Bioinformatics Solutions
Using R and Bioconductor

Dr. Alla Bulashevska
AG Prof. Jens Timmer

14.07.2009

• C and Perl were the languages of choice for the first generation of bioinformaticians: massive amounts of sequence data

• R is well suited to scientific challenges in functional genomics which employs technologies to measure abundances of biomolecules under different conditions

• Since 2001 Bioconductor project

Packages for analyzing high-dimensional (time series) data obtained from high-throughput functional genomics assays:

  e.g. expression microarrays, metabolic profiling:

  • Identifying Interesting Genes with siggenes
  • R package limma: Linear Models for Microarray Data
  • Reverse Engineering Genetic Networks using GeneNet
  • Package GAGE: Generally Applicable Gene Set Enrichment
Identifying Interesting Genes with *siggenes*

- Common task in microarray experiments: identification of genes whose expression values differ substantially between groups or conditions - **differentially expressed genes (degs)**
- **multiple testing** problems in which thousands of hypotheses are tested simultaneously
  - Testing statistic and corresponding p-value are computed for each gene
  - Raw p-values are adjusted for multiple testing:
    - **Significance Analysis of Microarrays (SAM)**
      observed test statistics are plotted against expected under null hypotheses; points that differ from diagonal correspond to degs
    - **Empirical Bayes Analysis of Microarrays (EBAM)**
      Models the distribution of observed test statistics as mixture of two components: one for degs and the other for not

R package **limma**: Linear Models for Microarray Data

- A number of summary statistics are computed for each gene and each contrast:
  - **moderated t-statistic**: ratio of log2-fold change to its standard error
    - Standard errors have been moderated across genes
  - **p-value**: obtained from *moderated t-statistic* after adjustment for multiple testing using Benjamini and Hochberg's method to control **False Discovery Rate (FDR)**

- **B-statistic**: log-odds that the gene is differentially expressed
  - B=0 corresponds to 50/50 chance
- **moderated F-statistic**
  - for each gene combines the *t-statistics* for all contrasts tests whether any of the contrasts are non-zero
R Package **GAGE**: Generally Applicable Gene Set Enrichment

- **Gene Set Analysis (GSA)** focuses on sets of related genes
- Incorporates prior knowledge of biological pathways in form of gene sets
- Small coordinated gene expression changes in a pathway can have major biological effect even if these changes are not statistically significant for any individual gene
- Uses all data instead of prefiltering for a short list of strongly differentially expressed genes
- **GAGE** can handle multiple microarray datasets with different sample sizes, experimental designs, profiling techniques

Reverse Engineering Genetic Networks using **GeneNet**

- Output - graph where each gene corresponds to a node and edges depict dependencies between nodes
- Based on **Graphical Gaussian Model (GGM)** which represent multivariate dependencies by means of *partial correlation*
- Model selection - assigning statistical significance to edges using local False Discovery Rate
- Exploratory approach that may help to identify „hubs“ or **clusters of genes** that are functionally related or co-regulated
**Bioconductor** is more than just a repository of biology-related packages!!!

The project has driven a number of technological innovations:

• Management of package dependence hierarchies

• Interfaces between R and other software systems

• Biological metadata (e.g. genome annotations) packages

• More structured data formats than basic data types of R

• Common data structures that allow efficient exchange of data and computational results between different packages
  (Example is ExpressionSet - a class for storage data and info on a microarray experiment)