GeneChip microarrays – Signal intensities, RNA concentrations and probe sequences

Article *in* Journal of Physics Condensed Matter · April 2006

DOI: 10.1088/0953-8984/18/18/S04

CITATIONS
24

READS
368

2 authors:

Hans Binder
University of Leipzig
191 PUBLICATIONS **6,795** CITATIONS
SEE PROFILE

Stephan Preibisch
Max-Delbrück-Centrum für Molekulare Medizin
75 PUBLICATIONS **29,196** CITATIONS
SEE PROFILE

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **18** (2006) S537–S566 doi:10.1088/0953-8984/18/18/S04

GeneChip microarrays—signal intensities, RNA concentrations and probe sequences

Hans Binder[1] **and Stephan Preibisch**

Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Haertelstraße 16-18, Germany

E-mail: binder@izbi.uni-leipzig.de

Received 24 July 2005, in final form 12 December 2005
Published 19 April 2006
Online at stacks.iop.org/JPhysCM/18/S537

Abstract

GeneChip microarrays consist of hundreds of thousands of oligonucleotide probes. The transformation of their signal intensities into RNA transcript concentrations requires the knowledge of the response function of the measuring device. We analysed the 'apparatus' function of perfect match (PM) and mismatched (MM) oligonucleotide probes of GeneChip microarrays after changes of the target concentration using the results of a spiked-in experiment. In agreement with previous studies we found that a competitive two-species *Langmuir*-adsorption model describes the probe intensities well. Each PM and MM probe is characterized by two hybridization constants which specify the propensity of the probe to bind specific and non-specific transcripts. The affinity for non-specific hybridization is on average equal for PM and MM. The purine–pyrimidine asymmetry of base pair interaction strengths, however, causes a characteristic PM–MM intensity difference, the sign of which depends on the middle base of the probe. The affinity for specific hybridization of the PM exceeds that of the MM on average by nearly one order of magnitude because the central mismatched base only weakly contributes to the stability of the probe/target duplexes. For the first time we differentiate between the free energy parameters related to the 64 possible middle-triples of DNA/RNA oligomer duplexes with a central Watson–Crick pairing and a central mismatched pairing. Both the PM and MM probes respond to the concentration of specific transcripts, which can be estimated from the PM and MM probe intensities using the *Langmuir*-model. The analysis of the PM–MM intensity difference provides at least no loss of accuracy and precision of the estimated concentration compared with the PM-only estimates which in turn outperform the MM-only estimates. The results show that the processing of the PM–MM intensity difference requires the consideration of a background term due to non-specific hybridization, which is, however, reduced by nearly one order of magnitude when compared with the respective background of the

[1] Author to whom any correspondence should be addressed.

PM and MM probes. The calculation of the sequence-specific affinity constants using a positional-dependent nearest-neighbour model opens up the possibility to estimate target concentrations beyond the training set of several hundred of spiked-in probes.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

One central methodical task of the microarray technique is the parallel and exact probing of the abundance of thousands of different target-RNA sequences in a few or even only one experiment. The transformation of the signal intensities of the probe spots of the chip into concentration values requires the knowledge of the response function of the measuring device. The basic idea of the functioning of gene microarrays is based on the sequence specificity of probe–target interactions combined with fluorescence detection. In reality, this straightforward principle is opposed by the complexity of the experimental system due to imperfections of chip fabrication and RNA preparation, due to the non-linearity of the probe response, the detection limit of the optical scanner and especially due to problems which are inherently connected with the high throughput character of the method. Note that the intention to screen wide genomic areas at once necessarily results in cross hybridizations between the probes and non-complementary and thus non-specific RNA sequences which can add considerable uncertainty to the data.

The accompanying paper deals with selected aspects of microarray hybridization on a theoretical basis [1]. The present paper supplements this approach and analyses publicly available chip data of a special calibration experiment to specify the hybridization model in terms of the binding isotherm and the sequence dependence of the binding parameters. These tasks are essential prerequisites for developing adequate analysis algorithms of microarray data.

The paper divides into three parts. (i) The methodical section introduces the data used, the models and analysis algorithms. (ii) The second section focuses on the response function of the microarray probes which relates the measured intensity to the concentration of the RNA target. This analysis is partly in parallel with previous work [2, 3]. This way it reviews the state of the art and illustrates the reliability of the approach. More importantly, the performed fits of the intensity data accomplish a necessary precondition for the next step: namely, they provide the set of affinity constants which are further analysed in part (iii). Here we study the relations between the affinities of probes with and without mismatched base-pairings for specific and non-specific hybridization, their sequence dependence and finally we estimate the apparent binding strengths of canonical Watson–Crick (WC; for example C • g) and of self-complementary (SC; for example C • c*) base pairs in the hybrid duplexes (upper and lower case letters refer to the nucleotides in the DNA probe and RNA target strains, respectively; the asterisk denotes labelling). This section continues our previous work on physico-chemical aspects of the GeneChip technology [4–8].

This study for the first time reports sequence-specific interaction free energies for matched and mismatched base pairings which were extracted from GeneChip microarray intensity data. The data discriminate between the contribution of specific and non-specific hybridization to the observed intensities. The results provide new insights into practical issues such as the intensity relations between perfect matched and mismatched GeneChip probes as a function of the middle triple, i.e. the three central bases in the probe sequence. Related results of previous work refer basically only to non-specific binding [9, 10].

2. Chip data and signal processing

2.1. Chip data

The present analysis uses data obtained by the GeneChip microarray technology which is being widely applied in many areas of biological and medical research. These high density oligonucleotide microarrays enable the analysis of the transcripts of up to about 45 000 genes in parallel. The probes are oligonucleotides of 25 bases in length [11, 12]. There are two types of probe: the so called perfect match (PM) reference probes that exactly match a 25meric target sequence taken from the gene of interest, and so-called mismatched (MM) partner probes. Their sequence differs from the PM probes only by a single base in the centre of the sequence, which is replaced by the complement. The MM probes intend to correct the PM intensity for the amount of non-specific hybridization because they are relatively insensitive for target binding. In addition a number of (usually 11) PM/MM probe pairs are taken from the same gene and form a so-called probe set. The summarization of their signals into one expression measure intends to increase the accuracy compared with chips which use only one probe per gene.

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 $N_{\mathrm{pair}} = 11 \times 42 = 462$ 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, [S]. Fourteen ($N_{\mathrm{conc}} = 14$) different concentrations ranging from [S] = 0 pM (i.e. non-specific RNA only, see below) to 512 pM were used. 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) and three replicates of each condition. Non-specific hybridization was taken into account by adding a complex human RNA background in equal amounts to all hybridization solutions ([NS] $\approx$ constant). It was extracted from a HeLa cell line not containing the spiked-in transcripts. The PM and MM probe intensities were corrected for the optical background before further analysis using the algorithm provided by MAS 5.0 which subtracts the mean intensity averaged over the 2% of probes possessing the smallest intensity on each chip from the intensity of all other probes [13].

2.2. Hybridization model and probe intensities

The *Langmuir*- and *Sips*- (also called *Langmuir–Freundlich*-) models of adsorption on solid surfaces provide an adequate starting point to describe the hybridization of oligonucleotide probes on microarrays. The *Langmuir*-model assumes a limited number of energetically identical binding sites whereas the more general *Sips*-model in its original meaning corresponds to sorption onto sites with a symmetrical quasi-*Gaussian* distribution of the adsorption free energies [14]. The two-species *Sips*-isotherm in the context of the microarray experiment assumes that two different RNA species compete for the adsorption sites on the chip which are provided by each oligonucleotide probe. The following equation relates the fluorescence intensity emitted from a probe spot to the concentration of specific and non-specific transcripts, [S] and [NS], respectively (see the accompanying paper [1]),

$$I_p^{\mathrm{P}} = I_p^{\mathrm{P,max}} \frac{(X_p^{\mathrm{P,S}})^{a_p^{\mathrm{P}}} + X_p^{\mathrm{P,NS}}}{1 + ((X_p^{\mathrm{P,S}})^{a_p^{\mathrm{P}}} + X_p^{\mathrm{P,NS}})}. \tag{2.1}$$

The index 'p' ($p = 1 \ldots N_{\mathrm{pair}}$) assigns the probe pair number and the superscripts 'P = PM, MM' and 'h = S, NS' specify the probe type and the type of the transcripts, respectively. The

$X_p^{\mathrm{P,h}}$ characterize the binding strength of the DNA probes for duplex formation with the RNA transcripts. The binding strength is directly related to the respective association constant $K_p^{\mathrm{P,h}}$ and the RNA concentration [h] according to the mass action law, $X_p^{\mathrm{P,h}} \equiv K_p^{\mathrm{P,h}}[\mathrm{h}]$.

Only the concentration of specific transcripts, $[S]$, is explicitly known in the spiked-in experiment. The intensity of each probe is consequently described by the four parameters, $K_p^{\mathrm{P,S}}$, $X_p^{\mathrm{P,NS}}$, a_p^{P} and $I_p^{\mathrm{P,max}}$. The number of parameters can be reduced by additional constraints, for example by the assumption of common values of the intensity asymptote, $I_p^{\mathrm{P,max}}$, and the *Sips*-exponent, a_p^{P}, for all probes (see the next section).

With $a_p^{\mathrm{P}} = 1$, equation (2.1) simplifies to the two-species *Langmuir*-isotherm which is identical with the previously used function [2, 3]

$$I_p^{\mathrm{P}} = A_p^{\mathrm{P}} \frac{B_p^{\mathrm{P}} \cdot [\mathrm{S}]}{1 + B_p^{\mathrm{P}} \cdot [\mathrm{S}]} + D_p^{\mathrm{P}}. \tag{2.2}$$

The model parameters in this equation are combinations of the physically motivated parameters $K_p^{\mathrm{P,S}}$, $X_p^{\mathrm{P,NS}}$ and $I_p^{\mathrm{P,max}}$, according to

$$A_p = \frac{I_p^{\mathrm{P,max}}}{1 + X_p^{\mathrm{P,NS}}}, \qquad B_p^{\mathrm{P}} = \frac{K_p^{\mathrm{P,S}}}{1 + X_p^{\mathrm{P,NS}}} \qquad \text{and} \qquad D_p^{\mathrm{P}} = I_p^{\mathrm{P,max}} \frac{X_p^{\mathrm{P,NS}}}{1 + X_p^{\mathrm{P,NS}}}.$$

We will use equation (2.1) for further analyses because it directly provides the probe-specific sorption strengths of specific and non-specific RNA fragments which, in turn, are functions of the respective probe sequence.

2.3. The effect of the sequence on the sensitivity of the probes

Let us define a probe-specific incremental contribution of the binding constants,

$$\begin{aligned} Y_p^{\mathrm{P,S}} &\equiv \log K_p^{\mathrm{P,S}} - \langle \log K_p^{\mathrm{P,h}} \rangle_\Sigma \qquad \text{and} \\ Y_p^{\mathrm{P,NS}} &\equiv \log X_p^{\mathrm{P,NS}} - \langle \log X_p^{\mathrm{P,NS}} \rangle_\Sigma = \log K_p^{\mathrm{P,NS}} - \langle \log K_p^{\mathrm{P,NS}} \rangle_\Sigma, \end{aligned} \tag{2.3}$$

with $\log A \equiv \log_{10} A$. The angular brackets, $\langle \ldots \rangle_\Sigma$, denote averaging over the ensemble of considered spiked-in probes. Note that the second equation holds because the concentration of non-specific transcripts, [NS], is a constant and thus it cancels out in the right part. The $Y_p^{\mathrm{P,h}}$ characterize the sensitivity of the probe for specific (h = S) and non-specific (NS) transcripts [4], i.e. their ability to detect a given amount of RNA fragments.

The sensitivity can be additively decomposed into a specific free energy contribution of each base at each position of the probe sequence, ξ_p^{P}, according to

$$Y_p^{\mathrm{P,h}} = -\sum_{k=1}^{\mathrm{LP}} \sum_{\mathrm{B=A,T,G,C}} (\Delta\varepsilon_k^{\mathrm{W}}(\mathrm{B})(\delta(\xi_{p,k}^{\mathrm{P}}, \mathrm{B}) - f_k^{\Sigma}(\mathrm{B}))). \tag{2.4}$$

Here $\delta(x, y)$ is the delta function ($\delta = 1$ for $x = y$ and $\delta = 0$ otherwise) and $f_k^{\Sigma}(\mathrm{B})$ is the fraction of base B at position k of the sequence within the chosen ensemble of probes. The positional-dependent single-base related terms, $\Delta\varepsilon_k^{\mathrm{W}}(\mathrm{B})$, can be assigned to W = WC and SC pairings depending on the chosen probes (PM or MM) and conditions (h = S or NS) (see below). The $\Delta\varepsilon_k^{\mathrm{W}}(\mathrm{B})$ should be interpreted as 'apparent' (or effective) *Gibbs* free energy values in contrast to the respective 'intrinsic' free energy of adsorption. The latter value refers exclusively to the binding reaction between sorbent and sorbate whereas the former one in addition includes effects such as electrostatic and entropic blocking, competitive complex formation and the fluorescence 'strength' of each base (see [1, 6] for details).

2.4. Least-square fits

The parameters of the intensity model were estimated by non-linear least-square fits of equation (2.1) to the experimental intensity data as a function of the spiked-in concentration to minimize the sum $SQI_p^{\mathrm{P}} = \sum_{[\mathrm{S}]} \omega_p^{-2} (I_{p,\mathrm{exp}}^{\mathrm{P}} - I_{p,\mathrm{calc}}^{\mathrm{P}})^2$. The weighting factor, ω_p^2, was estimated as a function of signal intensity using the error function $\omega_p^2 = E_1 + E_2/(I_p^{\mathrm{P}}) + E_3/(I_p^{\mathrm{P}})^2$ [6]. It accounts for the increase of signal error at small intensities in a logarithmic scale. The constants E_i ($i = 1 \ldots 3$) 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 [6].

The least-square fits adjust the two binding parameters in equation (2.1), $K_p^{\mathrm{P,S}}$ and $X_p^{\mathrm{P,NS}}$, for each PM and MM probe. The asymptotic intensity value $I_p^{\mathrm{P,max}}$ and/or the *Sips*-exponent a_p^{P} were either optimized also for each probe or globally for the whole ensemble of spiked-in probes. The latter approach provides only one common value of I^{max} and/or of a^{Sips} for all probes. Hence, the total number of adjustable parameters to fit the intensities of all intensity data of the PM or MM probes varies between $N_{\mathrm{parm}} = 4 \cdot N_{\mathrm{pair}}$ (i.e., $K_p^{\mathrm{P,S}}$, $X_p^{\mathrm{P,NS}}$, $I_p^{\mathrm{P,max}}$ and a_p^{P}), $N_{\mathrm{parm}} = 3 \cdot N_{\mathrm{pair}} + 1$ (i.e., $K_p^{\mathrm{P,S}}$, $X_p^{\mathrm{P,NS}}$ and one global value of I^{max} or a^{Sips}) and $N_{\mathrm{parm}} = 2 \cdot N_{\mathrm{pair}} + 2$ (i.e., $K_p^{\mathrm{P,S}}$, $X_p^{\mathrm{P,NS}}$ and one global value of I^{max} and a^{Sips}).

