ChipCheck—A Program Predicting Total Hybridization Equilibria for DNA Binding to Small Oligonucleotide Microarrays

Karsten H. Siegmund,[†] Ulrich E. Steiner,*[,‡] and Clemens Richert*[,†,‡]

Institute for Organic Chemistry, University of Karlsruhe (TH), D-76128 Karlsruhe, Germany, and Department of Chemistry, University of Constance, D-78457-Konstanz, Germany

Presented here is the program ChipCheck that allows the computation of total hybridization equilibria for hybridization experiments involving small oligonucleotide arrays. The calculation requires the free energies of binding for all pairs of probes and targets as well as total strand concentrations and probe molecule numbers. ChipCheck has been tested computationally on microarrays with up to 100 spots and 42 target strands (4200 binding equilibria). It arrives at solutions through iterations employing the multidimensional Newton method. While currently running in simulation mode only, an extension of the approach to the exhaustive analysis of chip results is being outlined and may be implemented in the future. The output displays the extent of correct and cross hybridization both graphically and numerically. In principle, calculating total hybridization equilibria allows for eliminating noise from DNA chip results and thus an improvement in sensitivity and accuracy.

INTRODUCTION

Oligonucleotide microarrays, also known as "DNA chips", allow for massively parallel hybridization experiments.[1−4] High density chips now routinely present 100 000 and more probe sequences on their surface.[5] Gene expression profiling is the most common application for DNA chips,[6] but diagnostics,[7] SNP genotyping,[8,9] and sequencing by hybridization are applications that have also been envisioned.[10] Since the results of experiments with DNA microarrays generate large data sets, a fair number of algorithms have been developed to aid the handling and analysis of these data.[11,12] To review (or even fully list) them is beyond the scope of this article. However, it should be mentioned that many of them are proprietary and that their foci include visualization and quantification of data,[13] encoding biological knowledge,[14] feature extraction,[15] array size determination for sufficient specificity,[16] nonparametric statistical techniques,[17] statistical significance,[18,19] and eliminating artifacts due to saturation.[20] To the best of our knowledge, few of the algorithms attempt to solve the problem of cross hybridization between partially matched strands[21] or mismatched probe data,[22,23] and no publicly accessible algorithm undertakes the calculation of total hybridization equilibria.

In an ideal hybridization experiment, only duplexes between strands fully complementary according to the Watson−Crick base pairing rules should form, and all duplexes should have the same stability, so that signal intensities obtained from each spot can be directly translated into relative amounts of the complementary strand. Unfortunately, real life hybridization experiments suffer from several violations of these principles of an ideal chip assay. Partial complementarity between mismatched probes and target strands can lead to false positive results, and low G/C-content of duplexes can lead to false negative results. For a simple system, some of the errors resulting from the former violation are schematically shown in the lower part of Figure 1. A high concentration of strands with partial complementarity to a probe can lead to strong false positive signals. Binding of strands with partial complementarity can also enhance the signal on a spot that does engage in correct hybridization, leading to an incorrectly strong signal. Unfortunately, the extent to which these signal distortions occur at a given spot cannot be settled once and for all during chip design, since it depends on the concentration of all other strands with partial complementarity to a given probe and thus varies with the composition of a sample. In traditional Southern Blots, low level false positive signals are treated as noise, and the same is true for DNA chip results without special data treatment.[24] In fact, many of the programs developed for analyzing chip experiments try to pick correct from false signals via statistical analysis[25] or defining a cutoff level below which signals cannot be interpreted. The latter measure makes the detection of low abundance mRNA (often encoding critical, low copy number proteins) impossible in experiments with large arrays.

To deconvolute, rather than statistically sort the information gained through hybridizations on real life DNA chips, one needs to know the binding constants for the strands involved. The prediction of binding equilibria for oligonucleotide duplexes in solution[26] has become increasingly accurate in recent years. The scope of "nearest neighbor" predictions, i.e., predictions that do not just add increments to the total binding constant for each base pair but modulate the increment for each base pair according to what the neighboring base pair is, has been greatly expanded. Through a series of experimental studies involving UV-melting curves,

* Corresponding authors phone: ++49 (0)7531 88 3570; fax: +49 (0)-7531 88 3014; e-mail: Ulrich.Steiner@uni-konstanz.de (U.E.S.); phone: ++49 (0)721 608 2091; fax: ++49 (0)721 608 4825; e-mail: cr@rrg.uka.de (C.R.).

[†] University of Karlsruhe (TH).

[‡] University of Constance.

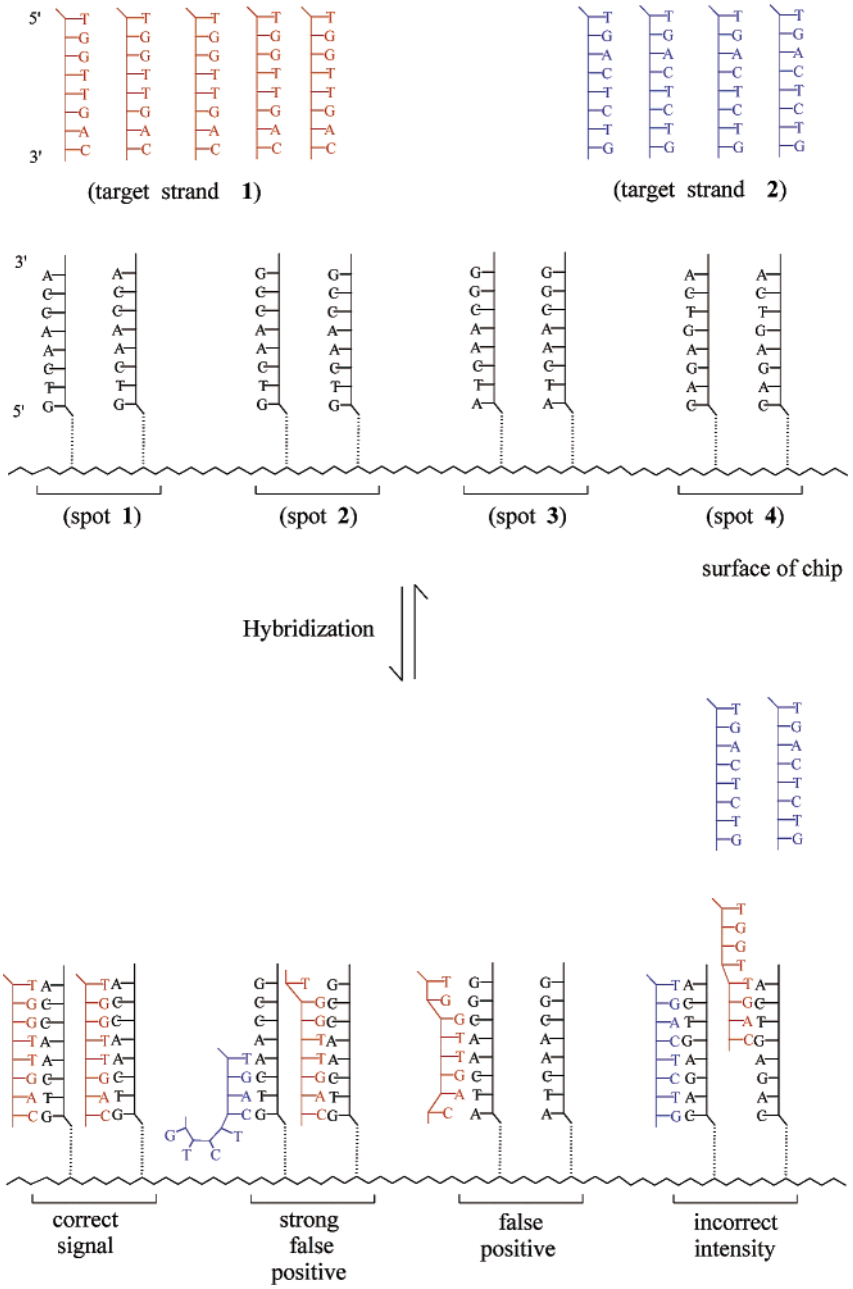


Figure 1. Cartoon of a hybridization on a DNA chip producing both correct and false signals.

