Published online February 2, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 2 ***681–690***
DOI: 10.1093/nar/gkh196

Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray

Inhan Lee*, Alan A. Dombkowski[1] and Brian D. Athey

Michigan Center for Biological Information and Department of Psychiatry, University of Michigan, 3600 Green Court, Suite 700, Ann Arbor, MI 48105 and [1]Institute of Environmental Health Science, Wayne State University, Detroit, MI 48201, USA

Received October 3, 2003; Revised November 21, 2003; Accepted December 8, 2003

ABSTRACT

Sequence-specific oligonucleotide probes play a crucial role in hybridization techniques including PCR, DNA microarray and RNA interference. Once the entire genome becomes the search space for target genes/genomic sequences, however, cross-hybridization to non-target sequences becomes a problem. Large gene families with significant similarity among family members, such as the P450s, are particularly problematic. Additionally, accurate single nucleotide polymorphism (SNP) detection depends on probes that can distinguish between nearly identical sequences. Conventional oligonucleotide probes that are perfectly matched to target genes/genomic sequences are often unsuitable in such cases. Carefully designed mismatches can be used to decrease cross-hybridization potential, but implementing all possible mismatch probes is impractical. Our study provides guidelines for designing non-perfectly matched DNA probes to target DNA sequences as desired throughout the genome. These guidelines are based on the analysis of hybridization data between perfectly matched and non-perfectly matched DNA sequences (single-point or double-point mutated) calculated *in silico*. Large changes in hybridization temperature predicted by these guidelines for non-matched oligonucleotides fit independent experimental data very well. Applying the guidelines to find oligonucleotide microarray probes for P450 genes, we confirmed the ability of our point mutation method to differentiate the individual genes in terms of thermodynamic calculations of hybridization and sequence similarity.

INTRODUCTION

The ability of DNA microarrays to analyze tens of thousands of genes in one assay has led to continuous improvements in microarray techniques (1–4) and to software that analyzes microarray data. While researchers have begun to pay close attention to the quality of microarray data (5–10), the problem of cross-hybridization has received little consideration.

The success of microarrays in gene expression profiling depends on the specificity between the selected probes and the target genes. All DNA microarrays operate on the principle of DNA hybridization between complementary target and probe sequences. In cross-hybridization, an expressed gene hybridizes with probes designed for other genes as well as with its own designated probe, introducing noise. Experiments with cDNA microarrays have raised cross-hybridization concerns (11,12). Although oligonucleotides have the advantage of greater specificity than cDNA microarrays, they too are subject to some degree of cross-hybridization. In particular, genes of highly similar sequences such as those in a large gene family are very difficult to distinguish in microarray experiments (13,14).

An even greater problem related to cross-hybridization arises when microarrays are employed to discriminate single nucleotide polymorphisms (SNPs). In order to enhance specificity, new probes other than conventional linear oligonucleotides have been developed, such as structured DNA probes (molecular beacons) (15) or gold nanoparticle probes (16). Another approach includes the preparation of target genes by pooling two separate PCRs (17). Unfortunately, all of these approaches require new materials and/or complex processes. A simpler, more promising approach by Guo *et al.* (18) involves the introduction of artificial mismatches in the probes to enhance the discrimination of SNPs. However, they introduced artificial nucleotides, which cannot utilize the hybridization differences between natural nucleotides, for mismatch pairs. Also their dataset was limited in size and only short length oligos were used in the experiments. Thus, it is difficult to extend their findings to general SNP probe selection.

In theoretical terms, the thermodynamics of nucleic acid hybridization has been well established (19,20) and generally accepted in predicting melting temperature (T_m). The T_m of two-nucleotide hybridizations can be obtained from several public websites as well as calculated using either nearest-neighbor thermodynamics (21,22) or the simple %GC-content method (23,24). The thermodynamic parameters of single

*To whom correspondence should be addressed. Tel: +1 734 615 5918; Fax: +1 734 998 8571; Email: inhan@umich.edu base pair mismatches have also been determined experimentally (25–29).

Nearest neighbor calculation is known to predict experimental results well. By definition, thermodynamic parameters depend on the sequence content of nearest neighbor, 5′/3′ direction, and the chain terminal sequences, not on the position of oligonucleotides. A longer oligo is thought to have higher calculated T_m because of the additive calculations for enthalpy and entropy changes (see equations in Materials and Methods). However, it is much faster computationally to consider one sequence at a time instead of two consecutive sequences, which include directional information. Also there has been no quantitative analysis of calculated T_m and oligo length relationships as far as we know. These considerations are especially relevant when we want to introduce mismatch pairs in the probe sequences.

In this study, we calculated hybridization T_m of oligos extracted from human mRNA sequences based on nearest neighbor calculations. Then we statistically analyzed calculated T_m values of human mRNA sequences in terms of a single nucleotide and its position, the most relevant information in designing probes, rather than considering two consecutive nucleotides and their directions. This study was based on attempts to (i) find guidelines for introducing a mismatch pair that can discriminate between two similar oligonucleotide pairs with the same calculated hybridization T_m and (ii) determine the potential of a new probe design including point mutated probes for oligonucleotide microarrays that would minimize cross-hybridization across the entire genome. Our primary focus was on the guidelines, since the number of possible mismatch pairs increases dramatically as the length of a gene or the number of mutations increases, and consideration of all possible mismatch pairs for an entire genome is impractical. The T_m variance was examined using point mutation of DNA oligos based on theoretical calculations. Calculated ΔT_m dependencies on oligo length, mutation site, mutation sequence and distance between two mutation sites could be studied using a more extensive dataset than that obtained from experiments. New insight into mismatch pairs could be obtained from these *in silico* experiments.

MATERIALS AND METHODS

T_m of an oligonucleotide bound to a complementary nucleotide is the temperature at which 50% of duplex strands are separated. Since it is at equilibrium and the reaction is intermolecular, T_m depends on oligonucleotide concentration as well as Na+ concentration. A nearest neighbor model may calculate T_m from the following equation

$$T_m = T°\cdot\Delta H°/(\Delta H° - \Delta G° + R\cdot T°\ln[C/4]) + 16.6\cdot\log_{10}\{[Na^+]/(1 + 0.7[Na^+])\} - 269.3 \qquad \mathbf{1}$$

where $\Delta G°$ is the standard free energy, $\Delta H°$ the enthalpy, $T°$ the temperature, R the gas constant (1.987 cal/kmol), and C the total oligonucleotide concentration for non-self-complementary cases if the strands are in equal concentration (19). $\Delta G°$ and $\Delta H°$ are for the sums of all nearest neighbor and chain end contributions in coil-to-helix transition. At Na^+ concentration of 1 M, equation **1** becomes

$$T_m = \Delta H°/(\Delta S° + R\ln[C/4]) \qquad \mathbf{2}$$

where $\Delta S°$ is the predicted entropy change. $C/4$ becomes C for self-complementary oligonucleotide duplexes (20). We used equation **2** for T_m calculations with the thermodynamic parameters in the table in the Supplementary Material, which shows references where they were found. Since the parameters for the mismatched pairs are not applicable to terminal or penultimate position, we did not calculate mismatch pairs for those positions with this method.

