

Bioconductor

**Course in Practical Microarray Analysis
Berlin 27.-30.1.2003**

**Slides ©2002 Sandrine Dudoit, Robert Gentleman.
Adapted by Wolfgang Huber.**

Statistical computing

Applications

- statistical design and analysis:
 - technology development and validation, data pre-processing, estimation, testing, clustering, prediction, etc.
- integration with biological information resources (in house and external databases)
 - gene annotation (Unigene, LocusLink)
 - graphical (pathways, chromosome maps)
 - patient data, tissue banks

Outline

- Overview of Bioconductor packages
 - Biobase
 - annotate
 - marrayClasses, ...Input, ...Norm, ...Plots
 - affy
 - vsn
- Dynamic statistical reports using Sweave:
'reproducible analyses'

Bioconductor

- Bioconductor is an open source project to design and provide high quality software and documentation for bioinformatics.
- Current focus: microarrays and gene (transcript) annotation
- Mostly R, but other languages/platforms also possible
- Open to (your?) contributions / feedback
- Software and documentation are available from www.bioconductor.org.

Bioconductor packages

- General infrastructure
 - `Biobase`
 - `annotate`, `AnnBuilder`
 - `tkWidgets`
- Pre-processing for Affymetrix data
 - `affy`, `vsn`.
- Pre-processing for cDNA data
 - `marrayClasses`, `marrayInput`, `marrayNorm`,
`marrayPlots`, `vsn`.
- Differential expression
 - `eddi`, `genefilter`, `multtest`, `ROC`.
- etc.

How to use

- Short courses
- Vignettes
 - Problem-oriented "How-To"s
- R demos
 - e.g. `demo(marrayPlots)`
- R help system
 - interactive with browser or printable manuals;
 - detailed description of functions and examples;
 - E.g. `help(maNorm)`, `? marrayLayout`.
- Search Mailing list archives; Google
- Post to mailing list

All on WWW.

Biobase

contains class definitions and infrastructure classes:

- **phenoData**: sample covariate data (e.g. cell treatment, tissue origin, diagnosis)
- **miame** (minimal information about marray experiments)
- **exprSet**: matrix of expression data, phenoData, miame, and other quantities of interest.
- **aggregate**: an infrastructure to put an aggregation procedure (cross-validation, bootstrap) on top of any analysis

exprSet

Basic data structure for storing results from a **series of microarrays**: intensities, patient (sample) data, gene identifiers .

Transparent subsetting w.r.t. genes and samples.

Slots:

exprs: matrix

phenoData: contains dataframe with patient data

annotate

Goal: associate experimental data with available meta data, e.g. gene annotation, literature.

Tasks:

associate vendor identifiers (Affy, RZPD, ...) to other identifiers

associate transcripts with biological data such as chromosomal position of the gene

associate genes with published data (PubMed).

produce nice-to-read tabular summaries of analyses.

PubMed

www.ncbi.nlm.nih.gov

- For any gene there is often a large amount of data available from PubMed.
- We have provided the following tools for interacting with PubMed.
 - `pubMedAbst`: defines a class structure for PubMed abstracts in R.
 - `pubmed`: the basic engine for talking to PubMed.
- **WARNING**: be careful you can query them too much and be banned!

PubMed: high level tools

- `pm.getabst`: obtain (download) the specified PubMed abstracts (stored in XML).
- `pm.titles`: select the titles from a set of PubMed abstracts.
- `pm.abstGrep`: regular expression matching on the abstracts.

Data rendering

- A simple interface, [11.htmlpage](#), can be used to generate a webpage for your own use or to send to other scientists involved in the project.
- open-source: modify to your needs / taste!

