Sequence specific sensitivity of oligonucleotide probes

H. Binder [1]*, T. Kirsten [1], M. Löffler [1], P. Richter [2], P. Stadler [2]

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig,[2] Institute for Informatics, University of Leipzig; * corresponding author: binder@izbi.uni-leipzig.de

Summary

Gene expression analysis by means of high-density-oligonucleotide-array chips is based on the sequence specific binding of mRNA to oligonucleotide probes and its measurement using fluorescence labels. The sensitivity of a probe to detect a certain amount of target mRNA considerably varies among the about 400.000 probes, which are covalently attached to a typical Affymetrix® chip. We present a model which describes the probe intensities as a function of their sequence. The effect of a given base at each position of the probe sequence was specified in terms of a general set of nearest neighbour sensitivity parameters. Consequences of our results for gene expression analysis and chip design were discussed.

Background

The high-density-oligo-Nucleotide-array (HDONA) chip technique uses sequence specific binding (hybridisation) of complementary target mRNA to DNA oligonucleotide probes, which are attached to the chip surface in a well-defined geometrical arrangement. The integral fluorescence intensity of the probe arrays is directly related to the amount of bound, fluorescently labelled mRNA, which in turn serves as a measure of the mRNA concentration in the studied sample solution and thus of the expression degree of a given gene. Several factors such as the binding affinity between target and probe, fluorescence effects, the performance of the detector and of the imaging system affect the measured intensity.

Motivation

The understanding and consideration of these factors in terms of a probe specific sensitivity and the development of suited correction methods represents an essential pre-requisite for the adequate analysis of HDONA chips, which is not solved at present. Existing methods of gene expression (GE) data analysis either completely ignore specific probe sensitivities [1] or they are considered in an empirical fashion without including sequence information [2]. A first attempt to describe the sensitivity as a sum of base and position dependent terms was recently published [3].

Results

In the first step we analysed the intensities of perfect match (PM) and mismatch (MM) probes of Affymetrix chips as function of simple sequence characteristics (middle base, middle triple, number of C, G, A and T per probe, etc.). The results clearly indicate a systematic relationship between the chosen sequence characteristics and the respective intensity, which reflects the probe specific sensitivity. The results suggest that the sensitivity represents a function of the complete base sequence of the probe.

In a second step we correlate the probe intensities and the respective hybridisation affinities referring to oligonucleotides in solution using RNA/DNA heteroduplex free energies which were calculated by means of a nearest neighbour (NN) base pair model [4]. The results show that the NN model must be extended for analysis of HDONA chip data.

The third step includes the least squares fit of the normalized intensity, $Y(PM) = \log PM - \langle \log PM \rangle_{ps}$ ($\langle \ldots \rangle_{ps}$ denotes averaging over the probe set) of the 200.000 PM probes per chip to a modified NN model which in addition considers the position of the bases along the respective probe sequence:

$$Y_{calc} = \sum_{k=1}^{24} \sum_{b,b'=A,T,G,C} A_k(b,b') \cdot \left(\delta(b, S_k^{PM}) \cdot \delta(b', S_{k+1}^{PM}) - f_{k,ps}(b,b') \right) \quad .$$

Here δ denotes the Kronecker delta (δ(x,y)=1 if x=y and δ(x,y)=0 if x≠y) and and S_k^{PM} is the base at position k of the PM probe sequence. $f_{k,ps}(b,b')$ is the fraction of nearest neighbours bb' at position k in the probe set. This analysis provides a set of sensitivity coefficients, $A_k(b,b')$, depending on the neighboured bases bb' (e.g., AA, AT, AG, etc.) and their position, k, along the sequence of the oligomer probe (k=1 is defined as the free end, not attached to the solid support).

