Sensitivity of Microarray Oligonucleotide Probes: Variability and Effect of Base Composition

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

Interdisciplinary Centre for Bioinformatics, Institute for Medical Informatics, Statistics and Epidemiology, and Bioinformatics group, Department of Computer Science, University of Leipzig, Kreuzstrasse 7 b, D-04103 Leipzig, Germany

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

The optimization of both probe design and analysis algorithms for microarray experiments requires improved understanding and predictability of oligonucleotide hybridization behavior. Our physicochemical theory of GeneChip probe sensitivities divides the probe intensity into an averaged intensity value which serves as a relative measure of the RNA target concentration and the sensitivity of each probe. The sensitivity decomposes into additive terms because of specific and nonspecific hybridization, saturation, the heterogeneous distribution of labels, and intramolecular folding of target and probe. The observed heterogeneity of probe sensitivities is mainly caused by variations of the probe affinity for target binding owing to sequence differences between the probes. The sensitivity values are therefore analyzed in terms of simple molecular characteristics, which consider the base composition and sequence of the probes. We found that the mean sensitivity, averaged over all probes of a chip containing a certain number of bases of one type, strongly increases with an increasing number of C nucleotides per oligomer, whereas A nucleotides show the opposite tendency. These trends are asymmetrical with respect to the number of G and T nucleotides, which have a much weaker, and perhaps a somewhat opposite, effect in probes of intermediate and high sensitivity. The middle base systematically affects the relationship between the sensitivities of perfect match (PM) and mismatch (MM) probes. MM probes are, on the average, more sensitive than the respective PM probes if the middle base is a purine in the PM probe of the respective probe pair. For pyrimidines, this relationship reverses. This purine−pyrimidine asymmetry is partly related to the effect of labeling.

Introduction

Microarray chip technology is revolutionizing biology by empowering researchers in the collection of large-scale information on gene expression. It is based on the sequence-specific binding of RNA fragments to oligonucleotide probes that are attached to the chip surface in a well-defined geometrical arrangement and its measurement using fluorescence labels.[1] The integral fluorescence intensity of the probe arrays is related to the amount of bound fluorescently labeled RNA, which, in turn, serves as a measure of the RNA concentration in the sample solution and, thus, of the expression degree of a given gene. Physicochemical factors are of central importance for the understanding of microarray hybridization behavior.

The optimization of probe design and appropriate analysis algorithms require an improved understanding of the hybridization behavior of oligonucleotides. One key issue in microarray technology is how to select oligonucleotide probes with high sensitivity (signal intensity per RNA) and specificity (ratio of specific to nonspecific hybridization). A second, closely related key question addresses the analysis of microarray intensity data in terms of reliable measures of the expression degree of the genes of interest.

Deterministic models based on a molecular description of hybridization show great potential in terms of usefulness for the prediction of the probe sensitivities and, thus, for improvements over existing expression measures based on statistical models. However, only a few studies have addressed sequence-specific effects on the measured intensities of microarray probes. For instance, nonlinearities in the probe responses and sequence effects in the behavior of mismatched probes were discussed in refs 2−5. Matveeva et al. analyzed correlations between the predicted energetics of probe−target duplexes and target self-structures on one hand and microarray data on the other hand.[6] Naef and Magnasco[3] and Mei et al.[7] proposed models which describe the affinity of a probe as the sum of position-dependent base-specific contributions. Zhang et al.[8] applied a position-dependent nearest-neighbor model for RNA/DNA duplexes formed on microarrays. Free energies for RNA/DNA duplex formation are explicitly considered in a recent model of microarray hybridization.[9] Despite this recent progress, it seems that the system producing the measured intensities is presently too complex to be fully described with relatively simple physical models. One idea to overcome the problem suggests a combination of deterministic and stochastic aspects.[10]

The presented paper is aimed at establishing a physicochemical theory of microarray probe sensitivity, to evaluate its predictions using experimental chip data and, finally, to answer the question of how the base composition affects the sensitivity of oligonucleotide probes as one prerequisite for further development of adequate deterministic models. We make use of

* Corresponding author. Interdisciplinary Centre for Bioinformatics of Leipzig University, Kreuzstr. 7b, D-4103 Leipzig, Germany. E-mail: binder@izbi.uni-leipzig.de. Fax: ++49-341-1495-119.

† Interdisciplinary Centre for Bioinformatics.

‡ Institute for Medical Informatics, Statistics and Epidemiology.

§ Bioinformatics group, Department of Computer Science.

10.1021/jp049593g CCC: $27.50
Published on Web 10/27/2004

the fact that each GeneChip microarray provides hybridization data for about 250 000 oligomer sequences at once, which can be very useful for extracting sequence-related factors that influence the hybridization efficiency.

The paper is organized as follows: In the theoretical section, we develop the concept of GeneChip probe sensitivities in terms of normalized intensities. In the second part, we analyzed microarray intensity data taken from Affymetrix GenChips within the light of the theoretical predictions. In the last part, the sensitivity values are analyzed in terms of simple sequence characteristics such as the base composition and the nucleotides in the middle of the oligomer sequence. The accompanying paper addresses the issue of molecular interactions in terms of base pairing, nearest-neighbor stacking contributions, and the effect of labeling on duplex stability.[11]

Theory

Normalized Intensities of Microarray Probes. The GeneChip technology of Affymetrix uses short 25-mer oligomers of which the sequence refers to the consensus sequence of the respective target genes.[12] Between 11 and 20 different reporter sequences for each gene form a so-called probe set. For each target sequence, a pair of probes is present on the chip to quantify the extent of nonspecific binding. One type of probes, the so-called PM probes, perfectly matches the target sequence in terms of Watson−Crick (WC) base pairs. The second type, the so-called mismatch (MM) probes, is created by replacing the (13^{th}) middle base of the respective PM probe with the respective complementary base.

The oligonucleotides are attached to the quartz surface of the chip in spot-like probe locations where probes of different sequences refer to different spots. The photolithographic technology of chip production presently allows one array to hold about 10^5-10^6 different probe spots per square centimeter (i.e., at an extremely high packing density). A typical Affymetrix GeneChip, such as the human genome chip, HG U133, contains more than 22 000 probe sets with nearly 250 000 different perfect match (PM) and mismatch (MM) probe sequences. RNA fragments with fluorescently labeled uracil (u*) and cytosine (c*) bases bind to the oligonucleotide probes during hybridization. The fluorescence intensity of the probe-bound RNA is measured by means of an imaging system (scanner, detector, and imaging software). The intensity of each array defines the respective probe intensity.

Let us define the sensitivity as the deviation of the intensity for each probe from the mean over the probe set in a logarithmic ($\log A \equiv \log_{10} A$) scale

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

where I^{PM} and I^{MM} are the intensities of perfect match (PM) and mismatch (MM) reporter probes, respectively, which were corrected for the optical background level. The broken brackets $\langle \cdots \rangle_{\mathrm{set}}$ denote arithmetic averaging over the respective probe set. Consequently, the mean value of Y^{P} averaged over the probe set (and, of course, also over the entire chip) vanish (i.e., $\langle Y^{\mathrm{P}} \rangle_{\mathrm{set}} = 0$ and $\langle Y^{\mathrm{P}} \rangle_{\mathrm{chip}} = 0$).

The intensity of a probe depends on the amount of bound RNA and on its fluorescence yield according to

$$I^{\mathrm{P}} = D_{\mathrm{chip}} \cdot (N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{P}}\xi^{\mathrm{T}}) \cdot F(\xi^{\mathrm{T}}) + \sum_{\xi \neq \xi^{T}} N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{P}}\xi) \cdot F(\xi)) \qquad (2)$$

D_{chip} denotes a proportionality factor depending on signal processing and sample preparation. It is assumed to be a constant for each chip. $N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{P}}\xi^{\mathrm{T}})$ is the amount of target RNA (in moles) which binds to the probe. The sequences of target and PM probe, ξ^{T} and ξ^{PM}, are complementary (i.e., their sequences match via Watson−Crick (WC) base pairs). Consequently, $N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{PM}}\xi^{\mathrm{T}})$ quantifies the amount of specific binding. For MM probes, $N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{MM}}\xi^{\mathrm{T}})$ gives the amount of target RNA which specifically binds to the respective mismatch probe. The sum in eq 2 considers non specific hybridization events. It runs over all RNA sequences different from the target that bind to the probe. The fluorescence term, $F(\xi)$, defines the fluorescence yield of one mole of RNA of sequence ξ.

Binding, Saturation, and Folding. The relationship between free and bound RNA of sequence ξ is characterized by the binding constant

$$K^{\mathrm{b}}(\xi^{\mathrm{P}}\xi) = c^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{P}}\xi) \cdot [c^{\mathrm{f,unfold}}_{\mathrm{RNA}}(\xi) \cdot c^{\mathrm{f,unfold}}_{\mathrm{P}}(\xi^{\mathrm{P}})]^{-1} \qquad (3)$$

in accordance with the mass action law. The different c values denote the respective concentrations in appropriate units. The superscripts f and b discriminate between free and bound species (e.g., free RNA in the sample solution and free probes on the chip). The second superscript "unfold" considers the fact that only unfolded probes and targets are able to hybridize. In other words, intramolecularly folded species must first unfold into an extended conformation before duplex formation according to the scheme

folded probe ⇔ unfolded probe ⇔ duplex ⇔
unfolded RNA transcript ⇔ folded RNA transcript

