Base Pair Interactions and Hybridization Isotherms of Matched and Mismatched Oligonucleotide Probes on Microarrays

Hans Binder,* Stephan Preibisch, and Toralf Kirsten

Interdisciplinary Centre for Bioinformatics, University of Leipzig, Haertelstr. 16-18, D-04107 Leipzig, Germany

The microarray technology enables the expression degree of thousands of genes to be estimated at once by the measurement of the abundance of the respective messenger RNA. This method is based on the sequence specific binding of RNA to DNA probes and its detection using fluorescent labels. The raw intensity data are affected by the sequence-specific affinity of probe and RNA for duplex formation, by the background intensity due to nonspecific hybridization at small transcript concentrations and by the saturation of the probes at high transcript concentration owing to surface adsorption. We address these issues using a binding model which describes specific and nonspecific hybridization in terms of a competitive two-species *Langmuir* isotherm and DNA/RNA duplex formation in terms of sequence-specific, single-base related interactions. The GeneChip microarrays technology uses pairs of so-called perfect match (PM) and mismatch (MM) oligonucleotide probes to estimate the amount of nonspecific hybridization. The mean affinity of the probes decrease according to PM(specific) > MM(specific) ≫ PM(nonspecific) ≈ MM(nonspecific). The stability of specific and nonspecific DNA/RNA duplexes is mainly determined by Watson Crick (WC) pairings. Mismatched self-complementary pairings in the middle of the MM sequence only weakly contribute to the duplex stability. The asymmetry of base pair interaction in the DNA/RNA hybrid duplexes gives rise to a duplet-like symmetry of the PM − MM intensity difference at dominating nonspecific hybridization and a triplet-like symmetry at specific hybridization. The signal intensities of the PM and MM probes and their difference are assessed in terms of sensitivity and specificity. The presented results imply the refinement of existing algorithms of probe level analysis to correct microarray data for nonspecific background intensities and saturation on the basis of the probe sequence.

Introduction

Gene expression microarray chips consist of DNA oligomers with up to several hundreds of thousands of different sequences that are immobilized onto a support such as glass, silicon, or nylon membrane in a spot-like arrangement. They provide a powerful functional genomics technology, which permits the expression profiling of tens of thousands of genes in parallel.[1,2] The working principle of this technology is based on duplex formation (hybridization) between target messenger RNA extracted from cell lines or tissues on one hand and complementary DNA nucleotide strands grafted to the chip (the reporter or probe molecules) on the other hand. Formed duplexes are detected using fluorescent or radionucleotide labels. Each spot on the chip consists of oligomers of one sequence. It is therefore representative for a certain gene and probes the abundance of the respective RNA transcript. The microarray technology also allows high-throughput genotyping using SNP-microarrays which detect specific probe/target duplexes between complementary DNA strands.

Different types of DNA arrays are designed for RNA profiling, which differ by the type of probes (cDNA or synthetic oligonucleotides) and by the DNA density on the array (see e.g. ref 3). So-called high-density-oligonucleotide-arrays (HDONA) are produced by a photolithographic technology, which allows synthesis of oligonucleotide sequences on the chip surface in an extremely high density. This way 10^5-10^6 different probe spots can be localized on one microarray of an area of about one squared centimeter.[4] The probe intensity, i.e., the integral fluorescence intensity of each probe spot, is related to the amount of bound, fluorescently labeled RNA, which in turn serves as a measure of the concentration of complementary RNA in the sample solution used for hybridization and thus of the expression degree of the respective gene.

HDONA arrays of the so-called GeneChip type (Affymetrix Inc., Santa Clara) use so-called probe sets of 11 (in some cases up to 20) different 25meric reporter sequences for each gene.[4] The processing of a set of several fluorescence intensities per gene is expected to improve the reliability of the method. Note that the sample RNA is cleaved into fragments with a length of several dozen nucleotides before hybridization. The RNA fragments referring to different regions of the target gene are expected to bind virtually independently to the oligonucleotide probes of one set.

The target RNA for each probe constitutes only a fraction of the total RNA in the sample solution used for hybridization. A considerable amount of RNA involves other sequences than the intended target. Unfortunately, these nonspecific transcripts compete with the target RNA for duplex formation with the probes. This way they also contribute to the signal intensities due to nonspecific binding. The lack of specificity raises a serious problem for the analysis of microarray data because the residual "chemical background" intensity is not related to the expression degree of the gene of interest and therefore distorts the signal of specifically bound target RNA.

To deal with this problem, each probe sequence on GeneChip micoarrays is present in two modifications

* Corresponding author. E-mail: binder@izbi.uni-leipzig.de. Fax: ++49-341-97-16679.

(1) Schena, M. *Bioessays* **1996**, *18*, 427.

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

(3) Hong, B.-J.; S−J, O.; Youn, T.-O.; Kwon, S.-H.; Park, J.-W. *Langmuir* **2005**, *21*, 4257.

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

called perfect match (PM) and mismatch (MM) probes. The sequence of the PM is taken from the gene of interest, and thus, it is complementary to a 25mer 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 sequence where the "middle base" is replaced by its complement. The MM are designed as reporters for nonspecific hybridization that measure the intensity of the "chemical background", i.e., of bound RNA not referring to the target gene. The MM signal provides a potential correction of the respective PM intensity for the contribution of nonspecific binding.

The idea behind the pairwise design of probes is based on the assumption that nonspecific transcripts bind with virtually identical affinity to the PM and MM probes of one pair, whereas the target RNA is expected to hybridize the MM with considerable less affinity due to the mismatched base pairing in accordance with "conventional" hybridization thermodynamics of DNA/RNA duplexes in solution.[5] The lower stability of mismatched probe/target duplexes predicts a systematically equal or smaller intensity of the MM probes. It was however found, that a considerable fraction of the MM probes fluoresces with higher intensity than the paired PM.[6] Consequently, subtracting MM from PM intensities as a way of correcting the PM intensities for nonspecific binding seems not always appropriate.[7,8] As a consequence, the "mysterious" MM were either completely ignored in signal analysis algorithms [9,10] or they are considered in an empirical fashion to exclude "bad" probes from the analysis.[4]

Hence, one important question for GeneChip data analysis is how to include the MM intensities adequately. This more technical issue, in turn, requires the detailed understanding of the basic rules of oligonucleotide duplex formation on microarrays and, in particular, of the hybridization mechanism of matched and mismatches microarray probes with specific and nonspecific RNA transcripts on the level of base pairings. The effect of competitive hybridization of specific and nonspecific RNA fragments on the thermodynamically attainable performance of DNA chips can be quantified in terms of the hybridization (or binding) isotherms of the PM and MM probe spots. The isotherms provide a basic characteristic of the probes because they relate the degree of the hybridization to the bulk RNA composition and thus to the expression degree. The instrumental response characteristic must in addition consider the effect of selective labeling which produces the fluorescence intensity measured by the detector.

The present paper addresses these issues in terms of a hybridization model, which explicitly considers the RNA concentration and the amount of specific and nonspecific transcripts in the sample solution on one hand and the sequence of the oligonucleotide probes and especially their middle base on the other hand. The theoretical results are compared with microarray intensity data, which were taken from a calibration experiment provided by Affymetrix.

(5) Sugimoto, N.; Nakano, S.; Katoh, M.; Matsumura, A.; Nakamuta, H.; Ohmichi, T.; Yoneyama, M.; Sasaki, M. *Biochemistry* **1995**, *34*, 11211.

(6) Naef, F.; Lim, D. A.; Patil, N.; Magnasco, M. *Phys. Rev. E* **2002**, *65*, 4092.

(7) Irizarry, R. A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y. D.; Antonellis, K. J.; Scherf, U.; Speed, T. P. *Biostat* **2003**, *4*, 249.

(8) Naef, F.; Hacker, C.; Patil, N.; Magnasco, M. *Genome Biol.* **2002**, *3*, research0018.1.

(9) Irizarry, R. A.; Bolstad, B. M.; Collin, F.; Cope, L. M.; Hobbs, B.; Speed, T. P. *Nucl. Acids. Res.* **2003**, *31*, e15.

(10) Huber, W.; von Heydebreck, A.; Sueltmann, H.; Poustka, A.; Vingron, M. *Bioinformatics* **2002**, *1*, 1.

Microscopic Model of Hybridization on Microarrays

Binding Affinity and the Intensity of Oligonucleotide Probes. Gene expression analysis by means of high-density-oligonucleotide-array (HDONA) chips is based on the sequence specific binding of RNA fragments to oligonucleotide probes and its measurement using fluorescent labels. Affymetrix uses short 25mers as perfect match (PM) probes, the sequence, ξ^{PM}, of which is complementary to a fragment of the consensus sequence of the respective target gene, ξ^{T}.[4] The probe and target sequences are given by strings of $N_b = 25$ letters (A, T, G or C), e.g., $\xi^{PM} = 3'$-ACCCAG...T-$5'$ and $\xi^{T} = 3'$-u*gggu*c*...a-$5'$ (uppercase letters refer to the DNA probe, lower case letter refer to the RNA-target, and the asterisk denotes labeling).

The PM probe intends to bind the target RNA via the Watson Crick (WC) pairings A−u*, T−a, G−c*, and C−g. The respective association constant for duplex formation, $K_p^{PM,S} = K_p^b(\xi_p^{PM}\ \xi_p^T)$ (the index p denotes the probe number, the first superscript specifies the probe type, P = PM or MM; the second superscript indicates the type of hybridization, h = S and NS for specific and nonspecific, respectively), quantifies the strength of specific binding between target and probe according to the binding reaction $(\xi^{PM}\xi^T) \leftrightarrow \xi^{PM} + \xi^T$. The association constant for duplex formation of the mismatch (MM) probe with the target, $K_p^{MM,S} = K_p^b(\xi_p^{MM}\ \xi_p^T)$, characterizes the affinity of target RNA for specific binding despite the fact that the middle base of the MM probe disables WC pairings. Instead, the 13th base is assumed to form the respective self-complementary (SC) pair with the target RNA, A−a, T−u*, G−g, or C−c*.

The sample solution used for hybridization usually contains a large number of RNA fragments with sequences differing from that of the target, i.e., $\xi \neq \xi^T$. Also these nonspecific RNA fragments bind in significant amounts to the probes despite the fact that probe and DNA only partly match each other via WC pairings. The respective association constants, $K_p^P(\xi_p^P\xi)_{\xi \neq \xi^T}$ quantify the affinity of the probe (P = PM, MM) for duplex formation with nonspecific RNA fragments of sequence $\xi \neq \xi^T$ according to the reaction $(\xi^P\xi) \leftrightarrow \xi^P + \xi$. The mean binding affinity of the probe for nonspecific hybridization is given by the concentration-weighted average over the binding constants of this "cocktail" of RNA sequences, $K_p^{P,NS} = \langle K_p^P(\xi_p^P\xi)\rangle_{\xi \neq \xi^T} \equiv \sum_{\xi \neq \xi^T} c_{RNA}(\xi) \cdot K_p^P(\xi_p^P\xi)/\sum_{\xi \neq \xi^T} c_{RNA}(\xi)$. The nonspecific fragments are expected to bind with lower affinity to the probe compared with the target RNA owing to the smaller number of WC pairings. The ratio $r_p^P = K_p^{P,NS}/K_p^{P,S} < 1$ specifies the mean relative binding strength of the probe for nonspecific hybridization compared with that of specific binding with the target sequence, ξ^T.[11]

The amount of probe-bound RNA is detected by means of fluorescent labels, which are linked to the uracyls (u*) and cytosines (c*). The respective fluorescence intensity per probe spot measured by the detector can be described by[11]

$$I_p^P \approx F_{chip}N_p^{F,S}K_p^{P,S}[c_{RNA}^S + c_{RNA}^{NS}r_p^P r_p^{F,P}]S_p^P \qquad (1)$$

if one neglects the optical background. Essentially, four factors affect the signal intensity of a particular probe according to eq 1:

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

(i) The binding "strength" (or affinity) of the DNA probe for duplex formation with the RNA fragments upon hybridization determines the amount of RNA that binds to the probe. It is characterized by the binding constant of specific hybridization, $K_p^{P,S}$, and the mean relative strength of nonspecific binding, r_p^P. In eq 1 the binding equilibria between the probe and all relevant nonspecific RNA sequences are replaced by one equilibrium between the probe and a characteristic nonspecific transcript, which is characterized by the mean binding constant $K_p^{P,NS}$. In other words, the cocktail of nonspecific RNA fragments is assumed to act like a single species in accordance with previous treatments of cross hybridization.[12] Equation 1 further considers saturation of the probes with specific and nonspecific RNA fragments, which both compete for the free binding sites provided by the monomeric oligomers according to a competitive two-species Langmuir isotherm, $S_p^P = (1 + K_p^{P,S}[c_{RNA}^S + c_{RNA}^{NS} r_p^P])^{-1}$ (see also eq 1).

