Microarray Oligonucleotide-probes

David P. Kreil[1] Roslin R. Russell[2] Steven Russell[2]

[1]Chair of Bioinformatics, Department of Biotechnology,
University of Natural Resources and Applied Life Sciences, Vienna
Muthgasse 18, A-1190 Vienna, Austria

[2]Department of Genetics,
University of Cambridge
Downing Street, Cambridge CB23EH, U.K.

Contact: MethEnz05@Kreil.Org
R.Russell@gen.cam.ac.uk
S.Russell@gen.cam.ac.uk

Abstract

Oligonucleotide probes are increasingly the method of choice for many modern DNA microarray applications. They provide higher target-specificity, probe selection gives improved experimental control of hybridization properties, and targetting of specific gene subsequences allows the better discrimination of highly similar targets such as splice-variants or gene families. Only recently has there been substantial progress in dealing with the complexities of probe set design and probe-specific signal interpretation. After a discussion of advantages and disadvantages of oligonucleotide probes in comparison to amplicons, this manuscript therefore focusses on recent advances and remaining key challenges in probe design and computational data analysis for spotted and in-situ synthesized oligonucleotide microarray technologies. Both experimental questions and computational aspects are addressed. Experimental issues discussed include the choice of an optimal number of probes per target and probe lengths and their influence on bias and random measurement noise, effects of different probe or substrate modifications and laboratory protocols on signal specificity and sensitivity. Computational topics include practical considerations and a case study in probe-sequence design, the exploitation of probing multiple target regions, and the modelling of probe-sequence specific signals. The current state-of-the-art of the field is examined, and principled thermodynamic probe-design criteria are proposed that are based on the free energy of the probe-target complex at the hybridization temperature, rather than its melting temperature. Lastly, we note and discuss an emerging trend in recent computational work towards a focus on signal interpretation rather than probe-sequence design.

DNA microarray technology is pervading many aspects of the life-sciences. From humble beginnings, detecting the expression of a few tens of genes, entire eukaryotic genomes can now be interrogated (Schena *et al.*, 1995; Bertone *et al.*, 2004). While the technology can be used for a variety of applications (Hanlon and Lieb, 2004; MacAlpine and Bell, 2005; Pinkel and Albertson, 2005), its main use is still in gene transcript expression profiling. Often, microarrays are used to screen for genes involved in a particular biological process of interest; however, larger datasets of comprehensive transcript coverage measured under a variety of conditions have considerable potential for much wider, systems-level analysis, *e.g.*, *via* the detection of co-regulated groups of genes (Saidi *et al.*, 2004; Lee and Batzoglou, 2003; Ihmels *et al.*, 2002; Ihmels *et al.*, 2004). Sensitive pattern-detection tools require particularly accurate data so that biologically meaningful signatures can be distinguished from confounding experimental effects, which can only partly be removed at analysis-stage (Kreil and Russell, 2005). At present, unfortunately, hybridization signal levels measured are not easily related to absolute quantities of target transcripts. This chapter outlines the advantages and challenges of using oligonucleotide-probes for transcript expression profiling, discusses typical considerations and practical aspects in probe-sequence design, and highlights recent developments in the modelling of hybridization behaviour that are of relevance for probe design and the interpretation of hybridization signals. Whilst recognising that there are many sources of bias and noise in microarray data, developing an understanding of probe hybridization behaviour will be instrumental in achieving a quantitative view of the transcriptome.

The case for oligonucleotide-probes

There are two types of common DNA microarray probes: oligonucleotides and double-stranded amplicons (Schena *et al.*, 1996; Johnston *et al.*, 2004). Amplicon probes have particularly high sensitivity and, for some applications, their relatively large tolerance to small transcript-sequence variations can be helpful – *e.g.*, transparently tolerating naturally occurring polymorphisms. This same property, however, makes amplicon-probes less well suited for the discrimination of very similar targets, such as alternative-splicing variants, or families of paralogous genes. With all amplicon-based probes, moreover, the technical problems associated with PCR-amplification of thousands of clones are not easily overcome (Hegde *et al.*, 2000, and Burr *et al.*, this volume). Consequently, some laboratories report that only 66–79% of probes were not contaminated and matched their respective targets (Hager, this volume). Nevertheless, probing species without a fully sequenced genome, comparing highly related strains, or exploiting specialized cDNA-libraries, indicate the use of amplicon arrays (Suchyta *et al.*, 2003; Diatchenko *et al.*, 1996).

With the increased experimental control available with oligonucleotide-probes and because of the challenges of manufacturing amplicon-probes of uniform and validated quality, many modern microarray applications use synthesized oligonucleotide-probes. Either multiple shorter probes *per* target are employed, as with Affymetrix chips (Lockhart *et al.*, 1996), or longer oligonucleotide-probes are used, typically 35–70-mers (Kane *et al.*, 2000; Hughes *et al.*, 2001; Nuwaysir *et al.*, 2002). Oligonucleotide-probes overcome many of the difficulties of amplicons and show increased target sequence discrimination (Duggan *et al.*, 1999; Relogio *et al.*, 2002). Moreover, one can ensure uniform probe concentrations, hybridization affinities, and minimal cross-hybridization. Consequently, very clean arrays can be achieved.

Considerations for oligonucleotide-probe design.

Many issues affect probe design, no matter whether arrays are to be produced commercially or in-house. The number of probes *per* target and their lengths must be chosen first, and the following section discusses tradeoffs that must be made due to technical limitations of production platforms. Additional complexities, deferred to later sections, include the discrimination of multiple splice-variants of a transcript, issues of target secondary-structure, and the detection of RNA degradation. Even without considering these, thorough studies regarding optimal choice of probe length or the number of probes *per* target are difficult, and few systematic comparisons exist.

Number of probes per target and probe lengths

There are two properties of the microarray measurement process that one wishes to **maximize** for transcript expression profiling: 1) the *sensitivity*, a measure of how little is lost of the signal reflecting specific hybridization between the probe and its target; and 2) the *specificity*, a measure of how little non-specific hybridization there is of the probe with molecules other than its target. At the same time, one aims to **minimize** two other properties of the measurement process: 1) the random signal variation or *noise*, often expressed as coefficient of variation (*CV*), the standard deviation divided by the mean of multiple measurements; and 2) the *bias*, the systematic deviation of the measurement from the true signal due to probe-specific or other confounding technical effects.

Sensitivity generally increases with probe length since the binding-energy for longer probe-target hybrid complexes is typically higher; 60-mers, for example, detect targets with eightfold higher sensitivity than 25-mers (Chou *et al.*, 2004). The specificity of very short probes decreases with diminishing probe length because of the increasing chance of a random match to non-target sequences. On the other hand, the specificity of very long probes decreases with growing probe length because the chance that a fragment of the probe matches an unwanted target increases with probe length. The fact that biological nucleotide sequences are not random further contributes to this because different targets can share domains of high sequence similarity.

Noise can be reduced by increased binding-energy *via* greater probe length and by making multiple measurements *per* target – be that through replicate or multiple probes. Bias of individual probes should ideally be measured or modelled and removed. The bias of a set of probes for a given target can also be reduced through the combination of multiple probes of random bias, where the average bias decreases with the number of probes.

Although experimental comparisons are difficult because optimal hybridization protocols differ for probes of different lengths and binding affinities, qualitative trends observed for 25–500-mers (Fig. 1) suggested that, depending on how one weighs optimization criteria, probe lengths between 50 and 150 may yield a good compromise (Fig. 2). For the shorter probes, replicates from different arrays or multiple probes should be used to reduce noise, and probe bias compensated, *e.g.*, by the computational means discussed later.