The relationship between the amount of folded and unfolded species is characterized by the equlibrium constants $K^{\mathrm{fold}}_{\mathrm{i}} = c^{\mathrm{free,fold}}_{\mathrm{i}}/c^{\mathrm{free,unfold}}_{\mathrm{i}}$ with i = P, RNA. Making use of the conditions of material balance, $N^{\mathrm{tot}}_{\mathrm{i}} = N^{\mathrm{b}}_{\mathrm{i}} + N^{\mathrm{f}}_{\mathrm{i}}$ and $N^{\mathrm{f}}_{\mathrm{i}} = N^{\mathrm{f,fold}}_{\mathrm{i}} + N^{\mathrm{f,unfold}}_{\mathrm{i}}$ (i = P, RNA), the concentrations in the denominator of eq 3 rewrite into

$$c^{\mathrm{f,unfold}}_{\mathrm{RNA}}(\xi) \propto [N^{\mathrm{tot}}_{\mathrm{RNA}}(\xi) - N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{P}}\xi)] \cdot [1 + K^{\mathrm{fold}}_{\mathrm{RNA}}(\xi)]^{-1}$$

$$c^{\mathrm{f,unfold}}_{\mathrm{P}}(\xi^{\mathrm{P}}) \propto [N^{\mathrm{tot}}_{\mathrm{P}}(\xi^{P}) - \sum_{\xi} N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{P}}\xi)] \cdot [1 + K^{\mathrm{fold}}_{\mathrm{P}}(\xi^{\mathrm{P}})]^{-1} \qquad (4)$$

where the Ns are the moles of the respective species. The sum in the second equation considers the fact that specific and nonspecific binding compete for the oligomers of the probe. Let us discuss two limiting cases, namely (i) a large excess of free RNA transcripts ($\sum N^{\mathrm{b}}_{\mathrm{RNA}} \ll N^{\mathrm{tot}}_{\mathrm{RNA}}$) and (ii) a large excess of probes ($\sum N^{\mathrm{b}}_{\mathrm{RNA}} \ll N^{\mathrm{tot}}_{\mathrm{P}}$). Rearrangement of eq 3 after the insertion of eq 4 provides for the special case i

$$N^{\mathrm{b}}_{\mathrm{RNA}}(\xi^{\mathrm{P}}\xi) \approx N_{\mathrm{P}}(\xi^{\mathrm{P}}) \cdot c_{\mathrm{RNA}}(\xi) \cdot K^{\mathrm{b+f}}(\xi^{\mathrm{P}}\xi) \cdot [1 + \sum_{\xi'} c_{\mathrm{RNA}}(\xi') \cdot K^{\mathrm{b+f}}(\xi^{\mathrm{P}}\xi')]^{-1}$$

with

$$K^{\mathrm{b+f}}(\xi^{\mathrm{P}}\xi) = K^{\mathrm{b}}(\xi^{\mathrm{P}}\xi) \cdot \{[1 + K^{\mathrm{fold}}_{\mathrm{p}}(\xi^{\mathrm{P}})][1 + K^{\mathrm{fold}}_{\mathrm{RNA}}(\xi)]\}^{-1} \qquad (5)$$

The effective binding constant, $K^{\mathrm{b+f}}_{\mathrm{p}}$, considers duplex formation and the folding−unfolding equilibrium as well (see also ref 13). It shows that the binding affinity of the RNA fragment for a given probe decreases if the RNA and/or the probe tend to fold intramolecularly. Limiting case ii gives rise to an equation analogous to eq 5 in which N_{P} is however replaced by

c_{RNA} and vice versa. In this case, one obtains $N^{b}_{RNA} \propto c_{RNA}$ (i.e., no saturation). Experiments in which microarray chips are treated with increasing amounts of RNA show, however, a clearly saturation-like behavior.[2,4,14] This result is incompatible with the limiting case ii but in agreement with i, which predicts that N^{b}_{RNA} asymptotically levels off to $N^{b}_{RNA} \rightarrow N_P$ with increasing RNA concentrations at $c_{RNA} \gg 1/K^{b+f}_{p}$.

Fluorescence Labeling. The fluorescence intensity of an RNA fragment is proportional to the number of labeled nucleotides per sequence, N_{u^*} and N_{c^*} (the indices c* and u* refer to the biotinylated bases cytosine and uracil), the fluorescence yield per label, ϕ, and the efficiency of labeling, p_{lab}, in a first-order approximation

$$F(\xi) \approx p_{lab} \cdot [\phi_{c^*} \cdot N_{c^*}(\xi) + \phi_{u^*} \cdot N_{u^*}(\xi)] \approx p_{lab} \cdot \phi \cdot N^{F}(\xi) \quad (6)$$

Nucleotides can quench the emission of fluorescent probes with an efficiency depending on the base type,[15] and thus on the base to which a label is attached. The effect is, however, relatively small, and moreover, for labeled bases, it also depends on their neighbors. The approximation on the right-hand side of eq 6 assumes, therefore, $\phi_{c^*} \approx \phi_{u^*}$ (with $N^{F} = N_{c^*} + N_{u^*}$).

A more detailed view on the fluorescence intensity of target−probe duplexes assumes a binominal probability distribution of n_F labels among a sequence with N^F biotinylated and, thus, potentially labeled bases

$$B(n_F, N^F, p_{lab}) = \binom{N^F}{n_F} [p_{lab}]^{n_F} (1 - p_{lab})^{N_F - n_F}$$

The presence of labels might affect the binding equilibrium (eq 3). The substitution $F(\xi) \cdot K^{b+f}(\xi^P, \xi) \approx \phi \cdot \sum [B(n_F, N_F, p_{lab}) \cdot n_F \cdot K^{b+f}(\xi^P, \xi, n_F)]$ considers this fact (see eqs 5 and 6). The sum runs over the number of labeled bases from $n_F = 0$ to $N^F(\xi)$. The modified binding constant, $K^{b+f}(\xi^P, \xi, n_F) \approx K^{b+f}(\xi^P, \xi) \cdot (K_F)^{n^F}$, accounts for the alteration of the binding affinity between the probe and the target by a constant factor K_F per label. With this approximation, eq 6 rewrites into

$$F(\xi) \approx \phi [\sum_{n_F=0}^{N^F} B(n_F, N^F, p_{lab}) \cdot n_F \cdot K_F^{n_F} + p_{lab} \cdot N^{F}_{ex}] \quad (7)$$

The second term in eq 7 in addition takes into account that the length of the RNA fragment usually exceeds the length of the target sequence. The number of u* and c* outside the respective 25-mer is denoted by N^{F}_{ex}. In the limiting case of a high affinity penalty per label, $K_F \ll 1$, only targets with a single label, or even without a label in the target sequence, contribute to the brightness, $F(\xi) \approx \phi \cdot p_{lab} \cdot K_F$ for $N^{F}_{ex} = 0$ and $F(\xi) \approx \phi \cdot p_{lab} \cdot N^{F}_{ex}$ for $N^{F}_{ex} > 0$, respectively. Equation 6 refers to the limiting case of a small affinity penalty, $K_F \approx 1$ (i.e., $F(\xi) \approx \phi \cdot p_{lab} \cdot (N^F + N^{F}_{ex})$).

The Sensitivity of Oligomer Probes. The probe intensity becomes, after insertion of eqs 5 and 6 into eq 2

$$I^P \approx F_{chip} \cdot N^F(\xi^T) \cdot K^{b+f}(\xi^P \xi^T) \cdot c^{tot}_{RNA} \cdot S^P[K^{b+f}(\xi^P \xi^T) \cdot c^{tot}_{RNA}] \cdot [x^S(\xi^T) + \Delta^P(r^F)] \quad (8)$$

with the chip-specific constant

$$F_{chip} = N_p(\xi^P) \cdot D_{chip} \cdot p_{lab} \cdot \phi,$$

the saturation term

$$S^P[K^{b+f}(\xi^P \xi^T) \cdot c^{tot}_{RNA}] = \{1 + K^{b+f}(\xi^P \xi^T) \cdot c^{tot}_{RNA} \cdot [x^S(\xi^T) + \Delta^P(1)]\}^{-1},$$

the total RNA concentration $c^{tot}_{RNA} = \sum_\xi c_{RNA}(\xi)$
the fraction of target RNA $x^S(\xi^T) = c_{RNA}(\xi^T)/c^{tot}_{RNA}$
the fraction of mismatched RNA $x^{NS}(\xi)|_{\xi \neq \xi^T} = c_{RNA}(\xi)/c^{tot}_{RNA}$
the relative contribution of nonspecific hybridization ($a = 1, r^F$)

$$\Delta^P(a) = \sum_{\xi \neq \xi^T} x^{NS}(\xi) \cdot r^P(\xi \xi^P \xi^T) \cdot a$$

and the ratios
$r^P(\xi \xi^P \xi^T) = K^{b+f}(\xi^P \xi)/K^{b+f}(\xi^P \xi^T)$ and
$r^F(\xi \xi^T) = N^F(\xi)/N^F(\xi^T)$.

After the substitution of eq 8 into eq 1, one obtains the sensitivity of each probe

$$Y^P \approx Y^P_S + Y^P_F + Y^P_{NS} - Y^P_{sat} - Y^P_{fold} - Y^T_{fold} \quad (9)$$

as a sum of terms describing

$$Y^P_S = \Delta \log[K^b(\xi^P \xi^T)] \quad \text{(specific hybridization)}$$

$$Y^P_F = \Delta \log[N^F(\xi^T)] \quad \text{(fluorescence)}$$

$$Y^P_{NS} = \Delta \log[x^S + \Delta^P(r^F)] \quad \text{(nonspecific hybridization)}$$

$$Y^P_{sat} = \Delta \log\{1 + K^{b+f}(\xi^P \xi^T) \cdot c^{tot}_{RNA} \cdot [x^S + \Delta^P(1)]\} \quad \text{(saturation)}$$

$$Y^P_{fold} = \Delta \log[1 + K^{fold}(\xi^P)] \quad \text{(folding of the probe)}$$

$$Y^T_{fold} = \Delta \log[1 + K^{fold}(\xi^T)] \quad \text{(folding of the target)}$$

and the definition
$\Delta \log[A] \equiv \log(A) - \langle \log(A) \rangle_{set}$.

The sensitivity, Y^P, provides a measure of the (logarithmic) intensity of a given probe compared to the mean intensity of all probes of the respective probe set. It specifies its ability to detect a certain amount of RNA in the sample solution used for hybridization. The sensitivity decomposes into additive terms caused by different effects (see eq 9). The first term in eq 9, Y^P_S, describes the sensitivity due to specific hybridization at ideal conditions if the binding of the target RNA to the probe is not perturbed by nonspecific binding, intramolecular folding, or saturation effects. Note that it is independent of the RNA concentration, c^{tot}_{RNA}, and of the chip-specific factor, F_{chip}, because the transformation according to eq 1 cancels out all factors that are common for the respective probe set and chip.

Mean Intensity and Transcript Concentration. Usually, the target RNA concentration is a priori unknown. The set average of the logarithmic intensity can be used as an intrinsic, approximative measure of the target RNA concentration c^S_{RNA} according to

$$\langle \log I^P \rangle_{set} \approx \log(c^S_{RNA}) + \langle Z(\xi^P \xi^T) \rangle_{set} + \log F_{chip}$$

with

$$c^S_{RNA} \equiv x^S \cdot c^{tot}_{RNA}$$

and

$$Z(\xi^P\xi^T) \equiv \log\{N^F(\xi^T)\cdot K^{b+f}(\xi^P\xi^T)\cdot S^P[K^{b+f}(\xi^P\xi^T)\cdot c_{RNA}{}^{tot}]\cdot [1 + \Delta^P(r^F)/x^S]\} \quad (10)$$

The chip specific constant, $\log F_{chip}$, can be eliminated by subtracting the chip average, $\delta\langle\log(I^P)\rangle_{set-chip} \equiv \langle\log(I^P)\rangle_{set} - \langle\log(I^P)\rangle_{chip}$. The set-average of the probe-specific effects, $\langle Z(\xi^P\xi^T)\rangle_{set}$, gives rise to a set-specific variation of the intensity estimate, $\langle\log I^P\rangle_{set}$, around the logarithmic concentration value.