Most hybridization T_m calculations were done using OMP software (DNA Software, Inc., Ann Arbor, MI; http://www.dna-software.com). The software simulates and predicts nucleic acid hybridization in solution and produces structural and thermodynamic parameters. OMP predicts nucleotide acid structures utilizing dynamic programming methods (30) and calculates their reactions in equilibrium based on nearest neighbor calculations, which determine 'thermal' T_m using equation **2**. OMP calculates 'actual' T_m considering competing secondary structures of oligonucleotides. For reproducibility, we used 'thermal' T_m in this paper.

We selected oligos of various lengths (20, 30, 40, 50, 60, 70 and 80mer: shorter oligos are subsets of longer ones) from human mRNA sequences: EphB1 mRNA (*EPHB1*, gi: 4758283) and IgG Fc binding protein mRNA (*FCγBP*, gi: 4503680). We set marks at every 100th position from the 3′-end and copied 20 sequences from the marks to the 5′-end to obtain 20mer oligonucleotides providing a varied sub-sample of all possible oligos. Since longer oligonucleotides were prepared in a similar way, all the shorter ones were subsets of the longer ones. Point-mutated oligos (single or double) were systematically selected from these original oligo sequences. The hybridization T_m of original and mutated oligos with the complementary target sequence was calculated. The OMP parameters used for the calculations were oligonucleotide concentration 1 mM, Na^+ concentration 0.5 M, and assay temperature 37°C. Each original sequence mutated to three other sequences. After the hybridization T_m was calculated, T_m differences between the original pair and the mutated ones were calculated and the histograms of ΔT_m obtained for categories such as position, sequence change, distance between mutation sites and GC content. Each histogram step is 1°C and mean and standard deviation were calculated from the histogram values. Exponents shown in Figures 3 and 5 were obtained using power fitting in Microsoft Excel.

In the probe selection program, we used an oligo probe and a corresponding sequence of genes for OMP calculation to reduce computation time. When the OMP input contains nucleotides A and B, the OMP output reports each calculated concentration and T_m of ensembles such as folded A, A–A dimer, folded B, B–B dimer, A–B dimer and random coils. Hybridization percentage is defined as the percentage of A–B dimer among the ensembles at the assay temperature. Parameters for probe design were probe length = 60mer, assay temperature = 80°C, Na^+ concentration = 0.5 M, target and probe concentrations = 1 mM, probe folding T_m < 60°C, hybridization percentage of probe and target gene >95%, and hybridization percentage of probe and non-target gene <20%.

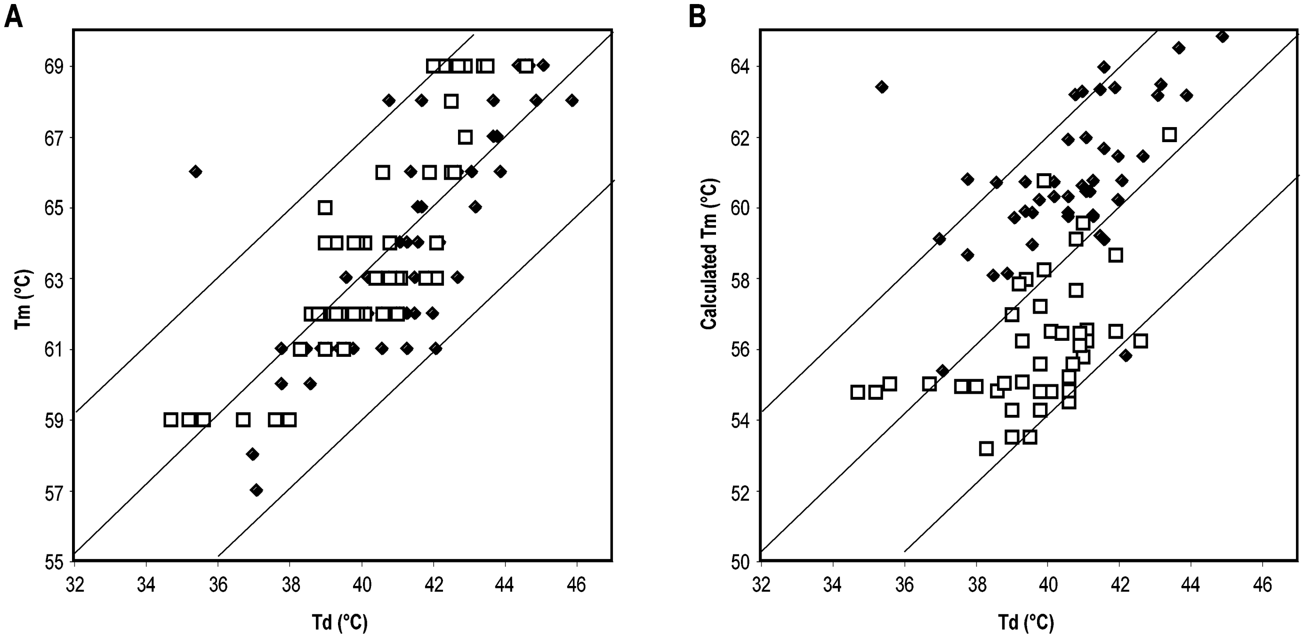

Figure 1. Correlation between theoretical T_m and experimental T_d from Urakawa *et al.* (31). (**A**) T_m was calculated using OMP software. (**B**) T_m was calculated using equation **2** and the parameters in the table in Supplementary Material. Filled diamonds are from spm (oligo1) and its mismatched pairs and open squares are from npm (oligo2) and its mismatched pairs in Urakawa *et al.* (31).