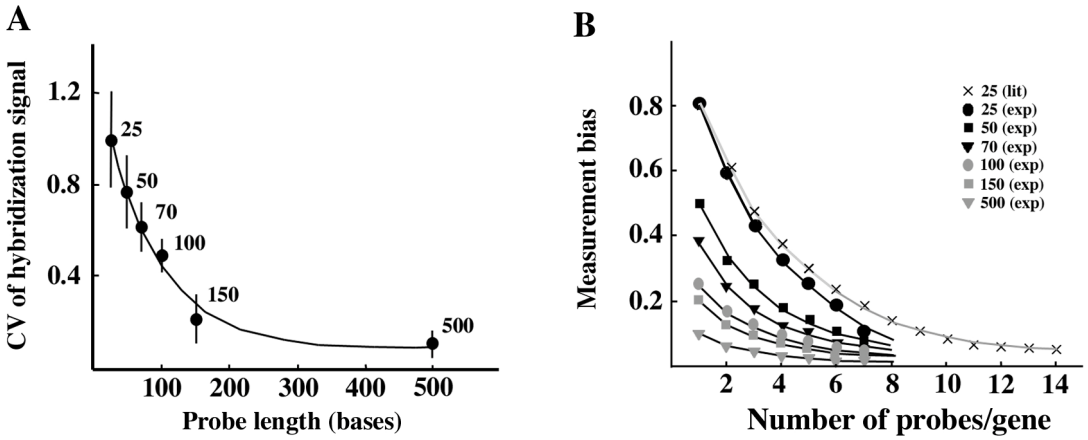


Figure 1: Noise and bias. A) Noise (CV) reduces with probe length. B) Bias decreases with increased probe lengths (see symbols in legend) and with larger numbers of probes *per* gene. Bias was assessed as the average deviation between the robust means of the signals of random probe *subsets* and the robust mean of the signals of the *full* probe set (Chou *et al.*, 2004). (Redrawn after Chou *et al.*, .)

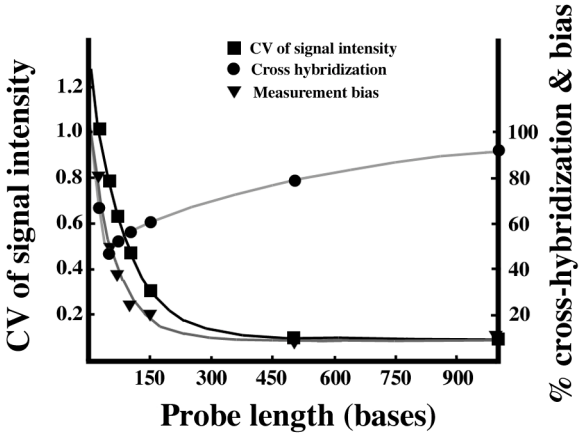


Figure 2: Optimization by multiple criteria: three of the four criteria discussed are shown. The fourth, sensitivity, increases monotonically with probe length over the studied range (Chou *et al.*, 2004). (Redrawn after Chou *et al.*, .)

A very different type of 'specificity' measure, indicative of detection performance for single-nucleotide polymorphisms (SNPs), is obtained in studies comparing hybridization signals from 'perfect match' (PM) probes and 'mismatch' (MM) probes featuring a single-base mismatch (Relogio *et al.*, 2002). In transcript expression profiling, however, sensitivity to SNPs is detrimental because of natural sequence polymorphisms in samples.

Microarray production and hybridization protocols

Hybridization behaviour of probes, measurement sensitivity, and specificity strongly depend on microarray production and hybridization protocols. The steric hindrance of

the solid support to which probes are attached, *e.g.*, reduces hybridization more than twofold and, together with electrostatic effects, may render the terminal bases close to the support effectively invisible (Shchepinov *et al.*, 1997). Terminal uncharged amphiphilic spacer groups with 30–60 carbon-carbon bonds provide a remedy. Length 40 seems optimal, as longer spacers reduce the effective concentration of the probe by allowing diffusion. Thus creating sufficient space between substrate and probes to mitigate sterical and electrostatic effects on hybridization yet also limiting probe diffusion, such optimal spacer groups could achieve a 150-fold increase in hybridization sensitivity while making the entire probe available for specific hybridization (Shchepinov *et al.*, 1997). Slide substrate coatings with a 'gel-like' spacing effect provide an alternative, *e.g.*, GE Healthcare CodeLink or FullMoonBioscience PowerMatrix substrates (Ramakrishnan *et al.*, 2002; Le Berre *et al.*, 2003).

Relevant hybridization protocol parameters include hybridization temperature and duration, hybridization buffer composition and additives (such as formamide), and the stringency of washes. In the manufacture of spotted arrays, other protocol parameters of relevance include spotting buffer and microarray slide substrate chemistries, probe concentrations, and ambient temperature and humidity (Kreil *et al.*, 2003; Auburn *et al.*, 2005). The protocols employed should be validated for sensitivity *and* specificity, *e.g.*, by spike-in experiments, which are also easily assessed visually, an efficiency factor in protocol-parameter screens. Typical results reflect the confounding effects of experimental conditions on hybridization behaviour: False-colour hybridization images of a spotted probe (Supplement) showed the clear detection of a difference in transcript expression, yet a change in just the microarray manufacturing chemistry made the same probe hybridize non-specifically in both channels. While not suitable for immediate visual assessment, an optimization of the detection sensitivity for differential expression provides an alternative quantitative assay less dependent on a set of spike probes being representative of an array. Multiple replicate hybridizations and the use of dye-swaps for multi-channel systems ensure that differential signals cannot occur by chance and that a maximization of the number of difference calls corresponds to optimal hybridization conditions, once not-detected spots are correctly accounted for (Kreil *et al.*, in preparation).

In order for the measurement process to be both sensitive and specific for, ideally, *all* the probes on the array, one requires probes of uniform hybridization properties *at the reaction conditions employed*. This is the objective of probe-sequence design.

Practical considerations in probe-sequence design, a case study

Probe-sequence design is complex for a variety of reasons, ranging from trivial technical nuisances to difficult theoretical problems that are the subject of active research. The prediction of actively transcribed genome regions, for instance, is still far from exhaustive, and hence the mixture of target transcripts that need to be discriminated in a sample is not fully known. Already a prediction of the properties of a candidate probe under competitive hybridization with a known mixture of different transcripts is a difficult thermodynamic modelling challenge. As a consequence, more or less crude approximations and heuristics are often employed, with modelling replaced by sequence-similarity searches and alignments, and *ad-hoc* rules of thumb regarding probe sequence-complexity and secondary structure. These simplified approaches, however, are typically not sufficiently reliable on their own and usually require experimental

validation of candidate probes in a final probe-selection step. Approaches exploiting the different hybridization kinetics of specific and non-specific binding to detect non-specific probes are only applicable to probes for sufficiently highly expressed targets (Dai *et al.*, 2002). Comprehensive experimental validation is hence rarely performed for reasons of complexity and cost.

While a plethora of probe-design software is available (*cf.* Supplement), it is clear that designs that are less crude approximations to proper thermodynamic modelling are better predictors of oligonucleotide hybridization performance (Luebke *et al.*, 2003). The common limitations of readily available software can roughly be classified into the following categories:

1. Limited validity of heuristics and approximations *in lieu* of proper thermodynamic modelling; *e.g.*, sequence-similarity or 'consecutive matches' to probe complements as indicative of cross-hybridization; or sequence palindromes as indicative of hairpin secondary-structure instead of proper thermodynamic hybridization and folding models;
2. Replacement of parameter optimization by (a) acceptance of any solution with parameters meeting specified thresholds; or (b) requiring users to fix the parameter for design runs; *e.g.*, accepting any probe with binding-energy in a given range; or requiring users to specify both probe-length and energy thresholds rather than inferring one from the other;
3. Lack of support for dealing with similarities in biological sequences and the consequential difficulty of designing specific probes for groups of similar sequences, *e.g.*, alternative splice-forms or paralogous genes;
4. Technical issues such as malfunctions or undocumented software requirements.

This section does not attempt a comprehensive discussion but instead highlights *typical* consequences of problems encountered using the design of a genome-scale transcript array for *Drosophila melanogaster* (Kreil *et al.*, in preparation) as a case study, covering:

- Consequences of the choice of probe design software
- Construction of a set of target transcript sequences
- Choosing design parameters
- Searching for 'optimal' probes
- Post-processing, *e.g.*, to account for under-prediction of transcripts from the genomic sequence

Although critical at all phases of employment, technical issues are collected in the Supplement because of their often transient nature.

