Hybridization biases of microarray expression data

A model-based analysis of RNA quality and sequence effects

Mario Fasold

February 11, 2013

Abstract

Modern high-throughput technologies like DNA microarrays are powerful tools that are widely used in biomedical research. They target a variety of genomics applications ranging from gene expression profiling over DNA genotyping to gene regulation studies. However, the recent discovery of false positives among prominent research findings indicates a lack of awareness or understanding of the non-biological factors negatively affecting the accuracy of data produced using these technologies. The aim of this thesis is to study the origins, effects and potential correction methods for selected methodical biases in microarray data.

The two-species Langmuir model serves as the basal physicochemical model of microarray hybridization describing the fluorescence signal response of oligonucleotide probes. The so-called hook method allows to estimate essential model parameters and to compute summary parameters characterizing a particular microarray sample. We show that this method can be applied successfully to various types of microarrays which share the same basic mechanism of multiplexed nucleic acid hybridization.

Using appropriate modifications of the model we study RNA quality and sequence effects using publicly available data from Affymetrix GeneChip expression arrays. Varying amounts of hybridized RNA result in systematic changes of raw intensity signals and appropriate indicator variables computed from these. Varying RNA quality strongly affects intensity signals of probes which are located at the 3' end of transcripts. We develop new methods that help assessing the RNA quality of a particular microarray sample. A new metric for determining RNA quality, the degradation index, is proposed which improves previous RNA quality metrics. Furthermore, we present a method for the correction of the 3' intensity bias. These functionalities have been implemented in the freely available program package *AffyRNADegradation*.

We show that microarray probe signals are affected by sequence effects which are studied systematically using positional-dependent nearest-neighbor models. Analysis of the resulting sensitivity profiles reveals that specific sequence patterns such as runs of guanines at the solution end of the probes have a strong impact on the probe signals. The sequence effects differ for different chip- and target-types, probe types and hybridization modes. Theoretical and practical solutions for the correction of the introduced sequence bias are provided.

Assessment of RNA quality and sequence biases in a representative ensemble of over 8000 available microarray samples reveals that RNA quality issues are prevalent: about 10% of the samples have critically low RNA quality. Sequence effects exhibit considerable variation within the investigated samples but have limited impact on the most common patterns in the expression space. Variations in RNA quality and quantity in contrast have a significant impact on the obtained expression measurements.

These hybridization biases should be considered and controlled in every microarray experiment to ensure reliable results. Application of rigorous quality control and signal correction methods is strongly advised to avoid erroneous findings. Also, incremental refinement of physicochemical models is a promising way to improve signal calibration paralleled with the opportunity to better understand the fundamental processes in microarray hybridization.