calorimetry, and NMR-based structure elucidation, refined parameter sets have been generated.[27−31] Recently, this work has included generating parameter sets for duplexes that contain mismatches[32−34] and dangling ends.[35] The computer program HyTher,[36,37] by Drs. Peyret and SantaLucia, Wayne State University, for example, is accessible free of charge for calculations via the Internet. It calculates, among other things, hybridization thermodynamics for any given duplex and can also perform "primer walks".

We have an interest in base pairing fidelity and chemical approaches to improving it, particularly at the termini of duplexes. So-called "molecular caps" have been generated that, when covalently appended to the terminus of a hybridization probe, increase base pairing fidelity for the terminal and penultimate base pair.[38,39] Further, we have taken the molecular cap approach to custom-made oligonucleotide arrays.[40] When performing hybridization experiments in solution, we noticed that mismatches at the termini of short duplexes (which are more sensitive to mismatches than longer duplexes) induced melting point depressions as small as 1.9 °C.[38] In fact, there are reports of duplexes with mismatches that are more stable than their fully complementary counterparts.[41] This prompted us to scrutinize the fidelity of hybridization-based molecular recognition events on a more systematic level. Here we present the results of the initial phase of this project, the computer program "ChipCheck", which will be accessible free of charge on the Internet. The program calculates total hybridization equilibria for experiments with small DNA microarrays.

MATERIALS AND METHODS

General. The ChipCheck algorithm was implemented in the Perl scripting language that can be run on a variety of platforms. The program was developed and tested on personal computers running under Debian GNU/LINUX with Athlon AMD chipsets (1800+) and 512 MB RAM or under

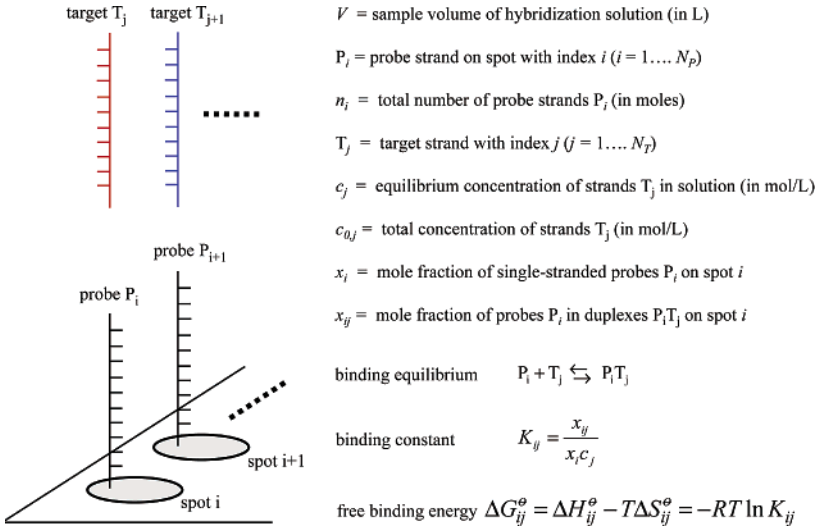


Figure 2. Compilation of terms and equations employed in this work.

SuSe Linux with an Intel PIII Processor and 256 MB RAM. Data for calculations can be submitted to ChipCheck via the Internet at the following URL http://chipcheck.chemie.uni-karlsruhe.de/chipcheck/ or http://chip.chemie.uni-karlsruhe.de/chipcheck/. The Web page and output of the scripts was written in HTML 4.01 Transitional and was checked using the tool available at http:// validator.w3.org/. All thermodynamic data employed for calculations were extracted semiautomatically from HyTher at http:// ozone2.chem.wayne.edu/ using Module 1. Possible duplex formation between target strands in solution and intramolecular folding were not considered in the standard version of ChipCheck. A trial version of an extended program termed "ChipCheck_Sol" that does include binding equilibria for cross hybridization between target strands can be accessed via the ChipCheck homepages.

Model Calculations. The sequences chosen for *Example 1* are derived from those tested earlier experimentally in these laboratories.[38,39] Mismatches were introduced at selected positions to test for cross hybridization, as described in the text (*vide infra*). For the hybridization buffer, conditions close to those favored by us[40] were entered. A total number of 10^{12} probe molecules of a given sequence was chosen, since, with the microarrays currently employed by us,[40] this number per spot was realistic, based on the probe density reported in other work with immobilized oligonucleotides, where high coverage and hybridization efficiencies had been achieved.[42] Numbers of probe strands that are typical for high-density microarrays (spot size approximately 50 μm) may, of course, also be entered when submitting data sets for calculations with ChipCheck. Experimental variables, like the concentration of monovalent ions and magnesium ions, do affect the thermodynamics of duplex formation and can be entered when generating thermodynamic parameters with HyTher.

For *Example 2*, sequences were provided by Drs. Pankratz and Bauer (Forschungszentrum Karlsruhe, http:// www.fzk.de/microarray) from a high-density chip set of 2 $\times$ 16 300 spots per chip for gene expression studies on zebrafish. Since the probe density on the surface has not been determined quantitatively, an immobilization of 20% of the strands spotted during the immobilization procedure (0.5 nL of 40 μM solutions) was assumed. In the simulation of the hybridization, only the regions of the target strands that are complementary to the probes were considered. The concentration of the target strands entered are based on the results of actual hybridization studies with the zebrafish chips. The concentrations were obtained by converting the mass values obtained from the analysis software of the collaborators (in ng cDNA) into molar concentrations using an average length of the cDNAs of 400 nt and an average mass of 308.8 Da per residue. The sequences of *Example 3* were chosen to contain both variants of 5′-TGGTTGAC-3′ and a number of fully unrelated sequences. Conditions for the virtual hybridization were the same as those for *Example 1*.