The relevance of the improvement of the least-square fit by increasing the number of parameters used in model 2 compared with the original model 1 can be evaluated in terms of the ratio of the sum of squared residuals,

$$FI^{1\to 2} = \frac{\sum SSQI_p^1/(N_{\mathrm{data}} - N_{\mathrm{parm}}^1 + 1)}{\sum SSQI_p^2/(N_{\mathrm{data}} - N_{\mathrm{parm}}^2 + 1)} \tag{2.5}$$

where the sums run over all considered spiked-in probes and concentrations, $N_{\mathrm{data}} = N_{\mathrm{pairs}} \cdot N_{\mathrm{conc}}$. The significance level of the improvement is given by the F-statistics of $FI^{1\to 2}$ with the respective degrees of freedom.

The positional-dependent coefficients of the SB model, $\Delta\varepsilon_k^{\mathrm{W}}(\mathrm{B})$, were determined by multiple linear regression which minimizes $SQY^{\mathrm{P,h}} = \sum_{p=1}^{N\mathrm{pair}} (SQI_p^{\mathrm{P}})^{-1} (Y_{p,\mathrm{exp}}^{\mathrm{P,h}} - Y_{p,\mathrm{calc}}^{\mathrm{P,h}})^2$ (see equations (2.3) and (2.4)).

3. Hybridization isotherms of matched and mismatched microarray probes

3.1. Adjustment of the intensity asymptote

Figure 1 shows the intensities of the 462 spiked-in PM and MM probes and of the PM–MM intensity difference as a function of the spiked-in concentration, [S]. The progressive addition of specific transcripts causes the systematic increase of the intensity. The curves in figure 1 illustrate this behaviour for three individual probe pairs. The background level and the slope considerably differ between the chosen examples. The data clearly show that the intensities of the probes vary over more than three orders of magnitude even at constant [S] especially in the range of dominating non-specific hybridization. The PM–MM intensity difference is either positive at all concentration values or significantly negative below a certain concentration of specific transcripts (see lower panel in figure 1). The system obviously produces a very heterogeneous picture which requires systematic analysis in terms of an appropriate hybridization model.

In a first attempt, the concentration-dependent intensity data of all spiked-in probes are fitted using equation (2.1) by adjusting the three probe-specific parameters, $K_p^{\mathrm{P,S}}$, $X_p^{\mathrm{P,NS}}$ and

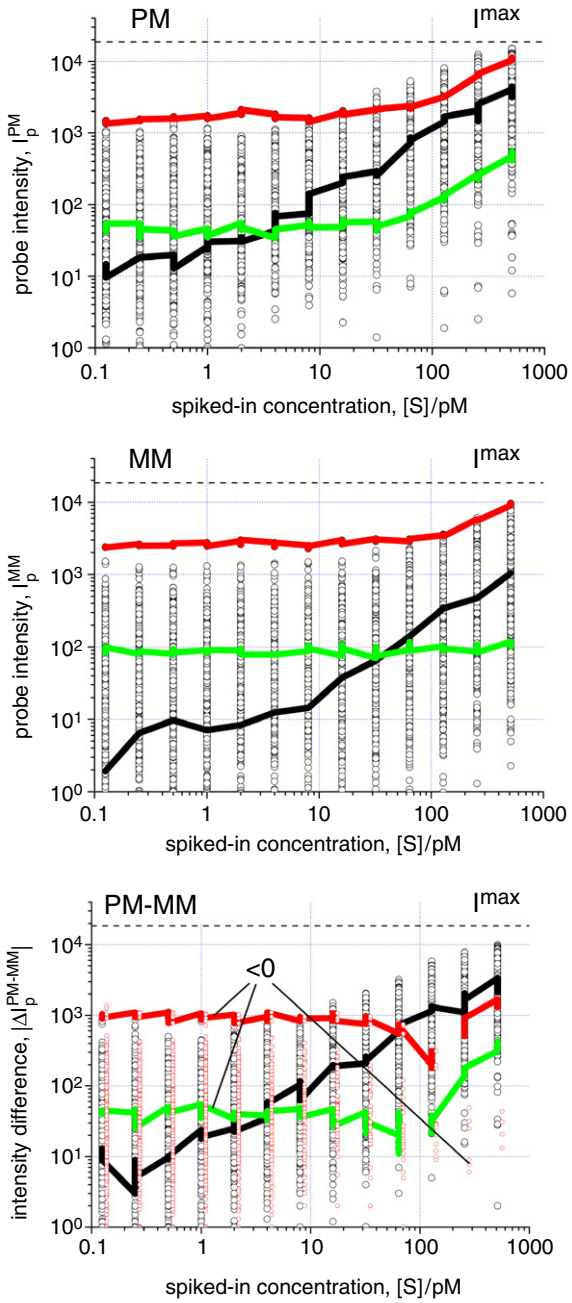

Figure 1. Signal intensities of perfect match (PM) and mismatch (MM) probes and of the absolute value of their difference, $|\Delta I^{\mathrm{PM-MM}}| = |I^{\mathrm{PM}} - I^{\mathrm{MM}}|$ (PM–MM) as a function of the spiked-in concentration of specific RNA transcripts, [S]. The 462 'spiked-in' probes cover a wide intensity range of more than three orders of magnitude even for constant [S]. Three of them are explicitly highlighted to illustrate their completely different behaviour (see the curves: black: 200665_s_at; red: 204836_at; green: 212827_at). Note that the PM–MM difference becomes negative for about 40% of all probe pairs in the absence of specific transcripts (small circles in the lower panel) and for two of the selected examples at small [S] as indicated by '<0'. The fraction of negative PM–MM values continuously decreases with [S] and almost completely vanishes at [S] > 128 pM. The maximum intensity level estimated from the *Langmuir*-fits, $I^{\max}$, is shown by the horizontal line.

$I_p^{\mathrm{P,max}}$, and leaving $a_p^{\mathrm{P}} = 1$ constant. The maximum intensity value of each probe, $I_p^{\mathrm{P,max}}$, is expected to depend on the number of labelled bases within the respective complementary RNA fragment. To check this prediction we correlate the obtained $I_p^{\mathrm{P,max}}$-data with the number of labelled cytosines and uracyls in the RNA sequences. Here we simply assume 65mers which contain the complementary probe region plus the 20 adjacent bases on both sides. No significant correlation was detected (data not shown).

Moreover, the obtained coefficient of variation of the $I_p^{\mathrm{P,max}}$ values is about 0.1–0.2, which corresponds to a degree of scattering of 10%–20% of the mean value of the maximum intensity. Theoretical considerations predict however a coefficient of variation due to labelling of only a few per cent (see [1] and also [4]). The discrepancy between the measured and expected scattering width lets us conclude that the variation of $I_p^{\mathrm{P,max}}$ is caused by other effects of stochastic and/or systematic, but unknown origin. Peterson *et al* [15] showed that the scattering of the asymptotic intensity might be an artefact due to a limited concentration window of the data.

In a second attempt we fitted the intensity data using a constant value of the maximum intensity value which was set to the log-average of the $I_p^{\mathrm{PM,max}}$-data obtained in the previous analysis, $\log I^{\max} = \langle \log I_p^{\mathrm{PM,max}} \rangle$. The comparison of the goodness of fit of both models provides $FI^{1\to 2} \approx 1.05$ (see equation (2.5)) and a significance level of $p \approx 0.04$ using the respective F-statistic. In view of this relatively weak improvement of the fit and because of the unknown physical meaning of a varying probe-specific maximum intensity parameter we applied the simpler second model which assumes the global mean, $I^{\max}$, as a common value for all PM and MM probes (see the horizontal dashed line in figure 1). This assumption differs from previous approaches to fit microarray intensity data which used a probe-dependent asymptote of the hybridization isotherm [2, 3].

3.2. Sips versus Langmuir model

In the next step of data analysis we compared the fits of the *Langmuir*-model which uses a constant exponent $a_p^{\mathrm{P}} = \text{constant} = 1$ with that of the *Sips*-model which freely adjusts the exponent for each probe. The improvement of the quality of the fits was characterized by $FI^{1\to 2} \approx 1.1$–1.2 (equation (2.5)) which provides p-values in the range $p \approx 10^{-4}$–10^{-8}. Hence, the consideration of the *Sips*-exponent slightly but significantly improves the fit of the measured hybridization isotherms.

In its original meaning this *Sips*-exponent is inversely related to the width of a *Gaussian* free energy distribution of heterogeneous adsorption sites. Also free energy distributions of non-*Gaussian* shape caused, for example, by the truncation of the probes, and also other effects related to homogeneous adsorption such as electrostatic and entropic blocking, give rise to *Sips*-like isotherms with exponents less than unity [1]. Figure 2 shows the frequency distribution of the *Sips*-exponent obtained from the fit of equation (2.1) to the intensity data of the PM and MM probes. The distributions are centred about their mean, which is nearly equal for the PM and MM probes, $\langle a^{\mathrm{PM}} \rangle \approx \langle a^{\mathrm{MM}} \rangle \approx 0.9$ (see the vertical bars in figure 2). The frequency of the *Sips*-exponents of the PM reveals a pronounced peak in the range $a_p^{\mathrm{PM}} \approx 0.65$–$1.0$ whereas the MM-exponents are more evenly distributed.

The *Sips*-exponent of the PM and MM shows a relatively weak positive correlation with the respective constant of specific hybridization (see figure 2, lower panel). In other words, a stronger affinity constant, $K_p^{\mathrm{P,S}}$, on the average is paralleled by a bigger a_p^{P}-value. Note that both parameters, $K_p^{\mathrm{P,S}}$ and a_p^{P}, determine the slope of the isotherm at a given concentration. Given a limited concentration range of the data, a steeper slope of the isotherm can be adjusted by bigger values of $K_p^{\mathrm{P,S}}$ and/or a_p^{P} as well. Hence, the exponent should be tentatively judged

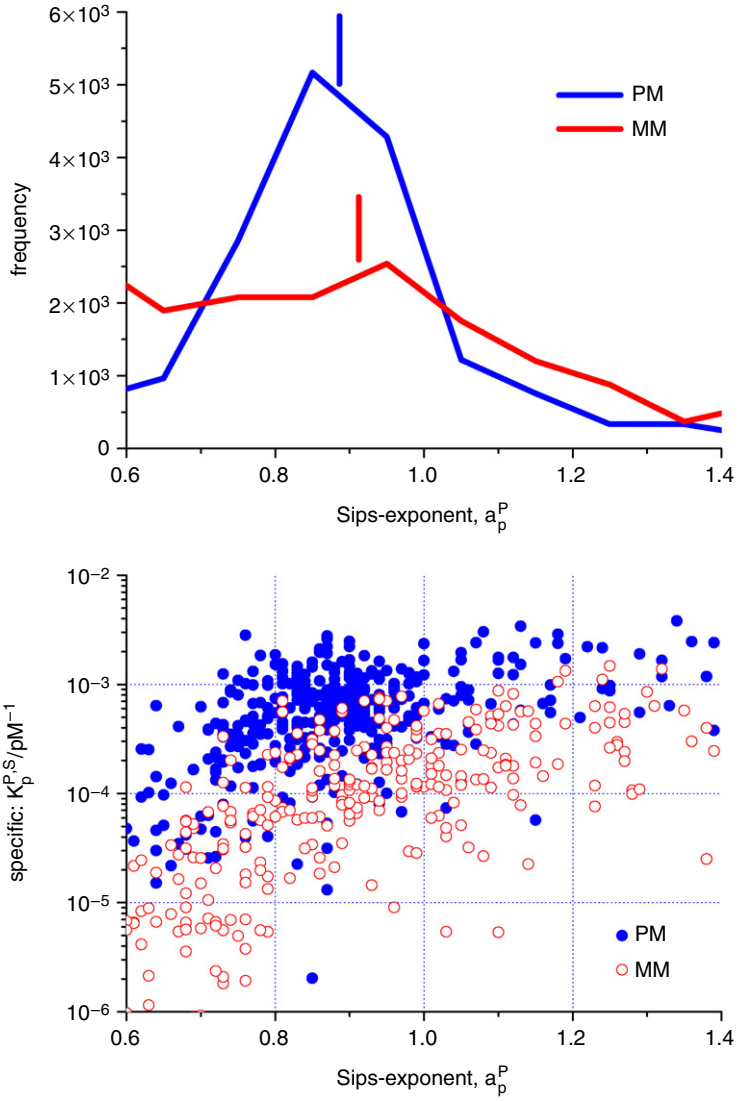

Figure 2. Frequency distribution of the *Sips*-exponent of the PM and MM probes (upper panel). The vertical bars indicate the respective mean value of the exponent, $\langle a^{\mathrm{P}} \rangle$ (P = PM, MM). The lower panel shows the correlation between the binding strength of specific hybridization (y-axis) and the *Sips*-exponent (x-axis).

as an additional fit parameter without clear physical meaning despite the arguments in favour of the *Sips*-model (see [1]). The question about the physical significance of the *Sips*-exponent obtained from microarray data requires additional work.

The present paper focuses on the analysis and interpretation of the model parameters obtained from the fit of an appropriate hybridization isotherm in terms of the respective probe sequence. In the following sections we use the results obtained from the fit of the simpler, more parsimonious (i.e., without unnecessary parameters) *Langmuir*-model with a common intensity asymptote of all probes in view of the arguments given above. This approach seems justified because: (i) the applied *Langmuir*-model provides acceptable fits of the intensity data (see

next section); (ii) the *Sips*-model and the assumption of a probe-dependent intensity asymptote only slightly improve the goodness of the fit; (iii) both the *Sips*- and the *Langmuir*-models provide similar values of the affinity parameters $K_p^{\mathrm{P,S}}$ and $X_p^{\mathrm{P,NS}}$ (not shown). Thus, the basic conclusions about the effect of the sequence of a probe on its affinity and intensity is virtually not affected by the choice of the model. Note that a recent detailed comparison of several adsorption models including the *Sips*- and *Langmuir*-isotherms has shown that the *Langmuir*-model appropriately describes microarray intensity data taken from the HG-U95 spiked-in experiment in agreement with our results [3].