Choice of software, and construction of the target transcript set

OligoArray 2.1 ('OA2', `http://berry.engin.umich.edu/oligoarray2_1/`) uses relatively few heuristic shortcuts (Rouillard *et al.*, 2003): Cross-hybridization and self-folding are assessed using full two-state thermodynamic models employing the `mfold` algorithm, allowing for mismatches, bulges, loops, and hairpins (SantaLucia,

1998; Zuker, 2003). Otherwise, OA2 is comparable to many other probe-design tools in using BLAST sequence-similarity search (Altschul *et al.*, 1997) together with heuristics to screen non-target transcripts for potential cross-hybridization. Access to the software source-code (unpublished) allowed the verification of the implementation and was most valuable for dealing with technical issues as they emerged. The sources of a revised version will be published later this year (J.-M. Rouillard, *pers.comm.*, 2005).

As OA2 has no concept of 'related' sequences and treats all predicted stable hybridizations to non-target transcripts equally, duplicate and very similar sequences had to be removed in the construction of a 'non-redundant' set of target transcript sequences, using tools like `nrdb90` or `CD-HI` (Holm and Sander, 1998; Li *et al.*, 2001). For compatibility to common labelling methods, design was restricted to the 1500 base 3'-regions of targets and sense probes had to be built for the labelled (anti-sense) targets derived by reverse transcription from the (sense) mRNAs in samples (Marko *et al.*, 2005).

Choice of design parameters, search for 'optimal' probes

OA2 execution parameters provide thresholds for the acceptance of probe candidates. The probe candidate closest to the 3'-terminal of the target sequence that passes all criteria is selected: probe length and probe-target melting-temperature T_m within given ranges, no stable probe secondary-structure (self-folding), GC content in range (which we did not restrict), no tandem repeats, and a minimal number of predicted stably hybridizing non-target transcripts. Accepted probe lengths were set to 65–69 as pilot experiments had demonstrated a good compromise between sensitivity and specificity with the protocols employed in our laboratory.

OA2 default parameters for 45–47-mers permit 85°C≤T_m≤90°C, tolerate stable cross-hybridization only for T_x<65°C, and stable probe secondary-structure for T_s<65°C. Examining the T_m values of all candidate probes in the 1500 base 3'-regions of target sequences yielded a set of T_m values *per* sequence. Some target sequences had extreme probe-candidate T_m distributions, with min(Q3(T_m))=76.3 and max(Q1(Tms))=97; Q1/3 denoting the first and third quartiles, respectively. On the other hand, most targets had melting-temperatures in a common range, with (Q1,median,Q3)(median(T_m))=(87.0,89.0,90.5). This was well matched to the suggested tolerated T_m interval of 85–90°C: More than 90% of target sequences were covered with at least 25% of candidate probes *per* target having a T_m in this interval. For our target set, the optimal 5°C range maximizing coverage for 45–47-mers was 86.6–91.6°C.

In contrast, for 65–69-mers, the extremes were min(Q3(Tms))=81.5 and max(Q1(Tms))=100.6, while (Q1,median,Q3)(median(Tms))=(91.7,93.3,94.7). Less than half the target sequences, however, were covered with at least 25% of candidate probes having a T_m in the default interval 85–90°C, severely reducing the number of probe candidates that could be considered. Shifting the 5°C window to 90.6–95.6°C (a 5.6°C offset to OA2 defaults), however, could achieve coverage of 94% of all target sequences with at least 25% of candidate probes in range (Fig. 3). Thus, for most target sequences a large number of probe candidates could be considered, increasing the likelihood that a specific probe with no cross-hybridization could be found. For a small number of target sequences (6%), however, probe-design meeting these parameter thresholds was difficult.

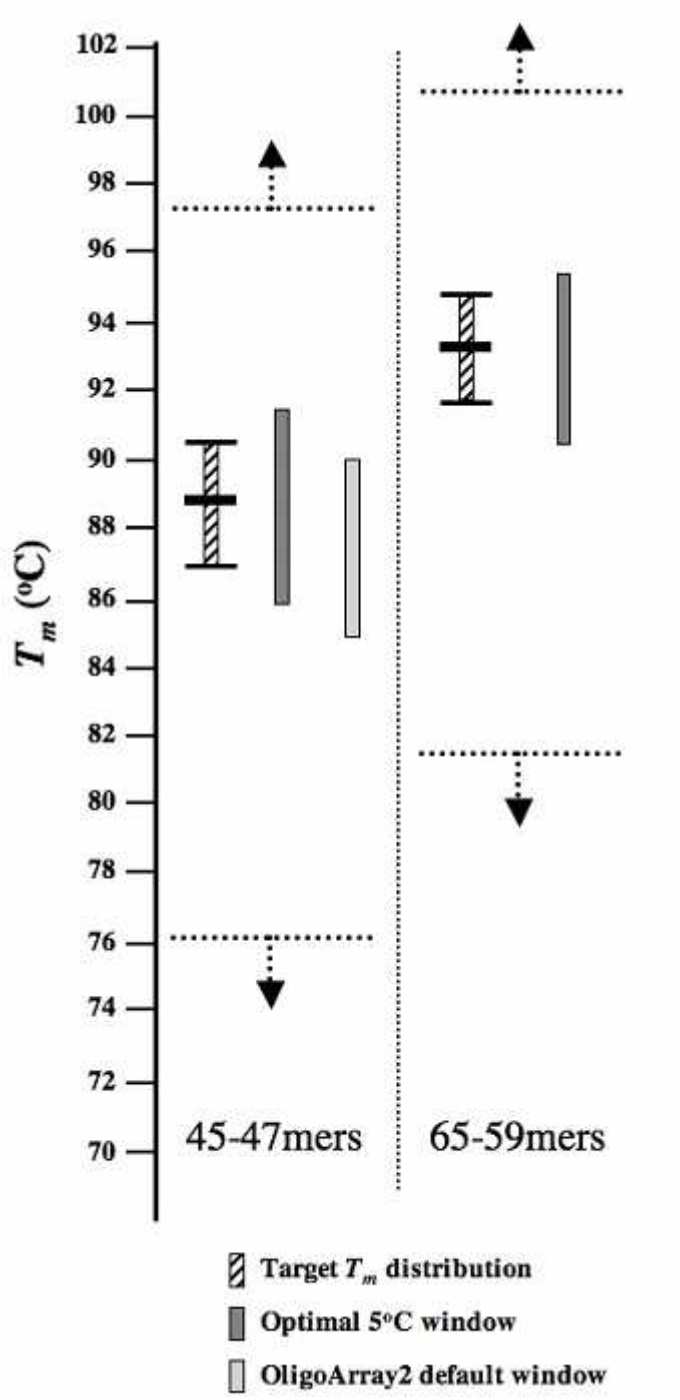


Figure 3: Choice of parameter thresholds for oligonucleotide-probe design. A design parameter selection well-matched to the parameter distribution in the probe candidates extends the search space of acceptable candidates, increasing the likelihood that specific probes can be found. The right panel considers 65–69-mers in comparison to 45–47-mer probes (the OA2 default probe-lengths) in the left panel. The distribution of the median T_m of all possible candidate probes *per* target is shown. The median of these for the set of target transcripts is marked by a black bar; the whiskers indicate first and third quartiles. The OA2 default T_m window is reasonable for 45–47-mers but needs adjustment for different probe lengths. While a 5°C window can be found to suit most targets, there typically are targets with unusual properties requiring reruns

with different parameter sets. This is illustrated for the most extreme cases: 75% of probe candidates for these transcripts have a T_m beyond the values indicated by the dotted-lines. (Predicted T_m values as calculated by OA2.)

The T_x and T_s thresholds were conservatively adjusted by 4°C from 65°C to 69°C, leaving a margin of 1.6°C for the effect of T_m overestimation at large temperatures (Rouillard *et al.*, 2003), also matching the observed shift of the 'optimal' T_m window. Targets with no satisfactory probes were rerun with increasingly relaxed parameters.