RESULTS

In a hybridization experiment with an oligonucleotide microarray, target strands in the hybridization solution form duplexes with probe strands immobilized on the chip surface. The terms used in this work, together with some fundamental relationships, are shown graphically in Figure 2. Throughout this work, the term "probe" is used for strands immobilized on a (virtual) chip surface, and the term "target" is used for strands introduced with the (virtual) solution to be analyzed. For calculating the total hybridization equilibria, the affinity of each target for each probe has to be known. The affinity is best expressed in the binding constant, which has a fundamental relationship with the free energy of binding. If standard entropy ($\Delta S°$) and standard enthalpy ($\Delta H°$) of duplex formation are known, then the binding constant can be calculated for any given temperature and hence the fraction of bound and single stranded probes and free target strands. The difficulty in computing the concentrations for a DNA chip experiment lies in the competition between the individual binding events. A mass distribution for the target strand has to be found that satisfies the constraints of all binding constants defined by the affinity for the probes on all spots and the total number of target molecules available.

As shown in Figure 2, the process of hybridization is described by the equilibrium

$$P_i + T_j \leftrightarrows P_iT_j$$

with an equilibrium constant K_{ij}, and employing the law of

mass action, one obtains

$$K_{ij} = \frac{x_{ij}}{x_i c_j} \tag{1}$$

Fortunately, standard entropy, standard enthalpy, and standard free enthalpy of binding for any given sequence combination can be calculated with programs such as HyTher,[36,37] which is accessible free of charge on the Internet at http://ozone2.chem.wayne.edu/, both for the solution case and with a linear correction for microchip setups. On the basis of observed linear correlations between the thermodynamic parameters of hybridization in solution and on a chip, it is possible to adapt the estimated thermodynamic parameters to the situation on the chip and obtain estimates of the binding constants K_{ij}. With these, a solution to the matrix of equations resulting from the competing binding equilibria had to be found.

If n_i, $c_{0,j}$, K_{ij}, and the volume (V) are given, a solution for x_{ij}, x_i, and c_i, i.e., ($N_p + N_p \times N_T + N_T$) variables total, can be developed based on the following equations:

$$x_{ij} = K_{ij} x_i c_j \tag{1'}$$

obtained by rearranging eq 1

$$\sum_j x_{ij} + x_i = 1 \tag{2}$$

representing the mass balance for probe P_i, and

$$\sum_i n_i x_{ij} + V c_j = V c_{0,j} \tag{3}$$

representing the mass balance for target T_j. Altogether the number of equations matches the number of variables and a unique solution can be obtained, in general. The x_i may be eliminated by combining eqs 1′ and 2:

$$x_i = \frac{1}{1 + \sum_j K_{ij} c_j} \tag{4}$$

Substituting (4) in eq 1′ yields

$$x_{ij} = \frac{K_{ij} c_j}{1 + \sum_l K_{il} c_l} \tag{5}$$

while rearrangement of eq 3 gives

$$c_j = c_{0,j} - \frac{\sum_i n_i x_{ij}}{V} \tag{6}$$

An iterative procedure cycling through eqs 5 and 6 to determine the correct x_{ij} and c_j was initially considered straightforward. It turned out, however, that this procedure is not convergent if V is so small that the c_j become much smaller than the $c_{0,j}$, i.e., if depletion effects of the sample solution become significant. Therefore, a mathematically more robust strategy had to be implemented to solve the equations in the general case. To this end, eq 5 for the x_{ij} was substituted into eq 3 to yield $c_j N_T$ coupled equations of the form

$$f_j \equiv c_j \left(1 + \frac{1}{V} \sum_i \frac{n_i K_{ij}}{1 + \sum_l K_{il} c_l} \right) - c_{0,j} = 0 \tag{7}$$

The main computational challenge was to develop a rugged iteration algorithm that would produce satisfactory, single solutions to the matrix of equations for all binding equilibria. Different iteration approaches were tested, and several sources of unsuccessful calculations were identified. When algorithms proceeding in increasingly small steps toward the solution were employed, the convergence criterion could be met, simply because the steps became too small to cause a noticeable change in the total result (the calculation would "fall asleep"). This problem was best solved by combining a few steps of conventional iteration with subsequent steps with the multidimensional Newton algorithm or using an optimized Newton algorithm exclusively. Another problem encountered was the generation of results that contained negative concentrations for target strands. In some instances, seemingly satisfactory solutions were obtained for all but a few strand concentrations, which were negative. This made too little difference to the total result to force the algorithm out of the negative concentration range for this very strand. To prevent negative concentrations, which are physically not meaningful, the program was modified to set all negative concentration values encountered to a specific fraction of the value of the last iteration.

Finally, a serious problem was encountered when longer DNA sequences were tested, such as those of *Example 2*, discussed below. The melting points of these sequences, more then 40 nucleotides in length, are approximately 90 °C in standard buffers, and hybridization temperatures more than 40 °C below this temperature lead to binding constants $>10^{40}$ M^{-1} in HyTher calculations. Accordingly, since one mole of probe strands is 6.022×10^{23} strands, and a small fraction of 1 L is typically used as the hybridization solution, not a single target molecule would remain in solution to fulfill the binding constant, if less than one full equivalent of the target strand was present in the experiment. Calculations trying to find the theoretical tiny fraction of a single molecule remaining unbound are inherently difficult to set up with reasonable convergence criteria and lead to results that are physically not meaningful. The overly large constants cause the numeric derivation of the functions, needed for the Newton method, to fail due to rounding. The small variations made to some of the smaller binding constants were lost by rounding during the calculations, causing the derivative to become zero and producing nonmeaningful results or no result at all. This problem was addressed by determining and coding the analytical derivative of the functions used to calculate the equilibrium.

Ideally, calculations (and experiments) with duplexes of such high stability should not be performed at temperatures too far away from the melting point. Instead they should be performed at a temperature where reasonable binding energies are found ($<10^{20}$ M^{-1}), even for long duplexes. It should be noted, in this context, that the ΔG values are not independent of temperature and have to be recalculated (in

Matrix of Individual Hybridizations

Sequence on Spot (3′ to 5′, left to right)	Signal for sequences bound (%) (Sequence of target strands: 5′ to 3′, top to bottom) tgcttgac	agcttgac	tagttgac	tgattgac	tgagtgac	tgcttgac	tgcttaac	tgcttgaa	Σ Signal (% coverage of spot)
accaactg	51.162	23.145	0.424	0.003	0.106	0.042	0.003	6.867	81.752
tccaactg	25.049	47.575	0.543	0.015	0.052	0.021	0.001	3.362	76.618
atcaactg	1.039	1.039	19.608	0.052	0.008	0.001	0.000	0.140	21.887
actaactg	0.242	0.107	0.021	30.687	0.038	0.000	0.000	0.032	31.128
acctactg	0.146	0.066	0.003	0.020	63.426	0.037	0.000	0.020	63.717
accatctg	0.499	0.226	0.004	0.000	0.050	49.168	0.064	0.067	50.077
accaattg	0.249	0.111	0.002	0.000	0.001	0.092	27.211	0.035	27.701
accaactt	11.042	4.995	0.092	0.001	0.023	0.009	0.051	43.341	59.553

Figure 3. Screenshot of a ChipCheck output of a calculation on *Example 1* showing the matrix of individual hybridizations between target and probe strands.

a trivial calculation) from the respective ΔH and ΔS values at every new temperature.

