

R/BioC Exercises: meta-data and pathways

Perry Moerland

April 23, 2010

☞ Information on how to log on to a PC in the exercise room and the UNIX server can be found here: <http://bioinformaticslaboratory.nl/twiki/bin/view/BioLab/EducationBioinformaticsII>. Don't forget to enable X11 forwarding when starting up PuTTY.

☞ These exercises depend on Chapters 7 and 8 (Gentleman et al. 2005; Carey et al. 2005). References to equations are with respect to these chapters. Chapters 19-22 offer interesting reading on working with graphs but are optional (Gentleman et al. 2005; Huber et al. 2005; Carey et al. 2005; Gentleman et al. 2005).

Introduction

Hexachlorobenzene (HCB) is a persistent environmental pollutant with toxic effects in man and rat. We'll study gene expression in rats that were fed a diet supplemented with 0, 150, or 450 mg HCB/kg for 4 weeks (6 rats for each group). Liver gene expression was studied using the Affymetrix rat RG-U34A GeneChip (Ezendam et al. 2004). We use this dataset to illustrate various aspects of working with meta-data and pathways in Bioconductor.

The data set was normalized with RMA using the following commands (to save computing time there is no need to try this yourself):

```
> library(affy)
> fns <- list.celfiles(path = "Data", full.names = TRUE)
> data <- ReadAffy(filenamees = fns)
> eset <- rma(data)
> save(eset, file = "ezendam.Rdata")
```

Exercise 1: Copy the experimental design `ezendam.txt` and the normalized data `ezendam.Rdata` from `/data/home/stud00/MetaPathways` to your home directory. Note that one of the rats in the 450 mg group died before the end of the experiment and should be excluded from the experiment. This is indicated with “not done” in `ezendam.txt`. The experiment can be analyzed using linear models as you have learnt last week. Adapt the analysis of section 23.8 (Gentleman et al. 2005) for the purpose of making all pair-wise comparisons between the 0 (control), 150 (low), and 450 (high) mg groups.

As in section 23.8 it is assumed that you stored the result of the call to `eBayes` in variable `fit2`. This variable is a list in which the component `fit2$F.p.value` combines the three

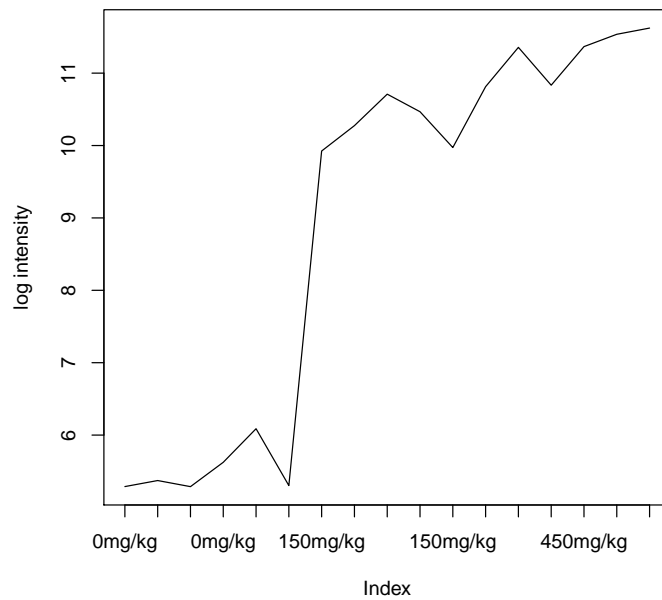
pair-wise comparisons (“low vs control”, “high vs control”, and “high vs low”) into the p -value of a F -test. This tests whether a gene is differentially expressed in at least one of the three comparisons.

