Introduction To Bioconductor

Sandrine Dudoit, Robert Gentleman, and Rafael Irizarry

Bioconductor Workshop
Fred Hutchinson Cancer Research Center
December 4-6, 2002

© Copyright 2002, all rights reserved
Bioconductor Basics

• Bioconductor (www.bioconductor.org) is a software project aimed at providing high quality, innovative software tools appropriate for computational biology
• We rely mainly on R (www.r-project.org) as the computational basis
• we welcome contributions
Some basics

• for microarray data analysis we have assembled a number of R packages that are appropriate to the different types of data and processing

• some issues:
  – data complexity
  – data size
  – data evolution
  – meta-data
Software Design

- to overcome complexity we use two strategies: Abstract Data Types and object oriented programming
- to deal with data evolution we have separated the biological meta-data from the experimental data
Pedagogy

• among the many choices we made in the Bioconductor project is to try and develop better teaching materials

• in large part this is because we are between two disciplines (Biology and Statistics) and most users are familiar with only one of these
Vignettes

• we have adopted a new type of documentation: the vignette

• a vignette is an integrated collection of text and code – the code is runnable and using Sweave it is possible to replace the code with its output

• these documents are short and explicit directions on how to perform specific tasks
Vignettes – HowTo’s

• a good way to find out how to use Bioconductor software is to read the relevant Vignette
• then extract the code (\texttt{tangleToR}) and examine it
• HowTo documents are shorter (one or two pages)
• please write and contribute these
Vignettes

• in Bioconductor 1.1 we introduced two new methods to interact with Vignettes

• openVignette() – gives you a menu to select from

• vExplorer() – our first attempt at turning Vignettes into interactive documents
Bioconductor packages
Release 1.1, Nov. 18, 2002

- General infrastructure:
  Biobase, rhdf5, tkWidgets, reposTools.
- Annotation:
  annotate, AnnBuilder → data packages.
- Graphics:
  geneplotter, hexbin.
- Pre-processing for Affymetrix oligonucleotide chip data:
  affy, CDF packages, vsn.
- Pre-processing for cDNA microarray data:
  marrayClasses, marrayInput, marrayNorm, marrayPlots, vsn.
- Differential gene expression:
  edd, genefilter, multtest, ROC.
Outline

• Biobase and the basics

• annotate and AnnBuilder packages

• genefilter package

• multtest package

• R clustering and classification packages
**Biobase: exprSet class**

- **exprs**: Matrix of expression measures, genes x samples
- **se.exprs**: Matrix of SEs for expression measures
- **phenoData**: Sample level covariates, instance of class `phenoData`
- **annotation**: Name of annotation data
- **description**: Object of class MIAME
- **notes**: Any notes
> golubTest
Expression Set (exprSet) with
  7129 genes
  34 samples
    phenoData object with 11
variables and 34 cases
  varLabels
    Samples: Samples
    ALL.AML: ALL.AML
    BM.PB: BM.PB
    T.B.cell: T.B.cell
    FAB: FAB
    Date: Date
    Gender: Gender
    pctBlasts: pctBlasts
    Treatment: Treatment
    PS: PS
    Source: Source

Typing the name of the data set produces this output
exprSet

- the set is closed under subsetting operations (either x[,1] or x[1,]) both produce new exprSets
- the first subscript is for genes, the second for samples
- the software is responsible for maintaining data integrity
exprSet: accessing the phenotypic data

- phenotypic data is stored in a special class: `phenoData`
- this is simply a dataframe and a set of associated labels describing the variables in the dataframe
Annotation packages

• One of the largest challenges in analyzing genomic data is associating the experimental data with the available metadata, e.g. sequence, gene annotation, chromosomal maps, literature.

• The `annotate` and `AnnBuilder` packages provides some tools for carrying this out.

