oligoExpress – exploiting probe level information in Affymetrix GeneChip expression data

Jan Budczies, Joachim Grün

Oligene GmbH, Schumannstr.20/21, Campus Charité Berlin-Mitte, 10117 Berlin, Web: www.oligene.de, Email: budczies@oligene.de

We present oligoExpress, a database system for management and analysis of Affymetrix gene expression data. Often, the evaluation of Affymetrix expression data starts with summary of the probe level measurements to a matrix of expression values (e.g. GCOS signals) that is used as input for all further analyses. However, a lot of other useful information can be extracted from the probe levels measurements. Examples are detection p-values, signal log ratios (SLRs), and change p-values, all introduced by the chip manufacturer (cf. Affymetrix, 2002, Statistical Algorithms Description Document). As an exhaustive exploitation of probe level information allows conducting some statistics on the chip, it can be helpful to keep the number of expensive external replications small. Three different kinds of summaries can be generated from the probe level data: measurements of a single chip, results of pairwise chip comparison, and results of comparisons between biological sample groups. We have developed an `R` application that copes with the enormous amount and the different kinds of the summary data and collects them in a database.

Our software integrates data processing and annotation in an automated workflow that uses the raw data files (cel-files) and a sample annotation file as input. The sample annotation file is an Excel table with names of the cel-files as row names and an arbitrary number of attribute columns. Each attribute corresponds to a (biological) property of the samples and has a value or is “not applicable” for each of the samples. Data processing starts with absolute analysis (signals, detection calls) of each of the chips and comparative analysis (SLRs, change calls) of each possible pair of chips. For absolute analyses the implementation of Affymetrix algorithms in the package `affy` is employed. For comparative analysis we have implemented the corresponding algorithms ourselves, as they have not been available under `R` up to now. As final step of data processing, analysis of biological groups, as they are stored in the annotation file, is performed. For probe set and gene annotation we make use of the package `annaffy`. All data are stored in a relational database that is defined and accessed via the RODBC interface. Microsoft Access is employed as test target database in our first applications, but usage of all other databases with ODBC interface (e.g. MySQL, Oracle) is straightforward.

One of the most common goals of DNA microarray experiments is the detection of differentially expressed genes between two states or types of cells, for example cells from healthy and diseased tissues. As an application of the oligoExpress database system, we have evaluated different procedures for the detection of differential gene expression. The Latin Square data set was downloaded from the Affymetrix homepage and prepared as oligoExpress database. The Latin Square data consist of 3 technical replicates of 14 hybridizations of 42 spiked transcripts in a complex human background. Different procedures including summary statistics for signals, detection calls, signal log ratios and change calls were checked for their performance. Results, recorded in terms of sensitivity and specificity, demonstrated the power of summary statistics based on signal log ratios for the detection of differential transcripts.