(ii) The fluorescence "strength" (or yield) of the hybridized RNA determines the emitted intensity per bound transcript. It is roughly related to the amount of labeling, which is given by the mean number of fluorescently labeled cytosines and uracyls in the sequence of the respective fragment of specific target RNA, $N_p^{F,S} = N_p^{c*} + N_p^{u*}$. The ratio $r_p^{F,P} = N_p^{F,P,NS}/N_p^{F,S}$ specifies the relation between the amount of labeling of probe-bound nonspecific and specific RNA-fragments. Note that the specific target RNA fragment is identical for the PM and MM probes of one probe pair, whereas the nonspecific RNA effectively differs by one base (see below).

(iii) The total concentration of RNA fragments in the sample solution used for hybridization, $c_{RNA}^{tot} = \sum_\xi c_{RNA}(\xi)$, is directly related to the amount of binding according to the mass action law. It splits into the concentration of target RNA (specific transcripts), $c_{RNA}^S = c_{RNA}(\xi^T)$, and into the concentration of nonspecific transcripts involving other sequences than the intended target, $c_{RNA}^{NS} = c_{RNA}(\xi \neq \xi^T)$.

(iv) The chip specific constant F_{chip} considers the detection "strength" of the technique. It considers aspects of chip fabrication such as the number and density of oligonucleotides per probe spot, the sensitivity of the imaging system and factors due to the performance of the experiment, e.g., the yield of labeling.

Note that the microarray experiment intends to measure the expression degree of the target gene in terms of c_{RNA}^S, the concentration of specific transcript. Signal analysis consequently requires the correction of the measured intensity for the effect of labeling, the chip specific constant, and most importantly, for saturation and nonspecific hybridization.

Mean Binding Isotherm and the Free Energy of Duplex Formation. Let us split the log intensity of each probe into a mean value, $\langle \log I_p^P \rangle_\Sigma$, averaged over an appropriate ensemble of probes Σ referring to one concentration of specific transcripts (i.e., c_{RNA}^S = const) on one hand, and an incremental contribution, which reflects the individual properties of the selected probe, $Y_p^P \equiv \Delta \log I_p^P$, on the other hand, i.e.

$$\log I_p^P = \langle \log I_p^P \rangle_\Sigma + Y_p^P \qquad (2)$$

In this work, we use two options for ensemble averaging. First, in the so-called spiked-in experiment (see below), RNA transcripts of selected probes were titrated onto a series of chips in well-defined concentrations, $c_{RNA}^S =$

$c_{spiked-in}$. In this case, the respective ensemble of probe intensities referring to explicitly known concentration, $c_{spiked-in} = const (\Sigma = spiked-in)$, were taken from different chips. Alternatively, one can pool all probes of a probe set (Σ = set) together because they refer to one gene and consequently to one target concentration with $c_{RNA}^S = const.$, which is however a priori unknown. In this case averaging was performed over probes from one chip.

The mean intensity can be described by an effective binding isotherm adapted from eq 1

$$\langle \log I^P \rangle_\Sigma \approx$$
$$\log F + \log K_0^{P,S} + \log[c_{RNA}^S + c_{RNA}^{NS} r_0^P] + \log S_0^P$$

with

$$\log F = \log F_{chip} + \log N_0^F,$$
$$S_0^P = (1 + K_0^{P,S}[c_{RNA}^S + c_{RNA}^{NS} r_0^P])^{-1},$$
$$\text{and } r_0^P = K_0^{P,NS}/K_0^{P,S} \quad (3)$$

The effective constants $\log K_0^{P,h} = \langle \log(K_p^{P,h}) \rangle_\Sigma$ (h = S, NS) and $\log N_0^F = \langle \log (N_p^{F,S}) \rangle_\Sigma$ represent mean values over all considered probes. Equation 3 assumes that the log-intensity average is a function of these effective values and that hybridized PM and MM probes are equally labeled on the average.

The binding constant of the probes provides the respective Gibbs free energy of duplex formation

$$\Delta G_p^{P,S} = \mu_{duplex}^0(\xi_p^P \xi_p^T) - (\mu_{DNA}^0(\xi_p^P) + \mu_{RNA}^0(\xi_p^P))$$

and

$$\Delta G_p^{P,NS} = \langle \mu_{duplex}^0(\xi_p^P \xi) - (\mu_{DNA}^0(\xi_p^P) + \mu_{RNA}^0(\xi)) \rangle_\xi$$

where the μ^0 denotes the respective standard chemical potentials of the reactants and of the duplex. The free energy of duplex formation can be decomposed into a sum of base and positional dependent contributions

$$\Delta G_p^{P,h} = -RT \ln(WK_p^{P,h}) = RT \ln 10 \sum_{k=1}^{N_b} \epsilon_k^{P,h}(\xi_{p,k}^P)$$
$$\text{with h = S,NS} \quad (4)$$

where $\xi_{p,k}^P$ denotes the nucleotide base at position k of the probe sequence, R and T are the gas constant and the temperature, respectively, and W is the cratic contribution accounting for the mixing entropy.[13]

The free energy terms can be further split into a base independent mean value averaged over the chosen ensemble of probes and into a base dependent contribution in analogy with eq 2

$$\epsilon_k^{P,h}(B) = \epsilon_{0,k}^{P,h} + \Delta \epsilon_k^{P,h}(B) \quad \text{with B = A,T,G,C}$$

and

$$\sum_{k=1}^{N_b} \epsilon_{0,k}^{P,h} = -\log K_0^{P,h} + const \qquad (5)$$

In general, the hybridization at the surface of a DNA chip differs from the Langmuir scenario in that both the adsorbates (the targets) and the surface (the probe layer)

(12) Halperin, A.; Buhot, A.; Zhulina, E. B. *Biophys. J.* **2004**, *86*, 718.

(13) Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry*; W. H. Freeman and Company: New York, 2002; Vol. 1.

are charged. As a result, the free energy of duplex formation incorporates electrostatic terms, which depend on the amount of bound RNA.[12,14] For this situation, the binding constant has to be supplemented by a concentration dependent exponential factor, which considers the progressive depletion of the free adsorbate near the surface owing to electrostatic repulsion between bound and free species. This effect gives rise to a saturation-like behavior where further binding with increasing bulk concentration of the adsorbate is effectively hampered by always bound species. Despite these limitations, we will use the Langmuir form as a good approximation because it provides a satisfactory description of the used experimental data (see below and also refs 15−17). The resulting binding constants (and free energies) must be interpreted as apparent values that include the electrostatic contribution.

The competitive two-species Langmuir isotherm assumes two discrete energetic states for specific and nonspecific hybridization (see above). The explicit consideration of a continuous distribution of binding free energies due to the heterogeneity of RNA sequences can be achieved by the replacement $cK \rightarrow (cK)^a$ (with the exponent $a < 1$) in the respective Langmuir-type isotherm.[18] Note however that even the most critical application of the used Langmuir form to the average over all probes (eq 3) actually provides a good description of the experimental data (see below). We therefore judge this simpler Langmuir version as the adequate approach in this work.

Sensitivity of the Oligonucleotide Probes. The incremental contribution to the intensity

$$Y_p^P = \log I_p^P - \langle \log I_p^P \rangle_\Sigma, \quad P = PM \text{ or } MM \quad (6)$$

defines the sensitivity of the respective probe, which, in a first order approximation, characterizes its ability to detect a certain amount of RNA independently of the experimental conditions given by the chip specific factor F_{chip}. Note that the transformation according to eq 6 cancels out all factors to the intensity, which are common for the chosen ensemble of probes. Our definition of the sensitivity for the special case of oligonucleotide probes on GeneChip microarrays is adapted from the general definition of the IUPAC for analytical techniques, which identifies the sensitivity with the measured response per concentration increment (see ref 19 and references therein).

Insertion of eq 6 into eq 1 shows that the probe sensitivity additively decomposes into terms due to the binding affinity and fluorescence[11]

$$Y_p^P \approx Y_p^{P,b} + Y_p^{P,F}$$

$$Y_p^{P,b} = \log(K_p^{P,S}[c_{RNA}^S + c_{RNA}^{NS} r_p^P r_p^{P,F}]S_p^P) -$$
$$\langle \log(K_p^{P,S}[c_{RNA}^S + c_{RNA}^{NS} r_p^P r_p^{P,F}]S_p^P) \rangle_\Sigma$$

$$Y_p^{P,F} = \log(N_p^{F,S}) - \langle \log(N_p^{F,S}) \rangle_\Sigma \quad (7)$$

Positional Dependent Single Base (SB) Model of the Sensitivity. Positional dependent SB models were recently used to predict microarray probe intensities.[20,21] In our notation, the SB model decomposes the sensitivity of each

probe into a sum of sensitivity contributions $\sigma_k^P(\xi_{p,k}^P)$, depending on the base at position $k = 1, ..., N_b$ of the probe sequence, $\xi_{p,k}^P$

$$Y_p^{P,SB} = \sum_{k=1}^{N_b} \sum_{B = A,T,G,C} \sigma_k^P(B)(\delta(B,\xi_{p,k}^P) - f_k^\Sigma(B)) \quad (8)$$

Here δ denotes the Kronecker delta ($\delta(x,y) = 1$ if $x = y$ and $\delta(x,y) = 0$ if $x \neq y$). The term $f_k^\Sigma(B)$ is the fraction of base B at position k in the considered ensemble of probes, or, in other words, the probability of occurrence of letter B at position k in the ensemble. The SB sensitivity contributions at a given position spread symmetrically about zero; that is, they are restricted to the condition

$$\sum_{B = A,T,G,C} \sigma_k^P(B) = 0 \text{ for all positions k} = 1, ..., N_b \quad (9)$$

The position-averaged SB sensitivity

$$\sigma^P(B) = \langle \sigma_k^P(B) \rangle_k \equiv \frac{1}{N_b} \sum_{k=1}^{N_b} \sigma_k^P(B) \quad (10)$$

characterizes the mean contribution of base B to the sensitivity independently of its position along the sequence. The mean over all bases in terms of absolute values

$$\sigma^P = \langle |\sigma_k^P(B)| \rangle_{k,B} = \frac{1}{4N_b} \sum_{k=1}^{N_b} \sum_{B = A,T,G,C} |\sigma_k^P(B)| \quad (11)$$

can be interpreted as a measure of the variability of the sensitivity of the probes due to sequence specific effects (see also eq 6).

The intensity of a selected probe represents the superposition of the respective ensemble averaged intensity and of the sequence specific contribution given by the SB sensitivity model (see eqs 6 and 8)

$$\log I_p^P \approx \langle \log I^P \rangle_\Sigma + \sum_{k=1}^{N_b} \sigma_k^P(\xi_{p,k}^P) \quad (12)$$

In the general case, both the mean intensity and the SB contributions are functions of the RNA target concentration. Let us neglect saturation for the sake of simplicity. Then insertion of eq 3 into 12 provides

$$\log I_p^P \approx \log F + \log K_0^{P,h} + \log c_{RNA}^h + \sum_{k=1}^{N_b} \sigma_k^{P,h}(\xi_{p,k}^P) \quad (13)$$

in the limiting case of specific (h = S for $c_{RNA}^S \gg c_{RNA}^{NS} r_0^P$) and of nonspecific (h = NS for $c_{RNA}^S \ll c_{RNA}^{NS} r_0^P$) hybridization, respectively. Equation 13 shows that the fit of the SB model to the sensitivities of an appropriately chosen ensemble of probes provides estimates of the SB sensitivity parameters, which characterize specific and nonspecific DNA/RNA probe/target duplexes.

Fluorescence Contribution. The sensitivity of each probe divides into two additive contributions according to eq 7 due to (i) the binding "strength" of the RNA for duplex

(14) Vainrub, A.; Pettitt, B. M. *Phys. Rev. E* **2002**, *66*, art. no. 041905.
(15) Hekstra, D.; Taussig, A. R.; Magnasco, M.; Naef, F. *Nucl. Acids. Res.* **2003**, *31*, 1962.
(16) Burdon, C. J.; Pittelkow, Y. E.; Wilson, S. R. *Stat. Appl. Gen. Mol. Biol.* **2004**, *3*, 35.
(17) Burdon, C. J.; Pittelkow, Y. E.; Wilson, S. R. *arXiv: q-bio. BM/0411005 v1* **2004**.
(18) Sips, R. *J. Phys. Chem.* **1948**, *16*, 490.
(19) Etkins, R.; Edwards, P. *Clin Chem* **1997**, *43*, 1824.

(20) Naef, F.; Magnasco, M. O. *Phys. Rev. E* **2003**, *68*, 11906.
(21) 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.

formation with the probe and (ii) the fluorescence "strength" of bound RNA. A relatively high binding strength consequently represents a necessary but not sufficient condition of highly sensitive probes. In addition the bound RNA must emit light with sufficiently high intensity, which in turn depends on the amount of labeling of the probe/RNA duplex.

