PROBING GENE EXPRESSION – SEQUENCE SPECIFIC HYBRIDIZATION ON MICROARRAYS

Hans Binder

Interdisciplinary Centre for Bioinformatics, University of Leipzig, Haertelstr. 16-18, D-04107 Leipzig, Germany, www.izbi.de, binder@izbi.uni-leipzig.de

Abstract: *Background:* DNA microarrays are routinely used to monitor the transcript levels of thousands of genes simultaneously. However, the array design, hybridization conditions, and oligodeoxyribonucleotide probe sequence impact the performance of the DNA microarray platform and must be considered by data analysis.
Results: We analyzed the signal intensities of GeneChip microarrays in terms of a microscopic binding model. It considers specific and non-specific transcripts, which both compete for duplex formation with perfect match (PM) and mismatch (MM) oligonucleotide probes. Intensity simulations enable us to judge the accuracy and precision of gene expression measures. The accuracy of the estimated fold changes ranks according to PM-MM>PM>MM whereas the precision decreases with PM≥MM>PM-MM where PM-MM denotes the respective intensity difference.
Conclusions: MM probes possess the potency to correct the intensity of the respective PM probe for the non-specific background. The middle base related bias of the MM intensity must however be considered by improved algorithms of data analysis. Moreover, the knowledge of base pair interactions suggests to substitute the complementary mismatches on GeneChips by alternative rules of MM design.

Keywords: DNA/RNA duplex stability, perfect match and mismatch probes, gene expression

Running head: Hybridization on microarrays

1. Introduction

The gene chip microarray technology empowers researchers in the collection of large-scale data on gene expression. The method is based on the selectivity of the hybridization reaction between target RNA transcribed from the gene of interest and complementary DNA probes grafted on the chip. The formation of probe/target duplexes is a complex process governed by an intricate interplay between several effects such as binding and saturation, surface electrostatics and non-equilibrium thermodynamics (Halperin et al., 2004; Hekstra et al., 2003; Held et al., 2003; Naef and Magnasco, 2003; Vainrub and Pettitt, 2002; Zhang et al., 2003). The proper interpretation of microarray data in terms of gene expression requires the detailed understanding of the hybridization mechanism on the level of base pairings at different concentrations of target RNA.

A typical GeneChip microarray such as the human genome HG-U133 chip consists of nearly 500,000 probe spots on an area of about 1.5 squared centimetres. Each spot is formed by 25meric DNA oligonucleotides of

(almost) one sequence which are grafted with the 3'-end at the glass support. The sequence of these perfect match (PM) probes corresponds to a 25meric fragment in the consensus sequence of the target gene. The DNA oligomers are expected to capture the complementary messenger RNA by sequence specific duplex formation, and, this way, to probe its abundance. The amount of bound RNA is detected using fluorescent labels. Consequently each probe spot gives rise to an intensity value which, in the ideal case, is directly related to the concentration of target RNA.

The sample solution represents a complex mixture of RNA fragments of different length and sequence. Consequently, the total RNA concentration can be split into two fractions, namely that of target RNA, which is specific (S) for a given probe, and that of non-specific (NS) RNA fragments, i.e. $c_{RNA}=c_{RNA}^{S}+c_{RNA}^{NS}$. Unfortunately, also the latter RNA can possess a non-negligible affinity for duplex formation with the probe oligomers. This NS-hybridization is problematic for chip analysis because it adds a "chemical" background intensity, which is not related to the expression degree of the target gene To deal with this problem, each PM probe is paired with a so-called mismatch probe (MM) on microarrays of the GeneChip-type (Affymetrix, 2001). The MM-sequence is identical with that of the respective PM probe except for the base in the middle of the oligomer, which is replaced by its complement to prevent S-hybridization. This way, the MM probe intends to measure the amount of NS-hybridization, and thus to provide a correction of the PM intensity for the chemical background.

The lower binding affinity of the MM probes predicts a systematically smaller spot intensity if compared with that of the respective PM, i.e., $I^{MM} < I^{PM}$. Figure 1 correlates the MM with the PM intensities of a typical

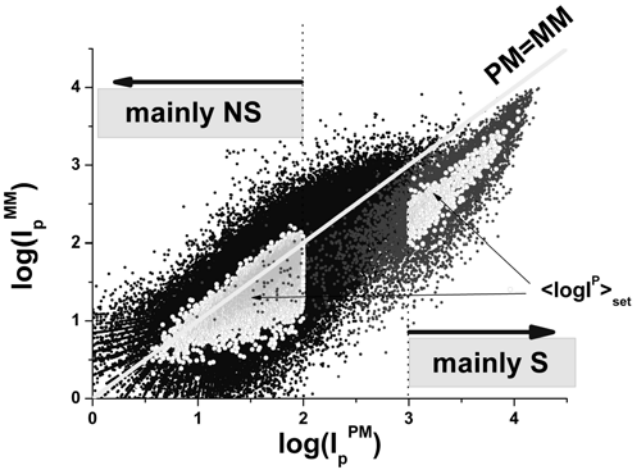

Figure 1: PM/MM intensity correlation plot of probe pairs taken from probe sets which meet the condition $<\log I^{PM}>_{set}<2$ (left data cloud, the open symbols show the set averages) and $<\log I^{PM}>_{set}>3$ (right data cloud). The former and latter data refer to predominantly non-specifically (NS) and specifically (S) hybridized probes, respectively. Note the small amount of bright MM in the S-subset (<5%) and the high amount in the NS-subset (>40%).