Nonspecific Versus Specific Hybridization. Microarray hybridization experiments intend to measure the expression degree given as the concentration of the target $c^S_{RNA} \equiv c_{RNA}(\xi^T)$. A suitable data analysis must, therefore, correct the total fluorescence intensity for the additive contribution due to nonspecific hybridization Δ^P (see eq 8).

By nonspecific binding, we imply the lower-affinity mismatched duplexes involving sequences other than the intended target. The smaller number of WC pairs gives rise to a weaker binding affinity and thus to $r^P(\xi\xi^P\xi^T) < 1$ (see eq 8). Let us approximate the mismatch term in eq 8 by a simple product of effective values according to

$$\Delta^P(r^F) \approx \Delta^P(1) \approx x^{NS}\cdot r^P(\xi^P) \quad (11)$$

where $x^{NS} = (1 - x^S)$ is the total fraction of mismatched RNA. The ratio $r^P(\xi^P) = K^{eff}(\xi^P)/K^{b+f}(\xi^P\xi^T)$ characterizes the mean decrease of the binding affinity of the probe for mismatched RNA compared with that for the target. The effective binding constant, $K^{eff}(\xi^P)$, is mainly determined by the number of remaining WC pairs between the probe and the RNA fragments. Furthermore, it seems safe to assume that the number of labeled bases is, on the average, similar for matched and mismatched fragments, $r^F(\xi\xi^T) \approx 1$ (see eq 8).

In the limit of a high fraction of target RNA ($x^S \rightarrow 1$) and/or low affinity for mismatched RNA ($r^P(\xi^P) \rightarrow 0$), the term Y^{NS}_P vanishes in eq 9. In the absence of target RNA, one obtains $Y^{NS}_P(x^S \rightarrow 0) \approx \Delta\log(r^P(\xi^P))$. The intermediate case ($0 < x^S < 1$) provides $Y^{NS}_P(r_x) \approx \Delta\log[1 - r_x\cdot r^P(\xi^P)]$ where $r_x = (1 - x^S)/x^S$ is a constant for each probe set. Note that the term Y^{NS}_P vanishes independently of the fraction of target RNA for $r^P(\xi^P) \approx$ constant. This assumption appears reliable in a first-order approximation, because the mean binding affinity of probes taken from a cocktail of RNA fragments with a broad distribution of base composition is directly related to its binding strength in terms of WC pairs with the target, which is, in turn, related to its affinity constant, $K^{b+f}(\xi^P\xi^T)$.

The Effect of Saturation. Equation 9 predicts that all terms contributing to the probe sensitivity are independent of the RNA concentration, except the saturation term, Y^P_{sat}. Figure 1 illustrates the effect of saturation with increasing concentrations of specific transcripts using a simple model calculation (see the caption of Figure 1 for details). At small RNA target concentrations, $[x^S + \Delta^{NS}]\cdot c^{tot}_{RNA} \approx [c^S_{RNA} + \text{constant}] \ll 1/K^{b+f}(\xi^P\xi^T)$, the probes are far from saturation, and one gets $Y^P_{sat} \approx 0$. With increasing c^S_{RNA}, the probes progressively saturate with bound RNA fragments. At $[c^S_{RNA} + \text{constant}] \approx K^{b+f}(\xi^P\xi^T)$, about one-half of the oligomers of the respective probe become saturated, accompanied by a drop of the sensitivity to about 50% of its initial value. At high specific RNA content, $[c^S_{RNA}] \gg 1/K^{b+f}(\xi^P\xi^T)$, all sensitivity terms vanish except the fluorescence term (i.e., $Y^P \rightarrow Y^P_F$), meaning that the probe loses its sensitivity for RNA binding. The remaining variation of fluorescence intensity is solely due to differences of the number of labeled bases between the probes (which is not considered in Figure 1).

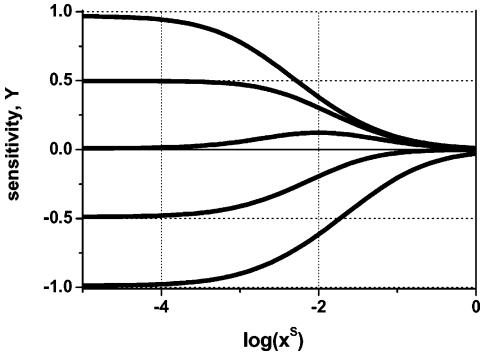

Figure 1. The effect of saturation on the probe sensitivities as a function of the fraction of specific transcripts. Saturation progressively decreases the apparent sensitivity values of the probes with an increasing fraction of specific transcripts. We defined a model set of five probes which are characterized by sensitivity values $Y^i_S = -1.0, -0.5, 0.0, +0.5$, and $+1.0$ ($i = 1...5$) referring to specific hybridization. Then, the curves are calculated by means of $Y^i = \log I^i - \langle\log I\rangle_{set}$ with $\log I^i = Y^i_S - \Delta\log(1 + \Delta_{sat})$, the set average $\langle\log I\rangle_{set} = (1/5)\sum_i \log I^i$ (see also eq 9) and $\log(\Delta_{sat}) \approx Y_S + \log\kappa + \log(x^S + (1 - x^S)\cdot r_b)$ (see eq 8, $\log\kappa = \log[c^S_{RNA}\cdot K^{b+f}(\xi^P\xi^T)] = 2$; $r_b = 10^{-3}$). The contribution of fluorescence emission, Y_F, is not considered.

Note that saturation causes a nonlinear relation between $\delta\langle\log(I^P)\rangle_{set-chip}$ and the relative transcript concentration $\log(c^S_{RNA})$ (see eq 10). Upon saturation of all probes of a set, the concentration dependence cancels out, and eq 10 transforms into a set-specific constant (i.e., $\langle\log(I^P)\rangle_{set} = \langle Z\rangle_{set} + \log(F_{chip}) = \text{constant}$).

Relation to Thermodynamics. The Gibbs free energies of duplex formation and folding are related to the respective equilibrium constants by

$$G^b(\xi^P\xi) = -RT\ln[K^b(\xi^P\xi)\cdot W]$$

$$G^{fold}(\xi) = -RT\ln[K^{fold}(\xi)] \quad (12)$$

where W is a factor in concentration units that accounts for the change of ideal mixing entropy (the so-called cratic contribution to the entropy; see ref 16, pp 283). R and T denote the gas constant and the absolute temperature, respectively. On the other hand, the sensitivity of an oligonucleotide probe, Y^P, is also related to the respective equilibrium constants of binding and folding according to eq 8. It appears appropriate to scale the sensitivity (see eqs 9 and 11) in energy units in analogy to eq 12

$$\Delta G^{app} = -(RT\ln 10)\cdot Y^P \approx \Delta G^{b,S} - [\Delta G^{fold}(\xi^P) + \Delta G^{fold}(\xi^T)] + \Delta\zeta^{NS} + \Delta\zeta^F - \Delta\zeta^{sat}$$

with the definition

$$\Delta A \equiv A - \langle A\rangle_{set} \quad (13)$$

Here, ΔG^{app} defines the difference between the apparent free energy of hybridization of the considered probe and the respective set average, $\langle\ldots\rangle_{set}$. The factor $\ln 10 = \log_e 10 \approx 2.3$ considers the different bases of the logarithmic scales for Y and G (see eqs 1 and 12). The first terms at the right-hand side of eq 13 provide the free-energy difference due to specific binding, $\Delta G^{b,S} = -(RT\cdot\ln 10)\cdot Y^P_S$, and to folding of the probe and target. The last three terms consider the effects of nonspecific hybridization, $\Delta\zeta^{NS} = -(RT\cdot\ln 10)\cdot Y^P_{NS}$ (see eq 11), of labeling, $\Delta\zeta^{fluor} = -(RT\cdot\ln 10)\cdot Y^P_F$, and of saturation, $\Delta\zeta^{sat} = -(RT\cdot\ln 10)\cdot Y^P_{sat}$. The scaling factor between the sensitivity

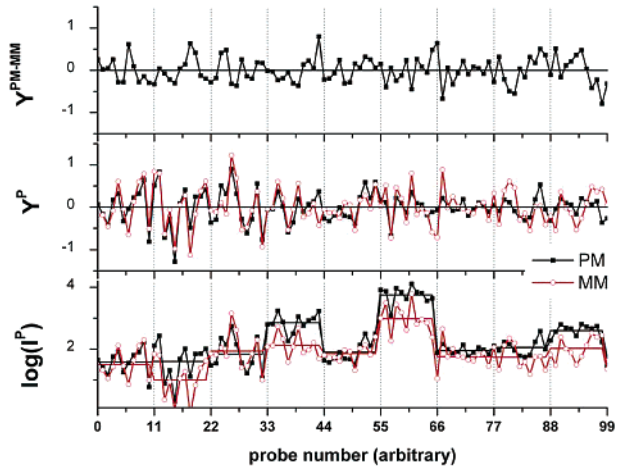

Figure 2. Typical log 10 intensity values, $\log(I^P)$, the respective sensitivities, Y^P (P = PM, MM) and their difference, Y^{PM-MM}, of 99 perfect match (PM) and mismatch (MM) probes of an Affymetrix HG U133 chip. The data between two vertical dotted lines refer to one probe set containing 11 probes. The respective mean logarithmic intensities averaged over each probe set, $\langle \log I^{PM} \rangle_{set}$ and $\langle \log I^{MM} \rangle_{set}$, are shown by horizontal lines together with the logarithmic intensities (part below).

and free energy is $RT \cdot \ln 10 \approx 6$ kJ/mol for typical hybridization temperatures (40 °C).

Chip Data and Processing. Intensity data of the human chips are taken from 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 (see also ref 14 for a description of the previous HG U95 Latin Square experiment). The HG U133-LS experiment is a calibration data set in which transcripts referring to 42 genes ($42 \times 11 = 462$ probes) are spiked onto 14 different arrays at 14 concentrations corresponding to all cyclic permutations of the series (0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512) pM in a complex human background extracted from a HeLa cell line not containing the spikes. Each condition was realized in triplicate.

PM and MM intensities are background-corrected using the algorithm provided by *MAS 5.0*.[12] We also analyzed the base composition characteristics of two additional chip types corresponding to the human genome, HG U95Av2, and the mouse genome, MG U74Av2. The results agree with those obtained from the HG U133 chips (data not shown). All chip analyses are performed using the gene expression data warehouse platform of IZBI (see http://www.izbi.de).

