BIOINFORMATICS AND DATA SCIENCE - 2024/5

Module code: BMSM031

Module Overview

This is an introductory course on bioinformatics and data science aimed at medical and life sciences undergraduate and postgraduade students that did not have previous exposure to quantitative methodologies. There are no pre-requisites in terms of advanced algebra, calculus, probability theory, statistics or computer programming.

Computer programming (coding) and data science are salient skills needed for securing academic and industry jobs. Worldwide, the bioinformatics job market has experienced robust growth over the last decade. In the UK, there is chronic deficit of bioinformaticians as documented by the government's "Full review of the Shortage Occupation List"

Bioinformatics and Data Science underlies many academic disciplines and score highly among the most useful skills you will learn in your degree. In the Life and Health Sciences, Bioinformatics and Data Science is essential for epidemiologists, geneticists, biologists, and biomedical scientists to convert research questions into testable statistical models, and to produce interpretable, reproducible, and valid results.

This module introduces both facets of study design and data analysis with the aim of enabling graduate students to independently make scientific investigations in a coherent and reproducible way, and to apply notions of causality, statistical inference, and artificial intelligence. It also provides practical hands-on experience conducting computational genomics and gene expression analyses, two bioinformatics skills in high demand worldwide

Module provider

School of Biosciences

Module Leader

ROCCO Andrea (Biosciences)

Number of Credits: 15

ECTS Credits: 7.5

Framework: FHEQ Level 7

Module cap (Maximum number of students): 40

Overall student workload

Independent Learning Hours: 92

Lecture Hours: 11

Seminar Hours: 3

Practical/Performance Hours: 22

Guided Learning: 11

Captured Content: 11

Module Availability

Semester 2

Prerequisites / Co-requisites

None.

Module content

Indicative content includes:

1) Introduction to programming in R: - Introduction to R environment and RStudio - Data types - Input and Output of Data - Exploring, formatting and manipulating data - Plotting and Saving Plots. - Conditionals and Loops - Functions

2) Introduction to statistics and data analysis: - Descriptive statistics - Statistical Inference - Multiple testing correction - Linear regression

3) Computational genomics: - Sequence databases - Sequence alignment - Blast Search - Sequence annotation

4) Gene expression analysis: - Pre-processing of RNA-Seq data - Differential gene expression analysis of RNA-Seq data - Pathway and Gene ontology enrichment

5) Introduction to Artificial Intelligence: - Clustering and unsupervised learning - Classification and supervised learning - Overfitting and Model Complexity

6) Introduction to causal inference - Definition of causality as a concept - Difficulties with causal inference - The counterfactual framework - Definitions of treatment effect

7) Design of Observational and Experimental Studies: - Measure of effect and measures of occurrence. - Experimental studies: Blocked designs, Randomized clinical trials, Adaptive clinical trials - Observational studies: Cohort, Case-control and Cross-sectional

Assessment pattern

Assessment type	Unit of assessment	Weighting
Oral exam or presentation	Seminar: Study design and analysis plan	30
Coursework	Written report and source code	70

Alternative Assessment

The alternative assessment for ‘Seminar: Study design and analysis plan’ is a ‘Video recording of their seminar presentation’.

Assessment Strategy

The assessment strategy is designed to provide students with the opportunity to demonstrate they have achieved the learning outcomes by testing their ability to:

1) Evaluate methodologies and study designs that address concrete research questions;

2) Design, develop and implement a data analysis plan;

3) Demonstrate effective communication and computer programming skills

Thus, the module has both summative and formative assessments.

Summative assessment for this module consists of:

1) Seminar (30%). The student will make an oral (PowerPoint) presentation of a study design and analysis plan to answer a research question. During the seminar, the student will obtain feedback from their peers and will be assessed on i) communication skills and on ii) the coherence with which the study design and the analysis plan addresses the research question.

2) Report (70%). At the end of the module, the student will submit a mini-report with the outcome of the mini-project together with the source code of the analyses. The report will be assessed for i) the coherence of analysis and results, ii) the ability to display data (tables and graphs), iii) the ability to interpret results and iv) the understanding of limitations in statistical methodology and study design. The source code will be assessed for good coding practices like modularity and encapsulation, pertinent use of software packages, good model building skills (like controlling batch effects) and adequate exploration of algorithms parameter space.

