Topics in Applied Statistics: Statistical Genetics - Bioinformatics

Instructors: P.PAPASTAMOULIS

Course Code: 61234
Course Type: Elective of Course Group 1
Course Level: Graduate (MSc)
Year of Study: A’
Semester: 2nd
ECTS: 3
Language: English

Course Description

Modern biology is a data-rich science. This course will expose the students to high-throughput biological datasets (such as microarrays, RNA-Seq, ChIP-Seq) and present the main inferential tools to deal with challenges they impose to the statistician. These methods include techniques for:

- controlling the False Discovery Rate in multiple testing (such as the Benjamini-Hochberg procedure)
- modelling high-throughput count data (multifactorial designs, generalized linear models)
- performing differential expression analysis in microarray and RNA-Sequencing data
- taking into account heterogeneity in sizeable data (mixture models)
- fitting (frequentist or Bayesian) models specifically designed for estimating gene and tran- script expression given a known genome/transcriptome annotation and (big) datasets of short nucleotide reads

Prerequisites

This course is tailored to a statistically trained audience. More specifically:

- Prerequisites
- Estimation/Hypothesis Testing theory
- (Generalized) Linear Models

Some basic knowledge on:

- Computational Statistics
- Bayesian Inference
- R programming

Students will also benefit from the following courses (not required):

- Bayesian Statistics
- Statistical Learning
- Statistics for Big Data

Target Learning Outcomes

After completing the course, the students will:

- know the basic statistical challenges in bioinformatics
- properly deal with large scale hypothesis testing
- learn many novel statistical ideas and methods developed in the last 20 years, such as hybridizations of Bayesian and frequentist data analysis
- put their hands on many different types of data that modern biologists have to deal with, including microarrays, RNA-Seq, chip-Seq and single cell measurements
- know how to apply the relevant methods using R and Bioconductor.

Recommended Bibliography

- Holmes, Susan and Wolfgang Huber. Modern Statistics for Modern Biology. Cambridge University Press, 2019
- Efron, Bradley. Large scale inference: Empirical Bayes Methods for Estimation, Testing and Prediction. Cambridge University Press, 2010
- Gentleman, Robert, et al., eds. Bioinformatics and computational biology solutions using R and Bioconductor. Springer Science & Business Media, 2006
- McLachlan, Geoffrey and David Peel. Finite Mixture Models. Wiley Series in Probability and Statistics, 2000
- Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B, 1995
- Dudoit, Sandrine and Shaffer, Juliet Popper and Boldrick, Jennifer C. Multiple hypothesis testing in microarray experiments. Statistical Science, 2003
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expres- sion analysis of digital gene expression data. Bioinformatics, 2010
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 2014
- Li, B., Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 2011
- Glaus, P, Honkela, A, Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics, 2012
- Hensman, J, Papastamoulis, P, Glaus, P, Honkela, A, Rattray, M. Fast and accurate approximate inference of transcript expression from RNA-seq data. Bioinformatics, 2015
- Lönnstedt, Ingrid and Speed, Terry. Replicated Microarray data. Statistica sinica, 2002
- Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology, 2004

Teaching and Learning Activities

The computational aspects of this course will be implemented in R, a free software environment for statistical computing and graphics. R can be downloaded at https://www.r-project.org and installed on all types of environments (Windows, Mac, Linux). The R programming language will be enhanced by the specialized method packages from the Bioconductor project https://www.bioconductor.org, such as limma, DeSeq2, edgeR, BitSeq, rsem-EBSeq. Supplementary command line tools (such as Bowtie2) will also be used.

Assessment and Grading Methods

There will be a total of 2 homework assignments that will contribute ≈ 50% in the final grade. The remaining ≈ 50% will be determined by the final exam.