Data Analyses and Discussion

Sensitivities and Set-Averaged Mean Intensities of PM and MM Probes. The lower panel of Figure 2 shows typical baseline-corrected PM and MM signal intensities, $\log I^{PM}$ and $\log I^{MM}$, taken from 9 subsequent probe sets of an HG U133 Affymetrix chip. Each of the chosen sets contains 11 probes. The respective set averages, $\langle \log I^{PM} \rangle_{set}$ and $\langle \log I^{MM} \rangle_{set}$, are shown by horizontal lines. Note that the log 10 probe intensities scatter around the respective set average which, in turn, fluctuates around the chip averages, $\langle \log I^{PM} \rangle_{chip}$ and $\langle \log I^{MM} \rangle_{chip}$, respectively. The probe sensitivity provides the logarithmic intensity relative to the respective set average which fluctuates around zero (see eq 1 and Figure 2, middle panel). Note the relatively high degree of correlation between the sensitivities of PM and MM probes. The sensitivity difference between the PM and MM intensities, $Y^{PM-MM} \equiv Y^{PM} - Y^{MM}$, is shown in the upper panel of Figure 2.

For an overview of all of the intensity data of one chip, we plot the logarithmic intensities of the PM and MM probes as a function of the respective set average, $\langle \log I^P \rangle_{set}$ (see upper row of panels in Figure 3). Each probe intensity decomposes into two contributions according to the sensitivity concept, the respective set-averaged mean intensity, $\langle \log I^{PM} \rangle_{set}$, and the probe sensitivity, Y^P. The second row of panels in Figure 3 shows the sensitivity values of the PM and MM probes of the chip as a function of the relative set average, $\delta\langle \log I^{PM} \rangle_{set-chip}$. The data cloud shows a pear-like shape where the scatter width of the sensitivity data narrows with increasing $\delta\langle \log I^{PM} \rangle_{set-chip}$. The

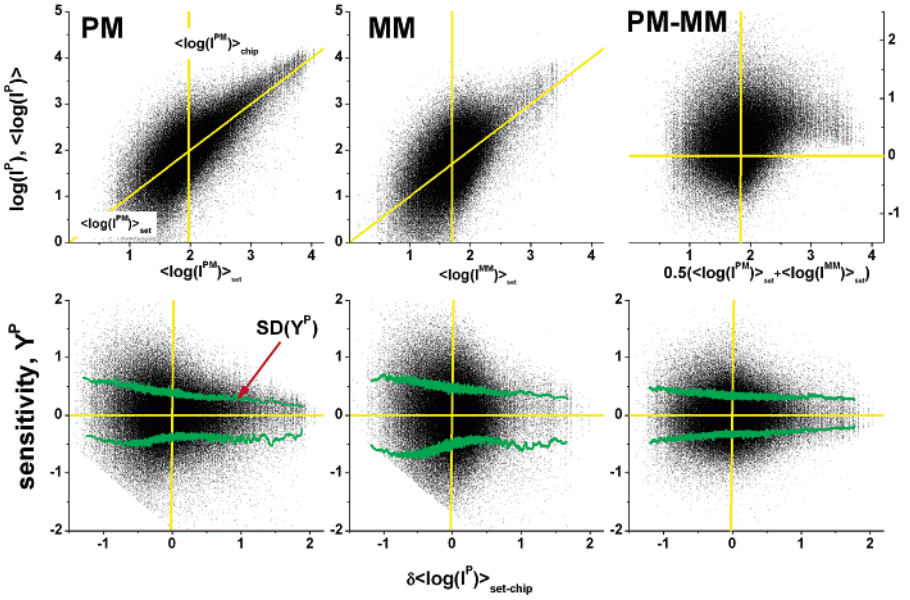

Figure 3. Scatter plot of PM, MM, and PM − MM log-intensity values of all probes of an Affymetrix HG U133 chip as a function of the set average of the respective intensity, $\langle \log I^P \rangle_{set}$ (P = PM, MM, PM + MM), (panel above) and the respective sensitivities as a function of set average of the intensity relative to the chip average, $\delta\langle \log I^P \rangle_{set-chip}$ (panel below). The vertical lines refer to the chip averages of the respective PM and MM intensities, $\langle \log I^{PM} \rangle_{chip} = 1.97$ and $\langle \log I^{MM} \rangle_{chip} = 1.75$, respectively. The standard deviations of the sensitivities, SDs, are separately calculated for $Y^P > 0$ and $Y^P < 0$ as running averages over the subsequent 200 probes along the abscissa (see scatter curves in the part below).

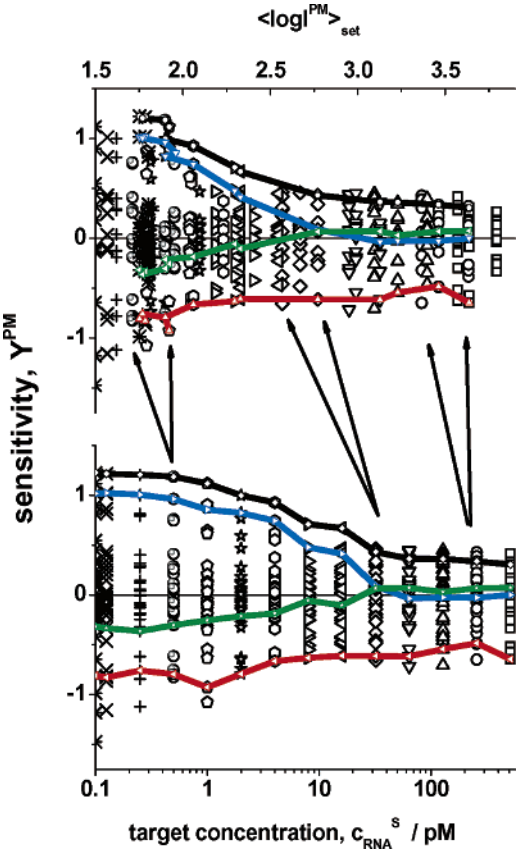

Figure 4. PM sensitivities of three spiked-in transcripts (203508_at, 204513_s, 204563_at) of the HG U133-LS experiment as a function of the target concentration (panel below) and of the set-averaged intensity (panel above). Each concentration splits up into a range of $\langle \log I^{PM} \rangle_{set}$ values as indicated for selected concentrations by the arrows between the two panels. Each symbol type refers to one spiked-in concentration. The sensitivities of selected probes are connected by lines. The standard error of the data, SE $\approx$ SD/$\sqrt{2}$ < 0.03, was estimated from three replicates (see Appendix). Error limits are smaller than the symbols.

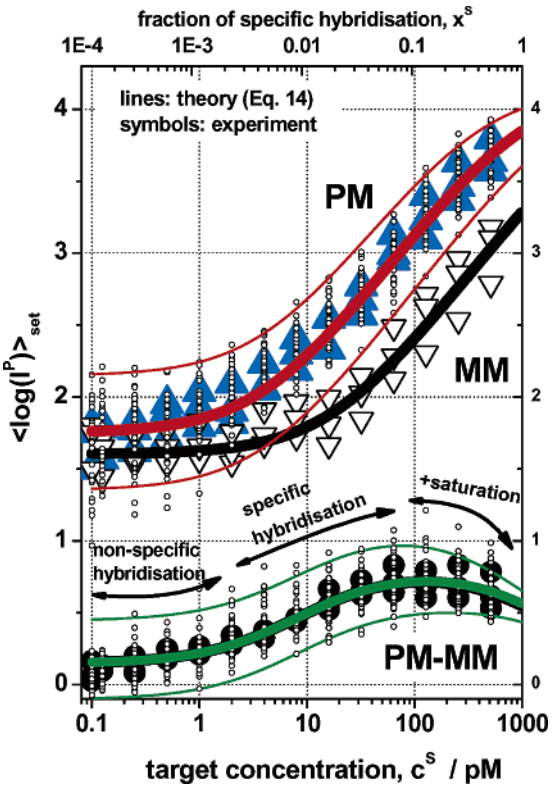

Figure 5. Set-averaged mean intensity of PM and MM probes of three spiked-in transcripts (large symbols: 203508_at, 204513_s, 204563_at) as a function of transcript concentration. The thick lines are calculated using eq 14 with log κ^P/r^P = 0.0/0.004 (P = PM) and −0.8/0.018 (P = MM) and log F_{chip} = 4.15. The abscissa above is the respective fraction of specific RNA transcripts $x^S = c^S_{RNA}/c^{tot}_{RNA}$ where the total RNA concentration was set arbitrarily to c^{tot}_{RNA} = 1000 pM. The part below shows the log-intensity difference between the PM and respective MM data. The small symbols are the set-averaged mean intensity data of the remaining 39 spiked-in transcripts of the HG U133-LS experiment (PM and PM − MM, MM data are omitted for clarity). The thin lines forming the envelope of these data are calculated with log κ^{PM} + 0.4/log κ^{PM} − 0.4 = +0.4/−0.4 and unchanged remaining parameters.

abscissa provides a measure of the apparent expression degree of the genes considered by the respective probe sets.

The Effect of Transcript Concentration. The HG U133-LS experiment enables us to estimate the relationship between the probe sensitivity and the RNA concentration of the spiked-in genes, on one hand, and the respective set-averaged probe intensity on the other hand. In particular, we make use of these data to analyze the effect of nonspecific hybridization and of saturation, which both depend on the concentration of specific transcripts.

Figure 4 shows the sensitivities of the probes of three selected probe sets as a function of c^S_{RNA}, the concentration of specific spiked-in RNA (part below). The traces of selected individual probes (see lines in Figure 4) show similar courses to the model curves shown in Figure 1. This qualitative agreement suggests that saturation significantly affects the probe sensitivities with increasing transcript concentration.

The translation of the abscissa units from concentration into set-averaged intensities, $\langle \log I^P \rangle_{set}$, gives rise to a range of $\langle \log I^P \rangle_{set}$ values for each c^S_{RNA} owing to the variation of the set-specific constant, $\langle Z \rangle_{set}$ (see arrows in Figure 4 and eq 10). The uncertainty of $\langle \log I^P \rangle_{set}$ relative to the logarithmic concentration scale is roughly $\delta|\langle \log I^P \rangle_{set}|_{c=constant} \approx \delta\langle Z \rangle_{set} \approx 0.3$ for the three considered probe sets.

Moreover, an inspection of both panels of Figure 4 reveals that the concentration and intensity scales are related nonlinearly with each other. To obtain more quantitative information, we directly correlate the transcript concentration with the respective set-averaged intensity (see Figure 5). Making use of eqs 8 and 11, one obtains an equation, which correlates both values

$$\langle \log I^P \rangle_{set} \approx \log(F_{chip}) + \log \kappa^{P,S} + \log[x^S + (1 - x^S)\cdot r^P] - \log\{1 + \kappa^{P,S}\cdot[x^S + (1 - x^S)\cdot r^P]\}$$

with

$$\log \kappa^{P,S} = \langle \log[K^b(\xi^P \xi^T)\cdot c_{RNA}{}^{tot}] \rangle_{set} \qquad (14)$$

The fit to selected experimental set averages of PM and MM probe intensities shows that eq 14 describes the effects of saturation and nonspecific hybridization well (compare lines and symbols in Figure 5). Estimates of the parameters log $\kappa^{P,S}$, F_{chip}, and r^P are given in the caption of Figure 5.

