Data Analysis and Modeling Techniques
Bioinformatics Tools for Microarray Analysis

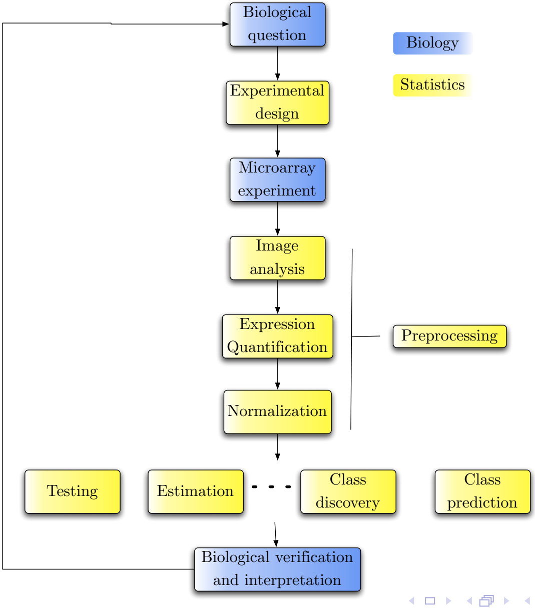
Haibe-Kains B[1,2]   Bontempi G[2]

[1]Microarray Unit, Institut Jules Bordet

[2]Machine Learning Group, Université Libre de Bruxelles

November 27, 2006

Microarray Analysis Design



Bioinformatics Softwares

- **R** is a widely used open source language and environment for statistical computing and graphics
  - Software and documentation are available from `http://www.r-project.org`

- **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data
  - Software and documentation are available from `http://www.bioconductor.org`

Bioconductor Goals

- Provide access to a wide range of powerful statistical and graphical methods for the analysis of genomic data
- Facilitate the integration of biological metadata in the analysis of experimental data: e.g. literature data from PubMed, annotation data from LocusLink
- Allow the rapid development of extensible, scalable, and interoperable software
- Promote high-quality documentation and reproducible research
- Provide training in computational and statistical methods for the analysis of genomic data

BioC and Microarray Experiment

- Two objects represent a microarray experiment :
  - **exprSet** object
  - **phenoData** object
- You can extract from these objects almost all the information about a set of microarray experiments

exprSet

- exprs : matrix of expression levels
- se.exprs : standard errors for gene expressions
- phenoData : phenotypic and/or experimental information
- annotation : base name for the associated annotation
- description : description of the experiment (MIAME)
- notes : set of notes

BioC and Microarray Experiment

exprSet Object : Example

Example of exprSet object :

```
> library(Biobase)
> data(sample.exprSet)
> sample.exprSet
Expression Set (exprSet) with
500 genes
26 samples
phenoData object with 3 variables and 26 cases
varLabels
sex: Female/Male
type: Case/Control
score: Testing Score
```

BioC and Microarray Experiment

phenoData Object and Example

phenoData

- pData : dataframe with phenotypic and/or experimental info
- varLabels : list of labels for the pData variables

Example of phenoData object :

```
> pData(sample.exprSet)[1:3, ]
     sex    type score
A Female Control  0.75
B   Male    Case  0.40
C   Male Control  0.73
```

BioC Functions for Affymetrix© Data

- Most of the functions are in the **affy** package

```
> library(affy) #load the library
> library(help=affy) #help about the library contents
```

- Raw data from Affymetrix© platform are in text files called CEL
- CEL files contain the expressions and the position at the probe level
- To be used by the functions in the affy package, the CEL files have to be read in an **AffyBatch** object

Affy BioC Functions

Suite

AffyBatch structure (see `?AffyBatch`)

- cdfName : object of class `character` representing the name of CDF file associated with the arrays in the `AffyBatch` (e.g. hgu133plus2)
- exprs : object of class `matrix` inherited from `exprSet`. The matrix contains one probe per row and one chip per column
- phenoData : object of class `phenoData` inherited from `exprSet`
- annotation : object of class `character` identifying the annotation that may be used for the chips
- description : object of class `MIAME` (Minimal Information About Microarry Experiment)

Affy BioC Functions

Suite

- AffyBatch creation (see `?read.affybatch`)

```
abatch <- read.affybatch(filenames, phenoData,
 description, verbose=TRUE)
```

