Seminar III: R/Bioconductor
August-December 2009

Leonardo Collado Torres

Graduated from the Undergraute Program on Genomic Sciences (LCG),
UNAM, Cuernavaca, Mexico

lcollado@lcg.unam.mx

http://www.lcg.unam.mx/~lcollado/index_en.php

August 12, 2009

Assistants: Alejandro Reyes areyes@lcg.unam.mx, José Reyes jreyes@lcg.unam.mx and Víctor Moreno jmoreno@lcg.unam.mx

Abstract

The course Seminar III: R/Bioconductor will be taught at the *Licenciatura de Ciencias Genómicas* (LCG) at UNAM on Fridays from 12:00 to 14:00. This course will give an overview of Bioconductor which is a set of public tools, built on top of R and developed for the study of genomic sciences. To take this course it's a prerequisite to have a basic understanding of Statistics and R. A basic R introduction course was taught to the sixth LCG generation which is public at http://www.lcg.unam.mx/~lcollado/E/ (Spanish only). The order in which the material will be covered and the project associated to this course will be integrated and directly related to the course *Bioinformatics and Statistics I* from the LCG.

The oficial page of the course is http://www.lcg.unam.mx/~lcollado/B/. There you can find the presentations, code associated to the presentations, exercises, expected answers, supporting material, and the data sets that we'll use.

For any doubt, question, suggestion or to ask for an advice session, please do so through the forum of the National Node of Bioinformatics (NNB) located here.

Contents

1 Objectives

- Introduce the students to the world of Bioconductor so that they'll use the most updated set of tools for genomics built on `R`.
- Expand their knowledge and skills to make plots with three variables and of genomic order.
- Learn how to import public data sets into `R` using Bioconductor.
- Learn the basics to analyze microarrays using Bioconductor.
- Manage sequences and the software developed for the analysis of data derived from high throughput sequencing methods using Bioconductor.
- Build the bases for reproducible research and practice sharing data through a Bioconductor package.
- Train and prepare the future heirs of `R` and Bioconductor at LCG. In a year the current assistants will be the professors and some of the students will become assistants.

2 Project

During the current semester, the students will develop in teams a project involving `Perl`, `MySQL` and `PHP` for the course of *Bioinformatics and Statistics I*. The project of Seminar III: R/Bioconductor will consist doing an statistical analysis of the data from the other project and on building an experimental data package for Bioconductor. Said package will have to comply with all the prerequisites that Bioconductor outlines. These include a documentation of the *vignette* type built with `Sweave` and LaTeX. This file, written in English[1], has to explain how the idea/question was conceived that motivated the project, how to get the data, the analysis done including `R` code, graphs and conclusions. The data can be imported from the `MySQL` data base using the `RMySQL` package. The question(s) that motivated the project can be simple but the data mining process and/or the statistical analysis should not be trivial. You will have to present this project on the final website of the other project either on the *vignette* format or in a format of your choice.

In other words, the project will be a real exercise that you will contribute to the international community through Bioconductor.

[1] Everything has to be in English including the names of the variables

3 A sample class

A *normal* class will follow this dynamic. On the first minutes, the assistants or the professor will ask one or several students what subject(s) they found interesting that were discussed that week on the Bioconductor mailing list or from a Bioconductor related paper. Meanwhile, one of two students will prepare themselves[2] and then will briefly talk about a Bioconductor package[3]:

- describing for what it is used
- showing some images that can be derived from its use
- what they found appealing from the package
- for what type of analysis workflow it is used
- which other Bioconductor packages complement it or may partially overlap its functionality

The student(s) that gave the brief talk will hand in a *vignette* file in English with the previous information which we'll share through the offical course page. Then we'll move unto the class subject which in general will include a description of a package, examples and a practice lab. Finally, the students will have to do a more complete/advanced practice which they will most likely finish as homework.

All the classes will be video recorded and the official language inside the classes will be English.[4]

4 Evaluation

Although it might seem very strict, we prefer to leave it as clear as possible at the beginning. As long as you hand in a minimum of 10 homeworks and your project, your grade will depend on four factors:

Participation 20 %

Your class participations, reading the Bioconductor mailing list and/or finding papers of interest, asking questions inside our forum.