GeneChip microarray in a logarithmic scale. More than 40 % of the data points are found above the diagonal referring to so-called "bright" MM with a larger fluorescence intensity if compared with their PM counterpart (Naef et al., 2002). This result implies that conventional hybridization theory is simply inadequate, and particularly, that the basic mechanism of MM hybridization is not understood yet. As a consequence, many algorithms of gene expression analysis simply ignore MM intensity data or they are considered in an empirical fashion to exclude "bad" probes from the analysis (see (Irizarry et al., 2003) for an overview).

This study deals with basic issues of the GeneChip technology, such as the systematic effect of the probe sequence, and of matched and mismatched base pairings on the signal intensity, which at present are still unsolved. This sequence specific view is expected to improve data analysis as well as chip design.

2. Data

We have used microarray data from a calibration experiment provided by Affymetrix (http:// www.affymetrix.com/ support/ technical/ sample_data/ datasets.affx). In this experiment specific transcripts of 42 genes referring to 462 probes were titrated in definite concentrations onto a series of chips in three replicates to study the relation between the ("spiked-in") RNA concentration and the intensity of the respective "spiked-in" probe. For a more detailed description of the HG U133 Latin Square data set see for example ref. (Binder and Preibisch, 2005). The Affymetrix technology uses fragmented biotin-labelled RNA for hybridization which is obtained by reverse transcription of the extracted RNA into cDNA (mRNA → cDNA) and subsequent in vitro transcription (cDNA → cRNA), biotinylation and fragmentation.

3. Results and Discussion

3.1. Probe intensities at specific and at non-specific hybridization

Besides the PM/MM pairs the GeneChip technology uses a second redundancy of probe design to get independent estimates of the expression degree of each gene. Usually, eleven PM/MM probe pairs referring to different regions of the same gene, and thus to a common concentration value of the target RNA, are collected into so-called probe sets. The set-averaged mean log intensity provides a rough measure of the concentration of S-transcript according to $\langle \log I^{P} \rangle_{set} \propto [\log c_{RNA}{}^{S} + Z^{set}]$, where Z^{set} is a set-specific constant, which scatters with a standard deviation of $\sim \pm 0.5$ about its chip average (Binder et al., 2004).

Figure 1 selects two subsets of probe intensities meeting the conditions of $\langle \log I^{PM} \rangle_{set} < 2$ and $\langle \log I^{PM} \rangle_{set} > 3$, respectively. The former one includes probes referring to relatively small concentrations of S-transcripts and thus to the limiting case of dominating NS-hybridization. The respective data cloud nearly symmetrically spreads about the diagonal with a relative large fraction of bright MM ($I^{MM} > I^{PM}$) of more than 40%. Contrarily, the data cloud formed by the second ensemble of probes is clearly shifted away from the diagonal

with a tiny fraction of bright MM of less than 5%. These probes correspond to the limiting case of dominating S-hybridization with a relatively high concentration of S-transcripts. Hence, the effect of bright MM is related to NS-hybridization. S-hybridization nearly exclusively produces bright PM, $I^{PM}>I^{MM}$, as expected by hybridization theory.

3.2. Binding model of duplex formation

The fluorescence intensity per probe spot can be described by (Binder et al., 2004)

$$I^P \approx F_{chip} \cdot c_{RNA} \cdot K^{P,S} \cdot \left[x^S + \left(1 - x^S\right) \cdot r^P \right] \cdot S^P \qquad (1)$$

if one neglects the optical background. The binding "strength" (or affinity) of the DNA probe for duplex formation with the RNA is characterized by the binding constants of S- and NS-hybridization, $K^{P,h}$ (h=S, NS; $r^P = K^{P,NS}/K^{P,S}$ denotes their ratio) and the saturation term $S^P = \left(1 + K^{P,S} \cdot c_{RNA} \cdot \left[x^S + \left(1 - x^S\right) \cdot r^P \right]\right)^{-1}$.

The fraction of target RNA is $x^S = c_{RNA}{}^S/c_{RNA}$, and the fraction of NS-RNA, involving other sequences than the intended target, is $x^{NS} = (1 - x^S)$. The chip specific constant F_{chip} specifies the detection "strength" of the technique. It includes besides other factors the amount of labelling.

3.3. PM/MM-trajectories of individual probes

Each probe is characterized by a "PM/MM-trajectory", which describes the intensity change upon increasing content of S-transcripts ($0 \le x^S \le 1$) in the $\log I^{MM}$ versus $\log I^{PM}$ correlation plot. Figure 2 shows the experimental intensity data of six selected probes together with fits by means of Eq. (1) (compare curves and symbols). The trajectory, typically, "starts" near the diagonal line in the absence of S-transcripts (i.e. $I^{PM} \approx I^{MM}$ for $x^S = 0$), "moves"

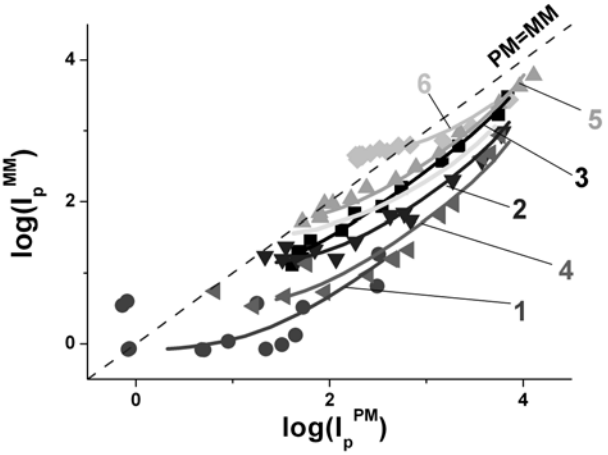

Figure 2: PM/MM trajectories of selected spiked-in probes (see Table 1 for probe # and sequence). The curves are calculated using Eq. (1) and the parameters listed in Table 1.