Both, the binding affinity and the fluorescence yield are functions of the base composition of the probe. It appears therefore appropriate to split the SB sensitivity into a Gibbs free energy and a fluorescence contribution

$$\sigma_k^{P,h}(\xi_{p,k}^P) = -\Delta\epsilon_k^{P,h}(\xi_{p,k}^P) + \Delta\varphi_k^{P,h}(\xi_{p,k}^P) \quad \text{with h = S,NS} \tag{14}$$

The former term, $\Delta\epsilon_k^{P,h}(\xi_{p,k}^P)$, characterizes the relative binding strength of the base at position k of the probe sequence ξ_p^P (see also eq 5). The fluorescence contribution, $\Delta\varphi_k^{P,h}(\xi_{p,k}^P)$, considers the fact that not every base fluoresces owing to the specificity of labeling for cytosines and uracyls. The free energy and fluorescence contributions are assumed to meet the symmetry condition (eq 9).

The fluorescence intensity of a RNA target fragment is related to the number of labeled bases, c* and u*. It, in turn, depends on the number of complementary bases, B = G and A, in the PM probe sequence if one assumes binding via WC pairs. In the Supporting Information (S1) we show that the positional dependent SB sensitivity contributions of labeled bases are enhanced whereas the contributions of nonlabeled bases are decreased by a constant, positional independent increment Δ^F (see eq A3 in the Supporting Information)

$$\Delta\varphi_k^{PM,S}(B) = \Delta\varphi_k^{WC}(B) \approx \begin{cases} +\Delta^F \text{ for B = A,G} \\ -\Delta^F \text{ for B = T,C} \end{cases} \text{ with}$$

$$\Delta^F \approx \left(1 + \sqrt{\frac{N_b^{RNA}}{N_b}} - 1\right)\frac{2}{\ln 10 N_b^{RNA}} \tag{15}$$

This result assumes direct proportionality between the emitted fluorescence intensity and the number of potentially labeled bases in the target sequence, $I^P \propto N_p^F$. The fluorescence effectively increases the single base sensitivity of labeled A−u* and G−c* WC base pairs and decreases the sensitivity of nonlabeled T−a and C−g pairs in a symmetrical fashion. For self-complementary pairs (A−a, T−u*, G−g and C−c*) the relation reverses, i.e., $\Delta\varphi_k^{WC}(B) = -\Delta\varphi_k^{SC}(B)$.

The increment Δ^F depends on the total sequence length of the RNA fragments, N_b^{RNA}, and on the length of the probe oligomers, $N_b = 25$, which is explicitly considered in the SB model. One obtains $\Delta^F \approx 0.04$ if the target length exactly matches the probe ($N_b^{RNA} = 25$). The fluorescence term remains nearly constant for longer target sequences up to $N_b^{RNA} = 50$ and then it progressively decreases to $\Delta^F \approx 0.03$ for $N_b^{RNA} = 65$ and to values less than 0.02 for $N_b^{RNA} > 100$ nucleotides. Hence, a value of $\Delta^F \approx 0.04$ can be judged as an upper limit of the fluorescence contribution to the positional dependent sensitivity.

The comparison between the binding data of labeled and nonlabeled oligonucleotides shows that labeling (i.e., the covalent linkage of biotinyl residues with attached fluorescent labels to the nucleotide bases) slightly but significantly decreases the binding strength of a nucleotide base by a reduced free energy increment of less than 0.05

(see eq 4,[22]). Hence, the fluorescence strength and the change of the free energy contribution owing to labeling obviously compensate each other at least partially with respect to their effect on the SB sensitivity.

Data Processing and Parameter Estimation

Chip Data. Microarray intensity data are taken from the Affymetrix human genome Latin Square (HG U133-LS) data set available at http://www.affymetrix.com/support/technical/sample_data/datasets.affx. These data are obtained in a calibration experiment, in which specific RNA transcripts referring to 42 genes (and thus to $N_{data} = 11 \times 42 = 462$ PM/MM probe pairs) were titrated in definite concentrations onto microarrays of the Affymetrix HG U133 type to study the relation between the probe intensity and the respective ("spiked-in") RNA concentration. Fourteen different concentrations ranging from 0 pM (i.e., no specific transcripts) to 512 pM were used for each probe. The experiment further uses 14 different arrays for all cyclic permutations of the spiked-in concentrations and spiked-in genes (the so-called Latin Square design). Nonspecific hybridization was taken into account by adding a complex human RNA background extracted from a HeLa cell line not containing the spiked-in transcripts to all hybridization solutions. The PM and MM probe intensities were corrected for the optical background before further analysis using the algorithm provided by MAS 5.0.[4]

Least-Squares Fits. The sensitivity coefficients of the SB model, $\sigma_k^P(B)$, were determined by means of multiple linear regression which minimizes the sum of weighted squared residuals between measured and calculated sensitivities,[23] SSQR $= \sum_{p=1}^{Ndata} \omega_p^{-2}(Y_p^P - Y_p^{P,SB})^2$. The sum runs over all considered probes N_{data}. The resulting system of linear equations was solved by means of single value decomposition (SVD[24]), which guarantees the solution that meets the symmetry condition (eq 9).

The weighting factor, ω_p^2, was estimated using the error model described in the Supporting Information (S2), $\omega_p^2 = \text{var}(\log(I_p)) = a + b/(I_p^P) + c/(I_p^P)^2$. It accounts for the increase of signal error at small intensities in a logarithmic scale. The constants a, b, and c consider the noise level of the binding equilibrium, of a probe-specific stochastic term, and of the optical background, respectively. They were estimated using a set of more than 3000 oligonucleotide probes present as replicates on each HG U133 chip.

Results

Binding Isotherms and Signal Intensities of Individual Probes. The spiked-in LS data set provides PM and MM intensities of 42 selected probe sets as a function of the concentration of specific target RNA in a constant background of nonspecific hybridization. The concentration dependence of the intensity of six selected probe pairs is shown in Figure 1. The courses are well described by eq 1 (compare lines and symbols, note the logarithmic scale). Accordingly, each curve is characterized by two model parameters, the affinity constant for specific binding, $K \equiv K_p^{P,S} \times$ pM and the effective affinity ratio, $r \equiv c_{RNA}^{NS} r_p^P r_p^{F,P} \approx c_{RNA}^{NS} K_p^{P,NS}/K_p^{P,S} \times 10^{-3}$ pM^{-1} which provides a measure of the intensity ratio due to nonspecific and specific hybridization at $c_{RNA}^S = 1$ pM. Typically, the mean

(22) Binder, H.; Kirsten, T.; Hofacker, I.; Stadler, P.; Loeffler, M. *J. Phys. Chem. B* **2004**, *108*, 18015.

(23) Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill: New York, 1969.

(24) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambrigde University Press: New York, 1989.
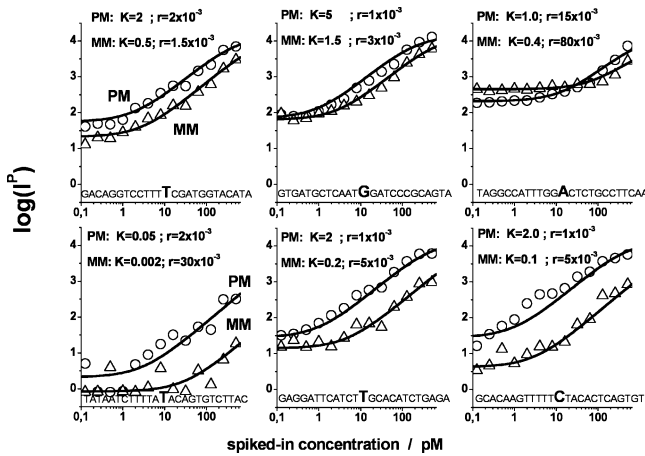
Figure 1. Binding isotherms of six selected PM/MM probe pairs (see circles and triangles, respectively) showing the log-intensity as a function of the target concentration. The data are taken from the LS data set. The curves are calculated using the binding model (eq 1) with $K = K_p^{P,S} \times$ pM and $r = c_{RNA}^{NS} r_p^P = c_{RNA}^{NS} K_p^{P,NS}/K_p^{P,S} \times 10^{-3}$ pM^{-1} (the parameter values are given within the Figure) and $\log(FN_p^F) = 4.15$. The sequence of the respective PM probe is given within each panel with an enlarged middle letter. Note the heterogeneous behavior of the different probes especially in the limit of small target concentrations.

affinity of nonspecific hybridization is two to 3 orders of magnitude smaller than the affinity of the probes for specific transcripts (see the data given in Figure 1). On the other hand, the binding constant for specific association of the PM exceeds that of the MM by a factor between about two and twenty. The PM intensity of all considered examples is therefore distinctly higher than that of the respective MM probe at high specific transcript concentrations.

The relation between the PM and MM intensities is however more heterogeneous in the limit of dominating nonspecific hybridization (i.e., at small spiked-in concentrations). This result indicates that the affinity of the PM probes for nonspecific transcripts is either higher, equal, or even smaller compared with that of the respective MM. Note also that the binding affinities vary by nearly 3 orders of magnitude between the different probes especially in the limit of small specific transcript concentrations. In addition, the PM curves are shifted by different degrees relative to the MM curves. This result indicates a puzzling relation between the affinities of the PM and MM probes due to the mismatched base pair in the middle of the probe sequence. The binding constant of each individual probe is directly related to the strengths of the base pairings in the respective DNA probe/RNA transcript duplex and thus it depends on the sequence of the probe. The consideration of the sequence given in Figure 1 provides however no simple explanation of the observed intensity courses.

In summary, the dependence of the signal intensity of PM and MM probes on the concentration of target RNA can be well approximated by binding isotherms of the Langmuir type which are characterized by the binding constants of specific and nonspecific hybridization for each individual probe. The whole ensemble of about 250,000 PM and MM probes on the HG U133 chip consequently requires the knowledge of nearly 10^6 affinity constants to predict their intensity as a function of the concentration of target RNA. The determination of this rather high number of constants by model fits of the binding isotherms appears hardly to realize because one needs spiked-in data for each probe.

Mean Binding Isotherms. The mean PM and MM probe intensities, which are log-averaged over the en-
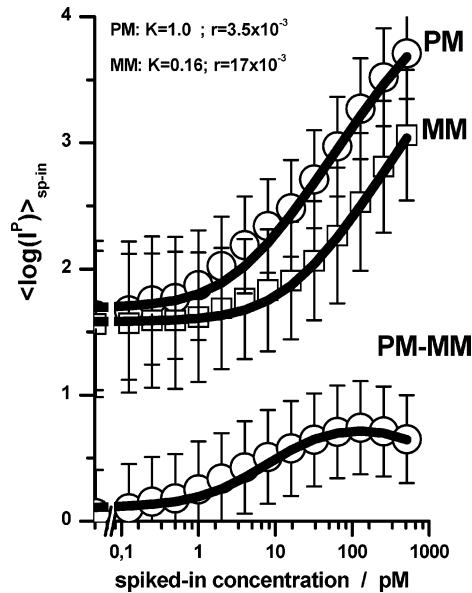

Figure 2. Mean binding isotherms averaged over all different 462 PM and MM probes of the LS data set (each probe intensity was considered in triplicate). The panel below shows the log-intensity difference, $\langle \log I^{PM-MM} \rangle_{sp-in} = \langle \log I^{PM} - \log I^{MM} \rangle_{sp-in}$. The error bars refer to the standard deviation of the individual probe intensities. The curves are calculated by means of eq 3 with $K = K_0^{P,S} \times$ pM and $r = c_{RNA}^{NS} r_0^P \times 10^{-3}$ pM^{-1} (see figure for parameter values) and $\log(F) = 4.15$.

semble of spiked-in genes are shown in Figure 2 as a function of the spiked-in concentration (symbols). These data illustrate the basic effect of the transcript concentration on the probe intensities. The PM and MM probes on the average possess similar intensities at small specific transcript concentrations, i.e., at dominating nonspecific hybridization. In the intermediate concentration range, both the PM and MM intensity almost linearly increases with increasing specific transcript concentration. The vertical shift between the PM and MM data reflects the mean affinity difference for specific hybridization between the PM and MM probes. Upon further increasing concentration, the experimental PM intensity data progressively deviate in the negative direction from the linear relationship owing to the onset of saturation. Note that the mean MM intensity values are considerably less affected by this effect.

The mean intensities are well described by eq 3 with Σ = sp-in (see lines in Figure 2). It turns out that the mean binding constant of the PM probes for target RNA exceeds that of the MM on the average by a factor of $K_0^{PM,S}/K_0^{MM,S} \approx 6-7$. The effective affinity of the PM for specific binding is by $2-3$ orders of magnitude stronger than that for nonspecific binding ($r_0^{PM} = 0.0035$). On the other hand, the mean affinity of PM and MM probes for nonspecific binding is equal in magnitude.

