Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays

Hans Binder*, Stephan Preibisch, Toralf Kirsten

Interdisciplinary Centre for Bioinformatics, University of Leipzig

* corresponding author: Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Haertelstr. 16-18, binder@izbi.uni-leipzig.de, fax: ++49-341-97-16679

Supplementary Material

S1: Single base contribution of fluorescence emission

The length of the RNA fragments, N_b^{RNA}, typically exceeds the length of the 25meric oligomer probes. Consequently also labelled bases which dangle outside of the probe/target duplex potentially contribute to the measured fluorescence intensity in addition to labels attached to the 25meric target region. Let us denote the number of bases outside of the respective 25meric duplex by N_b^{out} for a RNA fragment of total length $N_b^{RNA} = N_b + N_b^{out}$. The respective number of labelled bases inside and outside of the 25mer is $N_p^{F,in}(\xi^T)$ and $N_p^{F,out}(\xi^{T,out})$, respectively, where $\xi^{T,out}$ is the subsequence of the target RNA exceeding the probe on both sides. The fluorescence intensity of a RNA fragment is related to the number of labelled c* and u*, which is given by the number of complementary G and A of the target gene according to

$$N_p^{F,S} = N_p^{F,in}(\xi^T) + N_p^{F,out}(\xi^{T,out}) = N_p^{u*}(\xi^T + \xi^{T,out}) + N_p^{c*}(\xi^T + \xi^{T,out}) = N_p^{A}(\xi^P + \xi^{P,out}) + N_p^{G}(\xi^P + \xi^{P,out})$$

if one assumes exclusively WC pairings. The contribution to the sensitivity of a selected probe owing to the number of potentially labelled bases per target, $N_p^{F,S}$, is (see Eqs. 1 and 7)

$$Y_p^{P,F} \equiv \log N_p^F - \left\langle \log N_p^F \right\rangle_{set} = \Delta_p^F \cdot \left[\sum_{k=1}^{N_b} \sum_{B=A,G} \left(\delta(B,\xi_k^P) - f_k^{set}(B) \right) - \sum_{k=1}^{N_b} \sum_{B=T,C} \left(\delta(B,\xi_k^P) - f_k^{set}(B) \right) \right] \qquad (A1)$$

$$\text{with} \quad \Delta_p^F = \frac{\log N_p^F - \left\langle \log N_p^F \right\rangle_{set}}{\delta N_p^{F,in}} \quad \text{and} \quad \delta N_p^{F,in} = N_p^{F,in} - \left\langle N_p^{F,in} \right\rangle_{set} = \sum_{k=1}^{N_b} \sum_{B=A,T,G,C} \left(\delta(B,\xi_k^P) - f_k^{set}(B) \right)$$

where averaging was performed over the probe set ($\Sigma \equiv$ set).

The coefficient Δ_p^F specifies the contribution of fluorescence labelling per potentially labelled base in the considered target sequence of length N_b. Effectively each labelled base pair increases and each nonlabelled pair decreases the sensitivity by Δ_p^F. With $\delta N_p^F = \delta N_p^{F.in} + \delta N_p^{F,out}$ ($\delta N_p^{F,i} = N_p^{F,i}$ -

$<N_p^{F,i}>_{set}$, i = in, out) and $<\delta N^F>_{set}=0$ one obtains the following approximation for Δ_p^F in the limit of small $\delta N_p^F/<N_p^F>_{set} << 1$, which is justified for sequence lengths $N_b^{RNA} > 20$,

$\Delta_p^F = [\log(1+\delta N_p^F/<N_p^F>_{set}) - <\log(1+\delta N_p^F/<N_p^F>_{set})>_{set}]/\delta N_p^{F,in}$

$\approx (\ln 10 \cdot \delta N_p^{F,in} \cdot <N_p^F>_{set})^{-1} (\delta N_p^F - <\delta N^F>) = (1 + \delta N_p^{F,out}/\delta N_p^{F,in})/(\ln 10 <N_p^F>_{set})$.