The concentration range can be roughly divided into three regions according to the course of the curves. In the limit of low concentration of specific transcripts ($c^S_{RNA} \to 0$ or $x^S \ll (1 - x^S)\cdot r^P$; see eq 14), the probe intensity is dominated by nonspecific hybridization. It levels off to a constant value $\langle \log I^P \rangle \propto \langle \log(\kappa^{P,S}\cdot r^P) \rangle = \langle \log(\kappa^{P,NS}) \rangle \propto \log K^{eff}(\xi^P)$ with decreasing concentration (see eq 14). The difference of the PM and MM logarithmic intensities, $\langle \log I^{PM-MM} \rangle = \log(\kappa^{PM,NS}/\kappa^{MM,NS}) \approx$

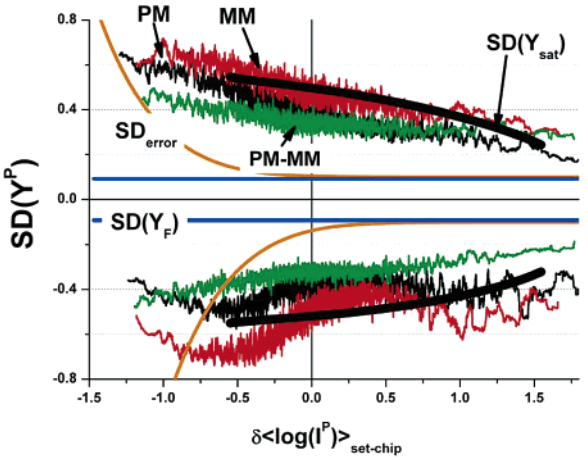

Figure 6. Enlarged view of the standard deviations shown in Figure 3 for the PM, MM, and PM − MM sensitivities. The SD_{error} curves are taken from the error analysis described in the Appendix for $Y = \pm 0.25$ (see Figure 13). $SD(Y_F)$ is the expected standard deviation of the sensitivity due to fluctuations of the fluorescence intensity, which are caused by the heterogeneous distribution of labeled nucleotides in a 25-meric target fragment. The $SD(Y_{sat})$ curves show the effect of saturation on the distribution width of the sensitivity values. They are calculated by means of $SD(Y_{sat})$ with $Y_{sat} \approx (Y - \log(1 + 10^{Y+\langle \log I \rangle}))$ (see also Theory section).

0.1, provides a measure of the logarithmic ratio of the effective binding constants for nonspecific hybridization of the PM and MM probes. The relatively small value shows that nonspecific RNA fragments, on the average, possess similar affinities for PM and MM probes. This result confirms the initial intention to use MM probes to correct the PM intensities for contributions due to nonspecific hybridization. Note that the total chip average of the logarithmic intensity difference over all probe pairs, $\langle \log I^{PM-MM} \rangle_{chip} \approx 0.2$, is close to the low-concentration limit of the mean over the spiked-in probes. We conclude that most of the probes of the considered chips are nonspecifically hybridized.

Specific hybridization progressively dominates the observed intensity in the intermediate concentration range of spiked-in transcripts. With $x^S \gg (1 - x^S) \cdot r^P$ and $(\kappa^{P,S} \cdot x^S) \ll 1$, eq 14 transforms into a linear relationship between the probe intensity and the specific transcript concentration, $\langle \log I^P \rangle \propto \log(c^S_{RNA}) \propto \log(x^S)$. The vertical shift between the PM and MM curves increases with increasing x^S and levels off to $\log(\kappa^{PM,S}/\log \kappa^{MM,S}) \approx 0.85$, which provides a measure of the logarithmic ratio of the effective binding constants for specific hybridization of the PM and MM probes (see also the PM−MM panel in Figure 5). It clearly indicates that the PM probes, on the average, possess a stronger affinity compared with those of the respective MMs.

At higher specific transcript concentration, the experimental PM intensity data systematically deviate in the negative direction from linearity, indicating the onset of saturation which is characterized by the condition $(\kappa^{P,S} \cdot x^S) \approx 1$. On the average, the MM intensities are considerably less affected by saturation because of their smaller binding affinity, $\kappa^{MM,S} < \kappa^{PM,S}$.

The Variability of Sensitivity Data: Stochastic and Systematic Effects. The sensitivity data scatter around the abscissa forming a pear-like data cloud (see Figure 3). The total scatter width was estimated as a function of $\delta\langle \log I^{PM} \rangle_{set-chip}$ by the squared running mean of 200 subsequent sensitivity values along the abscissa (see the scatter curves in Figures 3 and 6). This analysis provides a measure of the variability of probes on each chip in terms of the standard deviation $SD(Y^P) = \pm(\langle (\pm Y^P)^2 \rangle)^{0.5}$. To account for asymmetry effects, we separately calculated $SD(Y^P)$ for positive and negative sensitivity values.

Let us divide the observed variability into two contributions from stochastic errors and systematic, probe-specific effects, $SD(Y^P)^2 = SD_{error}(Y^P)^2 + SD_{sys}(Y^P)^2$. The stochastic term is described well by an error model which was recently proposed for chip intensity data[17,18] (see Appendix). The estimated stochastic error is relatively small in the asymptotic limit of large abscissa values, but it considerably increases with decreasing $\langle \log I^{PM} \rangle_{set}$ (see Figure 3 and also Figure 13 in the Appendix). This trend partially explains the increased variability of the sensitivity at small abscissa values. Comparison of $SD(Y^P)$ and $SD_{error}(Y^P)$ leads, however, to the conclusion that the total variability of the sensitivity cannot be explained by stochastic factors, because $SD(Y^P) \gg SD_{error}(Y)$.

The remaining variability of the sensitivity obviously reflects systematic effects which are related to the binding affinity of the probes and to fluorescence emission. The latter contribution can be estimated by means of

$$SD(Y_F) \approx \sqrt{\sum_{N_F=1}^{N_b} B(N^F, N_b, p)(\log N^F - \langle \log N^F \rangle_{set})^2} \approx \frac{SD(N^F)}{\ln 10 \cdot \langle N^F \rangle} = \frac{1}{\ln 10}\sqrt{\frac{(p-1)}{p \cdot N_b}} \approx (\ln 10 \cdot \sqrt{N_b})^{-1}$$

where $B(N^F, N_b, p)$ is the binominal distribution of N^F potentially labeled nucleotides with a probability of occurrence $p = 0.5$ in a sequence of length N_b. This simple approach provides $SD(Y_F) \approx 0.092$ for $N_b = 25$ and $SD(Y_F) \approx 0.055$ for $N_b = 65$. The latter estimation assumes that labeled nucleotides outside of the target region of the RNA fragments also contribute to the fluorescence intensity (vide supra). The former value giving the standard deviation of the 25-mer might be viewed as the upper limit of the inherent scattering width because of the heterogeneous distribution of labeled nucleotides. It is clearly smaller than the observed variability, which is obviously dominated by variations of the binding affinity due to sequence specific effects.

The decreasing scatter width with increasing mean intensity can be partly explained by saturation using the simple model described already (see $SD(Y_{sat})$ in Figure 6). Note that the standard deviation of the PM sensitivities, $SD(Y^{PM})$, decreases from values of approximately 0.6 to 0.2 over the considered range of $\delta\langle \log I^{PM} \rangle_{set-chip}$. The width of scattering and, consequently, the respective standard deviation of the MM sensitivities is slightly larger [$SD(Y^{MM}) = 0.75-0.2$], whereas that of the difference sensitivity, Y^{PM-MM}, is clearly smaller [$SD(Y^{PM-MM}) = 0.5-0.2$]. The latter result reflects a high degree of correlation between the PM and MM sensitivities. In addition, nonspecific hybridization also seems to increase the scatter width of the sensitivity data in the range of low mean intensity values. This trend can be attributed to the relatively low affinity of the chemical background (see text to follow).

Chip Averaged Mean Sensitivity as a Function of the Base Composition of the Probes. The variability of the binding affinity between the probes and related saturation effects depend on molecular interactions between probe and target, and thus, they are functions of the base composition of the oligonucleotides. In a next step, we therefore analyzed the sensitivities of PM and MM probes as functions of simple sequence characteristics such as the number of each base A, T, G, or C per probe (see Figure 7). The mean sensitivity, averaged over all probes con-

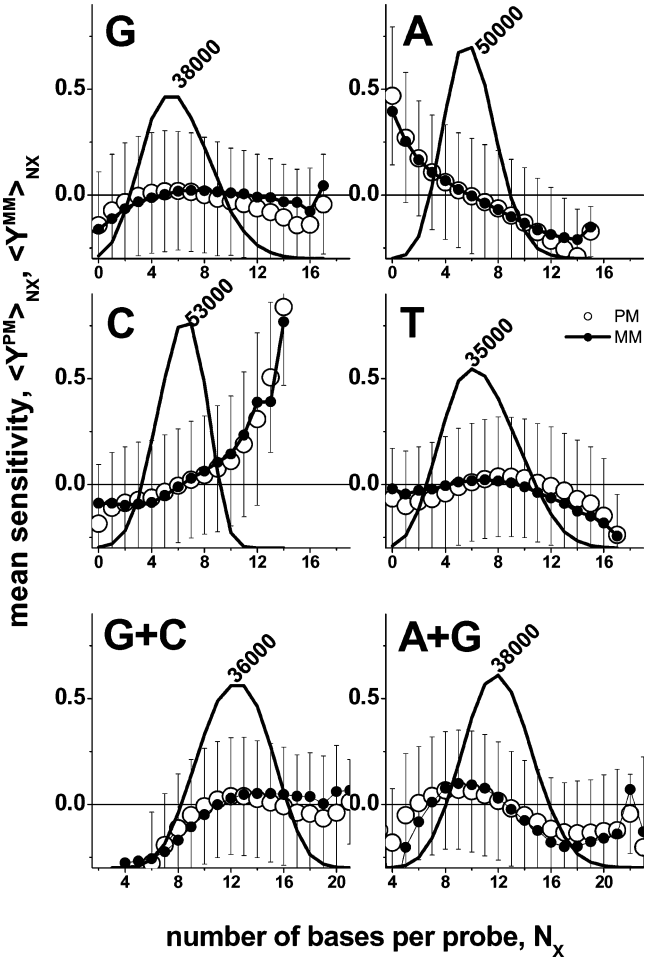

Figure 7. Mean sensitivities of PM (○) and MM (●) probes as a function of the number of one (X = A, T, G, or C, see figure) or two (X = G + C or A + G) nucleotide bases per probe sequence, N_X (see figure). The averages were taken over all PM and MM probes of a HG U133 chip. The error bars indicate the respective variability in terms of the standard deviation. The bell-shaped curves are the number distributions of probes containing N_X bases of letter X. Their maximum value is given within the figure. Note that the area under the distribution is the total number of probes per chip (~248 000). The horizontal line at $Y^P = 0$ provides the mean sensitivity of each probe set.

taining a certain number of bases of one type, strongly increases with an increasing number of C nucleotides per 25-mer by more than one order of magnitude, whereas A residues give rise to the opposite tendency. On the other hand, the probe sensitivity is nearly independent of the number of G and T except a slight decrease at higher numbers of G/T. This tendency can be attributed to the depletion of C with an increasing number of nucleotides other than C. Interestingly, there are only tiny differences between the behavior of the mean PM and MM sensitivities (compare solid and open circles in Figure 7).