• These are very likely to change, evolve and improve, so please check the current documentation - things may already have changed!
Annotation packages

• Annotation data packages;
• Matching IDs using environments;
• Searching and processing queries from WWW databases
  – LocusLink,
  – GenBank,
  – PubMed;
• HTML reports.
WWW resources

• Nucleotide databases: e.g. GenBank.
• Gene databases: e.g. LocusLink, UniGene.
• Protein sequence and structure databases: e.g. SwissProt, Protein DataBank (PDB).
• Literature databases: e.g. PubMed, OMIM.
• Chromosome maps: e.g. NCBI Map Viewer.
• Pathways: e.g. KEGG.
• Entrez is a search and retrieval system that integrates information from databases at NCBI (National Center for Biotechnology Information).
NCBI Entrez

**annotate: matching IDs**

Important tasks

- Associate manufacturers probe identifiers (e.g. Affymetrix IDs) to other available identifiers (e.g. gene symbol, PubMed PMID, LocusLink LocusID, GenBank accession number).
- Associate probes with biological data such as chromosomal position, pathways.
- Associate probes with published literature data via PubMed.
<table>
<thead>
<tr>
<th><strong>annotate</strong>: matching IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affymetrix identifier</strong></td>
</tr>
<tr>
<td><strong>HGU95A chips</strong></td>
</tr>
<tr>
<td><strong>LocusLink, LocusID</strong></td>
</tr>
<tr>
<td><strong>GenBank accession #</strong></td>
</tr>
<tr>
<td><strong>Gene symbol</strong></td>
</tr>
<tr>
<td><strong>PubMed, PMID</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Chromosomal location</strong></td>
</tr>
</tbody>
</table>
Annotation data packages

- The Bioconductor project has started to deploy packages that contain only data. E.g. `hgu95a` package for Affymetrix HGU95A GeneChips series, also, `hgu133a`, `hu6800`, `mgu74a`, `rgu34a`.
- These data packages are built using `AnnBuilder`.
- These packages contain many different mappings to interesting data.
- They are available from the Bioconductor website and also using `update.packages`.

Annotation data packages

• Maps to GenBank accession number, LocusLink LocusID, gene symbol, gene name, UniGene cluster.
• Maps to chromosomal location: chromosome, cytoband, physical distance (bp), orientation.
• Maps to KEGG pathways, enzymes, Gene Ontology Consortium (GO).
• Maps to PubMed PMID.
• These packages will be updated and expanded regularly as new or updated data become available.
hu6800 data package

A function to return a vector of rda file names

Annotation data file for hu6800 on ACCNUM
Annotation data file for GObyNum on AFFYCOUNTS
Annotation data file for hu6800 on CHR
Annotation data file for hu6800 on CHRLOC
Annotation data file for hu6800 on CHRORI
Annotation data file for hu6800 on ENZYME
Annotation data file for hu6800 on ENZYME3AFFY
Annotation data file for hu6800 on GENENAME
Annotation data file for hu6800 on GO
Annotation data file for GObyNum on GO2AFFY
Annotation data file for GObyNum on GO2ALLAFFY
Annotation data file for hu6800 on GRIFF
Annotation data file for hu6800 on LOCUSID
Annotation data file for hu6800 on MAP
Annotation data file for hu6800 on PATH
Annotation data file for GObyNum on PATH2AFFY
Annotation data file for hu6800 on PMID
Annotation data file for hu6800 on PMID2AFFY
Annotation data file for hu6800 on SUMFUNC
Annotation data file for hu6800 on SYMBOL
Annotation data file for hu6800 on UNIGENE
annotate: matching IDs

• Much of what `annotate` does relies on matching symbols.
• This is basically the role of a `hash table` in most programming languages.
• In R, we rely on `environments` (they are similar to hash tables).
• The annotation data packages provide R environment objects containing `key` and `value` pairs for the mappings between two sets of probe identifiers.
• Keys can be accessed using the R `ls` function.
• Matching values in different environments can be accessed using the `get` or `multiget` functions.
**annotate**: matching IDs

E.g. *hgu95a* package.

- To load package `library(hgu95a)`
- For info on the package and list of mappings available
  ```
  ? hgu95a
  hgu95a()
  ```
- For info on a particular mapping
  ```
  ? hgu95aPMID
  ```
annotate: matching IDs

> library(hgu95a)
> get("41046_s_at", env = hgu95aACCNUM)
[1] "X95808"
> get("41046_s_at", env = hgu95aLOCUSID)
[1] "9203"
> get("41046_s_at", env = hgu95aSYMBOL)
[1] "ZNF261"
> get("41046_s_at", env = hgu95aGENENAME)
[1] "zinc finger protein 261"
> get("41046_s_at", env = hgu95aSUMFUNC)
[1] "Contains a putative zinc-binding motif (MYM) | Proteome"
> get("41046_s_at", env = hgu95aUNIGENE)
[1] "Hs.9568"
annotate: matching IDs