RESULTS

Correlations between T_m calculations and experimental data

Published data exist on the dissociation temperature (T_d) of perfect-match and single-base-mismatched nucleic acid duplexes (31). Even though T_d depends on time while T_m does not, we decided to compare them since the experiments were performed with a fixed setup and we assume one base mismatch will not change the kinetics dramatically. Both nearest neighbor parameters and the experimental data were obtained from melting profiles. T_m of two perfectly matched oligonucleotides and their all-possible single-base-mismatched pairs was calculated and compared with their data (Fig. 1). The sequence of oligo1 was TCGCACATCA GCGTCAGTT and that of oligo2 was CCCCTCTG CTGCACTCTA. T_m of Figure 1A was obtained from OMP calculation and that of Figure 1B was calculated using equation **2** and the table in Supplementary Material. Oligonucleotide concentration for these calculations was 100 nM, which was the default parameter of OMP. Both T_m calculations are about 20°C higher than the T_d data. We could reduce the calculated T_m close to T_d by setting the oligonucleotide concentration at 1 pM. However, in comparing T_d and T_m, the absolute value is less important than overall correlation, so we kept the default oligo concentration. In OMP calculations (Fig. 1A), both oligo1 (filled diamond) and oligo2 (open square) show very close linear correlation between T_m and T_d except in one case, while in-house calculations (Fig. 1B) show a more dispersed relation between T_m and T_d and differences between oligo1 and oligo2. The one data point far off from the linear relationship in Figure 1 is the [G, T] mismatch at the 7th position from the 5′-end of oligo1. Excluding this far-off data point, the correlation coefficients R^2 of oligo1 and oligo2 in Figure 1A are 0.73 and 0.72, respectively, while those in Figure 1B are 0.46 and 0.25, respectively. Considering that the experimental standard deviation was from 0.4 to 2°C, the linear correlation between nearest neighbor calculations and experimental data seems to be very high. Since OMP shows better linear correlation with single-mismatched pair oligos, T_m differences by point mutation calculated using OMP will be presented.

ΔT_m by single-point mutation

Two human mRNA sequences were used as templates for oligo generation. The sequences, EphB1 mRNA (*EPHB1*, gi: 4758283) and IgG Fc binding protein mRNA (*FCγBP*, gi: 4503680), have no significant similarity as assessed by the BLAST 2 program (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html) using default parameters. Their properties, obtained from the PROLIGO oligo parameter calculation website (http://proligo2.proligo.com/Calculation/calculation.html), and T_m calculated from the following equation are shown in Table 1

$$T_m = 81.5 + 16.6 \cdot \log_{10}\{[Na^+]/(1 + 0.7[Na^+])\} + 0.41(\%G + C) - 500/D - P \quad 3$$

where D is duplex length and P is percent mismatch, and we used [Na+] = 0.5 M (19). These target genes were selected to demonstrate that our oligo probe design approach is independent of the sequence content of target genes. Oligo probes of various lengths were selected from each target gene as described in Materials and Methods.

The hybridization T_m for oligos of various lengths and their complementary target sequences was computed and defined as the original T_m ($T_{m\ original}$) (Fig. 2A). The T_m calculations were performed as detailed in Materials and Methods. The average hybridization T_m for each oligo length based OMP calculation

Table 1. General properties of *EPHB1* and *FCγBP*

Gene	Accession	Length	GC content (%)	MW (g/mol)	T_m (°C)[a]
EPHB1	NM_004441.1	3871	54	1194541	98.0
FCγBP	NM_003890.1	16382	63	5057549	102.0

[a] Calculated from equation **3**.

is roughly 4–5°C higher with *FCγBP* than that with *EPHB1* (Fig. 3A), consistent with the T_m differences between the two entire genes based on %GC calculation, whose values are presented in Table 1. Note that overall temperatures are roughly 10°C higher than calculated conditions at [Na$^+$] = 100 mM and at the same oligo concentration of 1 mM. T_m values generally depend on GC content and secondary structures. The standard deviation of oligo hybridization T_m from *FCγBP* is smaller than that from *EPHB1* in calculation, possibly because the number of oligos obtained from *FCγBP* is much greater due to the longer sequence.

The mutated T_m ($T_{m\ \mathrm{mutated}}$) was defined as the hybridization T_m between a point-mutated strand and the reverse complementary sequences of the original target strand (Fig. 2A). The T_m variance by a mutated sequence, ΔT_m is as follows

$$\Delta T_m = T_{m\ \mathrm{mutated}} - T_{m\ \mathrm{original}} \qquad \mathbf{4}$$

Interestingly, calculated ΔT_m for the mutated oligo probes in both genes has the same dependency on oligo length with virtually the same standard deviation as indicated by the error bars (Fig. 3B). The dependency can be described as

$$|\Delta T_m| \propto L^{\alpha} \qquad \mathbf{5}$$

where $|\Delta T_m|$ is the absolute value of the mean ΔT_m and L is the oligo length. Both exponents α of *EPHB1* and *FCγBP* oligos are –1.87. All standard deviations in Figure 3 diminish as the oligo length increases. Since the calculated ΔT_m of the two mRNAs showed the same α and standard deviation despite the content differences in the two genes, they suffice for our study. The combined results will be presented.

We investigated the calculated ΔT_m of each mutation site for each oligo probe to assess the nature of the large standard deviation shown in Figure 3 and to provide guidelines for oligo probe design. The mutation site is defined as the position from the end of the oligo, ignoring 5′ and 3′ direction. Position 1 refers to the two end positions of each oligo, position 2 the next positions toward the center, and so forth (Fig. 2B). Mutation site dependency of oligo length 20 is shown in Figure 4A. A slight change in calculated T_m with relatively small standard deviation is observed at position 1, and $|\Delta T_m|$ increases with larger standard deviation as the position number increases up to four. There is no mutation site dependency from position 4 towards the center. This trend is independent of the specific nucleotide mutation (Fig. 4B, standard deviations not shown for data clarity). When the mutation site is at position 2, two groups appear: practically no change in T_m (mutations from either A or T), and the beginning of significant T_m change (mutations from either C or G). Many mutations present changes in T_m and distinctively larger standard deviations at position 3, while mutations from T undergo little T_m change. However, no mutation shows position dependency except the three end positions. ΔT_m dependency on nucleotide mutation of oligo 20mer, excluding the last three end positions, is shown in Figure 4C. When A or T are mutated to G, $|\Delta T_m|$ is the lowest. $|\Delta T_m|$ by mutation from G is larger than in any other cases. Although the mutation from G to C yields the greatest change in calculated T_m, the standard deviation is much larger than the mutation from G to T, which presents a similar T_m change (clearer for oligo length 30 and longer). All these features hold true for other oligo lengths as well. The experimental data from Urakawa *et al.* (31) excluding the last three end positions are shown in Figure 4D for comparison. Color and shape are coded as in Figure 4B. Since the lengths of oligo1 and oligo2 described in Figure 1 are different, we display all data at each position from the 5′-end instead of position definition in Figure 2B. Except for the one data point, indicated with an arrow, which was the far-off value in Figure 1, data fall within the range of the Figure 4B trend. Blue and red triangles (mutation from A and T to G) present the least T_m changes; orange shapes (mutations from G) represent most T_m changes. ΔT_d differences among positions of the same color and shape are around 2°C, while differences among specific sequences at certain positions (8 and 13 positions from 5′-end) can exceed 6°C.

