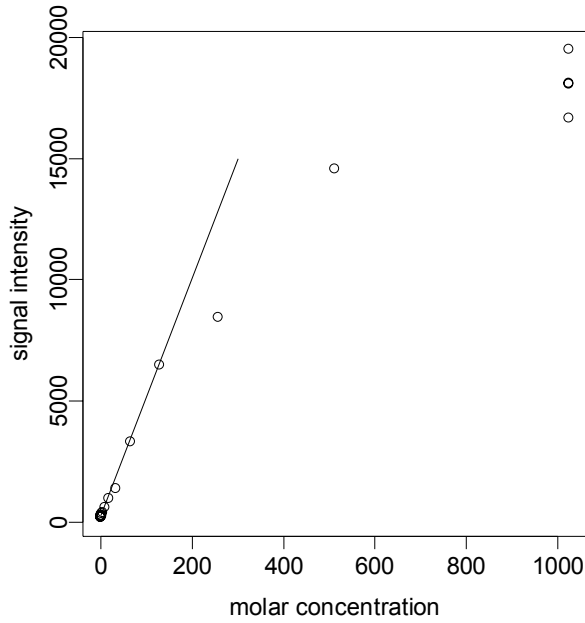


Figure 1. The dependency of the probe signal intensity (in device units) from the molar concentration of DNA transcript (in pmol). DNA transcript 684_at; probe [517:489]; first group of experiments. Probe [517:489] is specific to the transcript 684_at. The dependency can be approximated by linear function for low DNA concentrations.

4. EXPERIMENTAL BINDING COEFFICIENTS

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

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

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]:

$$\begin{aligned} & \text{minimize } \sum \varepsilon_k^2 \text{ in (9),} \\ & \text{subject to: } b_j \geq 0. \end{aligned} \quad (10)$$

The program for solving the problem (10) was implemented as a combination of C++ and Matlab code. It was used to calculate a complete set of binding coefficients for 409,600 probes and 16 targets (thirteen known targets and three complex targets). 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 random binding model (8).

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

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 quadratic programming problem (10) solution gave the best optimization:

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

Next, the probes specific, 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%. Almost identical relationship is observed when using Smith-Waterman [Smith and Waterman, 1981] alignment score with various parameters instead of common substring length.

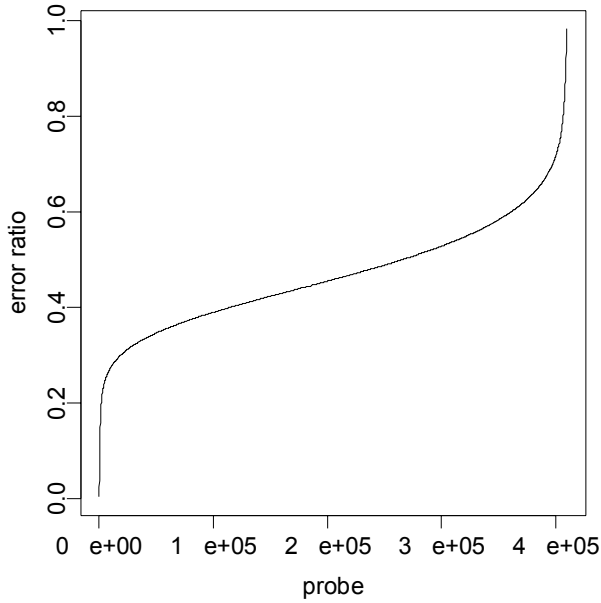


Figure 2. Sorted error ratios $\sum \mathcal{E}_k^2 / \sum \hat{\mathcal{E}}_k^2$ calculated for 409,600 probes.

5. ESTIMATED BINDING COEFFICIENTS

As suggested by above results, even modest similarities 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 all together contribute to the binding coefficient:

$$b = \sum_a n_a c_a + \varepsilon, \quad (11)$$

where b is 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).