Remember that you are testing many different hypotheses, one for each probeset on the array, and that you have to correct for multiple testing. We do this with the Benjamini-Hochberg correction:

```
> F.adj.p.value <- p.adjust(fit2$F.p.value, method = "BH")
```

Let us have a closer look at the most differentially expressed gene, that is, the one with the smallest p -value.

```
> plot(exprs(eset)[order(F.adj.p.value)[1], ], type = "l", ylab = "log intensity",  
+       xaxt = "n")  
> axis(1, at = 1:ncol(exprs(eset)), labels = targets[, "ExperimentalFactor[Dose]"])
```



This plot clearly shows that this gene has low expression in the control condition and high expression in the other two conditions. You can check other genes to see other expression patterns. When looking at `fit2$genes` you might understand that interpretation of these results is quite difficult. In the rest of the exercises, you will see how you can add detailed annotation to the results.

Meta-data

For many array platforms, Bioconductor provides annotation files that map chip specific probe labels to different targets such as gene ID or chromosomal location. Also for our Affymetrix rat RG-U34A there is an annotation package available that you can load in the standard way:

```

> library(rgu34a.db)
> ls("package:rgu34a.db")

 [1] "rgu34a"                "rgu34a_dbconn"        "rgu34a_dbfile"
 [4] "rgu34a_dbInfo"        "rgu34a_dbschema"     "rgu34aACCNUM"
 [7] "rgu34aALIAS2PROBE"    "rgu34aCHR"           "rgu34aCHRENGTHS"
[10] "rgu34aCHRLOC"        "rgu34aCHRLOCEND"     "rgu34aENSEMBL"
[13] "rgu34aENSEMBL2PROBE" "rgu34aENTREZID"      "rgu34aENZYME"
[16] "rgu34aENZYME2PROBE"  "rgu34aGENENAME"      "rgu34aGO"
[19] "rgu34aGO2ALLPROBES"  "rgu34aGO2PROBE"      "rgu34aMAP"
[22] "rgu34aMAPCOUNTS"    "rgu34aORGANISM"      "rgu34aORGPKG"
[25] "rgu34aPATH"          "rgu34aPATH2PROBE"    "rgu34aPFAM"
[28] "rgu34aPMID"          "rgu34aPMID2PROBE"    "rgu34aPROSITE"
[31] "rgu34aREFSEQ"        "rgu34aSYMBOL"        "rgu34aUNIGENE"
[34] "rgu34aUNIPROT"

```

i From BioC release 2.1 onwards the old style annotation packages have been replaced by packages based on SQLite. These are easily recognized by the extension `.db`. The old style packages will be deprecated in future Bioconductor releases. In general, operations on the new annotation packages are slower unless one uses directed SQL queries. The new annotation packages are more memory efficient and more powerful. See <http://www.bioconductor.org/packages/bioc/html/AnnotationDbi.html> for more information. In order to not diverge from the book too much, we will stick to using the new annotation packages in the old way, but you are welcome to explore the rich features of the database style annotation.

Exercise 2: Most of the mappings are from probeset to target. As an example, look at the probeset `E00778cds_s_at` in some more detail and find its name, symbol, chromosomal location, and the KEGG pathways in which it is implicated (remember *help*). Also try to find similar annotation in Ensembl using functions from the `biomaRt` package. What is the advantage of the latter approach?

Some meta-data also has reverse mappings (also see *revmap* in new style annotation) which is actually the preferred format when constructing gene sets: we just want to know which genes are part of a set. Examples are `rgu34aPATH2PROBE` and `rgu34aGO2PROBE`

Exercise 3: Map one of the pathways found in the previous exercise to its probeset members. Is probeset `E00778cds_s_at` indeed part of this pathway?

The previous exercise only gives the KEGG ID of a pathway. More pathway-specific information can be obtained with the `KEGG.db` package.

Exercise 4: Extract the names for all pathways found in Exercise 2 (hint: *mget*). For one of these pathways go to <http://www.genome.jp/kegg/pathway/map/mapXXXXX.html> where you have to replace `XXXXX` with the KEGG ID of the pathway you are interested in.¹ Can you find the enzyme code corresponding to `E00778cds_s_at` in this map?