The binominal distribution $B(N^F, N_b^{tot}, p) = \binom{N_b^{tot}}{N^F} p^{N^F} (1-p)^{N_b^{tot}-N^F}$ specifies the probability to find N_p^F potentially labelled nucleotides among a total sequence length of the target fragment of N_b^{RNA} nucleotides where $p \approx 0.5$ is the probability for a uracyl or a cytosine at any position of the target sequence. After substitution of the set average by the overall mean of the number of labels per target by $<N_p^F>_{set} \approx <N_p^F>_{binom} = p \cdot N_b^{RNA}$ one gets the relative fluorescence contribution per sequence position

$$\Delta_p^F \approx \left(1 + \frac{\delta N_p^{F,out}}{\delta N_p^{F,in}}\right) \cdot \Delta_0^F \quad \text{with} \quad \Delta_0^F = \left(\ln 10 \cdot p \cdot N_b^{RNA}\right)^{-1} \approx \frac{2}{\ln 10 \cdot N_b^{RNA}} \tag{A2}$$

Its mean value,

$$\Delta^F \equiv \left\langle \Delta_p^F \right\rangle_{chip} \approx \left(1 + \sqrt{\frac{N_b^{out}}{N_b}}\right) \cdot \Delta_0^F = \left(1 + \sqrt{\frac{N_b^{RNA}}{N_b} - 1}\right) \cdot \Delta_0^F \quad , \tag{A3}$$

provides the average contribution per considered base within a probe sequence of lenth N_b as a function of the total length of the RNA fragment, N_b^{RNA}. The mean incremental contributions are approximated in Eq. A3 by the standard deviation of the binominal distribution according to $<\delta N>_{set} \approx p \cdot N^{0.5}$.

Bases in the probe sequence referring (B=A,G) and not-referring (B=T,C) to labels in the complementary target sequence add and subtract the constant contribution Δ^F to the sensitivity, respectively.

S2: Signal and sensitivity error of single Affymetrix GeneChips

The weighting factor for the least squares fits of the positional dependent sensitivity models is given by the variance of the experimental sensitivity data, $\omega^{P}{}_{p}{}^{2} \approx \mathrm{var}(Y^{P}{}_{p})$ [1]. It can be estimated for each probe from chip replicates using standard error analysis. The SB sensitivity contributions are partly obtained from least square fits of the sensitivity data of single chips. We therefore developed a method, which estimates $\mathrm{var}(Y^{P}{}_{p})$ for each individual chip using selected probe intensities.

The variance of the sensitivity can be directly related to the variance of the respective signal intensity according to Eq. 6, $\mathrm{var}(Y^{P}{}_{p}) \approx \mathrm{var}(\log(I_{p}{}^{P})) + \mathrm{var}(\langle\log(I_{p}{}^{P})\rangle) \approx \mathrm{var}(\log(I_{p}{}^{P}))\,(1+(N_{probe}-1)^{-1}) \approx \mathrm{var}(\log(I_{p}{}^{P}))$ where $N_{probe} = 11 - 20$ is the number of probes per probe set. For the estimation of the chip-specific value of $\omega^{P}{}_{p}{}^{2}$ we make use of the fact that a considerable number of PM and MM probes are present as replicates on each Affymetrix© chip. We identified repeated probes by comparison of all sequences present on the chip. For example, the human HG U133 chip contains 3463 probes in duplicate (2x), 725 in triplicate (3x), 186 fourfold (4x), 77 fivefold (5x), 37 sixfold (6x), 7 sevenfold (7x), 2 ninefold (9x) and one each for 12x, 16x and 20x. We calculated the variance, $\mathrm{var}_{exp}(\log(I_{p}{}^{P}))$ (P=PM, MM) and log-averaged mean intensity, $\langle I_{p}{}^{P}\rangle = \exp(\langle \ln I_{p}{}^{P}\rangle_{replicate})$, for each of these groups of replicates for a selected chip.