ΔT_m by two-point mutation

Single-point mutations were compared with two-point mutations of oligos from *EPHB1* mRNA. Figure 5 presents calculated ΔT_m as a function of oligo length excluding mutations of the last three end positions. To simplify the variables, only 20 positions from the 3′-end of oligonucleotides are included in the two-point mutation data. The values of 2 × ΔT_m by single-point mutation are also presented to show that calculated ΔT_m by two-point mutation is not simply additive of ΔT_m by single-point mutation calculation, indicating synergetic effects in two-point mutation. The exponent α in equation **5** of the single-point mutation without the last three end positions is slightly reduced at –1.93 from Figure 3B. The exponent in the two-point mutation is –1.55, showing a more sensitive dependency on length.

We define the reference position as the smaller position of the two mutation sites in a two-point mutation. The distance is defined as one plus the number of nucleotides between them (Fig. 2C). The ΔT_m dependency of oligo 20mer on the reference position and distance is presented in Figure 6. By definition, increasing distance refers to the second mutation site moving towards the other oligo-chain end until the second site reaches the same position number as the reference position. Open shapes represent ΔT_m values of mismatch pairs when at least one of the mutation sites is at the last three end positions; filled shapes indicate both mutation sites are at inner positions. The greatest $|\Delta T_m|$, along with the smallest standard deviation, is observed when two consecutive sites (distance of one) are mutated at position six or greater. For a fixed distance between two mutation sites, there is no position

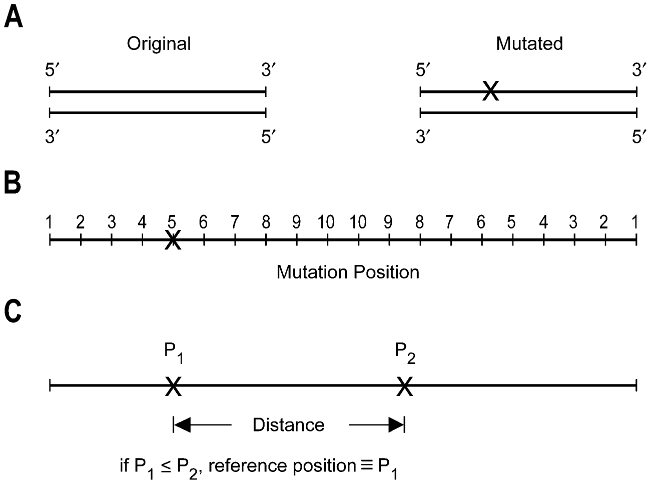

Figure 2. Schematic diagrams of oligonucleotides. (**A**) Original T_m is defined as the hybridization T_m of the perfect match pair. The point-mutated site is described in × and the mutated T_m is defined as their hybridization T_m. Three possible mutations occur at each site. (**B**) The mutation site is defined as the position from the end of the oligos, ignoring 3′ and 5′ direction. A single-point mutation is performed at position 5. (**C**) In the two-point mutations, the reference position is defined as a smaller position number between two mutation sites. The distance is defined as 1 plus the number of nucleotides between them.

dependency except with the last three positions (reference position ≤ 3) and at reference–distance pairs 4–1 and 4–2. Also, for fixed reference positions, very little distance dependency is observed except when the pairs are nearest neighbors or when both mutation sites are among the last four end positions. Note that the last fourth end positions do not show differences from the inner ones when each mutation occurs at each other's chain side (1–16, 2–15, 3–14). On the other hand, when both mutations occur on the same side of chain, mutations at 3–2, 3–3 and 4–2 show distance dependency. These features also hold true for longer oligo lengths. Mutations at 4–3, 5–1 and 5–2 show moderate reference position or distance dependencies, which disappear at oligo lengths longer than 30.

With two-point mutations, the ΔT_m dependency on the specific nucleotide mutation is similar to that observed with single-point mutations. When both mutations are from G, $|\Delta T_m|$ is the largest, whereas when both mutations are from A or T to G, $|\Delta T_m|$ is the least. Some distinctive features related to the distance between mutation sites are observed in the double mutations. Figure 7 shows specific sequence changes correlated with differences in distances between mutation sites in 20mer oligos. We varied the distance between mutation sites from one to four while fixing one mutation site position at ten. Nucleotide-specific mutations that provided a significantly larger $|\Delta T_m|$ in the nearest neighbor position as compared to distances greater than one are shown to the left of the {C to T, C to T} double mutation in Figure 7. The double mutations showing a relatively larger $|\Delta T_m|$ in the next nearest neighbor mutation cases as compared to the other positions are shown on the right side of the figure. There are three mutation categories where two consecutive mismatches (filled diamonds in the figure) provide significantly larger $|\Delta T_m|$: (i) both mutations are from A or T and at least one of them goes to G, except for {A to G, A to G} mutations; (ii) both mutations are from C, except {C to T, C to T} mutations; and (iii) one mutation is from C to T and the other is from G. Note that mutations from both A to G and C to T, which are exceptional cases of the above categories, lead to [G, T] mismatches. On the other hand, when both the mutations are from G, mismatches at the next nearest neighbor positions (filled rectangular) provide relatively larger $|\Delta T_m|$ than any other mutation distances.

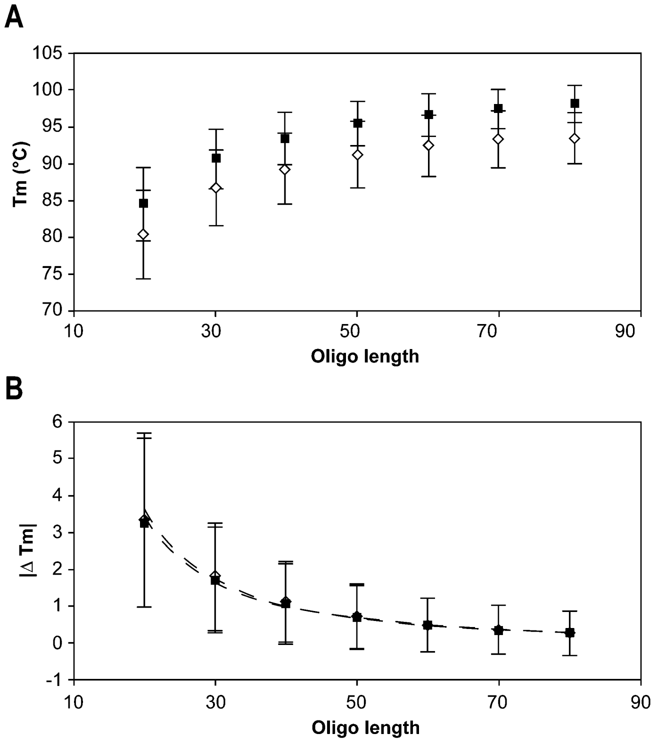

Figure 3. Hybridization T_m dependency (**A**) and T_m variance (ΔT_m) dependency (**B**) on oligonucleotide length due to single-point mutation based on OMP calculation. Open diamonds represent data obtained from oligo sequences of EphB1 mRNA. Filled squares represent data obtained from IgG Fc binding protein mRNA. Shorter oligos are subsets of longer ones. Dotted lines in B are fitted graphs providing the exponent α in equation **5** in the main text. One original oligo pair in (A) produces (3 × oligo length) mutated pairs in (B).