Homeworks 30 %

Every homework shall have a 9 am deadline[5] on the next Friday. You have

[2] Turn on, connect their lap and set up their presentation

[3] Without repeating them

[4] If you need some English lession ask Iliana

[5] Server time!!

to do them individually unless specified and late homeworks will only be accepted up to one week after the due date. They shall be portable, meaning that the code doesn't depend on your folder structure and the data is available online[6]. For each homework you'll hand in two files: the `pdf` file generated using `Sweave` and LaTeX; the .R script created with `Stangle`. These files have to be named like this: *username_XX_descrip* where *XX* is the homework number and *descrip* is whatever you want. For example: `lcollado_01_review.pdf` and `lcollado_01_review.R`.

Presenting a Bioconductor package 10 %
This presentation has to meet the requirements mentioned at *A sample class.* It consists of a brief 5 min talk and the *vignette* file with the information (in English).

The project 40 %
Create an experimental data package for Bioconductor and use the data for a statistical analysis. It has to meet their requirements[7] and the *vignette* file that in reality is a report. Meaning that it has:

- An abstract-
- An introduction explaining where you got the idea/question and what it is-
- Describe how you did the data mining process. How you got the data and why you chose x, y, z source/variable.
- The analysis of your data inclusing the `R` code, graphs and results[8].
- Conclusions-

5 *Tentative* class calendar

14 Aug Class I

Review Initiating the course, `R` review and the apply function family

1. Course description including the project and evaluation.
2. How to look for help in `R`
3. Exercise using `for` and a couple of graphs.

[6] Either on a website, through a package like `biomaRt` or simply on your `public_html` folder at the LCG server

[7] Everything (variables, functions, text, etc) in English

[8] Don't forget your interpretation!!

4. Apply function family.

21 Aug Class II

Bioconductor and documentation Promoting reproducible research

1. Intro to Bioconductor
2. Help inside Bioconductor: *mailing lists*
3. Basic package installation
4. Basis for doing reproducible research and examples of the *vignette* style
5. `Sweave` as an interface of `R` with LaTeX
6. Short introduction to LaTeX and Beamer

28 Aug Class III

Plots Advanced plotting

1. Overview of the graphs you can make using `lattice`
2. Some graph examples using `Plotrix`

4 Sept Class IV

Public Data biomart, GEO and ArrayExpress

1. Exploring `biomaRt`
2. Basic `GEOquery` usage
3. Usage of `ArrayExpress`
4. A series of examples

11 Sept Class V

Interacting with `MySQL` `RMySQL` usage and overview of `annotationdbi`

1. Installing `RMySQL`
2. Connecting to a data base using `RMySQL`
3. Using `R` to construct `MySQL` queries
4. Quick overview of the `RSQLite` package
5. Description of the `annotationdbi` package

18 Sept Class VI

***Genomic* graphs** Visualizing loads of data at once

1. Overview of the `GenomeGraphs` package

2. Linking `GenomeGraphs` to `biomaRt`
3. Interacting with *Genome Browsers* like the one from `UCSC` through `rtracklayer`

25 Sept Class VII

Microarrays The first Bioconductor stronghold

1. Basis of linear regressions
2. Basic correlations
3. Using `limma` to find diferentially expressed genes
4. `affy` package

2 Oct Class VIII

Sequence analysis The basic tools

1. Using `IRanges`
2. Generating *views*
3. Manipulating sequences using `Biostrings`
4. Alingning and mapping using `Biostrings`

9 Oct Class IX

`R` **and *HTS***[9] Quality control and some analysis

1. Quality control of `Solexa` data using `ShortRead`
2. An HTS case: `chipseq` workflow

16 Oct Class X

Bioc packages Building your `Bioc` package

1. Basic structure of an `R` package
2. Requirements for an experimental data package in Bioconductor
3. `Qt` plotting demo

23 Oct Class XI

GOs GO analysis

1. Using `BLAST`
2. Several GO analysis using `R`

30 Oct Class XII

Statistics misc

1. Lowess and loess
2. Multiple testing corrections

6 Nov Class XIII

Microarrays II A more detailed session

1. `multtest` package
2. Getting into the detail

13 Nov Class XIV

HTS: a case *E. coli* transcriptome

1. Detailing an analysis workflow

20 Nov Class XV

Undefined class Guest?

1. Hoping to get a guest from abroad.

27 Nov Class XVI

Undefined class Open to suggestions

1. to be decided

30 Nov - 4 Dec First Exam Week

Advice Wrapping up your analysis

1. Advice sessions to detail the statistical analysis of your project
2. Advice to build your Bioconductor data package

7-11 Dec Second Exam Week

Projects Hand in and evaluation

1. Hand in your project[10]
2. Project evaluation
3. Checking and correcting the project
4. Send to Bioconductor[11]

[10] Most likely on Monday

[11] Most likely on Friday