> get("41046_s_at", env = hgu95aCHR)
[1] "X"
> get("41046_s_at", env = hgu95aCHRLOC)
[1] "66457019@X"
> get("41046_s_at", env = hgu95aCHRORI)
[1] "-@X"
> get("41046_s_at", env = hgu95aMAP)
[1] "Xq13.1"
> get("41046_s_at", env = hgu95aPMID)
[1] "10486218" "9205841" "8817323"
> get("41046_s_at", env = hgu95aGO)
[1] "GO:0003677" "GO:0007275"
annotate: database searches and report generation

- Provide tools for searching and processing information from various biological databases.
- Provide tools for regular expression searching of PubMed abstracts.
- Provide nice HTML reports of analyses, with links to biological databases.
annotate: WWW queries

• Functions for querying WWW databases from R rely on the `browseURL` function

`browseURL("www.r-project.org")`
annotate: GenBank query


• Given a vector of GenBank accession numbers or NCBI UIDs, the `genbank` function
  – opens a browser at the URLs for the corresponding GenBank queries;
  – returns an `XMLdoc` object with the same data.

`genbank("X95808", disp="browser")`

`genbank(1430782, disp="data", type="uid")`
annotate: LocusLink query

www.ncbi.nlm.nih.gov/LocusLink/

- **locuslinkByID**: given one or more LocusIDs, the browser is opened at the URL corresponding to the first gene.

  locuslinkByID(“9203”)

- **locuslinkQuery**: given a search string, the results of the LocusLink query are displayed in the browser.

  locuslinkQuery(“zinc finger”)
  http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi?Q=zinc finger&ORG=Hs&V=0
**annotate**: PubMed query


- For any gene there is often a large amount of data available from PubMed.
- The `annotate` package provides the following tools for interacting with PubMed
  - `pubMedAbst`: 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!
annotate: pubMedAbst class

Class structure for storing and processing PubMed abstracts in R

• authors
• abstText
• articleTitle
• journal
• pubDate
• abstUrl
annotate: high level tools for PubMed query

- **pm.getabst**: download the specified PubMed abstracts (stored in XML) and create a list of `pubMedAbst` objects.
- **pm.titles**: extract the titles from a set of PubMed abstracts.
- **pm.abstGrep**: regular expression matching on the abstracts.
annotate: PubMed example

```r
pmid <- get("41046_s_at", env=hgu95aPMID)
pubmed(pmid, disp="browser")


absts <- pm.getabst("41046_s_at", base="hgu95a")
pm.titles(absts)
pm.abstGrep("retardation",absts[[1]])
```
annotate: PubMed example

RGui - [R Console]

Slot "articleTitle":
[1] "Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can

Slot "journal":
[1] "DMB Hem"

Slot "pubdate":
[1] "Apr 1997"

Slot "abstract":
[1] "No URL Provided"

[[3]]
An object of class "pubMedEntry"
Slot "authors":
[1] "S M SM van der Horst" "H IH Scholten" "F F H Hu" "C C Philippe" "R R R Roeloffs"

Slot "abstract":
[1] "In several families with non-specific X-linked mental retardation (XLN) linkage analyses have assigned the underlying gene defect to tsi

Slot "articleTitle":

Slot "journal":
[1] "Hum Mol Genet"

Slot "pubdate":

Slot "abstract":
[1] "No URL Provided"

> pm1.titles(absbs)
[[1]]
[1] "Cloning and mapping of members of the XYN family."

> pm.abstGrep("retardation", absbs[[1]])
[1] TRUE FALSE TRUE

R 1.5.1 - A Language and Environment
annotate: data rendering

• A simple interface, ll.htmlpage, can be used to generate an HTML report of your results.
• The page consists of a table with one row per gene, with links to LocusLink.
• Entries can include various gene identifiers and statistics.
### BioConductor Gene Listing