Table 1: Binding constants of selected probes (see trajectories in Fig. 2) and of the mean PM/MM trajectories (see Fig. 3).

	PM sequence	**# bases**		**PM**		**MM**	
#	**Selected probes**	**C**	**A**	**$logK^{PM,S}$**	**$logr^{PM}$**	**$logK^{MM,S}$**	**$logr^{MM}$**
1	TATAATCTTTTA**T**ACAGTGTCTTAC	4	7	-1.3	-2.7	-2.7	-1.5
2	GAGGATTCATCT**T**GCACATCTGAGA	5	7	0.3	-3.0	-0.7	-2.3
3	GACAGGTCCTTT**T**CGATGGTACATA	5	6	0.3	-2.7	-0.3	-2.8
4	GCACAAGTTTTT**C**TACACTCAGTGT	6	6	0.3	-3.0	-1.0	-2.3
5	GTGATGCTCAAT**G**GATCCCGCAGTA	7	6	0.7	-3.0	0.2	-2.5
6	TAGGCCATTTGG**A**CTCTGCCTTCAA	7	5	0.0	-1.8	-0.4	-1.1
	Middle base averages (PM)			**$logK_B^{PM,S}$**	**$logr_B^{PM}$**	**$logK_B^{MM,S}$**	**$logr_B^{MM}$**
B=	**A**			-0.15	-2.45	-0.7	-1.8
	T			-0.05	-2.45	-0.8	-2.05
	G			0.0	-2.45	-0.8	-1.5
	C			+0.20	-2.45	-0.9	-1.75
	Standard deviation			0.14	0.0	0.08	0.23
	Total mean			**$logK_0^{PM,S}$**	**$logr_0^{PM}$**	**$logK_0^{MM,S}$**	**$logr_0^{MM}$**
				0.0	-2.45	-0.8	-1.8

towards bright PM (i.e. $I^{PM}>I^{MM}$) with increasing x^S and, finally, the trajectory returns back in direction of the diagonal at high x^S values owing to saturation.

Each PM/MM-trajectory is characterized by four model parameters: the affinity constant for S-binding, $K^{P,S}$ and the effective affinity ratio r^P for the P=PM and MM probes (see Tab. 1). It turns out that the S-binding constants of the PM exceed that of the MM by a factor between about two and twenty. Therefore, the PM intensity of all considered probes is distinctly higher than that of the respective MM probe at larger S-transcript concentrations. The relation between PM and MM intensities however is more heterogeneous in the limit of dominating NS-hybridization. The trajectories can start on both sides of the diagonal line in Fig. 2. This result indicates that the affinity of the PM probes for NS-transcripts is either higher or smaller compared with that of the respective MM.

3.4. Mean PM/MM trajectories

The PM/MM trajectories of individual probes are well described by the suggested binding isotherms (Eq. (1)). To generalize these results in terms of mean trajectories we calculated the "total" average over the log-intensities of all 462 spiked-in probes using also all three available replicates at each concentration $<logI^P> \equiv <logI^P>_{sp\text{-}in}$ (P=PM, MM) as well as partial averages over subsets of probes with the common middle base B=A,T,G,C at position k=13 of their sequence, $logI_B^P \equiv <logI_p^P>_B$. The respective trajectories

characterize the average intensity relation between the PM and MM probes (see symbols in Fig. 3). Also the mean intensities are well approximated by Eq. (1) (see lines) where the probe-specific binding constants are substituted by effective values. They can be interpreted as log-averages over the considered ensemble of probes (i.e., $\log K^{P,h} \rightarrow \log K_0^{P,h} \approx \langle \log K^{P,h} \rangle_{sp\text{-}in}$ and $\log K^{P,h} \rightarrow \log K_B^{P,h} \approx \langle \log K^{P,h} \rangle_B$). The mean binding constant of the PM probes for target RNA, $K_0^{PM,S}$, exceeds that of the MM almost by one order of magnitude (see Tab. 1). On the other hand, the mean binding constant of the probes for non-specific binding is by two-three orders of magnitude weaker than that for specific binding ($\log r_0^{PM} = -2.45$, $\log r_0^{MM} = -1.8$).

The middle-base specific PM/MM trajectories diverge in a systematic fashion from each other. For example, the trajectories of the purine middle bases B=A, G start in the range of bright MM (i.e. $I^{PM} < I^{MM}$) at small intensities (and dominating NS-hybridization) in contrast to that of the pyrimidines B=T, C. The trajectories of G and T however merge with increasing x^S (at higher intensities). This behavior indicates that S- and NS-RNA are binding differently to the probes as a function of their middle base. Note that the mean S-binding constant of the PM is decreasing according to C>T≈G>A (see $\log K_B^{PM,S}$ in Tab. 1) in contrast to that of the MM, which is a constant almost.