Employment and post-processing

The design-runs for the parameter sets considered were executed on a distributed collection of computers using Grid Engine (`http://gridengine.sunsource.net/`). Accounting for possible under-predication of transcripts from the genomic sequence, all OA2-selected probes were screened to exclude predicted stable hybridizations to any genomic DNA sequence (BLAST search of both genomic sequence strands plus standard OA2 heuristics and `mfold` thermodynamic calculation). To partly compensate the lack of support for transcript groups, when no specific probe was found, probes only predicted to cross-hybridize to alternative splice-forms of the target gene were chosen over probes with predicted cross-hybridization to transcripts of different genes.

Conclusion

It is noteworthy that there are no tools presently capable of automatically selecting a uniform range of thermodynamic properties allowing high specificity for most probes and delivering a probe set appropriately dealing with families of paralogous genes and alternative-splicing variants. Combining results from multiple OA2-runs with carefully selected parameters, however, a 'state-of-the-art' probe set could be obtained, with probes for more than 90% of all targets meeting all design criteria. For about 4% of targets, however, probes were predicted to cross-hybridize with transcripts from non-target genes, most likely orthologues. Support for orthologues and splice-forms, and more automated parameter selection could considerably simplify employment. Lastly, while OA2 gives very little control of probe placement, this property is a particular strength of oligonucleotide arrays.

Locations of probe-sequence target regions, discrimination of highly similar targets

The ability of probing specific target regions can be exploited, *e.g.*, to test for RNA integrity. Probe-location dependent trends in signals from multiple probes for an abundant transcript indicate RNA degradation. While bacterial RNA is degraded in 3'-to-5' direction, eukaryotic RNA is degraded by exonuclease digestion from the 5'-end (Brown, 2002). Transcript secondary structure may, however, hide particular target regions and hence affect probe hybridization signal (Ratushna *et al.*, 2005), a complication that should be considered (but usually isn't) in probe design and signal interpretation.

In contrast to the 3'-bias of many commonly used labelling methods, modern protocols can provide labelled full-length targets (Castle *et al.*, 2003; Johnson *et al.*, 2003). Calculated placement of probes then allows the discrimination of highly similar tar-

gets, like families of paralogous genes or alternative splice-forms. The latter are of particular interest in the quest for understanding complex eukaryotes. Alternative splice-forms are predicted for ~50% of all human genes, comprising a complex variety of transcriptional constructs hard to distinguish with microarrays (Lander *et al.*, 2001; Shai, 2004). While exon-junction spanning probes (Kane *et al.*, 2000) can improve discrimination (Fig. 4) they can also cross-hybridize with alternative splice-forms due to construction constraints. Moreover, 5% of human splice acceptor sites have NAGNAG motifs (N being a SNP), parts of which are received by some splice-forms (Hiller *et al.*, 2004). Probes for expression analysis must tolerate both this alternative motif inclusion and the motif degeneracy while being highly specific to the target splice-form queried.

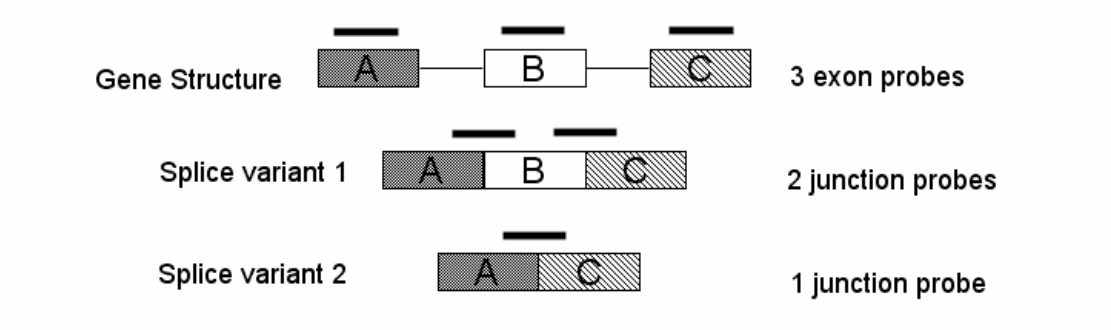


Figure 4: Exon and exon-junction probes. Black bars indicate probe locations. Direct measurement of splice-variant 2 requires exon-junction probes.

Clearly, signal interpretation constitutes a critical and challenging aspect of these microarray applications, which is reflected in a wide range of approaches. Adding gene-structure specific effects *per* splice-form in a linear model of effects (Li and Wong, 2001), specific splice-forms could be discriminated for genes of known structure (Wang *et al.*, 2003). GenASAP could deduce splicing events from exon and exon-junction probe data by fitting a Bayesian generative linear model for single-cassette exon inclusion/exclusion using structured variational expectation-maximization (Shai, 2004). Comparing samples against their mixture and introducing (unknown) probe- and splice-form-specific affinities in a linear model of effects (Li and Wong, 2001), differences in splicing patterns between samples could be detected (Le *et al.*, 2004). Present approaches to analysing such complex datasets do not explicitly model cross-hybridization. With the severity of such probe-level effects, however, further progress is expected from including individual probe characteristics into the modelling process. The subsequent sections give an overview of our current state-of-the-art understanding of probe behaviour.

In-situ synthesis *vs* deposition of pre-synthesized oligonucleotides

Both robotic deposition of pre-synthesized oligonucleotides on arrays (Auburn *et al.*, 2005) and *in-situ* synthesis of probes each have their advantages and disadvantages. Using fixed-mask lithography, an approach pioneered by Affymetrix (Lockhart *et al.*, 1996), oligonucleotide synthesis is achieved by repeated cycles of base additions with different masks for light-directed deprotection of terminal hydroxyl groups (Pirrung,

2002). The typical coupling efficiency of only 92–94% *per* step (McGall *et al.*, 1997), however, limits the technology to short probes (Fig. 5), although improved photosensitive groups exist (Pirrung, 2000). Typically, 11–14 probes of 25 bases are used *per* target for transcript expression profiling. Fixed mask lithography produces ~1,300,000 probes/chip, making this the technology of choice for extremely high numbers of probes. On the other hand, while well-suited for industrial production of standard arrays, making small numbers of specialized arrays is uneconomical.

Very high density arrays can flexibly be produced by *in-situ* synthesis *via* digital micro-mirror device (DMD) lithography, yielding ~400,000 features/array. Since improvements in photosensitive deprotection efficiencies from 95% to 98% giving stepwise synthesis yields of up to 96% (Singh-Gasson *et al.*, 1999; Nuwaysir *et al.*, 2002; Buhler *et al.*, 2004), arrays for transcript expression profiling are offered with 60-mer probes that typically employ 5 or more probes *per* target (*cf.* Nimblegen arrays, Scacheri *et al.*, this volume). *In-situ* synthesis by ink-jet deposition can flexibly produce high density arrays of ~40,000 spots of excellent spot morphologies. Coupling efficiencies of up to 98% allow higher-yield synthesis of 60-mer probes (Hughes *et al.*, 2001; Lausted *et al.*, 2004). Typically one probe is used *per* target for transcript expression profiling.

As an alternative to *in-situ* synthesis, **pre-synthesized** oligonucleotide-probes can be spotted at high density, giving arrays of ~40,000 probes. Compared to *in-situ* synthesis, pre-synthesized probes can be produced at much higher purity and yield.
A coupling efficiency of >99% can be achieved in synthesis, and purification of the final product is possible by one or multiple rounds of reverse-phase high-performance liquid-chromatography (RP-HPLC), which works well for shorter oligonucleotides, and/or polyacrylamide gel electrophoresis (PAGE). Typically, 50–70-mers are used for transcript expression profiling, with one probe *per* target. Spotted arrays also allow more complex designs in which probes for multiple targets are spotted as composite probes for multiplexed target measurements or normalization purposes (Shmulevich *et al.*, 2003; Yang *et al.*, 2002).