In a first-order approximation, one expects similar changes with the number of letters for complementary bases, because the binding strengths of unlabeled A−u and T−a (and of C−g and G−c) pairs differ only slightly[19] (see also the accompanying paper[11]). In contrast, the chip data shown in Figure 7 reveal a strong asymmetry between the behavior of A−u* and T−a pairs, as well as of G−c* and C−g pairs (the asterisk indicates the label). These differences can be partly attributed to biotinylation and fluorescence labeling of the u* and c* of the RNA fragments.[3] Obviously, labeling of the target, on the average, reduces the affinity of the respective base pairs and, consequently, also the sensitivity of the probe.

This effect also becomes evident if one plots the mean sensitivity as a function of the A/G content (i.e., of the number of potentially labeled base pairs G−c* and A−u* per probe sequence ($N^F = N_{G+A} = N_G + N_A$) (see Figure 7)). The sensitivity slightly increases up to $N_{G+A} \approx 8$, presumably because the number of potentially emitting nucleotides increases with N_{G+A}. The mean sensitivity, however, starts to decrease with further increasing the argument at $N_{G+A} > 9$. The potential increase of intensity is overcompensated by a weakening of the binding affinity, presumably because of labeling ($K_F < 1$; see eq 7).

The behavior of the mean intensity as a function of the number of G and C residues ($N_{G+C} = N_G + N_C$) per probe sequence gives further evidence of the asymmetry of G−c* and C−g pairs. As expected, the mean sensitivity increases up to $N_{G+C} \approx 11$, owing to the increasing amount of C. The mean sensitivity remains, however, nearly constant with further increasing G/C content. Note that the number of adjacent G and C residues also increases with increasing G/C content of the probe. Hence, GC and CG nearest neighbors along the probe sequence obviously diminish the sensitivity. Interestingly, the mean fraction of GC couples per position present in all probes of Affymetrx chips is considerably smaller, by a factor of 4, than its expected value in the case of randomly distributed letters (~250 000/16 ≈ 16 000). The manufacturer is obviously aware of the discussed effect.

The Correlation Between the Composition and Sensitivity. To get further insights into the effect of base composition on the sensitivity of the probes, we calculated the mean number of each base letter per probe divided by the mean number of base X per chip $\langle N_X\rangle_{1000}/\langle N_X\rangle_{chip}$ (X = A, T, G, C), and correlated these values with the respective sensitivities (Figure 8). The angular brackets, $\langle\cdots\rangle_{1000}$, denote running averages over 1000 subsequent probes along the abscissa. Systematic deviations of the ratio $\langle N_X\rangle_{1000}/\langle N_X\rangle_{chip}$ from unity indicate a nonrandom base composition of the respective probes. The results clearly show that PM and MM probes of weak sensitivity contain a relatively high fraction of A and T, whereas the fraction of G and C is depleted. Note that C gives rise to a similar effect as G at $Y^P < 0$. Also, A and T behave in a symmetrical fashion. In contrast, at $Y^P > 0$, all letters asymmetrically affect the probe sensitivities. For example, the probes enrich with C with increasing sensitivity over the whole Y^P range, whereas the content of G depletes at higher sensitivity values, presumably because the number of unfavorable GC couples increases (see also ref 11).

Interestingly, the mean number of T residues changes in a more complicated fashion near $Y^P \approx 0$. The increase of the sensitivity in the intermediate Y^P range is accompanied by a marked accumulation of T, whereas at higher and lower Y^P values, the T content shows the opposite tendency (i.e., it decreases with increasing sensitivity). These trends indicate that base-specific effects are related to more detailed sequence characteristics such as nearest-neighbor or triple interactions, which are analyzed in detail in the accompanying paper.[11]

The composition dependence of PM and MM probe sensitivities is very similar (Figure 8). The respective plot of $\langle N_X\rangle_{1000}/\langle N_X\rangle_{chip}$ versus Y^{PM-MM} roughly looks like the mirror image of the respective plot of the base composition as a function of Y^{MM}. This effect can be trivially explained by the wider scattering width of the MM sensitivity values about the origin. As a result, the A/T and G/C content increases/decreases with the increasing sensitivity difference of the probe pairs.

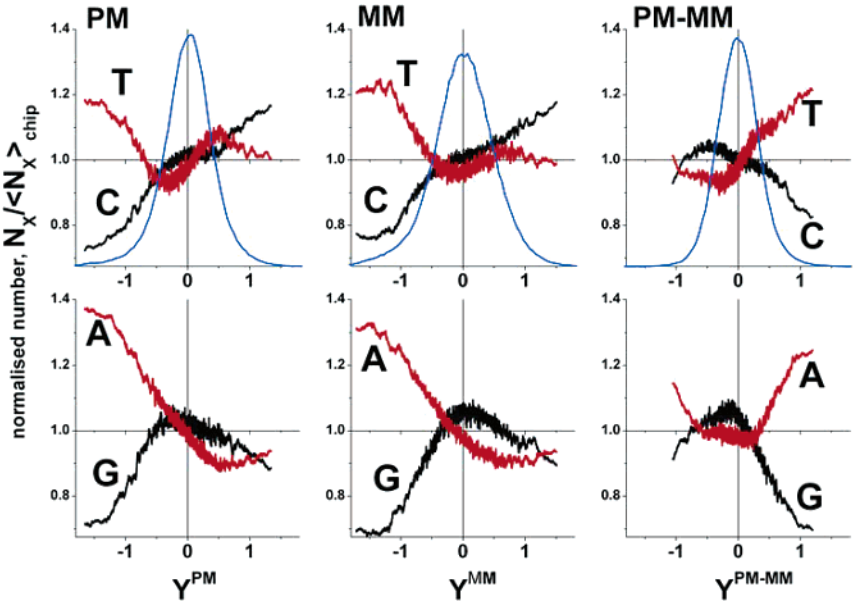

Figure 8. Normalized mean number of bases X = A, T, G, C (see figure) per sequence, $N_X \equiv \langle N_X\rangle_{1000}$, as a function of the sensitivity of the probes Y^P for PM (left panel) MM (middle) and the difference PM − MM (right panel). N_X was calculated as a running average over 1000 probes (see text) and normalized with respect to the mean number of the respective letter per chip, $\langle N_X\rangle_{chip}$. The horizontal line at $N_X/\langle N_X\rangle_{chip} = 1$ refers to the chip average of the respective base. The probes on the HG 133 chip contain, on the average, $\langle N_A\rangle_{chip} = 5.9 \pm 1.9$ adenines, $\langle N_T\rangle_{chip} = 6.8 \pm 2.0$ thymines, $\langle N_C\rangle_{chip} = 6.2 \pm 1.8$ cytosines, and $\langle N_G\rangle_{chip} = 6.1 \pm 2.5$ guanines. The bell-shaped curve in the part above shows the respective probability distribution of the probes.

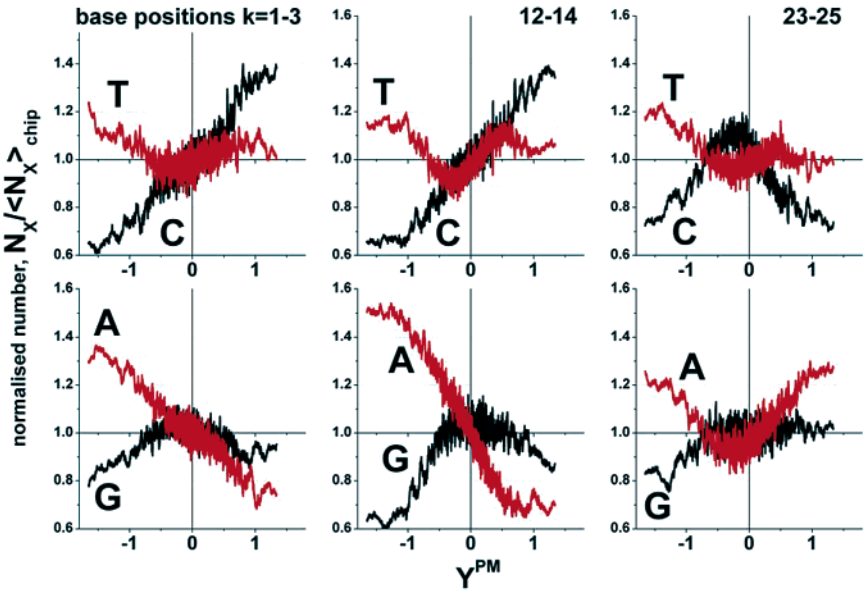

Figure 9. Normalized mean number of bases X = A, T, G, C (see figure) at positions $k = 1-3$ (left panel), 12−14 (middle panel), and 23−25 (right panel) of the PM probe sequence. Position $k = 1$ refers to the 3′ end, which is attached to the chip, whereas $k = 25$ is the free 5′ end of the probe. See legend of Figure 8 and text for details.

Position-Dependent Effect of Base Composition. Microarray oligomer probes are fixed at the quartz surface with the 3′ end (base position $k = 1$), whereas the free 5′ end ($k = 25$) faces away from the chip. This asymmetry also implies a sensitivity profile along the sequence, because entropic factors are expected to locally modulate the binding affinity between the probe and the target. Figure 9 shows the normalized base composition, $\langle N_X\rangle_{1000}/\langle N_X\rangle_{chip}$, for three sequence ranges of the oligomers referring to the fixed end ($k = 1-3$), to the middle ($k = 12-14$), and to the free end ($k = 23-25$) of the probes.

