# Interactions in Oligonucleotide Hybrid Duplexes on Microarrays

**Hans Binder,*,† Toralf Kirsten,† Ivo L. Hofacker,‡ Peter F. Stadler,†,§ and Markus Loeffler†,||**

*Interdisciplinary Centre for Bioinformatics, University of Leipzig, Institute of Theoretical Chemistry and Structural Biology, University of Vienna, Bioinformatics group, Department of Computer Science, and Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Kreuzstrasse 7b, D-4103 Leipzig, Germany*

*Received: January 29, 2004; In Final Form: August 23, 2004*

We investigated Affymetrix GeneChip intensity data in terms of chip-averaged sensitivities over all perfect match (PM) and mismatch (MM) probes possessing a common triple of neighboring bases in the middle of their sequence. This approach provides a model-independent estimation of base-specific contributions to the probe sensitivities. We found that fluorescent labels attached to nucleotide bases forming Watson–Crick (WC) pairs in most cases decrease their binding affinity and, thus, decrease the sensitivity of the probe. Single-base-related mean sensitivity values rank in ascending order according to C > G ≈ T > A. The central base of PM and MM probes mainly forms WC pairings in duplexes with nonspecific transcripts, which obviously dominate the chip-averaged sensitivity values. Linear combinations of the triple-averaged probe sensitivities provide nearest-neighbor (NN) sensitivity terms, which rank in a similar order as the respective NN free-energy terms obtained from previous thermodynamic studies on the stability of RNA/DNA duplexes in solution. Systematic deviations between both data sets can be mostly attributed to the labeling of the target RNA in the chip experiments. Our results provide a set of molecular NN and single-base-related interaction parameters which consider specific properties of duplex formation in microarray hybridization experiments.

## Introduction

Target binding to high-density oligonucleotide microarrays used for gene expression experiments is governed by the molecular interactions in the hybrid duplexes formed by RNA fragments and DNA probes. The knowledge of the details of the DNA/RNA hybridization behavior on a molecular level and its estimation by means of effective parameters represents one prerequisite for selecting optimal probe sequences from target genes for newly designed chips. Especially short oligonucleotides might be ineffective as RNA binders as a result of relatively weak interactions between probe and target. Existing methods for chip design mostly involve thermodynamic criteria based on interaction parameters referring to hybrid duplexes in solution for the optimization of probe sequences (see refs 1, 2 and references therein). Recent analyses show that several factors, such as the presence of fluorescent labels, modifies the stability of RNA/DNA duplexes on microarrays compared with duplexes in solution.[3,4] The understanding of the hybridization properties of microarray probes presumably requires a modified view of the molecular interactions in DNA/RNA duplexes, which takes into account labeling and also, possibly, effects due to the fixation of the probes at the quartz surface.

Available microarray intensity data are directly related to the binding affinity of the individual probes.[4] They therefore provide valuable information about molecular interactions in RNA/DNA duplexes, which can be used to extract relevant interaction parameters. In this work, we make use of two types of redundancies in the design of Affymetrix GeneChip microarrays, which were created to improve the reliability of the method.[5,6]

First, so-called probe sets consisting of 11–20 different reporter probes for each gene allows us to estimate the sensitivity of a probe as the deviation of its intensity from the respective set average in a logarithmic scale.[4] The sensitivity of a microarray oligonucleotide probe characterizes its ability to detect a certain amount of RNA transcripts independently of the conditions of sample preparation, hybridization, and measurement of the fluorescence intensity. It is mainly determined by the affinity of a particular DNA probe to bind RNA fragments via complementary Watson–Crick (WC) pairs.

Second, each probe is present in pairs of so-called perfect match (PM) and mismatch (MM) modifications. The sequence of the PM is taken from the gene of interest, and thus, it is complementary to a 25-mer in the RNA target sequence. The sequence of the MM is identical with that of the PM probe except the position in the middle of the oligomer where the middle base is replaced by its complementary base. The pairwise design of probes intends to measure the amount of nonspecific hybridization and, by this way, to correct the PM intensities. An important question for GeneChip data analysis is how to include the MM intensities adequately. One prerequisite for solving this issue is the detailed study of the effect of the MM base in probe–target duplexes on the signal intensity.

In the accompanying paper,[4] we found that the middle base systematically shifts the PM and MM probe sensitivities relative to another. Also, other studies reported that the strength of base-pair interaction in the middle of the oligonucleotide affects the affinity of the probes for target binding to an extraordinary extend.[3,7] In addition, stacking interactions between nearest

* Corresponding author. E-mail: binder@izbi.uni-leipzig.de. Fax: ++49-341-1495-119.

† Interdisciplinary Centre for Bioinformatics.

‡ Institute of Theoretical Chemistry and Structural Biology.

§ Department of Computer Science.

|| Institute for Medical Informatics, Statistics and Epidemiology.

10.1021/jp049592o CCC: $27.50
Published on Web 10/27/2004

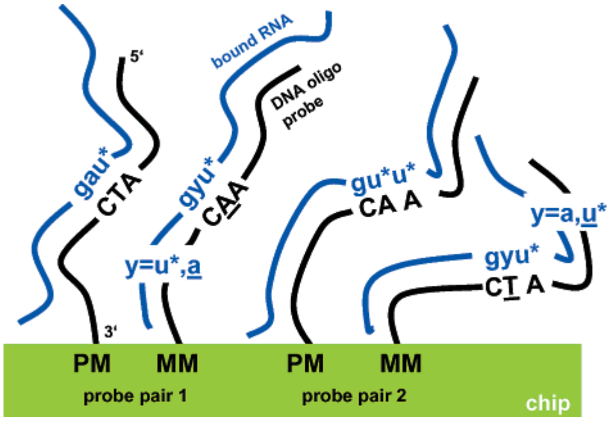

**Figure 1.** Schematic representation of RNA/DNA probe duplexes on GeneChip microarrays. A probe pair consists of a PM and MM probe. In the MM sequence, the middle base is replaced by the respective complementary base. Upper case letters indicate the middle triple of the DNA probe, which is paired with the respective complementary bases of the RNA target except the middle base of the MM (lower case letters; the asterisk indicates fluorescence labeling). We assume that the central base of the MM probes forms complementary WC pairs in duplexes with nonspecific transcripts and SC pairs in duplexes with specific transcripts (see text). The middle letters of PM and MM probes of the same pair are different. PMs and MMs with common middle bases refer to different probe pairs.

neighbors within the sequence of the probe and the target are known to influence the stability of the duplexes.[8−10] It seems, therefore, reasonable to consider the middle base and also its nearest neighbors and to study the probe sensitivities as a function of the middle triple XYZ (X, Y, Z = A, T, G, C; see Figure 1 for illustration) (i.e., of the nucleotide bases at position $k = 12-14$ of the probe sequence).

One key issue of probe design and chip data analysis addresses the relationship between the base composition of a probe and its affinity for target binding. Matveeva et al. showed that thermodynamic evaluations of the oligomer−target duplex and oligomer self-structure stabilities based on sequence information can facilitate probe design.[1] In our previous publication, we analyzed probe sensitivities as a function of simple sequence characteristics.[4] The results reveal, for example, a direct correlation between the number of C or A bases in the probe sequence and the sensitivity.

Our present work is aimed at characterizing the interactions between DNA and labeled RNA in terms of sequence-related parameters referring either to PM or MM probes. We make use of the fact that a typical Affymetrix GeneChip contains, on the average, more than 3500, but at minimum, more than 1000 different probes with a common middle triple. Averaging of the sensitivity values over these ensembles of probes with common triples XYZ to a large extent reduces the specific effect of the sequence outside of the middle triple (i.e., for base positions $k = 1-11$ and $15-25$). The chip-averaged sensitivities for each middle triple of the PM and MM probes, by this way, allow the detailed characterization of the binding affinity as a function of the triple sequence. We derived here nearest-neighbor (NN) interaction parameters from the triple averages and compared them with the NN free-energy terms of duplex formation in solution.[10,11]

Our model-independent approach complements previous studies which analyze the intensity of microarray probes in the framework of base- and position-dependent models.[3,7,12,13] An alternative model which considers the positional dependence of NN sensitivity terms was recently presented.[14]

## Methods and Microarray Data

**Triple Averages of the Probe Sensitivity.** We defined the sensitivity of perfect match (PM) and mismatch (MM) probes of Affymetrix GeneChips as their normalized intensity in a log10 scale

$$Y^{P} = \log I^{P} - \langle \log I^{P} \rangle_{\text{set}};\ P = \text{PM, MM} \qquad (1)$$

where the angular brackets $\langle \cdots \rangle_{\text{set}}$ denote arithmetic averaging over the probe set of 11−20 probes referring to one target gene.[4] The probe intensities are corrected for the optical background using the algorithm provided by *MAS 5.0*.[6] So-called chip averages of middle triples, $\langle Y^{P}(\text{XYZ}) \rangle_{\text{chip}}$, are calculated over all probe sequences with a common middle triple given by the bases XYZ (X, Y, Z = A, T, G, C) at position $k = 12-14$ of the probe sequence.

Most of the presented triple averages are mean values of the chip averages over the $N_{\text{chip}} = 42$ GeneChips provided by the Affymetrix human genome HG U133 Latin square (HG U133-LS) data set available at http://www.affymetrix.com/ support/technical/sample_data/datasets.affx

$$\langle Y^{P}(\text{XYZ}) \rangle = \frac{1}{N_{\text{chip}}} \sum_{i=1}^{N_{\text{chip}}} \langle Y^{P}(\text{XYZ}) \rangle_{\text{chip,i}}$$

In addition, we analyzed chip data taken from two different types of Affymetrix GeneChips referring to the human (HG U95Av2) and mouse (MG U74Av2) genome. The results only differ insignificantly from results obtained from the HG U133 chips presented here (not shown). All chip analyses are performed using the gene expression data warehouse platform of IZBI (see www.izbi.de).

