Supporting information

Non-specific hybridization scaling of microarray expression estimates – a physico-chemical approach for chip-to-chip normalization

Hans Binder, Jan Brücker, Conrad Burden

Hook-analysis

The so-called hook method (see [1-3] for a detailed description) applies to microarrays of the GeneChip-type containing pairs of perfect match (PM) and mismatch (MM) probes to estimate the abundance of each transcript. The middle base of the 25meric probes either matches or mismatches the target sequence giving rise to a weaker specific binding constant of the MM. The hook-method independently analyzes the intensity data of each GeneChip microarray using the two-species Langmuir hybridization isotherm which assumes competitive binding of non-specific and specific transcripts to each probe as described above. The method processes the PM and MM probe intensities (I^{PM} and I^{MM}, respectively) using the transformation

$$\begin{aligned}\Delta &= \log I^{PM} - \log I^{MM} \quad ,\\ \Sigma &= \tfrac{1}{2}\left\langle \log I^{PM} + \log I^{MM} \right\rangle_{set}\end{aligned} \quad . \tag{1}$$

The angular brackets $<\ldots>_{set}$ denote averaging over the probe set collecting several probes which address the same transcript. Smoothing of the Δ-versus-Σ plot provides the so-called hook-curve. It enables decomposition of the probe intensities into contributions due specific and non-specific hybridization by simple graphical analysis and subsequent correction of the intensities for sequence specific effects using the positional-dependent nearest neighbour model [4]. The corrected intensities are re-plotted into Δ-versus-Σ coordinates and smoothed to obtain the corrected version of the hook curve which allows identification of absent and present probes using a simple break criterion (see below). In the next step the hook curve is analyzed in terms of the two-species Langmuir binding model which predicts the following parametric equations for the Δ- and Σ-coordinates

$$\begin{aligned}&\Delta(R) = \Delta^{start} + \log\left\{\frac{(R+1)}{\left(R\cdot 10^{-\alpha}+1\right)}\right\} - \log\left\{\frac{B^{PM}(R)}{B^{MM}(R)}\right\}\\ &\text{and} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad .\\ &\Sigma(R) = \Sigma^{start} + \tfrac{1}{2}\log\left\{(R+1)\cdot\left(R\cdot 10^{-\alpha}+1\right)\right\} - \tfrac{1}{2}\log\left\{B^{PM}(R)\cdot B^{MM}(R)\right\}\end{aligned} \tag{2}$$

with the saturation terms $B^{PM}(R) = 1 + 10^{-\left(\beta - \frac{1}{2}\Delta^{start}\right)}(R+1)$ and $B^{MM}(R) = 1 + 10^{-\left(\beta + \frac{1}{2}\Delta^{start}\right)}\left(R\cdot 10^{-\alpha}+1\right)$ and the S/N-ratio of the PM-probes, $R \equiv X^S / X^N$, as the argument. The parameter couples (Σ^{start}, Δ^{start}) and (β, α) characterize the position and the geometrical dimensions of the hook-curve in terms of the coordinates of their starting point and their width and height, respectively. On the other hand, these parameters are related to well-defined hybridization characteristics of the selected chip

$$\alpha = \log\frac{s}{n} \quad , \quad \beta = \tfrac{1}{2}\log n - \left\langle \log X^{PM,N} \right\rangle_{chip} \qquad \text{and} \qquad (3)$$

$$\Delta^{start} = \log n \quad , \quad \Sigma^{start} = \log I_{max} - \beta$$

Here, s and n are the “PM/MM”-gain parameters which are defined as the mean, chip-averaged ratios of the binding constants of the PM and MM probes for specific and non-specific hybridization,

$$s = \left\langle \frac{K^{PM,S}}{K^{MM,S}} \right\rangle_{chip} \quad \text{and} \quad n = \left\langle \frac{K^{PM,N}}{K^{MM,N}} \right\rangle_{chip} \quad , \qquad (4)$$

respectively. I_{max} is the maximum intensity reached at complete saturation of the probes with bound transcripts. The limiting values of the hook-coordinates at vanishing and infinite argument are

$$\Delta(0) \approx \Delta^{start} \quad , \quad \Sigma(0) \approx \Sigma^{start} \quad \text{and} \quad \Delta(\infty) \approx 0 \quad , \quad \Sigma(\infty) \approx \log I_{max} \qquad (5)$$

in the limit of weak nonspecific binding (X^N<<1) and of a common intensity asymptotes of the PM and MM upon saturation, respectively.

Hence, the simple relation between the geometrical dimensions of the hook curve on the one hand and basic hybridization characteristics of the selected chip on the other allows the straightforward evaluation of the particular hybridization by visual inspection of the corresponding hook curve. For example, its width β simply reflects the mean binding strength of non-specific hybridization in a chip-specific fashion.

In the final step of the hook analysis the sequence-corrected probe-level intensity data are corrected for the non-specific background and for saturation effects and then summarized for each probe set to get one transcript-related estimate of the specific binding strength.