Three model calculations performed with ChipCheck will be discussed here briefly. The full list of free enthalpies of binding for the model calculation, as obtained from HyTher, are available as Supporting Information and may also be obtained from the ChipCheck homepage. The first model calculation (*Example 1*) involved DNA octamers derived from 5′-TGGTTGAC-3′ as the parent sequence immobilized on the virtual chip. These are very short sequences, where single mismatches can be expected to cause a more significant change in duplex stability than in longer sequences. To generate a demanding situation for binding selectivity, neighboring sequences on the virtual chip were chosen to differ by no more than a single base from each other. The strands offered in the virtual solution contained all eight complementary octamers.

ChipCheck offers four output options after a data set has been entered. The first two display the matrices of ΔG values and binding constants for each probe/target pair. The next two show the results from total hybridization calculations, either on the level of each individual hybridization or on the level of total signals for a given spot. Figure 3 shows an excerpt from a screenshot taken from the third of the four output options for the octamer case outlined above and equimolar concentrations of all target strands. Every row represents the binding to one spot, drawn out horizontally into the contributions from the individual target strands. The extent to which a given target covers the probe strands on the surface is displayed both numerically and via a color code where yellow codes for the fully covered spot and black for the absence of hybridization. It can be discerned that even at this, the most favorable case in terms of the relative concentration of the target strands, a measurable level of cross hybridization is observed for most combinations. Cross hybridization between mismatched strands is particularly strong when the mismatches are at or near the terminus (upper left-hand corner of the matrix).

Figure 4a shows the result of the simulation for the octamer case with equimolar concentrations of all target strands in a more condensed form, where all signals from

a)

Total Hybridization Results per Spot

(Total/Best Binder/∑Mishybridizations)

Probes on Spot (3′ to 5′, left to right)	Signal for sequences bound (%) Σ Signal (% coverage of spot)	Signal for strongest hybridisation (%)	Σ Signal of other hybridisations (%)
accaactg	81.75	51.16 tggttgac	30.59
tccaactg	76.62	47.57 aggttgac	29.04
atcaactg	21.89	19.61 tagttgac	2.28
actaactg	31.13	30.69 tgattgac	0.44
acctactg	63.72	63.43 tggatgac	0.29
accatctg	50.08	49.17 tggtagac	0.91
accaattg	27.70	27.21 tggttaac	0.49
accaactt	59.55	43.34 tggttgaa	16.21

b)

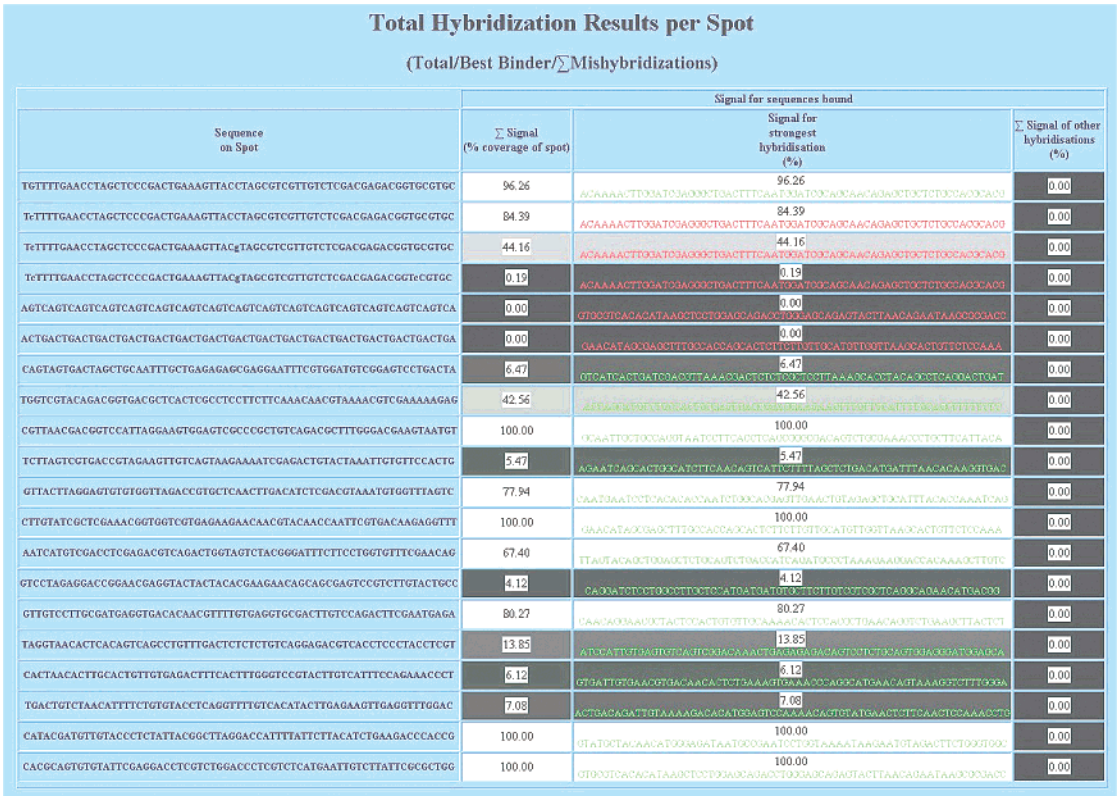
Total Hybridization Results per Spot

(Total/Best Binder/∑Mishybridizations)

Probes on Spot (3′ to 5′, left to right)	Signal for sequences bound (%) Σ Signal (% coverage of spot)	Signal for strongest hybridisation (%)	Σ Signal of other hybridisations (%)
accaactg	81.73	51.22 tggttgac	30.51
tccaactg	76.60	47.61 aggttgac	29.00
atcaactg	21.84	19.62 tagttgac	2.22
actaactg	0.63	0.35 tgattgac	0.28
acctactg	0.91	0.40 tgatgac	0.52
accatctg	50.05	49.19 tggtagac	0.86
accaattg	27.70	27.21 tggttaac	0.49
accaactt	59.54	43.35 tggttgaa	16.19

Figure 4. Screenshots of ChipCheck outputs of calculations on *Example 1* showing total hybridizations per spot (a) with equimolar amounts of the target strands and (b) with target strands 4 and 5 depleted by 3 orders of magnitude.

strands other than the most tightly binding one are added and presented in the rightmost column of the table. Figure 4b shows the results of a simulation for the same probe/target combination where the concentration of the strands complementary to the fourth and fifth probe was dropped



Total Hybridization Results per Spot

(Total/Best Binder/∑Mishybridizations)