**Golub et al. data, genes with permutation \text{maxT} adjusted p-value < 0.01**

**Locus Link Genes**

<table>
<thead>
<tr>
<th>LocusID</th>
<th>Gene name</th>
<th>Chromosome</th>
<th>ALL mean</th>
<th>AML mean</th>
<th>t-statistic</th>
<th>raw p-value</th>
<th>adj p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7720</td>
<td>X5153.5</td>
<td>7</td>
<td>-0.295</td>
<td>1.59</td>
<td>-10.6</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1471</td>
<td>Y2789_1</td>
<td>20</td>
<td>-0.81</td>
<td>2.08</td>
<td>-9.78</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>9084</td>
<td>M351.50</td>
<td>15</td>
<td>0.498</td>
<td>1.24</td>
<td>-8.03</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4008</td>
<td>M30938</td>
<td>8</td>
<td>-0.392</td>
<td>1.1</td>
<td>-7.78</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>5324</td>
<td>L5209_5</td>
<td>11</td>
<td>-0.182</td>
<td>1.36</td>
<td>-7.57</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>6929</td>
<td>M31323</td>
<td>19</td>
<td>0.685</td>
<td>-0.391</td>
<td>7.52</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>5926</td>
<td>X74262</td>
<td>1</td>
<td>0.689</td>
<td>0.665</td>
<td>7.42</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>7155</td>
<td>Z51015</td>
<td>3</td>
<td>0.54</td>
<td>0.945</td>
<td>7.25</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>16099</td>
<td>L47735</td>
<td>5</td>
<td>0.794</td>
<td>0.779</td>
<td>7.21</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4502</td>
<td>U2276d_6</td>
<td>6</td>
<td>1.86</td>
<td>0.297</td>
<td>7.28</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>65107</td>
<td>H16112-H171612</td>
<td>1</td>
<td>1.91</td>
<td>0.889</td>
<td>7.11</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>90</td>
<td>H41432_1</td>
<td>1</td>
<td>0.43</td>
<td>-0.771</td>
<td>7.05</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>5923</td>
<td>L14187_1</td>
<td>13</td>
<td>-0.438</td>
<td>-3.3</td>
<td>7.02</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>596</td>
<td>U17936_5</td>
<td>14</td>
<td>-0.097</td>
<td>-1.07</td>
<td>7.07</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>7830</td>
<td>H31321</td>
<td>6</td>
<td>1.82</td>
<td>1.07</td>
<td>7.05</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4036</td>
<td>U10136_7</td>
<td>5</td>
<td>0.71</td>
<td>1.51</td>
<td>-6.97</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>7541</td>
<td>Y12070</td>
<td>1</td>
<td>-0.157</td>
<td>0.692</td>
<td>-6.96</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>7989</td>
<td>X74601</td>
<td>4</td>
<td>0.611</td>
<td>-0.183</td>
<td>6.85</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>6976</td>
<td>Y20789_5</td>
<td>4</td>
<td>-0.371</td>
<td>0.322</td>
<td>6.97</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>6709</td>
<td>P5243</td>
<td>8</td>
<td>0.413</td>
<td>0.982</td>
<td>6.66</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1723</td>
<td>U26265_6</td>
<td>10</td>
<td>-0.209</td>
<td>1.16</td>
<td>6.85</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>3292</td>
<td>H42759_9</td>
<td>7</td>
<td>0.64</td>
<td>0.504</td>
<td>6.82</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>915</td>
<td>M2397_1</td>
<td>19</td>
<td>-0.901</td>
<td>0.351</td>
<td>6.79</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1599</td>
<td>M2315_1</td>
<td>17</td>
<td>1.21</td>
<td>2.12</td>
<td>-6.77</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>6993</td>
<td>M12999_9</td>
<td>14</td>
<td>1.13</td>
<td>0.132</td>
<td>5.76</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>957</td>
<td>X62654_7</td>
<td>10</td>
<td>0.513</td>
<td>1.33</td>
<td>-6.76</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>514</td>
<td>X07745_5</td>
<td>2</td>
<td>-0.939</td>
<td>0.533</td>
<td>-6.74</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1424</td>
<td>M31217_7</td>
<td>12</td>
<td>0.106</td>
<td>-0.683</td>
<td>6.71</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>7226</td>
<td>U62136_5</td>
<td>12</td>
<td>-0.153</td>
<td>0.92</td>
<td>6.68</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>3650</td>
<td>X13949_5</td>
<td>4</td>
<td>-0.541</td>
<td>4.33</td>
<td>6.61</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>3265</td>
<td>X13949_4</td>
<td>14</td>
<td>-0.666</td>
<td>3.28</td>
<td>6.61</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
</tbody>
</table>
**annotate**: chromLoc class

Location information for **one** gene

- **chrom**: chromosome name.
- **position**: starting position of the gene in bp.
- **strand**: chromosome strand +/-.
**annotate:** chromLocation class

Location information for a set of genes

- **species:** species that the genes correspond to.
- **datSource:** source of the gene location data.
- **nChrom:** number of chromosomes for the species.
- **chromNames:** chromosome names.
- **chromLocs:** starting position of the genes in bp.
- **chromLengths:** length of each chromosome in bp.
- **geneToChrom:** hash table translating gene IDs to location.