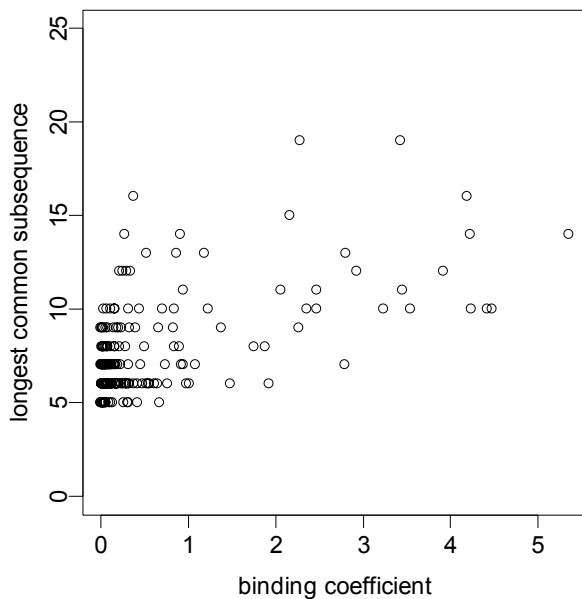


Figure 3. Binding coefficients and longest common substrings lengths for the 304 top probes and transcript 684_at are 61% correlated.

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 definition of binding patterns set 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 HG-U95A microarray is 25 nucleotides, there are only 25 types of binding patterns in the set. If the binding coefficients are known for a set of DNA sequences and a set of probes, equation (11) can now be rewritten as:

$$b_{ij} = \sum_l n_{ijl} c_l + \varepsilon_{ij}, \quad (12)$$

where b_{ij} is 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. Optimization problem to find match contribution in this case will look like:

$$\begin{aligned} & \text{minimize } \sum \varepsilon_{ij}^2 \text{ in (12),} \\ & \text{subject to: } c_l \geq 0, \end{aligned} \quad (13)$$

$$\text{additional condition: } c_{l+1} \geq c_l. \quad (14)$$

We used experimental values of binding coefficients for 304 probes, selected above to calculate the contributions of matches of various lengths into 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). 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, which therefore cannot be compensated statistically.

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

One could expect faster growth of match contribution function with the increase of 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, and because of possible non-linearity their binding coefficients may be underestimated.

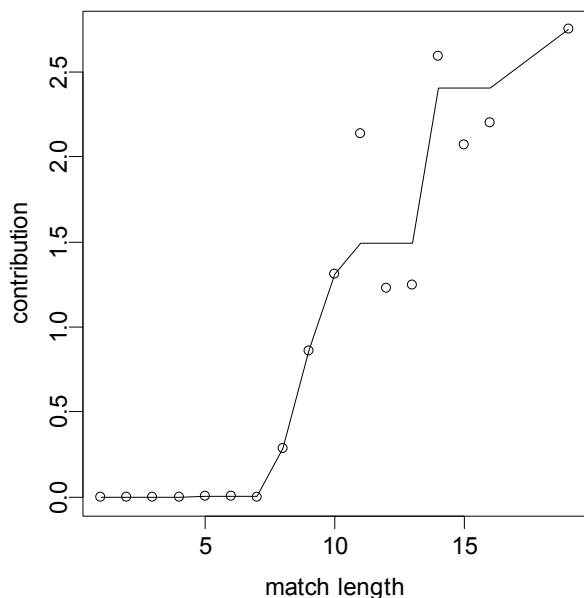


Figure 4. Contributions of perfect matches of different length into the binding coefficient, calculated for transcript 684_at and the top 304 probes. Dots are the solution of problem (13) with no additional condition; line is the solution of the same problem with additional condition (14).

Calculated match contributions were substituted back into (12) to obtain estimated binding coefficients that were afterwards compared with experimental binding coefficients obtained in previous section. Figure 5 illustrates the results of this comparison. Method based on the use of binding patterns performs better than the method using just best match scores. We expect that this method can be further improved by using 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.