The data shown in Fig. 3 are averaged over the limited ensemble of 462 spiked-in probes. To generalize these results for the whole set of 250.000 PM and MM probes of the chip we correlate the intensities of the PM which possesses a common middle base with their paired MM probe intensities (and complementary middle base, see Fig. 4). The data cloud for B=A is clearly shifted towards bright MM compared with that for T. The same tendency was obtained for G and C (not shown here). The systematic trend due to the different middle bases can be filtered out more clearly, if one calculates running averages over 1000 subsequent probes along the axes (see lines in Figs. 4 and 5). So it turns out that the curves referring to the whole set of

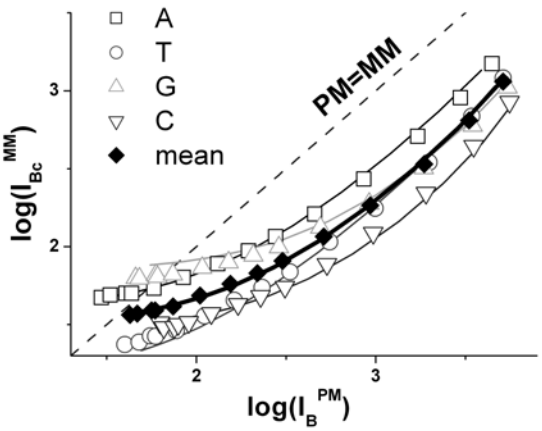

Figure 3 : Mean PM/MM trajectories averaged over all spiked-in probe pairs with a common PM-middle base and their total mean.

probes show virtually the same features as the respective curves for the spiked-in probes (compare Figs. 5 and 3).

3.5. Sequence specific binding: single-base model and probe sensitivity

The observed intensities are functions of the affinity for DNA/RNA duplex formation, which in turn depends on the sequences of the 25meric probe and of the bound RNA fragments. Note that the trajectories of most of the selected probes in Fig. 2 and Tab. 1 are shifting systematically towards higher intensities (and $K^{P,S}$) with increasing C and decreasing A content (see columns "# bases" in Tab. 1). For a more detailed description we used the positional dependent single-base (SB) model, which approximates the deviation of the probe intensity from its set average by a sum of base-specific terms according to

$$Y^P = \log I^P - \left\langle \log I^P \right\rangle_{set} \approx \sum_{k=-N_{out}+1}^{N_b+N_{out}} \sigma_k^P(\xi_k^P) \quad , \quad P = PM, MM \qquad , \qquad (2)$$

where ξ_k^P is the base (A,T,G or C) at position k of the probe sequence taken from the target gene. Equation (2) defines the sensitivity of the probe,

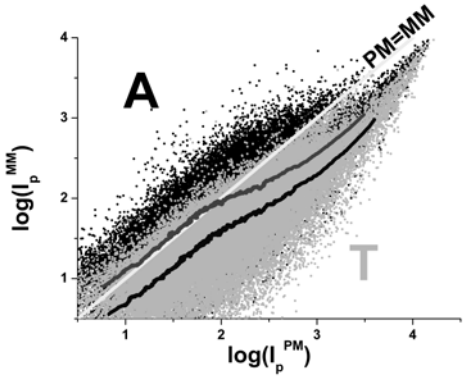

Figure 4: PM/MM correlation plot of probe pairs with PM-middle bases A and T. Both data clouds are shifted in vertical direction to each other. The lines are running averages through the respective clouds.

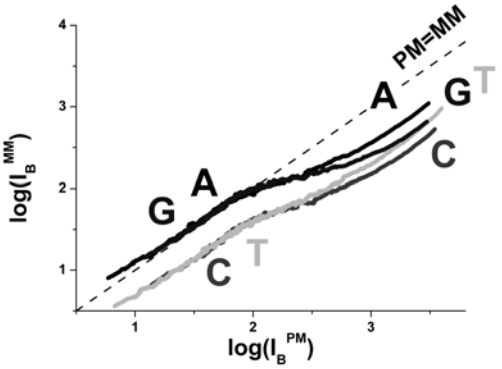

Figure 5: Running averages for all four PM-middle bases. Note the correspondence with the middle-base averages over the spiked in probes (Fig. 3).

Y^P, 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 and the total RNA concentration, F_{chip} and c_{RNA}, respectively (see Eq.(1)).

Note that it is by the DNA-probe sequence only that we identify the SB sensitivity. It consequently refers to matched and/or to mismatched pairings with the RNA in the respective duplex. Moreover, the length of the RNA fragments typically exceeds the length of the 25meric probes. Hence, also bases which dangle outside of the target sequence can affect the binding affinity, because they modify the propensity of the RNA fragments for intramolecular folding. In addition, also fluorescently labelled bases outside of the target region contribute to the measured fluorescence intensity. The model, therefore, considers the next N_{out}=20 bases, which precede and follow the probe sequence of N_b=25 nucleotides in the sequence of the target gene.