Each HG U133 Affymetrix chip contains, on the average, nearly 4000 different probes with a common middle triple. It is therefore reasonable to assume that averaging reduces the specific effect of the sequence outside of the middle triple (i.e., for base positions $k = 1-11$ and $15-25$). On the other hand, the distribution of the nucleotide bases in the ensemble of probes with a common middle triple can significantly deviate from the random distribution. In the Appendix, we show that this effect gives rise to a systematic bias of the triple averages, which can be considered by a correction factor $F_{\text{triple}} \approx 1.2$ for the triple averages and $F_{\text{NN}} \approx F_{\text{SB}} \approx 1.1$ for the derived nearest-neighbor and single-base terms.

**Binding Affinity of PM and MM Probes for Specific and Nonspecific Transcripts.** The normalized intensity, $Y^{P}$, defines the sensitivity of a given probe.[4] It is directly related to the binding constant between target RNA and DNA oligonucleotide probes, $K^{P,S} = K^{P,S}(\xi^{P}\xi^{T})$ (the superscripts, P = PM, MM, and T, differentiate between PM and MM probes and the target; see also eq 2), to nonspecific hybridization and terms which consider fluorescence emission, saturation of the probes with bound targets, and the folding propensity of probe and target

$$Y^{P} \approx Y_{S}^{P} + Y_{NS}^{P} + Y_{F}^{P} - Y_{\text{sat}}^{P} - Y_{\text{fold}}^{P} - Y_{\text{fold}}^{T} \qquad (2)$$

with

$$Y_{S}^{P} = \Delta \log[K^{P,S}(\xi^{P}\boldsymbol{\xi}^{T})]$$

$$Y_{NS}^{P} \approx \Delta \log[x^{S} + (1 - x^{S})\cdot r^{P}(\xi^{P})]$$

and the definition $\Delta \log(A) \equiv \log(A) - \langle \log(A) \rangle_{\text{set}}$ (see the accompanying paper[4] for a detailed description). The ratio

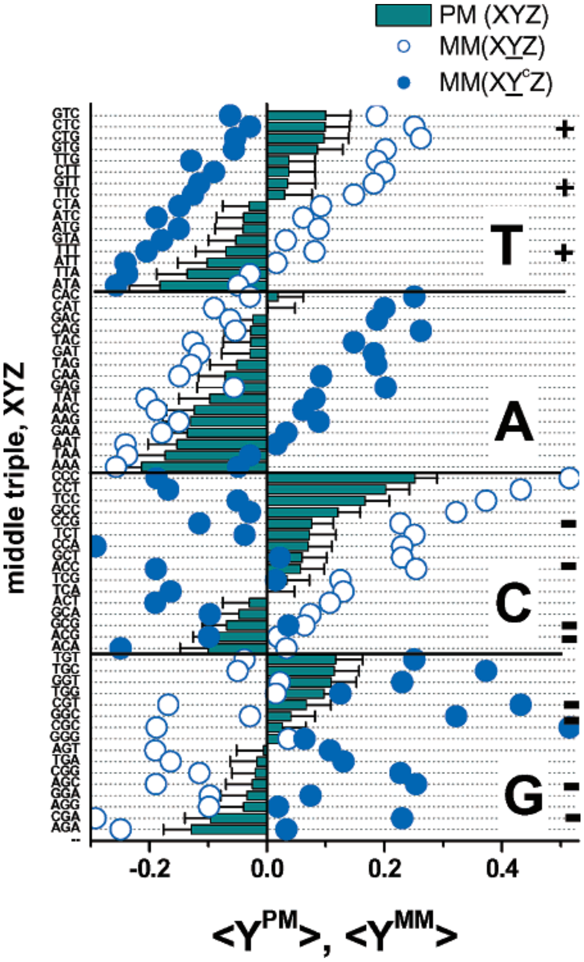

**Figure 2.** Mean sensitivities of all PM and MM probes with a given middle triple of letters (see left axis and figure for assignment). The data reflect the base-specific interactions in the middle of the target−probe duplexes which systematically affect the probe sensitivities. The combination of middle triples XYZ/X$\underline{Y}^c$Z refers to PM/MM probe pairs ($Y^c$ denotes the complementary nucleotide base of Y; underlined letters denote the mismatched middle base of the MMs). The probes are grouped for each middle letter of the PM sequence and ranked with increasing $\langle Y^{PM}\rangle$ within each group. The error bars are the standard errors referring to the averaging over 42 chips provided by the HG U133-LS experiment. The mean number of probes per triple on an HG U133 chip is 3900 with a standard deviation of ±1380. The signs − and + indicate triples with only 1100−1400 and with 6000−7400 probes, respectively.

$r^P(\xi^P) = \langle K^{P,NS}(\xi^P\xi)\rangle|_{\xi\neq\xi^T}/K^{P,S}(\xi^P\xi^T)$ specifies the relationship between the affinity of specific and nonspecific binding (see text to follow).

We expect that the difference between the sensitivities of the PM and MM probes of one pair

$$Y^{PM-MM}_{pair} \equiv Y^{PM} - Y^{MM} \approx Y^{PM-MM}_{S} + Y^{PM-MM}_{NS} \qquad (3)$$

essentially cancels out the contributions due to fluorescence and folding, because both probe sequences refer to one target. This expectation is confirmed by the observation that the mean standard deviation of the difference $\langle Y^{PM-MM}_{pair}(XYZ)\rangle$, $SD(Y^{PM-MM})_{triple} \approx 0.28$, is markedly smaller compared with the standard deviations of the individual PM and MM sensitivities, $SD(Y^{PM})_{triple} \approx 0.41$ and $SD(Y^{MM})_{triple} \approx 0.47$ (see also text to follow). Hence, systematic factors such as specific binding outside the middle triple, nonspecific binding, fluorescence, and folding indeed affect $Y^{PM}$ and $Y^{MM}$ in a similar fashion. The variability due to these effects is considerably reduced in the sensitivity difference $Y^{PM-MM}$. A similar tendency is expected for the triple averaged PM−MM differences of PM and MM probes with identical middle triples (see eq 5) and for the PM−PM and MM−MM differences of probes with complementary middle bases (see eqs 6 and 7).

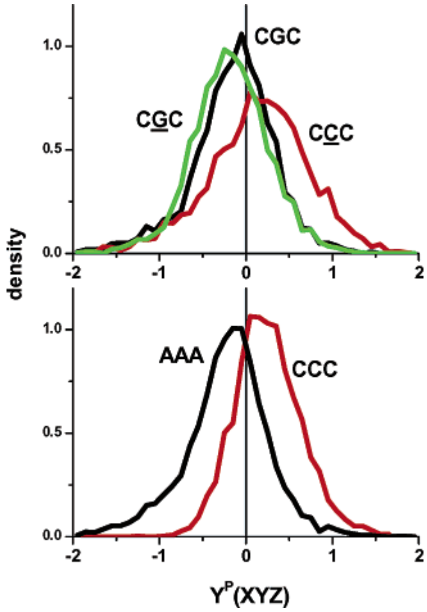

**Figure 3.** Probability density distribution of the sensitivity of probes with certain middle triples. PM probes with the middle triples AAA and CCC (panel below); PM probes with the triple CGC and MM probes with C$\underline{G}$C and C$\underline{C}$C (panel above).

The sample solution used for hybridization contains a total RNA concentration of $c^{tot}_{RNA}$. Only a fraction $x^S = x^S(\xi^T)$ refers to target RNA, whereas $x^{NS}(\xi \neq \xi^T) = 1 - x^S$ defines the fraction of nonspecific transcripts involving sequences other than the intended target. The binding constant of the PMs, $K^{PM,S}(\xi^{PM}\xi^T)$, quantifies the affinity of specific binding between the target and the probe with sequences $\xi^T$ and $\xi^{PM}$, respectively. Both sequences are complementary in terms of WC base pairs. The binding constant of the MMs, $K^{MM,S}(\xi^{MM}\xi^T)$, characterizes the binding affinity of target RNA, which specifically binds to the respective MM probe despite the fact that the middle base disables WC pairing with the target. Instead, the 13th base is assumed to form the respective SC pair, $\underline{A}$−$\underline{a}$, $\underline{T}$−$\underline{u}$*, $\underline{G}$−$\underline{g}$, or $\underline{C}$−$\underline{c}$* (see Figure 1: uppercase letters refer to the DNA probe and lower case letters to the RNA oligonucleotide; potentially labeled bases are indicated by asterisks).

