BBS BIOINFORMATICS MINOR

Overview

Welcome to the Part II BBS Bioinformatics Minor. Bioinformatics is an interdisciplinary field that uses computational approaches to process biological data. With the biological and biomedical sciences becoming more data-driven than ever before, bioinformatics is becoming central to research and work in these areas. The BBS Bioinformatics minor subject consists of two sections:

1. A 2-week **foundations** block that introduces the fundamental bioinformatic concepts and methodologies used to analyse biological data. The material covered in this block is non-examinable
2. A 6-week **core material** block that covers specific bioinformatics and machine learnings related concepts. The material covered in this block is examinable.

The course is aimed at students coming from the biomedical sciences who have had little exposure to computational biology fundamentals. As such it provides data science foundation sessions that go over programming, data visualisation and manipulation, and basic statistics, all of which will be used throughout the course.

Topics throughout the course are introduced through a set of lectures that introduce theoretical concepts, and practicals which provide some hands-on practice using real biological datasets.

We look forward to welcoming you to the course in January.

Learning Aims & Objectives

Aims

The course has the following aims:

1. To introduce students to common methods used in Bioinformatics and their application to biological data.
2. To develop students' skills in basic data science methods and provide exposure to the computational, statistical and machine learning knowledge required to begin to analyse biological data and systems.
3. To provide an introduction to omics applications, focusing on the use of next generation sequencing and Genome Wide Association Studies (GWASs).
4. To introduce gene ontologies and gene-set enrichment analysis used to link downstream analysis results back to the underlying biology.

Objectives

At the end of the course, students will be able to:

1. Be able to interpret scripts and output from common bioinformatics related software packages.
2. Understand the different stages involved in processing and analysing omics data and be familiar with software packages used to perform such analysis.
3. Explain basic bioinformatics concepts and use gene ontologies and gene-set enrichment analysis methods to map results of bioinformatic analysis back to their biological function.

After attending this module, students will not be independent in the analysis of complex biological data but will have acquired the critical thinking needed to understand what the analysis of genomic data entails, what are the strengths and weaknesses of different analysis strategies, and will be equipped with a basic set of bioinformatics skills that will enable them to explore and interpret genomic data, as well as other types of biological data, available in the public domain.

Teaching and Assessment

Teaching

The course consists of a set of lectures that introduce theoretical concepts, and practicals which provide hands-on practice using real biological datasets. Large group tutorials will be provided for the different course blocks to provide support to the students and give them the opportunity to interact with the tutors and discuss questions they might have. Furthermore, the students will have access to helpdesks on the course VLE where they can ask questions anytime throughout the course. There will be opportunities for one-to-one discussion with tutors throughout the course.

Students will be given access to the online training environment which provides a pre-configured environment with the software used throughout the course. As such there will be no need to install software in advance. Software installation instructions will be provided. In addition, IT support will also be provided in case students would like to install the software on their computer.

Examination

Assessment will be via a 3-hour written exam paper.

Changes for 24/25

N.B. The structure of the exam will be different for the 24/25 academic year compared with previous years. For the 24/25 exam there will be six questions, each corresponding to one of the six topics covered in the core materials block. Students will be expected to answer all six questions, with each question expected to take students approximately 30 minutes to complete. All questions will be weighted equally.

Guidance will be provided on the VLE on how past paper exam questions will map onto this year’s exam question structure, so no student will be disadvantaged by this change in terms of access to appropriate past paper material.

Course Structure

Lecture Timetable

Block	Week	Date	Time	Topic	Location *
Foundations	0	Tue Jan 21	3:00 – 5:00pm	Unix	Craik-Marshall
	1	Mon Jan 27	3:00 – 5:00pm	R	Craik-Marshall
		Tue Jan 28	3:00 – 5:00pm		Titan
	2	Thu Jan 30	4:00 – 5:00pm	Statistics	Biffen
		Mon Feb 3	3:00 – 5:00pm		Craik-Marshall
		Tue Feb 4	3:00 – 5:00pm		Craik-Marshall
Core Materials	3	Thu Feb 6	4:00 – 5:00pm	Unsupervised Machine Learning	Titan
		Mon Feb 10	3:00 – 5:00pm		Titan
		Tue Feb 11	3:00 – 5:00pm		Titan
	4	Thu Feb 13	4:00 – 5:00pm	Sequence Alignment & Phylogenetics	Biffen
		Mon Feb 17	3:00 – 5:00pm		Craik-Marshall
		Tue Feb 18	3:00 – 5:00pm		Craik-Marshall
	5	Thu Feb 20	4:00 – 5:00pm	NGS	Biffen
		Mon Feb 24	3:00 – 5:00pm		Craik-Marshall
		Tue Feb 25	3:00 – 5:00pm		Craik-Marshall
	6	Thu Feb 27	4:00 – 5:00pm	Differential Expression & Gene Set Enrichment Analysis	Biffen
		Mon Mar 3	3:00 – 5:00pm		Craik-Marshall
		Tue Mar 4	3:00 – 5:00pm		Craik-Marshall
	7	Thu Mar 6	4:00 – 5:00pm	Supervised Machine Learning	Titan
		Mon Mar 10	3:00 – 5:00pm		Titan
		Tue Mar 11	3:00 – 5:00pm		Titan
	8	Thu Mar 13	4:00 – 5:00pm	GWAS	Biffen
		Mon Mar 17	3:00 – 5:00pm		Craik-Marshall
		Tue Mar 18	3:00 – 5:00pm		Craik-Marshall

