

Course Title Methods in Bioinformatics

IBM1081

2nd Semester 2015

Thursday 14:00 - 18:00 (14:00-15:00 T/ 15:00 - 18:00 P)

Location: LED Sala 3 (FMRP)

Instructor: Professor Houtan Noushmehr

Course Description

This course will focus on the analysis and visualization of microarray data. The general aim is to introduce students to the various techniques and issues involved with analyzing high dimensional multivariate expression data. Examples will be taken from the literature. Additionally, students will visualize the results in a modern statistical scripting software, R. Topics include:

- * Utilizing packages and libraries to run functions and plot data in R.
- * Becoming familiar with statistical inference issues and modern statistical methods for analyzing biological data.
- * Detecting and attributing sources of data variability (separating signal from noise) and normalizing these differences with an introduction to some commonly employed methods.
- * Determining differentially expressed genes with relevant statistical tests and controlling for false positive discovery.
- * An introduction to linear and non-linear dimension reduction methods.
- * An introduction to common pattern recognition (clustering) and classification techniques.

Lecture schedule

<u>Lecture</u>	<u>Topic</u>
Lecture 1 (06/08)	Introduction to gene expression technology/Introduction to R
Lecture 2 (13/08)	Introduction to R (cont.)
Lecture 3 (20/08)	Data visualizations, outliers, and missing data
Lecture 4 (27/08)	Replicate analysis and power
Lecture 5 (17/09)	Normalization and Bioconductor
Lecture 6 (24/09)	Detecting differentially expressed genes
Lecture 7 (08/10)	Paper presentations
Lecture 8 (15/10)	Paper presentations
Lecture 9 (22/10)	Multiple testing adjustments
Lecture 10 (29/10)	Dimensionality reduction
Lecture 11 (05/11)	Cluster Analysis
Lecture 12 (12/11)	Classification
Lecture 13	Preparation for final project

(19/11)	
Lecture 14 (26/11)	Project presentations

**Point
Distribution**

<u>Assignments</u>	<u>Points</u>
Lab assignments	40 (5/lab)
Homework assignments	105 (35/assignment)
Paper presentation	45
Final project	65
Total	255

**Course
Policies and
Procedures**

The first part of class will be dedicated to lecture topics. Many of the topics will come directly from the literature. The second part of class will require the completion of a lab assignment. These are relatively short assignments designed to help students get familiar with the R language and implement basic microarray analyses. Many of the functions will be given in the lecture notes, so the lab will focus on utilizing this code to address various problems. Each student will be responsible for presenting a paper and completion of a final project. A sign-up format will be implemented so that each lecture will include a few presentations from students. All presentations should include 15-20 Power Point slides (depending on the work).

Textbook(s)

There is no required text for this course, however, some good references for specific lecture topics are listed below.

The Analysis of Gene Expression Data, Giovanni Parmigiani (Editor), Elizabeth S. Garrett (Editor), Rafael A. Irizarry (Editor), Scott L. Zeger

Statistical Analysis of Gene Expression Microarray Data, T. P. Speed (Editor)

A Biologist's Guide to Analysis of DNA Microarray Data, Steen Knudsen (Author)

Introductory Statistics with R, Peter Dalgaard

**Obtaining
data sets
(alternative to
getting from
course
website)**

Download the text files from the following sites

Golub et al. AML/ALL data (delete A/P calls) - Affymetrix data

http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43

links: Train dataset (text); Test dataset (text)

Eisen et al. Large B-cell Lymphoma data (NA values) - cDNA data

<http://rana.lbl.gov/EisenData.htm>

links: Gene Expression Data

Alon et al. Colon Cancer data - Affymetrix data

<http://www.sph.uth.tmc.edu/hgc/Downloads.asp>

links: GEColon.DAT

Spellman et al. yeast cell cycle data (missing values) - cDNA data

<http://genome-www.stanford.edu/cellcycle/data/rawdata/>

links: Tab delimited data

or

Download zip files from Bioconductor and install packages

Bioconductor (<http://www.bioconductor.org/>)

Under <Software>

<Experimental Data>

yeastCC: Spellman et al. yeast cell cycle data - cDNA data

golubEsets: Golub et al. AML/ALL data - Affymetrix data

colonCA: Alon et al. Colon Cancer data - Affymetrix data

in R, access data objects as follows:

load package from local drive (*all commands below are given in R*)

```
> library(yeastCC)
```

```
> library(Biobase)
```

```
> library(annotate)
```

```
> data(yeastCC)
```

```
> dat <- exprs(yeastCC)
```