The ratio $r^P(\xi^P) = \langle K^{P,NS}(\xi^P\xi)\rangle|_{\xi\neq\xi^T}/K^{P,S}(\xi^P\xi^T)$ specifies the mean binding affinity of nonspecific transcripts relative to the affinity of specific binding.[4] The binding constant of nonspecific hybridization, $\langle K^{P,NS}(\xi^P\xi)\rangle|_{\xi\neq\xi^T}$, represents the concentration-weighted average over the binding constants of an ensemble of RNA sequences, $\xi \neq \xi^T$, which only partly match the probe sequence $\xi^P$ by complementary bases. The stability of these duplexes is mainly determined by the number of remaining WC base pairings between the probe and bound RNA according to the conventional theory of hybridization on microarrays.[15] The nonspecific transcripts are taken from a cocktail of RNA fragments with a broad distribution of base composition, which enables WC pairings with the middle bases of the PM and the MM probes despite the fact that the respective bases are complementary (see Figure 1). In other words, we assume that the hybridization solution contains a sufficiently large number of different sequences that partially match the probe sequences

TABLE 1: Mean Probe Sensitivities as a Function of the Middle Base and Triple

| probe (triple) | central base pair | A, $\underline{A}^c$ | T, $\underline{T}^c$ | G, $\underline{G}^c$ | C, $\underline{C}^c$ |
|---|---|---|---|---|---|
| PM(XYZ)[a]<br>$\langle Y^{PM}(XYZ)\rangle$ | WC | A−u*<br>−0.08 ± 0.05 | T−a<br>−0.01 ± 0.05 | G−c*<br>+0.01 ± 0.1 | C−g<br>+0.05 ± 0.1 |
| MM(X$\underline{Y}$Z)[a]<br>$\langle Y^{MM}(X\underline{Y}Z)\rangle$ | WC/SC | $\underline{A}$−u*/$\underline{A}$−a<br>−0.14 ± 0.1 | $\underline{T}$−a/$\underline{T}$−u*<br>+0.12 ± 0.1 | $\underline{G}$−c*/$\underline{G}$−g<br>−0.11 ± 0.1 | $\underline{C}$−g/$\underline{C}$−c*<br>+0.21 ± 0.2 |
| PM(XYZ)-MM(X$\underline{Y}$Z)[b]<br>$\langle Y^{PM-MM}\rangle$ | WC−WC/SC | A−u* − $\underline{A}$−y<br>u*/a → u*<br>+0.02 ± 0.02 | T−a − $\underline{T}$−y<br>a/u* → a<br>−0.07 ± 0.02 | G−c* − $\underline{G}$−y<br>c*/g → c*<br>+0.08 ± 0.04 | C−g − $\underline{C}$−y<br>g/c* → g<br>−0.10 ± 0.04 |
| PM(XYZ)-MM(X$\underline{Y}^c$Z)[b]<br>$\langle Y_{pair}^{PM-MM}\rangle$ | WC−WC/SC | A−u* − $\underline{T}$−y<br>$\underline{T}$ → A<br>−0.12 ± 0.04 | T−a − $\underline{A}$−y<br>$\underline{A}$ → T<br>+0.08 ± 0.03 | G−c* − $\underline{C}$−y<br>$\underline{C}$ → G<br>−0.13 ± 0.1 | C−g − $\underline{G}$−y<br>$\underline{G}$ → C<br>+0.11 ± 0.1 |
| PM(XYZ)-PM(XY$^c$Z)[b]<br>$\langle Y^{PM-PM}\rangle$ | WC−WC | A−u* − T−a | T−a − A−u*<br>+0.17 ± 0.1 | G−c* − C−g | C−g − G−c*<br>+0.08 ± 0.2 |
| MM(X$\underline{Y}$Z)-MM(X$\underline{Y}^c$Z)[b]<br>$\langle Y^{MM-\overline{MM}}\rangle$ | WC/SC−WC/SC | $\underline{A}$−y − $\underline{T}$−y | $\underline{T}$−y − $\underline{A}$−y<br>+0.26 ± 0.05 | $\underline{G}$−y − $\underline{C}$−y* | $\underline{C}$−y − $\underline{G}$−y<br>+0.32 ± 0.2 |

[a] Combination of middle bases in probe/target duplexes of PM and MM probes. The middle triple is given by the bases XYZ with X, Y, Z = A, T, G, C. The superscript c indicates the complementary nucleotide base, $Y^c$ = A, T, G, C for Y = T, A, C, G. Watson−Crick base pairs in the DNA/RNA duplexes are denoted by Y−y$^c$ where uppercase letters refer to the DNA probe and lower case letters refer to the RNA target. The DNA letters of the MM probes are underlined. They are assumed to form self-complementary (SC) base pairs of the type $\underline{Y}$−y with specific RNA transcripts (S) or, alternatively, WC pairs with nonspecific transcripts (NS) of the type $\underline{Y}$-y$^c$. The labeled nucleotides of the target are indicated by the asterisk, i.e., y* = c*,u*. [b] Combination of middle pairs which refer to the considered sensitivity differences of PM and MM probes shown in Figures 5 and 6. The representation of the form y$^c$/y → y$^c$and Y$^c$ → Y indicates the effective change in the middle triples referring to the respective sensitivity differences. [c] The values are triple-averaged sensitivities over all probes with the respective middle base (see first row) and the respective standard deviation.

via WC pairings including their central bases. The binding constant of nonspecific hybridization is therefore related to WC pairings for PM and MM probes as well. This interpretation is compatible with our observation that the relationship between the base composition and the sensitivity of a probe is virtually identical for PMs and MMs.[4]

Let us neglect saturation and folding for the sake of simplicity. Then, the triple-averaged sensitivity of the MM probes is given by the concentration-weighted mean affinity over central SC and WC pairs owing to specific and nonspecific hybridization, respectively, whereas only WC pairs contribute to the respective triple averages of the PM probes

$$\langle Y^{PM}(XYZ)\rangle \approx \langle Y_S^{PM} + Y_{NS}^{PM}\rangle_{XYZ} =$$
$$\langle \Delta\log(K^{PM,S}) + \Delta\log[x^S + (1-x^S)\cdot r^{PM}]\rangle_{XYZ} \propto$$
$$\Delta G^{WC}(XYZ)$$
$$\langle Y^{MM}(XYZ)\rangle \approx \langle Y_S^{MM} + Y_{NS}^{MM}\rangle_{XYZ} =$$
$$\langle \Delta\log(K^{MM,S}) + \Delta\log[x^S + (1-x^S)\cdot r^{MM})\rangle_{XYZ} \propto$$
$$[x_{eff}^S\cdot\Delta G^{SC}(X\underline{Y}Z) + (1-x_{eff}^S)\cdot\Delta G^{WC}(X\underline{Y}Z)] \qquad (4)$$

where $\Delta G^{WC}(XYZ) = \Delta G^{WC}(X\underline{Y}Z)$ and $\Delta G^{SC}(X\underline{Y}Z)$ are the relative binding affinities of the respective triples in terms of the Gibbs free energy referring to central WC (Y−y$^c$) and SC ($\underline{Y}$−y) pairs, respectively (see Figure 1, underlined letters indicate the central base of the MM). The free-energy terms are weighted by the effective fraction of specific transcripts, $x_{eff}^S$. The considered averages consequently provide a measure of the relative affinity of the respective trimeric fragments for duplex formation in terms of WC and SC pairs within longer oligonucleotide sequences.

## Results and Discussion

**Triple-Averaged Probe Sensitivities.** Figure 2 shows the mean sensitivities of PM and MM probes which have been averaged over all sequences with a common middle triple XYZ (see also footnotes in Table 1 for further conventions of sequence assignments and probe−target base pairings). The data are grouped for each middle base, Y = A, T, G, C and ranked with respect to the PM values. The respective PM averages reveal an increasing sensitivity level according to Y = A < T ≈ G < C. Note that the mean PM sensitivities with the middle letter A are always negative, whereas those with middle letter C are preferentially positive. Maximum and minimum mean sensitivities of $\langle Y^{PM}(CCC)\rangle = +0.25 \pm 0.04$ and $\langle Y^{PM}(AAA)\rangle = -0.21 \pm 0.05$ are found for CCC and AAA homotriples, respectively. The uncertainty of the data was estimated in terms of the standard error of the 42 spiked-in probes.

Note also that within each group the minimum PM sensitivities are found for AYA (Y = A, T, G, C) triples whereas the largest sensitivities refer to CYC triples except for Y = G. For the middle base G, the maximum values are found for TGT ($\langle Y^{PM}(TGT)\rangle = +0.12 \pm 0.05$); whereas CGC gives rise to a distinctly decreased mean sensitivity ($\langle Y^{PM}(CGC)\rangle = +0.03 \pm 0.04$). The latter result confirms the unfavorable effect of neighboring G and C on the sensitivity stated previously.[4]

In Figure 2, each PM sensitivity with the middle triple XYZ is directly compared with the sensitivity of two different MM probes, namely the MM probe with the same middle triple X$\underline{Y}$Z ($\langle Y^{MM}(X\underline{Y}Z)\rangle$) and the MM probe with the middle triple X$\underline{Y}^c$Z ($\langle Y^{MM}(X\underline{Y}^cZ)\rangle$, $Y^c$ denotes the complementary base for Y). The latter MM probe belongs to the same probe pair as the considered PM probe (see also Figure 1 for illustration). Note that the sequence of these MM probes agrees with that of the respective PM probe except for the middle base. In contrast, the sequence of the former MM probe typically differs from that of the respective PM probe except for the middle triple. Table 1 gives an overview of the respective base pairings in the middle of the probe−target duplex.

The probability density distribution of the probe sensitivities referring to one middle triple is well described by a Gaussian shape of similar width for the PM and MM averages of all considered middle triples. Some examples are shown in Figure 3. The distribution width is characterized by a standard deviation of $SD(Y^P)_{triple} < 0.5$, which seems to be large even when compared with the maximum difference observed between the mean sensitivity values of CCC and AAA, $\langle \Delta Y^{CCC-AAA}\rangle =$

$\langle Y^{\mathrm{PM}}(\mathrm{CCC})\rangle - \langle Y^{\mathrm{PM}}(\mathrm{AAA})\rangle < 0.6$. Application of the two-sample t-test to the difference between the triple averages for the 42 chips of the LS experiment provides a significance level of 0.02 for the mean difference $\langle \Delta Y^{\mathrm{CCC-AAA}}\rangle$. Analyses of the differences between the other triples provide similar results. Hence, differences between the triple averages greater than 0.02 might be judged as significant. Note that the standard error of the individual triple values, SE($Y$) ≈ 0.05, is more than twice as large as the standard error of the difference, SE($\Delta Y$) ≈ 0.02, owing to correlations between the intensity values on each chip, which contribute to SE($Y$) but not to SE($\Delta Y$). Consequently, SE($Y$) represents a maximum error estimate of the triple averages (see error bars in Figure 2).

