

An Open Source Client-Server System for the Analysis of Affymetrix Microarray Data

Lars Martin Jakt Mitsuhiro Okada
mjakt@cdb.riken.jp mitu@cdb.riken.jp

Shin-Ichi Nishikawa
nishikawa@cdb.riken.jp

Riken Center for Developmental Biology & The Foundation for Biomedical Research
and Innovation, 2-2-3 Minatojima-minamimachi, Kobe 650-0047, Japan

Keywords: microarray, Affymetrix, client-server, gui

1 Introduction

Microarray data is rich in nature and can often be used to approach a large number of biological questions. This is especially true when data from a broad range of cell types or tissues can be compared easily. We believe that the current limiting factor in the exploitation of microarray data is not technical, but rather the number of biological questions asked of the data. This is mostly a function of the number of biologists that can be coerced into spending time inspecting and questioning the data produced. In order to address this problem we have created a data analysis system that allows the simultaneous analysis of Affymetrix [1] type data by an arbitrary number of researchers. Our system provides tight integration with genomic data from the Ensembl [2] project along with a range of statistical query methods accessed through an easily used graphical environment running on remote clients. We are planning to release this system under the GPL and are actively looking for future collaborators and contributors to the project.

2 System Architecture

Our system consists of 3 basic parts, a PostgreSQL data base server, a server process and a graphical client program. The PostgreSQL server contains all the expression and annotation data and is used by the server process for database lookups on the annotation tables. The server process maintains the expression data in memory and is primarily used for performing statistical analyses on the expression data in addition to facilitating communication between the client process and the backend database server. The graphical client process connects to the server over a network, displays data provided by the server and provides interfaces for the end user to specify database lookups and statistical queries on the expression data.

3 The Expression Data

Affymetrix chips measure gene activity using short pairs of oligonucleotide probes. Each pair contains one member with perfect complementarity (PM) to the target sequence and one member with a single base mismatch (MM) which is used as a background control with the difference between these (PM-MM) being taken as an indicator of the expression level. Probe pairs are arranged into probe sets containing between 11 to 16 probe pairs (depending on chip type) with each gene represented by 0 or

more probe-sets on the chip. Most currently available analysis programs use these 16 difference values to calculate an aggregate expression value which is used in subsequent analysis. Our system, however, treats the individual probe pairs as replicates and considers the individual probe pair profiles across the experimental data. This provides the user with an immediate impression of the quality of the data whilst also providing convenient statistical measures of the reliability of the data (see Figure 1).

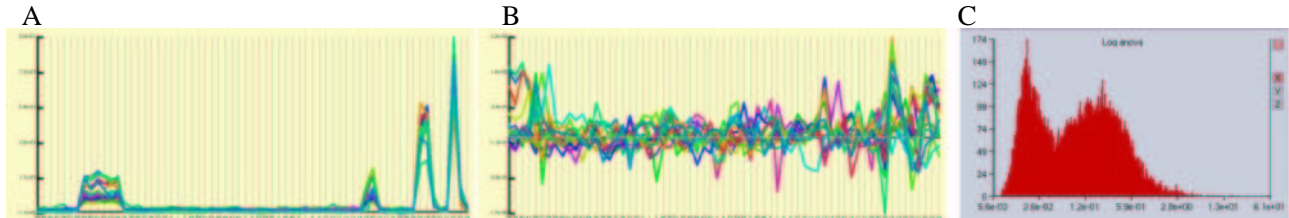


Figure 1: Expression (y-axis) of hemoglobin (a) and gata1 (b) across 65 different cell sources (x-axis). Note the close co-variation of the probe pair profiles in a) as opposed to b). c) Distribution histogram of the log anova score (variation between cell sources as opposed to variation within) for 12,352 probe sets.

4 Genomic Integration

In order to provide annotation and a visualisation of the location of the regions targeted by the probe-sets we compared the probe set sequences provided by Affymetrix with the Ensembl mouse genome sequence. The resulting matches are stored in the database and are used to link the probe set identities to Ensembl gene identities. This allows probe sets to be selected on the basis of the Ensembl annotation, or on the basis of their genomic location. The probe-set matches and the Ensembl gene predictions are displayed by the client program allowing the user to view the organization of the relevant genomic locus, select probe sets with matches in the locus and to select and obtain protein, transcript and genomic sequences from the locus (see Figure 2).

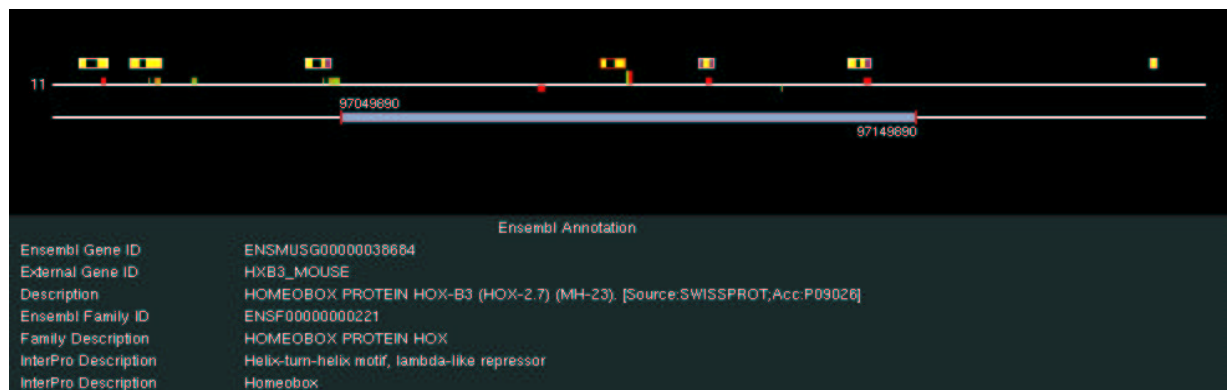


Figure 2: The locus of the current probe set. The chromosome identity is indicated to the left and the specific locations are indicated on the lower white line. Sequence features are indicated above or below the upper white line indicating the relevant strand. Blast matches to probe set sequences are displayed as small rectangles immediately abutting the line, whereas the intron exon structure of genes are shown as a series of yellow (coding) and purple (non-coding) boxes bounded by an open white rectangle.

References

- [1] <http://www.affymetrix.com/>
- [2] <http://www.ensembl.org/>