For the position in the middle of the sequence, one observes a considerably wider gap between the local concentrations of A and G residues in the range of small sensitivities compared with that for positions near the 3′ and 5′ ends of the 25-mer. Hence, the specific effect of adenines, namely the correlation between weak sensitivity values and a high local concentration of A, is obviously maximum in the middle of the sequence. Interestingly, near the free end at $k = 23-25$, the nucleotide C depletes in the range of high sensitivities, whereas A considerably accumulates at $Y^P > 0$. These trends are in contrast to the monotonic alterations of the composition of these bases observed for positions $k = 1-3$ and 12−14 throughout the whole Y^P range. At the free end, this relationship reverses. Here, the C nucleotides, on the average, even seem to destabilize the duplexes on a relative scale, whereas the enrichment of A correlates with higher sensitivities.

These results show that the composition dependence on the sensitivity changes along the sequence. The puzzling relationship

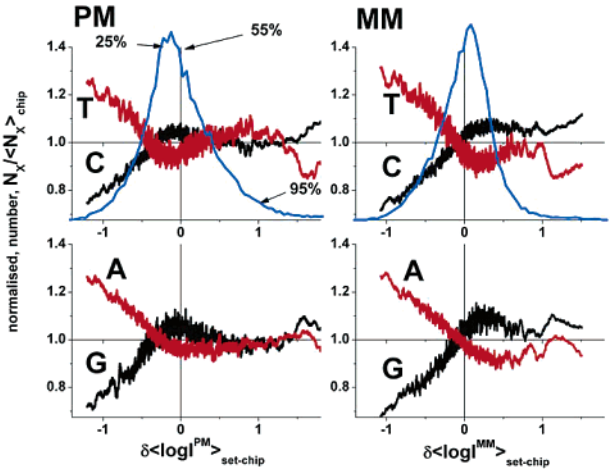

Figure 10. Normalized mean number of bases X = A, T, G, C (see figure) per sequence, $N_X \equiv \langle N_X \rangle_{1000}$, as a function of the normalized set-averaged intensity $\delta\langle \log I^{P} \rangle_{set-chip}$ for PM (left panel) and MM probes (right panel). See also legend of Figure 8 for further details. The bell-shaped curve in the part above shows the respective probability distributions of the probes. The percent values refer to the number of probes with abscissa values smaller than the indicated point (see arrow).

between the mean probability of the appearance of the nucleotides and the sensitivity suggests a more detailed description in terms of base- and position-dependent models. A few first approaches using position-dependent single-base[3,7] and nearest-neighbor[8,20] models were recently published.

Correlations Between the Base Composition and the Set-Averaged Intensity. Figure 10 shows the relative base composition as a function of the normalized set average of the intensity $\delta\langle \log I^{PM} \rangle_{set-chip}$, which was introduced as a relative measure of the target concentration (see eq 10). Note that the running average of $\langle N_X \rangle_{1000}$ over 1000 probes refers to 40−90 probe sets. At small abscissa values, $\delta\langle \log I^{PM} \rangle_{set-chip} < -0.25$, we observe a similar composition dependence as in Figure 8. Namely, probe sets with high A/T and low G/C contents accumulate at the low-intensity and -sensitivity end of the respective *y*-axis. This result shows that the base composition of both the respective probe sets and the individual probes shows a similar trend. There is, however, no rationale for assuming a correlation between the transcript concentration and the base composition of the respective probe set. The accumulation of probe sets which contain probes with an extraordinarily high A and/or low C content presumably reflects a sort of chemical background due to the relatively low affinity of the probes of the respective probe sets.

The G/C content was recently used as a measure of the level of nonspecific hybridization to correct raw intensity values.[21] If one defines the background level via a threshold value in the range of low intensity, then this threshold intensity correlates with the G/C content according to our results. At higher probe intensities, the C and/or A content of a probe seems, however, to provide more suitable measures for estimating their affinity for, for example, nonspecific binding, because G and/or T shows a more puzzling behavior.

The Systematic Bias Between PM and MM Probe Sensitivities is Related to the Middle Base. Figure 11 correlates MM with PM sensitivities for PM probes with a common middle base at position $k = 13$ of the sequence. There is a clear preference of PM with the middle bases G and A to be paired with relatively sensitive MM ($Y_p^{PM} < Y_p^{MM}$), in agreement with previous results.[3] The relationship reverses for middle bases C and T ($Y_p^{PM} > Y_p^{MM}$). Figure 11 further indicates a strong correlation between MM and PM sensitivities. Linear regressions of $Y^{MM} \approx (s \cdot Y^{PM} + \delta)$ to the data shown in Figure 11 provides slopes of $0.97 > s > 0.80$ and vertical shifts in the upward direction of $\delta = +0.16 \pm 0.03$ for middle bases G and A, and a similar downward shift for C and T.

The preference of PM probe sequences with middle bases G and A for relatively sensitive MM and vice versa for middle C and T for less sensitive MM becomes evident in the plots of the sensitivity difference, Y^{PM-MM}, as a function of the relative mean intensity per set, $\delta\langle \log I^{P} \rangle_{set-chip}$. Figure 12 shows the respective scatter plots separately for each middle base. All data clouds referring to a certain middle nucleotide show the typical pear-like shape. The respective probability density distributions given in the part above reveal a slightly wider spread of the

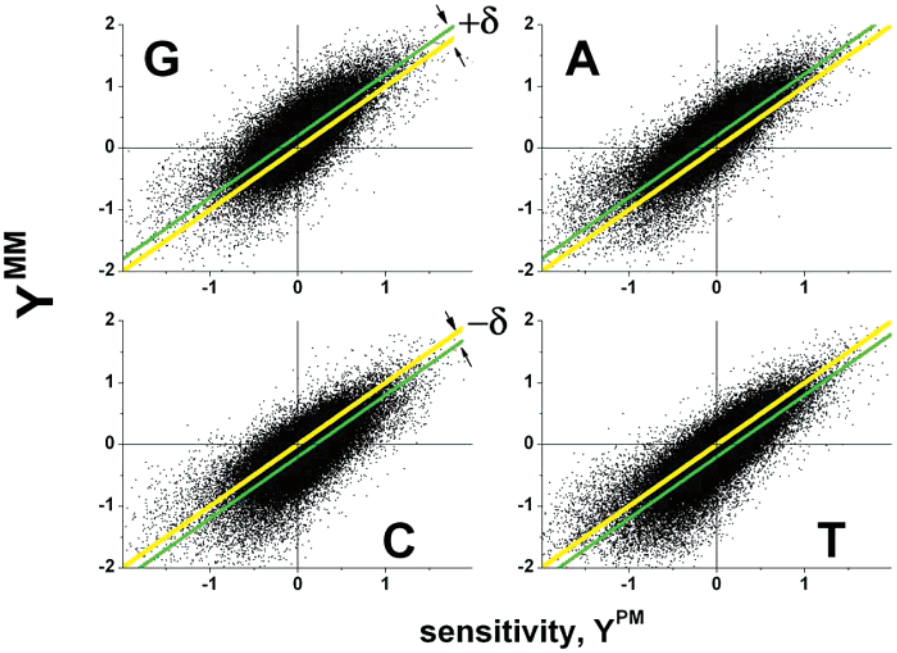

Figure 11. Correlation plot of MM versus PM sensitivities. Each of the panels considers only PM probes with the middle base G, A, C, or T (see figure). The diagonal lines correspond to $Y^{PM} = Y^{MM}$, i.e., data points above the diagonal refer to relatively sensitive MMs, $Y^{PM} < Y^{MM}$ (preferentially for middle bases G and A, whereas C and T are biased toward $Y^{PM} > Y^{MM}$). The thin lines correspond to $Y^{MM} = Y^{MM} \pm \delta$ (see text).

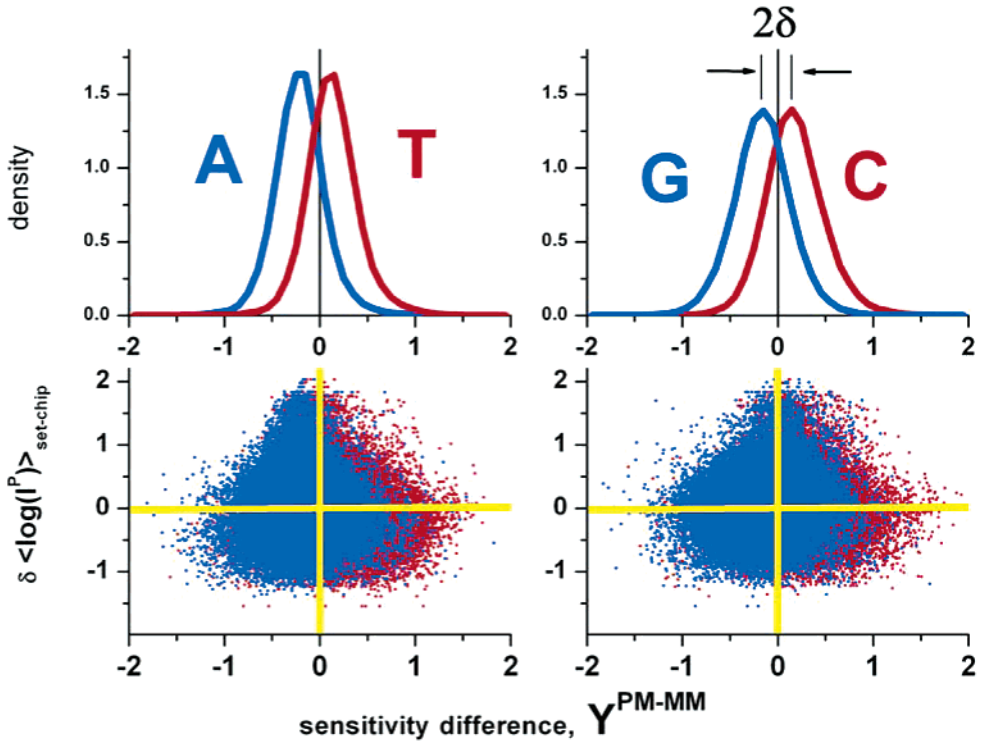

Figure 12. Scatter plot of the relative mean intensity of the probe set, $\delta\langle \log I^{PM+MM}\rangle_{set-chip} = 0.5\langle \log I^{PM} + \log I^{MM}\rangle_{set-chip}$, versus the sensitivity difference, Y^{PM-MM} (panel below) for probes with middle bases A and T (left) and G and C (right). The panel above shows the respective probability densities of Y^{PM-MM}, which is defined as the fraction of data points per abscissa increment, $\rho = N_{total}^{-1}\, \delta N/\delta Y^{P}$. See figure for assignments.

probes with middle bases C and G compared with those with A and T. The maxima of the probability density distribution for both middle bases G and A are shifted by a common increment of $2\delta \approx 0.3 \pm 0.05$ relative to those with C and T (see Figure 12).

