Adequate usage of Affymetrix background probes on Exon and Gene 1.0 ST arrays

Jan Brücker[1], Peter F. Stadler[1,2,3], Hans Binder[1]
[1]*Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany*
[2]*Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany*
[3]*Santa Fe Institute, Santa Fe, USA*
Email: {bruecker | binder}@izbi.uni-leipzig.de

Problem

The correction of microarray data for the non-specific background is probably the most critical task to extract proper expression measures. Previously we reported an appropriate correction method based on probe design of older GeneChip generations, which use paired perfect match (PM) and mismatch (MM) probes [BKP08, BP08]. Newer chip generations are designed as PM only arrays requiring new methods of background correction. The poster addresses the central issue of selecting a suited set of background probes to train background models. As a second question we judge the quality of positional dependent sensitivity models to describe the background intensities.

Background

Control Probes

Newer Chip Generations of Affymetrix GeneChips, like Human Exon 1.0 ST and Human Gene 1.0 ST abdicate upon Mismatch (MM) probes. Instead they introduced two sets of control probes, both designed to bind transcripts only non-specifically. One set is build on sequences that have no close matches in the human genome. The sequences are taken from bacterial genomes. The other set is build on the sequences of intronic regions of the human genome.

Calibration

Unfortunately, there is no simple, linear relation between measured probe intensities and the respective specific transcript concentrations. Calibration methods aim at correcting the intensities to establish the linear relation between input and output variables of the microarray.

Non-specific Background

All probes on a microarray are effected by transcripts that bind non-specifically. Therefore, calibration methods take into account the non-specific binding to a probe. The quality of the background correction strongly depends on the chosen set of background probes, that is used for training the model.

“Hook” Calibration

We recently published a calibration method for PM/MM Gene Arrays [BKP08, BP08]. An essential step of this “Hook”-Method is to model the non-specific binding in a probe-sequence dependent manner.

Conclusion

The usage of Affymetrix background control probes to estimate sequence dependent background of the probes is suboptimal. A suited set of background probes can be selected by plotting the log-averaged intensity difference between the probe sets and the respective background-control probes as a function of the set averaged probe intensities (hook plot). The relatively small number of ~20.000 probes for either intron or antigenomic probe sets are insufficient to calculate a reliable estimate of the 16 * 24 = 384 parameters of a NN model (or the 64 * 23 = 1427 parameters of a NNN model). The hook method usually finds more than half of a chip’s probes to be unspecific and therefore supplies sufficient data to estimate the parameters of even higher-order models.

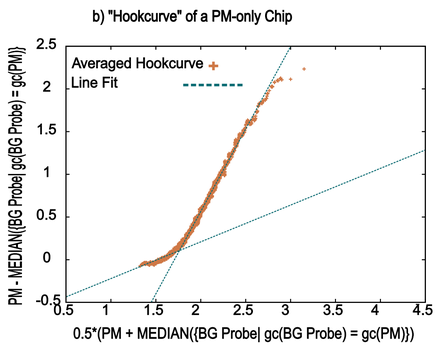
C-rich probes are problematic in terms of the positional dependent sensitivity model. We will consider this deficiency to improve the model by appropriate sequence terms.

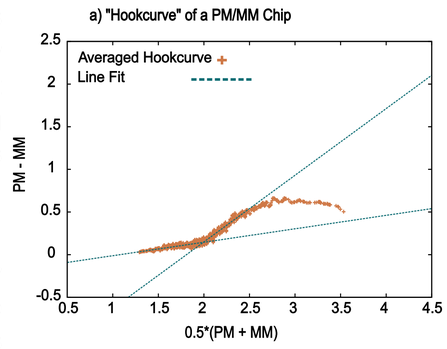
The estimation of the non-specific background thus represents the first step to adapt successful calibration strategies for PM/MM-chips to PM-only arrays.

Selecting Background Probes

Hook Plot for PM/MM Chips (a)

For training a background model, as for PM/MM Gene Arrays, the hook algorithm [BKP08, BP08], identifies probe sets that appear to bind only, or almost only non-specifically. The basic idea is that those probe sets have a rather low intensity and that PM and MM probe sets bind on the average the same amount of random transcripts, i.e. PM-MM ≈ 0. We identify those probe sets by finding the kink in the hook-plot that divides the non-specific “background” probe sets from the specific sets.