BioConductor Gene Listing

Top 100 genes orderd by t-test

Locus Link Genes

locus link	Effect Size	p-value	adjusted p-value	v1	v2	v3	v4	nv1	nv2	nv3	nv4
NA	0.032	1.2e-10	1e-06	15	15	15	15	15	15	15	15
NA	0.032	1.2e-10	1e-06	15	15	15	15	15	15	15	15
81822	0.032	1.2e-10	1e-06	15	15	15	15	15	15	15	15
50683	0.032	1.2e-10	1e-06	15	15	15	15	15	15	15	15
NA	0.032	1.2e-10	1e-06	15	15	15	15	15	15	15	15
NA	0.0313	1.3e-10	1.1e-06	15	15	15	15	15	15	15	15
25348	0.0313	1.3e-10	1.1e-06	15	15	15	15	15	15	15	15
7	0.032	3.3e-08	0.00029	15	15	15	15	15	15	15	15
24472	-1.48	1.4e-06	0.012	12	11	11	11	14	14	15	14
NA	0.0308	5.2e-06	0.045	15	15	15	15	15	15	15	15
12	0.0283	5.6e-06	0.048	15	15	15	15	15	15	15	15
NA	0.0279	6.9e-06	0.059	15	15	15	15	15	15	15	15
24472	-1.2	7.9e-06	0.067	13	13	12	12	15	15	15	14
19	-0.92	1.1e-05	0.091	11	11	10	10	12	12	13	12
8	1.26	1.1e-05	0.094	9.8	8.9	14	15	9.6	9.4	9	9.2
NA	0.895	1.6e-05	0.13	12	12	15	15	12	12	12	12
70940	0.709	1.7e-05	0.14	10	11	10	10	11	11	12	12

Data packages

The Bioconductor project develops and deploys packages that contain only data.

Available: Affymetrix hu6800, hgu95a, hgu133a, mgu74a, rgu34a, KEGG, GO

These packages contain many different mappings between relevant data, e.g.

KEGG: EnzymeID - GO Category

hgu95a: Affy Probe set ID - EnzymeID

Update: simply by R function `update.packages()`

dataset: hgu95a

maps to LocusLink, GenBank, gene symbol, gene Name.

chromosomal location, orientation.

maps to KEGG pathways, to enzymes.

data packages will be updated and expanded regularly as new or updated data become available.

Diagnostic plots and normalization for cDNA microarrays

(S Dudoit, Y Yang, T Speed, et al)

- **marrayClasses** :
 - class definitions for microarray data objects and basic methods
- **marrayInput**:
 - reading in intensity data and textual data describing probes and targets;
 - automatic generation of microarray data objects;
 - widgets for point & click interface.
- **marrayPlots**: diagnostic plots.
- **marrayNorm**: robust adaptive location and scale normalization procedures.

`marrayPlots` package:

`vExplorer()`

marrayInput package

- Start from
 - image quantitation data, i.e., output files from image analysis software, e.g., `.gpr` for GenePix or `.spot` for Spot.
 - Textual description of probe sequences and target samples, e.g., `gal` files, `god` lists.
- `read.marrayLayout`, `read.marrayInfo`, and `read.marrayRaw`: read microarray data into R and create microarray objects of class `marrayLayout`, `marrayInfo`, and `marrayRaw`, resp.

marrayNorm package

normalization for a batch of arrays

- simple global scaling methods
- intensity or A -dependent location normalization (`maNormLoess`);
- pin- oder platewise

vsn package

normalization for a batch of arrays

- for each array and/or color, estimate calibration offset and scaling factor
- variance stabilizing transformation

With Affymetrix data: combine with affy package

Multiple hypothesis testing

- Bioconductor R `multtest` package
- Multiple testing procedures for controlling
 - FWER: Bonferroni, Holm (1979), Hochberg (1986), Westfall & Young (1993) maxT and minP.
 - FDR: Benjamini & Hochberg (1995), Benjamini & Yekutieli (2001).
- Tests based on t - or F -statistics for one- and two-factor designs.
- Permutation procedures for estimating adjusted p-values.
- Documentation: tutorial on multiple testing.

Sweave

- The Sweave framework allows dynamic generation of statistical documents intermixing documentation text, code and code output (textual and graphical).
- Fritz Leisch's **Sweave** function from R **tools** package.
- See ? Sweave and manual
<http://www.ci.tuwien.ac.at/~leisch/Sweave>

Sweave input

Source: a text file which consists of a sequence of documentation and code segments ('chunks')

- **Documentation chunks**

- start with @
- can be text in a markup language like LaTeX.

- **Code chunks**

- start with <<name>>=
- can be R or S-Plus code.

- **File extension:** .rnw, .Rnw, .snw, .Snw.

Sweave output

After running Sweave and Latex, obtain a single document, e.g. pdf file containing

- the documentation text
- the R code
- the code output: text and graphs.

The document can be **automatically regenerated** whenever the data, code or text change.

Ideal medium for the communication of data **analyses** that want to be **reproducible** by other researchers: they can read the document and at the same time have the code chunks executed by their computer!

Sweave