The sensitivity coefficients of the SB model, $\sigma_k^P(B)$, have been determined by means of multiple linear regression of the Y^P-values of selected subsets of PM- and MM-probes referring predominantly to S- and NS-hybridization. In accordance with our previous results we collect all probe pairs of the chip meeting the condition $\langle \log I^{PM} \rangle_{set} > 3$ and $\langle \log I^{PM} \rangle_{set} < 2$ into the former and latter subset, respectively (see Fig. 1).

The shapes of the sensitivity profiles of the PM probes of both subsets, and of the NS-hybridized MM probes, are very similar (see Fig. 6). In particular, the profiles for B=C, A show the typical parabola-like shape within the region of the probe sequence ($1 \le k \le 25$). They are showing maximum and minimum in the middle of the sequence, respectively, whereas the sensitivity

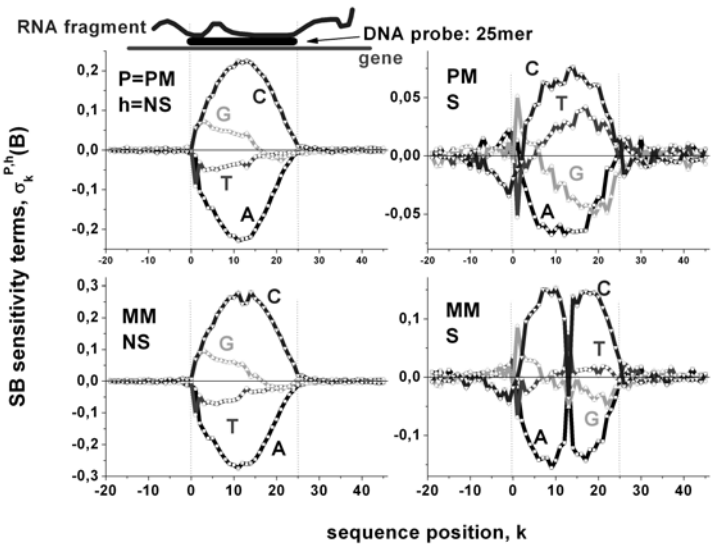

Figure 6: Single-base-sensitivity profiles of PM and MM probes in the limit of specific (S) and non-specific (NS) hybridization. The profiles consider 65 positions and extend to 20 bases before and after the 25meric probe sequence (see the cartoon). Note the "dent" in the middle of the MM-S profiles for B=C and A.

contributions for B=T, G change almost monotonously (see also (Binder et al., 2003; Binder et al., 2005; Mei et al., 2003; Naef and Magnasco, 2003).

In contrast, the profiles for B=A, C of S-hybridized MM distinctly differ in the middle of the sequence from the other profiles considered. Namely, the sensitivity contribution of the middle base markedly drops to tiny values near zero. Hence the mismatched middle base of the MM probes on average provides only a weak base-specific contribution to the probe intensity within the limit of S-hybridization. On the other hand, the remaining sequence positions at $k \neq 13$ show similar sensitivity profiles for the PM and MM probes on all conditions.

Finally, the small sensitivity contributions outside of the target region at $k<1$ and $k>25$ indicate that these positions only weakly contribute to the probe intensity in a base specific fashion.

3.6. Base pairings in probe/target duplexes

We analysed PM and MM probe intensities using two approaches: first, by averaging over all probes with a common middle base and the analysis of the respective PM/MM trajectories in terms of the binding model and, second, by the fit of the probe sensitivities by the sum of SB terms which explicitly extract the relative contribution of the middle base to duplex stability. Both independent approaches are complementing each other. Note that the middle base-specific binding strength and the respective SB sensitivity term both characterize the effective interactions of the middle base in the RNA/DNA oligonucleotide duplexes, i.e., $\log K_B^{P,h} \approx \sigma_{13}^{P,h}(B)$.

The results give rise to the following interpretation in terms of the base pairings that stabilize the DNA/RNA duplexes (see Fig. 7). The PM probes "per definition" form exclusively Watson Crick (WC) pairs with the complementary sequence of the target RNA. The central WC pair of the PM,

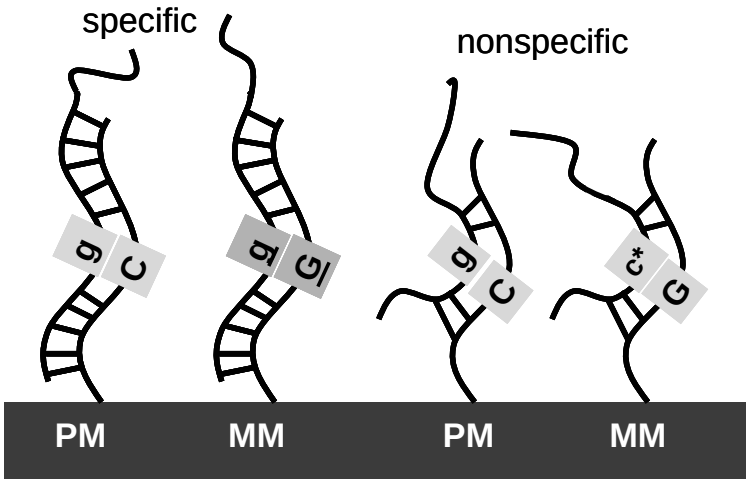

Figure 7: Base pairings in the middle of duplexes between DNA probes and RNA fragments. The NS-duplexes are stabilized by a smaller number of WC pairings compared with the S-duplexes. The middle base of the MM forms a SC pairing upon S-hybridization. Note the reversal of the WC pair in the NS-duplexes of the PM and MM probes.