The mean increment between the sensitivities of PM probes in the middle of the sequence and the respective MM sensitivities provide a first rough measure of the mismatch effect. For pyrimidines (C and T), the increment is positive, $\delta \approx +0.15$. This relationship reverses for purines (G and T), which provide $\delta \approx -0.15$.

Summary and Conclusions

Our sensitivity concept of microarray oligomer probes divides the probe intensity into two additive contributions: (i) the set-averaged intensity value, which serves as a relative measure of the target concentration and (ii) the sensitivity of a probe, which characterizes its ability to detect a certain amount of RNA in microarray hybridization experiments. We defined the sensitivity as the deviation of the background-corrected intensity of a probe from the mean over the respective probe set in a logarithmic scale. Theoretical considerations based on physicochemical principles show that the sensitivity can be decomposed into terms based on specific and nonspecific hybridization, saturation, the heterogeneous distribution of labels, and the intramolecular folding of target and probe.

The sensitivity of the probes of typical GeneChips varies by more than two orders of magnitude. This range is mainly caused by the sequence-specific binding affinity between the DNA oligonucleotide probes and the RNA fragments. The number of the light-emitting fluorescently labeled nucleotide bases per probe only weakly affects the sensitivity. Effects such as saturation of the probes with bound RNA and folding of probe and target are expected to make the probes insensitive. The former effect significantly increases with RNA concentration, giving rise to a nonlinear relationship between the set-averaged intensity and the amount of specific transcripts.

We analyzed the sensitivities of perfect match (PM) and mismatch (MM) probes of Affymetrix GeneChips as a function of simple sequence characteristics and discovered the following:

(1) The sensitivity of PM and MM probes increases with an increasing number of C nucleotides but decreases with an increasing number of A per probe sequence. These trends are asymmetrical with respect to the number of G and T, which have a much weaker, and perhaps even opposite, effect in the range of intermediate and high sensitivity values.

(2) Complementary bases similarly affect probes of weak sensitivity. We conclude that probes of weak sensitivity can be identified by their G/C and/or T/A content. For probes of high sensitivity, an analogous conclusion seems not to be as simple, because G and C (and A and T) affect the sensitivity in a different fashion.

(3) The relationship between the base composition and sensitivity is virtually identical for PM and MM probes.

(4) The middle base systematically influences the relationship between PM and MM sensitivities. The MMs are, on the average, more sensitive than the PMs in probe pairs with purines (G and A) in the middle of the PM sequence. For pyrimidines (C and T), this relationship reverses. This purine−pyrimidine asymmetry is possibly related to the effect of labeling.

The results clearly indicate a systematic relationship between the chosen sequence characteristics and the sensitivity of the probes. The sensitivity of a particular probe is governed by an intricate interplay between different effects such as base-specific binding, folding, labeling, and saturation, which requires further studies in terms of molecular models.

Acknowledgment. We thank Dr. Ivo Hofacker (University of Vienna) and Dr. Peter Richter (University of Leipzig) for

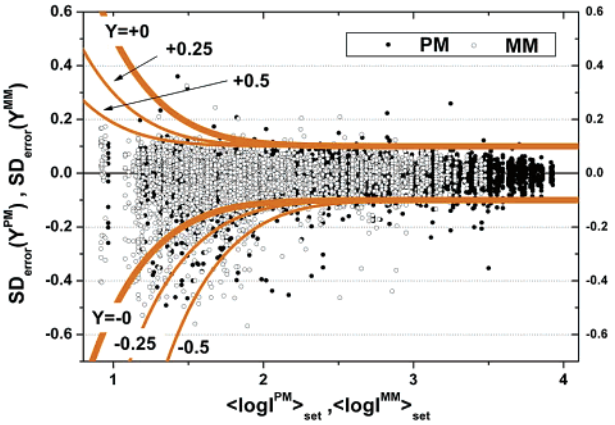

Figure 13. Standard deviation of PM and MM sensitivities of the spiked-in genes taken from three replicates of the HG U133-LS experiment as a function of the respective set-averaged intensities. The sign of the SD values agrees with the sign of the sensitivity values. The lines are calculated according to the error model (see eq A2) for $Y = \pm 0, \pm 0.25, \pm 0.5$ using $a_c^2 = 0.1$, $b_c^2 = 25$.

discussion of aspects of the paper. The work was supported by the Deutsche Forschungsgemeinschaft under Grant BIZ 6/1.

Appendix

Sensitivity Error of Affymetrix GeneChips. The signal intensity can be written to a good approximation as $I' \approx \langle I\rangle \cdot \exp(\ln 10 \cdot e_\sigma) + \langle \boldsymbol{\beta}\rangle + e_\beta$ where $\langle \boldsymbol{\beta}\rangle$ denotes the mean background noise and $\langle I\rangle$ is the mean background-corrected signal intensity[17,18] (see also ref 22). The symbols e_σ and e_β denote normally distributed error terms with mean 0 and variance s_σ^2 and s_β^2, respectively. After background correction and logarithmic transformation of the intensity, one obtains for the variance in the asymptotic approximation[18]

$$\mathrm{var}_{\mathrm{mod}}(\log I^{\mathrm{P}}) \approx a^2 + \frac{b^2}{\langle I^{\mathrm{P}}\rangle^2} \qquad \text{(A1)}$$

with $a^2 \approx s_\sigma^2$ and $b^2 \approx s_\beta^2$.

The variance of the sensitivity (eq 1) can be directly related to the variance of the respective signal intensity by means of $\mathrm{var}(Y^{\mathrm{P}}) \approx \mathrm{var}[\log(I^{\mathrm{P}})] + \mathrm{var}[\langle \log(I^{\mathrm{P}})\rangle_{\mathrm{set}}] \approx \mathrm{var}[\log(I^{\mathrm{P}})][1 + (N - 1)^{-1}] \approx \mathrm{var}[\log(I^{\mathrm{P}})]$ where $N = 11-20$ is the number of probes per probe set. After rearrangement of eq 1 into $\log I^{\mathrm{P}} = \langle \log I^{\mathrm{P}}\rangle_{\mathrm{set}} + Y^{\mathrm{P}}$ and insertion into eq A1, one obtains the standard deviation of the sensitivity as a function of the sensitivity value and the set average of the intensity

$$\mathrm{SD}_{\mathrm{error}}^{\mathrm{mod}}(Y^{\mathrm{P}}) \approx \mathrm{SD}_{\mathrm{error}}^{\mathrm{mod}}(\log I^{\mathrm{P}}) \approx \pm\sqrt{a^2 + \frac{b^2}{10^{2(\langle \log I^{\mathrm{P}}\rangle_{\mathrm{set}} \pm Y^{\mathrm{P}})}}} \qquad \text{(A2)}$$

Note that $\mathrm{SD}_{\mathrm{error}}^{\mathrm{mod}}(Y^{\mathrm{P}})$ is asymmetrical with respect to the sign of Y^{P}. We, therefore, define the sign of the standard deviation to agree with the sign of the sensitivity.

Figure 13 shows the sign-dependent standard deviation of PM and MM sensitivities of the spiked-in genes of the HG U133-LS data set, which were determined from the triplicate chip experiments (see Chip Data and Processing section). The respective model curves refer to $\mathrm{SD}_{\mathrm{error}}(Y^{\mathrm{P}})$ (see caption of Figure 13). The increase of signal error is clearly evident at small mean intensity values. At higher mean intensity values, the error of the logarithmically transformed data levels off into a constant.

References and Notes

(1) Lipshutz, R. J.; Fodor, S. P. A.; Gingeras, T. R.; Lockhart, D. J. *Nat. Genet.***1999**, *21*, 20.
(2) Chudin, E.; Walker, R.; Kosaka, A.; Wu, S.; Rabert, D.; Chang, T.; Kreder, D. *GenomeBiology* **2001**, *3*, 1465.
(3) Naef, F.; Magnasco, M. O. *Phys. Rev. E* **2003**, *68*, 11 906.
(4) Hekstra, D.; Taussig, A. R.; Magnasco, M.; Naef, F. *Nucleic Acids Res.* **2003**, *31*, 1962.
(5) Naef, F.; Lim, D. A.; Patil, N.; Magnasco, M. *Phys. Rev. E* **2002**, *65*, 4092.
(6) Matveeva, O. V.; Shabalina, S. A.; Nemtsov, V. A.; Tsodikov, A. D.; Gesteland, R. F.; Atkins, J. F. *Nucleic Acids Res.* **2003**, *31*, 4211.
(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*, 11237.
(8) Zhang, L.; Miles, M. F.; Aldape, K. D. *Nat. Biotechnol.* **2003**, *21*, 818.
(9) Held, G. A.; Grinstein, G.; Tu, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7575.
(10) Wu, Z.; Irizarry, R. A. Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Microarrays. In *RECOMB'04*; San Diego, CA, 2004.
(11) Binder, H.; Kirsten, T.; Hofacker, I.; Stadler, P.; Loeffler, M. Interactions in Oligonucleotide Hybrid Duplexes on Microarrays. *J. Phys. Chem. B* **2004**, *108*, 18015.
(12) *Affymetrix Microarray Suite*, Version 5.0; Affymetrix, Inc.: Santa Clara, CA, 2001.
(13) Mathews, D. H.; Burkard, M. E.; Freier, S. M.; Wyatt, J. R.; Turner, D. H. *RNA* **1999**, *5*, 1458.
(14) Technical Note In *New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays*; Affymetrix, Inc.: Santa Clara, CA, 2001.
(15) Marras, S. A. E.; Kramer, F. R.; Tyagi, S. *Nucleic Acids Res.* **2002**, *30*, e122.
(16) Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry*; W. H. Freeman and Company: New York, 2002; Vol. 1.
(17) Durbin, B. P.; Hardin, J. S.; Hawkins, D. M.; Rocke, D. M. *Bioinformatics* **2002**, *18* (Suppl. 1), S105.
(18) Rocke, D. M.; Durbin, B. *J. Comput. Biol.* **2001**, *8*, 557.
(19) Sugimoto, N.; Nakano, S.; Katoh, M.; Matsumura, A.; Nakamuta, H.; Ohmichi, T.; Yoneyama, M.; Sasaki, M. *Biochemistry* **1995**, *34*, 11211.
(20) Binder, H.; Kirsten, T.; Loeffler, M.; Stadler, P. Sequence specific sensitivity of oligonucleotide probes. In *Proceedings of the German Bioinformatics Conference*; Munich, Germany, October 12−14, 2003.
(21) Wu, Z.; Irizarry, R. A.; Gentleman, R.; Murillo, F. M.; Spencer, F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays; Department of Biostatistics, Johns Hopkins University, 2003.
(22) Naef, F.; Hacker, C.; Patil, N.; Magnasco, M. *GenomeBiology* **2002**, *3*, research 0018.1-0018.11.