Course guide
200631 - ADO - Omics Data Analysis

Last modified: 19/04/2022

Unit in charge: School of Mathematics and Statistics
Teaching unit: 1004 - UB - (ENG)Universitat de Barcelona.

Degree: MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Optional subject).

Academic year: 2022 **ECTS Credits:** 5.0 **Languages:** English

LECTURER

Coordinating lecturer: DIEGO GARRIDO MARTÍN

Others: Segon quadrimestre:
DIEGO GARRIDO MARTÍN - A
SANTIAGO RIOS AZUARA - A

PRIOR SKILLS

The course assumes no prior knowledge more than the usual of a student in a Master's Degree of Statistics.
However a good attitude toward biology (specifically Molecular biology) and a good knowledge of the R programming language can help to get the most out of the course.

Ideally this course would be taken after an introduction to bioinformatics as part of a bioinformatics oriented curriculum. However, given that currently there is no guarantee that ideally the two subjects are relatively independent so that, although it is interesting to have completed "Fundamentals of Bioinformatics" to have some familiarity with the problems that can be solved using the techniques developed here, is not considered essential.

REQUIREMENTS

The course assumes basic levels of statistics similar to those that can be achieved in the first semester of the Master. Students should be familiar with the concepts of hypothesis testing and statistical significance, analysis of variance and basic techniques of multivariate statistics such as principal component and cluster analysis.
Concepts necessary to follow the course can be found for example in the text "Applied Statistics for Bioinformatics using R" available on the R website (cran.r-project.org/doc/contrib/Krijnen-IntroBioInfStatistics.pdf) or Data Analysis for the Life Sciences (http://rwdc2.com/files/rafa.pdf)

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
5. CE-1. Ability to design and manage the collection of information and coding, handling, storing and processing it.
6. CE-2. Ability to master the proper terminology in a field that is necessary to apply statistical or operations research models and methods to solve real problems.
7. CE-3. Ability to formulate, analyze and validate models applicable to practical problems. Ability to select the method and / or statistical or operations research technique more appropriate to apply this model to the situation or problem.
8. CE-5. Ability to formulate and solve real problems of decision-making in different application areas being able to choose the statistical method and the optimization algorithm more suitable in every occasion.
Translate to english
9. CE-6. Ability to use appropriate software to perform the necessary calculations in solving a problem.
10. CE-9. Ability to implement statistical and operations research algorithms.

Transversal:
1. ENTREPRENEURSHIP AND INNOVATION: Being aware of and understanding how companies are organised and the principles that govern their activity, and being able to understand employment regulations and the relationships between planning, industrial and commercial strategies, quality and profit.

2. SUSTAINABILITY AND SOCIAL COMMITMENT: Being aware of and understanding the complexity of the economic and social phenomena typical of a welfare society, and being able to relate social welfare to globalisation and sustainability and to use technique, technology, economics and sustainability in a balanced and compatible manner.

3. TEAMWORK: Being able to work in an interdisciplinary team, whether as a member or as a leader, with the aim of contributing to projects pragmatically and responsibly and making commitments in view of the resources that are available.

4. EFFECTIVE USE OF INFORMATION RESOURCES: Managing the acquisition, structuring, analysis and display of data and information in the chosen area of specialisation and critically assessing the results obtained.

TEACHING METHODOLOGY

Student participation will be implemented in three ways
- Through its active participation in the discussions raised (online) in the form of debates (at least one for each part of the course).
- By submitting small exercises sugegsted in class qith fortnightly periodicity.
- With the completion and submission of two assignments (eg: the analysis of a microarray dataset and a second one such as the analysis of an NGS dataset).