**Sensitivity Differences of PM and MM Triple Averages.** Both types of mean MM sensitivities deviate from the respective PM sensitivity into opposite directions, and in most cases, they differ even in sign (compare open and solid symbols with the bars in Figure 2). The differences between the mean PM and MM sensitivities

$$\langle Y^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle = \langle Y^{\mathrm{PM}}(\mathrm{XYZ})\rangle - \langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y}Z})\rangle$$

and

$$\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle = \langle Y^{\mathrm{PM}}(\mathrm{XYZ})\rangle - \langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y}^{c}Z})\rangle \quad (5)$$

reveal this behavior more clearly (Figure 4). The sign of $\langle Y^{\mathrm{PM-MM}}\rangle$ strongly correlates with the middle base of the PM probe. As a rule of thumb, the middle letters A and G give rise to positive mean sensitivity differences between PMs and MMs with common middle triples, XYZ/X$\underline{\mathrm{Y}}$Z, whereas T and C cause negative values of $\langle Y^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle$ (see open symbols in Figure 4 and Table 1). The relationship reverses for the sensitivity differences, $\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle$, between the PMs and MMs of one probe pair with middle triples XYZ/X$\underline{\mathrm{Y}}^{c}$Z (note that the figure reverses sign for $\langle Y^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle$ for direct comparison with $\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle$). This result is compatible with the preference of middle letters A and G for sensitive MMs, $Y^{\mathrm{PM}} < Y^{\mathrm{MM}}$, and vice versa, the preference of T and C for sensitive PMs, $Y^{\mathrm{PM}} > Y^{\mathrm{MM}}$.[4]

In addition, we calculated the sensitivity difference between the PMs with complementary middle bases

$$\langle Y^{\mathrm{PM-PM}}(\mathrm{XYZ})\rangle = \langle Y^{\mathrm{PM}}(\mathrm{XYZ})\rangle - \langle Y^{\mathrm{PM}}(\mathrm{XY^{c}Z})\rangle \quad (6)$$

and the respective sensitivity difference between the MMs (see Figure 5 and Table 1)

$$\langle Y^{\mathrm{MM-MM}}(\mathrm{X\underline{Y}Z})\rangle = \langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y}Z})\rangle - \langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y}^{c}Z})\rangle \quad (7)$$

Only two middle letters, T and C, have been considered in Figure 5 owing to the symmetry $\langle Y^{\mathrm{P-P}}(\mathrm{XYZ})\rangle = -\langle Y^{\mathrm{P-P}}(\mathrm{XY^{c}Z})\rangle$.

The sign of the PM−PM and MM−MM sensitivity differences correlate with the middle letter in a similar fashion as the PM−MM heterodifferences (Figure 4). Moreover, comparison of Figures 5 and 6 reveals a similar effect of the adjacent bases on the calculated sensitivity differences. For example, the triples CCC and GCG provide the largest and smallest values for all considered sensitivity differences with middle letter C. The sensitivity values are obviously related to the base in the middle of the probe−target duplex (see Table 1 for an overview and assignments). Most of the positive $\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle$ and $\langle Y^{\mathrm{P-P}}(\mathrm{XYZ})\rangle$ data correspond to pyrimidines (C and T), whereas negative differences are found for purines (G and A).

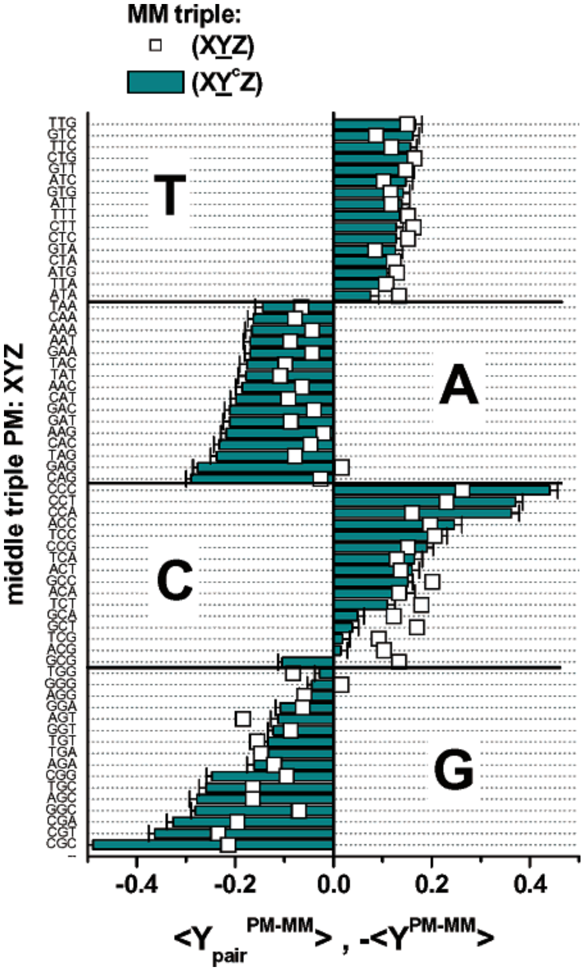

**Figure 4.** Mean sensitivity difference of PM and MM probes, $\langle Y^{\mathrm{PM-MM}}\rangle$ and $\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}\rangle$ (see eq 5). The sign of the data strongly correlates with the middle base. The PM/MM couples refer either to one probe pair with the combination of middle letters XYZ/X$\underline{\mathrm{Y}}^{c}$Z (bars) or to different probe pairs with PMs and MMs possessing identical middle triples (XYZ/X$\underline{\mathrm{Y}}$Z, squares). The middle triples are grouped for each PM middle letter (see figure) and ranked with respect to $\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}\rangle$. Note that the figure shows negative values of $\langle Y^{\mathrm{PM-MM}}\rangle$ for direct comparison with $\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}\rangle$. The error bars refer to the standard error of 42 chips. See text and Table 1 for further explanation.

**Molecular Interactions in Probe−Target Duplexes.** The sensitivity difference between the PMs with complementary middle bases, $\langle Y^{\mathrm{PM-PM}}(\mathrm{XYZ})\rangle$ (eq 6), directly compares the strengths of complementary WC base pairs in DNA/RNA hybrid duplexes (see Figure 5 and Table 1). For example, $\langle Y^{\mathrm{PM-PM}}(\mathrm{XCZ})\rangle$ compares the strength of C−g with that of G−c* in the respective triples. The middle-base-related sign of the $\langle Y^{\mathrm{PM-PM}}(\mathrm{XYZ})\rangle$ data therefore reflects the following relationships between the interaction strengths: T−a > A−u* and, partially, C−g > G−c*. The asymmetry can be, at least partially, explained by the biotinilation and labeling of the pyrimidines c* and u* in the RNA sequence. The labels obviously hamper the formation of WC pairs.

This suggestion is further confirmed by the observation that the sensitivity difference reverses and becomes negative for nonlabeled middle pairs (i.e., C−g < G−c*) if both nearest neighbors became labeled in the triples GCG/c*gc*, TCT/u*gu*, TCG/u*gc*, and GCT/c*gu*. In other words, in these triples, the G−c* pair adjacent to the central C−g obviously decreases the sensitivity if two labeled bases flank the nonlabeled central base in the target sequence. On the other hand, the opposite

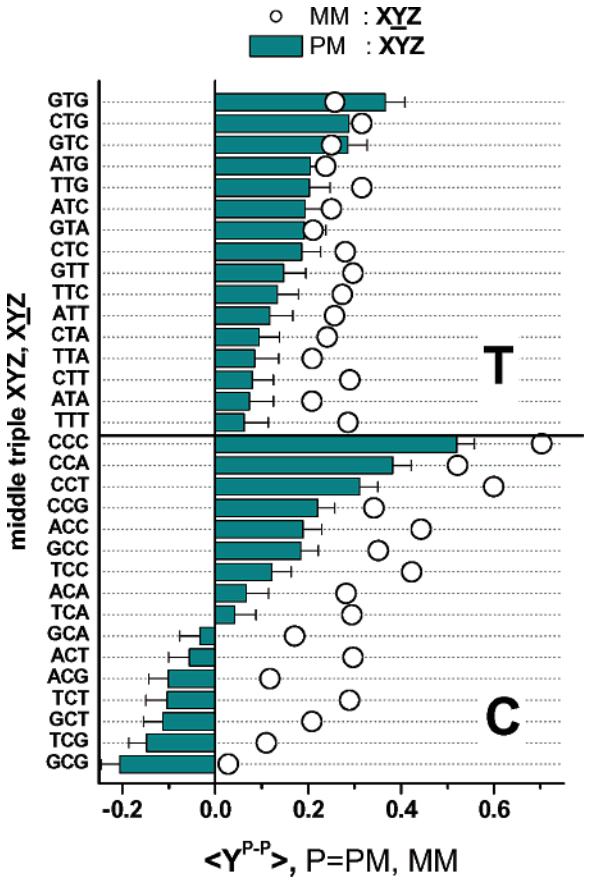

**Figure 5.** Mean sensitivity difference of PM and of MM probes with middle triples XYZ and XY$^c$Z (see eqs 6 and 7). The differences $\langle Y^{\mathrm{PM-PM}}(\mathrm{XYZ})\rangle$ compare the strength of WC base pairs of complementary letters, e.g. T−a with A−u* (panel above) and C−g with G−c* (panel below). Only middle letters T and C are considered because of symmetry reasons. The PM data are ranked by ascending sensitivity difference for each middle letter.

tendency is observed for labeled G−c* pairs adjacent to the nonlabeled middle pair in GTG/c*ac*. Neighboring TG behaves obviously differently compared with adjacent CG and CT.

The sensitivity differences between the MMs with complementary middle bases, $\langle Y^{\mathrm{MM-MM}}(\mathrm{X\underline{Y}Z})\rangle$, depend in a very similar fashion on the middle base as the respective PM difference (see Figure 5). This agreement can be rationalized if the central base in the MM−transcript duplex mainly forms WC pairings with the bound RNA as in the case of the PMs. Note that the formation of central WC pairings by the MMs is expected if nonspecific hybridization dominates duplex formation ($x_{\mathrm{eff}}^{\mathrm{S}} \ll 1$, see eq 4). From the chip-averaged intensity difference between PM and MM probes, we concluded that most of the probes of the considered chips are nonspecifically hybridized (see ref 4) in agreement with our interpretation of the sensitivity difference of the MM probes. Alternatively, a similar behavior of MM and PM sensitivity differences turns out if the binding strength of the central SC pair in specific duplexes is relatively weak ($|\Delta G^{\mathrm{SC}}(\mathrm{XYZ})| \ll |\Delta G^{\mathrm{WC}}(\mathrm{XYZ})|$, see eq 4). In this case, the SC pair virtually does not contribute to the respective triple-averaged sensitivity value.