¹http://www.genome.jp/dbget-bin/get_pathway?org_name=rno&mapno=XXXXX gives the detailed pathway for rat instead of the non-species specific pathway.

Exercise 5: Often several probesets map to the same gene, for example, to interrogate splice variants. Determine if there are any other probesets that map to the same Entrez Gene ID as E00778cds_s_at.²

Exercise 6: Investigate the gene expression pattern (from `ezendam.Rdata`) for the probesets found in the previous question graphically (by now you might have guessed that the answer to previous question should have been “yes”). What is the correlation between the gene expression patterns? What does this mean?

Exercise 7: The `rgu34probe` package provides probe sequence information for our rat chip. What is the 25-mer of the first probe of probeset E00778cds_s_at ((Carey et al. 2005), section 8.4)? Retrieve the annotation for this probeset from Ensembl http://www.ensembl.org/Rattus_norvegicus/Location/Genome?ftype=ProbeFeature;fdb=funcgen;ptype=pset;id=E00778cds_s_at;. Can you explain the correspondence between the probe interrogation position (from `rgu34probe`) and the genomic location at Ensembl?

Exercise 8: In Exercise 4 you found the enzyme code corresponding to E00778cds_s_at. Are there other probesets mapped to this particular enzyme code? Are these all included in the KEGG pathway you selected? Why or why not? Can you explain why these probesets are mapped to the same enzyme code? What can you say about the rank of these probesets in the ordered list of p -values of the F -test?

Gene set enrichment

We now go back to the outcome of the linear model analysis. Instead of analysing this list gene by gene, we take the *gene set* perspective explained during the pathway lecture. In the remainder of these exercises, you will learn how to construct gene sets for rat and how to analyze these gene sets in the context of the hexachlorobenzene experiment. Three sets of gene sets will be constructed:

- Chromosomes and cytogenetic bands (`g1`)
- KEGG pathways (`g2`)
- GO categories (`g3`)

Exercise 9: Make a set of gene sets that lists for each of the rat chromosomes and cytogenetic bands which genes are part of it. The suitable environments for this are `rgu34aCHR` and `rgu34aMAP`, respectively. Hint: you can use `split` (`help(split,package=base)`) to transform the resulting mapping from probeset to target (=chromosome/cytoband) in a mapping from target to probesets and compare it with the outcome of using database style `revmap`.

Exercise 10: Make also the sets of gene sets for KEGG and GO. Suitable environments here are `rgu34aPATH2PROBE` and `rgu34aGO2ALLPROBES`. What is the difference between the latter and `rgu34aGO2PROBE`?

²You might use http://www.affymetrix.com/support/help/IVT_glossary/index.affx and http://en.wikipedia.org/wiki/Five_prime_untranslated_region to make sense of the nomenclature of Affymetrix probesets.

As you will have noticed the sets of gene sets are lists with each component tagged by an ID. For GO, the elements of a component have names that correspond to GO evidence codes ((Gentleman et al. 2005), Table 7.2).

In the previous section, you have studied the KEGG pathways in quite some detail. Let us now have a closer look at the GO categories.

Exercise 11: First map probeset E00778cds_s_at to its GO identifiers using `rgu34aGO`. Investigate the first GO ID in more detail using the `GOTERM` environment from the `GO.db` package. Then extract all GO terms in `rgu34aGO` that E00778cds_s_at belongs to. Hint: *sapply*.

Exercise 12: The number of GO gene sets constructed in Exercise 10 is large. A first way of pruning is to eliminate all probe sets with evidence code `IEA`. These are uncurated electronically inferred and, therefore, less trustworthy. Next, some probesets occur in the same GO category with different evidence codes. Remove these duplicates. Finally, remove all GO categories with less than five elements left. How many GO gene sets are left?

The `limma` package offers an elegant way of testing whether a set of genes is enriched for differential expression. It is based on the location tests explained during the pathway lecture. Read `help(geneSetTest)` carefully before you start with the next exercise.