Point-mutated probe sample

We developed software based on the thermodynamic calculation and sequence similarity test to design oligo probes that hybridize with only one specific gene among the entire genome. However, using oligo probes perfectly matched to the target gene made it difficult to select probes for genes in large families with significant similarity, such as the P450s. To avoid cross-hybridization in a microarray experiment, it is desirable to have a probe hybridize with its complementary target at a T_m that is significantly higher than the T_m of the same probe matched with any other gene product in the transcriptome. It is also desirable to have all probes in one microarray hybridize with their targets at similar T_m values. Based on the above results, we selected point mutations to design probes that distinguish similar genes. One example is P450 *CYP21A2* human mRNA (gi: 20522237, accession:

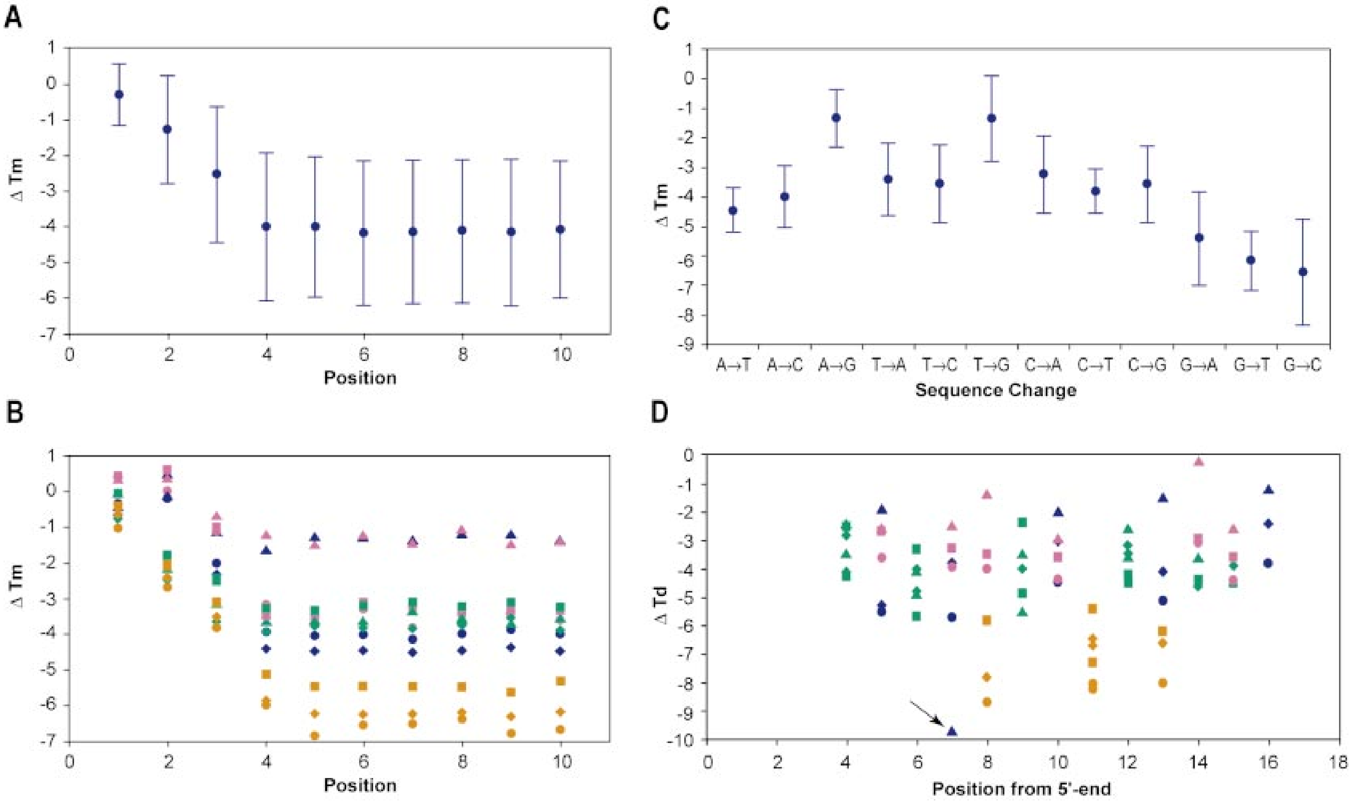

Figure 4. Calculated ΔT_m dependency on nucleotide mutation and position in 20mer oligonucleotides. (**A**) Average ΔT_m values of all sequence changes at each position are shown with standard deviations. (**B**) Average ΔT_m values of each sequence change are shown. Data point colors represent the original sequences: blue, A; pink, T; green, C; and orange, G, while shapes represent the mutated sequences: square, A; diamond, T; circle, C; and triangle, G (blue diamonds, from A to T; blue circles, A to C; blue triangles, A to G; pink squares, T to A; pink circles, T to C; pink triangles, T to G; green squares, C to A; green diamonds, C to T; green triangles, C to G; orange squares, G to A; orange diamonds, G to T; and orange circles, G to C). (**C**) ΔT_m dependency on nucleotide in 20mer oligonucleotides excluding the three end points. (**D**) All ΔT_d values of each sequence change from Urakawa *et al.* (31) are shown at each position from the 5′-end excluding the three end positions. Data from oligo 1 and oligo 2 in Figure 1 are not distinguished. Color and shape representations are as in (B). A data point indicated with an arrow is a [G, C] mismatch case, which will be ignored in the comparison.

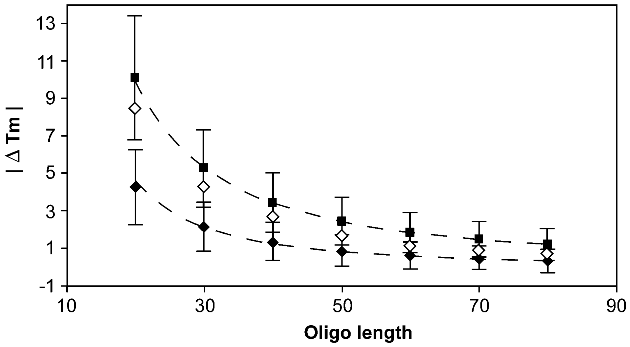

Figure 5. Calculated ΔT_m dependency on oligonucleotide length by single- and two-point mutations excluding the three end points. Filled diamonds represent single-point mutations, filled squares two-point mutations, and open diamonds represent 2 × ΔT_m by single-point mutations. Dotted lines are fitted graphs.