3.3. Langmuir-isotherms of PM and MM probes

The fit of the concentration dependence of the spiked-in data by means of the *Langmuir*-model provides the four binding parameters $K_p^{\mathrm{PM,S}}$, $X_p^{\mathrm{PM,NS}}$, $K_p^{\mathrm{MM,S}}$, $X_p^{\mathrm{MM,NS}}$ for each PM/MM probe pair. The common value of the intensity asymptote, $I^{\max}$, was used to normalize the probe intensities according to $\theta_p^{\mathrm{P}} = I_p^{\mathrm{P}}/I^{\max}$. The normalized intensity, θ_p^{P}, defines the surface coverage of the probe, i.e., the fraction of DNA oligomers which are 'occupied' by bound RNA fragments.

Figures 3 and 4 replot the intensity data shown in figure 1 into θ-versus-$K_p^{\mathrm{P,S}} \cdot [\mathrm{S}]$ (left panel) and θ-versus-$X_p^{\mathrm{P}} \equiv K_p^{\mathrm{P,S}} \cdot [\mathrm{S}] + X_p^{\mathrm{P,NS}}$ (right panel) coordinates. The chosen representations normalize the behaviour of the probes along the ordinate according to their surface coverage and along the abscissa according to their binding strength for specific transcripts (left part of figure 4) and to their overall binding strength including that for the non-specific transcripts (figure 3 and right part of figure 4). The normalized intensity data of both the PM and MM probes group well along the hyperbolic master curve provided by the *Langmuir*-isotherm, $y = x/(1+x)$ (see the curves in the figures). Note that the overall binding strength of the MM is on the average nearly one order of magnitude weaker than that of the PM (i.e., $X_p^{\mathrm{MM}} \ll X_p^{\mathrm{PM}}$) at higher abscissa values. As a consequence, the MM intensities show a less pronounced saturation behaviour compared with that of the PM.

Figure 3 (upper panel) illustrates the different saturation behaviour of the PM and MM probes for a single probe pair. The surface coverage of the selected PM probe clearly exceeds $\theta^{\mathrm{PM}} > 0.1$ at the biggest spiked-in concentrations whereas that of the MM reaches at maximum only $\theta^{\mathrm{MM}} \sim 0.07$. The analysis of all spiked-in data shows that about 18% of the PM-data refer to a surface coverage larger than $\theta^{\mathrm{PM}} > 0.1$ whereas only 9% of the MM exceeds 0.1 (figure 4, right panel). The effect of the varying binding affinity of probes of different sequence on their *Langmuir*-type behaviour is discussed in [16].

Figure 3 (lower panel) shows all spiked-in data of the PM at the smallest and highest concentrations, $[\mathrm{S}] = 0$ and 512 pM, respectively. Both data sets overlap at intermediate abscissa values. Hence, the affinity for non-specific hybridization can exceed that for specific binding even at high concentrations of specific transcripts owing to large differences between the binding constants of the individual probes.

The abscissa in the left part of figure 4 normalizes only the strength of specific hybridization of the different probes, $X_p^{\mathrm{P,S}}$. This representation filters out the effect of the non-specific background on the individual probe intensities. The data clearly split up at small abscissa values. There is no master curve with the argument $X_p^{\mathrm{P,S}}$ which uniquely describes all the data at small [S]. The limiting value of the hybridization isotherm at vanishing $[\mathrm{S}] \ll 1/K_p^{\mathrm{P,S}}$ characterizes the limiting surface coverage due to non-specific hybridization, $\theta_p^{\mathrm{P}}|_{[\mathrm{S}]\to 0} \to X_p^{\mathrm{P,NS}}/(1 + X_p^{\mathrm{P,NS}}) \approx X_p^{\mathrm{P,NS}}$, because the probes are nearly exclusively occupied by non-specific transcripts at these conditions. About 99% of the PM and MM probes of the spiked-in data possess a limiting surface coverage of less than 2.5×10^{-2} as indicated by the

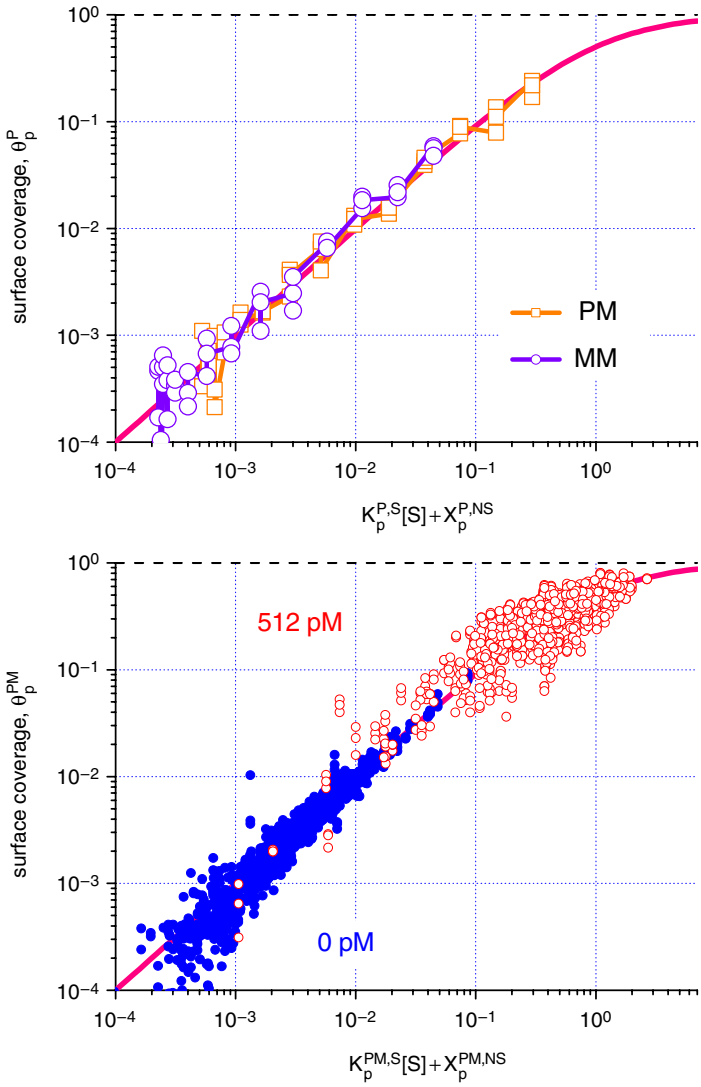

Figure 3. Surface coverage as a function of the 'total' binding strength of one selected PM/MM-probe pair at all 14 spiked-in concentration (upper panel) and of all 462 PM probes of the spiked-in experiment at two selected spiked-in concentrations ([S] = 0 pM and 512 pM, lower panel). The total binding strength, $X = K_p^{P,S}[S] + X_p^{P,NS}$, was obtained by fits of the *Langmuir*-isotherm to the respective intensity data (see text). All data points group well along the hyperbolic *Langmuir*-master curve, $\theta = X/(1+X)$ (see line). Note the similar behaviour of the PM and MM (upper panel) and the partial overlap of the data points referring to the completely different [S]-values (lower panel). All data conditions are realized in triplicate.

range between the two master curves which are calculated with $X_p^{P,NS} = 0$ and 2.5×10^{-2} (see left part of figure 4).

3.4. The PM–MM difference

The MM probes were designed with the intention of measuring the amount of non-specific hybridization, which contributes to the PM intensities. The results presented in the previous

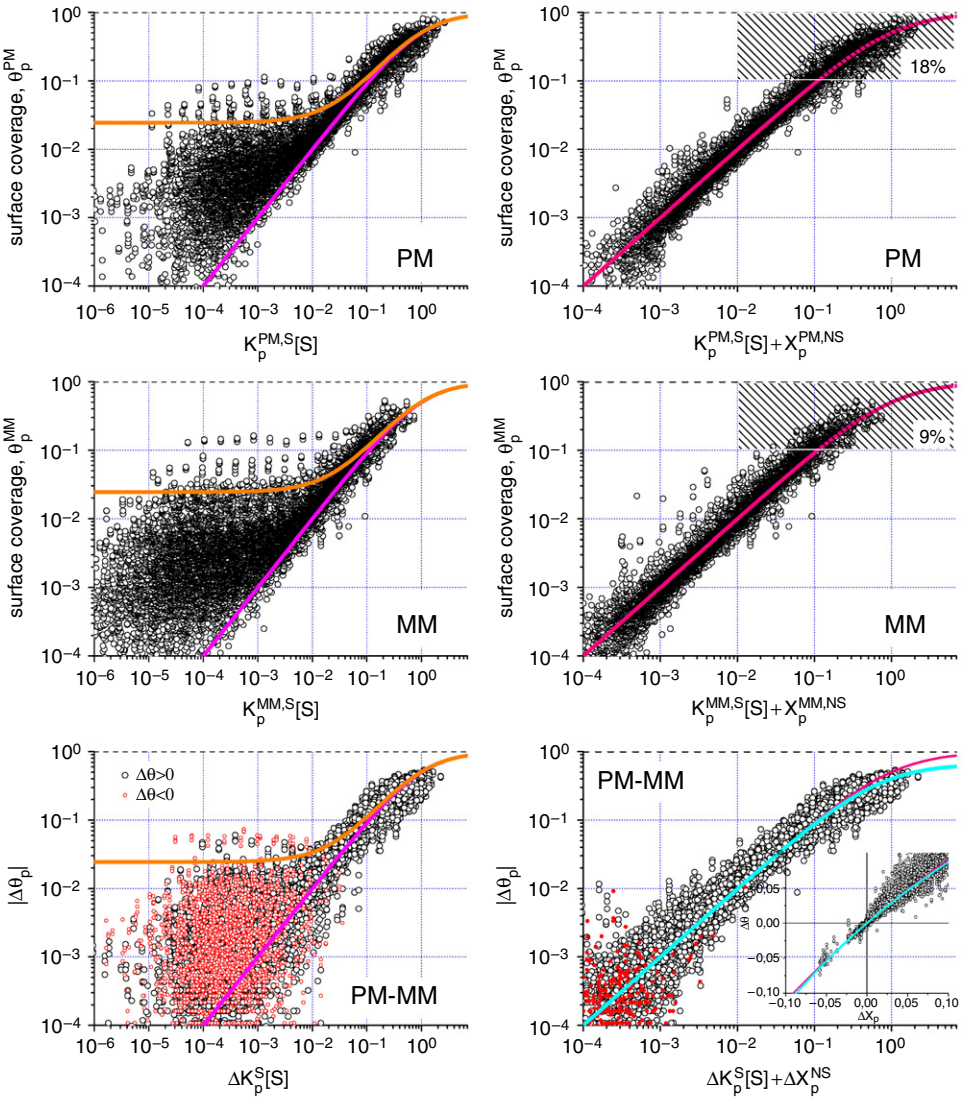

Figure 4. Surface coverage of PM and MM probes ($\theta_p^{\mathrm{P}} = I_p^{\mathrm{P}}/I^{\max}$, P = PM, MM, see equation (2.1)), and of their difference, $|\Delta\theta_p^{\mathrm{PM-MM}}| = |\theta_p^{\mathrm{PM}} - \theta_p^{\mathrm{MM}}|$, as a function of the 'specific' binding strength (left part) and of the 'total' binding strength (right part) of all spiked-in probes. The data arrange well along the *Langmuir*-master curve in the right part of the figure (see lines). Note that about 18% of all PM data but only about 9% of all MM data refer to a surface coverage bigger than $\theta > 0.1$ (see hatched areas). The two master curves in the left part refer to two values of the limiting non-specific coverage ($\theta^{\mathrm{P}}|_{[\mathrm{S}]\to 0} \approx X^{\mathrm{P,NS}} = 0.025$ and 0.0). The difference between the coverages of the PM and MM probes of each pair, $\Delta\theta_p = \theta_p^{\mathrm{PM}} - \theta_p^{\mathrm{MM}}$, becomes negative for a fair number of probe pairs (red circles). The inset in the right part shows the same data in a linear scale. An additional master curve, $X/(1 + 1.5 \cdot X)$, is plotted in the right lower panel which better fits the intensity data in the range of saturation (see text).

section show that the almost identical sequences of the PM and MM probes of one pair bind non-specific transcripts with essentially identical affinity. The subtraction of the MM from the PM intensity is therefore expected to remove this 'chemical background'.

The PM–MM intensity difference directly provides the difference of surface coverage, $\Delta\theta_p = \frac{I_p^{\mathrm{PM}} - I_p^{\mathrm{MM}}}{I^{\max}}$, which is given to a good approximation for $X_p^{\mathrm{P,h}} < 1$ by (see equation (2.1))

$$\begin{aligned}\Delta\theta_p \equiv \theta_p^{\mathrm{PM}} - \theta_p^{\mathrm{MM}} &= \frac{X_p^{\mathrm{PM}}}{1 + X_p^{\mathrm{PM}}} - \frac{X_p^{\mathrm{MM}}}{1 + X_p^{\mathrm{MM}}} \\ &= \frac{X_p^{\mathrm{PM}} - X_p^{\mathrm{MM}}}{1 + (X_p^{\mathrm{PM}} + X_p^{\mathrm{MM}}) + X_p^{\mathrm{PM}} \cdot X_p^{\mathrm{MM}}} \approx \frac{\Delta X_p}{1 + \Sigma X_p} \\ &\text{with } \Delta X_p \equiv X_p^{\mathrm{PM}} - X_p^{\mathrm{MM}} = \Delta K_p^{\mathrm{S}} \cdot [\mathrm{S}] + \Delta X_p^{\mathrm{NS}}, \\ &\Sigma X_p \equiv X_p^{\mathrm{PM}} + X_p^{\mathrm{MM}} = \Sigma K_p^{\mathrm{S}} \cdot [\mathrm{S}] + \Sigma X_p^{\mathrm{NS}}, \\ &\Delta K_p^{\mathrm{S}} \equiv K_p^{\mathrm{PM,S}} - K_p^{\mathrm{MM,S}} \qquad \text{and} \qquad \Sigma K_p^{\mathrm{S}} \equiv K_p^{\mathrm{PM,S}} + K_p^{\mathrm{MM,S}}. \end{aligned} \tag{3.1}$$

Equation (3.1) predicts a nearly *Langmuir*-type behaviour for $0 < \Delta X_p, \Sigma X_p \ll 1$ if one plots $\Delta\theta$ as a function of the difference ΔX, i.e., $\Delta\theta = \Delta X/(1 + \Delta X)$ (see figure 4, panel below). This approximation partly fails at abscissa values near unity where it underestimates saturation. The modified function $\Delta\theta = \Delta X/(1 + 1.5 \cdot \Delta X)$ better fits the data (compare the curves in the right, lower panel in figure 4).