Sensitivity Profiles along the Probe Sequence. The nearly 250,000 PM and MM probe sensitivities per chip were analyzed in terms of the position dependent single base (SB) model (eq 8) in correspondence with recent studies.[20,21] This approach quantifies the individual, sequence-specific intensity of each probe as the deviation from the respective set-average. Accordingly, the formation of probe-target hybrid duplexes is described by four SB sensitivity parameters for each position of the 25meric probes. The set of positional dependent sensitivity coefficients providing the optimal fit of the sensitivity values of all PM and MM probes of the HG U133 chip are shown in Figure 3. The PM sensitivity profiles of base C and A change in a parabola-like fashion along the probe sequence
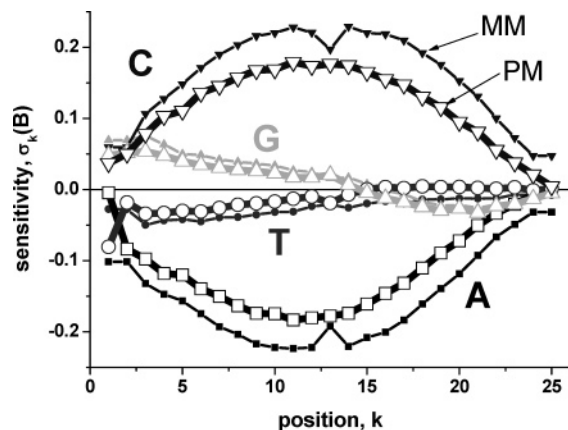
Figure 3. Single base sensitivity profiles of PM and MM oligo probes. The profile of each base (see figure) was obtained by the least-squares fit of eq 8 to the sensitivities of all 248,000 probes using all 42 HG U133 chips of the LS experiment.

being maximum and minimum in the center at $k = 13$, respectively. The substitution of an A by a C at position $k = 13$ is expected to enhance the probe sensitivity by the factor $\sim 10^{0.4} = 2.5$. Note that the intensity of a poly C probe is about 10^5 times higher than that of a 25meric poly A.

Contrarily, the sensitivity terms for G and T monotonically change along the sequence. Differences between the base specific sensitivities almost completely vanish at the free 5′ end of the probe at $k = 25$, whereas the sensitivity of G is considerably larger than that for T at the 3′ end, which is attached to the glass slide. Note that the bases G and T provide only tiny contributions to the positional dependent base sensitivity in the center of the sequence at $k = 13$. The base-specific sensitivity profiles are nearly equal for PM and MM probes except the small "dents" in the middle of the MM sequence for A and C and their slightly larger absolute values.

The position dependence of the sensitivity terms can be rationalized by a gradient of the base specific contribution to the free energy of base-pair interactions along the sequence. For example, the higher flexibility of the oligonucleotide chain near its free end is expected to reduce the base specificity owing to entropic effects. On the other hand, it should be taken into account that the positional dependent SB contributions are mean parameters, which are averaged over all individual DNA/RNA duplexes of one spot. Each microscopic state contributes to the SB sensitivity with a weight according to the probability of occurrence of the respective base pairing in the dimers. Consequently also "zippering effects", e.g., target/probe duplexes which look like a partly opened double-ended zipper,[25] and/or shorter probe lengths with less than 25 bases due to imperfect synthesis [26,27] potentially cause a gradient of sensitivities along the sequence because the probability of paired bases is expected to decrease in an asymmetrical fashion in the direction toward the 3′ and 5′ ends of the oligonucleotide probe.

Effect of Specific and Nonspecific Hybridization on the Sensitivity Profile. The LS experiment enables us to study the effect of the probe sequence on the sensitivity as a function of transcript concentration. Figure 4 shows the log intensities (panel above) and the respective sensitivities (panel below) of the PM and MM probes of
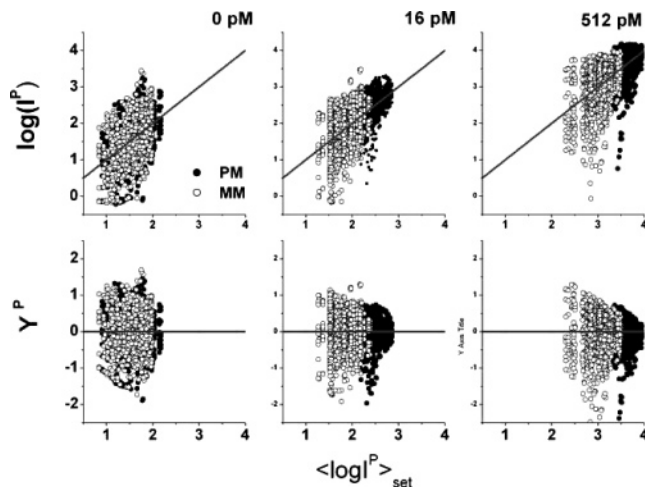
(25) Deutsch, J. M.; Liang, S.; Narayan, O. *arXiv: q-bio. BM/0406039 v1* **2004**.
(26) Jobs, M.; Fredriksson, S.; Brookes, A. J.; Ulf, L. *Anal. Chem.* **2002**, *74*, 199.
(27) McGall, G. H.; Barone, A. D.; Diggelman, M.; Fodor, S. P. A.; Gentalen, E.; Ngo, N. *J. Am. Chem. Soc,* **1997**, *119*, 5081.

Figure 4. Intensities (panel above) and sensitivities (eq 6, panel below) of the spiked-in probes at three different concentrations of specific target RNA (see figure) as a function of the set averaged probe intensity. The data clouds shift toward higher abscissa values with increasing target concentration.
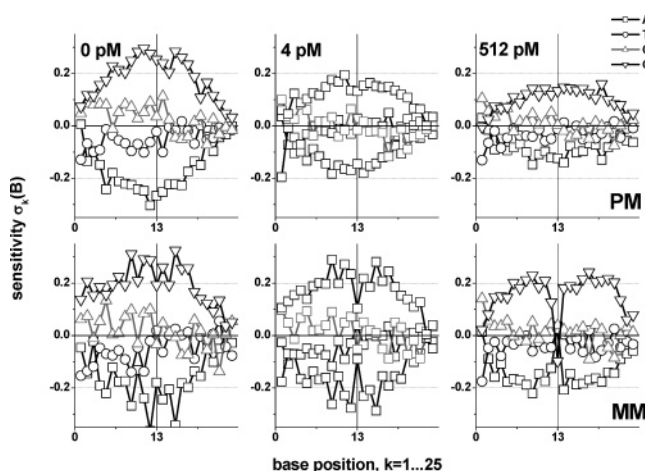

Figure 5. Single base related sensitivity profiles of PM and MM oligo probes referring to three concentrations of specific target RNA (see figure). The profile of each base (see figure) was obtained by the least-squares fit of eq 8 to the sensitivities of the 462 different spiked-in probes for each concentration (each condition was realized in triplicate). Note that the width between the maxima and minima of the profiles of C and A, respectively, decreases with increasing transcript concentration owing to a decreased variability of sequence specific affinity.

the spiked-in genes at selected concentrations as a function of the set averaged intensity, $\langle \log I_p^P \rangle_{set}$. With increasing spiked-in concentration, the data clouds shift in the direction of higher abscissa values. The progressive shift between the PM and MM values reflects the higher affinity of the PM probes for specific binding (see also Figure 2). Note that the intensity values increase with increasing concentration of specific transcripts, whereas the respective sensitivity is virtually independent of the amount of spiked-in transcripts.

The sensitivity data are fitted by means of the SB model for each concentration. Note that the spiked-in data set of 3(number of genes per concentration) × 14(number of concentrations) × 11(number of probes per set) = 462 probes enables the determination of the 100 positional dependent sensitivity coefficients, $\sigma_k^P(B)$ ($k = 1,...25$; B = A,T,G,C) for P = PM and MM probes. The respective sensitivity profiles (Figure 5) are distinctly more noisy owing to the relatively small number of used intensity data than the profiles which have been obtained by the
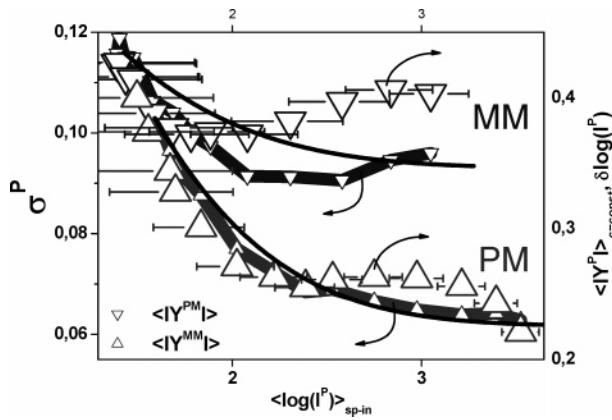
Figure 6. Variability of the sensitivity as a function of set averaged intensity of the spiked-in probes at different concentrations of specific transcripts. The thick lines (left ordinate) are the mean absolute SB sensitivity terms, σ^P (eq 11) whereas the symbols denote mean absolute sensitivity values, $\langle|Y^P|\rangle$ (right ordinate). The thin lines are calculated by means of the variability model, eq 16 ($\delta \log(I^P)$, right ordinate), using the mean binding isotherms (Figure 2) and $\delta \ln K^{P,S}/r_\Delta^P = 0.52/1.70$ and 0.80/1.25 for P = PM and MM, respectively. The increase of the variability with decreasing abscissa values can be explained by a higher variability of nonspecific binding as indicated by $r_\Delta^P > 1$.

fit of all, nearly 250,000 probes pairs per chip (compare with Figure 3).

The comparison of the sensitivity profiles reveals that their distribution width about the abscissa (see, e.g., the difference $\sigma_{13}^P(C) - \sigma_{13}^P(A)$ in the middle of the sequence) progressively decreases with increasing transcript concentration. We calculated the base and position averaged absolute value of the sensitivity terms, σ^P (eq 11), to quantify the observed tendency (see Figure 6, thick lines, left ordinate). In addition, we determined the mean absolute sensitivity value, $\langle|Y^P|\rangle_{c=\text{const}}$, for each concentration (see symbols in Figure 6), which characterizes the variability of the probe intensities with respect to their set average. Note that σ^P specifies the variability of the probe sensitivity due to the heterogeneity of the sequence in contrast to $\langle|Y^P|\rangle_{c=\text{const}}$, which includes also sequence-independent contributions. The parallel change of σ^P and $\langle|Y^P|\rangle_{c=\text{const}}$ indicates that the increase of variability with decreasing transcript concentration is related to the sequence and thus to changes of the effective affinity of target/probe duplex formation.

The sensitivity is directly related to the variability of the probe intensity in a logarithmic scale, $\delta \log I^P$ (see eq 6). Differentiation of eq 1 at $c_{\text{RNA}}^S = \text{const}$ and $F = \text{const}$ and assuming $S^P = 1$ for sake of simplicity provides $\delta \log I^P$ as a function of the variability of the effective binding constant of specific and nonspecific transcripts in a logarithmic scale, $\delta \ln K_p^{P,h} \approx \delta K_p^{P,h}/K_0^{P,h}$, i.e.

$$|Y^P| \propto \delta \log I^P \approx \frac{\sqrt{a + (\delta I^P|_{c=\text{const}})^2}}{\ln 10 \langle I^P \rangle}$$

with

$$\delta I^P|_{c=\text{const}} = F K_0^{P,S}(c_{\text{RNA}}^S + c_{\text{RNA}}^{NS} r_0^P r_\Delta^P)\delta \ln K_p^{P,S},$$

$$F = F_0\langle N_p^{F,S}\rangle, \quad \text{and} \quad r_\Delta^P = \frac{\delta \ln K_p^{P,NS}}{\delta \ln K_p^{P,S}} \quad (16)$$

Note that the form of eq 16 is compatible with the error model used for the weighting factor of the least-squares

fits with $b = 0$ and $c = \delta I^2$ (see above). The additive term "a" refers to fluctuations of the optical background, of the concentration and composition of the RNA as well as of the chip-specific factor. The thinner lines in Figure 6 are calculated by means of eq 16. Their courses reasonably agree with the experimental data.

The logarithmic scale of the binding constant used in eq 16 is directly related to the free energy of binding (see eq 4). Hence, eq 16 is justified if one assumes variations of the probe intensity, which linearly scale with the free energy of duplex formation. The ratio r_Δ^P specifies the variability of the binding affinity of nonspecific transcripts relative to that of specific ones. Note that $K_p^{P,NS}$ represents an effective binding constant referring to a cocktail of RNA fragments which bind nonspecifically to the probes in contrast to $K_p^{P,S}$, which is the binding affinity of the single target sequence. It is therefore reasonable to assume $\delta \ln K_p^{P,NS} > \delta \ln K_p^{P,S}$, i.e., a higher variability of the affinity for nonspecific transcripts due to their more heterogeneous base composition. Note that the error model considers only "stochastic" effects in replicated measurements, whereas the variability data shown in Figure 6 (and eq 16) in addition include systematic contributions due to variations of the affinity between probes of different sequences.