NM_000500.4). The test database was downloaded from the NCBI RefSeq human mRNA site (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_prot). *CYP21A2* has high similarity with P450 pseudogene (*CYP21A1P*) on chromosome 6 (gi: 20270487, accession: NG_001111.1) among the test database. A BLAST search (with word size 8) reported most sequences in *CYP21A2* have 97–100% identity with *CYP21A1P*. One lowest similarity report was 92% identity among 155 sequences. The only non-reported part was positions between 828 and 854 from the 5′-end, which are only 26 sequences surrounded by 98% identity sequences. Moreover, these 26mer reported sequence similarities with other genes in the test database. Our probes were 60mer, with one of the selection criteria being over 95% hybridization at assay temperature (80°C). Varying the probe size from 60mer should not improve the selection much.

One probe perfectly matching a target sequence has cross-hybridization potential with *CYP21A1P* and more than 10 other genes, among them RAP1, GTPase activating protein 1 (*RAP1GA1*) mRNA (gi: 20270487, accession: NG_001111.1) and human double homeobox, 2 (*DUX2*) mRNA (gi: 21687002, accession: NM_012147.2), based on a BLAST search of the test database and T_m calculations. Using three-point mutation, the probe was redesigned to avoid cross-hybridization. Calculated hybridization T_m and hybridization percentage at 80°C of both the probes and target/non-target genes are in Table 2. Bold letters are the mutation sites; lower case letters represent mutation sequences. At assay temperature 80°C, both perfect match and three-point mismatch

Table 2. Calculated hybridization properties between oligo probes and target/non-target genes within the test database

Oligo probes	Probe 1 Hybridization T_m (°C)	Hybridization percentage[a] (%)	Probe 2 Hybridization T_m (°C)	Hybridization percentage[a] (%)
Target gene				
CYP21A2	93	100	86	100
Non-target genes				
CYP21A1P	82	69	73[b]	4[b]
RAP1GA1	79	39	74[b]	6[b]
DUX2	80	46	74[b]	11[b]

Probe 1: CAGGCCATAGAGAAGA**G**GGATCACATCGT**G**GAGATGCAGCTGAG**G**CAGCACAAGGAGAGC; Probe 2: CAGGCCATAGAGAAGA**c**GGATCACATCGT**t**GAGATGCAGCTGAG**t**CAGCACAAGGAGAGC. Bold letters are the mutation sites; lower case letters represent mutation sequences.

[a] Hybridization percentage of target-probe dimer among all possible states (monomer, homodimer and heterodimer) in equilibrium at 80°C.

[b] Expected not to contribute to cross-hybridization at assay temperature.

probes will hybridize 100% with the target gene if there are no other competing genes. However, calculated hybridization T_m values of a perfect match probe with *CYP21A1P*, *RAP1GA1* and *DUX2* are around 80°C, leaving open the possibility of cross-hybridization. Meanwhile, the mismatch probe lowered the calculated hybridization T_m values with *CYP21A1P*, *RAP1GA1* and *DUX2* (and all other non-target genes) ~5–10°C. All cross-hybridizations at assay temperature 80°C are expected to disappear.

DISCUSSION

Nearest neighbor T_m calculations, especially using OMP, showed an excellent linear correlation with experimental T_d for single-mismatched pairs, even though we found it difficult to set the right salt and oligonucleotide concentration parameters. We calculated T_m of 50mer oligonucleotides with 80 and 74% sequence similarity in addition to other pairs shown in Kane *et al.* (13). Their hybridization condition was under 42°C and our T_m calculation at 150 mM salt and100 pM oligo concentration (rough estimate) for an 80% similarity pair was 47°C, while that for a 74% similarity pair was 24°C. It is tempting to say that an 80% similarity pair has a chance for hybridization while a 74% similarity one does not, as indicated by the experimental results. However, when we increased the oligo concentration at 100 nM, the calculated T_m values of 80 and 74% similarity pairs were 67 and 52°C, respectively. The dependable interpretation should be that calculated T_m of an 80% similarity pair of that specific sequence arrangement differs from that of a 74% similarity pair large enough to be easily distinguishable in the hybridization experiments. Since our purpose is to find the guidelines for introducing mismatch pairs to distinguish two similar sequences, the relative T_m values at fixed parameter conditions is all that is needed. Absolute calculated T_m can be adjusted according to each experimental condition. Even though all nearest neighbor parameters were obtained from oligos <20mers, an experimental study reported that single base mismatch results from long DNA (373 bp) agreed well with studies of short oligos (28), and Mfold server, one of the commonly-used oligonucleotide hybridization web services using nearest neighbor parameters (http://www.bioinfo.rpi.edu/applications/mfold/), recommends that results from oligo lengths <100 bp are reliable.

As noted in the Introduction, if hybridization T_m of longer oligos is higher because of the additive nature of enthalpy and entropy changes in calculation, the effect of one point mutation on the T_m will be smaller in longer oligos. Figure 3B quantitatively describes the change of T_m by single point mutation in relation to length. Surprisingly, calculated T_m differences by point mutation do not depend on the original mRNA's GC content. There is also little dependency of ΔT_m on the oligo probe's GC content (data not shown). Statistical data show that one point mutation of oligo 60mer does not change $|\Delta T_m|$ much, while that of oligo 20mer can significantly change $|\Delta T_m|$. This is key data in support of matched sequences with non-target genes from a similarity test as candidates for mutations. When we design long oligo of 50mer or 60mer, the perfect 50 or 60 consecutively match sequences will be searched. Even if the sequence similarity test finds 20 consecutive sequence matches with other genes, one point mutation can dramatically reduce the hybridization T_m with non-target genes, while maintaining a similar T_m for the target gene. However, the large standard deviation of $|\Delta T_m|$ especially in shorter oligos, raises the following issues.

Figure 4 provides crucial guidelines for increasing $|\Delta T_m|$ using a single-point mutation in terms of a single nucleotide and its position, which are the most relevant indicators to probe design. We left the 5′ and 3′ directions out of the guidelines based on analysis of the data. Experimental data in other work asserts that a maximum destabilizing effect of a mismatch can be obtained at the center of an oligo (18). However, our data indicate that the position of the mutation is less significant than the identity of the nucleotides involved in the mutation, excluding the three positions at either end of the oligo. Mutating a G at position 5 will, on average, produce a significantly greater ΔT_m than a mutation from A or T to G at position 10, which is at the center of a oligo 20mer. This follows from the nature of nearest neighbor calculation except that more position dependency was shown near the oligo terminals than predicted by the nearest neighbor calculation. Figure 4D, adapted from experimental data in Urakawa *et al.* (31) and displayed in a different color and shape for each sequential change, clearly supports our assertion. Interestingly, the 11th positions from the 5′-end of both oligo1 and oligo2 were G. Without considering the sequences, position 11 from the 5′-end can be misinterpreted as a critical position for mutations. However, our study is limited to properties in solution. There was a report of microarray signal