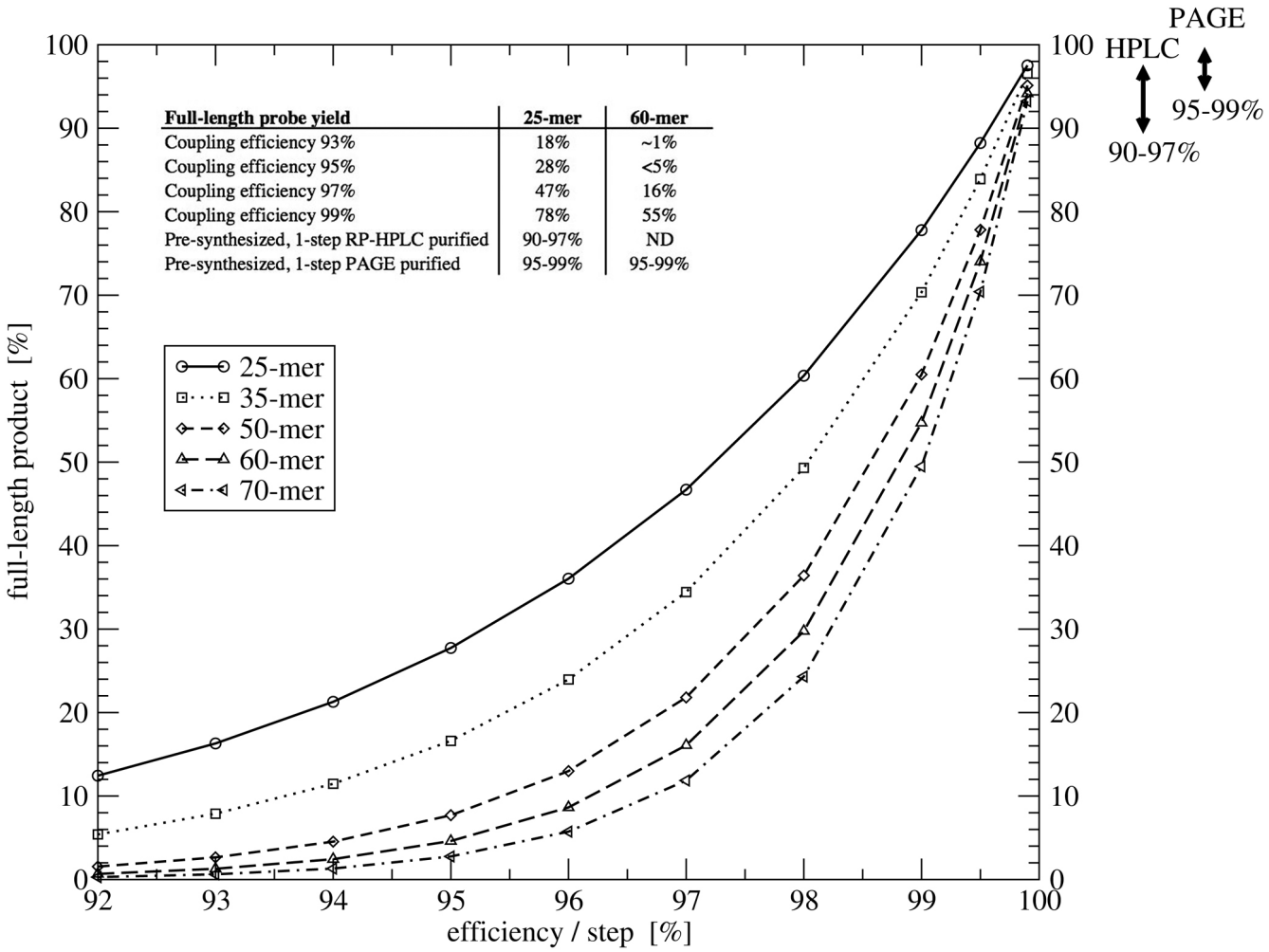


Figure 5: Response of synthesis yields to varying coupling efficiencies for oligonucleotides of different lengths. Photolithographic and ink-jet synthesis typically achieve efficiencies of 94–96% and ~97% *per* step, respectively. Pre-synthesized nucleotides are made with efficiencies >99%, and subsequent purification steps are feasible, of particular relevance for longer oligonucleotide-probes. A single round of RP-HPLC obtains 90–97% of full-length product, PAGE yields 95–99% purity.

Since probes containing mixtures of prematurely terminated oligonucleotides reduce measurement specificity at optimal hybridization conditions (Jobs *et al.*, 2002) and purification steps are expensive, many laboratories spot probes with 5'-terminal amino-groups onto aldehyde substrates. Only full-length probes bind to the substrate covalently while prematurely terminated oligonucleotides are washed off. Increased probe purity extremely simplifies thermodynamic modelling.

Thermodynamic modelling of microarray probe hybridization

Microarray specific effects

While the thermodynamics of nucleic acid hybridization in solution has long been an area of extensive research (Dimitrov and Zuker, 2004; SantaLucia and Hicks, 2004), only the recent popularization of microarrays has brought the more convoluted issue

of hybridization behaviour of oligonucleotides tethered to a solid support into the focus of current research. The solid support can interfere with target molecule binding sterically and chemically. Even with gel-like substrate coatings or spacers attached to probes reducing this effect, it was surprising that models for hybridization behaviour in solution could directly be applied for pre-synthesized probes attached to a gel substrate, once a linear correction was applied to thermodynamic parameters (Table 1); this even unaffected by fluorescent end-labels (Fotin *et al.*, 1998).

Thermodynamic parameter	*Linear correction for microarrays*
ΔH^0	$\Delta H^0_{\text{array}} = \Delta H^0_{\text{solution}} - 24$
ΔD^0	$\Delta D^0_{\text{array}} = \Delta S^0_{\text{solution}} - 70$
ΔG^0, original paper, slope constrained $= 1$	$\Delta G^0_{\text{array}} = \Delta G^0_{\text{solution}} - 3.2$
ΔG^0, original paper, slope unconstrained	$\Delta G^0_{\text{array}} = 1.1\,\Delta G^0_{\text{solution}} - 3.2$??
ΔG^0, recalculated, slope unconstrained	$\Delta G^0_{\text{array}} = 0.78\,\Delta G^0_{\text{solution}} - 1.0$
ΔG^0, `HyTher`	$\Delta G^0_{\text{array}} = 0.85\,\Delta G^0_{\text{solution}} - 2.33$

Table 1: Linear corrections to thermodynamic parameters for oligonucleotide-probes attached to a solid support. The first alternative formula for ΔG^0 gives the relationship published in the original paper by Fotin *et al.* (1998), where the slope has been assumed to be one. The next line shows regression results without this constraint, as published. This does not, however, fit the data in Table 3 of the Fotin *et al.* paper (J. SantaLucia, Jr., *pers.comm.*, 2005). The formula labelled 'recalculated' was obtained by linear least-squares regression from the original table data (Fotin *et al.*, 1998), while the last line shows the correction suggested by `HyTher` (`http://ozone2.chem.wayne.edu/`).

The situation for probes from manufacturing processes giving mixtures of prematurely terminated oligonucleotides is more complicated. For a long time, therefore, probe-sequence specific variation in signal intensity from such arrays was not understood. Sequence-specific probe bias, particularly strong for short sequences, was reduced by combining measurements from multiple probes, yet without exploiting probe-sequence information (Li and Wong, 2001; Bolstad *et al.*, 2003). Recently, however, *empirical* models of sequence-specific binding with position-specific weights have been introduced: The predicted contributions of probe regions to the overall binding strength are attenuated depending on their positions along the probe-sequence.

For data from Affymetrix chips, Zhang *et al.* (2003) successfully fit the signal intensities of a particular probe *i* for a target *j* as sum of contributions from specific and non-specific binding to the probe plus a global background constant *B*:

$$I_{ij} = \frac{N_j}{1+\exp(E_{ij})} + \frac{N^*}{1+\exp(E^*_{ij})} + B \quad .$$

N_j is the number or target molecules, N^* the number of molecules binding non-specifically to (all) probes. For a probe-sequence (b_1,b_2,…,b_k,…,b_{25}), the free-energy terms for specific and non-specific binding, $E_{ij} = \sum_{k=1}^{25} \omega_k \, \varepsilon(b_k, b_{k+1})$ and $E^*_{ij} = \sum_{k=1}^{25} \omega^*_k \, \varepsilon^*(b_k, b_{k+1})$, are parameterized by empirical base-pair stacking energies $\varepsilon/\varepsilon^*$ and position-dependent weights ω_k/ω_k^*. This simple model fitted probe signal levels well, removing probe-sequence specific bias, apparently of particular relevance for low-intensity signals. The probe centre gave the largest contribution to binding (Fig. 6). The empirical base-pair stacking energies, however, can vary considerably between different chip designs (data from `http://odin.mdacc.tmc.edu/~zhangli/PerfectMatch/`), reflecting the empirical nature of the model.