$B \bullet b^c$ (lower case letter refer to RNA; the superscript denotes the complement, e.g. C•g), is replaced by the respective self complementary (SC) pair, $\underline{B}^c \bullet \underline{b}^c$ (e.g. $\underline{G} \bullet \underline{g}$) in the respective MM/target duplex. The shift of the trajectories into the range of bright PM at dominating S-hybridization indicates that the SC pairing of the MM is considerably weaker than the respective WC pairing of the PM. The middle base averaged binding constants, $K_B^{P,S}$, reflect the relative strength of the respective WC pairing. Its values reveal a purine-pyrimidine asymmetry according to C•g > G•c*≈ T•a > A•u* (the asterisk denotes labeling).

On the other hand, the "NS-background" represents a mixture of RNA fragments with a broad distribution of base compositions, which enables the formation of a sufficient number of WC pairings which stabilize the NS-duplexes. The middle bases on average are assumed to form WC pairings. These reverse direction for each PM/MM pair: $B \bullet b^c$ for the PM becomes $B^c \bullet b$ for the MM. The probe pairs split into two fractions with purine (A,G) middle bases of the PM and preferentially bright MM ($I^{MM}>I^{PM}$) and with pyrimidines (C,T) in the middle and the reverse intensity relation ($I^{PM}>I^{MM}$) due to the purine/pyrimidine asymmetry of interaction strengths.

3.7. Simulated intensity data

To illustrate the effect of the probe sequence on the intensity we used a synthetic, randomly generated "target gene" of 3000 nucleotide bases. The intensity of all possible PM and MM probes was calculated by means of the following equations adapted from Eqs. (1) and (2)

$$I_p^P \approx K_0^{P,S} \cdot \left[x^S \cdot 10^{Y_p^{P,S}} + \left(1 - x^S\right) \cdot 10^{Y_p^{P,NS}} \cdot r_0^P \right] \cdot S_p^P \quad with \quad P = PM, MM \quad ; \tag{3}$$

$$Y_p^{P,h} = \sum_{k=p-12}^{p+13} \sigma_k^{P,h}(\xi_k^P) \quad , \quad h = NS, S \quad ;$$

$$S_p^P = \left(1 + c_{RNA} \cdot K_0^{P,S} \cdot \left[x^S \cdot 10^{Y_p^{P,S}} + \left(1 - x^S\right) \cdot 10^{Y_p^{P,NS}} \cdot r_0^P \right]\right)^{-1}$$

The PM probe sequence refers to a sliding window of 25 positions, which moves along the gene sequence, ξ_p (p=1,…3000). For the respective MM sequence, the middle base at k=p was replaced by its complement. The model parameters, namely the total binding constants and the SB sensitivity contributions, were taken from the fits of the mean PM/MM trajectories and from the fits of the SB model to the experimental sensitivity data (see above).

Figure 8 shows the calculated intensities as a function of sequence position. The PM and MM intensities are correlated in Fig. 9. Note that this correlation plot shows essentially the same characteristic features as the plot of the experimental data (compare with Fig. 1). In particular, the data shift towards "bright" PM with increasing x^S and, finally, they turn back to the diagonal line owing to saturation.

Let us at first neglect saturation (S_p^P=1). In this special case the simulated PM and MM intensity data vary by about four orders of magnitude due to differences in their sequence within the limits of NS- (x^S=0, left panel of Fig. 8) and S-hybridization (middle panel of Fig. 8). Note that neighboring probes with the indices p and p+1 are shifted by only one base each to another. Both,

the intensity and the sensitivity of the probes smoothly change along the target gene as a consequence.

The comparison of the respective Y_p^{PM}-courses shows that the sensitivity of the PM is invariant for changes of the fraction of specific transcripts. This property reflects the constant, i.e. middle base independent ratio of the S- and NS-binding constants, i.e. $r_B^{PM} \approx$const (Tab. 1). The MM sensitivity reveals a more complex behavior. Firstly, the main course of Y_p^{MM} changes parallel to that of Y_p^{PM} because both sequences are identical for all positions $k \neq 13$. Secondly, the individual MM values scatter however about the PM sensitivities due to their complementary middle bases (see Y^{PM-MM}). Thirdly, and most interestingly, the scattering pattern is different for NS- and S-hybridized MM owing to the different sensitivities of the middle base (see also