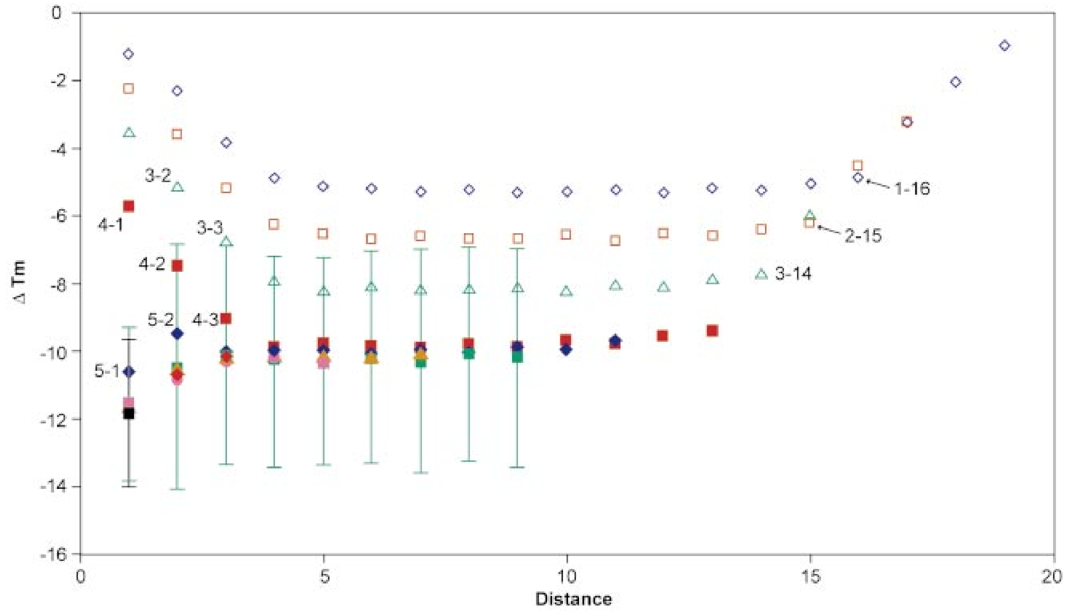

Figure 6. Calculated ΔT_m dependency on the reference position and the distance between two mutation sites by two-point mutations of 20mer oligonucleotide. Open blue diamonds represent reference position 1, open red squares reference position 2, open green triangle position 3, filled red squares position 4, filled blue diamonds position 5, filled green squares position 6, filled orange triangles position 7, filled pink circles position 8, filled red diamonds position 9 and filled black square position 10. Only reference positions 6 and 10 display standard deviations. Certain reference position and distances, explained in the main text, are annotated.

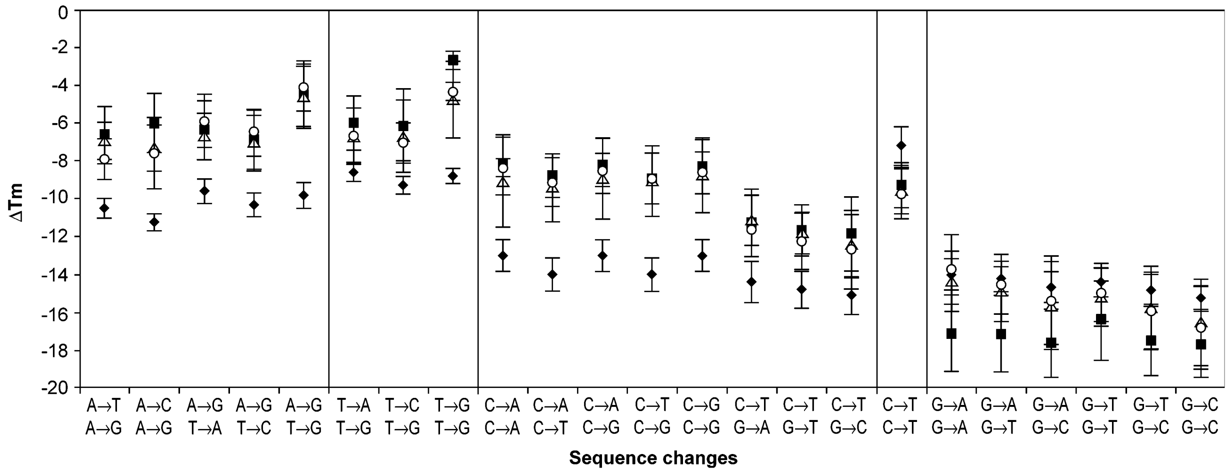

Figure 7. Specific sequence changes correlated with significantly different calculated ΔT_m of 20mer oligonucleotides in the distance between mutation sites at fixed position 10. The distances from position 10 are one (filled diamond), two (filled square), three (open triangle) and four (open circle).

dependency on mismatch position of the 60mer oligos attached on the surface (4). Even though sequence effects were not considered and it was not clear how many genes were used for the mismatch position experiments, there seem to be distinctive differences between positions close to the surface and away from the surface. Note that the T_d measurement experiments were also done on the surface-attached oligos (about 20mers) and agreed very well with the calculations in relative terms. Surface effects seem to be more important as oligo length increases, which is beyond the scope of this paper.

For each sequence, we suggest that the following changes produce the greatest ΔT_m: A to T, T to A, C to T and G to T. We selected G to T as the best change because the standard deviation is much smaller, even though the mean ΔT_m value of G to C is larger (Fig. 4C). Similarly, mutations from T to A have a smaller standard deviation than T to C, which is obvious in oligo lengths 30mer and longer. Summarizing the results from single-point mutation, the design guidelines for the mutated probes are as follows: (i) find the sequence similarity area with other non-target genes; (ii) exclude the last three end positions; (iii) search for a G mutation; and (iv) avoid mutations from A to G or T to G.

Among the sequence changes in Figure 4, A to G and T to G show significantly smaller $|\Delta T_m|$ than others, while G to T and

G to C show significantly larger $|\Delta T_m|$ than most others. Mutation from A to G leads to [G, T] mismatches, which are among those most commonly observed in DNA (32,33). Thermodynamic studies of [G, T] mismatches have shown them to be stable duplexes (26–28). Similarly, mismatches [G, A] resulting from mutation T to G have also been reported as stable duplexes (26,27). On the other hand, G to T mutation leads to [C, T] mismatches that various thermodynamic studies report as one of the most unstable pairs (26–28). Mismatches [C, C] due to G to C mutations have also been reported to significantly destabilize the duplexes (26,29). Our *in silico* results correlate well with previous experimental findings.