*Not all locations have been finalized.

Topics

Foundations - Unix

Dr Bajuna Saleha

In this practical we will explore the basic structure of the Unix command line and how we can interact with it using a basic set of commands. You will learn how to navigate the filesystem, manipulate text-based data and combine multiple commands to quickly extract information from large data files. You will also learn how to write scripts and use programmatic techniques to automate task repetition.

Foundations - R

TBD

In these practicals we will learn about the basic programming concepts in R that also form the basis of any programming language. We also learn about data types and data structures and how we can use these to read and store data. We will then learn about the umbrella package tidyverse, as well as two popular packages; the ggplot2 package which allows us to visualise data professionally and the dplyr package which is used to manipulate data effectively.

Foundations - Statistics

Dr Vicki Hodgson

Statistics is an important component to the analysis of data. We will learn about Linear Models, an approach for modelling the relationship between a single scalar response and one or more explanatory variables (simple linear regression and multiple linear regression, respectively). In these sessions we will focus on approaches for estimating coefficients, interpretation of coefficients and model selection approaches.

Core Materials – Unsupervised Machine Learning

Dr Soumya Banerjee

We will discuss, compare, and contrast several methods to extract patterns and structures from data. First, we will learn how to use cluster analysis to identify subgroups in our dataset. We will define different similarity measures and explore several algorithms, focusing mainly on the k-means algorithm. We will also learn how to obtain representations of our data in a smaller dimension so we can visualise dependencies, identify structures and obtain new variables that retain most of the information in our experiments.

Core Materials – Sequence Alignment and Phylogenetics

Dr Katy Brown

In these sessions we will cover the principles of sequence alignment, a fundamental step in many bioinformatics analyses. The lectures will cover different sequence alignment strategies and how they work, how sequence alignment is carried out in practice and some downstream applications. We will also cover basic phylogenetic analysis and how sequence alignments are used in this context. In the practical students will generate their own alignments and explore some potential applications.

Core Materials – Next Generation Sequencing

Dr Sergio Martinez-Cuesta

The lecture will provide an overview of the Next Generation Sequencing (NGS) technology, including library preparation and sequencing by synthesis. We will examine key file formats, learn how to assess sequencing data quality, and understand how reads aligned to a reference genome can be used to infer genomic variants and interactions between proteins and DNA. The practical will provide a hands-on opportunity to perform quality control, alignment, variant calling and prediction of functional effects of variants.

Core Materials – Differential Expression and Gene-set Enrichment Analysis

Dr Katy Brown

In these sessions we will discuss how differential gene expression (DGE) analysis is carried out, including various quality control steps and different algorithms. We will also discuss downstream analyses with the resulting lists of differentially expressed genes, focussing on gene set enrichment analysis.

Core Materials – Supervised Machine Learning

Dr Soumya Banerjee

We will introduce the concept of supervised learning: the task of learning a function that maps an *input* to an *output* (categorical or continuous label). We will introduce basic terminology, and link foundational statistical approaches, e.g. linear and logistic regression, to classical approaches such as k nearest neighbours, support vector machines and decision trees.

Core Materials – Genome Wide Association Studies

Dr Ruhina Laskar

The lecture and practicals will provide an overview of genome-wide association studies, introducing key concepts and considerations when performing a GWAS and interpreting results. We will go through step-by-step exercises on the quality control of genotype data and how to conduct genetic association analyses. This will be performed using PLINK and R. We will also cover applications of GWAS, and prediction of disease risk using polygenic risk scores.