Plot for PM-only Chips (b)

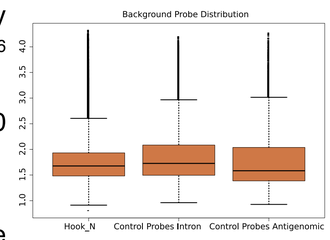
For PM-only chips we can create an analogous plot by substituting the intensity of a MM probe by the median intensity of all background probes with the same GC-content as the PM probe. Also this plot shows a kink at low mean intensities which we use as a simple criterion to select the probes which are predominantly hybridized non-specifically in analogy with the PM/MM hook-plots. Particularly, we consider the probes with intensities smaller than that cut-off as background probes.

Judging the Quality of Background Probes

Intensity Distributions

In this section we compare the properties of the three potential sets of background probes to judge their quality as intensity correction term:

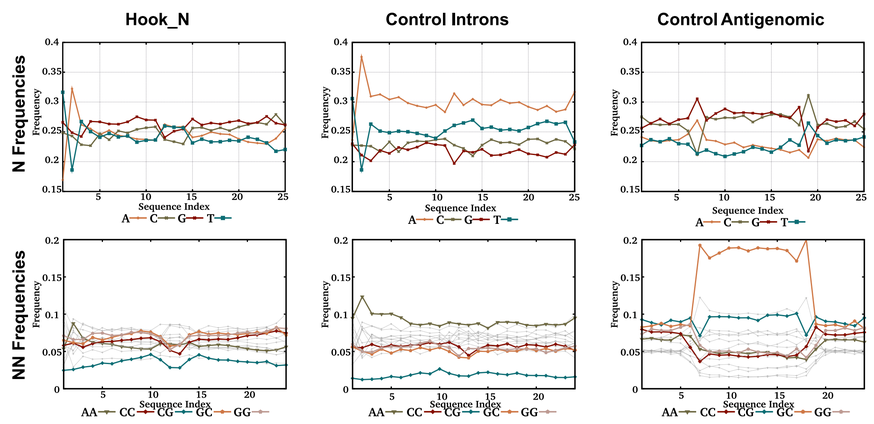
a) selected using the hook method (usually 50-75% of the chip’s probes i.e. several 10^6 probes for exon arrays)
b) bacterial-genome (antigenomic) ~ 17000 probes
c) intronic probes ~ 21000 probes



The intensities of the probes selected by the hook method show the lowest variance among the three studied sets.

Base Frequencies

To detect possible biases in the background probe sets we calculated the position dependent frequencies of single bases and of nearest neighbor motifs in all sets of background probes. The plots show that the Affymetrix Control probe sets have tendencies to over - or underrepresent bases and



motifs at certain positions. Especially GC pairs become highly over-represented in the center positions of the probe sequences of the antigenomic control set.

Base Sensitivities of Background Probes

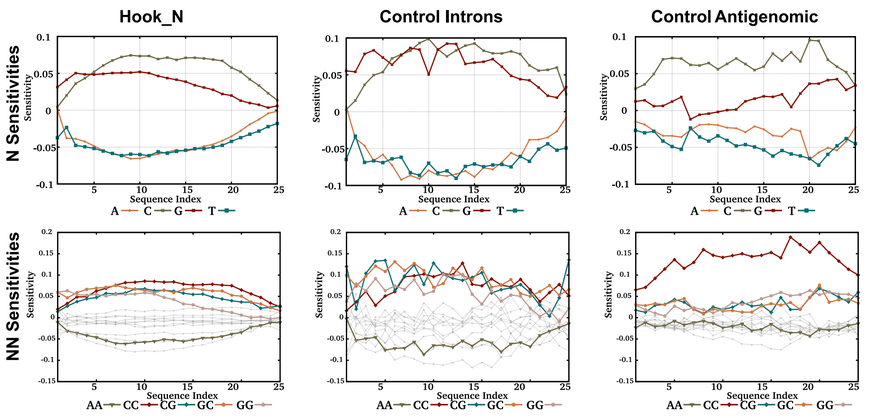
We used a model that decomposes the sensitivity Y_p of each probe into a sum of sensitivity contributions depending on the base at position of the $k = 1, ..., N_b$ probe sequence $\xi_{p,k}$ [BPK05].