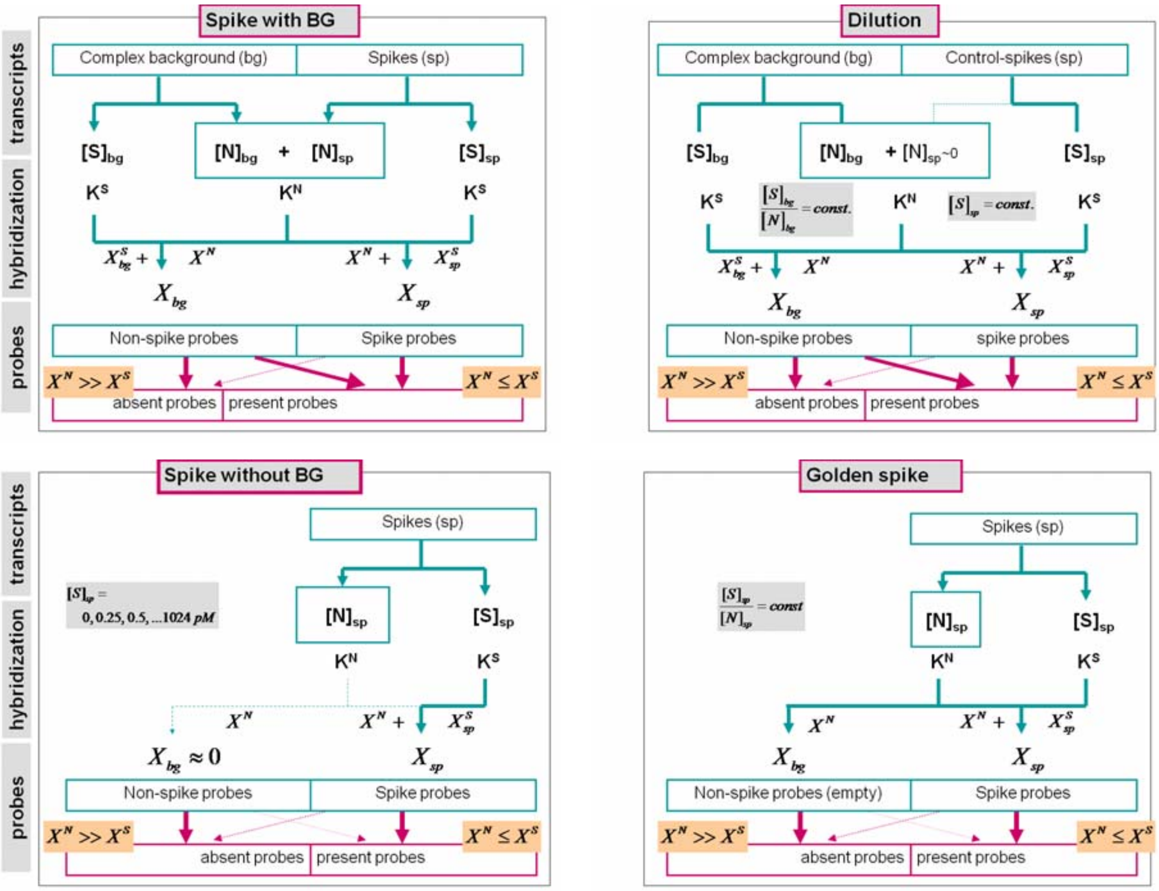
(1) Binder, H.; Preibisch, S.; Berger, H. *Methods in Molecular Medicine* **2008**, *in press*, (preprint: www.izbi.de/izbi/working_papers.php).

(2) Binder, H.; Preibisch, S. *Algorithms for Molecular Biology* **2008**, *3:12*.

(3) Binder, H.; Krohn, K.; Preibisch, S. *Algorithms for Molecular Biology* **2008**, *3:11*.

(4) Binder, H.; Kirsten, T.; Loeffler, M.; Stadler, P. *Proceedings of the German Bioinformatics Conference* **2003**, *2*, 145.

Benchmark experiments – schematic overview



Schematic representation of the benchmark hybridizations analysed in this study: The *spike-with-BG* describes the general case: Both, the complex background and the spikes give rise to non-specific and specific hybridizations of all probes. The amount of hybridization is related to the binding strength which is the product of the respective concentration times the respective binding constant, $X^h = K^h \cdot [h]$ (h= N, S). The total binding strength of each probe additively decomposes into specific and non-specific contributions. The relation between the specific and non-specific binding strength determines whether a probe is absent or present. The "spike" probes are specific, i.e. complementary to the spikes. For the non-specific hybridization we use the one-species approximation which assumes one mean background contribution for all probes. The general spike-with-BG case applies to the LS+BG- and RNA/DNA-experiments. In the *LS-without-BG* and *Golden-spike* experiments only spikes are added. In the former experiment non-specific hybridization contributes only weakly to the hybridization of the probes. The non-spike probes are essentially "empty", i.e. not hybridized. Contrarily, in the Golden-spike experiment the spikes give rise to marked non-specific signals of these empty probes. In the *dilution experiment* the complex background hybridizes all probes.