When two mutation sites are introduced into oligo pairs, much larger $|\Delta T_m|$ can be achieved than with single-point mutations with stronger length dependency. This implies that careful selection of mutation sites will allow us to distinguish between two similar genes. Figure 6 provides a concise guideline for this selection. Once the cases which will produce a large $|\Delta T_m|$ using two-point mutations are selected from Figure 6, an important factor in achieving the greatest T_m change is to consider specific nucleotides such as including mutations from G. Additional information regarding the relation between distance and specific nucleotides can be obtained from Figure 7. Although we only find a trend of distances 1 and 2 producing the largest $|\Delta T_m|$ values in Figure 6, Figure 7 demonstrates that specific nucleotide mutations have a significant effect. When both mutations are from G, next nearest neighbor pairs will provide the largest T_m change. Nearest neighbor mutation pairs provide the greatest $|\Delta T_m|$ in all other cases, excluding the C to T, C to T pairs.

Our approach to using strategic point mutations for increased oligo probe specificity could greatly improve microarray results, particularly with gene families that present probe design challenges. We have identified guidelines for the introduction of point mutations in probe design. Using P450 *CYP21A2*, we demonstrate that application of the point mutation guidelines can improve probe specificity in the challenging case where significant sequence similarity exists between the target gene and non-target genes. An additional benefit in regard to mismatch pair probes is that we could reduce target-probe hybridization T_m if the experimental preference is for a narrower T_m range, such as around 85°C in the example of Table 2. Calculating the ΔT_m for all possible mutation pairs for all potential probes of a target gene is impractical due to the exponential growth of the number of probes as the length of a target gene increases. Therefore, practical guidelines are presented for point mutation selections that require little computational overhead. In addition to improving probe specificity among gene family members, this approach promises to improve the design of oligo probes for SNP detection (18). There are several points to be considered, however. All these calculations are based on solution equilibrium, while microarray probes are on a solid platform. Surface effects need to be added. Secondly, all calculations are hybridization of two oligos, not of a gene and an oligo, or multiple genes and an oligo. Current computational models cannot deal with these factors. Finally, our results and guidelines are based on theoretical modeling; therefore, laboratory validation is imperative, even though independent experimental data have already proved the value of our guidelines for single-point mutations (31).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank DNA Software, Inc., for providing OMP and Donald Hicks for the scientific description of OMP. This work is supported by grant N002831 from Michigan Economic Development Corporation—Life Sciences Corridor Fund and in part by grant ES06639 from the National Institute of Environmental Health Sciences.

REFERENCES

1. Fodor,S.P., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.
2. 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**, 467–470.
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. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
4. 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. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
5. Finkelstein,D., Ewing,R., Gollub,J., Sterky,F., Cherry,J.M. and Somerville,S. (2002) Microarray data quality analysis: lessons from the AFGC project. Arabidopsis Functional Genomics Consortium. *Plant Mol. Biol.*, **48**, 119–131.
6. Hess,K.R., Zhang,W., Baggerly,K.A., Stivers,D.N. and Coombes,K.R. (2001) Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol.*, **19**, 463–468.
7. Kerr,M.K. and Churchill,G.A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, **77**, 123–128.
8. Mills,J.C. and Gordon,J.I. (2001) A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Res.*, **29**, e72.
9. Tseng,G.C., Oh,M.K., Rohlin,L., Liao,J.C. and Wong,W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.
10. Wang,X., Ghosh,S. and Guo,S.W. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.*, **29**, e75.
11. Xu,W., Bak,S., Decker,A., Paquette,S.M., Feyereisen,R. and Galbraith,D.W. (2001) Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana. Gene*, **272**, 61–74.
12. Evertsz,E.M., Au-Young,J., Ruvolo,M.V., Lim,A.C. and Reynolds,M.A. (2001) Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques*, **31**, 1182–1192.
13. 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**, 4552–4557.
14. 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.
15. 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. USA*, **96**, 6171–6176.
16. Taton,T.A., Mirkin,C.A. and Letsinger,R.L. (2000) Scanometric DNA array detection with nanoparticle probes. *Science*, **289**, 1757–1760.

17. Germer,S., Holland,M.J. and Higuchi,R. (2000) High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.*, **10**, 258–266.
18. Guo,Z., Liu,Q. and Smith,L.M. (1997) Enhanced discrimination of single nucleotide polymorphisms by artificial mismatch hybridization. *Nat. Biotechnol.*, **15**, 331–335.
19. Wetmur,J.G. (1991) DNA probes: applications of the principles of nucleic acid hybridization. *Crit. Rev. Biochem. Mol. Biol.*, **26**, 227–259.
20. SantaLucia,J.,Jr (1998) A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
21. Crothers,D.M. and Zimm,B.H. (1964) Theory of the melting transition of synthetic polynucleotides: evaluation of the stacking free energy. *J. Biol. Chem.*, **9**, 1–9.
22. Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
23. Howley,P.M., Israel,M.A., Law,M.F. and Martin,M.A. (1979) A rapid method for detecting and mapping homology between heterologous DNAs. Evaluation of polyomavirus genomes. *J. Biol. Chem.*, **254**, 4876–4883.
24. Wallace,R.B., Shaffer,J., Murphy,R.F., Bonner,J., Hirose,T. and Itakura,K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res.*, **6**, 3543–3557.
25. Tibanyenda,N., De Bruin,S.H., Haasnoot,C.A., van der Marel,G.A., van Boom,J.H. and Hilbers,C.W. (1984) The effect of single base-pair mismatches on the duplex stability of d(T-A-T-T-A-A-T-A-T-C-A-A-G-T-T-G). d(C-A-A-C-T-T-G-A-T-A-T-T-A-A-T-A). *Eur. J. Biochem.*, **139**, 19–27.
26. Werntges,H., Steger,G., Riesner,D. and Fritz,H.J. (1986) Mismatches in DNA double strands: thermodynamic parameters and their correlation to repair efficiencies. *Nucleic Acids Res.*, **14**, 3773–3790.
27. Ikuta,S., Takagi,K., Wallace,R.B. and Itakura,K. (1987) Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs. *Nucleic Acids Res.*, **15**, 797–811.
28. Ke,S.H. and Wartell,R.M. (1993) Influence of nearest neighbor sequence on the stability of base pair mismatches in long DNA; determination by temperature-gradient gel electrophoresis. *Nucleic Acids Res.*, **21**, 5137–5143.
29. Peyret,N., Seneviratne,P.A., Allawi,H.T. and SantaLucia,J.,Jr (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G and T.T mismatches. *Biochemistry*, **38**, 3468–3477.
30. Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
31. Urakawa,H., Fantroussi,S.E., Smidt,H., Smoot,J.C., Tribou,E.H., Kelly,J.J., Noble,P.A. and Stahl,D.A. (2003) Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **69**, 2848–2856.
32. Wang,R.Y., Kuo,K.C., Gehrke,C.W., Huang,L.H. and Ehrlich,M. (1982) Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. *Biochim. Biophys. Acta*, **697**, 371–377.
33. Modrich,P. and Lahue,R. (1996) Mismatch repair in replication fidelity, genetic recombination and cancer biology. *Annu. Rev. Biochem.*, **65**, 101–133.