

Seminar III: R/Bioconductor

Leonardo Collado Torres

`lcollado@lcg.unam.mx`

Bachelor in Genomic Sciences

`www.lcg.unam.mx/~lcollado/`

August - December, 2009

Public Data

Intro

RMySQL

AnnotationDbi

Install

► Packages:

```
> install.packages("RMySQL")  
> source("http://bioconductor.org/biocLite.R")  
> biocLite("AnnotationDBI")
```

RMySQL

- ▶ Installing this package is not that simple.
- ▶ In Windows, you will need to download a libmysql.dll file and copy it to the RMySQL folder.
- ▶ On Linux, you need to install the mysql libraries¹.
- ▶ For now use Montevalban :)

¹Probably the dev ones too for compiling

Explore the pkg

- ▶ You can explore quite a bit of the package using
> `help(package = "RMySQL")`
- ▶ Its **important** to close every connection you open.

Sample session

- ▶ A sample session would look be similar to this. First we connect to the database.

```
> con <- dbConnect(MySQL(), user = "lcollado",  
+   password = "LOL", dbname = "BPdB",  
+   host = "kabah.lcg.unam.mx")
```

- ▶ Then we do some queries, list the tables, download some tables into data frames, etc.

```
> dbListTables(con)  
> df <- dbReadTable(con, "tablename")
```

- ▶ And we end by closing the connection:

```
> dbDisconnect(con)
```

Queries vs downloading tables

- ▶ SQL is faster for doing joins between tables, so you might want to use queries then.
- ▶ If all the info is on one table, and you are in local network with the database server, you might prefer to download the table. Then use a `tapply` or other functions if you have a grouping variable.

Exercise I

- ▶ Use **RMySQL** to access your own database from the Bioinformatics and Statistics I course.
- ▶ Use a query to retrieve data and make a plot :)
- ▶ If you don't have your own database, let me know.

SQLite

- ▶ An alternative for MySQL that seems to be faster is SQLite.
- ▶ If you are using that SQL language, you might want to install the RSQLite R package :)

Intro

- ▶ Its an alive package :) Meaning that its under intensive development.²
- ▶ This package is the key interface to access the gamma of annotation packages.
- ▶ It allows you to retrieve data from them using some R functions, or by directly using SQL queries.

²Regretably, the latest version has an installation bug.

A lab

<http://bioconductor.org/workshops/2009/BioC2009/labs/annotations/AnnotationDbi.pdf>

- ▶ Check the above lab :) Its a very complete description of the AnnotationDbi package.
- ▶ Read the document and pay attention to exercises 1, 2, 3, and 5. If you have a working installation, then try to do them :)

SessionInfo

```
> sessionInfo()
```

```
R version 2.10.0 Under development (unstable) (2009-07-25 r48998)  
i686-pc-linux-gnu
```

```
locale:
```

```
[1] LC_CTYPE=en_US.UTF-8  
[2] LC_NUMERIC=C  
[3] LC_TIME=en_US.UTF-8  
[4] LC_COLLATE=en_US.UTF-8  
[5] LC_MONETARY=C  
[6] LC_MESSAGES=en_US.UTF-8  
[7] LC_PAPER=en_US.UTF-8  
[8] LC_NAME=C  
[9] LC_ADDRESS=C  
[10] LC_TELEPHONE=C  
[11] LC_MEASUREMENT=en_US.UTF-8  
[12] LC_IDENTIFICATION=C
```

```
attached base packages:
```

SessionInfo

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