Function buildChromClass
geneplotter: cPlot
geneplotter: alongChrom
geneplotter: alongChrom
Gene filtering

• A very common task in microarray data analysis is **gene-by-gene selection**.

• Filter genes based on
  – data quality criteria, e.g. absolute intensity or variance;
  – subject matter knowledge;
  – their ability to differentiate cases from controls;
  – their spatial or temporal expression pattern.

• Depending on the experimental design, some highly specialized filters may be required and applied sequentially.
Gene filtering

• **Clinical trial.** Filter genes based on association with survival, e.g. using a Cox model.

• **Factorial experiment.** Filter genes based on interaction between two treatments, e.g. using 2-way ANOVA.

• **Time-course experiment.** Filter genes based on periodicity of expression pattern, e.g. using Fourier transform.
The `genefilter` package provides tools to sequentially apply filters to the rows (genes) of a matrix.

There are two main functions, `filterfun` and `genefilter`, for assembling and applying the filters, respectively.

Any number of functions for specific filtering tasks can be defined and supplied to `filterfun`.

E.g. Cox model p-values, coefficient of variation.
**genefilter**: separation of tasks

1. Select/define functions for specific filtering tasks.
2. Assemble the filters using the `filterfun` function.
3. Apply the filters using the `genefilter` function → a logical vector, **TRUE** indicates genes that are retained.
4. Apply that vector to the `exprSet` to obtain a microarray object for the subset of interesting genes.
**genefilter**: supplied filters

Filters supplied in the package

- **kOverA** – select genes for which k samples have expression measures larger than A.
- **gapFilter** – select genes with a large IQR or gap (jump) in expression measures across samples.
- **ttest** – select genes according to t-test nominal p-values.
- **Anova** – select genes according to ANOVA nominal p-values.
- **coxfilter** – select genes according to Cox model nominal p-values.
genefilter: writing filters

• It is very simple to write your own filters.
• You can use the supplied filtering functions as templates.
• The basic idea is to rely on lexical scope to provide values (bindings) for the variables that are needed to do the filtering.
**genefilter**: How to?

1. First, build the filters
   
   ```r
   f1 <- anyNA
   f2 <- kOverA(5, 100)
   ```

2. Next, assemble them in a filtering function
   
   ```r
   ff <- filterfun(f1,f2)
   ```

3. Finally, apply the filter
   
   ```r
   wh <- genefilter(exprs(DATA), ff)
   ```

4. Use `wh` to obtain the relevant subset of the data
   
   ```r
   mySub <- DATA[wh,]
   ```
golubEsets

• now we will spend some time looking at filtering genes according to different criteria
golubEsets

• are there genes that are differentially expressed by Sex?
• if so on which chromosomes are they?
• are there any genes on the Y chromosome that are expressed in samples from female patients?
Differential gene expression

• Identify genes whose expression levels are associated with a response or covariate of interest
  – clinical outcome such as survival, response to treatment, tumor class;
  – covariate such as treatment, dose, time.

• Estimation: estimate effects of interest and variability of these estimates.
  E.g. slope, interaction, or difference in means in a linear model.

• Testing: assess the statistical significance of the observed associations.
Acknowledgements

- **Bioconductor core team**
  - Ben Bolstad, Biostatistics, UC Berkeley
  - Vincent Carey, Biostatistics, Harvard
  - Francois Collin, GeneLogic
  - Leslie Cope, JHU
  - Laurent Gautier, Technical University of Denmark, Denmark
  - Yongchao Ge, Statistics, UC Berkeley
  - Robert Gentleman, Biostatistics, Harvard
  - Jeff Gentry, Dana-Farber Cancer Institute
  - John Ngai Lab, MCB, UC Berkeley
  - Juliet Shaffer, Statistics, UC Berkeley
  - Terry Speed, Statistics, UC Berkeley
  - Yee Hwa (Jean) Yang, Biostatistics, UCSF
  - Jianhua (John) Zhang, Dana-Farber Cancer Institute
- Spike-in and dilution datasets:
  - Gene Brown’s group, Wyeth/Genetics Institute
  - Uwe Scherf’s group, Genomics Research & Development, GeneLogic.
- GeneLogic and Affymetrix for permission to use their data.