Sequence on Spot	Signal for sequences bound: ∑ Signal (% coverage of spot)	Signal for strongest hybridisation (%)	∑ Signal of other hybridisations (%)
TGTTTTGAACCTAGCTCCCGACTGAAAGTTACCTAGCGTCGTTGTCTCGACGAGACGGTGCGTGC	96.26	96.26	0.00
TcTTTTGAACCTAGCTCCCGACTGAAAGTTACCTAGCGTCGTTGTCTCGACGAGACGGTGCGTGC	84.39	84.39	0.00
TcTTTTGAACCTAGCTCCCGACTGAAAGTTACgTAGCGTCGTTGTCTCGACGAGACGGTGCGTGC	44.16	44.16	0.00
TcTTTTGAACCTAGCTCCCGACTGAAAGTTACgTAGCGTCGTTGTCTCGACGAGACGGTcCGTGC	0.19	0.19	0.00
AGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCA	0.00	0.00	0.00
ACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGACTGA	0.00	0.00	0.00
CAGTAGTGACTAGCTGCAATTTGCTGAGAGAGCGAGGAATTTCGTGGATGTCGGAGTCCTGACTA	6.47	6.47	0.00
TGGTCGTACAGACGGTGACGCTCACTCGCCTCCTTCTTCAAACAACGTAAAACGTCGAAAAAGAG	42.56	42.56	0.00
CGTTAACGACGGTCCATTAGGAAGTGGAGTCGCCCGCTGTCAGACGCTTTGGGACGAAGTAATGT	100.00	100.00	0.00
TCTTAGTCGTGACCGTAGAAGTTGTCAGTAAGAAAATCGAGACTGTACTAAATTGTGTTCCACTG	5.47	5.47	0.00
GTTACTTAGGAGTGTGTGGTTAGACCGTGCTCAACTTGACATCTCGACGTAAATGTGGTTTAGTC	77.94	77.94	0.00
CTTGTATCGCTCGAAACGGTGGTCGTGAGAAGAACAACGTACAACCAATTCGTGACAAGAGGTTT	100.00	100.00	0.00
AATCATGTCGACCTCGAGACGTCAGACTGGTAGTCTACGGGATTTCTTCCTGGTGTTTCGAACAG	67.40	67.40	0.00
GTCCTAGAGGACCGGAACGAGGTACTACTACACGAAGAACAGCAGCGAGTCCGTCTTGTACTGCC	4.12	4.12	0.00
GTTGTCCTTGCGATGAGGTGACACAACGTTTTGTGAGGTGCGACTTGTCCAGACTTCGAATGAGA	80.27	80.27	0.00
TAGGTAACACTCACAGTCAGCCTGTTTGACTCTCTCTGTCAGGAGACGTCACCTCCCTACCTCGT	13.85	13.85	0.00
CACTAACACTTGCACTGTTGTGAGACTTTCACTTTGGGTCCGTACTTGTCATTTCCAGAAACCCT	6.12	6.12	0.00
TGACTGTCTAACATTTTCTGTGTACCTCAGGTTTTGTCACATACTTGAGAAGTTGAGGTTTGGAC	7.08	7.08	0.00
CATACGATGTTGTACCCTCTATTACGGCTTAGGACCATTTTATTCTTACATCTGAAGACCCACCG	100.00	100.00	0.00
CACGCAGTGTGTATTCGAGGACCTCGTCTGGACCCTCGTCTCATGAATTGTCTTATTCGCGCTGG	100.00	100.00	0.00

Figure 5. Screenshot of an output of a total hybridization calculation performed on *Example 2*. Total hybridizations per spot are shown with low level binding by cross-hybridizing strands highlighted in red. Mismatched nucleotides in probe sequences were entered as lower case letters, solely to highlight their position.

by 3 orders of magnitude. It can be discerned that now the hybridization of strands other than the fully complementary one dominates the distribution of sequences bound to these spots. To visually highlight this fact, the sequence of these strongest binding mismatched strands are displayed in red, rather than the green reserved for the case where fully complementary strands dominate the binding at an individual spot.

A model calculation involving 20 longer probe strands (65mers) and 15 target strands of the same length was also performed (*Example 2*). The target strands all have fully complementary partners on the virtual chip, but five of the probes (spots 1−5) do not have fully complementary target strands in the virtual solution. The hybridization temperature was set to 75 °C, i.e., well under the average melting point of the fully complementary duplexes. Also, the concentrations of target strands were chosen based on the results of an actual chip hybridization experiment performed by Drs. Pankratz and M. Bauer, our collaborators, with a high-density chip. The condensed form of the results from the ChipCheck simulation is shown in Figure 5. In all cases, the hybridization is dominated by one sequence in solution. For spots 2−6, these are mismatched sequences, since the solution is devoid of complementary targets. Spots 2−4 bear sequences similar to that of spot 1, except that one, two, or three mismatches were introduced. Spots 5 and 6 are repeats of tetramers AGCT and ACTG, respectively, and show hybridization of the nonsequence-related target strands present on a very low level (<0.01% coverage of immobilized strands).

To test the performance of ChipCheck on a larger data set, a simulation was then run on 100 virtual spots and 42 target strands. Again, octamers were used, with part of the sequences being derived from 5′-TGGTTGAC-3′. The run time for this example (4200 binding equilibria) on a personal computer with an AMD chipset (1800+) with 512 MB RAM was 3.5 s, and results for the individual hybridization matrix were obtained in under 10 s as a remote user, including the loading of the html page. The result of the simulation is shown graphically in Figure 6. The numerical information was left out to improve the visibility but may be obtained on the HyTher homepage. Even at the limited information level of this condensed form of the output, it can be discerned that noticeable false positive results are found in the lower half of the matrix, where no pairs of complementary probes/targets are found. Once these hot spots for false positives are identified, they may be eliminated by using other probes or, where appropriate, through chemical modifications.[40]

As a very first example of a simulation of experimental hybridization results, we have performed a ChipCheck calculation for the microarray data presented in Figure 2 of ref 40. For this, HyTher thermodynamic values with linear correction for microchips were used for the unmodified strands, and the $\Delta\Delta H$ and $\Delta\Delta S$ values for the chl modification reported in Table 4 of ref 38 added to the ΔH and ΔS values of the unmodified, perfectly matched duplex to obtain values for the modified probe. Further, a density of 3.6×10^{11} molecules/cm^2 was assumed, based on literature values for aldehyde-bearing glass slides and amino-terminal DNA.[43,44] A hybridization temperature of 60 °C was entered to compensate for the heating during the sonication step in the experimental procedure. Figure 7 shows a comparison of the experimental results (total fluorescence intensity integrals for each spot pair) and the results obtained by simulation. It can be discerned that the ChipCheck calculation slightly under-