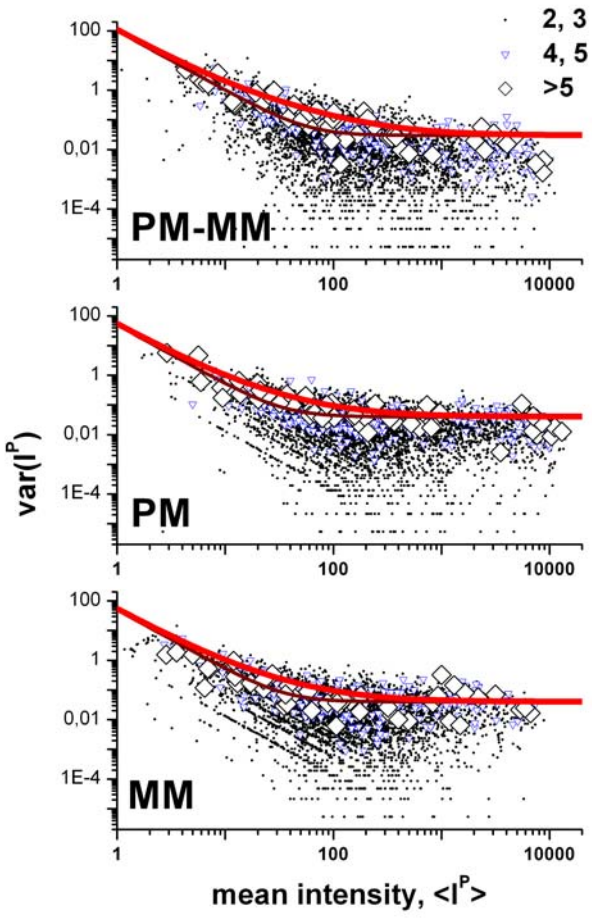

Figure S2: Log-log plot of the variance of the intensities of replicate PM and MM probes present on a HG U133 chip as a function of the mean intensity $\langle I^{P}\rangle$ averaged over probes present in duplicate and triplicate (see legend in the Figure: 2,3…small points), four- and fivefold (4,5…small triangles) and more than fivefold (>5…rhombes). The lines are calculated according to the error model (Eq. A5) with a/b/c = 0.04/5/50 (thick line) and 0.04/0/50 (thin line) for P=PM and MM. The panel above shows the respective analysis of log-intensity differences, PM-MM. In this case the variance is given by $\mathrm{var}(\log I^{PM-MM}) = \mathrm{var}(\log I^{PM} - \log I^{MM})$ and the squared mean intensity by $\langle I^{PM+MM}\rangle^{2} = \langle I^{PM}.I^{MM}\rangle^{-1} = \exp[\langle \ln(I^{PM} + I^{MM})\rangle$. The thick and thin lines are calculated with a/b/c = 0.03/10/100 and 0.03/0/100, respectively.

The uncorrected signal intensity can be rewritten according to Eq. 1 (in the original article) in a simplified version as $I_p^P* \approx (<F_{chip} \cdot N^F \cdot c_{RNA}> + e_F) \cdot \exp[(<\ln K_p^P> + e_G)] + (<\beta_p^P> + e_B)$ where the angular brackets, <…>, denote means over replicated probes. The e_i (i=F, G, B) are error terms and β_p^P is the optical background of each probe, which is not related to hybridization. With $<I_p^P> \approx <F_{chip} \cdot N^F \cdot c_{RNA} \cdot K_p^b>$ one obtains the background-corrected intensity

$$I_p^P \approx I_p^P* - <\beta_p^P> \approx (<I_p^P> + e_F) \cdot \exp(e_G) + e_B \quad . \qquad \text{(A4)}$$

The constant F_{chip} depends on the yield of labelling (fraction of labelled uracyls and cytosines), on the number of oligos per spot and on the efficiency of the detector and of the imaging system (see ref. [2] for details). Consequently the first error term, e_F, considers effects such as variations of the labelling efficiency, of the number of oligos per probe spot and of their density, of the RNA concentration and the noise of the detector and of the imaging system. The exponential term, ε_G, can be rationalized as the error of the free energy of duplex formation, $\Delta G_p^b \propto -\ln K_p^P$, which is related, e.g., to incorrect sequences of individual oligos in each probe spot due to imperfect synthesis and/or to non-equilibrium effects of target binding. The last error, e_B, considers the noise of the detector and of the imaging system in the absence of hybridization.