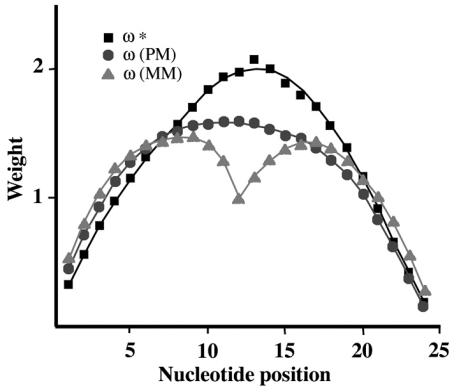


Figure 6: The position-specific weights in a position-dependent nearest-neighbour model. The centre part of an Affymetrix probe gives the strongest contribution to binding. The curve for the mismatch probes (MM) reflects destabilization from the central mismatch base. (Redrawn after Zhang *et al.*, 2003.)

Naef and Magnasco (2003) use position-dependent affinities A_k in modelling probe-specific signal intensities for Affymetrix chips,

$$\ln\left(\frac{I_{ij}}{\underset{i}{\text{median}}\, I_{ij}}\right) = \sum_{k=1}^{25} A_k(b_k),$$

giving position-dependent scores for each of the four bases. Figure 7A shows the distinct base-specific profiles. The destabilizing effects of in-sequence labels indicate possible advantages of labelling target sequences outside the probe binding regions. Overall, probe centres contributed most to overall binding.

GC-RMA (`www.bioconductor.org`) adopted the Naef and Magnasco model and in combination with data from non-specific hybridization predicts probe signals corrected for background and bias. Affinities obtained for G and T (Fig. 7B) showed somewhat different behaviour to that observed earlier, as can be expected for an empirical model, yet the predominant contribution to binding was again from the centres of the probes (Naef and Magnasco, 2003; Wu *et al.*, 2003).

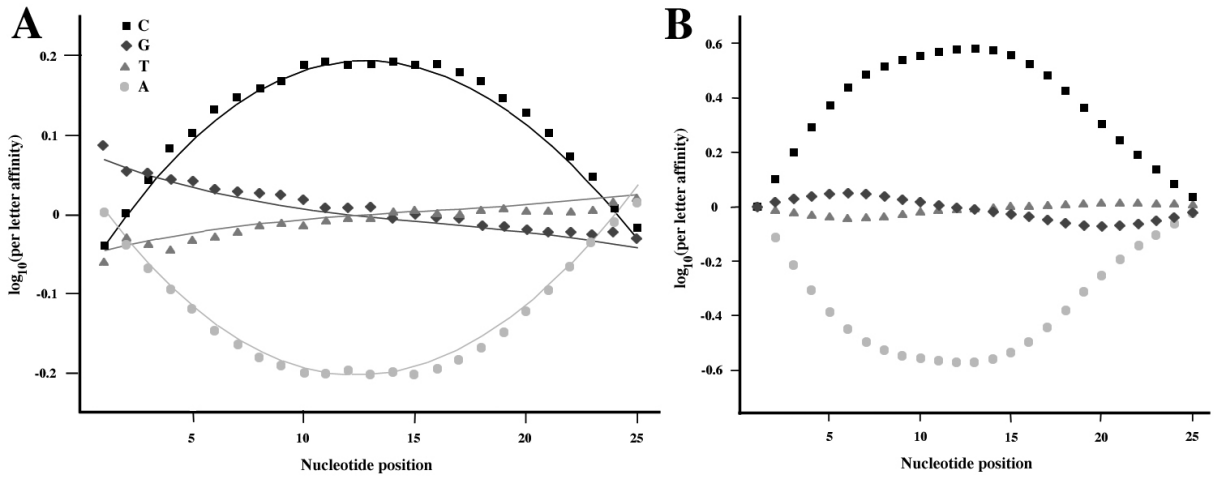


Figure 7: Position-specific affinities. (A) The position-specific affinities for each of the four bases from the model of Naef and Magnasco (2003). A/T and C/G asymmetries are due to labelled pyrimidines U/C impeding binding for A/G. Positions are in synthesis order, with 1 denoting the 3'-terminal attached to the chip. (Redrawn from Naef and Magnasco, 2003.) (B) The same model parameters but as obtained by Wu *et al.*(Wu *et al.*, 2003) Note the differences for G/T in comparison with panel (A). (Redrawn from Wu *et al.*, 2003.)

Common to all these approaches is the apparent attenuated influence of terminal probe regions. For the improvement of microarray manufacture and/or signal modelling, one wonders what could be its physical cause. At the 5'-end, one may well see the result of diminishing synthesis-yield through premature termination (Naef and Magnasco, 2003), while the reduced effect of bases in the 3'-terminal region could be due to steric hindrance of the solid support or overly dense population by short oligonucleotides (J. SantaLucia, Jr., *pers.comm.*, 2005).

Models for hybridization in solution

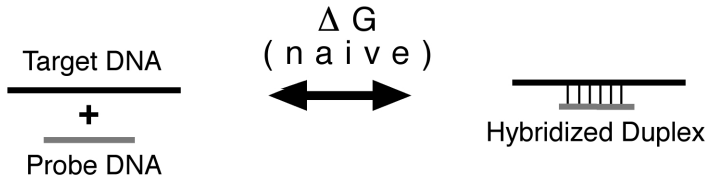
Even predicting hybridization in solution is a very complex modelling problem that is an area of active research (Dimitrov and Zuker, 2004; SantaLucia and Hicks, 2004). A hybridized complex or a folded structure actually assembles cooperatively in three-dimensional space, dynamically interacting with multiple other nucleic acid molecules and smaller molecules in solution as well as the solvent itself. In dependence on the temperature, what nucleic acids are present and at what concentrations, and the concentrations of salt-ions and other buffer components (like formamide), the nucleic acids can form a variety of heterogeneous complexes while at the same time folding within themselves. Therefore, to infer the concentration of a particular target transcript from microarray probe hybridization intensity, a fairly detailed understanding of the binding behaviour of the probe and its potential binding partners is required. To make modelling tractable, several approximations are necessary. A focus on secondary-structure elements is justified because tertiary structure is a much weaker, second-order effect. The strong Watson-Crick interactions further allow the 'discrete pairing approximation': positions in a sequence are either paired or not, rendering structure prediction suitable for dynamic-programming algorithms, which have

brought structure prediction for nucleic acids of up to 10,000 bases within reach for modern desktop computers (SantaLucia and Hicks, 2004).

The most common additional approximation in predicting microarray probe hybridization is looking at only one or two molecules at a time. The calculations for the hybridization of two molecules are typically much simplified further by assuming a 'two-state model', where the two molecules are either in a 'bound state', or not. To model the properties of the binding process under the two-state approximation, only the differences of thermodynamic parameters between the two states need to be calculated. For such computations, corresponding rules have been derived from the measurement of thermodynamic properties of selected nucleotides with purposefully designed sequences and structures, which contained basic reoccurring motifs (SantaLucia, 1998). An important part of this rule-set is formed by the Unified Watson-Crick Base-Pair Nearest-Neighbour parameters obtained by multiple-linear regression of measurements from several laboratories (SantaLucia, 1998) used by most microarray probe-design tools. State-of-the-art algorithms for the prediction of folding or hybridization structures of minimal and near-minimal energy use these parameters together with the corresponding rule-set for more complicated structural motifs like mismatched pairs, bulges, hairpins and various loops, and dangling ends (SantaLucia, 1998). Tools such as `mfold` (Zuker, 2003), `HyTher` (`http://ozone2.chem.wayne.edu/`), and `ViennaRNA` (Hofacker, 2003) can more accurately assess regions of non-target transcripts that are suspected of non-specific hybridization to a probe. Traditionally, these regions are selected by sequence-similarity and heuristics, however, the development of tools that can identify regions in a longer target DNA that will hybridize with a shorter probe by direct thermodynamic calculation (SantaLucia and Hicks, 2004) will soon make this inaccurate heuristic approximation unnecessary (M. Zuker, *pers.comm.*, 2004).