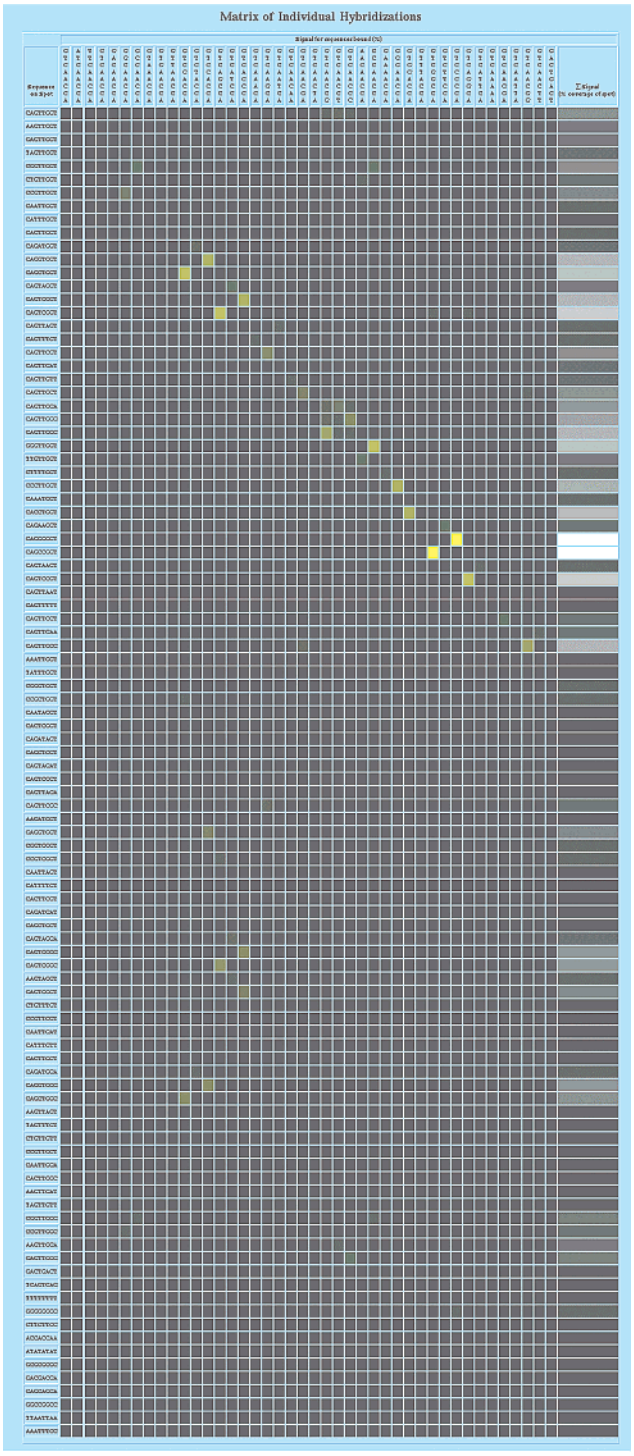


Figure 6. Screenshot of the ChipCheck output of a total hybridization calculation on *Example 3* featuring 100 probe spots and 42 target strands. To improve the readability, the numerical information representing the extent of spot coverage was omitted. The full data set, with better readability, can be created on the ChipCheck Internet page using the *Example 3* file.

estimates the extent of cross hybridization observed for the probe strand with a single mismatch (MM).

As an outlook, the following provides a brief discussion of the possibility of reaching a further goal, i.e., assigning gene expression levels and specific probe-gene binding constants from the signal data obtained from an actual chip experiment (analysis instead of simulation). We assume that the signal intensities S_i from the probes i have been

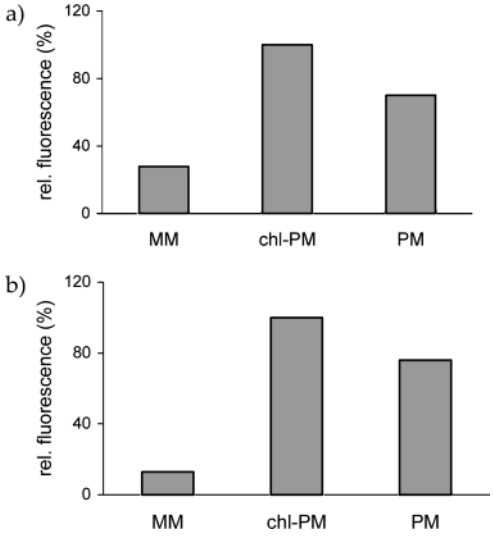


Figure 7. Comparison of experimental (a) and simulated (b) hybridization results with the sequences 5′-TAGTTGACTGCGAT-3′ (MM), 5′-chl-TGGTTGACTGCGAT-3′ (chl-PM), and 5′-TGGTTGACTGCGAT-3′ (PM) as immobilized probes and 5′-Cy3-ATCGCAGTCAACCA-3′ as target. The integration area of the fluorescence values reported in Figure 2 of ref 40 are plotted in (a), whereas (b) shows the relative fluorescence intensities expected based on the ChipCheck simulation. See ref 40 for experimental details.

normalized such that they represent the total hybridization of all targets to a particular probe i.

$$\sum_j x_{ij} = S_i \tag{8}$$

Eliminating x_{ij} via eq 5 yields

$$S_i = \sum_j \frac{K_{ij} c_j}{1 + \sum_l K_{il} c_l} \tag{9}$$

For each chip experiment, expressions 7 and 9 represent a total of $N_T + N_P$ equations.

If the binding constants K_{ij} are known, $2N_T$ equations are needed to determine all the c_j and $c_{0,j}$. If $N_P > N_T$, the data from one chip yield sufficient information to solve the problem and the redundancy in information can be used to check for the statistical significance of the parameter values.

If the binding constants K_{ij} are unknown, one would need $2N_T + N_T N_P$ equations to determine the c_j, $c_{0,j}$, and all K_{ij}. However, only $N_T + N_P$ equations are available with the data from one chip. If some variation in the relative concentrations of the genes in the sample can be achieved, e.g. by changing the conditions of gene expression, additional information is gained. With N_C as the number of chips used to evaluate N_C different samples, i.e., samples that differ systematically in the $c_{o,j}$ of the genes and are not just replicates of the same experiment, the number of parameters $\{c_{0,j}, c_j\}_{1\ N_c}$, K_{ij} increases to $2N_T N_C + N_T N_P$. However, the total information from the probe signals increases faster. With N_C different chip experiments on N_C different samples, the number of equations becomes $N_C(N_T + N_P)$. Hence for

$$N_C > \frac{N_P N_T}{N_P - N_T} \gtrsim N_T \tag{10}$$

the number of equations exceeds the number of unknown parameters. If we consider the typical practical situation in which sets of N_S ($\lesssim$20) probes are selected to match single genes and if we neglect cross hybridization of the probes in one set to other genes, then each set of probes represents an isolated problem of N_S probes and $N_T = 1$ targets. In this case it follows from eq 10 that as few as $N_C \geq 2$ different chip experiments are sufficient to determine all the desired parameters.

DISCUSSION

In its current form, ChipCheck can be employed to evaluate the performance to be expected theoretically of oligonucleotide arrays in hybridization experiments of moderate complexity. Thus, it may assist the design of new microarrays. This may be particularly important for the design of arrays with a high demand on fidelity, such as chips for diagnosis, SNP genotyping, and sequencing by hybridization. For these applications, resolving a single nucleotide with good signal-to-noise is often critical. The simulations will provide information on the fidelity to be expected in a given hybridization and allow optimizations by eliminating probes that can be expected to give high levels of false positive or false negative results.

A lack of reproducibility is a very serious concern in the DNA chip community.[45,46] The difficulties in obtaining reliable, reproducible results from microarray hybridization experiments have led to some disenchantment.[47] Due to the complexity of the DNA chip hybridization experiments, it is difficult to test all factors (including strand concentration, probe density and integrity, counterions, salt concentration, temperature) affecting absolute and relative signal intensities experimentally. ChipCheck may be employed to study these effects *in silico*, allowing for more efficient use of the experimental resources and time. Most DNA chips currently employed are larger than what has been tested with ChipCheck thus far. However, it is often the custom designed, focused chips of smaller spot numbers that are employed when fidelity is particularly important. The examples tested demonstrate that ChipCheck can function properly for such examples.

