Chapter XIV
The Affymetrix GeneChip® Microarray Platform

Djork-Arné Clevert
Charité Universitaetsmedizin Berlin, Germany
and Johannes Kepler University Linz, Austria

Axel Rasche
Max-Planck-Institute for Molecular Genetics, Germany

ABSTRACT

Readers shall find a quick introduction with recommendations into the preprocessing of Affymetrix GeneChip® microarrays. In the rapidly growing field of microarrays, gene expression, especially the Affymetrix GeneChip arrays, is an established technology present on the market for over ten years. Used in biomedical research, the mass of information demands statistics for its analysis. Here we present the particular design of GeneChip arrays, where much research has already been invested and some validation resources for the comparison of the methods are available. For a basic understanding of the preprocessing, we emphasize the steps, namely: background correction, normalization, perfect match correction, summarization, and couple these with alternative probe-gene assignments. Combined with a recommendation of successful methods a first use of the new technology becomes possible.

INTRODUCTION

Microarrays are the state of the art tool for high-throughput analysis of gene expression. Microarrays allow one to monitor the expression of several thousand genes in parallel in a single experiment facilitating a broad view of the expression state. The genome-wide investigation is basis of the systems biology modeling concept.

The Affymetrix GeneChip® platform was one of the first commercial techniques available on the market. It comes up with a sophisticated design measuring expression of a single gene by several probes on the same chip and providing control sequences for every feature. Chips are available for many species, including popular model species and especially in the biomedical research, where the platform is established. Due to the design and the dissemination of the platform much research has been performed on the analysis of the generated data. It is the main leads and successful results that we wish to describe here.

The vast amount of digital and noisy data generated by microarrays requires statistics for its evaluation. Affymetrix provides basic applications for processing the data and collaborates with companies providing Affymetrix recommended software. On the other hand most of the independent research has been carried out on the R software environment with the BioConductor package collection for statistical computing (Gentleman et al., 2004; R Development Core Team, 2005). In an attempt to be concise we shall focus on R/BioC.

DESIGN OF THE PLATFORM

In the GeneChip approach the expression of a gene is measured by several probes. The probes are selected from the transcript sequence of the respective gene. The UniGene database is the reference for the gene sequence. To avoid cross hybridization between several genes, the sequence of the probes has to be chosen unique to the gene. The length of the probes is always 25 nucleotides.

A number of such probes collected in probe sets stands for independent measurements of the number of transcripts for the gene. The number of probes in a probe set varies between chip platforms. For example in the popular Human Genome U133 Plus 2.0 array there are eleven probes in each probe set.

With the advancement of the human genome sequence and transcript libraries the choice of probe sequences has to be updated from one chip platform to the next. The assignment of the probe sets to genes is updated quarterly and can be retrieved from the NetAffx service on the Affymetrix homepage.

In the classic chip designs, each probe is spotted with its perfect match (PM) sequence and the so-called mismatch (MM) sequence. In the mismatch sequence the 13th nucleotide is altered. The idea is, that the mismatch sequence measures the background expression. The perfect match signal then contains the background expression plus the gene expression. In the newer chips Affymetrix spares the space for additional probes and replaces the mismatches with GC-bins. For a given number of G or C nucleotides (between 0 and 25) the GC-bin contains 25mers unrelated to any gene sequence. The assumption is, that sequences with the same GC content show similar expression behaviour. To make the hybridization

Figure 1. Probe sequences are selected from the transcribed regions of the gene sequence