Part V. Analysis and presentation via web interfaces

genefilter, multtest, and annotate packages

Sandrine Dudoit and Robert Gentleman

© Copyright 2002, all rights reserved
Outline

- `genefilter` package
- `multtest` package
- `annotate` package
  - annotation data packages;
  - matching IDs using environments;
  - searching and processing queries from WWW databases
    - LocusLink,
    - GenBank,
    - PubMed;
  - HTML reports.
Combining data across arrays

Data on $G$ genes for $n$ arrays

$G \times n$ genes-by-arrays data matrix

$$M = \log_2(\text{Red intensity} / \text{Green intensity})$$
expression measure, e.g. RMA.
Combining data across arrays

... but the columns have structure, determined by the experimental design.
Combining data across arrays

- *cDNA array factorial experiment*. Each column corresponds to a pair of mRNA samples with different drug x dose x time combinations.
- *Clinical trial*. Each column corresponds to a patient, with associated clinical outcome, such as survival and response to treatment.
- *Linear models* and extensions thereof can be used to effectively combine data across arrays for complex experimental designs.
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
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.
1. First, build the filters
   \[
   \begin{align*}
   f1 & \leftarrow \text{anyNA} \\
   f2 & \leftarrow \text{kOverA}(5, 100)
   \end{align*}
   \]
2. Next, assemble them in a filtering function
   \[
   ff \leftarrow \text{filterfun}(f1, f2)
   \]
3. Finally, apply the filter
   \[
   wh \leftarrow \text{genefilter}(\text{exprs(DATA)}, ff)
   \]
4. Use \(wh\) to obtain the relevant subset of the data
   \[
   \text{mySub} \leftarrow \text{DATA}[wh,]
   \]
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.
Multiple hypothesis testing

- When testing for each gene the null hypothesis of no differential expression, e.g. using a t- or F-statistic, two types of errors can be committed.

- Type I error or false positive
  - say that a gene is differentially expressed when it is not,
  - reject a true null hypothesis.

- Type II error or false negative
  - fail to identify a truly differentially expressed gene,
  - fail to reject a false null hypothesis.
Multiple hypothesis testing

• Large multiplicity problem: thousands of hypotheses are tested simultaneously!
  – Increased chance of false positives.
  – E.g. chance of at least one p-value < \( \alpha \) for \( G \) independent tests is \( 1 - (1 - \alpha)^G \)
    and converges to one as \( G \) increases.
    For \( G=1,000 \) and \( \alpha = 0.01 \), this chance is 0.9999568!
  – Individual p-values of 0.01 no longer correspond to significant findings.

• Need to adjust for multiple testing when assessing the statistical significance of the observed associations.
Multiple hypothesis testing

• Define an appropriate Type I error or false positive rate.
• Develop multiple testing procedures that
  – provide strong control of this error rate,
  – are powerful (few false negatives),
  – take into account the joint distribution of the test statistics.
• Report adjusted p-values for each gene which reflect the overall Type I error rate for the experiment.
• Resampling methods are useful tools to deal with the unknown joint distribution of the test statistics.
multtest package

• Multiple testing procedures for controlling
  – Family-Wise Error Rate - FWER: Bonferroni, Holm (1979), Hochberg (1986), Westfall & Young (1993) maxT and minP;

• Tests based on t- or F-statistics for one- and two-factor designs.
• Permutation procedures for estimating adjusted p-values.
• Fast permutation algorithm for minP adjusted p-values.
• Documentation: tutorial on multiple testing.
multtest package

Sorted adjusted p-values for different multiple testing procedures
Golub et al. (1999) ALL AML data

- FWER control
  - solid lines
- FDR control
  - dashed lines
- PCER control
  - dotted lines
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` package 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!
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.
**annotate**: matching IDs

<table>
<thead>
<tr>
<th>Affymetrix identifier, HGU95A chips</th>
<th>“41046_s_at”</th>
</tr>
</thead>
<tbody>
<tr>
<td>LocusLink, LocusID</td>
<td>“9203”</td>
</tr>
<tr>
<td>GenBank accession #</td>
<td>“X95808”</td>
</tr>
<tr>
<td>Gene symbol</td>
<td>“ZNF261”</td>
</tr>
<tr>
<td>PubMed, PMID</td>
<td>“10486218”</td>
</tr>
<tr>
<td></td>
<td>“9205841”</td>
</tr>
<tr>
<td></td>
<td>“8817323”</td>
</tr>
<tr>
<td>Chromosomal location</td>
<td>“X”, “Xq13.1”</td>
</tr>
</tbody>
</table>
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.
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 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
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