The heterodifferences $\langle Y_{\mathrm{pair}}^{\mathrm{PM-MM}}(\mathrm{XYZ})\rangle$ highly correlate with the respective homodifferences $\langle Y^{\mathrm{P-P}}(\mathrm{XYZ})\rangle$ (see Figures 4−6). This correspondence can be simply explained if the central base pairings in the PM and MM probes are dominated by WC pairings, in agreement with our interpretation given already. Hence, the difference between the PM and MM sensitivities of one probe pair mainly reflects the Y−y$^c$ versus Y$^c$−y asymmetry of the WC pairings (e.g., C−g versus G−c*).

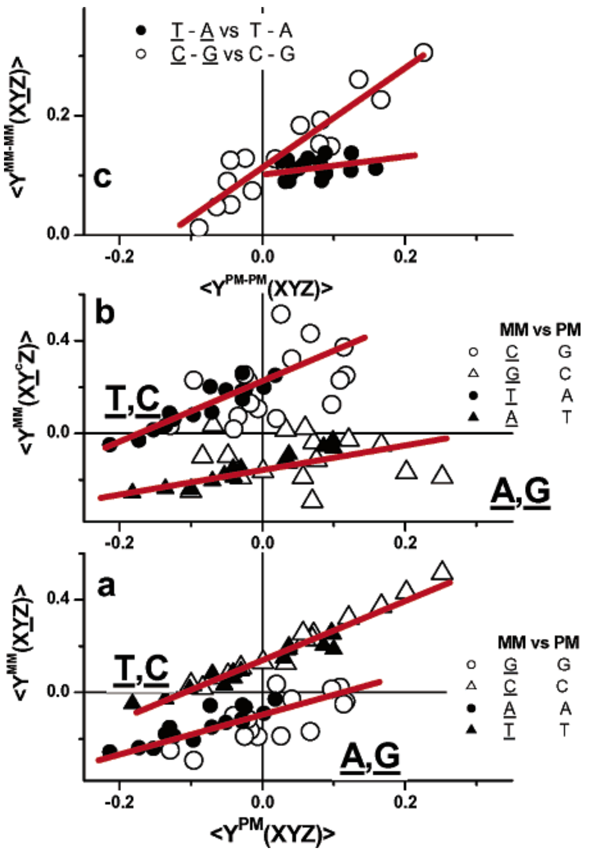

**Figure 6.** Correlation plots between the triple-averaged MM and PM sensitivities shown in Figure 2 (panels a and b) and between differences of MM and PM sensitivities with changed middle letter shown in Figure 5 (panel c). The degree of correlation reflects the effect of the nearest neighbors adjacent to the central base pair on the mean sensitivities. Panel a refers to the same middle base in the PM and MM sequences, whereas part b refers to the complementary middle bases. Panel c correlates differences between the complementary bases in PM and MM probes. The assignments of the symbols are given within the figure. Underlined letters specify the central base in the MM sequence (see also Table 1). The lines refer to linear fits to the data (see text).

**Correlation Between the Middle Triple Averages of PM and MM Sensitivities: The Effect of Mismatches.** The plot of ranked sensitivities shown in Figure 2 provides a first impression of the degree of correlation between the PM and MM averages for each middle letter. For example, $\langle Y^{\mathrm{PM}}(\mathrm{XYZ})\rangle$ and $\langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y}Z})\rangle$ run almost parallel for middle bases T, A, C, and partially G. The correlation between $\langle Y^{\mathrm{PM}}(\mathrm{XYZ})\rangle$ and $\langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y^c}Z})\rangle$ is, however, much weaker for the latter two middle letters G and C. Note that middle letters C and G generally give rise to the considerably wider scattering of all triple averages as a function of the nearest neighbors compared with the middle letters T and A (see Figures 1, 3, and 4).

Correlation plots more clearly reveal the characteristic relationships between the PM and MM averages (see Figure 6, parts a and b). The data can be divided into two groups dependent on the middle letter of the MM probes. For pyrimidines, T and C, the mean MM sensitivities always exceed those of the PMs in correspondence with the results of the previous section (i.e., $\langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y^c}Z})\rangle > \langle Y^{\mathrm{PM}}(\mathrm{XYZ})\rangle$ (with $\underline{\mathrm{Y}}^{\mathrm{c}} = \underline{\mathrm{T}}, \underline{\mathrm{C}}$). For purines, A and G, one obtains in nearly all cases the reverse relationship owing to the reversal of the respective substitution of the middle base (vide supra and Table 1).

Separate linear fits of the form $y = kx + \delta$ (with $x \equiv \langle Y^{PM}(XYZ)\rangle$ and $y \equiv \langle Y^{MM}(X\underline{Y}Z)\rangle/\langle Y^{MM}(X\underline{Y}^{c}Z)\rangle$) to each of the two groups of data provide slopes and intercepts of $k \approx 1.1 \pm 0.1$ and $\delta \approx +0.15/+0.20$ ($\pm 0.03$) for T and C, respectively, and $k \approx 0.6 \pm 0.1$ and $\delta \approx -0.10/-0.15$ ($\pm 0.03$) for A and G, respectively. In other words, the sensitivities of MMs of the first group with the central pyrimidines $\underline{T}$ (and $\underline{C}$), on the average, change in a similar fashion to the respective PM sensitivities (see panels a and b of Figure 6). Consequently, both the PM and MM sensitivities of this group are affected to a similar extent by factors which modulate the probe sensitivity, such as the nearest neighbors of the central base pair. The same conclusion can be derived from the sensitivities of the second group with central purines $\underline{A}$ and $\underline{G}$ in the MM sequences. The smaller slope, $k = 0.6$, however, indicates that the MM sensitivities with central $\underline{A}$ (and $\underline{G}$) depend to a considerably smaller extent on these factors, compared with the respective PM sensitivities.

In part c of Figure 6, we correlate the MM sensitivity difference between pyrimidine and purine middle bases, $\langle Y^{MM}(X\underline{Y}Z)\rangle - \langle Y^{MM}(X\underline{Y}^{c}Z)\rangle$ (with $\underline{Y} = \underline{T}, \underline{C}$), with the respective triple-averaged PM sensitivity differences, $\langle Y^{PM}(XYZ)\rangle - \langle Y^{PM}(XY^{c}Z)\rangle$ (Y = T, C). The slopes of the linear fits to the data show that the considered differences between the MMs with central $\underline{Y} = \underline{C}$ (and $\underline{Y}^{c} = \underline{G}$) are affected by their nearest neighbors in a similar fashion as the respective difference between the PMs with central C (and G, $k = 0.8$). In contrast, the difference between MMs with central $\underline{T}$ (and $\underline{A}$) to a considerably lesser degree correlates with the respective difference between PM pairs with central T (and A, $k = 0.2$).

**Nearest-Neighbor Interactions: Excess Sensitivity and 3′ → 5′ Asymmetry.** Stacking interactions between nearest neighbors within the RNA and DNA sequence make an important contribution to the stability of probe−target duplexes.[8−10] The interaction strength between adjacent nucleotide bases can be estimated on a relative scale by means of the so-called excess values making use of the averaged sensitivities of symmetrical triples XYX/YXY

$$\langle Y^{PM}_{exc}(XY)\rangle = \langle Y^{PM}_{exc}(YX)\rangle = (^{1}/_{4})\{\langle Y^{PM}(XYX)\rangle + \langle Y^{PM}(YXY)\rangle - [\langle Y^{PM}(XXX)\rangle + \langle Y^{PM}(YYY)\rangle]\} \quad (8)$$

The excess sensitivity provides a measure of the deviation of the cross interaction in the XY heterocouple of nearest neighbors from the additivity rule for the interactions in the respective homocouples, XX and YY.

Negative excess values are obtained for the symmetrical PM triples ACA and GCG (see panel a of Figure 7). This result shows that the heterocouples AC/CA and GC/CG decrease the sensitivity compared with the arithmetic mean of the sensitivities of the respective homocouples. In contrast, the triple TGT gives rise to a positive excess sensitivity, and thus, TG/GT heterocouples enhance the probe sensitivity on a relative scale.

The sensitivities of asymmetrical triples can be analyzed in terms of the 3′ → 5′ or left−right asymmetry

$$\langle Y^{PM}_{asym}(XY)\rangle = (^{1}/_{2})\{[\langle Y^{PM}(XXY)\rangle - \langle Y^{PM}(YXX)\rangle] + [\langle Y^{PM}(XYY)\rangle - \langle Y^{PM}(YYX)\rangle]\} \quad (9)$$

which provides a measure of the sensitivity change owing to the reversal of order of neighboring letters, YX → XY. Note that the left letter points toward the 3′ end of the DNA probe attached to the chip surface. The negative asymmetry value of the couple CG indicates a sensitivity gain of the probe if the sequence CG reverses into GC (see Figure 7). For the other NN couples, one obtains the following relationship between the sensitivities CA > AC, CT > TC, AT > TA, TG ≈ GT, and GA ≈ AG.