Exercise 13: Write a function that tests which of the KEGG pathways is enriched for differential expression in the F -test between the three different groups (`fit2$F`). Since you are testing for many different pathways, don't forget to adjust for multiple testing. What are the IDs and names of the ten most enriched pathways? Among the top pathways are some old acquaintances, can you explain this result given your answer to Exercise 8? You might want to read the original article (Ezendam et al. 2004) to see if they found similar pathways to be involved.

Exercise 14: In Exercise 8 you found all probesets with the same enzyme code as E00778cds_s_at. Investigate the expression profiles of these probesets in some more detail using hierarchical clustering, for example.

Exercise 15: Write a function that tests which of the GO categories is enriched for differential expression in the F -test between the three different groups (`fit2$F`). Since you are testing for many different pathways, don't forget to adjust for multiple testing. What are the IDs and names of the ten most enriched categories?

References

- Carey, V., R. Gentleman, W. Huber, and J. Gentry (2005). Chapter 21: Bioconductor software for graphs. See Gentleman, Carey, Huber, Irizarry, and Dudoit (2005).
- Carey, V., D. T. Lang, J. Gentry, J. Zhang, and R. Gentleman (2005). Chapter 8: Querying on-line resources. See Gentleman, Carey, Huber, Irizarry, and Dudoit (2005).

Ezendam, J., F. Staedtler, J. Pennings, R. J. Vandebriel, R. Pieters, J. H. Harleman, and J. G. Vos (2004). Toxicogenomics of subchronic hexachlorobenzene exposure in brown Norway rats. *Environ Health Perspect* 112, 782–791. Raw data for other organs is also available in ArrayExpress under accession number E-TOXM-15.

Gentleman, R., V. Carey, W. Huber, R. Irizarry, and S. Dudoit (Eds.) (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer.

Gentleman, R., V. Carey, and J. Zhang (2005). Chapter 7: Meta-data resources and tools in bioconductor. See Gentleman, Carey, Huber, Irizarry, and Dudoit (2005).

Gentleman, R., W. Huber, and V. Carey (2005). Chapter 19: Introduction and motivating examples. See Gentleman, Carey, Huber, Irizarry, and Dudoit (2005).

Gentleman, R., D. Scholtens, B. Ding, V. Carey, and W. Huber (2005). Chapter 22: Case studies using graphs on biological data. See Gentleman, Carey, Huber, Irizarry, and Dudoit (2005).

Huber, W., R. Gentleman, and V. Carey (2005). Chapter 20: Graphs. See Gentleman, Carey, Huber, Irizarry, and Dudoit (2005).

This document was generated with

```
> sessionInfo()
```

```
R version 2.10.1 (2009-12-14)  
i386-pc-mingw32
```

```
locale:
```

```
[1] LC_COLLATE=English_United Kingdom.1252  
[2] LC_CTYPE=English_United Kingdom.1252  
[3] LC_MONETARY=English_United Kingdom.1252  
[4] LC_NUMERIC=C  
[5] LC_TIME=English_United Kingdom.1252
```

```
attached base packages:
```

```
[1] stats      graphics  grDevices  utils      datasets  methods    base
```

```
other attached packages:
```

```
[1] bioDist_1.18.0      KernSmooth_2.23-3  annotate_1.24.1  
[4] GO.db_2.3.5         rgu34aprobe_2.5.0  KEGG.db_2.3.5  
[7] biomaRt_2.2.0       rgu34a.db_2.3.5    org.Rn.eg.db_2.3.5  
[10] RSQLite_0.8-4       DBI_0.2-5          AnnotationDbi_1.8.2  
[13] Biobase_2.6.1       limma_3.2.3
```

```
loaded via a namespace (and not attached):
```

```
[1] RCurl_1.3-1  tools_2.10.1 XML_2.8-1    xtable_1.5-6
```