We conclude that the inflation of the variability of the sensitivity (and the probe intensity) at small concentrations of specific transcripts (and at small set-averaged intensities) is partially caused by a higher variability of the binding affinity of nonspecific transcripts compared with that of specific ones. The higher variability of the sensitivity of the MM in the asymptotic range at higher abscissa values reflects the higher relative contribution of variations of the binding constant, $\delta K_p^{MM,S}/K_0^{MM,S} > \delta K_p^{PM,S}/K_0^{PM,S}$.

Sensitivity of Matched and Mismatched Base Pairings. To compare the position-dependent sensitivity profiles at different transcript concentrations, we separately plot their normalized values, $\sigma_k^P(B)_{\text{rel}} = \sigma_k^P(B)/\sigma^P$, in Figures 7 and 8 for P = PM and MM, respectively. The PM profiles of each base are virtually invariant with changing concentration of specific transcripts. Hence, nonspecific and specific hybridization can be well described by almost the same set of relative sensitivity terms in a first-order approximation. This result is confirmed by the observation that the sensitivity profiles of the reduced ensemble of spiked-in probes scatter about the respective $\sigma_k^P(B)_{\text{rel}}$ profile obtained from the full ensemble of all probes of the chip (see circles in Figure 7).

For the MM profiles, the results dramatically change at position $k = 13$ of the probe sequence, which refers to the mismatched self-complementary pairing with the target RNA sequence. The absolute value of the SB sensitivity contributions of the middle bases A and C progressively decreases with increasing concentration of specific transcripts (see Figure 9). Their specific contribution to the probe sensitivity almost completely vanishes at spiked-in concentrations greater than 128 pM. Note that the bases T and G provide only tiny values of the SB sensitivity terms at position $k = 13$ at all concentrations. Hence, the sensitivity of the MM probes is virtually invariant with respect to the mismatched base in the middle of the sequence if specific transcripts dominate hybridization. In other words, the middle base provides essentially no base-specific contribution to the stability of the duplex. On the other hand, the nearly linear relation between the MM probe intensity and the spiked-in concentration strongly indicates that the target RNA "specifically" binds to the MM probes (see Figure 2). This
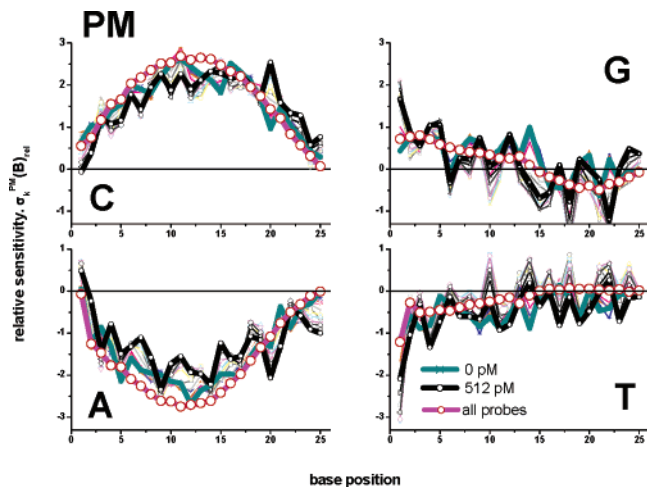
Figure 7. Single base related sensitivity profiles of PM oligo probes in relative units, $\sigma_k^P(B)_{rel} = \sigma_k^P(B)/\sigma^P$ referring to all 14 concentrations of specific target RNA of the LS experiment. The profiles of each base (see Figure) were obtained by the least-squares fit of eq 8 to the sensitivities of the 462 different spiked-in probes at each concentration. The circles refer to the profiles which are obtained using all PM probes of the chip (see also Figure 3). The profile of each base is virtually not affected by the concentration of specific transcripts.
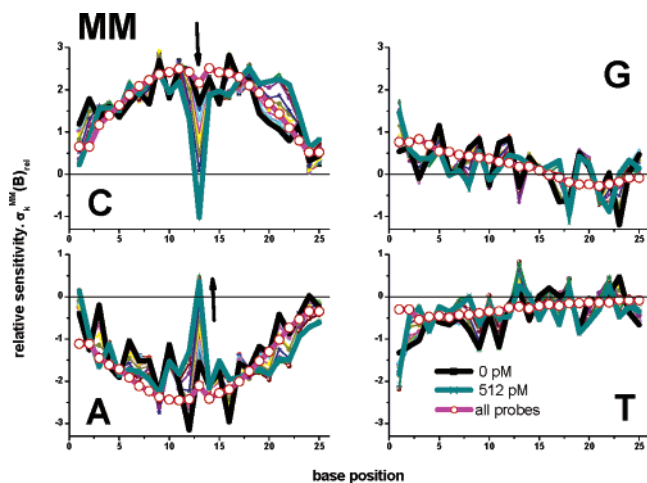

Figure 8. Single base related sensitivity profiles of MM oligo probes in relative units, $\sigma_k^P(B)_{rel} = \sigma_k^P(B)/\sigma^P$ referring to all 14 concentrations of specific target RNA of the LS experiment. See legend of Figure 7 for further details. The profile of each base is virtually not affected by the specific transcript concentration except the sensitivity term of the middle bases A and C. Their absolute value progressively decreases with increasing concentration of specific transcripts (see arrows).

result lets us conclude that specific binding to the MM probes is mainly driven by the remaining bases at positions $k = 1, ..., 12$ and $14, ..., 25$, which enable duplex formation via Watson−Crick base pairings.

Discussion

We studied the probe intensities of Affymetrix Gene-Chips as a function of the concentration of specific transcripts, the sequence of which completely matches the respective PM probe sequence by complementary bases. Specific hybridization is typically overlaid by nonspecific hybridization. Nonspecific RNA transcripts only partially match the probe sequence by WC pairings.

The concentration dependence of the signal intensity of each probe can be well described by a simple two-species Langmuir hybridization isotherm, which considers the
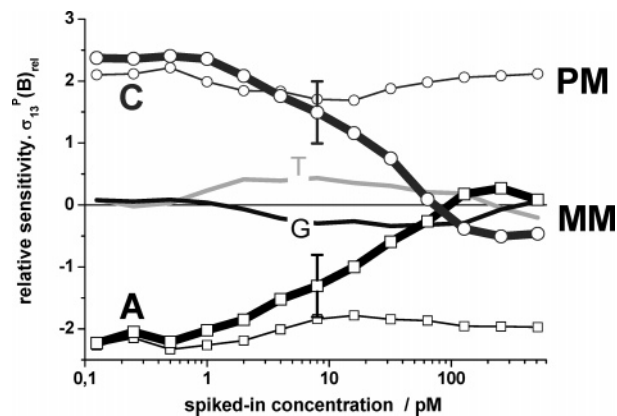

Figure 9. Single base related sensitivity terms of the middle base of PM (thin lines) and MM (thick lines) probes as a function of specific transcript concentration. The curves for T and G are almost identical for PM and MM probes. In contrast, the middle bases C and A progressively loose their sensitivity with increasing concentration of specific transcripts (see Figure for assignments).

binding equilibria between free and bound species of specific and nonspecific transcripts (see eq 1). In our approach, all free RNA fragments compete for duplex formation with the binding sites provided by the oligonucleotide probes. It turns out that the binding of nonspecific transcripts to MM probes is on the average characterized by a similar mean binding constant when compared with that of the PM probes, $K_0^{PM,NS} \approx K_0^{MM,NS}$. Contrarily, the affinity of the MM for specific transcripts is on the average nearly 1 order of magnitude smaller that that of the PM, $K_0^{PM,S} > K_0^{MM,S}$. The relations between the binding affinities can be summarized as PM(specific) > MM(specific) ≫ PM(nonspecific) ≈ MM(nonspecific).

The deviation of the intensity of an individual probe from its mean value over an appropriately chosen ensemble of probes in the logarithmic scale defines its sensitivity. It can be described as the sum of positional and base dependent terms, $\sigma_k^P(B)$ (see eq 8), in accordance with previous models.[20,21,28,29] Our results show that the PM-sensitivity profile is virtually independent of the concentration of target RNA, c_{RNA}^S. Hence, nonspecific hybridization dominating at small c_{RNA}^S values and specific hybridization dominating at high c_{RNA}^S values give rise to virtual identical profiles of the PM-sensitivity terms. This result surprises if one considers the large difference between the mean binding constant for specific and nonspecific hybridization of more than 2 orders of magnitude. Also the MM profiles closely resemble that of the PM for all sequence positions except the middle base. It turns out that the middle position of the MM only weakly contributes to its sensitivity.

The sensitivity terms, $\sigma_k^{P,h}(B)$, decompose into contributions due to the binding affinity, $\Delta\epsilon_k^{P,h}(B)$, and fluorescence emission, $\Delta\varphi_k^{P,h}(B)$, according to eq 14. The fluorescence provides only a relatively small contribution of $|\Delta\varphi_k^{P,h}(B)| \leq 0.04$ to the SB sensitivity terms at least in the middle of the sequence ($|\sigma_{13}^{P,h}(B)| < 0.15$ for B = C,A; see Figure 3). In other words, the observed probe sensitivity mainly reflects the sequence specific affinity for duplex formation, i.e., the propensity of the probe to bind

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

(29) Zhang, L.; Miles, M. F.; Aldape, K. D. *Nat. Biotechnol.* **2003**, *21*, 818.
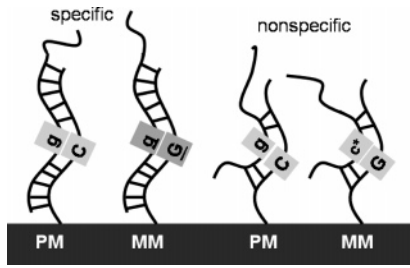
Figure 10. Base pairings in the middle of duplexes between DNA probes and RNA fragments. The nonspecific (NS) duplexes are stabilized by a smaller number of WC pairings compared with the specific (S) duplexes. The middle base of the MM forms a SC pairing upon specific hybridization. Note the reversal of the WC pair in the nonspecific duplexes of the PM and MM probes.

RNA fragments from the hybridization solution. Hence, the sensitivity terms can be interpreted to a good approximation by the respective incremental contributions to the interaction free energy, i.e., $\Delta\epsilon_{13}^{P,h}(B) \approx \sigma_{13}^{P,h}(B)$. In the following we discuss the obtained results using this approximation.

Base Pair Interactions in Specific Duplexes. The PM probe and the RNA target match each other via complementary Watson−Crick (WC) base pairs. The respective positional dependent free energy terms (see eq 5) consequently characterize the binding strength of WC pairings in the specific duplexes, i.e., $\epsilon_{0,k}^{PM,S} \approx \epsilon_{0,k}^{WC}$ and $\Delta\epsilon_k^{PM,S}(\xi_{p,k}^{PM}) \approx \Delta\epsilon_k^{WC}(\xi_{p,k}^{PM})$. Also the MM probes bind the specific transcripts via WC pairs except the middle base at position $k = 13$, which faces "itself" in a self-complementary (SC) pair (see Figure 10 for illustration). One can therefore expect that the positional dependent free energy terms of the PM and MM probes are nearly identical for k≠13 but different for $k = 13$. The binding strength of the middle base of the MM consequently refers to the SC pairing, i.e., $\epsilon_{0,13}^{MM,S} \approx \epsilon_{0,13}^{SC}$ and $\Delta\epsilon_{13}^{MM,S}(\xi_{p,13}^{MM}) \approx \Delta\epsilon_{13}^{SC}(\xi_{p,13}^{MM})$.

The fit of the SB model to the sensitivity data referring to large spiked-in concentrations provides estimates of the incremental free energy of duplex stabilization (see eq 13). We obtained similar values for PM and MM outside of the middle base as expected, $\sigma_k^{MM,S}(B)|_{k \neq 13} \approx \sigma_k^{PM,S}(B)|_{k \neq 13} \approx \Delta\epsilon_k^{WC}(B)|_{k \neq 13}$ (see Figures 5, 7, and 8). The relatively small contribution of the middle base of the MM, $|\sigma_{13}^{MM,S}(B)| \approx |\Delta\epsilon_{13}^{SC}(B)| \leq 0.05$, indicates that the SC pairings on the average have virtually lost their sensitivity. This result is compatible with $|\Delta\epsilon_{13}^{SC}(B)| \ll |\Delta\epsilon_{13}^{WC}(B)|$ for B = A and C and with $|\Delta\epsilon_{13}^{SC}(B)| \approx |\Delta\epsilon_{13}^{WC}(B)| \approx 0$ for B = G anf T. One obtains for the binding constant of the MM (see eqs 4 and 5)

$$-\log K_p^{MM,S} \approx \sum_{k \neq 13}^{N_b} \epsilon_k^{WC}(\xi_{p,k}^{PM}) + \epsilon_{13}^{SC}(\xi_{p,13}^{MM}) = -\log K_p^{PM,S} - \epsilon_{13}^{WC-SC}(\xi_{p,13}^{PM}) \quad (17)$$

Equation 17 shows that the log-difference of the binding constants of the PM and MM probes roughly estimates the free energy difference between the respective WC and SC pairs in the middle of the probe sequence, $\epsilon_{13}^{WC-SC} = -(\log K_p^{PM,S} - \log K_p^{MM,S})$.