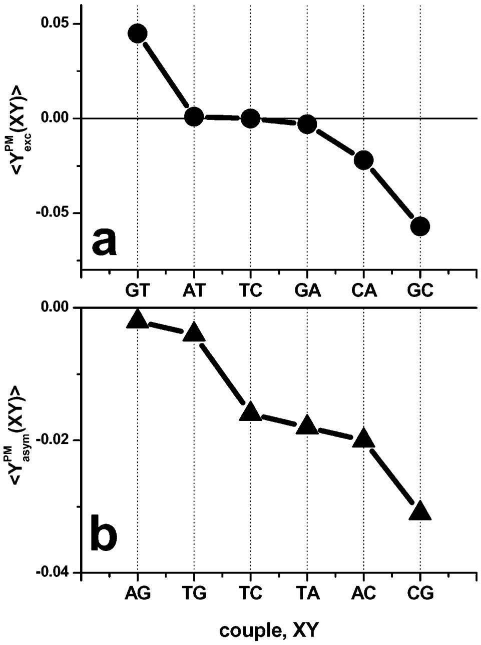

**Figure 7.** Excess sensitivities (part a) and left−right asymmetry (part b) of adjacent bases, XY. The values were derived from triple-averaged PM sensitivities using eqs 8 and 9, respectively.

**Nearest-Neighbor Sensitivity Terms.** The 64 triple-averaged sensitivity values, $\langle Y^{P}(XYZ)\rangle$, can be used to specify 16 NN sensitivity terms, $\langle Y^{P}(XY)\rangle$, in analogy with the NN energy contributions in models describing the stability of RNA/DNA oligonucleotide duplexes in solution (vide infra). For this purpose, we decompose the triple averages into two NN and two boundary terms, which consider the mean effect of the bases adjacent to the triple

$$\langle Y^{P}(XYZ)\rangle = \langle Y^{P}(XY)\rangle_{12,13} + \langle Y^{P}(YZ)\rangle_{13,14} + (^{1}/_{2})[\langle Y^{P}(X)\rangle + \langle Y^{P}(Z)\rangle] \quad (10)$$

According to eq 10, the triple data provide a system of 64 linear equations with 4 SB boundary terms and 32 NN terms referring to positions 12−13 and 13−14 of the sequence. The system of equations was solved by multiple linear regression using singular value decomposition.[16] The NN terms of the PM probes were aggregated into averages $\langle Y^{PM}(XY)\rangle = 0.5[\langle Y^{PM}(XY)\rangle_{12,13} + \langle Y^{PM}(XY)\rangle_{13,14}]$.

The upper panel of Figure 8 shows the NN sensitivity terms obtained from the PM and MM sensitivities. The maximum and minimum $\langle Y^{PM}(XY)\rangle$ values refer to XY = CC and AA homocouples, respectively. Note that the respective MM sensitivities, $\langle Y^{MM}(X\underline{Y})\rangle \equiv \langle Y^{MM}(X\underline{Y})\rangle_{12,13}$ and $\langle Y^{MM}(\underline{X}Y)\rangle \equiv \langle Y^{MM}(\underline{X}Y)\rangle_{13,14}$, are in a systematic fashion either bigger or smaller than the respective PM data. For pyrimidines in the middle of the sequence (Y = C, T), we obtained $\langle Y^{MM}(X\underline{Y})\rangle > \langle Y^{PM}(XY)\rangle$, whereas the PM values mostly exceed the respective MM sensitivities for purine middle letters, Y = G, A. This result corresponds to the classification of triple averages according to their middle base discussed already.

**Purine−Pyrimidine Asymmetry.** We found a strong middle-base-related purine−pyrimidine asymmetry of the triple averages. This trend can be partly related to the labels attached to the cytosines and uracils of the complementary RNA fragments

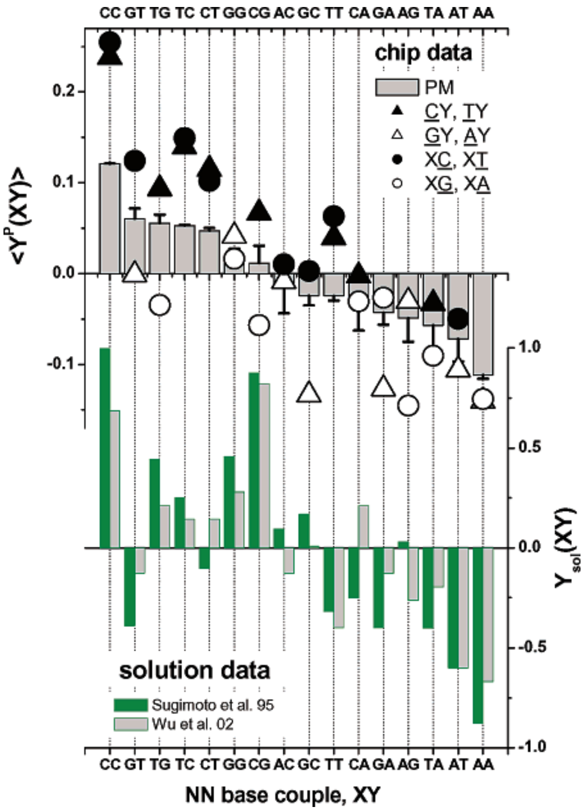

**Figure 8.** Nearest-neighbor pair sensitivities derived from triple-averaged PM and MM sensitivities of the HG U133-LS experiment (panel above, see eq 10) and the respective solution data which are calculated from NN free-energy terms (see eq 13) taken from refs 10, 11. Both sets of NN sensitivity terms roughly rank in the same order as a function of adjacent base couples. See the figure for assignment.

(see previous text). This purine−pyrimidine asymmetry can be further specified by the difference of sensitivities corresponding to nearest neighbors, which differ exactly by one pyrimidine according to

$$\langle Y^{\mathrm{PM}}_{\mathrm{RY}}(\mathrm{XY})\rangle = ({}^1/_2)\{\langle Y^{\mathrm{PM}}(\mathrm{XY})\rangle + \langle Y^{\mathrm{PM}}(\mathrm{YX})\rangle - [\langle Y^{\mathrm{PM}}(\mathrm{XY^c})\rangle + \langle Y^{\mathrm{PM}}(\mathrm{Y^cX})\rangle]\}$$

$$\langle Y^{\mathrm{MM}}_{\mathrm{RY,13}}(\mathrm{X\underline{Y}})\rangle = \langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y}})\rangle - \langle Y^{\mathrm{MM}}(\mathrm{X\underline{Y}^c})\rangle$$

and

$$\langle Y^{\mathrm{MM}}_{\mathrm{RY,WC}}(\mathrm{Y\underline{X}})\rangle = \langle Y^{\mathrm{MM}}(\mathrm{Y\underline{X}})\rangle - \langle Y^{\mathrm{MM}}(\mathrm{Y^c\underline{X}})\rangle \qquad (11)$$

with X = A, T, G, C; $\underline{\mathrm{X}}$ = $\underline{\mathrm{A}}$, $\underline{\mathrm{T}}$, $\underline{\mathrm{G}}$, $\underline{\mathrm{C}}$; Y = A, G; $\mathrm{Y^c}$ = T, C; $\underline{\mathrm{Y}}$ = $\underline{\mathrm{A}}$, $\underline{\mathrm{G}}$; and $\underline{\mathrm{Y}}^{\mathrm{c}}$ = $\underline{\mathrm{T}}$, $\underline{\mathrm{C}}$.

Note that eq 11 expresses the change of sensitivity upon the replacement of a nonlabeled by a labeled base pair in the respective couples. The PM sensitivity difference refers to two adjacent WC pairs, whereas the MM data split into two options. The first one, $\langle Y^{\mathrm{MM}}_{\mathrm{RY,13}}(\mathrm{X\underline{Y}})\rangle$, assesses the effect of a pyrimidine in the central sequence position, which forms an SC or a WC pair in specific and nonspecific duplexes, respectively. The second option, $\langle Y^{\mathrm{MM}}_{\mathrm{RY,WC}}(\mathrm{Y\underline{X}})\rangle$, estimates the sensitivity change due to the pyrimidine in the WC pairs adjacent to the central base.

The values of the three calculated differences are plotted in Figure 9 for G (left part) and A (right part) in the probe sequence. The negative values for most of the couples indicate that the replacement of a purine by a pyrimidine predominantly decreases the sensitivity of the probes. For example, the G in XY = CG diminishes the sensitivity by about −0.1 compared with CC, which refers to the decrease of intensity by about 20%. However, the sensitivity change after the substitution C → G considerably depends on the neighboring base and even becomes positive for XY = GG. On the other hand, the replacement T → A is much less sensitive to adjacent nucleotides.

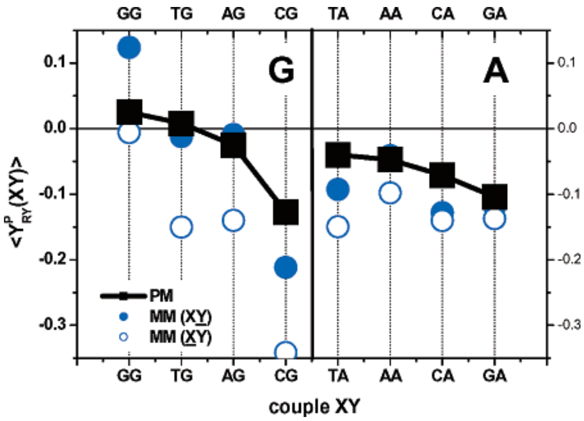

**Figure 9.** Sensitivity increment of purine−pyrimidine replacement for the couples of nearest neighbors, XY^c → XY. The sensitivity of base pairs with guanine strongly depends on their nearest neighbor, whereas pairs with adenine are relatively insensitive to adjacent bases. The values were derived from triple-averaged PM and MM sensitivities using eq 11.