Note that the *Langmuir*-isotherms of the PM and MM probes provide exclusively positive values of the surface coverage whereas their difference becomes negative at small abscissa values $\Delta X_p^{\mathrm{PM-MM}} < 10^{-2}$ for a fair number of probe pairs (see red symbols and insertion in figure 4).

3.5. Absolute RNA concentrations from microarray experiments

The hybridization isotherm relates the probe intensity to the absolute RNA concentration on the chip. The inversion of equation (2.1) with $a_p^{\mathrm{P}} = 1$ consequently enables the estimation of the RNA target concentration as a function of the probe intensities,

$$[\mathrm{S}]_p^{\mathrm{P}} = \frac{1}{K_p^{\mathrm{P,S}}} \left(\frac{I_p^{\mathrm{P}}}{1 - \theta_p^{\mathrm{P}}} - X_p^{\mathrm{P,NS}} \right); \qquad \mathrm{P} = \mathrm{PM}, \mathrm{MM}, \tag{3.2}$$

and of the PM–MM intensity difference (see equation (3.1))

$$[\mathrm{S}]_p^{\mathrm{PM-MM}} = \frac{\Delta I_p - (\Delta X_p^{\mathrm{NS}} - \Delta\theta_p \cdot \Sigma X_p^{\mathrm{NS}})}{(\Delta K_p^{\mathrm{S}} - \Delta\theta_p \cdot \Sigma K_p^{\mathrm{S}})}, \tag{3.3}$$

presuming that the binding parameters $K_p^{\mathrm{P,S}}$ and $X_p^{\mathrm{P,NS}}$ are known for each PM and MM probe. We used the results of the *Langmuir*-fits to calculate the target concentrations from the intensity data of the spiked-in probes.

The probes of each probe set refer to one target concentration. Consequently, the summarization of the apparent concentrations $[\mathrm{S}]_p^{\mathrm{P}}$ of the set into one value is expected to improve the precision of the estimated concentration value. Following previous studies in their basic ideas [2, 3] we tested several summarization algorithms: the arithmetic mean with different weighting functions and, alternatively, the median, $\mathrm{med}([\mathrm{S}]_p^{\mathrm{P}})_{\mathrm{set}}$, without and with the gradual removal of outliers using the Turkey biweight correction [17]. The latter approach provides the best results in terms of precision and accuracy, i.e., in terms of the minimum sum of squared residuals and of the systematic bias compared with the known spiked-in concentrations, respectively (see also next section). For a detailed comparison of different summarization algorithms which merge the probe-specific concentrations into one value per probe set see [3].

Figure 5 correlates the obtained concentration estimates for each probe set with the respective 'true' spiked-in concentrations. The PM and the MM probe intensities and also their difference are obviously suitable input data to estimate the concentration of specific transcripts. The obtained concentration estimates scatter reasonably about the diagonal line, which illustrates the respective target value of the concentration. The averaged concentration estimate of all considered spiked-in probe sets, $\langle [S]^P \rangle$, is shown as the black curve in figure 5. It closely fits to the target value nearly over the whole range of spiked-in concentrations.

3.6. Accuracy and precision

The degree of systematic agreement between the estimated concentrations and the respective true value, [S], defines the accuracy of the method. It reveals inconsistencies of the chosen *Langmuir*-model [1], of the weighting function used in the least-square fits (see above and [6]) and/or of the particular algorithm which aggregates the probe-specific concentrations into one set-value [3]. Also systematic errors of the spiked-in experiment can cause systematic deviations between the measured and calculated values. We used the ratio between the estimated mean and the target value, $\langle [S]^P \rangle / [S]$, as a measure of the accuracy for each spiked-in concentration (see figure 6, part (a)).

In contrast, the precision estimates the noise level of the calculated concentration owing to random errors. It is inversely related to the scattering width of the probe-set values about their mean and thus to the standard deviation of the concentration values, $\mathrm{SD}([S]^P)$. Part (b) of figure 6 shows the coefficient of variation, $\mathrm{CV}([S]^P) = \mathrm{SD}([S]^P)/[S]$, as a measure of the precision of the calculated concentrations.

Ideally, the accuracy and precision of the estimated concentration are best for $\langle [S]^P \rangle / [S] \approx 1$ and $\mathrm{CV}([S]^P) \approx 0$, respectively. It turns out that the accuracy (figure 6, part (a)) and precision (part (b)) of the calculated concentration are close to their ideal values in the intermediate concentration range, 10 pM $<$ [S] $<$ 100 pM. The data indicate a loss of accuracy as well as of precision at large and small [S]-values. In the former situation the probes become increasingly insensitive to changing RNA concentration owing to the saturation of the binding reaction. In the limit of small [S] essentially two effects cause the loss of accuracy and precision: firstly, specific binding progressively interferes with non-specific binding and, secondly, fluctuations of the optical background increasingly distort the obtained intensity values (see supplementary material in [6]).

The comparison of the three considered concentration estimates extracted from the PM and MM intensities and from their difference shows that the MM provide relatively imprecise concentrations especially in the limit of small [S] (figure 6, part (b)). On the other hand, the PM–MM and the PM-only measures are of similar quality.

Note that all concentration estimates consider the effect of non-specific hybridization and of saturation according to the *Langmuir*-model (see equations (2.1) and (3.1)–(3.3)). In contrast, several state-of-the-art algorithms of gene expression analysis ignore saturation and/or assume an equal affinity of PM and MM probes for non-specific hybridization [13, 18, 19]. To estimate the effect of these simplifications on the accuracy and precision of the obtained concentration values we applied equation (3.3) for three special cases. (i) Equal background of the PM and MM ($\Delta X_p^{\mathrm{NS}} = 0$) and no bright MM. The latter assumption was handled by the substitution $\Delta I_p^{\mathrm{P}} \rightarrow |\Delta I_p^{\mathrm{P}}|$. (ii) $\Delta X_p^{\mathrm{NS}} = 0$ as in (i), but no modification of ΔI_p^{P}, i.e., intensity values referring to bright MM are processed without modifications. (iii) $\Sigma K_p^{\mathrm{S}} = \Sigma X_p^{\mathrm{NS}} = 0$, i.e., saturation is ignored. The accuracy and precision of these approximations are compared with that of the *Langmuir*-model in parts (c) and (d) of figure 6.

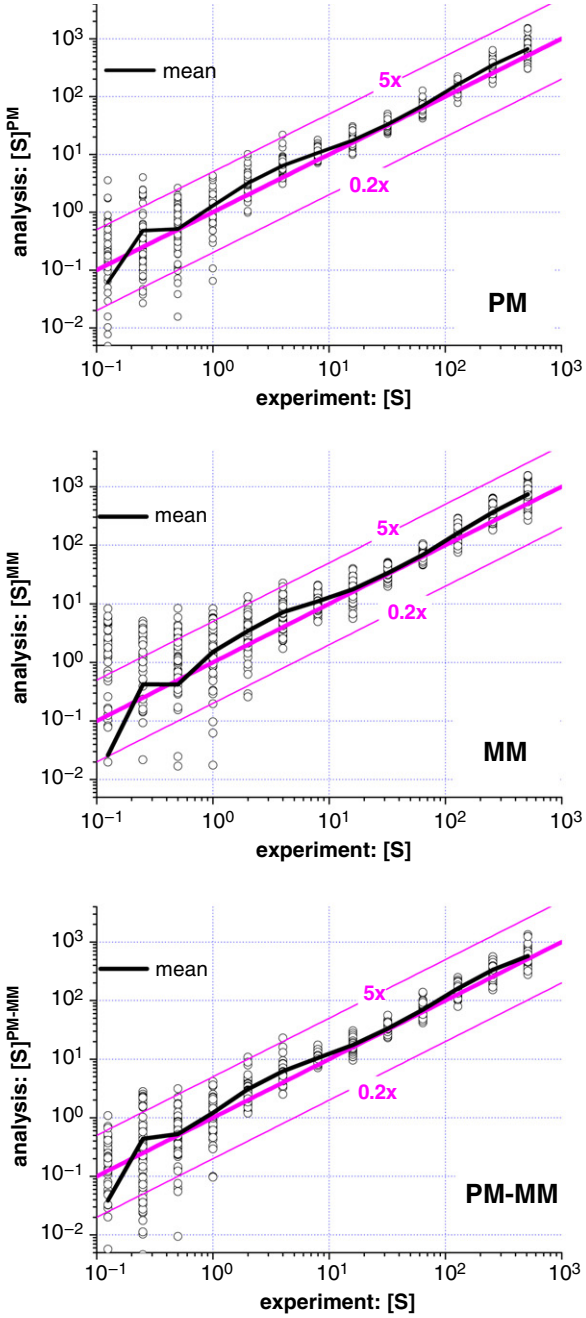

Figure 5. Correlation plot between the concentration estimates using the *Langmuir*-adsorption model and the 'true' concentrations for the spiked-in probes. The individual concentration values of each probe are aggregated into one value per probe set using their median combined with Turkey biweight outlier removal (see text). The diagonal line refers to $[S]^P = [S]$, i.e. the complete agreement between calculated and predetermined concentration values. The two satellite lines refer to the five-fold deviation from the diagonal, i.e., $[S]^P = 5 \cdot [S]$ and $[S]^P = [S]/5$. The concentration estimates are calculated using the PM and MM probe intensities of the spiked-in probes and their difference PM–MM (equations (3.2) and (3.3), respectively). The line shows the mean of all spiked-in probe sets. Note the excellent agreement between the estimated and the 'true' concentrations.

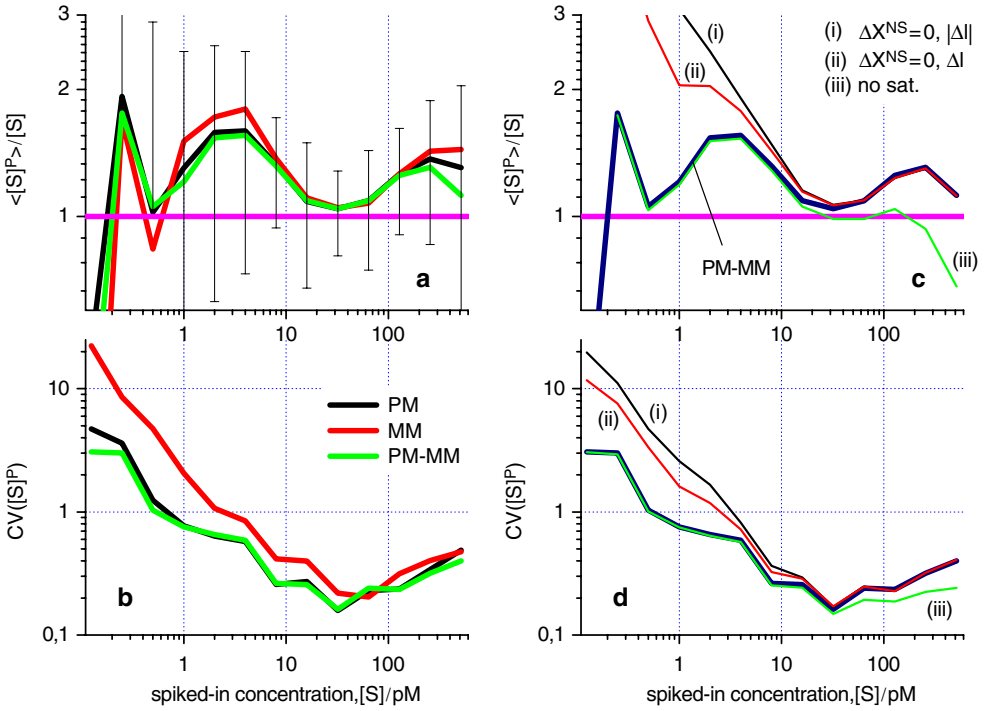

Figure 6. Accuracy (part (a)) and precision (part (b)) of the concentration estimate of the PM and MM intensities and of the PM–MM intensity difference as a function of the spiked-in concentration (see also figure 5). For the accuracy measure we use the ratio between the mean concentration estimate of all spiked-in probe sets and the 'true' spiked-in concentration. The precision was estimated using the coefficient of variation, $\mathrm{CV}([S]^P) = \mathrm{SD}([S]^P)/[S]$. The error bars in part (a) indicate the standard deviation of the PM data. Note that $\langle[S]^P\rangle/[S]$ and $\mathrm{CV}([S]^P)$ are minimum and thus the accuracy and the precision are maximum in the intermediate concentration range, $10 < [S] < 100$ pM. The PM and PM–MM estimates clearly outperform $[S]^{MM}$. Parts (c) and (d) illustrate the effect of neglecting the residual non-specific background (curves (i) and (ii)) and of saturation (curve (iii)) of the PM–MM intensity difference (see the text). The thick line was the PM–MM curve which was replotted from parts (a) and (b) for comparison, respectively. It was calculated using equation (3.3) and takes into account the residual background and saturation (see the text).

The neglection of the residual background according to approximations (i) and (ii) gives rise to considerably less accurate and less precise concentration estimates. This result clearly demonstrates that the subtraction of the MM intensity does not completely remove the background contribution from the respective PM intensity. This effect can be corrected by adequate analysis using equation (3.3). This way it outperforms methods (i) and (ii) which ignore the background at small [S]-values.

Neglecting saturation (case (iii)), as expected, systematically underestimates the concentration values at $[S] > 10$ pM. It is interesting to note that this trend already starts at relatively small concentrations owing to high affinity probes. Note that neglecting saturation seems to improve the precision. The respective $\mathrm{CV}([S]^P)$-values are clearly minimum among the tested algorithms at high concentration ($[S] > 50$ pM) (figure 6, part (d)). This effect can however be trivially explained by the decreased sensitivity of the probes for concentration changes upon saturation (see above and [4]).