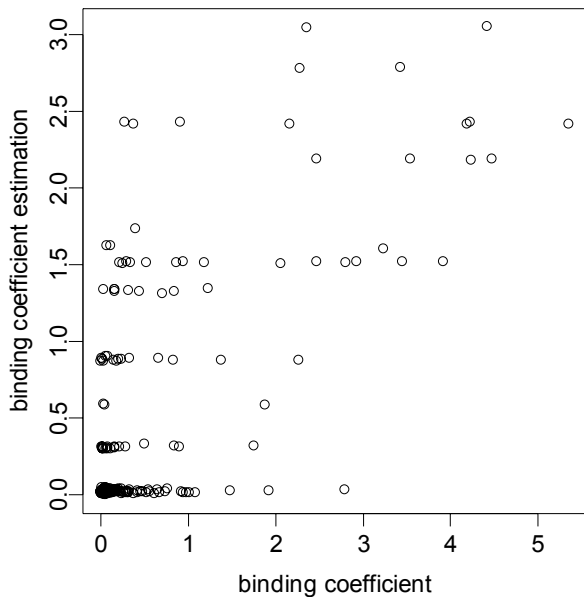


Figure 5. Estimated binding coefficients for the top 304 probes and transcript 684_at are 71% correlated with experimental binding coefficients.

6. DISCUSSION

Our results demonstrate that cross-hybridization can be a significant contribution 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 50% chance to have a 7-nucleotide match with any 25-nucleotide probe. This suggests that any gene, which presents in high abundance during hybridization, can affect the signal intensity on the half of the probes on the microarray. In seven of nineteen possible cases, 7-nucleotide match will cover the central nucleotide of 25-nucleotide probe, and thus, cross-hybridization affects PM probe and its corresponding MM probe differently. The ratio is even worse for longer matches. Therefore, one cannot always be sure about accuracy of results obtained using PM/MM methods.

The main benefit of using linear binding model suggested here is the opportunity to eliminate the effect of cross-hybridization. Once the binding

coefficients are determined (either experimentally or computationally), finding DNA concentrations in (7) from known the probe signal intensities becomes a trivial linear algebra problem that can be effectively solved computationally.

However, for the proper use of linear model the hybridizations should be performed at much lower target concentrations than those commonly used for hybridization, which may result in higher relative level of noise.

To adopt a typical microarray experiment with high DNA concentration a non-linear model with more than one parameter for each probe-target pair can be applied. Its disadvantage in comparison with linear model is that calculation of transcript concentrations from signal intensities can be a difficult mathematical problem requiring substantially longer computational time.

Nonetheless the use of a linear binding model can still be helpful when applied to analyzing microarray data obtained from experiments with high target concentrations. It can help to determine whether a signal corresponds to a low expressed gene or it is just a result of cross-hybridization

ACKNOWLEDGEMENTS

We would like to thank Seby Edassery and Peter Larsen from the Core Genomics Facility at UIC for their help during all stages of our work. We are also grateful to the members of CAMDA'02 Organizing Committee for the inspiration to start this work and for a very interesting dataset provided. This work is supported in part by a grant from The Whitaker Foundation (RG-00-0085).

REFERENCES

- J. Liang, S. Kachalo. (2002) Computational analysis of microarray gene expression profiles: clustering, classification and beyond. *Chemometrics and Intelligent Laboratory Systems* 62:199-216.
- Bertucci F., Viens P., Hingamp P., Nasser V., Houlgatte R., Birnbaum D. (2003) Breast cancer revisited using DNA array-based gene expression profiling. *Int. J. Cancer*. 20;103(5):565-71
- 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., Brown E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14(13):1649.
- Wodicka L., Dong H., Mittmann M., Ho M.H., Lockhart D.J. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15(13):1359-67

- Chakravarti A. (1999) Population genetics-making sense out of sequence. *Nat. Genet.* 21(1 Suppl):56-60
- Pollack J.R., Perou C.M., Alizadeh A.A., Eisen M.B., Pergamenschikov A., Williams C.F., Jeffrey S.S., Botstein D., Brown P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23(1):41-6
- Mei R., Galipeau P.C., Prass C., Berno A., Ghandour G., Patil N., Wolff R.K., Chee M.S., Reid B.J., Lockhart D.J. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* 10(8):1126-37
- 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
- McGall G.H., Fidanza J.A. (2001) Photolithographic synthesis of high-density oligonucleotide arrays. *Methods Mol. Biol.* 170:71-101
- 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
- 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
- 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
- Vernier P., Mastripiolito 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
- 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
- Affymetrix, Inc. (2001) New statistical algorithms for monitoring gene expression on GeneChip probe arrays. Technical report.
- J. C. G. Boot (1964) *Quadratic Programming*. North-Holland
- Smith T.F. and Waterman M.S (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.