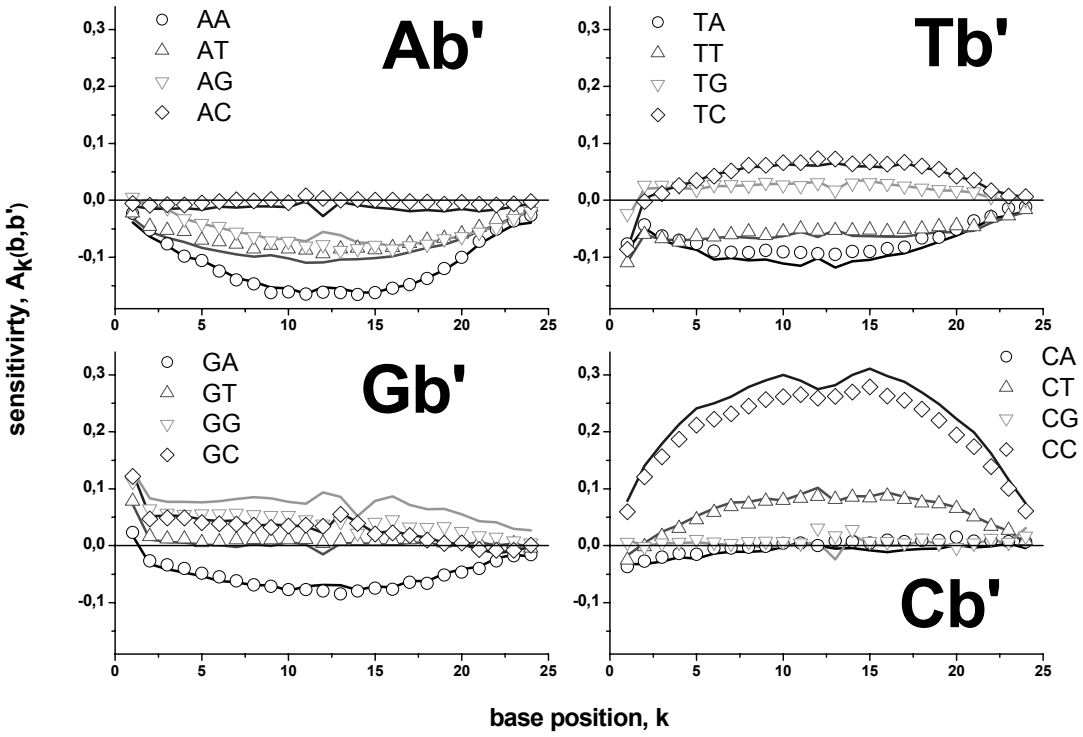

Figure 1: *Sensitivity coefficients of the modified NN model (b'=A,T,G,C). The symbols were obtained by fits of the normalised PM probe intensities of an Affymetrix MG_U74Av2 chip. The lines refer to the fit of the respective normalised MM intensities.*

Most of the obtained sensitivity parameters strongly vary as a function of the base position along the probe sequence (see Figure 1). For example, the sensitivity of the homo-couple of nearest neighbours CC is maximum in the middle of the probe sequence whereas the sensitivity of AA is minimum. In contrast, the sensitivity of GG is maximum at the free end of the probe whereas that of TT is minimum. The sensitivities of hetero-couples, e.g., AT, are well described by a linear combination of the sensitivities of the respective homo-couples, AA and TT in this example. The hetero-couples of nearest neighbours GC and CG deviate from this rule. GC and CG lower the probe intensities to an extraordinary extend. Its effect is maximum in the middle of the probe sequence. The analysis of the MM intensities by means of the same model provides virtually identical sensitivity coefficients compared with the PM data (compare symbols and lines). This result confirms the general character of the obtained parameter set.
The sensitivity parameters correlate with the respective free energy contribution to the NN model for RNA/DNA duplexes in solution. Systematic deviations can be attributed to the effect of

fluorescence labels bound to the target RNA, to folded secondary structures of target and probe prior to duplex formation, to cross hybridisation events and to the attachment of chip probes to the solid support. Consideration of the DNA sequence adjacent to the target motif shows only a weak effect on the observed intensities (e.g., due to bound fluorescence labels), and thus it can be neglected.
The relation between PM and MM probe intensities can be quantified in terms of the mismatch sensitivity change due to the replacement of the middle base in the PM probe by its complementary base in the MM probe. The mean mismatch contribution for G and A are compatible with strong MM intensities (PM<MM) whereas C and T give rise to strong PMs (PM>MM). The mismatch sensitivity change was further refined by consideration of the left and right neighbours of the middle base. According to this analysis the middle triples CGC and CCC provide the smallest and largest differences PM-MM, respectively.

Conclusions and outlook

The modelling of probe intensities as a function of sequence opens up new possibilities to improve GE analysis. Actual issues are the calculation of the mean expression degree for each probe set (referring to one gene), the consideration of MM probes and the selection of subsets of "good" and "bad" probes, which are suited for GE analysis to a different degree. Our results suggest that PM and MM probe intensities comprise similar information about mRNA target concentration, and thus both are suited to quantify the expression degree of a gene. The description of probe sensitivities as a function of their sequence also enables the estimation of the effect of SNPs on PM and MM intensities with consequences for special applications of gene chips. As a second consequence our results can be used for chip design. Consideration of sequence specific sensitivities enables selection of optimal probes from the sequences of target genes. Preliminary results of these issues will be presented.

This work was supported by DFG grant BIC-6 1/1.

1. Affymetrix. (2001). Affymetrix Microarray Suite 5.0, **In:** *User Guide, Affymetrix, Inc., Santa Clara, CA.*
2. Li, C., Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, ***Genome Biology 2****, 1-11.*
3. Naef, F., Magnasco, M. O. (2003). Solving the riddle of the bright mismatches: hybridization in oligonucleotide arrays, ***Physical Review E 68****, 11906.*
4. Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M., Sasaki, M. (1995). Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes., ***Biochemistry 34****, 11211-11216.*