The variance of log transformed and background corrected signal intensity is described to a good approximation by $\mathrm{var}_{mod}[\log(I_p^P)] \approx a + c/(<I_p^P>)^2$ with $a \approx s_G^2/(\ln 10)^2$ and $c \approx (s_F^2 + s_B^2)$ if one assumes exclusively normally distributed error terms with mean 0 and variance s_i^2 (i=F, G, B). This result agrees with a previously proposed error model of microarray intensity data [3].

Figure S2 compares experimental and theoretical variance data of PM and MM intensities and of their difference in a double-logarithmic scale. The model curves systematically underestimate the experimental variance data in the intermediate intensity range, $100 < I^P < 1000$ (see thin lines in Fig. S2). Considerable better agreement was achieved if one adds a term $\sim <I^P>^{-1}$ according to (see thick lines in Fig. 13)

$$\mathrm{var}_{mod}\left(\log I_p^P\right) \approx a + \frac{b}{<I_p^P>} + \frac{c}{<I_p^P>^2} \quad . \qquad \text{(A5)}$$

The additional term can be tentatively rationalized as non-Gaussian error terms, which contribute to e_F. Here we use Eq. A5 without further specification as an empirical measure to estimate the weighting factor in the sum of squared residuals in the least squares fits as a function of the signal intensity.

The analysis of the intensity difference, $\mathrm{var}(\log I^{PM} - \log I^{MM})$, as a function of $<I^{PM} \cdot I^{MM}>^{-1} = \exp[-<\ln(I^{PM} + I^{MM})>]$ provides similar plots as that for PM and MM (Fig. S2, panel above). The respective background error is however increased whereas the signal error decreases compared with the respective error data of PM and MM probes. This result is compatible with a uncorrelated background noise of PM and MM intensities. In this case one expects for the background error of the log-difference of PM and MM probes a standard deviation of $s_B^2(\text{PM-MM}) \approx s_B^2(\text{PM}) + s_B^2(\text{MM}) \approx 2s_B^2(\text{PM})$, where the arguments PM and MM refer to the log-transformed intensities of the respective probes. On the other hand, the signal error term, “a”, of $\mathrm{var}(\log I^{PM} - \log I^{MM})$ slightly reduces when

compared with that of the individual PM and MM probes. This result can be explained with correlations between the PM and MM intensities, which are discussed in the paper.

References

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

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

(3) Rocke, D. M.; Durbin, B. *J. Comput. Biol.* **2001**, *8*, 557.

S3: Overview of SB free energy parameters of DNA/RNA duolexes

Relations between the positional dependent SB free energy and fluorescence contributions of Watson-Crick (WC) and self complementary (SC) pairings in DNA/RNA oligonucleotide duplexes of the PM and MM microarray probes upon-specific (S) and non-specific (NS) hybridization [a]