Importantly, the most recent advances in thermodynamic computation now go beyond two-state models in the prediction of hybridization behaviour (Fig. 8).

2 State Model



N-State Model

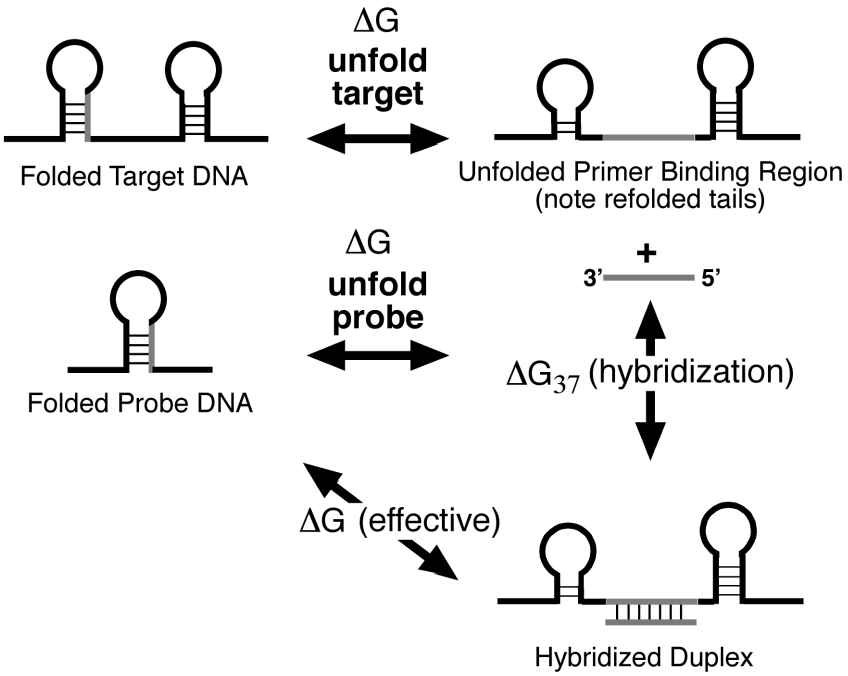


Figure 8: Multi-state coupled equilibria. A more realistic model allows much more accurate predictions of hybridization behaviour (Dimitrov and Zuker, 2004; Santa-Lucia and Hicks, 2004; Markham and Zuker, 2005). (Redrawn after SantaLucia and Hicks, 2004.)

Again, the same thermodynamic rule-set is used, but care has to be taken in order to avoid over-counting microstates: Although the experimental setup for the determination of the rule-set has been designed to minimize this effect, the parameters measured for the two-state model are for two *effective* states ('bound' and 'unbound'), each of which is actually a combination of multiple microstates. DNA Software's commercial OMP products account for this (SantaLucia and Hicks, 2004) and can provide correct multi-state modelling allowing multiple folding and binding events to be considered, including multiple simultaneous interactions *per* molecule. The improvements achievable by moving beyond two-state models can also be seen in DINAMelt, which for two molecules A and B models self-folding A_{self} and B_{self}, self-binding A-A and B-B, as well as hetero-duplex formation A-B (Dimitrov and Zuker, 2004; Markham and Zuker, 2005). DINAMelt calculates full partition sums (*i.e.*, accounting for all possible microstates), also taking care to avoid over-counting (N. Markham, *pers.comm.*, 2005). The multiple folding and binding events are modelled in competition to one another, giving temperature-dependent yields for each effective state.

While these methods are currently too slow to be used as primary screens of oligonucleotide-probe candidates during microarray design, they allow much more sophisticated evaluations of probe-sets.

Thermodynamic probe-design criteria

When aiming for uniform probe characteristics across a microarray, many probe-designs aim for uniform melting-temperatures T_m. These alone, however, only give information about the probes' behaviour at their respective melting-temperatures. Probes with the same T_m can behave quite differently at a reaction temperature $T_{hyb}<T_m$. For a given reaction temperature T_{hyb}, aiming for similar free-energies at T_{hyb} would hence actually result in more uniform hybridization (J. SantaLucia, Jr., *pers.comm.*, 2005). This can be improved on even further by accounting for competitive hybridization and actually calculating, for a target transcript, what proportion of molecules will be bound to its probe at T_{hyb}, aiming for uniformity across probes.

In screening probes, designs typically aim to avoid secondary structure. Clearly, strong secondary structure may render a probe inaccessible for its target. On the other hand, exploiting *competitive* hybridization, secondary structure can contribute much to the specific recognition of a probe's target. This is actually exploited by other experimental techniques like molecular beacons (Bonnet *et al.*, 1999). Using thermodynamic models for competitive hybridization, one can actually employ probe secondary-structure to adjust the level of specificity in target binding to that required (M. Zuker, *pers.comm.*, 2005), *e.g.*, highest for the discrimination of SNPs and highly similar targets, lower for transcript profiling transparently allowing for polymorphisms.

Outlook

With the increasing understanding of hybridization on microarrays, for many future microarray applications, the issue of probe design will yield to the task of probe-signal interpretation. Increasingly, modern methods leave little freedom in probe selection because probes have to target a very well defined region, *e.g.*, in probing particular gene regions to elucidate regulatory binding or splicing events. Many of these probes will show cross-hybridization or strong secondary structure, and probe-sets will display a wide spectrum of thermodynamic properties. To make the most of such data, a combination of experimental advances and sophisticated modelling will be instrumental. Repeated measurements under different hybridization conditions can, *e.g.*, discriminate specific from non-specific signal by exploiting hybridization kinetics (Dai *et al.*, 2002).

A further advance in quantitative microarray analysis has recently come with algorithms directly motivated by physical models. Application of the most elementary representation of surface adsorption, the Langmuir isotherm (Atkins and de Paula, 2004), could account for the nonlinearities observed at high signal intensity due to saturation of the probe with target molecules (Hekstra *et al.*, 2003) – not to be confused with saturation effects in the scanning of fluorescent images. Combination of such a Langmuir adsorption model with thermodynamic free-energy calculations was very successful, however, despite the significant improvements seen, systematic variation was still detectable in the data, highlighting the need for further studies (Held *et al.*, 2003).

The measurement process on microarrays is, over time, increasingly better understood and hence modelled. This correspondingly gives data that better reflect the true abundances of transcripts in samples, giving better detection characteristics in screens of samples for biological differences and providing a prerequisite for more sophisticated work in computational biology. While, overall, a lot of progress has been achieved, quantitative microarray analysis remains a challenging and active field of research.

Supplement

Further information is available at `www.flychip.org.uk/MethEnz2005/`.

Acknowledgments

We are grateful to Michael Zuker and John SantaLucia, Jr., for helpful discussions and advice. We also wish to thank Richard Auburn, Nicholas Markham, Lisa Meadows, and Andrew Thompson for helpful discussions and Jean-Marie Rouillard for the kind provision of the OligoArray 2.1 source code. The group of D. Kreil is funded by the Vienna Science and Technology Fund (WWTF), the Austrian Centre of Biopharmaceutical Technology (ACBT), Austrian Research Centres (ARC) Seibersdorf, and Baxter AG. The laboratory of S. Russell is funded by the Biotechnology and Biological Sciences Research Council, the UK Medical Research Council, and the Wellcome Trust.

References

Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25,** 3389-3402.

Atkins, P., and de Paula, J. (2004). "Atkins' Physical Chemistry." OUP, Oxford.

Auburn, R. P., Kreil, D. P., Meadows, L. A., Fischer, B., Matilla, S. S., and Russell, S. (2005). Robotic spotting of cDNA and oligonucleotide microarrays. *Trends Biotechnol* **23,** 374-9.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:,** 2242-2246.

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,** 185-93.

Bonnet, G., Tyagi, S., Libchaber, A., and Kramer, F. R. (1999). Thermodynamic basis of the enhanced specificity of structured DNA probes. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 6171-6.

Brown, T. A. (2002). "Genomes." BIOS Scientific, Oxford.