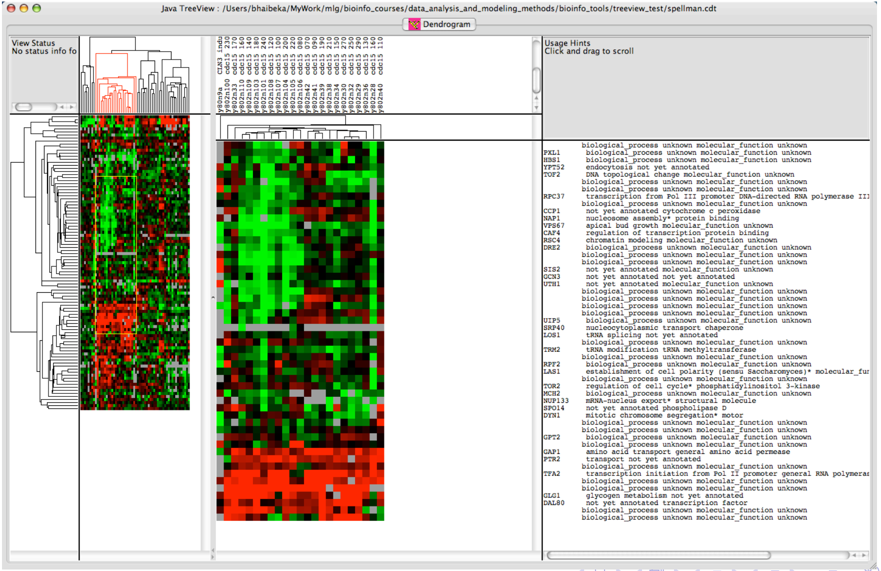
- Remarks :
  - the AffyBatch class is an extension of the exprSet class
  - `filenames` is an object of class `character` containing the whole paths to CEL files
  - all CEL files have to come from the same chip (e.g. hgu133plus2)

Clustering Softwares

- You can use R and the libraries **amap** and **ctc**
- **Java Treeview** is an open source software for clustering visualization
  - Software and documentation are available from `http://jtreeview.sourceforge.net`
- **Cluster3** is a open source clustering software with GUI
  - Software and documentation are available from `http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv`

Clustering Softwares

Java TreeView Snapshot



```
          biological_process unknown molecular_function unknown
PXL1       biological_process unknown molecular_function unknown
HBS1       biological_process unknown molecular_function unknown
YPT52      endocytosis not yet annotated
TOF2       DNA topological change molecular_function unknown
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
RPC37      transcription from Pol III promoter DNA-directed RNA polymerase III
          biological_process unknown molecular_function unknown
CCP1       not yet annotated cytochrome c peroxidase
NAP1       nucleosome assembly* protein binding
VPS67      apical bud growth molecular_function unknown
CAF4       regulation of transcription protein binding
RSC4       chromatin modeling molecular_function unknown
DRE2       biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
SIS2       not yet annotated molecular_function unknown
GCN3       not yet annotated not yet annotated
UTH1       not yet annotated molecular_function unknown
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
UIP5       biological_process unknown molecular_function unknown
SRP40      nucleocytoplasmic transport chaperone
LOS1       tRNA splicing not yet annotated
          biological_process unknown molecular_function unknown
TRM2       tRNA modification tRNA methyltransferase
          biological_process unknown molecular_function unknown
RPF2       biological_process unknown molecular_function unknown
LAS1       establishment of cell polarity (sensu Saccharomyces)* molecular_fur
          biological_process unknown molecular_function unknown
TOR2       regulation of cell cycle* phosphatidylinositol 3-kinase
MCH2       biological_process unknown molecular_function unknown
NUP133     mRNA-nucleus export* structural molecule
SPO14      not yet annotated phospholipase D
DYN1       mitotic chromosome segregation* motor
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
GPT2       biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
GAP1       amino acid transport general amino acid permease
PTR2       transport not yet annotated
          biological_process unknown molecular_function unknown
TFA2       transcription initiation from Pol II promoter general RNA polymeras
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
GLG1       glycogen metabolism not yet annotated
DAL80      not yet annotated transcription factor
          biological_process unknown molecular_function unknown
          biological_process unknown molecular_function unknown
```

Links

- Course web page : http://www.bioinfomaster.ulb.ac.be/cursus/index_html/en#DATANA
- Personal homepage : http://www.ulb.ac.be/di/map/bhaibeka/
- This presentation : http://www.ulb.ac.be/di/map/bhaibeka/bioinfo_courses/bioinfo_tools_pres_hkb.pdf

Thank you for your attention.