M-CHIPS database: Statistical access to experimental conditions

Kurt Fellenberg[1], Nicole Hauser[2], Judith Boer[3], Rosa Arribas-Prat[4], Tim Beissbarth[1,4], Benedikt Brors[1,2], Martin Vingron[1]

German Cancer Research Center (DKFZ), [1]Theoretical Bioinformatics, [2]Functional Genome Analysis, [3]Molecular Genome Analysis, and [4]Molecular Cell Biology, INF 280, 69120 Heidelberg, Germany

Introduction

We present here a database system that has been developed for several collaborating groups within the German Cancer Research Center. Most of these groups are working with radioactive hybridization to membranes on which cDNA fragments have been spotted. Therefore, our system is taylored for these data although it should be readily extensible to two-color hybridization on glass chips. A feature of these database is that data and, most important, annotations can be stored that stem from very different sources and have been obtained by different image analysis systems.

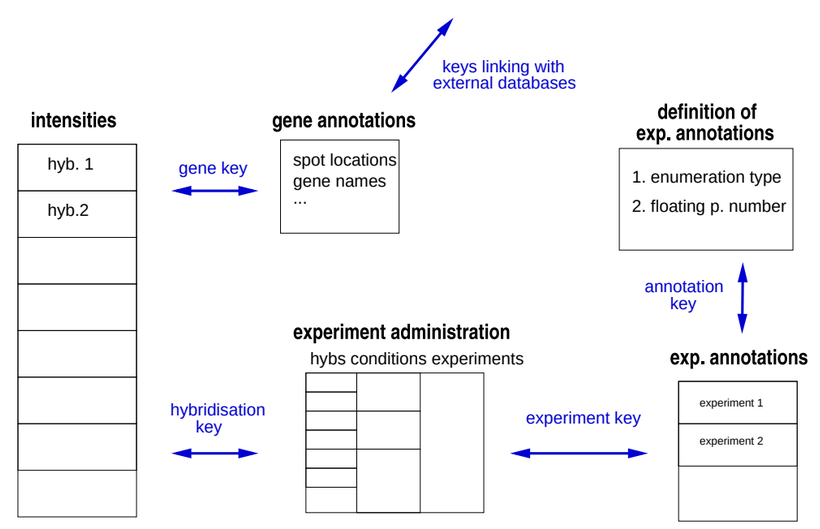
For storing large numbers of microarray data sets, several requirements must be met. A database to hold such data should allow multi-user access while providing data security, i.e. access control by separate read and write permissions on particular data sets granted to different users. It should handle transactions safely which means that inconsistencies resulting from simultaneous access must be prohibited. It should be flexible enough to allow easy addition of annotation categories, and it should scale up reasonably. We present a database system based on a PostgreSQL database server that meets all of these demands. It is equally important that the storage format meets the requirements of data analysis. To this end, we have developed a highly categorized annotation scheme that can be queried to find common annotations among a subset or cluster of experiments. The database is interfaced to M-CHIPS (Multi-Conditional Hybridization Intensity Processing Software), a MATLAB-based tool for pre-analysis processing of microarray intensity data (normalization, quality filtering) as well as for high-end analysis (i.e. clustering) and visualization.

Database Design

The database is made up of three sections: intensity values (which make up 99.8% of the storage space), annotations for genes and annotations for the experimental conditions. Each experiment can consist of several conditions, e.g. different agent concentrations, different time points of a time course, or multiple genotypes of an organism, each of which may have been used to derive several replicates of a hybridization. Consequently, the annotations for experimental conditions have been split into those constant within the entire experiment, those for a single condition and those for a single hybridization. Gene annotations comprise only the spot position on the array together with keys linking the database to sequence databases available online.

A highly categorized annotation scheme has been developed. Only values contained in a predefined list (enumeration-type values) or floating point numbers are allowed. This has the advantage that co-occurence of certain annotation values within a given cluster of experiments can be determined automatically. Thus, you will get an easy answer to the question ‘what makes this set of experiments so specific?’, even if there are hundreds of annotations. Problems resulting from misspelling or multiple identifiers for the same thing do not occur.