Buhler, S., Lagoja, I., Giegrich, H., Stengele, K. P., and Pfleiderer, W. (2004). New Types of Very Efficient Photolabile Protecting Groups Based upon the [2-(2-Nitrophenyl)propoxy]carbonyl (NPPOC) Moiety. *Helv. Chim. Acta* **87,** 620-659.

Castle, J., Garrett-Engele, P., Armour, C. D., Duenwald, S. J., Loerch, P. M., Meyer, M. R., Schadt, E. E., Stoughton, R., Parrish, M. L., Shoemaker, D. D., and Johnson, J. M. (2003). Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol* **4,** R66.

Chou, C. C., Chen, C. H., Lee, T. T., and Peck, K. (2004). Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res* **32,** e99.

Dai, H., Meyer, M., Stepaniants, S., Ziman, M., and Stoughton, R. (2002). Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays. *Nucleic Acids Res* **30,** e86.

Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D., and Siebert, P. D. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 6025-30.

Dimitrov, R. A., and Zuker, M. (2004). Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys J* **87,** 215-26.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nat Genet* **21,** 10-4.

Fotin, A. V., Drobyshev, A. L., Proudnikov, D. Y., Perov, A. N., and Mirzabekov, A. D. (1998). Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. *Nucleic Acids Res* **26,** 1515-1521.

Hanlon, S. E., and Lieb, J. D. (2004). Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr. Opin. Genet. Dev.* **14,** 697-705.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J. E., Snesrud, E., Lee, N., and Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. *Biotechniques* **29,** 548-50, 552-4, 556 passim.

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.

Held, G. A., Grinstein, G., and Tu, Y. (2003). Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 7575-7580.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. (2004). Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet* **36,** 1255-7.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res* **31,** 3429-31.

Holm, L., and Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14,** 423-9.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H., and Linsley, P. S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19,** 342-7.

Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20,** 1993-2003.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31,** 370-378.

Jobs, M., Fredriksson, S., Brookes, A. J., and Landergren, U. (2002). Effect of Oligonucleotide Truncation on Single-Nucleotide Distinction by Solid-Phase Hybridization. *Anal. Chem.* **74,** 199-202.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302,** 2141-4.

Johnston, R., Wang, B., Nuttall, R., Doctolero, M., Edwards, P., Lu, J., Vainer, M., Yue, H., Wang, X., Minor, J., Chan, C., Lash, A., Goralski, T., Parisi, M., Oliver, B., and Eastman, S. (2004). FlyGEM, a full transcriptome array platform for the Drosophila community. *Genome Biol.* **5**.

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28,** 4552-7.

Kreil, D. P., Auburn, R. P., Meadows, L., Russell, S., and Micklem, G. (2003). Quantitative microarray spot profile optimization: a systematic evaluation of buffer/slide combinations. *In* "German Conference in Bioinformatics", Munich, Germany.

Kreil, D. P., and Russell, R. R. (2005). There is no silver bullet--a guide to low-level data transforms and normalisation methods for microarray data. *Brief Bioinform* **6,** 86-97.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921.

Lausted, C., Dahl, T., Warren, C., King, K., Smith, K., Johnson, M., Saleem, R., Aitchison, J., Hood, L., and Lasky, S. R. (2004). POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biol.* **5,** R58.

Le Berre, V., Trevisiol, E., Dagkessamanskaia, A., Sokol, S., Caminade, A. M., Majoral, J. P., Meunier, B., and Francois, J. (2003). Dendrimeric coating of glass slides for sensitive DNA microarrays analysis. *Nucleic Acids Res* **31,** e88.

Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S. F., and Lee, C. (2004). Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res* **32,** e180.

Lee, S. I., and Batzoglou, S. (2003). Application of independent component analysis to microarrays. *Genome Biol* **4,** R76.

Li, C., and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 31-6.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17,** 282-3.

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,** 1675-80.

Luebke, K. J., Balog, R. P., and Garner, H. R. (2003). Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts. *Nucleic Acids Res* **31,** 750-758.

MacAlpine, D. M., and Bell, S. P. (2005). A genomic view of eukaryotic DNA replication. *Chromosome Res.* **13,** 309-326.

Markham, N. R., and Zuker, M. (2005). DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* **33,** W577-81.

Marko, N. F., Frank, B., Quackenbush, J., and Lee, N. H. (2005). A robust method for the amplification of RNA in the sense orientation. *BMC Genomics* **6,** 27.

McGall, G. H., Barone, A. D., Diggelmann, M., Fedor, S. P. A., Gentalen, E., and Ngo, N. (1997). The Efficiency of Light-Directed Synthesis of DNA Arrays on Glass Substrates. *J. Am. Chem. Soc.* **119,** 5081-5090.

Naef, F., and Magnasco, M. O. (2003). Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys* **68,** 011906.

Nuwaysir, E. F., Huang, W., Albert, T. J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J. P., Ballin, J., McCormick, M., Norton, J., Pollock, T., Sumwalt, T., Butcher, L., Porter, D., Molla, M., Hall, C., Blattner, F., Sussman, M. R., Wallace, R. L., Cerrina, F., and Green, R. D. (2002). Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res* **12,** 1749-55.

Pinkel, D., and Albertson, D. G. (2005). Comparative Genomic Hybridization. *Annu Rev Genomics Hum Genet*.

Pirrung, M. C. (2000). Production by Quantitative Photolithographic Synthesis of Individually Quality Checked DNA Microarrays. M. Beier, J.D. Hoheisel. *Chemtracts* **13,** 487-490.

Pirrung, M. C. (2002). How to Make a DNA Chip. *Angew. Chemie Int. Edn.* **41,** 1276-1289.

Ramakrishnan, R., Dorris, D., Lublinsky, A., Nguyen, A., Domanus, M., Prokhorova, A., Gieser, L., Touma, E., Lockner, R., Tata, M., Zhu, X., Patterson, M., Shippy, R., Sendera, T. J., and Mazumder, A. (2002). An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Res* **30,** e30.

Ratushna, V. G., Weller, J. W., and Gibas, C. J. (2005). Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics* **6,** 31.

Relogio, A., Schwager, C., Richter, A., Ansorge, W., and Valcarcel, J. (2002). Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res* **30,** e51.

Rouillard, J. M., Zuker, M., and Gulari, E. (2003). OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* **31,** 3057-62.

Saidi, S. A., Holland, C. M., Kreil, D. P., MacKay, D. J., Charnock-Jones, D. S., Print, C. G., and Smith, S. K. (2004). Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene* **23,** 6677-83.

SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 1460-1465.

SantaLucia, J., Jr., and Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* **33,** 415-40.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467-70.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 10614-9.

Shai, O. F., B. J. Morris, Q. D. Pan, Q. Misquitta, C. Blencowe, B. J. (2004). Probabilistic Inference of Alternative Splicing Events in Microarray Data. *In* "18th Annual Conference on Neural Information Processing Systems" (L. K. Saul, Y. Weiss, and L. Bottou, Eds.), Vol. 17. Neural Information Processing Systems Foundation.

Shchepinov, M. S., Case-Green, S. C., and Southern, E. M. (1997). Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res* **25,** 1155-61.

Shmulevich, I., Astola, J., Cogdell, D., Hamilton, S. R., and Zhang, W. (2003). Data extraction from composite oligonucleotide microarrays. *Nucleic Acids Res* **31,** e36.

Singh-Gasson, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnol.* **17,** 974-978.

Suchyta, S. P., Sipkovsky, S., Halgren, R. G., Kruska, R., Elftman, M., Weber-Nielsen, M., Vandehaar, M. J., Xiao, L., Tempelman, R. J., and Coussens, P. M. (2003). Bovine mammary gene expression profiling using a cDNA microarray enhanced for mammary-specific transcripts. *Physiol Genomics* **16,** 8-18.

Wang, H., Hubbell, E., Hu, J. S., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M. A., Ares, M., Kulp, D. C., and Haussler, D. (2003). Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics* **19 Suppl 1,** i315-22.

Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., and Spencer, F. (2003). A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *In* "Department of Biostatistics Working Papers". John Hopkins University, Baltimore, MD, USA.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method

addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30,** e15.

Zhang, L., Miles, M. F., and Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol* **21,** 818-21.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31,** 3406-15.