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

Here δ denotes the Kronecker delta $(\delta(x,y) = 1 \; if \; x = y, \; \delta(x,y) = 0 \; if \; x \neq y)$ The term $f_k^{\Sigma}(B)$ is the fraction of base B at position k in the considered ensemble of probes.

The sensitivity coefficients $\sigma_k(B)$ of the single base model were determined by means of multiple linear regression, which minimizes the sum of weighted squared residuals between measured and calculated sensitivities. The model was adjusted to higher terms, as NN (nearest neighbor) and NNN (next nearest neighbor).

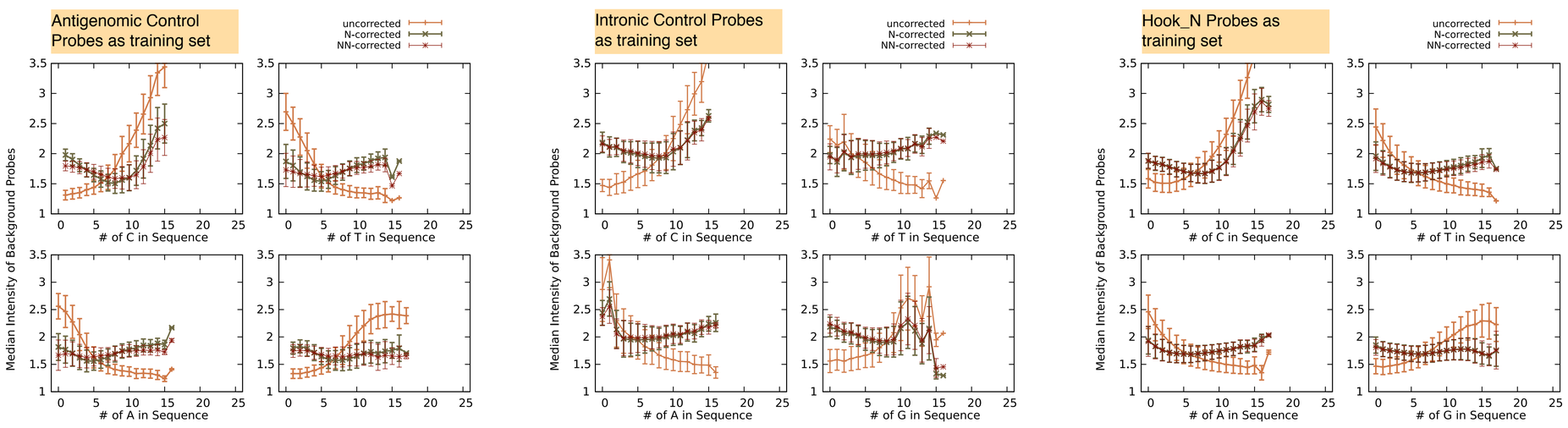
The antigenomic set differs on the level of base sensitivities: The intensities



are mainly governed by CC-motifs. The two other sets are more diverse with respect to their positional-dependent motif characteristics.

Sequence Correction

We tested how well the models trained on the three different sets of background probes (antigenomic-, intronic- and hook background) correct the sequence dependence of the background probes. We used the set of antigenomic background probes to test the correction terms. The positional dependent NN-model properly corrects the probe intensities for their base composition except for cytosine-rich probes with more than 10 C’s per probe. The hook training set corrects much of the base dependent bias.



UNIVERSITÄT LEIPZIG
Printed by Universitätsrechenzentrum Leipzig

References

[BKP08] Hans Binder, Knut Krohn, and Stephan Preibisch. ”Hook”-calibration of GeneChip-microarrays: chip characteristics and expression measures. *Algorithms Mol Biol*, 3:11, 2008.

[BP08] Hans Binder and Stephan Preibisch. ”Hook”-calibration of GeneChip- microarrays: theory and algorithm. *Algorithms Mol Biol*, 3:12, 2008.

[BPK05] H. Binder, S. Preibisch, and T. Kirsten. Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *Langmuir*, 21(20):9287–9302, Sep 2005.