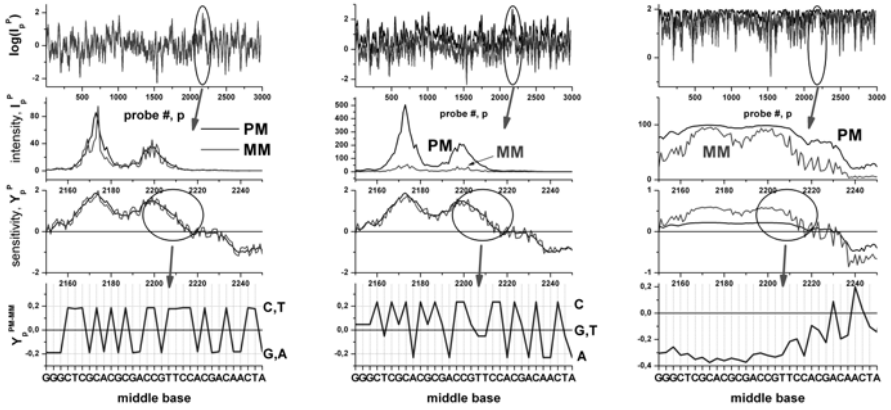

Figure 8: Simulated PM and MM intensity data, the respective sensitivities and the difference $Y^{PM-MM}=Y^{PM}-Y^{MM}$ (from top to bottom). The left and the middle panel refer to non-specific hybridization ($x^S=0$) and to a NS+S mixture with a fraction of specific transcripts of $x^S=0.03$, respectively, without considering saturation. The right panel considers saturation. Note the different scattering patterns of Y^{PM-MM} as a function of the middle base and of saturation.

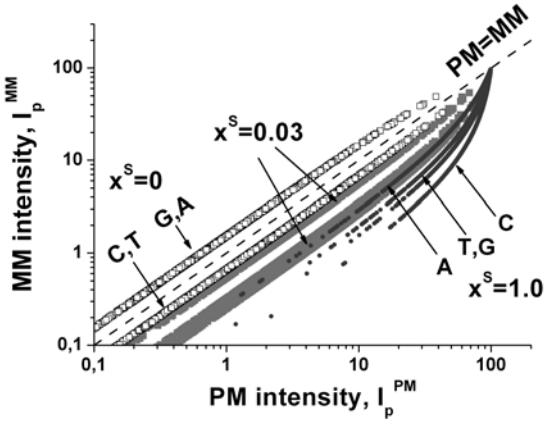

Figure 9: PM/MM- cor-relation plot of calculated intensities referring to three NS+S mixtures of different composition (see x^S values within the figure). Compare the simulated with the experimental data shown in Fig. 1. The branches refer to probe pairs with a common middle base of the PM (see figure).

Fig. 6). Note that the relative binding strength for NS- binding of the MM, $\log r_B^{MM}$, distinctly varies upon change of the middle base giving rise to the maximum standard deviation among the data considered (see Tab. 1).

The intensity data are strongly modified by the saturation of the probes with bound transcripts (see Fig. 8, right panel). This effect especially decreases the peak values of the PM intensity accompanied by a marked drop of the respective sensitivity. The effect of saturation also smoothes out the scattering of the MM sensitivity in the range of high intensities (see Y^{PM-MM} in Fig. 8).

3.8. Differential expression: accuracy and precision

The basic application of the GeneChip technology intends to estimate the level of differential gene expression in terms of the change of the RNA transcript concentration between different samples, e.g. between the sample of interest and an appropriately chosen reference. The respective ratio of target concentration, $R_{true} \equiv x^S(samp)/x^S(ref)$, defines the "true" fold change which an analysis algorithm aims to extract from the probe intensities. In the simplest approach, the intensities themselves provide the apparent fold changes in terms of the ratio $R_p^P \equiv I_p^P(samp)/I_p^P(ref)$ with P=PM, MM and PM-MM for PM-only, MM-only and $I_p^{PM-MM} = I_p^{PM} - I_p^{MM}$ difference estimates, respectively.

Our intensity simulation enables to judge the accuracy and precision of the apparent fold change by direct comparison with the true value. Note that R_p^P varies as a function of the probe sequence for a fixed R_{true}. In our notation, the precision specifies this variability in terms of the standard deviation, $SD(R_p^P*)$, of the relative apparent fold change $R_p^P* = R_p^P/R_{true}$ for all probes of the generated test gene (see previous section). On the other hand, the accuracy

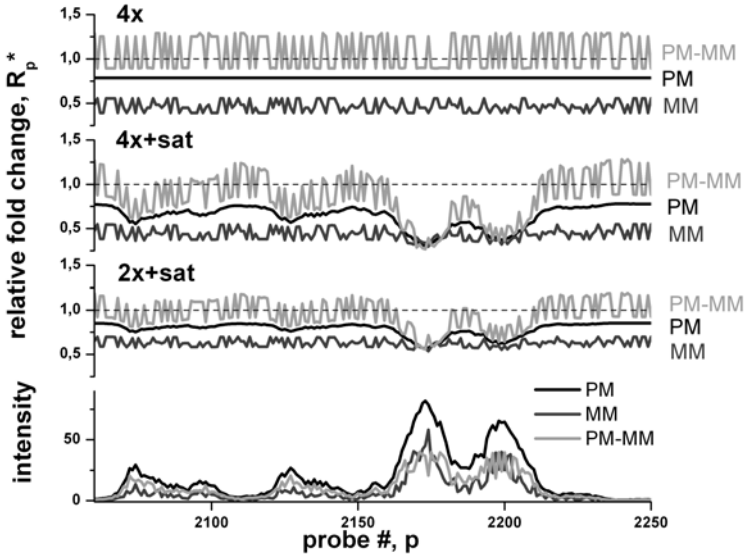

Figure 10: Simulated fold changes of PM, MM and PM-MM intensity measures. The data are normalized with respect to the "true" fold change of 4x and 2x, i.e. $R* = R/R_{true}$ (ideally =1). See Table 2 for assignments. The accuracy ("agreement with unity") ranks according to PM-MM>PM>MM whereas the precision ("scattering width about the mean") decreases with PM>MM>PM-MM.