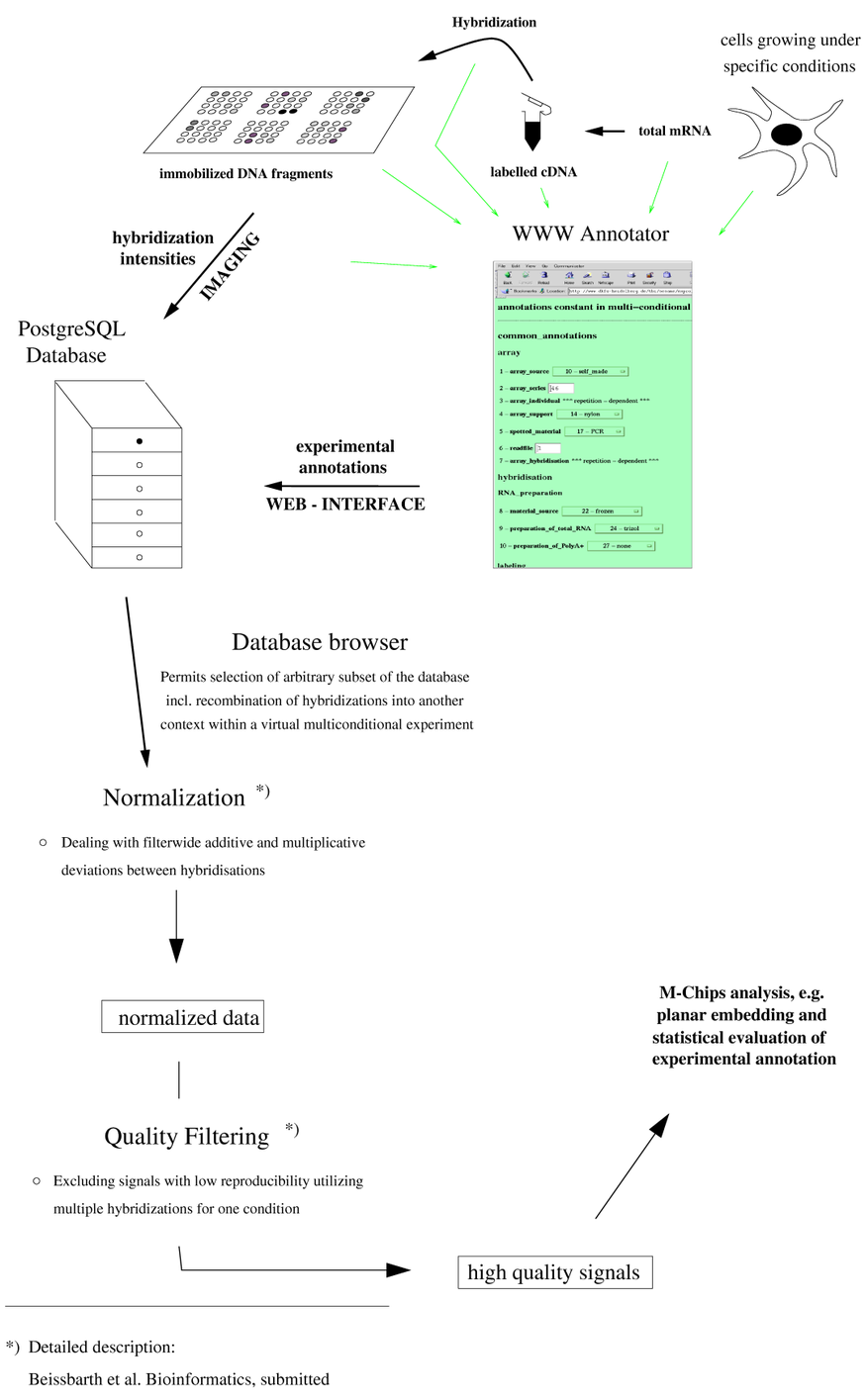
Rough sceme



Integration with M-CHIPS

The PostgreSQL database server is interfaced to M-CHIPS, a MATLAB-based tool for pre-analysis processing, visualization and high-end analysis like clustering and planar embedding. Annotations can be entered into the database via HTML forms that are compiled from the definition tables by CGI scripts (enabling fast addition of annotations or defined values). Intensities are loaded as flat files. Only raw intensities are put in there, background subtraction and normalization is performed from these values on the fly. Thus, improved normalization methods can easily replace old ones since only algorithms are exchanged, nothing needs to be recalculated. M-CHIPS performs also a quality filtering, which means that several quality measures are calculated; thresholds can be set to filter out spots with bad quality where the intensities are not reproducible.

For planar embedding, gene profiles (gene-wise intensity vector across all conditions in an experiment) and hybridization profiles (vector of all intensity values in a hybridization) are projected down to the same plane. Points lying close together in this biplot are assumed to have similar profiles. Given such a cluster of, let’s say, experiments, the genes that cause this similarity to appear are lying in the same direction from the centroid and thus may be easily identified. Furthermore, M-CHIPS will automatically query a cluster of experiments to find annotations that are over- or underrepresented in the cluster compared to the whole set of experiments.



*) Detailed description:
Beissbarth et al. Bioinformatics, submitted

Distribution of Annotation Values

Annotation value △ :
It’s frequency within the cluster is compared to it’s frequency in the whole set:

$$\frac{\#\text{ in cluster}}{\#\text{ all hybridisations of cluster}} \Big/ \frac{\#\text{ in whole set}}{\#\text{ all hybr. of whole set}} = \frac{5}{10} \Big/ \frac{8}{48} = 3\text{x overrepresented in cluster}$$