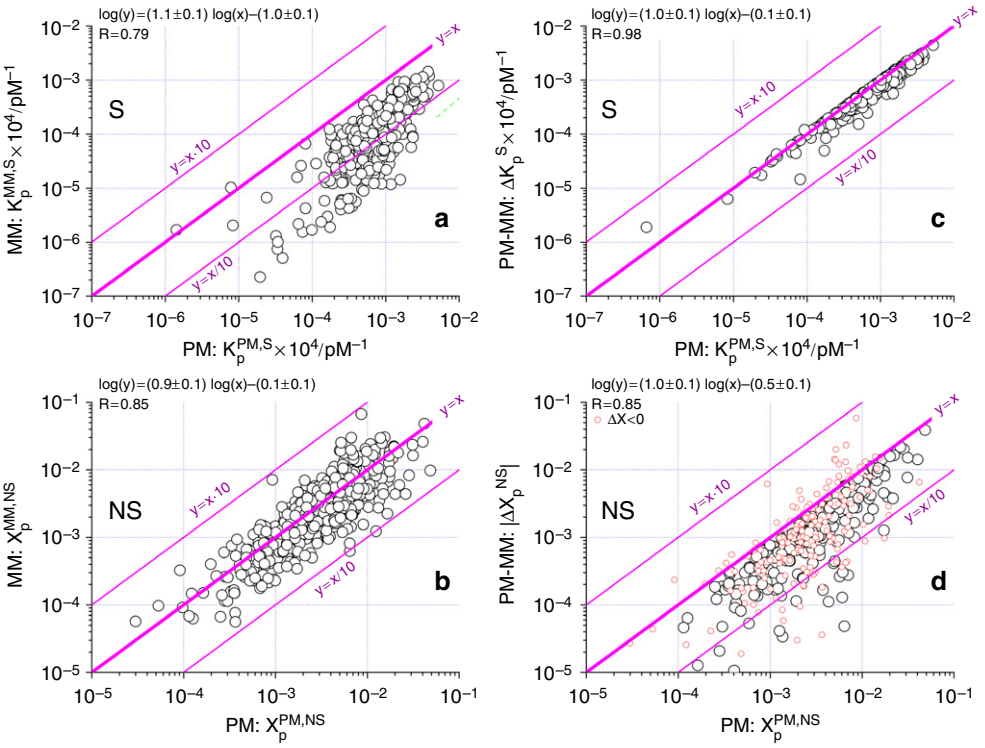

Figure 7. Correlation plot between the binding strengths of the PM and the MM probes (panels (a) and (b)) and of the PM probes and the PM–MM difference (panels (c) and (d)). The linear regression in the double-logarithmic plot provides slopes near unity and regression coefficients of $R \geqslant 0.79$ (see figure). The diagonal lines refer to $y = x$, $y = 10x$ and $y = 0.1x$ (see figure). They serve as a guide for the eye to compare the affinities. The binding constants of specific hybridization of the MM are clearly smaller than that of the PM for nearly all probe pairs (panel (a)). In contrast, the binding strength of non-specific hybridization of the MM are either smaller or bigger (about 40%) than that of the PM (panel (b)). For the PM–MM difference one obtains nearly as large values (specific binding, panel (c)) or distinctly smaller (non-specific binding, panel (d)) values compared with that of the PM.

4. Sequence-specific affinity of matched and mismatched DNA/RNA duplexes

4.1. The binding constants of the PM and MM probes

The sequences of the paired PM/MM probes are identical except the middle-base. One expects therefore a strong correlation between the binding affinities of the PM and MM probes. Indeed, the respective correlation plots of the binding constants of specific and non-specific hybridization reveal a high degree of correlation between both data sets (figure 7, parts (a) and (b)). Linear regression in the double-logarithmic scale provides slopes near unity and regression coefficients larger than $R > 0.79$ (see figure 7). The binding data for non-specific hybridization scatter about the diagonal line ($y = x$, figure 7, part (b)). This result illustrates that the insertion of a mismatched base on the average only weakly affects the propensity of the probe for non-specific hybridization. This result is further confirmed by the respective frequency distributions of the $X_p^{P,NS}$ which cover a broad, nearly identical range for the P = PM and MM probes (figure 8, lower panel).

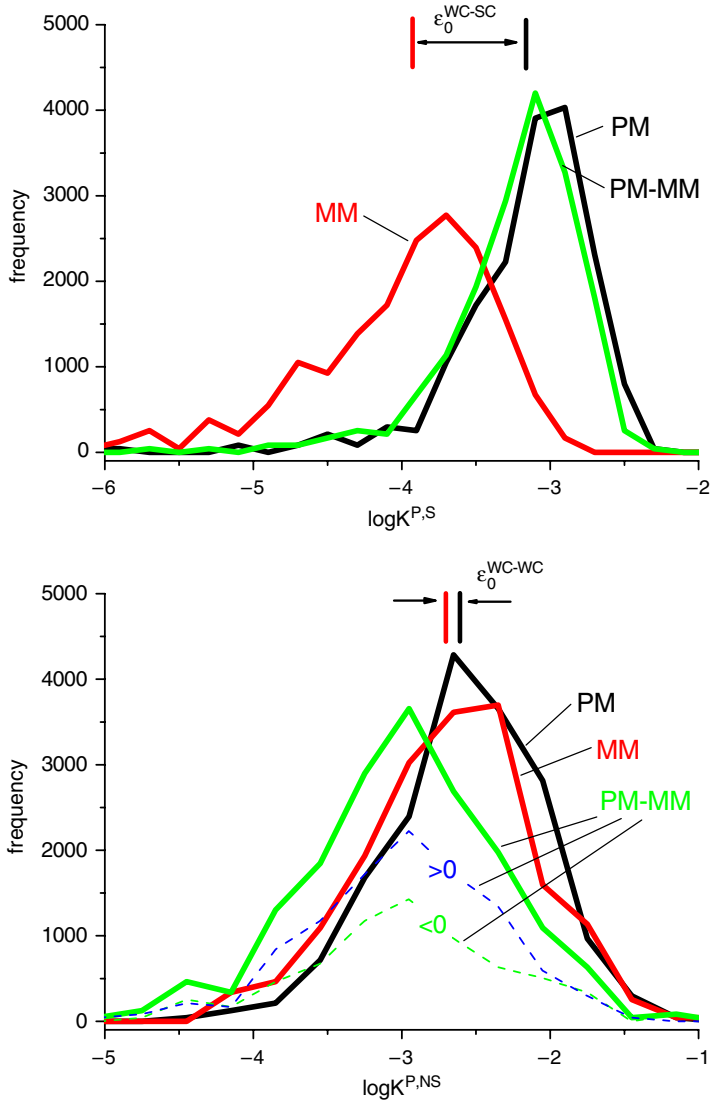

Figure 8. Frequency distribution of the logarithm of the binding constant of specific hybridization, $\log K_p^{\mathrm{P,S}}$ (upper panel), and of the binding strength of non-specific hybridization, $\log X_p^{\mathrm{P,NS}}$ (lower panel), which were obtained from the fit of the *Langmuir*-isotherm to the spiked-in data. The vertical lines indicate the mean values $\langle \log K_p^{\mathrm{P,S}} \rangle$ and $\langle \log X_p^{\mathrm{P,NS}} \rangle$ for P = PM, MM. Their differences, $\varepsilon_0^{\mathrm{WC-W}}$ (W = WC, SC), estimate the apparent mean free energy difference between WC pairings in non-specific duplexes of the PM and MM probes and between the WC pairing in the specific duplex of the PM and the self-complementary pairings in the specific duplexes of the MM (see text and figure 12 below for illustration). The log-distribution of the differences, $\log \Delta K_p^{\mathrm{S}} = \log(K_p^{\mathrm{PM,S}} - K_p^{\mathrm{MM,S}})$ and $\log \Delta X_p^{\mathrm{NS}} = \log(|X_p^{\mathrm{PM,NS}} - X_p^{\mathrm{MM,NS}}|)$ are indicated by 'PM–MM'. The affinity constants of specific hybridization provide nearly exclusively positive $\Delta K_p^{\mathrm{PM-MM,S}}$-values. In contrast, about 40% of the ΔX_p^{NS}-values are negative. The frequency distributions for $0 > \Delta X_p^{\mathrm{NS}}$ and $0 < \Delta X_p^{\mathrm{NS}}$ (see the dashed curves in the lower panel) reveal an almost identical shape with its maximum near $\sim \pm 10^{-3}$.

Contrarily, the specific binding data are exclusively located below the diagonal line (figure 7, part (a)). That means, the mismatched base of the MM causes a systematically reduced binding affinity for specific target RNA compared with the affinity of the PM. The respective frequency distribution of the specific binding constant of the MM is consequently shifted distinctly towards smaller values (figure 8, upper panel). The difference between their maxima of $\varepsilon_0^{\mathrm{WC-SC}} \sim -0.85$ and the offset of the linear regression line show that $K_p^{\mathrm{PM,S}}$ exceeds $K_p^{\mathrm{MM,S}}$ on the average by nearly one order of magnitude.

Parts (c) and (d) of figure 7 correlate the PM–MM difference of the binding constants (y-axis) with that of the PM (x-axis). The difference of the specific constants, $\Delta K^{\mathrm{S}} = K^{\mathrm{PM,S}} - K^{\mathrm{MM,S}}$, is dominated by $K^{\mathrm{PM,S}}$ owing to the relation $K^{\mathrm{PM,S}} \gg K^{\mathrm{MM,S}}$. The respective frequency distributions of ΔK^{S} and of $K^{\mathrm{PM,S}}$ are nearly identical (figure 8, upper panel). Hence, the intensity difference between the PM and MM responds nearly as sensitively to changes of the concentration of specific transcripts as the PM intensity.

The situation changes considerably for the affinity of non-specific transcripts. The subtraction of the MM values reduces the background term by roughly a factor of 3–5 compared with that of the PM. The PM–MM difference, $\Delta X^{\mathrm{NS}} = X^{\mathrm{PM,NS}} - X^{\mathrm{MM,NS}}$, correlates well with $X^{\mathrm{PM,NS}}$.

The shape of the frequency distribution of $\log|\Delta X_p^{\mathrm{NS}}|$ is virtually identical for $\Delta X_p^{\mathrm{NS}} > 0$ and $\Delta X_p^{\mathrm{NS}} < 0$ (see figure 8, lower panel). The maxima of both distributions are shifted with respect to the coordinate origin, $\Delta X_p^{\mathrm{NS}} = 0$. This result indicates that ΔX_p^{NS} does not randomly fluctuate about the origin. Instead, it divides into two populations centred about $\sim+10^{-3}$ and $\sim-10^{-3}$. The integral of the frequency distributions shows that about 40% of all probe pairs refer to $\Delta X_p^{\mathrm{NS}} < 0$, i.e. to 'bright' MM intensities which exceed the respective PM value at vanishing specific hybridization.

Hence, the absolute value of the limiting adsorption at vanishing specific hybridization obtained from the PM–MM intensity difference is reduced by about one half an order of magnitude compared with the non-specific background level of the individual PM and MM intensities despite the 'bright MM' effect (see also figure 8, lower panel). The original intention of using MM probes to correct the PM intensity for the chemical background seems to be justified in a first-order approximation.

4.2. Specific and non-specific binding

The sensitivity of a probe is directly related to the binding constant for specific transcripts, $K_p^{\mathrm{P,S}}$ whereas its specificity is inversely related to the ratio $r_p^{\mathrm{P}} = K_p^{\mathrm{P,NS}}/K_p^{\mathrm{P,S}} \propto X_p^{\mathrm{P,NS}}/K_p^{\mathrm{P,S}}$ (see the accompanying paper [1]). The quality of a probe in terms of sensitivity and specificity consequently increases in the $\log X_p^{\mathrm{P,NS}}$-versus-$\log K_p^{\mathrm{P,S}}$ plots shown in figure 9 with increasing abscissa and decreasing ordinate along the $y = -x$ diagonal. Note that the ordinate $\log X_p^{\mathrm{P,NS}} = \log(K_p^{\mathrm{P,NS}} \cdot [\mathrm{NS}])$ is directly related to $K_p^{\mathrm{P,NS}}$ because [NS] is virtually a constant in the spiked-in experiment (see also equation (2.1)).

The grey cross in figure 9 serves as a guide for the eye to identify 'good' probes, for example, in the lower-right quadrant. For the MM the number of probes inside this region distinctly decreases compared with that of the PM whereas the number of the respective data points for the PM–MM intensity difference increases. Hence the latter intensity measure outperforms the PM intensity which in turn outperforms the MM intensity in terms of the chosen criteria. Note that the PM–MM intensity difference provides a clearly better specificity (i.e., a smaller $\log \Delta K_p^{\mathrm{NS}} \propto \log \Delta X_p^{\mathrm{NS}}$) and a nearly as high sensitivity (i.e., $\log \Delta K_p^{\mathrm{S}}$) compared with those of the PM.

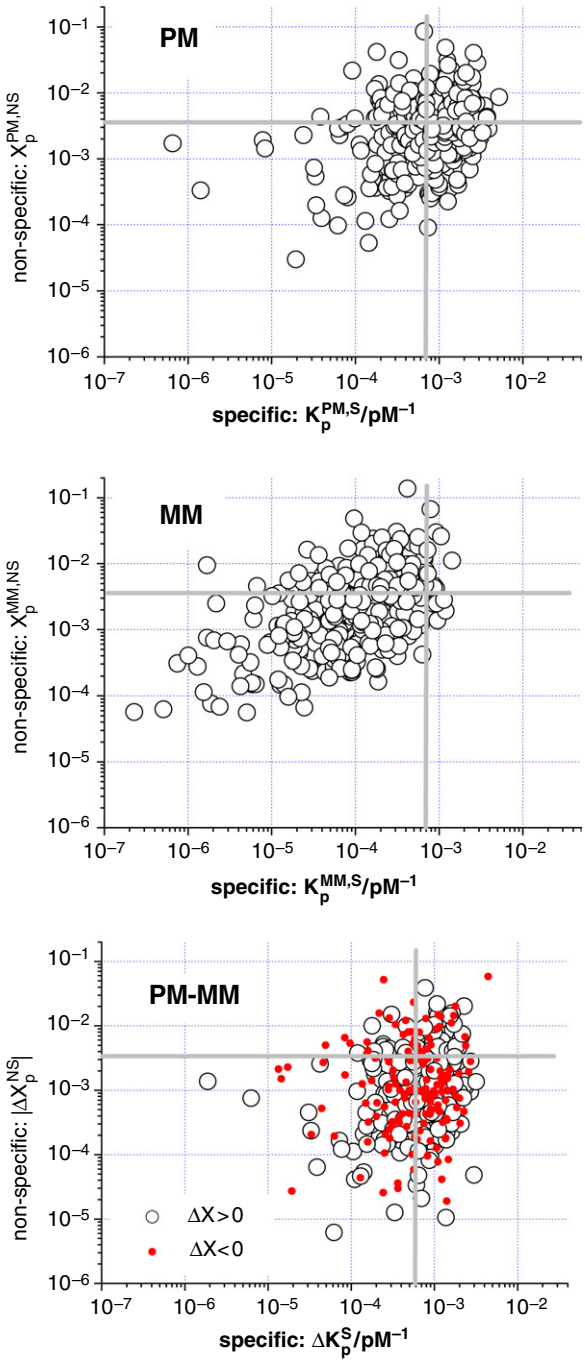

Figure 9. Correlation plot between the binding strengths of non-specific (y-axis) and of specific (x-axis) hybridization of the PM and MM probes and of the respective PM–MM differences obtained from the *Langmuir*-fits of the spiked-in data (large symbols). The absolute value of the negative ΔX_p^{NS} data are shown by solid circles in the panel below. The grey cross refers to the log-means, $\langle \log K_p^{\mathrm{PM,S}} \rangle$ and $\langle \log X_p^{\mathrm{PM,NS}} \rangle$, of all PM probes. It is drawn in all parts of the figure to serve as a guide for the eye to compare the position of the data clouds. Probes of 'good' quality are located in the lower-right quadrant (see text).