After decomposition of the free energy term according to eq 5 one obtains for the mean difference $\epsilon_{0,13}^{WC-SC} = -(\log K_0^{PM,S} - \log K_0^{MM,S})$ and $\Delta\epsilon_{13}^{WC-SC}(B) \approx \Delta\epsilon_{13}^{WC}(B) \approx$

$-\sigma_{13}^{PM,S}(B)$ for the base specific increment between the free energy of a WC and SC pairing in RNA-target/DNA-probe duplexes on the microarray. The fit of the total mean intensities of all spiked-in probes (Figure 2) provides $-\epsilon_{0,13}^{WC-SC} \approx 0.85 \pm 0.04$ (see also Table 1). This value well agrees with the mean reduced Gibbs free energy of a WC pair in DNA/RNA oligonucleotide duplexes in solution, $-\epsilon_{sol}^{WC} = 0.75-1.03$, which was estimated using literature data of the respective nearest neighbor free energy terms[5,30] (see footnote in Table 1). The agreement between the microarray and solution data can be rationalized if the mean free energy contribution of the SC pairs to duplex stability is on the average much weaker than the contribution of the WC pairs, $|\epsilon_{0,13}^{SC}| \ll |\epsilon_{0,13}^{WC}|$ and if the contribution of the WC pairs outside of the middle base is similar in specific and nonspecific duplexes.

Taking together we found that the SC pairs, on the average, only weakly contribute to the stability of probe-target duplexes. This result gives rise to $\epsilon_{13}^{WC-SC} \approx \epsilon_{13}^{WC}$ or equivalently $|\epsilon_{13}^{SC}| \ll |\epsilon_{13}^{WC}|$. In other words, the free energy difference between the WC and SC pairs roughly reflects the strength of the respective WC pairing in the duplexes.

Base Pair Interactions in Nonspecific Duplexes. By nonspecific binding, we imply the ensemble of lower affinity mismatched duplexes involving sequences other than the intended target. The fit of the SB model to the spiked-in data at small concentrations of specific transcripts provides positional dependent sensitivity terms, which to a good approximation agree with the respective free energy contributions to the stability of nonspecific duplexes (see above). For the PM and MM probes, we found very similar values, which in turn also agree with the respective PM data for the specific transcripts. The latter values have been assigned to WC pairings between the probe and the RNA fragments. This agreement confirms the expectation that the nonspecific duplexes are mainly stabilized by WC pairings. Consequently, also the middle base in the nonspecific dimers of the MM usually forms a WC pair in contrast to the respective specific duplexes where the middle base forms a SC pair (see Figure 10 for illustration). This assignment of base-pairings in nonspecific duplexes appears plausible because the "cocktail" of nonspecific RNA fragments in the hybridization solution usually contains enough sequences, which enable WC pairings with the central base of the MM and the complementary central base of the PM as well. Note that the middle base of the MM is per definition a "mismatched SC" pair only with respect to the respective target sequence that specifically hybridizes the probe but not with respect to nonspecific RNA fragments.

One obtains for the binding constant of the MM after consideration of the respective relations between the free energy parameters and of eqs 4 and 5

$$-\log K_p^{MM,NS} \approx \sum_{k \neq 13}^{N_b} \epsilon_k^{WC}(\xi_{p,k}^{PM}) + \epsilon_{13}^{WC}(\xi_{p,13}^{MM}) = -\log K_p^{PM,NS} - \epsilon_{13}^{WC-WC}(\xi_{p,13}^{PM}) \quad (18)$$

Equation 18 shows that the log difference between the binding constants of the PM and MM probes estimates the Gibbs free energy difference between their complementary WC pairs in the middle of the probe sequence. After decomposition of the energetic term according to eq 5, one obtains the base independent mean

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

Table 1. Middle-Base Related Free Energy and Fluorescence Contributions of Specific (S) and Nonspecific (NS) Hybridization on Microarrays and of DNA/RNA Duplexes in Solution (sol)[a]

	probe level	base pair level	PM middle base				mean[c]
			A	T	G	C	
NS	PM[b]	WC	A−u*	T−a	G−c*	C−g	
	MM[b]	WC	T−a	A−u*	C−g	G−c*	
	$\Delta\epsilon_{13}^{PM,NS}$	$\Delta\epsilon_{13}^{WC}$	**−0.20**	**+0.05**	**0.0**	**+0.25**	
	$\Delta\epsilon_{13}^{MM,NS}$	$\Delta\epsilon_{13}^{WC}$					
	$\Delta\epsilon_{13}^{PM-MM,NS}$	$\Delta\epsilon_{13}^{WC-WC}$	−0.25	+0.25	−0.25	+0.25	
	$\epsilon_{13}^{PM-MM,NS}$	ϵ_{13}^{WC-WC}	−0.15	+0.35	−0.15	+0.35	**0.05 ± 0.04**
	$\Delta\varphi_{13}^{PM,NS}$	$\Delta\varphi_{13}^{WC}$	**+0.04**	**−0.04**	**+0.04**	**−0.04**	
	$\Delta\varphi_{13}^{MM,NS}$	$\Delta\varphi_{13}^{WC}$					
	$\Delta\varphi_{13}^{PM-MM,NS}$	$\Delta\varphi_{13}^{WC-WC}$	+0.08	−0.08	+0.08	−0.08	
S	PM[b]	WC	A−u*	T−a	G−c*	C−g	
	MM[b]	SC	*T−u**	*A−a*	*C−c**	*G−g*	
	$\Delta\epsilon_{13}^{PM,S}$	$\Delta\epsilon_{13}^{WC}$					
	$\Delta\epsilon_{13}^{MM,S}$	$\Delta\epsilon_{13}^{SC}$	**+0.05**	**+0.05**	**−0.05**	**−0.05**	
	$\Delta\epsilon_{13}^{PM-MM,S}$	$\Delta\epsilon_{13}^{WC-SC}$	−0.25	0.0	0.05	+0.30	
	$\epsilon_{13}^{PM-MM,S}$	ϵ_{13}^{WC}	0.55	0.80	0.85	1.10	**0.85 ± 0.04**
	$\Delta\varphi_{13}^{PM,S}$	$\Delta\varphi_{13}^{WC}$					
	$\Delta\varphi_{13}^{MM,S}$	$\Delta\varphi_{13}^{SC}$	−0.04	+0.04	−0.04	+0.04	
	$\Delta\varphi_{13}^{PM-MM,S}$	$\Delta\varphi_{13}^{WC-SC}$	0	0	0	0	
sol		WC	A−u	T−a	G−c	C−g	
		$\Delta\epsilon_{sol}^{WC-WC\,e}$	−0.14/−0.23	0.14/0.23	−0.21/−0.18	0.21/0.18	
		$\epsilon_{sol}^{WC\,e}$	0.52/0.62	0.66/0.85	0.96/1.24	1.17/1.42	0.75/1.03
	$f_{13}^{sp-in}(B)$ [f]		0.22	0.35	0.23	0.19	

[a] The "probe-level" data are deduced from the combined analysis of probe intensities in terms of the SB and the intensity/binding models (see Figures 2 and 5). The bold type indicates independent parameters used in the model. The "base pair-level" data provide an interpretation of the "probe-level" data in terms of Watson−Crick (WC) and self-complementary (SC) base pairs (see text). All free energy terms are scaled with $\sim\{-(RT\ln10)^{-1}\}$. The resolution of the chosen energy parameters was arbitrarily set to ±0.05. See also Table S in the Supporting Information for definitions and relations between the parameters. [b] Pairings of the middle base in duplexes of oligo probes with specific and nonspecific RNA transcripts [c] "mean" values are averages over the base-specific values: $A_{mean} \equiv \langle A \rangle = (1/4)\sum_{B=A,T,G,C}A(B)$, i.e., $<\epsilon_{13}^{WC-WC}> = \epsilon_{0,13}^{WC-WC}$ and $<\epsilon_{13}^{WC-SC}> \approx \epsilon_{0,13}^{WC}$ for the free energy contributions in the limiting cases of nonspecific and specific hybridization, respectively. [e] Base-specific reduced free energy contribution of DNA/RNA duplex stability in solution. The mean value for each base, $\epsilon_{sol}^{WC}(B) = (-8RT\ln 10)^{-1}\sum_{X=A,T,G,C}(G(B,X) + G(X,B))$ with $T = 210$ K, was calculated using the respective nearest-neighbor terms taken from references 5 and 30. The difference is $\Delta\epsilon_{sol}^{WC-WC} = \epsilon_{sol}^{WC}(B) - \epsilon_{sol}^{WC}(B^c)$. The ϵ_{sol}^{WC} and $\Delta\epsilon_{sol}^{WC-WC}$ data should be compared with ϵ_{13}^{WC} and $\Delta\epsilon_{13}^{WC-WC}$, respectively (see text). [f] Fraction of PM probes with middle letter B within the ensemble of 462 spiked-in probes.

difference $-\epsilon_{0,13}^{WC-WC} = \log K_0^{PM,NS} - \log K_0^{MM,NS}$ and the difference of the base-dependent increment $\Delta\epsilon_{13}^{WC-WC}(B) = -\Delta\epsilon_{13}^{WC-WC}(B^c) \approx \sigma_{13}^{PM,NS}(B) - \sigma_{13}^{PM,NS}(B^c)$ (B^c denotes the complementary base of B, see also Table 1).

With eqs 17 and 18, one obtains for the ratio of the binding constants of nonspecific and specific hybridization of each PM/MM probe pair

$$\log r_p^{PM} = \log r_0^{PM} = \log K_0^{PM,NS} - \log K_0^{PM,S}$$

and

$$\log r_p^{MM} = \log r_0^{MM} - (\Delta\epsilon_{13}^{WC}(B) - \Delta\epsilon_{13}^{SC}(B)) \quad (19)$$

with $\log r_0^{MM} = \log K_0^{MM,NS} - \log K_0^{MM,S} = \log r_0^{PM} - \epsilon_{0,13}^{WC-SC}$. Note that the ratio of the PM, r_p^{PM}, is independent of the middle base because it forms WC pairings in specific and nonspecific duplexes as well. Consequently, the base-specific effect cancels out. Contrarily, the ratio for the MM depends on the middle base. Here the central WC pair in the nonspecific duplexes is replaced by a SC pairing in the specific dimers.

Rearrangement of eq 19 provides

$$\log K_p^{P,NS} \approx \log K_p^{P,S} + \log r_0^{PM} + \begin{cases} 0 & \text{for P = PM} \\ -\Delta\epsilon_{13}^{WC-SC}(B) & \text{for P = MM} \end{cases} \quad (20)$$

The second term is either a constant (P = PM) or it depends only on the middle base (MM). It consequently does not affect the obtained sensitivity profiles at all positions (PM) or at all positions except the middle base (MM) because the symmetry condition (eq 9) cancels out constant contributions. This result explains the very similar base and positional dependent SB sensitivity profiles of nonspecifically and specifically hybridized probes.

The stability of nonspecific probe/target duplexes of the PM and MM is governed by WC pairings according to this interpretation. Consequently, PM and MM probes with the same middle base are expected to hybridize with nonspecific transcripts on the average almost equally. For randomly distributed middle bases, one expects a vanishing mean difference, $\epsilon_{0,13}^{WC-WC}|_{random} = 0$. The fit of the total mean intensities of all spiked-in probes however provides $-\epsilon_{0,13}^{WC-WC} \approx 0.05 \pm 0.04$ (Figure 2, Table 1). This bias can be, at least partially, explained by a nonrandom distribu-

tion of middle bases for the probes on the chip according to

$$\log K_0^{P,NS} = \langle \log K_B^{P,NS} \rangle_{chip} =$$
$$\langle \log K_B^{P,NS} \rangle_{random} - \sum_{B=A,T,G,C} f_{13}^{chip}(B) \Delta\epsilon_{13}^{WC}(B)$$

where $f_{13}^{chip}(B)$ denotes the fraction of middle base B in all probes of the chip (see Table 1).

With $\langle \log K_B^{PM,NS} \rangle_{random} = \langle \log K_B^{MM,NS} \rangle_{random}$, one obtains after some rearrangements

$$-\epsilon_{0,13}^{WC-WC} = \log K_0^{PM,NS} - \log K_0^{MM,NS} =$$
$$- \sum_{B=T,C} \Delta f_{13}^{chip}(B) \, \Delta\epsilon_{13}^{WC-WC}(B)$$

with

$$\Delta f_{13}^{chip}(B) = (f_{13}^{chip}(B) - f_{13}^{chip}(B^c)) \quad \text{and}$$
$$\Delta\epsilon_{13}^{WC-WC}(B) = \Delta\epsilon_{13}^{WC}(B) - \Delta\epsilon_{13}^{WC}(B^c) \quad (21)$$

With the respective $f_{13}^{chip}(B)$ data (see Table 1) one obtains $-\epsilon_{0,13}^{WC-WC} \approx 0.04$ in agreement with the observed value. Note that in addition, also a nonrandom base distribution within the nonspecific RNA fragments in the hybridization solution can introduce an asymmetry between the respective PM and MM intensities.