The results of the model calculations presented above (*Examples 1* and *2*) already provide some insight. In the more demanding case, the hybridization where single mismatches have to be resolved (*Example 1*), the ChipCheck results demonstrate how severe the problem caused by false positives can be. If equimolar concentrations of all strands complementary to the probes on the surface are present, up to 38% of the strands bound to a given spot are from mishybridizations, even at equimolar concentrations of all target strands. If two target strands are present in much lower copy numbers, their signal is swamped out by the competing mismatch-bearing strands present in higher concentrations, leading to incorrectly high signals. In the case of the 65mers (*Example 2*), the fidelity is excellent for the strands without sequence homology, but spots with 1−3 mismatches again show strong signals (0.2−84.4% of probes bound, spots 2−4).

Currently, one possible weakness of the approach presented here may lie in the limited accuracy of thermodynamic data for DNA duplexes. While nearest neighbor calculations themselves necessarily involve a number of approximations, it has also been noted that the van't Hoff assumption of a stable ΔH over a range of temperatures may be violated for hybridization experiments performed far from the melting point.[48] Next, the size of the data sets (number of spot and target sequences per hybridization experiment) employed in simulations thus far has been small. Data sets closer to typical high density DNA chips will necessarily lead to increased run times and greater challenges for successful iterations. Since MELTING,[49] a program providing similar capabilities as HyTher, is freely available on the Internet (http://bioweb.pasteur.fr/seqanal/interfaces/melting.html), future versions of ChipCheck may contain a module for generating the thermodynamic parameters of duplexes itself, avoiding a bottleneck due to semiautomatic downloading of values from HyTher. In this upgraded version of ChipCheck, sequences and hybridization conditions would be the only data to submit to get a total hybridization result.

Equation 1 is actually the equivalent of the Langmuir adsorption isotherm and one should be aware of the high degree of idealization involved in this model. More realistic adsorption models, such as the Sips model have been suggested,[50] and may be easily implemented. However, the reduction in the binding constants resulting from the immobilization of probes in spots should be similar for all spots, so that the relative hybridization efficiencies determined are useful for practical considerations. The absolute values for the coverage of spots at any given temperature should be treated with great caution and confirmed experimentally.

As mentioned above, ChipCheck may be extended to include the capability to analyze rather than simulate the outcome of hybridizations on microarrays. Currently, the expression level of rarely expressed genes is often lost during the correction of chip results for background "noise".[51] For optimized chips, where many of the spots are used as internal controls, a dynamic range for linear detection of approximately 500-fold has been reported in "routine use" mode.[52] Data analysis involving total hybridization calculations for all matched duplexes and those formed through cross hybridization can be expected to yield more reliable results and higher sensitivity, not to mention more rigorous tests for statistical significance. With nearest neighbor parameters recursively optimized via tightly controlled experimental chip results, analyzed via the total hybridization approach, simulations and experiments should converge to high accuracy. Clearly, the very limited tests performed here provide little more than an approach to treating total hybridization equilibria. With further refined algorithms, most of the "noise" in DNA chip experiments should be removable via filtering. The expression of genes encoding low copy number proteins is critical, e.g. for cell signaling or regulation may then be detected on large microarrays with good fidelity and sensitivity; ... the door is just a crack open!

ACKNOWLEDGMENT

The authors thank M. Pankratz and M. Bauer (Forschungszentrum Karlsruhe, www fzk.de/microarray) for sharing experimental data and helpful discussions and A. Rapp and P. Grünefeld for help with computer issues. This work was supported in part by DFG (Grants RI 1063/2-1 and FOR 434).

Supporting Information Available: Data sets for model calculations (also available on the ChipCheck homepage (http://chipcheck.chemie.uni-karlsruhe.de/chipcheck/ or http://chip.chemie.uni-karlsruhe.de/chipcheck/)). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Fodor, S. P. A.; Read, J. L.; Pirrung, M. C.; Stryer, L.; Lu, A. T.; Solas, D. Light-directed, spatially addressable parallel chemical synthesis. *Science* **1991**, *251*, 767−773.

(2) DeRisi, J. L.; Iyer, V. R.; Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **1997**, *278*, 680−686.

(3) Southern, E. M.; Mir, K.; Shchepinov, M. Molecular interactions on microarrays. *Nat. Genet.* **1999**, *21*, 5−9.

(4) Lockhardt, D. J.; Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **2000**, *405*, 827−836.

(5) Duggan, D. J.; Bittner, M.; Chen, Y.; Meltzer, P.; Trent, J. M. Expression profiling using cDNA microarrays. *Nat. Genet.* **1999**, *21*, 10−14.

(6) Hedge, P.; Qi, R.; Abernathy, K.; Gay, S.; Dharap, S.; Gaspard, R.; Hughes, J. E.; Snesrud, E.; Lee, N.; Quackenbush, J. A concise guide to cDNA microarray analysis. *BioTechniques* **2000**, *29*, 548−562.

(7) Pomeroy, S. L.; Tamayo, P.; Gaasenbeek, M.; Sturla, L. M.; Angelo, M.; McLaughlin, M. E.; Kim, J. Y. H.; Goumnerova, L. C.; Black, P. M.; Lau, C.; Allen, J. C.; Zagzag, D.; Olson, J. M.; Curran, T.; Wetmore, C.; Biegel, J. A.; Poggio, T.; Mukherjee, S.; Rifkin, R.; Califano, A.; Stolovitzky, G.; Louis, D. N.; Mesirov, J. P.; Lander, E. S.; Golub, T. R. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **2002**, *415*, 436−442.

(8) Spencer, J.; Krugylak, L.; Stein, L.; Hsie, L.; Topalogou, T.; Hubbell, E.; Robinson, E.; Mittmann, M.; Morris, M. S.; Shen, N.; Kilburn, D.; Rioux, J.; Nusbaum, C.; Rozen, S.; Hudson, T. J.; Lipshutz, R.; Chee, M.; Lander, E. S. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **1998**, *280*, 1077−1082.

(9) Huber, M.; Mundlein, A.; Dornstauder, E.; Schneeberger, C.; Tempfer, C. B.; Mueller, M. W.; Schmidt, W. M. Accessing single nucleotide polymorphisms in genomic DNA by direct multiplex polymerase chain reaction amplification on oligonucleotide microarrays. *Anal. Biochem.* **2002**, *303*, 25−33.

(10) Southern, E. M. DNA chips: analysing sequence by hybridization to oligonucleotide on a large scale. *Trends Genet.* **1996**, *12*, 110−115.

(11) Quackenbush, J. Computational analysis of microarray data. *Nature Rev. Genet.* **2001**, *2*, 418−427.

(12) Ringner, M.; Peterson, C.; Khan, J. Analyzing array data using supervised methods. *Pharmacogenomics* **2002**, *3*, 403−415.

(13) Strehlow, D. Software for quantitation and visualization of expression array data. *BioTechniques* **2000**, *29*, 118−121.