Table 2: Accuracy (R*) and precision (SD) of PM, MM and PM-MM intensity measures for fold changes of gene expression (R_{true}). Saturation is neglected in one of the 4x samples and considered in the "+sat" samples. The fraction of specific transcripts is x_S=0.03 in the reference. See text.

	PM		MM		PM-MM	
R_{true}	R*	SD	R*	SD	R*	SD
4x	0.79	0.0	0.47	0.06	1.08	0.19
4x+sat	0.70	0.09	0.45	0.06	0.95	0.23
2x+sat	0.82	0.04	0.64	0.04	1.00	0.14

reflects the consistency between true and apparent fold changes in terms of $R^P*=<R_p{}^P*>_{gene}$, the averaged relative fold change. Ideally, $SD(R_p{}^P*)$ and R^P* adopt values near zero and unity respectively.

Figure 10 and Tab. 2 compare special situations referring to a "true" two- and fourfold concentration change (R_{true}=2 and 4). It clearly turns out that the PM-MM intensity difference provides the best accuracy with R* near unity. The subtraction of the MM intensity obviously provides a suitable correction of the PM data for the chemical background caused by non-specific hybridization. On the other hand, the PM-MM data are behaving relatively noisy giving rise to, by far, the worst precision. Saturation decreases both accuracy and precision (see "+sat" in Tab. 2).

In summary, the accuracy of the estimated fold changes ranks according to PM-MM>PM>MM, whereas the precision decreases with PM≥MM>PM-MM. The former result can be simply explained by the decreasing relative contribution of non-specific hybridization to the total signal intensity, which is minimum for PM-MM and maximum for MM. The latter trend is caused by the variability of the MM sensitivity owing to the changing affinity of the MM middle base in S- and NS-duplexes.

The potential accuracy-advantage of an analysis algorithm using the PM-MM difference is opposed by its low precision. Instead, a PM-only algorithm for extracting differential expression measures seems to afford a suited compromise between accuracy and precision in agreement with recent results (see (Irizarry et al., 2003) and references cited therein).

4. Conclusions: consequences for data analysis and chip design

NS hybridization considerably complicates the analysis of microarrays because it adds a background intensity not related to expression degree of the gene of interest. Probes with mismatched base pairings possess the potency to estimate the background level and, this way, to correct the intensity of the respective PM probe. We found that the intensity of complementary MM however introduces a systematic source of variation relative to the intensity of the respective PM probe owing to different base pairings in the NS-duplexes. In consequence, the naive correction of the PM signal by subtracting the MM intensity decreases the precision of expression measures. Our results imply improved algorithms of data analysis, which explicitly consider the middle-

base related bias of the MM intensities to reduce their systematic variability. Moreover, the knowledge of base pair interactions suggests to substitute the complementary mismatches on GeneChips by alternative rules of MM design.

5. Acknowledgments

The work was supported by the DFG under grant BIZ 6-1/2. I thank P. Stadler and M. Loeffler for the stimulating discussion and T. Kirsten and St. Preibisch for their help in practical questions.

6. References

Affymetrix. 2001. Affymetrix Microarray Suite 5.0. *In* User Guide. Affymetrix, Inc., Santa Clara, CA.

Binder, H., Kirsten, T., Loeffler, M., and Stadler, P. 2003. Sequence specific sensitivity of oligonucleotide probes. *Proceedings of the German Bioinformatics Conference* 2:145-147.

Binder, H., Kirsten, T., Loeffler, M., and Stadler, P. 2004. The sensitivity of microarray oligonucleotide probes - variability and the effect of base composition. *Journal of Physical Chemistry B.* 108(46) 18003-18014.

Binder, H., and Preibisch, S. 2005. Specific and non-specific hybridization of oligonucleotide probes on microarrays. *Biophys. J.* in press (see http://arxiv.org/ftp/q-bio/papers/0410/0410028.pdf).

Binder, H., Preibisch, S., and Kirsten, T. 2005. Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *http://www.arvix.org/abs/q-bio.BM/0501008*.

Halperin, A., Buhot, A., and Zhulina, E.B. 2004. Sensitivity, Specificity, and the Hybridization Isotherms of DNA Chips. *Biophys. J.* 86(2):718-730.

Hekstra, D., Taussig, A.R., Magnasco, M., and Naef, F. 2003. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucl. Acids. Res.* 31(7):1962-1968.

Held, G.A., Grinstein, G., and Tu, Y. 2003. Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl. Acad. Sci. USA* 100(13):7575-7580.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucl. Acids. Res.* 31(4):e15-.

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. Probe selection for high-density oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 100(20):11237-11242.

Naef, F., Lim, D.A., Patil, N., and Magnasco, M. 2002. DNA hybridization to mismatched templates: A chip study. *Phys. Rev. E* 65:4092-4096.

Naef, F., and Magnasco, M.O. 2003. Solving the riddle of the bright mismatches: hybridization in oligonucleotide arrays. *Phys. Rev. E* 68:11906-11910.

Vainrub, A., and Pettitt, B.M. 2002. Coulomb blockage of hybridization in two-dimensional DNA arrays. *Phys. Rev. E* 66:art. no. 041905.

Zhang, L., Miles, M.F., and Aldape, K.D. 2003. A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* 21:818-828.