**Comparison with Solution Data.** The sensitivity of oligonucleotide probes is scaled in terms of free energy according to $\Delta G = G - \langle G\rangle_{\mathrm{set}} = -RT{\cdot}\ln 10{\cdot}Y^{\mathrm{P}}_{\mathrm{p}}$ (see eq 13 in ref 4). After chip averaging for each triple, this equation rewrites into

$$\langle \Delta G(\mathrm{XYZ})\rangle_{\mathrm{chip}} = \langle G(\mathrm{XYZ})\rangle_{\mathrm{chip}} - \langle\langle G\rangle_{\mathrm{set}}\rangle_{\mathrm{chip}} = -RT{\cdot}\ln 10{\cdot}\langle Y^{\mathrm{P}}(\mathrm{XYZ})\rangle_{\mathrm{chip}} \qquad (12)$$

In other words, the chip average of the sensitivity for each triple provides a measure of the increment of free energy with respect to the chip-averaged free energy of duplex formation. Note that the chip average roughly agrees with the average over a reservoir of probes with equally and randomly distributed nucleotide bases (i.e., $\langle\langle\cdots\rangle_{\mathrm{set}}\rangle_{\mathrm{chip}} \approx \langle\cdots\rangle_{\mathrm{random}}$; see Appendix).

The stability of DNA/RNA oligonucleotide duplexes in solution is well described by means of NN models which decompose the free energy of duplex formation into a sum of NN terms, $G_{\mathrm{sol}}(\mathrm{XY})$, and an initiation energy (e.g., refs 10 and 17). These NN energies are functions of neighboring base pairs, XY (X, Y = A, T, G, C). For comparison of the chip data, $\langle Y^{\mathrm{P}}(\mathrm{XY})\rangle$, with the NN free-energy contributions obtained from studies on oligonucleotide duplex stability in solution, we transform the latter data according to the following (compare with eq 12)

$$Y_{\mathrm{sol}}(\mathrm{XY}) = \frac{-1}{RT\ln 10}[G_{\mathrm{sol}}(\mathrm{XY}) - \langle G_{\mathrm{sol}}(\mathrm{XY})\rangle_{\mathrm{total}}]$$

with

$$\langle G_{\mathrm{sol}}(\mathrm{XY})\rangle_{\mathrm{total}} = \frac{1}{16}\sum_{\mathrm{X,Y=A,T,G,C}} G_{\mathrm{sol}}(\mathrm{XY}) \qquad (13)$$

Figure 9 compares chip and solution NN sensitivities, which are ranked in the same order. This representation reveals that both types of data are well correlated. On the other hand, the

**TABLE 2: Single Base Sensitivity Terms[a]**

| probe, P | $\langle Y^{P}(C)\rangle$ | $r_G^P$ | $r_T^P$ | $r_A^P$ | $r_C^P - r_G^P$ | $r_T^P - r_A^P$ | $r_C^P - r_T^P$ | $r_G^P - r_A^P$ |
|---|---|---|---|---|---|---|---|---|
| PM | 0.036 | 0.15 | 0.15 | −1.70 | 0.85 | 1.85 | 0.90 | 1.85 |
| MM (WC) | 0.063 | 0.55 | 0.25 | −1.25 | 0.45 | 1.50 | 0.75 | 1.80 |
| MM | 0.10 | −0.55 | 0.60 | −0.70 | 1.55 | 1.25 | 0.40 | 0.15 |
| soln[10 b] | 0.39 | 0.53 | −0.47 | −1.07 | 0.47 | 0.6 | 1.47 | 1.60 |
| soln[11 b] | 0.32 | 0.42 | −0.47 | −0.95 | 0.58 | 0.48 | 1.47 | 1.37 |

[a] The ratios are defined as $r_x^P = \langle Y^P(X)\rangle/\langle Y^P(C)\rangle$; note that $r_C^P = 1$ and SE of the $r$ data is ±0.1. [b] Solution data are taken from refs 10, 11. Free-energy terms are converted into single-base sensitivity terms using eq 13.

solution and chip data for some of the labeled couples deviate markedly from each other on a relative scale. The largest relative difference is found for the couple CG.

The correlation plot between solution and chip NN sensitivity terms reveals three groups of data depending on the number of labels per couple (Figure 10). The chip sensitivity terms of double-labeled and nonlabeled couples correlate well with the respective solution terms with correlation coefficients of $r$ = 0.99 and 0.98, respectively. The sensitivities of double-labeled couples are systematically shifted toward smaller values. This trend reflects the effect of the label in WC pairs, which obviously hampers duplex formation between probe and target (see previous text). On the other hand, the correlation coefficient drops to 0.48 for the sensitivities of single-labeled couples. A single label gives rise to a more heterogeneous situation, because its effect on the respective NN sensitivity term strongly depends on the adjacent nonlabeled base pair.

**Single-Base Sensitivity Terms.** For a rough estimation of the effect of the individual bases on the probe sensitivity, it is appropriate to define the mean SB specific propensity for RNA/DNA duplex formation and

$$\langle Y^{PM}(X)\rangle = (^1/_8)\sum_{Y=A,T,G,C}[\langle Y^{PM}(XY)\rangle + \langle Y^{PM}(YX)\rangle]$$

$$\langle Y^{MM}(X)\rangle = (^1/_4)\sum_{Y=A,T,G,C}\langle Y^{MM}(X\underline{Y})\rangle$$

$$\langle Y^{MM}(\underline{Y})\rangle = (^1/_4)\sum_{X=A,T,G,C}\langle Y^{MM}(X\underline{Y})\rangle \qquad (14)$$

In an analogy with eq 11, we distinguish between the mean effect of a base in the PM sequence referring to WC pairs and the effect in the MM sequence for the mismatched central base and its nearest neighbor. The latter also forms WC pairings, whereas the mismatched base refers to WC and SC pairings as well (see eq 4).

Table 2 lists the chip sensitivities together with the respective solution data. The mean sensitivities are normalized with respect to $\langle Y^P(C)\rangle$ for comparison (see second column and footnote *a* in Table 2). The sensitivities of WC pairs in PM probes, $\langle Y^{PM}(X)\rangle$, group with ascending value according to C > G ≈ T > A. A similar series turns out if the respective base in a WC pair adjoins the central base in the MM probes. The sensitivity values of G and A are, however, increased on a relative scale. For the mismatched central bases $\underline{\mathrm{T}}$ and $\underline{\mathrm{G}}$, the SB sensitivities reverse order (i.e., the series of pair strengths changes into $\underline{\mathrm{C}} > \underline{\mathrm{T}} > \underline{\mathrm{G}} \approx \underline{\mathrm{A}}$. This difference possibly reflects the effect of SC pairing in the SB term of the mismatched bases, which represents the weighted average of WC and SC pairings (see eq 4).

Also, the solution data reveal an asymmetry of the strength of WC pairs. Table 1 shows that the mean sensitivity of nonlabeled G−c pairs is significantly smaller than that of C−g pairs (see also Figure 8). The relative change is $(r_C^P - r_G^P) \approx 0.5$ for nonlabeled and 0.85 for labeled pairs in PM probes. Consequently, about 40% of the relative sensitivity asymmetry between G−c* and C−g pairs can be attributed to the presence of the label. The relative effect of labeling further increases if one compares the relative sensitivity difference of T−a and A−u* pairs (compare $r_T^P - r_A^P$ for solution and chip data).

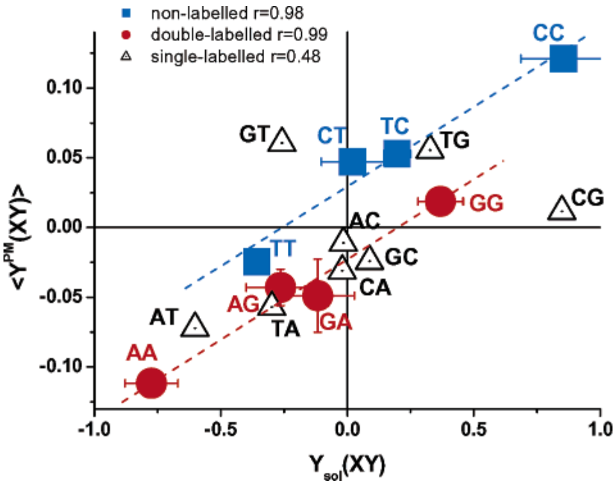

**Figure 10.** Correlation plot between the NN pair sensitivities derived from solution and microarray experiments shown in Figure 8. Note that subgroups of data referring to nonlabeled and double-labeled couples are well correlated but systematically shifted to each other, whereas single-labeled couples show a relatively heterogeneous distribution depending on the nearest neighbor. NN couples, referring to WC pairs without labels and with one and two labels are differently assigned (see figure). The respective correlation coefficient, *r*, is given within the figure. The solution sensitivities are arithmetic means of two data sets taken from ref 10 and 11 (see also Figure 8). The horizontal error bars refer to the differences between the respective values of both data sets. Error bars are omitted for single-labeled couples. The diagonal lines are drawn as a guide for the eye.

The normalized sensitivity differences listed in the last two columns of Table 2 estimate the increment of the binding affinity between pairs which are stabilized by three (C−g, G−c*) and two (T−a, A−u*) hydrogen bonds. This parameter can be interpreted as a rough measure of the effective strength of one H-bond in relative sensitivity units. The solution and chip data roughly agree for the purines, G and A, but considerably differ for the pyrimidines, C and T.