Middle Base Averaged Hybridization Isotherms. It is well established that the middle base systematically affects the relation between the PM and MM probe intensities,[6] which, in addition, changes as a function of specific transcript concentration.[31] This characteristic behavior can be understood in the light of the molecular hybridization theory presented in the preceding sections.

Particularly, the SB model predicts that the relation between the intensities of the PM and MM probes depends in a characteristic fashion on the middle base (see eqs 17 and 18). To further check this result, we calculated mean values over all spiked-in probes with a common middle base B as a function of transcript concentration. Equation 1 predicts for the middle base averaged log intensity the isotherm

$$\log I_B^P \equiv \langle \log I^P \rangle_B \approx \log F_0 + \log N_B^{F,S} + \log K_B^{P,S} +$$
$$\log[c_{RNA}^S + c_{RNA}^{NS} r_B^P r_B^{F,P}] - \log S_B^P$$

with

$$\log S_B^P \equiv \langle \log S^P \rangle_B \approx \log(1 + K_B^{P,S}[c_{RNA}^S + c_{RNA}^{NS} r_B^P]) \quad (22)$$

The effective binding constants in eq 22 are averages over all probes with the respective middle base B

$$\log K_B^{P,h} \equiv \langle \log K_p^{P,h} \rangle_B \approx \log K_0^{P,h} - \Delta\epsilon_{13}^{P,h}(B)$$
$$\text{with h = S and NS}$$

$$\log r_B^P \equiv \langle \log r_p^P \rangle_B \approx \log r_0^P - (\Delta\epsilon_{13}^{P,NS}(B) - \Delta\epsilon_{13}^{P,S}(B)) \approx$$
$$\begin{cases} \log r_0^{PM} & \text{for P = PM} \\ \log r_0^{MM} - \Delta\epsilon_{13}^{WC}(B) & \text{for P = MM} \end{cases}$$

with

$$B \equiv \xi_{p,13}^P = A, T, G, C \quad (23)$$

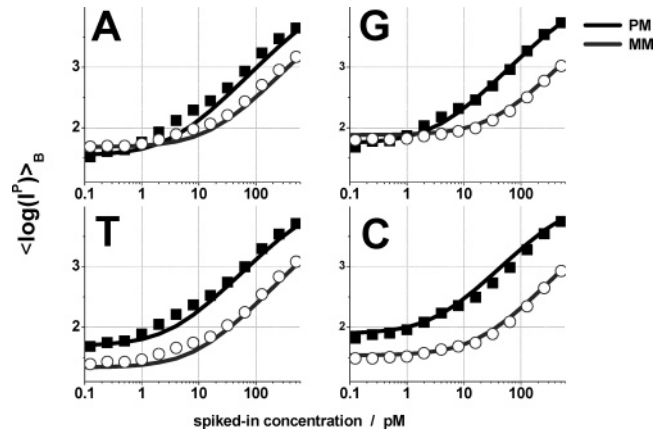Here we assume that the averaging ($\langle...\rangle_B$) cancels out all

Figure 11. Mean binding isotherms of probe pairs with middle base B = A, T, G, and C of the PM probe (see figure for assignment, squares and circles refer to PM and MM probes, respectively). The data represent averages over all spiked-in probes with the respective middle base. The curves are calculated using the binding model (eq 22) with the model parameters listed in Table 1, a fluorescence contribution of $\Delta^F = 0.04$ and $\log(F_0 N_B^F) = 4.15$.

positional dependent terms with $k \neq 13$, e.g. $\langle \sum_{k=1}^{N_b} \epsilon_k^{P,h}(\xi_{p,k}^P) \rangle_B \approx \epsilon_{13}^{P,h}(B)$. The mean effect of labeling is characterized by the equations

$$\log N_B^{F,S} \equiv \langle \log N_p^{F,S} \rangle_B \approx \log N_0^F + \Delta\varphi_{13}^{PM,S}(B)$$

and

$$\log r_B^{P,F} \equiv \langle \log r_p^{P,F} \rangle_B \approx (\Delta\varphi_{13}^{P,NS}(B) - \Delta\varphi_{13}^{P,S}(B)) \approx$$
$$\begin{cases} 0 & \text{for P = PM} \\ 2\Delta\varphi_{13}^{WC}(B) & \text{for P = MM} \end{cases} \quad (24)$$

Figure 11 compares the measured with the calculated middle-base averaged mean intensity values of the PM and MM probes as a function of the concentration of specific transcripts. The theoretical curves are calculated according to eq 22 using the mean affinity constants, $K_0^{P,h}$ and r_0^P, which were previously determined for the total average of the probe intensities (see Figure 2 and Table 1). The middle-base specific model parameters, $\Delta\epsilon_{13}^{P,h}(B)$ and $\sigma_{13}^{P,h}(B)$ are taken from the fits of the SB model (see Figures 3, 5, 7, and 8 and Table 1) in accordance with the results presented above (see legend of Figure 11). Hence, the curves are "synthesized" using the parameter estimates from the independent approaches of the SB model and the mean intensity fits and thus they represent rather a prediction than a fit. The agreement between calculated and measured isotherms confirms the consistency of the chosen formalism and illustrates the behavior of PM and MM intensities as a function of concentration.

The middle base averaged mean PM intensity exceeds the respective MM intensity over the whole concentration range of specific transcripts for pyrimidine middle bases C and T of the PM probes. Contrarily, for purine middle bases B = G and A, the PM and MM intensity courses intersect each other with $\log I_B^{MM} > \log I_B^{PM}$ in the limit of nonspecific hybridization and with the reverse relation, $\log I_B^{MM} < \log I_B^{PM}$, at higher concentrations of specific transcripts.

The middle base specific log-intensity differences, $\log I_B^{PM-MM} \equiv \log I_B^{PM} - \log I_{Bc}^{MM}$, changes from a characteristic

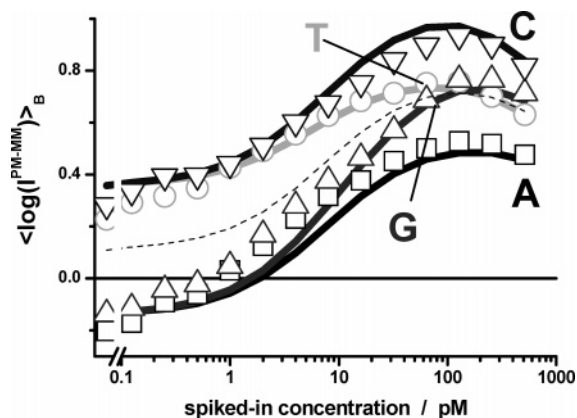(31) Binder, H.; Preibisch, S. *Biophys. J.* **2005**, *89*, 337.

Figure 12. Log-intensity difference of the middle base related mean binding isotherms shown in Figure 11 (see figure for assignments). The curves are calculated by means of the binding model (see legend of Figure 11 for details). The dotted curve is the mean difference averaged over all middle bases.

duplet-like pattern into a pattern of triplet-like symmetry at small and high concentrations of specific transcripts, respectively (see Figure 12). Note that the model curves excellently reproduce the features of the measured middle-base averaged intensities (compare lines and symbols in Figure 12). The duplet-like symmetry at dominating nonspecific hybridization reflects a pyrimidine−purine asymmetry of the interaction strength in WC pairings of DNA/RNA hetero duplexes. Particularly, the WC pairing in the middle of the sequence reverses between the PM and MM of one probe pair, i.e., B−b^c for PM transforms into B^c-b for the respective MM probe (e.g., G−$c^* \rightarrow$ C−g, see Figure 10 for illustration). The reversal of the base pairing is accompanied by the reversal of sign of the respective free energy difference if one compares pairs with complementary B and B^c in the middle of the PM sequence, i.e., $\Delta\epsilon_{13}^{WC-WC}(B^c) \approx -\Delta\epsilon_{13}^{WC-WC}(B)$. This symmetrical relation splits the respective affinities into two symmetric branches relative to the overall mean, namely for the purines G and A on the one hand and for the pyrimidines C and T on the other hand.

Our analysis shows that the mismatched SC pairs on the average only weakly contribute to the affinity between the MM probe and the respective RNA target. The different base pairings, namely the WC pair (B−b^c) for the PM and the SC pair (B^c−b^c) for the respective MM (e.g., G−$c^* \rightarrow C$−c^*) give rise to $\Delta\epsilon_{13}^{WC}(B) - \Delta\epsilon_{13}^{SC}(B^c) \approx \Delta\epsilon_{13}^{WC}(B)$. The triplet-like symmetry of the log-intensity difference at dominating specific hybridization consequently reflects the interaction strengths in the central WC pairings of specific duplexes which roughly divides into three states according to C > G ≈ T > A.

Background Correction: The PM − MM Difference. The MM probes were designed with the intention of measuring the amount of nonspecific hybridization, which contributes to the PM intensities. In particular, the almost identical sequence of the PM and MM probes of one pair is expected to bind nonspecific transcripts with essentially identical affinity. The subtraction of the MM from the PM intensity is therefore expected to remove this "chemical background". Making use of eqs 22−24 and 14, we obtain for the PM − MM difference of the middle base averaged intensities

$$I_B^\Delta = I_B^{PM} - I_{Bc}^{MM} \approx$$

$$F_0 N_B^F \frac{K_B^{PM,S}}{S_B^{PM}} (1 - E_B^S R_B^S)\{c_{RNA}^S + c_{RNA}^{NS} r_0^{PM} R_B\}$$

with

$$E_B^h = \frac{N_{Bc}^F}{N_B^F} \frac{K_{Bc}^{MM,h}}{K_B^{PM,h}} \approx$$

$$\exp(\ln 10(\epsilon_{0,13}^{PM-MM,h} - 2\sigma_{13}^{PM-MM,h}(B)))$$

$$R_B = \frac{(1 - E_B^{NS} R_B^S)}{(1 - E_B^S R_B^S)} \quad \text{and} \quad R_B^S = \frac{S_B^{PM}}{S_B^{MM}} \qquad (25)$$

Accordingly, the PM − MM intensity difference is linearly related to the fraction of specific transcripts and thus to the expression degree in an analogous fashion as the intensity of the single PM probes. The proportionality constant of the PM − MM difference is however reduced by the middle-base specific factor $(1 - E_B^S) = 0.70$ (B = A), 0.85 (T), 0.85 (G), and 0.90 (C) (using the data listed in Table 1 and $R_B^S = 1$) compared with the respective proportionality constant of the PM intensity.

Note also that subtracting the MM intensity from the PM signal only partly removes the "chemical background". Its relative contribution is reduced by the factor $R_B = 0.55$ (T) and 0.50 (C), and additionally reverses sign, $R_B = -0.25$ (A) and −0.20 (G), compared with the nonspecific contribution to the PM intensity. This result shows that the complementary middle letter of the PM and MM probes of one pair causes a base-specific bias of the affinity for nonspecific hybridization, which introduces a systematic source of variability between the PM and MM signals. The negative sign of the R_B values for B = A and G reflects the middle-base specific propensity for bright MM with purine middle bases in the limit of nonspecific hybridization (i.e., $I_B^\Delta < 0$ for B = A and G). The question whether this background term significantly affects gene expression measures obtained from additive intensity models[32,33] and suited correction algorithms will be separately addressed.

"Mysterious" MM. The pairwise design of PM/MM probes on GeneChip microarrays is based on three basic assumptions derived from conventional hybridization theory,[32] namely, (i) nonspecific binding is identical for PM and MM probes; (ii) the mismatch reduces the affinity of specific binding to the MM; (iii) the fluorescence response per bound transcript is identical for PM and MM and for specific and nonspecific hybridization as well. These assumptions seem to predict higher PM intensities compared with that of the MM for all probe pairs in contradiction to previous observations.[6] The "riddle of bright MM" for probe pairs with $I^{PM} < I^{MM}$ can be solved within the framework of conventional hybridization theory if one decomposes specific and nonspecific hybridization and analyzes the probe−target interactions on the level of base pairings and, in particular, as a function of the middle base. The explicit consideration of the strength of the central base-pairings in probe/transcript duplexes refines the picture and elucidates the origin of bright MM in terms of the pyrimidine/purine asymmetry of base pair interaction strengths. As a consequence, the first assumption modifies into "(i) nonspecific binding is on the average identical for PM and MM with a preference of probe pairs with a purine base in the middle of the PM sequence for bright MM and vice versa for pyrimidines". In conclusion, the "riddle of bright MM" is obviously due to some confusion about "what RNA hybridizes the probes".