Du-plex	Single base contributions			
	probe level	base pair level		
	Probe P= position	PM,MM k≠13	PM k=13	MM k=13
NS	base pairing	WC: B-b^c	WC: B-b^c	WC: B^c-b
	$\varepsilon_{0,k}^{P,NS} \approx$	$\varepsilon_{0,k}^{WC}$	$\varepsilon_{0,13}^{WC}$	$\varepsilon_{0,13}^{WC}$
	$\varepsilon_{0,k}^{PM-MM,NS} \approx$	0	$\varepsilon_{0,13}^{WC-WC} \approx -(\log K_0^{PM,NS} - \log K_0^{MM,NS})$	
	$\Delta\varepsilon_k^{P,NS}(B) \approx$	$\Delta\varepsilon_k^{WC}(B)$	$\Delta\varepsilon_{13}^{WC}(B)$	$\Delta\varepsilon_{13}^{WC}(B^c)$
	$\Delta\varepsilon_k^{PM-MM,NS}(B) \approx$	0	$\Delta\varepsilon_{13}^{WC-WC}(B) = -\Delta\varepsilon_{13}^{WC-WC}(B^c) \equiv \Delta\varepsilon_{13}^{WC}(B) - \Delta\varepsilon_{13}^{WC}(B^c)$ $C \approx T \approx -G \approx -A < 0$	
	$\Delta\varphi_k^{P,NS}(B) \approx$	$\Delta\varphi_k^{WC}(B)$	$\Delta\varphi_{13}^{WC}(B)$	$\Delta\varphi_{13}^{WC}(B^c) = -\Delta\varphi_{13}^{WC}(B)$
	$\Delta\varphi_k^{PM-MM,NS}(B) \approx$	0	$\Delta\varphi_{13}^{WC-WC}(B) \equiv \Delta\varphi_{13}^{WC}(B) - \Delta\varphi_{13}^{WC}(B^c) = 2\Delta\varphi_{13}^{WC}(B)$ $\lvert\Delta\varphi_{13}^{WC-WC}(B)\rvert \approx \lvert\Delta^F\rvert$; $G \approx A \approx -C \approx -T > 0$	
	$\sigma_k^{PM-MM,NS}(B) \approx$	0	$\Delta\varepsilon_{13}^{WC-WC}(B) - \Delta\varphi_{13}^{WC-WC}(B)$	
S	base pairing	WC: B-b^c	WC: B-b^c	SC: <u>B^c</u>-b^c
	$\varepsilon_{0,k}^{P,S} \approx$	$\varepsilon_{0,k}^{WC}$	$\varepsilon_{0,13}^{WC}$	$\varepsilon_{0,13}^{SC}$
	$\varepsilon_{0,13}^{PM-MM,S}$	0	$\varepsilon_{0,13}^{WC-SC} \equiv \varepsilon_{0,13}^{WC} - \varepsilon_{0,13}^{SC} \approx -(\log K_0^{PM,S} - \log K_0^{MM,S})$	
	$\Delta\varepsilon_k^{P,S}(B) \approx$	$\Delta\varepsilon_k^{WC}(B)$	$\Delta\varepsilon_{13}^{WC}(B)$	$\Delta\varepsilon_{13}^{SC}(B^c)$
	$\Delta\varepsilon_k^{PM-MM,S}(B) \approx$	0	$\Delta\varepsilon_{13}^{WC-SC}(B) \equiv \Delta\varepsilon_{13}^{WC}(B) - \Delta\varepsilon_{13}^{SC}(B^c) \approx \Delta\varepsilon_{13}^{WC}(B)$ $C > G \approx T > A$	
	$\Delta\varphi_k^{P,S}(B) \approx$	$\Delta\varphi_k^{WC}(B)$	$\Delta\varphi_{13}^{WC}(B)$	$\Delta\varphi_{13}^{SC}(B^c) = -\Delta\varphi_{13}^{WC}(B^c)$
	$\Delta\varphi_k^{PM-MM,S}(B) \approx$	0	$\Delta\varphi_{13}^{WC-SC}(B) \equiv \Delta\varphi_{13}^{WC}(B) - \Delta\varphi_{13}^{SC}(B^c) \approx 0$	
	$\sigma_k^{PM-MM,S}(B) \approx$	0	$\Delta\varepsilon_{13}^{WC-SC}(B)$	

[a] Single base related free energy (ε), fluorescence (φ) and sensitivity (σ) contributions to the probe intensities. The index k indicates the position of base B=A,T,G,C along the probe sequence. k=13 refers to the middle base whereas k≠13 refers to all positions outside the middle base. The superscript "c" denotes the complementary base, e.g., for B=A one gets of B^c=T. Single See text.