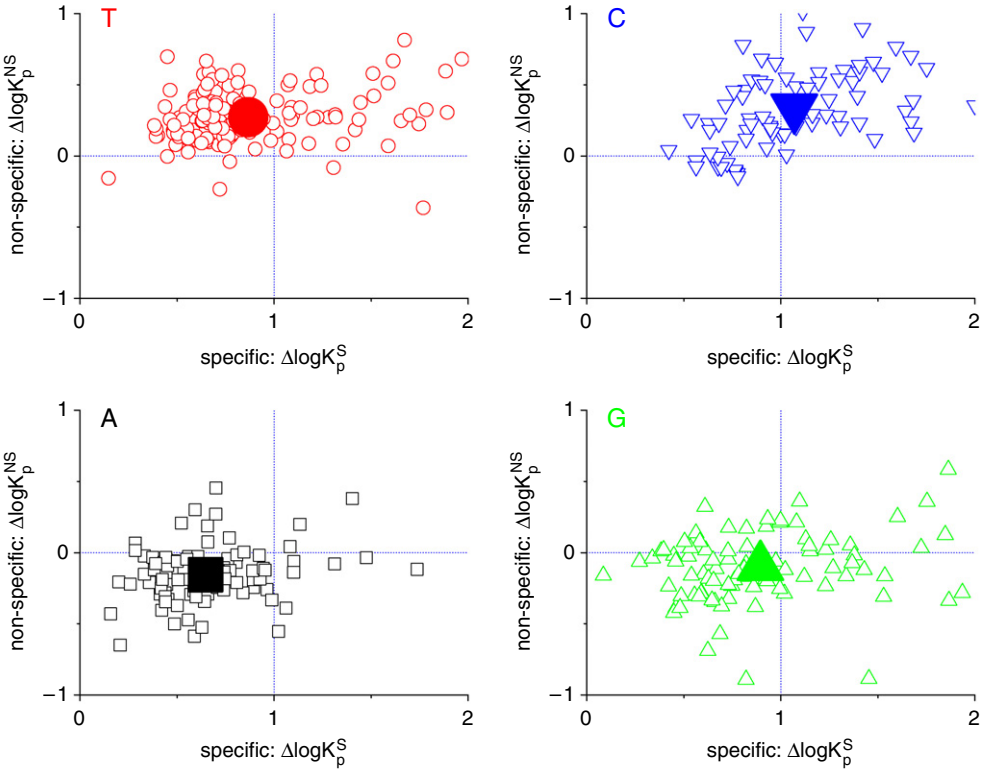

Figure 10. Correlation plot of the log-difference of the binding constants of non-specific (y-axis) and of specific (x-axis) hybridization between the PM and the MM probes, $\Delta \log K_p^{\mathrm{h}} = \log K_p^{\mathrm{PM,h}} - \log K_p^{\mathrm{MM,h}}$. Each panel shows the data for probe pairs with one common middle base of the PM (B = A, T, G, C; see the figure for assignments). The large solid symbols refer to the respective log-averages, $\langle \Delta \log K_p^{\mathrm{h}} \rangle_{\mathrm{B}}$. Note that pyrimidines (C, T) predominantly provide positive values along the ordinate whereas purines (A, G) are predominantly negative. Contrarily, the differences of the values referring to specific hybridization are always positive.

The chosen criteria allow the selection of high-quality probes provided that the constants for specific and non-specific binding are known. The question whether the binding constants and thus the probe quality can be predicted on the basis of their sequence will be addressed below.

4.3. The middle-base related bias of probe affinities

It is well established that the middle base at position $k = 13$ of the 25meric probe systematically affects the relation between the PM and MM probe intensities [20]. For probe pairs with single-ringed pyrimidines (C, T) in the middle of the PM sequence one finds a preference for 'bright' PM, $I^{\mathrm{PM}} > I^{\mathrm{MM}}$. For double-ringed purines (G, A) the relation reverses with the tendency for 'bright' MM.

To study the effect of specific and non-specific hybridization on the middle-base related bias we correlated the PM/MM log-ratio of the hybridization constants for non-specific with those for specific hybridization separately for all PM/MM-probe pair with the middle base B = A, T, G and C of the PM (figure 10). Note that the transformation

$\Delta \log(X_p^{\mathrm{NS}}) \equiv \log(X_p^{\mathrm{PM,NS}}/X_p^{\mathrm{MM,NS}})$ cancels out the concentration of non-specific transcripts, i.e., $\Delta \log(K_p^{\mathrm{NS}}) = \Delta \log(X_p^{\mathrm{NS}})$ (note that $X_p^{\mathrm{P,NS}} = K_p^{\mathrm{P,NS}}[\mathrm{NS}]$). The obtained data clouds are shifted relative to each other showing a characteristic pattern with respect to the coordinate axes. To reveal this trend more clearly, we replotted the mean values for each middle base, $\langle \Delta \log(K_p^{\mathrm{h}}) \rangle_{\mathrm{B}}$ (h = S, NS), into one coordinate system (figure 11, upper panel). It turns out that the mean log-difference of the non-specific association constants splits into a doublet along the ordinate according to $\mathrm{A} \approx \mathrm{G} < \mathrm{T} \approx \mathrm{C}$ (see the horizontal bars in figure 11) whereas the mean log-difference of the specific association constants reveals a triplet-like pattern along the abscissa, $\mathrm{A} < \mathrm{T} \approx \mathrm{G} < \mathrm{C}$ (see the vertical bars in figure 11).

The comparison of the binding constants of the PM and MM probes reveals a second interesting result: the $\Delta \log(K_p^{\mathrm{NS}})$-values are either positive (C, T) or negative (A, G) whereas the $\Delta \log(K_p^{\mathrm{S}})$ are always positive (figure 10). That means, the effect of 'bright' MM ($\Delta \log(K_p^{\mathrm{NS}}) < 0$) is exclusively related to non-specific binding whereas specific hybridization of the PM is always stronger than that of the respective MM. Note that $K_p^{\mathrm{PM,S}}$ exceeds $K_p^{\mathrm{MM,S}}$ on the average nearly by a factor of ten. This result seems to disagree with previous studies which however considers all probes of the chip without differentiation between specific and non-specific hybridization [9]. We recently showed that the fraction of specifically hybridized probes on these arrays is relatively small [7]. The high content of bright MM (about 40%) could be clearly assigned to probes which are predominantly hybridized with non-specific transcripts whereas specific hybridization gives exclusively rise to 'bright' PM (see [6, 7] for details).

4.4. Base-pairings in specific and non-specific duplexes of the PM and MM

The results reported in the previous section can be interpreted in terms of the base pairings that stabilize the DNA/RNA duplexes (see figure 12 for illustration and also [7, 8]). The PM probes 'per definition' form exclusively Watson–Crick (WC) pairs with the complementary sequence of the specific target RNA. The central WC pair of the PM, $\mathrm{B} \bullet \mathrm{b}^c$ (the left part of figure 12 shows for example $\mathrm{G} \bullet \mathrm{c}^*$), is replaced by the respective self-complementary (SC) pair, $\underline{\mathrm{B}}^c \bullet \underline{\mathrm{b}}^c$ (e.g. $\underline{\mathrm{C}} \bullet \underline{\mathrm{c}}^*$) in the respective MM duplex. This substitution is accompanied by the free energy change $\varepsilon_{13}^{\mathrm{WC-SC}}(\mathrm{B}) = \varepsilon_{13}^{\mathrm{WC}}(\mathrm{B} \bullet \mathrm{b}^c) - \varepsilon_{13}^{\mathrm{SC}}(\underline{\mathrm{B}}^c \bullet \underline{\mathrm{b}}^c)$ if one assumes that the free energy contribution of all other base pairings at positions $k = 1 \ldots 12$ and $14 \ldots 25$ remains unchanged.

The log-difference of the specific association constant consequently estimates the change of the apparent free energy upon the replacement of the WC pairing by the respective SC pairing, i.e. $-\langle \Delta \log K^{\mathrm{S}} \rangle_{\mathrm{B}} \approx \varepsilon_{13}^{\mathrm{WC-SC}}(\mathrm{B})$. The observed triplet-like relation of the free energy differences can be understood if one considers three energetic states for the WC pairing according to $\mathrm{C} \bullet \mathrm{g} > \mathrm{G} \bullet \mathrm{c}^* \approx T \bullet \mathrm{a} > \mathrm{A} \bullet \mathrm{u}^*$, and essentially one free energy of the SC pairing, $\underline{\mathrm{C}} \bullet \underline{\mathrm{c}}^* \approx \underline{\mathrm{G}} \bullet \underline{\mathrm{g}} \approx \underline{\mathrm{T}} \bullet \underline{\mathrm{u}}^* \approx \underline{\mathrm{A}} \bullet \mathrm{a}$ (see the right part of figure 12). Note that the stabilities of the WC pairings are asymmetrical with respect to purines and pyrimidines in the DNA sequence of the hybrid duplex: DNA • RNA pairings of the type pyrimidine • purine are more stable than purine • pyrimidine [5, 7, 21, 22].

The situation changes for the non-specifically hybridized probes. The 'NS-background' represents a mixture of RNA fragments with a broad distribution of base compositions, which enables on the average the formation of a sufficient number of WC pairings with the PM and the MM probes as well. The middle bases of the PM and of the MM are therefore mainly stabilized by WC pairings. Note however, that they reverse the direction if one compares the PM with the MM owing to the complementary character of their middle bases. The $\mathrm{B} \bullet \mathrm{b}^c$-pairing in the duplexes of the PM becomes $\mathrm{B}^c \bullet \mathrm{b}$ in the respective duplex of the MM. The left part of figure 12 illustrates this situation for $\mathrm{G} \bullet \mathrm{c}^* \rightarrow \mathrm{C} \bullet \mathrm{g}$.

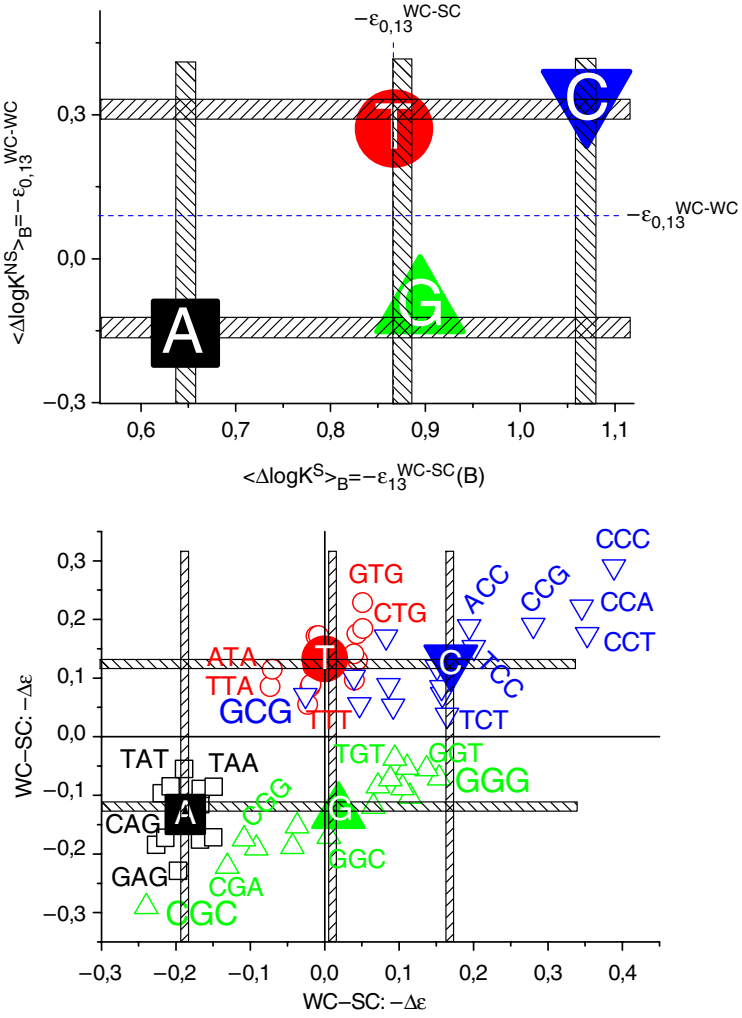

Figure 11. Correlation between the middle-base averaged PM–MM log-differences of the bindings strengths of non-specific and specific binding. The data in the upper panel are replotted from figure 10 into one coordinate system. This representation shows that the data sort according to the duplet C ≈ T > G ≈ A along the ordinate and according to the triplet C > G ≈ T > A along the abscissa as illustrated by the horizontal and vertical bars, respectively. Note that these data were obtained from the spiked-in data set of 462 probes. The lower panel shows the respective middle-base related data which were obtained from sub-ensembles of the full set of ∼250.000 probe pairs per chip (see text and also figure 14). They are shown as the incremental free energy difference, $\Delta\varepsilon_{0,13}^{\mathrm{WC-W}}$ (B) with W = WC, SC referring to the substitutions of complementary WC pairings ($\mathrm{B}^c \bullet \mathrm{b} \rightarrow \mathrm{B} \bullet \mathrm{b}^c$) and of a WC and a SC pairing ($\underline{\mathrm{B}}^c \bullet \underline{\mathrm{b}}^c \rightarrow \mathrm{B} \bullet \mathrm{b}^c$), respectively (solid symbols, see text). The open symbols are the respective middle-triple related free energy differences, $\Delta\varepsilon_{0,12}^{\mathrm{WC-W}}(\mathrm{B}'\mathrm{B}\mathrm{B}'')$ which consider additionally the nearest neighbours of the middle base. Selected middle triples are assigned in the figure.