> 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: 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**
**annotate**: WWW queries

- Functions for querying WWW databases from R rely on the `openBrowser` function

```r
openBrowser("www.r-roject.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


- **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.nih.gov/LocusLink/list.cgi?Q=zinc finger&ORG=Hs&V=0](http://www.ncbi.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

> pm: titles(absts)
[[1]]
[1] "Cloning and mapping of members of the XTM family."
[2] "Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which conf"

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

>
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 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>7291</td>
<td>X5738, et</td>
<td>7</td>
<td>0.295</td>
<td>1.59</td>
<td>-0.6</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1471</td>
<td>172981</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>5138</td>
<td>MG5100</td>
<td>15</td>
<td>0.496</td>
<td>1.24</td>
<td>-8.03</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4919</td>
<td>ML6030, et</td>
<td>8</td>
<td>0.284</td>
<td>1.1</td>
<td>-7.99</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>5024</td>
<td>LS2297, et</td>
<td>11</td>
<td>0.152</td>
<td>1.36</td>
<td>-7.97</td>
<td>2e-05</td>
<td>2e-04</td>
</tr>
<tr>
<td>6929</td>
<td>MG1038, et</td>
<td>19</td>
<td>0.655</td>
<td>0.391</td>
<td>7.95</td>
<td>2e-05</td>
<td>2e-04</td>
</tr>
<tr>
<td>5026</td>
<td>X74262, et</td>
<td>1</td>
<td>0.669</td>
<td>0.565</td>
<td>7.42</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>7165</td>
<td>ZI3911, et</td>
<td>3</td>
<td>1.54</td>
<td>3.945</td>
<td>7.28</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>25399</td>
<td>L47791, et</td>
<td>15</td>
<td>0.798</td>
<td>0.779</td>
<td>7.31</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4542</td>
<td>U2292</td>
<td>6</td>
<td>2.36</td>
<td>0.294</td>
<td>7.28</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>65310</td>
<td>HG1612</td>
<td>6</td>
<td>0.891</td>
<td>0.888</td>
<td>0.11</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>248233</td>
<td>MG1832, et</td>
<td>19</td>
<td>1.43</td>
<td>0.771</td>
<td>7.08</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>24598</td>
<td>MG1471, et</td>
<td>15</td>
<td>0.632</td>
<td>0.33</td>
<td>7.08</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>61754</td>
<td>Y13070, et</td>
<td>15</td>
<td>0.157</td>
<td>0.892</td>
<td>-6.96</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>7203</td>
<td>X74601, et</td>
<td>1</td>
<td>0.611</td>
<td>0.183</td>
<td>5.55</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>5026</td>
<td>T00799, et</td>
<td>4</td>
<td>0.371</td>
<td>0.32</td>
<td>6.67</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>67272</td>
<td>MG243, et</td>
<td>9</td>
<td>0.413</td>
<td>0.982</td>
<td>5.66</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>1783</td>
<td>U26261, et</td>
<td>19</td>
<td>0.229</td>
<td>1.16</td>
<td>8.85</td>
<td>1e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4902</td>
<td>U02799, et</td>
<td>7</td>
<td>0.64</td>
<td>0.504</td>
<td>6.92</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4933</td>
<td>MG2897, et</td>
<td>19</td>
<td>0.69</td>
<td>0.354</td>
<td>6.79</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>6199</td>
<td>MG3198, et</td>
<td>11</td>
<td>1.21</td>
<td>2.12</td>
<td>6.77</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>6933</td>
<td>MG1039, et</td>
<td>14</td>
<td>4.13</td>
<td>1.32</td>
<td>5.76</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>9357</td>
<td>MG1039, et</td>
<td>12</td>
<td>0.513</td>
<td>1.33</td>
<td>-6.76</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>4014</td>
<td>X07745, et</td>
<td>2</td>
<td>-0.939</td>
<td>-0.553</td>
<td>-6.74</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>12345</td>
<td>MG1231, et</td>
<td>12</td>
<td>0.103</td>
<td>-0.853</td>
<td>5.71</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>5026</td>
<td>MG1036, et</td>
<td>8</td>
<td>0.153</td>
<td>-0.92</td>
<td>6.68</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>3530</td>
<td>X18426, et</td>
<td>1</td>
<td>-0.541</td>
<td>-1.33</td>
<td>6.61</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
<tr>
<td>8435</td>
<td>D8762, et</td>
<td>3</td>
<td>0.086</td>
<td>0.065</td>
<td>6.61</td>
<td>2e-05</td>
<td>2e-05</td>
</tr>
</tbody>
</table>