LEARNING OBJECTIVES OF THE SUBJECT

Molecular Biology, along with Biomedicine (and at the same time Statistics), has received a great boost in recent years due to, among other reasons, the possibility of generating massive data, the best known of which is that of the human genome. Once the sequences of genomes has been available data generation has not stopped but, instead, has increased considerably. For example, microarray technology, only 10 years old, has allowed us to conduct experiments where simultaneous analysis can be performed on an individual with the goal of describing a certain pathological situation or to predict the evolution of a biological process.
The goal of this course is to present some of the problems that appear when using high throughput technologies and to show how to apply statistical methods to deal with these problems. This application can be separated into two aspects:
- On the one hand, there is the application of conventional statistical methods toward these new problems.
- On the other hand, there is the need to develop new methods and new tools in order to be able to manage this new data.
Both issues will be addressed in the course.

Skills to be acquired
Abilities acquired throughout this course will be:
- Knowledge of the different high-throughput data types and the techniques used to generate them.
- Knowledge of the methods for dealing with (collecting, preprocessing, analyzing, storing) high-performance data, giving special importance to the possibillity of carrying out a process of complete analysis: from generation up to obtaining results.
- Knowledge of the methods and of some of the existing tools for processing. Special importance will be given to the use of free and public software, especially the R language.

STUDY LOAD

Type	Hours	Percentage
Self study	80,0	64.00
Hours large group	30,0	24.00
Hours small group	15,0	12.00

Total learning time: 125 h

CONTENTS

1. Introduction to molecular biology, omics and high throughput technologies

Description:
1.1 Basic concepts of molecular biology
1.2 Methods for obtaining high throughput data
1.2.1 Overview
1.2.2 Gene expression microarrays
1.2.3 Other high throughput data (Next Generation Sequencing, Proteomics, Metabolomics, ')

Full-or-part-time: 6h
Theory classes: 3h
Practical classes: 3h

2. Analysis of microarray data

Description:
2.1 An overview of the analysis of microarray expression data
2.2 Reading and quality control of images.
2.3 Preprocessing: Normalization and filtering.
2.4 Detection of differentially expressed genes
2.4.1 Some issues: power analysis and multiple testing.
2.5 Pattern searching using cluster analysis
2.6 Molecular Diagnostics and classification methods.
2.6.1 Statistical problems which appear in building and validating classification models.
2.7 The gene ontology and its applications for biological interpretation.

Full-or-part-time: 20h
Theory classes: 10h
Practical classes: 10h

3. Analysis of other high-throughput data

Description:
3.1 NGS data analysis: Overview of data and technologies
3.2. Quality control and data preprocessing.
3.3 Differential expression analysis using NGS
3.4 Other types of studies: metagenomics, and exome variant analysis.

Full-or-part-time: 14h
Theory classes: 7h
Practical classes: 7h

GRADING SYSTEM

Continuous assessment will take place based on the participation of students in each of the activities described in the section Organization. The assessment of each of the activities will be:
· Class participation and discussion: 10%
· Completion of exercises in class: 30%
· Completion of the proposed continuous assessment tests: 60%

BIBLIOGRAPHY

Basic:
- Draghici, S. Statistics and data analysis for microarrays using R and bioconductor [on line]. 2nd ed. Chapman & Hall/CRC Mathematical & Computational Biology, 2012 [Consultation: 03/03/2021]. Available on: https://www.taylorfrancis.com/books/9780429130588.
- Tuimala, Jarno ; Laine, M. Minna. DNA microarray data analysis [on line]. 2nd ed. CSC, the Finnish IT center for Science, 2005Available on: https://www.researchgate.net/publication/261680899_DNA_Microarray_Data_Analysis_second_edition. ISBN 9525520129.
- Gibson, G. ; Muse, S.V. A Primer of genome science. 3rd ed. 2012.
- Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W. Bioinformatics and computational biology solutions using R and bioconductor. New York: Springer, 2005.
- Irizarry, R.A; Love, M.I. Data Analysis for the Life Sciences [on line]. 2015Available on: https://www.perlego.com/book/1573996/data-analysis-for-the-life-sciences-with-r-pdf.

RESOURCES

Other resources:
Aside from these books, there is a large quantity of free and high quality information on the Internet.
- The Wentian Li Portal: A portal with all kinds of information regarding microarray data analysis.
- StatWeb: Webpage with links to programs, groups, data, etc.