Formative assessment - Practical sessions will provide students with opportunities to assess their progress. The student will conduct analyses, discuss their results with colleagues and evaluate their colleagues' results. The student will also compare their results with solutions provided by the module convener. This will help the student assess their own progress.

Module aims

- Introduce students to the field of bioinformatics and data science.
- Develop an effective command of computer programming (coding) in R.
- Understand the inter-relation between experimental design and statistical methodology when addressing a research question.
- Understand the impact of sampling uncertainty and adequate statistical inference on the reproducibility of experimental and observational results.
- Independently construct study designs and analysis plans to answer research questions. This involves independently select, conduct, interpret and present results obtained by statistical and artificial intelligence methodologies.
- Evaluate statistical and bioinformatics analyses and critically interpret results presented in a scientific paper

Learning outcomes

		Attributes Developed
001	To develop, debug and test data analysis scripts and functions in the R programming language	KP
002	To select and apply the appropriate descriptive statistics for categorical and numeric variables, and critically interpret their results	KCP
003	To select and apply adequate statistical hypothesis test for questions involving measures of central tendency and 1-way association tables, and critically interpret their results	KCP
004	To display data graphically and interpret graphs	KP
005	To conduct linear regression analysis and critically interpret their results.	KCP
006	To understand the basic elements of causal inference, and know their importance for the construction of designs and interpretation of data analyses	KCPT

		Attributes Developed
007	To elicit hypotheses and construct effective study designs for observational and experimental studies	KCPT
008	To apply, interpret, and evaluate the results of supervised and unsupervised methods and understand the fundamentals of machine learning like model complexity, model selection, overfitting.	KCP
009	Apply and interpret results of methodologies to align and annotate genomes and to analyse gene expression	KCP

Attributes Developed

C - Cognitive/analytical

K - Subject knowledge

T - Transferable skills

P - Professional/Practical skills

Methods of Teaching / Learning

The learning and teaching strategy is designed to:

1) Practice both programming and application of data analysis methods to realistic problems;

2) Foster independent and critical thinking about data science and bioinformatics.

3) Communicate clearly and succinctly both verbally and in writing

The learning and teaching methods include:

1) Interactive active- learning sessions combining exposition with computer practicals -- provide opportunities for understanding critical concepts, hands-on programming, practical application of theoretical concepts, and face-to-face feedback and guidance. (33h).

2) Seminars (3h) -- provide opportunities for students to effectively communicate, ask questions, and give feedback.

3). Independent learning – Mini-project using a problem-based learning framework (114h)

Alternative learning and teaching methods for students that could not attend the course

Students can learn using the online content provided by the course. These include the capture content, power-points, books, practical exercises with answers, data and source code of the computer programs developed in the classroom/computer lab. In addition, students can book appointments during teachers office hours to discuss any relevant material.

Indicated Lecture Hours (which may also include seminars, tutorials, workshops and other contact time) are approximate and may include in-class tests where one or more of these are an assessment on the module. In-class tests are scheduled/organised separately to taught content and will be published on to student personal timetables, where they apply to taken modules, as soon as they are finalised by central administration. This will usually be after the initial publication of the teaching timetable for the relevant semester.

Reading list

https://readinglists.surrey.ac.uk
Upon accessing the reading list, please search for the module using the module code: BMSM031

Other information

The module will contribute to the five pillars of graduate learning at the university of surrey.

i) To enhance employability, students will be equipped with essential skills for the future, like communications and writing skills, the logic of scientific inference, principles of statistical analysis, artificial intelligence, causal inference and experimental design.

ii) The module will develop communication skills in seminars and group work during practical sessions in the computer laboratory. The classroom will be a safe space where students learn by giving and receiving feedback.

iii) To extend digital competencies, the student will acquire computer programming and big data analysis using statistical and artificial intelligence methods.

iv) To expand cultural and global capabilities, students will be exposed to data analysis examples from global health and one health issues across the world.

v) To improve resourcefulness and resilience, students will work independently to apply the learned methodologies in a mini-project, which will require self-regulation as the work is distributed across the duration of the module.

The module will also have an ethos of hybridity and flexible learning. Students will have online access to videos and materials covering the course content, extending the learning environment outside the computer laboratory and lecture room.

Programmes this module appears in

Programme	Semester	Classification	Qualifying conditions
Biomedical Science MSci (Hons)	2	Compulsory	A weighted aggregate mark of 50% is required to pass the module