(14) Moloshok, T. D.; Klevecz, R. R.; Grant, J. D.; Manion, F. J.; Speier, W. F.; Ochs, M. F. Application of Bayesian Decomposition for analysing microarray data. *Bioinformatics* **2002**, *18*, 566−575.

(15) Schadt, E. E.; Li, C.; Ellis, B.; Wong, W. H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* **2001**, *S37*, 120−125.

(16) Van Dam, R. M.; Quake, S. R. Gene expression analysis with universal n-mer arrays. *Genome Res.* **2002**, *12*, 145−152.

(17) Lazaridis, E. N.; Sinibaldi, D.; Bloom, G.; Mane, S.; Jove, R. A simple method to improve probe set estimates from oligonucleotide arrays. *Math. Biosci.* **2002**, *176*, 53−58.

(18) Zhang, L.; Wang, L.; Ravindranathan, A.; Miles, M. F. A new algorithm for analysis of oligonucleotide arrays: Application to expression profiling in mouse brain regions. *J. Mol. Biol.* **2002**, *317*, 225−235.

(19) Pan, W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **2002**, *18*, 546−554.

(20) Hsiao, L. L.; Jensen, R. V.; Yoshida, T.; Clark, K. E.; Blumenstock, J. E.; Gullans, S. R. Correcting for signal saturation errors in the analysis of microarray data. *BioTechniques* **2002**, *32*, 330.

(21) Li, C.; Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 31−36.

(22) Chu, T. M.; Weir, B.; Wolfinger, R. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math. Biosci.* **2002**, *176*, 35−51.

(23) Lemon, W. J.; Palatini, J. J. T.; Krahe, R.; Wright, F. A. Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics* **2002**, *18*, 1470−1476.

(24) Naef, F.; Hacker, C. R.; Patil, N.; Magnasco, M. Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.* **2002**, *3*, R0018.

(25) Brody, J. P.; Williams, B. A.; Wold, B. J.; Quake, S. R. Significance and statistical errors in the analysis of DNA microarray data. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12975−12978.

(26) Breslauer, K. J.; Frank, R.; Blsker, H.; Marky, L. A. Predicting DNA duplex stability from the base sequences. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 3746−3750.

(27) Marky, L. A.; Breslauer, K. J. Calorimetric determination of base-stacking enthalpies in double-helical DNA molecules. *Biopolymers* **1982**, *21*, 2185−2194.

(28) SantaLucia, J.; Kierzek, R.; Turner, D. H. Stabilities of consecutive A·C, C·C, G·G, U·C, and U·U mismatches in RNA internal loops − evidence for stable hydrogen-bonded U·U and $C\cdot C^{+}$ pairs. *Biochemistry* **1991**, *30*, 8242−8251.

(29) Marky, L. A.; Breslauer, K. J. Calculating thermodynamic data from transitions of any molecularity from equilibrium melting curves. *Biopolymers* **1987**, *27*, 1601−1620.

(30) Breslauer, K. J. Extracting thermodynamic data from equilibrium melting curves for oligonucleotide order−disorder transitions. *Methods Enzymol.* **1995**, *259*, 221−242.

(31) SantaLucia, J.; Allawi, H. T.; Seneviratne, P. A. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **1996**, *35*, 3555−3562.

(32) Allawi, H. T.; SantaLucia, J. Thermodynamics of internal C. T mismatches in DNA. *Nucleic Acids Res.* **1998**, *26*, 2694−2701.

(33) Allawi, H. T.; SantaLucia, J. Nearest-neighbor thermodynamics of internal A·C mismatches in DNA: Sequence dependence and pH effects. *Biochemistry* **1998**, *37*, 9435−9444.

(34) Peyret, N.; Seneviratne, P. A.; Allawi, H. T.; SantaLucia, J. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A·A, C·C, G·G, and T·T mismatches. *Biochemistry* **1999**, *38*, 3468−3477.

(35) Bommarito, S.; Peyret, N.; SantaLucia, J. Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res.* **2000**, *28*, 1929−1934.

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

(37) Peyret, N.; Seneviratne, P. A.; Allawi, H. T.; SantaLucia, J. Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A·A, C·C, G·G, and T·T Mismatches. *Biochemistry* **1999**, *38*, 3468−3477.

(38) Bleczinski, C. F.; Richert, C. Steroid-DNA interactions increasing stability, sequence-selectivity, DNA/RNA discrimination, and hypochromicity of oligonucleotide duplexes. *J. Am. Chem. Soc.* **1999**, *121*, 10889−10894.

(39) Mokhir, A. A.; Richert, C. Synthesis and monitored selection of 5′-nucleobase-capped oligodeoxyribonucleotides. *Nucleic Acids Res.* **2000**, *28*, 4254−4265.

(40) Dombi, K. L.; Griesang, N.; Richert, C. Oligonucleotide arrays from aldehyde-bearing glass with coated background. *Synthesis* **2002**, 816−824.

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

(42) Herne, T. M.; Tarlov, M. J. Characterization of DNA probes immobilized on gold surfaces. *J. Am. Chem. Soc.* **1997**, *119*, 8916−8920.

(43) Pirrung, M. C. How to Make a DNA Chip. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 1276−1291.

(44) Zammatteo, N.; Jeanmart, L.; Hamels, S.; Courtois, S.; Louette, P.; Hevesi, L.; Remacle, J. Comparison between different strategies of covalent attachment of DNA to glass surfaces to build DNA microarrays. *Anal. Biochem.* **2000**, *280*, 143−150.

(45) Piper, M. D. W ; Daran-Lapujade, P.; Bro, C ; Regenberg, B.; Knudsen, S.; Nielsen, J.; Pronk, J. T. Reproducibility of oligonucleotide microarray transcriptome analyses − An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **2002**, *277*, 37001−37008.

(46) Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C. A.; Causton, H. C.; Gaasterland, T.; Glenisson, P.; Holstege, F. C. P.; Kim, I. F.; Markowitz, V.; Matese, J. C.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. Minimum information about a microarray experiment (MIAME) − toward standards for microarray data. *Nature Genet.* **2001**, *29*, 365−371.

(47) Simon, R.; Radmacher, M. D.; Dobbin, K.; McShane, L. M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **2003**, *95*, 14−18.

(48) Vesnaver, G.; Breslauer, K. J. The contribution of DNA single-stranded order to the thermodynamics of duplex formation. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 3569−3573.
(49) Le Novère, N. MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics* **2001**, *17*, 1226−1227.
(50) Peterson, A. W.; Wolf, L. K.; Georgiadis, R. M. Hybridization of mismatched or partially matched DNA at surfaces. *J. Am. Chem. Soc.* **2002**, *124*, 14601−14607.
(51) Tran, P. H.; Peiffer, D. A.; Shin, Y.; Meek, L. M.; Brody, J. P.; Cho, K. W. Y. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res.* **2002**, *30*, e54.
(52) Lipshutz, R. J.; Fodor, S. P. A.; Gingeras, T. R.; Lockhart, D. J. High-density synthetic oligonucleotide arrays. *Nature Genet.* **1999**, *21s*, 20−24.