The three energetic states of the WC pairings (see above) produce a duplet for the difference, $\varepsilon_{13}^{\mathrm{WC-WC}}(\mathrm{B}) = \varepsilon_{13}^{\mathrm{WC}}(\mathrm{B} \bullet \mathrm{b}^c) - \varepsilon_{13}^{\mathrm{WC}}(\mathrm{B}^c \bullet \mathrm{b})$, because the free energy changes for the replacements C•g → G•c* and T•a → A•u* are nearly equal. The probe pairs consequently

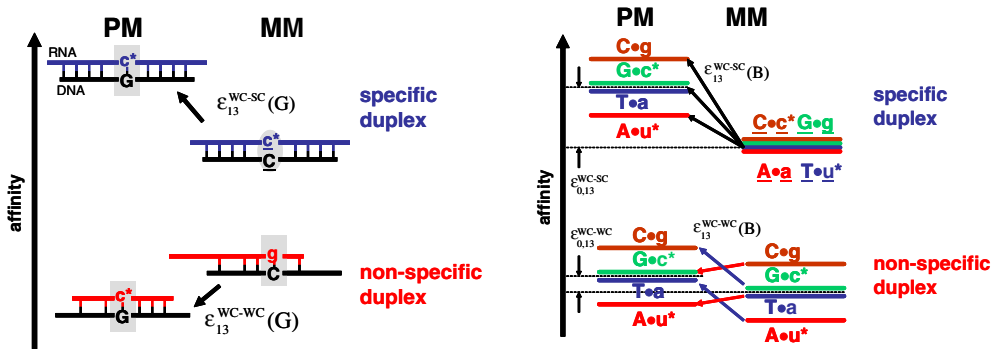

Figure 12. Schematic illustration of the base pairings in the specific (S) and non-specific (NS) duplexes of the PM and MM probes of a probe pair with a G in the middle of the PM sequence (left panel). The stability of the duplexes increases along the vertical affinity scale. The middle base forms the WC pairing G • c* in the specific and non-specific duplex of the PM. In the non-specific dimers of the MM the WC pair reverses direction, C • g. In the specific duplex of the MM the middle base forms the self-complementary pairing, $\underline{C} \bullet \underline{c}^*$. Note that the non-specific duplexes are less stable than the specific ones because they are stabilized by a smaller number of WC pairings as indicated by the smaller number of bonds. The right panel illustrates the stability of the respective duplexes with the middle bases A, T, G and C. The PM–MM difference of the specific and non-specific duplexes provides a triplet and a duplet of different values, respectively, as indicated by the arrows.

split into two fractions with purine (A, G) middle bases of the PM and preferentially bright MM ($\Delta \log K^{\mathrm{NS}} < 0$) and with pyrimidines (C, T) in the middle of the PM and the reverse relation ($\Delta \log K^{\mathrm{NS}} > 0$) due to the purine/pyrimidine asymmetry of interaction strengths in the DNA/RNA hybrid duplexes.

4.5. Positional-dependent base-pair interactions

The sequence of the 25meric probes mainly determines its affinity for duplex formation. Different bases along the sequence contribute differently to the free energy of duplex formation and thus to the binding constants of specific and non-specific hybridization. The contribution of a base pairing at position k of the sequence can be split into a mean, base-independent and into a base-specific, incremental term according to $\varepsilon_k^{\mathrm{W}}(\mathrm{B}) = \varepsilon_{0,k}^{\mathrm{W}} + \Delta\varepsilon_k^{\mathrm{W}}(\mathrm{B})(\mathrm{W} = \mathrm{WC}, \mathrm{SC})$, where the superscript specifies the type of the pairing. The incremental term, $\Delta\varepsilon_k^{\mathrm{W}}(\mathrm{B})$, was estimated from the affinity constants for specific (S) and non-specific (NS) binding of the PM and MM spiked-in probes using the positional-dependent single-base (SB) model (equations (2.3) and (2.4)). This analysis consequently provides four data sets, S–PM, NS–PM, S–MM and NS–MM (see figure 13). According to the interpretation given in the previous section, all terms refer to WC pairings except the middle position ($k = 13$) of the S–MM profiles which form SC pairings (see arrow in figure 13).

The obtained $\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{B})$-values vary along the sequence in a characteristic, parabola-like fashion for B = C, A and almost monotonically for B = G, T (figure 13). Note that the relative free energy contributions nearly completely vanish towards the free end of the probes at sequence position $k = 25$ owing to, for example, zippering effects. Previous studies report a similar behaviour of SB terms [6, 9, 23, 24]. The novelty of the presented results is twofold: (i) here we explicitly analysed the binding constants whereas previous studies used the intensities

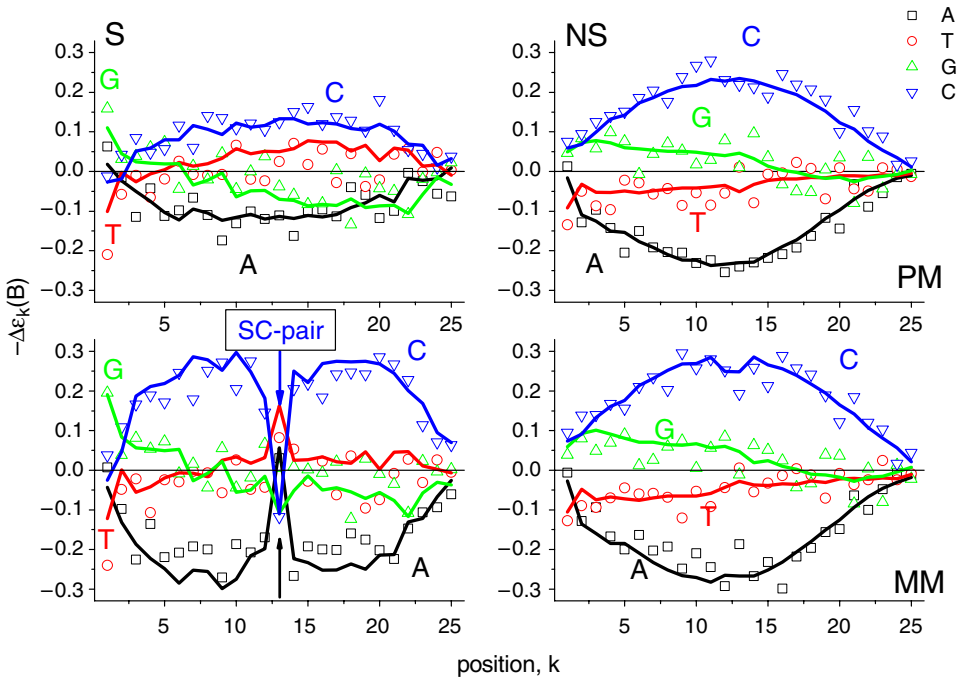

Figure 13. Positional-dependent single-base contributions to the apparent free energy of duplex formation of PM (upper panel) and MM (below) probes in specific (left) and non-specific (right) duplexes. Note the 'dent' in the middle of the specific C- and A-profiles of the MM probes. It is assigned to the respective SC pairings C–c* and A–a, respectively (see also figure 12 for explanation). The symbols are obtained from the fit of the SB model to the affinity constants, $K_p^{\mathrm{P,S}}$ and $X_p^{\mathrm{P,NS}}$ (P = PM, MM), of the *Langmuir* analysis of the spiked-in data (equation (2.3)). They are relatively noisy owing to the small data set of 462 spiked-in probe pairs. The lines refer to the profiles which are directly extracted from the probe intensities of selected subsets of all probes of the chip using a nearest-neighbour analysis (see equation (4.2) and the text).

of all or of subsets of the probes of the chip [9, 24]; (ii) accordingly, the SB terms were separately obtained for the four cases, S–PM, NS–PM, S–MM and NS–MM, whereas the previous profiles refer to weighted means, for example of S–PM and NS–PM [9].

The inspection of the S–MM profiles reveals that the free energy term referring to the SC pairing in the middle of the sequence distinctly deviates from the respective WC-values in the other profiles (see arrows in figure 13). Particularly, the free energy increments of the middle bases C and A markedly drop in an absolute scale to values near zero or even change their sign. That means, the binding affinity of the SC pairs is considerably less sensitive with respect to the middle base when compared with the WC pairings (see also [7] for a detailed discussion).

The results of the SB analysis confirm the results presented in the previous section. (i) Purines and pyrimidines are asymmetrical with respect to the binding free energy of the WC pairings, i.e., $-\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{C}) > -\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{G})$ and $-\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{T}) > -\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{A})$. (ii) The stability of the WC pairings increases according to $-\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{C}) > -\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{G}) \approx -\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{T}) > -\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{A})$ providing a triplet of energetic states. (iii) The stability of the mismatched SC pairing in the specific duplexes of the MM probes mostly lacks base specificity, $\Delta\varepsilon_k^{\mathrm{SC}}(\mathrm{C}) \approx \Delta\varepsilon_k^{\mathrm{SC}}(\mathrm{G}) \approx \Delta\varepsilon_k^{\mathrm{SC}}(\mathrm{T}) \approx \Delta\varepsilon_k^{\mathrm{SC}}(\mathrm{A})$. (iv) The remaining sequence positions at $k \neq 13$ show similar profiles for the PM and MM probes upon specific and non-specific hybridization.

The profiles of the PM for specific and non-specific hybridization (and also that of the MM for $k \neq 13$) are assumed to refer both to WC pairings. Detailed inspection reveals however subtle differences: for example, the profiles for T and G are shifted in vertical direction one to each other if one compares PM–NS and PM–S. Also the spread between the maximum and minimum is distinctly smaller in the S–PM profile. These differences possible reflect sequence-specific effects which are not adequately considered in the simple SB model such as special folding motifs and/or self-complementary dimers.

Note that the assumption of equal incremental free energy profiles, $\Delta\varepsilon_k^{\mathrm{WC}}(\mathrm{B})$, for specific and non-specific hybridization predict the strong correlation between the respective binding constants, $K_p^{\mathrm{P,S}}$ and $K_p^{\mathrm{P,NS}}$, respectively. The linear fits of the data in the $\log X_p^{\mathrm{P,NS}}$-versus-$\log K_p^{\mathrm{P,S}}$ scale however reveal only weak correlation with regression coefficients of $R < 0.2$ and <0.6 for the PM and MM probes, respectively (see figure 9). Our theoretical analysis of competing interactions on microarrays predicts a loss of correlation between non-specific and specific binding because the sensitivity of the probes for non-specific duplex formation is partly compensated by competing complexes such as folded monomers and partly self-complementary dimers [1]. The subtle differences between the respective free energy profiles possibly reflect such effects. The clarification of this issue requires further work to predict the probe quality from the sequence.

4.6. Nearest-neighbour interactions

Stacking interactions between nearest neighbours within the RNA and DNA sequences make an important contribution to the stability of probe–target duplexes [21, 25, 26]. The single-base related free energy parameters discussed in the previous section comprise the effect of neighbours in an averaged, base-independent fashion. They therefore should be judged as a rough, first-order approximation which can be improved by the explicit consideration of adjacent bases in the respective sequence. The data set of 462 spiked-in probes is unfortunately too small for the appropriate estimation of $16 \times 24 = 384$ positional-dependent nearest-neighbour (NN) parameters for the 25meric probes, $\Delta\varepsilon_k^{\mathrm{W}}(\mathrm{BB}')$ (BB' is the couple of adjacent bases at position k and $k+1$ of the probe sequence).

To extend the number of considered probes we make use of following result established previously (see [7, 8] for details): probe sets with a set-averaged intensity value of $\langle \log I^{\mathrm{PM}} \rangle_{\mathrm{set}} < 1.8$ are predominantly non-specifically hybridized whereas probes from sets with $\langle \log I^{\mathrm{PM}} \rangle_{\mathrm{set}} > 2.8$ are predominantly specifically hybridized. The respective number of relevant intensity values per chip is ∼90 000 for non-specifically, and ∼25 000 for specifically hybridized probes. Both ensembles of intensity data are sufficient to calculate the NN parameters for the four cases S–PM, NS–PM, S–MM, NS–MM.

A second problem arises from the fact that the concentration of specific transcripts is not explicitly known for these probe sets and thus the determination of the affinity constants by the fit of equation (2.1) cannot be applied. Alternatively we transform the probe intensities according to

$$Y_p^{\mathrm{P}} = \log I_p^{\mathrm{P}} - \langle \log I^{\mathrm{P}} \rangle_{\mathrm{set}} \tag{4.1}$$

in analogy with equation (2.3). Equation (4.1) defines the so-called empirical probe sensitivity which is directly related to the affinity of the considered probe [23]. Note that insertion of equation (2.1) into (4.1) and neglecting saturation cancels out the transcript concentrations because they are assumed to be a constant for each probe set. The application of the intensity transformation according to equation (4.1) in combination with the $\langle \log I^{\mathrm{PM}} \rangle_{\mathrm{set}}$-criterium (see above) consequently provides the sensitivities for specific and non-specific hybridization, $Y_p^{\mathrm{P,S}}$

and $Y_p^{\mathrm{P,NS}}$, respectively (see also equation (2.3)). Finally, the NN free energy terms were obtained by multiple linear regression of the NN model,

$$Y_p^{\mathrm{P,h}} = -\sum_{k=1}^{LP-1} \sum_{\mathrm{B,B'=A,T,G,C}} \left(\Delta\varepsilon_k^{\mathrm{WW'}}(\mathrm{BB'}) \cdot \left(\delta(\xi_{p,k}^{\mathrm{P}}, \mathrm{B}) \cdot \delta(\xi_{p,k+1}^{\mathrm{P}}, \mathrm{B'}) - f_k^{\Sigma}(\mathrm{B}, \mathrm{B'})\right)\right), \tag{4.2}$$

to the experimental sensitivity data in analogy with equation (2.4). The superscript $\mathrm{WW'} = \mathrm{WCWC}$ (for $k = 1 \ldots 24$); WCSC (for $k = 12$), SCWC (for $k = 13$) assigns the base pairings formed by the adjacent bases $\mathrm{BB'}$ at the respective positions in the DNA probe sequence in the hybrid duplexes.

The NN model consequently provides 16 profiles for all combinations of $\mathrm{BB'}$ for each of the considered hybridizations, S–PM, NS–PM, S–MM, NS–MM; i.e., a total number of 64 profiles with $k = 1 \ldots 24$ positional dependent NN free energy values each. Note that the NN analysis uses subsets of all probes of the chip according to the intensity criterion in contrast to the SB analysis which is based on the much smaller number of spiked-in intensities. To compare the results of both approaches we transform the NN free energy terms into SB values by appropriate averaging according to

$$\Delta\varepsilon_k^{\mathrm{W}}(\mathrm{B}) = \tfrac{1}{2}\left(\sum_{\mathrm{B'=A,T,G,C}} \Delta\varepsilon_{k-1}^{\mathrm{W',W}}(\mathrm{B'B}) + \sum_{\mathrm{B'=A,T,G,C}} \Delta\varepsilon_k^{\mathrm{W,W'}}(\mathrm{BB'})\right).$$

The results are shown in figure 13 (compare lines and symbols). The excellent agreement between both profiles confirms the reliability of the chosen approach. The NN free energy terms of the middle base were assigned to WC and SC pairings in analogy with the SB analysis (i.e., $\Delta\varepsilon_{12}^{\mathrm{WC,W}}(\mathrm{B'B})$ and $\Delta\varepsilon_{13}^{\mathrm{W,WC}}(\mathrm{BB'})$, W = WC, SC and B, B′ = A, T, G, C, see above).