Performance of Oligonucleotide Probes: Ideal Sensitivity and Specificity. The hybridization iso-

(32) Li, C.; Wong, W. H. *Genome Biol.* **2001**, *2*, 1.
(33) Li, C.; Wong, W. H. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 31.

therms of the DNA probes provide a natural starting point for the characterization of their performance. In the following, we will discuss the probe sensitivity and specificity as two important criteria, which can be derived from the isotherms to judge the quality of a probe as reporter for the concentration of specific target RNA in a complex mixture of RNA fragments.

The sensitivity characterizes the "detection strength" of a probe. Our definition of the sensitivity (eq 6) is motivated by practical reasons, which allow the calculation of its value for each GeneChip probe using its intensity with a minimum of assumptions and computational efforts. The respective values estimate the actual sensitivity in a relative scale under real conditions, which include specific and nonspecific hybridization and the degree of saturation as well. The sensitivity depends consequently also on the composition of the sample solution. We therefore introduce an ideal sensitivity, which estimates the potential detection strength of a probe for specific targets under ideal conditions, i.e., in the absence of nonspecific RNA fragments and saturation. The slope of the binding isotherms in the linear range at dominating specific hybridization provides a suited measure of this ideal value. For the middle base averaged isotherms, one obtains in the logarithmic scale (see eq 22)

$$Se_B^{P,S} \equiv \log\left(\frac{\partial I_B^P}{F_0 \partial c^S}\right)_{c^{NS}=0} = \log(N_B^{F,S} K_B^{P,S}) \qquad (26)$$

with P = PM, MM, Δ.

In particular, we are interested to compare the performance of the PM with that of the MM probes and with that of the PM − MM intensity difference, I_B^Δ (eq 25). The respective ideal sensitivity difference relative to that of the PM becomes with eqs 23−25

$$Se_B^{PM-MM,S} = Se_B^{PM,S} - Se_{Bc}^{MM,S} \approx - \epsilon_{13}^{WC-SC}(B)$$

$$Se_B^{PM-\Delta,S} = Se_B^{PM,S} - Se_B^{\Delta,S} \approx -\log(1 - E_B^S) \quad (27)$$

It turns out that the specific sensitivity of the PM distinctly exceeds that of the MM by $Se_B^{PM-MM,S} = 0.55$ (for B = A), 0.80 (T), 0.85 (G), and 1.10 (C). These values refer to an intensity difference between PM and MM probes of about 1 order of magnitude under ideal conditions. Contrarily, the sensitivity of the PM-MM intensity difference is nearly as large as that of the respective PM probe as indicated by the small difference $Se_B^{PM-\Delta,S} = 0.15$ (A), 0.08 (T), 0.07 (G), and 0.04 (C).

Specific and nonspecific RNA fragments compete for hybridization with the same probe. The specificity of a probe characterizes its selectivity, i.e., its power to decide between specific target RNA and the chemical background of nonspecific RNA fragments. We define the specificity as the log ratio of the probe response to specific and nonspecific hybridization in the absence of saturation, i.e., (see eq 22)

$$Sp_B^P \equiv -\log\left(\left|\frac{\partial I_B^P x^S = 1}{I_B^P(c^S = 0)\partial c^{NS}}\right|\right)_{c^{NS}=0} = -\log(|r_B^P r_B^{F,P}|) \qquad (28)$$

An ideal probe with a vanishing affinity for nonspecific binding consequently possesses a Sp_B^P value of infinity. Equations 23−25 provide the specificity difference between the PM and MM probes and between the PM intensity and the PM-MM intensity difference

$$Sp_B^{PM-MM} = Sp_B^{PM} - Sp_{Bc}^{MM} \approx - \epsilon_{13}^{WC-SC}(B) + 2\Delta\varphi_{13}^{WC}(B)$$

$$Sp_B^{PM-\Delta} = Sp_B^{PM} - Sp_B^\Delta \approx -\log |R_B| \qquad (29)$$

The specificity difference between the PM and MM probes reveals similar values as the respective sensitivity difference (compare eqs 27 and 29), $Sp_B^{PM-MM} = 0.63$ (A), 0.72 (T), 0.91 (G), and 1.02 (C). Accordingly, the MM specificity is distinctly smaller than that of the PM. Note that a specificity difference of about unity means that the affinity for specific binding to the PM exceeds that to the MM by 1 order of magnitude compared with the respective affinity for nonspecific binding. Contrarily, the negative values of $Sp_B^{PM-\Delta} = -0.61$ (A), -0.26 (T), -0.69 (G), and -0.30 (C) show that the specificity of the PM-MM intensity difference clearly outperforms the specificity of the PM.

Table 2 summarizes our evaluation of the different intensity measures based on eqs 27 and 29. These results might lead to the conclusion that the PM-MM intensity difference represents the optimal measure for specific RNA because it combines a nearly as high sensitivity with a distinctly better specificity compared with that of the PM on one hand but the much better sensitivity and specificity characteristics compared with that of the MM on the other hand.

In this study, we addressed selected aspects of the performance of GeneChip probes. In a more general context, the performance of microarray probes includes issues of chip design such as the optimization of the probe length, of the probe density on the chip, the selection of suited linker-groups which graft the probe on the support, and the optimization of the microstructure of the support (e.g., nylon or glass). Our study neglects the specific propensity of a selected probe and its complementary target sequence for intramolecular folding and/or for self-complementary dimerization. We expect that the middle-base related averaging "smoothes out" specific probe effects and thus this approach provides representative results for the mean affinities. The consideration of a specific probe sequence however requires an approach "beyond" the single-base or nearest-neighbor related description of oligonucleotide complex formation with the explicit treatment of the propensity for folding and self-complementary duplex formation.

Last but not least, the in situ photolithographic synthesis procedure of GeneChip oligonucleotide probes results in arrays in which the oligonucleotide features are heavily contaminated with truncated versions of the desired probe sequences. On one hand, this effect systematically decreases the average affinity of the oligonucleotide probes because their nominal length (e.g., 25 for GeneChip probes) overestimates their mean real length with consequences for the extracted thermodynamic parameters. On the other hand, the incomplete synthesis of the probes increasingly truncates the base positions toward their free end. We expect that our analysis of the systematic effect of the middle-base on the probe intensity is, if at all, only weakly affected by the truncation of the probes because their sequence remained mostly intact in the central sequence region and because the truncated probes of short length contribute with considerably smaller weigth to the spot intensity than longer oligonucleotides at small and intermediate concentrations of specific transcript owing to their exponentially decreasing binding constant.

Performance of the Microarray Experiment: Accuracy and Precision. The judgment of the performance of the probes also depends on the chosen experimental

Table 2. Summaries of the Performance of GeneChip Oligonucleotide Probes and of the Respective Differential Expression Measures[a]

		intensity measure		
		PM	MM	PM − MM
probe intensity	sensitivity	+	−	+
	specificity	−	−	+
differential expression	accuracy	±	±	+
	resolution	+	−	−

[a] "+" and "−" indicate good and bad performances according to eqs 27, 29, and 30. See text.

conditions and, in particular, on the RNA concentration and the RNA composition in the sample solution. The usual setup of the microarray experiment aims to estimate the differential expression in terms of a relative "fold" change, i.e., of the log-ratio of the transcript concentration of the sample of interest relative to that of an appropriately chosen reference sample, $DE^{true} = \log\{c_{RNA}^{S}(samp)/c_{RNA}^{S}(ref)\}$. Gene expression data analysis processes the respective probe intensities, $I_p^P(samp)$ and $I_p^P(ref)$, to provide an estimate of the differential expression DE (see, e.g., ref 9 for an overview).

The systematic deviation between this apparent and the true value, $\Delta DE = (DE - DE^{true})$, estimates the accuracy of the method. The specificity and the accuracy are closely related parameters because both depend on the relative contribution of nonspecific hybridization to the total intensity. In other words, a highly specific intensity measure is expected to provide also highly accurate DE values. Methods that use only the PM intensity typically underestimate the differential expression by more than 30%, i.e., $\Delta DE/DE_{true} > 0.3$,[9] partly because of incomplete background subtraction. Here one expects that, e.g., the PM − MM intensity difference provides a better alternative compared with PM-only measures of DE because of its higher specificity (see above).

The precision (or resolution) of gene expression analysis characterizes the confidence level of DE, i.e., the minimum difference between two DE values, which is judged as significant. The precision of an expression measure is inversely related to its variability, given, e.g., in terms of the standard deviation, SD(DE). Highly sensitive probes typically ensure a high precision, i.e., $SD(DE) \propto 1/Se^{P,S}$, because the relative error decreases with increasing intensity (see eq 16).

Our results predict a second interesting relation between the precision and the specificity of the probes besides this trivial effect. Note that the concentration of specific transcripts typically differs in the sample and the reference experiments, i.e., $c_{RNA}^{S}(samp) \neq c_{RNA}^{S}(ref)$. This change of c_{RNA}^{S} is accompanied by an alteration of intensity according to the hybridization isotherm (eq 22). The specificity can be interpreted as the variation of I_B^P referring to an increment of $\Delta c^S = 1$, if one neglects saturation for sake of simplicity (see eq 28). The Sp_B^P values can considerably vary as a function of the middle base. The middle base of the probes consequently introduces a systematic source of variability to the apparent differential expression between oligomers with different middle bases, which probe the RNA of the same gene. Note that the microarray probes are usually designed without special attention to their middle base. It seems appropriate to use the standard deviation of the specificity upon varying middle base as a measure of the precision of the apparent differential expression, i.e.

$$SD(DE) \propto SD(Sp^P) \equiv (1/4)\sqrt{\sum_{B = A,T,G,C}(Sp_B^P - \langle Sp_B^P\rangle_B)^2}$$
(30)

Equations 22−23 and 29 provide for the considered intensity measures $SD(Sp^P) \approx 0.0$ (for P = PM), 0.15 (MM) and 0.19 (Δ). Hence, the PM intensity should be judged as the best choice with respect to the precision of the differential expression because its specificity is invariant to changes of the middle base (see Table 2). Contrarily, the MM intensity and the PM − MM intensity difference introduce a considerable variability, which lowers the precision of the respective DE estimates. These findings agree with the results of recent statistical analyses, which show that expression measures based on MM or PM − MM intensities are less precise than that of PM-only estimates.[7] Hence, the good performance of the PM − MM intensity difference with respect to the sensitivity and specificity of the probes (see previous section) and the accuracy of the experiment must be relativized if one takes into account the resolution of the method. On the other hand, our results show that this effect possesses a systematic origin, which is mainly due to the change of base pair interactions in the middle of the probe sequence.

Taking together we emphasize that the performance of the microarray experiment depends on the performance of the chosen intensity measures, which in turn are related to the hybridization isotherms of the probes. The explicit consideration of sequence dependent factors in combination with the concentration dependence in more sophisticated analysis algorithms is expected to improve gene expression measures.

Summary and Conclusions

Our microscopic theory of hybridization explains the concentration dependence and the effect of the middle base on the intensity of perfect matched (PM) and mismatched (MM) microarray probes in terms of effective binding constants, which in turn depend on the base pair interactions in DNA/RNA oligonucleotide duplexes. We found that both PM and MM probes bind nonspecific RNA fragments on the average with similar affinity.

Both, the PM and MM probes respond to the concentration of specific transcripts and thus to the expression degree. The mean binding constant of the PM however exceeds that of the MM by nearly 1 order of magnitude. The markedly weaker binding affinity of the MM can be attributed to the self-complementary pairing of the middle base, which on the average only weakly contributes to the stability of the specific duplexes.

The pyrimidine/purine asymmetry of base pair interaction in the DNA/RNA hetero-duplexes splits the intensity difference between PM and MM probes at dominating nonspecific hybridization into two branches and at dominating symmetric hybridization into three branches. The former effect reflects the reversal of the central WC base pairing for each probe pair whereas the latter effect can be rationalized in terms of the relatively weak SC base pairings of the MM.

The free energy of duplex formation between target and probe mainly determines the observed intensities whereas the heterogeneity of fluorescence labeling provides only a second-order contribution.

The PM-MM intensity difference outperforms the PM intensity in terms of specificity because it largely removes the chemical background. On the other hand, the MM signal in the PM-MM difference lowers the precision of

differential gene expression measures owing to systematic effects of the middle base on the binding affinity of the MM.

In conclusion, hybridization on microarrays is in agreement with the basic rules of DNA/RNA hybridization in solution. The presented model implies the refinement of existing algorithms of probe level analysis to correct microarray data for nonspecific background intensities. In particular the results suggest the consideration of a middle-base specific correction term for the PM-MM intensity difference, which takes into account the fluctuations of the background intensity due to the reversal of the WC pairing in nonspecific duplexes.

Supporting Information Available: (i) The single base contribution to the fluorescence emission; (ii) the signal and sensitivity error of single Affymetrix GeneChips and, (iii) an overview of SB free energy parameters of RNA/DNA duplexes. This material is available free of charge via the Internet at http://pubs.acs.org or http://www.izbi.de.

LA051231S