## Summary and Conclusions

The relative contribution of matched and mismatched base pairings to the stability of DNA/RNA probe−target duplexes was estimated using mean sensitivity values averaged over all GeneChip microarray probes with a common middle triple. The sensitivities of PM and MM probes in a similar fashion depend on the middle triple. This agreement can be understood if the central base of both PM and MM probes mainly forms Watson−Crick pairs with bound RNA transcripts. Such behavior is expected if nonspecific hybridization dominates the respective

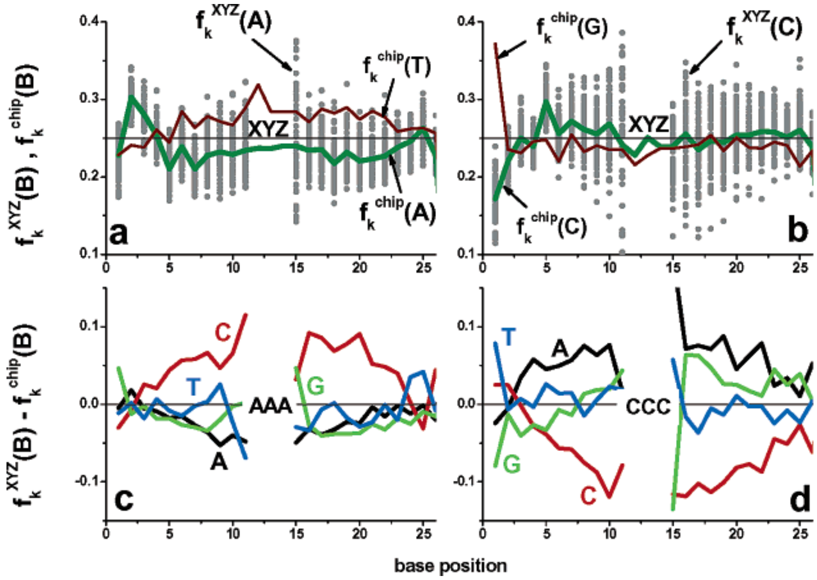

**Figure 11.** Probability profile, $f_k^{chip}(B)$, of the occurrence of base B = T, A (panel a) and G, C (panel b) in any probe of the HG U133 chip along the sequence (see lines). The points refer to the probability of occurrence, $f_k^{XYZ}(B)$, of base B = A (part a) and C (part b) in the subsets of probes with middle triple XYZ. Note that the data scatter about the respective chip average, $f_k^{chip}(B)$ with B = A and C, respectively. Panels c and d show the deviations, $f_k^{XYZ}(B) - f_k^{chip}(B)$, for the middle triples AAA and CCC. They represent the worst cases, i.e., the distributions with the maximum deviation from the mean.

chip averages of the sensitivity. Sensitivity differences between the PMs and MMs of one pair indicate a purine–pyrimidine asymmetry of interaction strengths in WC pairings according to which a C (or T) more strongly contributes to the binding affinity than a G (or A). This asymmetry can be partly attributed to the labeling in the complementary RNA sequence. Biotinyl residues with attached fluorescent labels obviously tend to reduce the sensitivity. The adjacent bases in each sequence considerably modify the sensitivity values, probably because of stacking interactions and steric effects. The triple-averaged probe sensitivities provide NN sensitivity terms, which rank in a similar order to the respective NN free-energy terms obtained from thermodynamic studies on the stability of RNA/DNA duplexes in solution. Systematic deviations between both data sets can be attributed mostly to the labeling of the target RNA in the chip experiments. The triple averages provide detailed information about sequence-specific effects in the middle of the probe sequence. In a forthcoming publication, we will address the effect of the complete base sequence on the probe sensitivities.

**Acknowledgment.** We thank the reviewers for the important ideas and suggestions, which significantly improved some of the presented results. The work was supported by the Deutsche Forschungsgemeinschaft under grant BIZ 6/1-1.

## Appendix

**Correction of the Chip-Averaged Triple Means for Nonrandom Base Distributions.** The probability profiles of the occurrence of base B = A, T, G, C at position $k$ in probes with the middle triple XYZ, $f_k^{XYZ}(B)$, significantly deviate from $f_k^{chip}(B)$, the probability of occurrence of base B at position $k$ in any probe of the chip (see Figure 11; compare the points in parts a and b with the lines for B = A and C, respectively). Note that $f_k^{chip}(B)$ only slightly deviates from the random value for equally distributed bases $f^{random} = 0.25$ for most of the positions $k$ (see lines in panels a and b of Figure 11).

The difference, $f_k^{XYZ}(B) - f_k^{chip}(B)$, provides a measure of the nonrandomness of probe composition relative to the chip average. The respective difference profiles for XYZ = AAA and CCC illustrate the worst cases (i.e., the triples providing the largest bias with respect to the mean; see Figure 11, panels c and d). The chip averages of the sensitivity, $\langle Y^P(XYZ)\rangle \equiv \langle Y^P(XYZ)\rangle_{chip}$, are weighted means referring to the probability profile, $f_k^{XYZ}(B)$, along the sequence. Consequently, the positive and negative deviations of $f_k^{XYZ}(C)$ effectively reduce the absolute values of the triple averages $\langle Y^{PM}(AAA)\rangle$ and $\langle Y^{PM}(CCC)\rangle$ compared with averages over a random base distribution, because cytosines most strongly contribute to the probe sensitivity compared with the other bases (see previous text and refs 3 and 7).

One can estimate the effect of nonrandomness of the base distribution on the sensitivity by

$$\Delta Y^{P}_{corr}(XYZ) \approx \sum_{\substack{k=1\\ k\neq 12..14}}^{25} \sum_{B=A,T,G,C} w_k \cdot \langle Y^{P}(B)\rangle \cdot (f_k^{XYZ}(B) - f(B)) \tag{A1}$$

where $\langle Y^P(B)\rangle$ is the single-base-related sensitivity term (see eq 14) and $f(B)$ is given by $f_k^{chip}(B)$ or $f^{random}$ if one chooses the mean base distribution of the chip or complete randomness as the reference state, respectively. The weighting function considers the positional dependence of single-base sensitivity terms. We used a parabola-like function, $w_k = 1 - [(k - 13)/12]^2$, which was derived from single-base model analyses and accounts for the decrease of single-base-related sensitivity terms toward the ends of the probe sequence.[4,3,7,13]

Equation A1 provides a first-order correction of the triple sensitivities for deviations from the mean distribution of bases along the probe sequences, $\langle Y^P(XYZ)\rangle_{random} \approx \langle Y^P(XYZ)\rangle_{chip} - \Delta Y^P_{corr}(XYZ)$. It turns out that the absolute sensitivity value of the triples CCC and AAA increases after correction by about 20%. Similar corrections are obtained for the other triples.

Linear regression of the corrected data, $\langle Y^P(XYZ)\rangle_{random}$, versus the uncorrected data, $\langle Y^P(XYZ)\rangle_{chip}$, reveals that the correction with respect to randomness ($f(B) = f^{random}$) can be simply considered by a multiplicative factor, $F_{triple} = 1.19 \pm$

0.02 ($r > 0.99$), which scales the raw data according to $\langle Y^{P}(XYZ)\rangle_{random} \approx F_{triple}\cdot\langle Y^{P}(XYZ)\rangle_{chip}$. The corrected triple averages were used to calculate corrected NN and SB terms by means of eqs 10 and 14, respectively. An analogous correlation analysis provides correction factors of $F_{NN} = 1.10 \pm 0.01$ and $F_{SB} = 1.13 \pm 0.04$ ($r > 0.99$) between the corrected and uncorrected NN and SB terms, respectively. Note that the NN and SB terms are calculated as linear combinations of the triple averages, which gives rise to the partial compensation of the correction terms and thus to the smaller correction factors for the derived data. The correction with respect to the base distribution of the chip ($f(B) = f_k^{chip}(B)$) provides a slight change of the correction factors by less than 5% of its value.

In summary, the systematic bias of the triple averages due to the nonrandom base distributions among the probes of one middle triple is nonneglible but relatively small. It can be considered by a correction factor $F_{triple} \approx 1.2$ for the triple averages and $F_{NN} \approx F_{SB} \approx 1.1$ for the derived nearest-neighbor and single-base terms.

## References and Notes

(1) Matveeva, O. V.; Shabalina, S. A.; Nemtsov, V. A.; Tsodikov, A. D.; Gesteland, R. F.; Atkins, J. F. *Nucleic Acids Res.* **2003**, *31*, 4211.

(2) Rouillard, J.-M.; Zuker, M.; Gulari, E. *Nucleic Acids Res.* **2003**, *31*, 3057.

(3) Naef, F.; Magnasco, M. O. *Phys. Rev. E* **2003**, *68*, 11906.

(4) Binder, H.; Kirsten, T.; Loeffler, M.; Stadler, P. *J. Phys. Chem. B* **2004**, *108*, 18003.

(5) Lipshutz, R. J.; Fodor, S. P. A.; Gingeras, T. R.; Lockhart, D. J. *Nat. Genet.* **1999**, *21*, 20.

(6) *Affymetrix Microarray Suite*, version 5.0; Affymetrix, Inc.: Santa Clara, CA, 2001.

(7) Mei, R.; Hubbell, E.; Bekiranov, S.; Mittmann, M.; Christians, F. C.; Shen, M.-M.; Lu, G.; Fang, J.; Liu, W.-M.; Ryder, T.; Kaplan, P.; Kulp, D.; Webster, T. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11 237.

(8) Gralla, J.; Crothers, D. M. *J. Mol. Biol.* **1973**, *73*, 497.

(9) Borer, P. N.; Dengler, B.; Tinoco, I., Jr.; Uhlenbeck, O. C. *J. Mol. Biol.* **1974**, *86*, 843.

(10) Sugimoto, N.; Nakano, S.; Katoh, M.; Matsumura, A.; Nakamuta, H.; Ohmichi, T.; Yoneyama, M.; Sasaki, M. *Biochemistry* **1995**, *34*, 11 211.

(11) Wu, P.; Nakano, S.; Sugimoto, N. *Eur. J. Biochem*. **2002**, *269*, 2821.

(12) Held, G. A.; Grinstein, G.; Tu, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7575.

(13) Zhang, L.; Miles, M. F.; Aldape, K. D. *Nat. Biotechnol.* **2003**, *21*, 818.

(14) Binder, H.; Kirsten, T.; Loeffler, M.; Stadler, P. Sequence specific sensitivity of oligonucleotide probes. In *Proceedings of the German Bioinformatics Conference*, Munich, Germany, Oct. 12-14, 2003.

(15) Li, C.; Wong, W. H. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 31.

(16) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge University Press: New York, 1989.

(17) Xia, T.; Santa Lucia, J. J.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. *Biochemistry* **1998**, *37*, 14 719.