4.7. Middle triples

In the next step we merged the NN parameters of the adjacent bases in the middle of the sequence at positions $k = 12$–14 to estimate the apparent free energy of the 64 possible 'middle triples', $\Delta\varepsilon_{12}^{\mathrm{W}}(\mathrm{B'BB''}) \approx 0.5 \bullet [\Delta\varepsilon_{12}^{\mathrm{WC,W}}(\mathrm{B'B}) + \Delta\varepsilon_{13}^{\mathrm{W,WC}}(\mathrm{BB'})]$. They refer either to a central W = WC or SC pairing. Figure 14 shows the obtained middle-triple related free energy contributions. The data are grouped for each middle base, B = A, T, G, C and ranked with respect to the interaction strength of the WC pairings.

The triples with a central WC pair, $\Delta\varepsilon_{12}^{\mathrm{WC}}(\mathrm{B'BB''})$, reveal an increasing interaction strength according to $\mathrm{B} = \mathrm{A} < \mathrm{T} \approx \mathrm{G} < \mathrm{C}$ (note the negative sign of the $\Delta\varepsilon$-scale and that the interaction strength is maximum for minimum $\Delta\varepsilon$-values). The incremental interaction terms for the middle letter A are always positive whereas those with middle letter C are preferentially negative in agreement with previous results which were obtained by an alternative averaging procedure [5]. Minimum and maximum values are found for CCC and AAA homo-triples, respectively.

The apparent incremental free energies of the triples with a central SC pair, $\Delta\varepsilon_{12}^{\mathrm{SC}}(\mathrm{B'\underline{B}B''})$, correlate well with the respective free energies of triples with a central WC pair for B = A, T, C but not for G. Self-complementary mismatches of the central guanine, $\underline{\mathrm{G}} \bullet \underline{\mathrm{g}}$, obviously disturb the stability of the adjacent WC pairings in a base-specific fashion. Note that SC mismatches of the central adenine, $\underline{\mathrm{A}} \bullet \underline{\mathrm{a}}$, give rise to virtually constant triple values, $\Delta\varepsilon_{12}^{\mathrm{SC}}(\mathrm{B'\underline{A}B''})$, which only weakly depend on their neighbours. Contrarily, the stability of the triples with a central $\underline{\mathrm{C}} \bullet \underline{\mathrm{c}}^*$ vary markedly as a function of the adjacent bases providing the weakest stability for $\mathrm{B'\underline{C}B''}$ with B′, B″ = A, G and the strongest interactions for $\mathrm{C\underline{C}C}$.

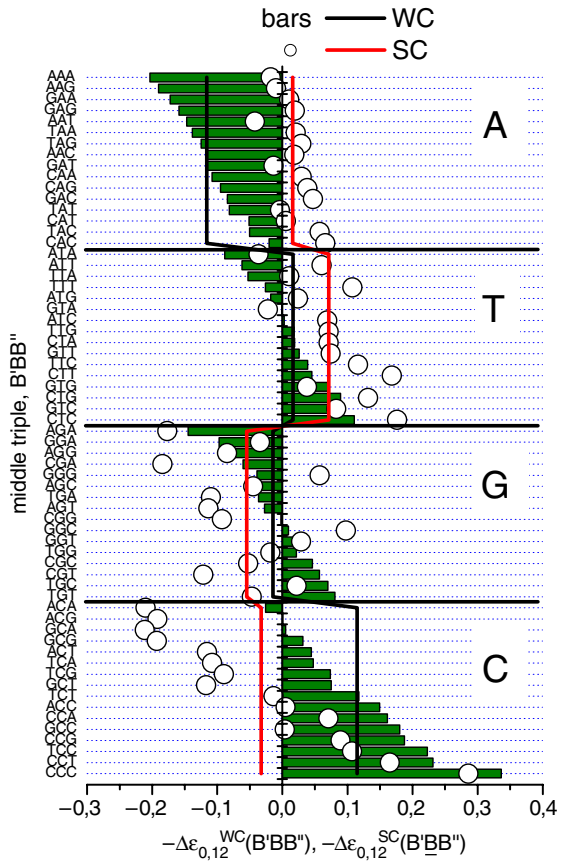

Figure 14. Middle-triple related apparent free energies referring to a central WC (bars) and a central SC (circles) pairing in the DNA/RNA duplexes of the PM and MM probes. The data are obtained by the fit of the nearest-neighbour model (equation (4.2)) to selected subsets of all PM and MM probes of a chip (see text). The data are grouped for each middle base and ranked with decreasing value of the 'WC' triples. The lines are the respective mean values of all triple data for a common middle base. The triple values widely scatter about the averages indicating the specific effect of the nearest neighbours owing to, for example, stacking interactions.

Interestingly, the incremental values are either almost identical for a number of WC and SC triples such as CCC/C$\underline{\text{C}}$C and AGA/A$\underline{\text{G}}$A, or they markedly differ for, for example, B′CB″/B′$\underline{\text{C}}$B″ with B′, B″ = A, G and GGG/G$\underline{\text{G}}$G. In the former case the relative stability of the triples is mainly determined by the neighbours at positions $k = 12$ and 14 and thus the triple virtually does not feel whether the central base forms a WC or SC pair. In the latter case the relative stability of the triple is distinctly changed by the replacement B $\rightarrow \underline{\text{B}}$. This situation is found for the central single-ringed C and adjacent double-ringed purines indicating that steric factors probably affect the relative stability of the mismatched SC pairing. However, this rule does not apply to the triples with a central adenine, indicating that the situation is more complex.

In summary, the nearest neighbours of a selected base obviously markedly modify its interactions within the probe/target duplexes. We previously showed that the NN interaction terms obtained from microarray data well correlate with the NN free energy contributions

which were extracted from DNA/RNA hybrid-duplexes in solution (see [5, 21, 27]). Systematic deviations between both data sets could be attributed to the effect of fluorescence labels attached to the microarray probes. Note that the previous data refer to neighboured WC pairings. The presented approach complements these results with the respective free energy parameters for adjacent WC and SC pairings. These data now allow us to estimate the effect of nearest neighbours on the middle-base related bias discussed above.

4.8. The middle-base related bias is affected by the neighbours

The affinity difference between middle triples with complementary central bases are calculated according to $\Delta\varepsilon_{12}^{\mathrm{WC-W}}(\mathrm{B'BB''}) \equiv \Delta\varepsilon_{12}^{\mathrm{WC}}(\mathrm{B'BB''}) - \Delta\varepsilon_{12}^{\mathrm{W}}(\mathrm{B'B^cB''})(\mathrm{W} = \mathrm{WC, SC})$ using the triple-data shown in figure 14. Note that the obtained $\Delta\varepsilon_{12}^{\mathrm{WC-W}}(\mathrm{B'BB''})$-values characterize the affinity difference between a PM probe with the middle triple $\mathrm{B'BB''}$ and the paired MM probe. Consequently the obtained data can be directly compared with the results of the SB analysis discussed above to assess the effect of nearest neighbours on the respective free energy difference (see figure 9, lower panel). It turns out that the triple values narrowly scatter about the SB data for $\mathrm{B} = \mathrm{T}$ and especially $\mathrm{B} = \mathrm{A}$. Hence, for a central adenine or thymine the respective free energy differences is well characterized by the respective SB value. In contrast, for a central guanine or cytosine the triple values widely spread about the SB data, especially along the abscissa. Consequently the SB values only crudely estimate the respective affinity difference between the respective PM and MM probes.

Note that negative ordinate values refer to 'bright' MM for which the MM intensity exceeds that of the PM at dominating non-specific hybridization (i.e., $I_p^{\mathrm{PM}} < I_p^{\mathrm{MM}}$) whereas positive ordinate values give rise to 'bright' PM (i.e., $I_p^{\mathrm{PM}} > I_p^{\mathrm{MM}}$) [7]. Interestingly, all triples with a central C or T are above, and all triples with a central A and G are below, the x axis. Hence, the consideration of nearest neighbours does not break the rule which has been established on the basis of the single-base analysis, namely that central pyrimidines (C, T) cause bright MM whereas central purines (A, G) give rise to bright PM. However, the degree of brightness varies considerably as a function of the adjacent bases. For example, CCC, CGC, GTG and GAG are more than two times as 'bright' compared with the overall mean whereas TCT and TGT possess tiny values near zero.

In summary, the consideration of the nearest neighbours significantly refines the interaction pattern in the considered duplexes. The NN model seems better suited to predict the probe affinities from their sequences to analyse microarray intensities beyond spiked-in data.

5. Summary and conclusions

We analysed the 'apparatus' function of GeneChip microarrays which is given by the intensity response of perfect matched and single-mismatched oligonucleotide probes to changes of the target concentration. In agreement with previous studies we found that the competitive two-species *Langmuir*-adsorption model well describes the probe intensities.

The knowledge of the hybridization isotherm of each probe in turn allows to extract absolute RNA concentrations from microarray intensity data. Note that this approach accounts for non-specific binding at small and for saturation at high intensities in a probe-specific fashion. The analysis of the PM–MM intensity difference provides at least no loss of accuracy and precision of the estimated concentration compared with the PM-only estimates which in turn outperform the MM-only results.

Each PM and MM probe is characterized by two hybridization constants which specify the propensity of the probe to bind specific and non-specific transcripts. Differences and

common properties of the PM and MM probes were extracted from the joint analysis of their hybridization constants.

- The binding affinities of the PM and of the MM strongly correlate for specific and for non-specific hybridization as well. The common sequence-region of the PM and MM dominates the affinity values.
- The affinity for non-specific hybridization is on the average equal for PM and MM. The purine–pyrimidine asymmetry of base pair interaction strengths causes however a characteristic PM–MM-intensity difference, the sign of which depends on the middle base of the probe. The processing of the PM–MM intensity difference requires the consideration of a background term due to non-specific hybridization which is reduced by nearly one order of magnitude when compared with the respective background of the PM and MM probes.
- Both the PM and MM probes respond to the concentration of specific transcripts. The affinity for specific hybridization of the PM however exceeds that of the MM on the average by nearly one order of magnitude because the central mismatched base only weakly contributes to the stability of the probe/target duplexes. The sensitivity of the PM–MM expression measure is nearly the same as that of the PM.
- The relation between the PM and MM intensities systematically depends on the middle base. This bias is different for specific and non-specific hybridization and it is strongly modulated by nearest neighbour of the middle base.
- The base-couple specific nearest-neighbour free energy contributions for specific and non-specific binding can be obtained from the chip data using a simple intensity criterion. The nearest-neighbour data provide the 64 possible middle-triples of DNA/RNA oligomer duplexes with a central WC and a central SC pairing which are published here for the first time (figure 14).

The two hybridization constants per probe used in the *Langmuir*-model can be predicted in a sequence-specific fashion using a sum of positional-dependent and base-specific nearest-neighbour free energy terms. We expect that this physical approach enables the development of new methods to correct raw intensity data from microarrays for the non-specific background and/or saturation effects and thus the proper estimation of the target concentrations beyond the training set of several hundreds of spiked-in probes. Note that the issues of probe-specific background intensity and saturation are currently not satisfactorily addressed by state-of-the-art preprocessing methods for GeneChip analysis.

Acknowledgment

The work was supported by the Deutsche Forschungsgemeinschaft under grant no. BIZ 6-1/3.

References

[1] Binder H 2006 *J. Phys.: Condens. Matter* **18** S491
[2] Hekstra D, Taussig A R, Magnasco M and Naef F 2003 *Nucleic Acids Res.* **31** 1962
[3] Burden C J, Pittelkow Y E and Wilson S R 2004 *Stat. Appl. Gen. Mol. Biol.* **3** 35
[4] Binder H, Kirsten T, Loeffler M and Stadler P 2004 *J. Phys. Chem.* B **108** 18003
[5] Binder H, Kirsten T, Hofacker I, Stadler P and Loeffler M 2004 *J. Phys. Chem.* B **108** 18015
[6] Binder H, Preibisch S and Kirsten T 2005 *Langmuir* **21** 9287
[7] Binder H and Preibisch S 2005 *Biophys. J.* **89** 337
[8] Binder H 2005 *Bioinformatics of Gene Regulation II* ed N Kolchanov and R Hofestaedt (Berlin: Springer Sciences and Business Media) p 451

[9] Naef F and Magnasco M O 2003 *Phys. Rev.* E **68** 11906
[10] Wu Z and Irizarry R A 2005 *Dept. of Biostatistics Working Paper* vol 73, p 1, John Hopkins University
[11] Lipshutz R J, Fodor S P A, Gingeras T R and Lockhart D J 1999 *Nat. Genet.* **21** 20
[12] Warrington J A, Dee S and Trulson M 2000 *Microarray Biochip Technology* ed M Schena (Natick, MA: Eaton) chapter 6 p 119
[13] Affymetrix 2001 *User Guide* (Santa Clara, CA: Affymetrix, Inc.)
[14] Sips R 1948 *J. Phys. Chem.* **16** 490
[15] Peterson A W, Heaton R J and Georgiadis R M 2001 *Nucleic Acids Res.* **29** 5163
[16] Carlon E and Heim T 2004 *Preprint* q-bio.BM/0411011 v1
[17] Affymetrix 2001 *New Statistical Algorithms for Monitoring Gene Expression on GeneChip Probe Arrays* Technical Note (Santa Clara, CA: Affymetrix)
[18] Li C and Wong W H 2001 *Proc. Natl Acad. Sci. USA* **98** 31
[19] Irizarry R A, Hobbs B, Collin F, Beazer-Barclay Y D, Antonellis K J, Scherf U and Speed T P 2003 *Biostatistics* **4** 249
[20] Naef F, Lim D A, Patil N and Magnasco M 2002 *Phys. Rev.* E **65** 4092
[21] Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M and Sasaki M 1995 *Biochemistry* **34** 11211
[22] Wu P, Nakano S and Sugimoto N 2002 *Eur. J. Biochem.* **269** 2821
[23] Binder H, Kirsten T, Loeffler M and Stadler P 2003 *Proc. German Bioinformatics Conf.* vol 2, p 145
[24] 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 and Webster T A 2003 *Proc. Natl Acad. Sci. USA* **100** 11237
[25] Gralla J and Crothers D M 1973 *J. Mol. Biol.* **73** 497
[26] Borer P N, Dengler B, Tinoco I Jr and Uhlenbeck O C 1974 *J. Mol. Biol.* **86** 843
[27] Wu P and Sugimoto N 2000 *Nucleic Acids Res.* **28** 4762