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 resources 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 via the Inspera platform.

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 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.

Department of Genetics Data Retention Policy (Examinations)

Data retained by the Department of Genetics, Downing Street, Cambridge and further information on the policy can be reviewed by following the link here.

Feedback

We would like to thank you for being part of the NST Part II BBS Bioinformatics 2024-25 course. We hope you enjoy the course and acquire new knowledge and skills in Bioinformatics. We kindly ask you to fill in the feedback surveys provided. Each survey should take around 2 minutes to complete. Your input is valuable to us as through your feedback we can understand what things we are doing well and what we need to improve.

If you would like to provide 1:1 feedback or discuss any concerns directly, please do get in touch with the course organiser or administrator and we will be happy to chat.

Course Structure

Lecture Timetable

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

To view the student online timetable on the **University of Cambridge Student Timetable 2024-2025**, please go to **www.timetable.cam.ac.uk**

(Follow NST Tripos >Biological and Biomedical Sciences >Part II >128-Bioinformatics). You will have to sign in using Raven.

Venues

- **Bioinformatics Training Room, 1st Floor, Craik-Marshall Building, 20 Downing Pl, Downing Site, Cambridge, CB2 3DT**
  (https://new.map.cam.ac.uk/?maplon=0.12227&maplat=52.20192&mapzoom=18&maplayers=Buildings%2CBuilding+Labels%2CUniversity+Sites&mapbasic=true&mapfeature=mfid171%2CBuildings).
  Access to the building is via Reception, which is not monitored. If you find yourself locked out, you can call the phone number for the Training Room 01223 (3)33614, and someone will come to let you in.

- **Titan Teaching Room 2, 2nd Floor, Cockcroft Building, New Museums Site, Pembroke Street, Cambridge CB2 3QH**
  (https://new.map.cam.ac.uk/?maplon=0.12054&maplat=52.20340&mapzoom=18&maplayers=Buildings%2CBuilding+Labels%2CUniversity+Sites&mapbasic=true&mapfeature=mfid258%2CBuildings)
  Please bring your own fully charged laptops. Only accessed via stairs.

- **University Centre, Cormack Room, Granta Place, Mill Lane, Cambridge, CB2 1RU**
  (https://new.map.cam.ac.uk/?maplon=0.11713&maplat=52.20137&mapzoom=18&maplayers=Buildings%2CBuilding+Labels%2CUniversity+Sites&mapbasic=true&mapfeature=mfid166%2CBuildings)
  Please bring your own fully charged laptops.

- **Student Services Centre, Lecture A (Arts School), New Museum Site, Bene't St, Cambridge CB2 3PT**
  (https://new.map.cam.ac.uk/?maplon=0.12000&maplat=52.20381&mapzoom=18&maplayers=Buildings%2CBuilding+Labels%2CUniversity+Sites&mapbasic=true&mapfeature=mfid1527%2CBuildings)
  Please bring your own fully charged laptops.

We are aware that there is a mix of venues which may be difficult to navigate. Please refer to the timetable and venues above, in the first instance, and make particular note of specific locations for various sessions dates.

General Information

Lectures and practical sessions will be delivered **in-person.** In the first instance please refer to the timetable for delivery method and location. You will be informed of any

changes to the arrangements on an ongoing basis throughout the term – through emails to the class.

Lectures

Lectures will be recorded and released on Moodle for future reference as soon after the session as possible. Please refer to the Recording Policy section (in Useful links below) for terms of use. Inevitably, recordings will differ in their quality: recordings made as a lecture is being delivered will be less perfect than a lecture that has been pre-recorded and will include lecturer miss-speaks and other glitches. It is important to keep this in mind when viewing lecture recordings.

Practical Sessions

Where possible, Practical material will be uploaded on the VLE in the relevant resources section before the timetabled session. Students may work through the material in their own time beforehand and during the timetabled practical session. Teaching staff will be available at the sessions to answer students' questions and help. **The practical sessions will not be recorded.**

Prior to the course, students enrolled in the course will receive access to the Online Training Environment where students can access and use pre-configured and installed software that is used for the practicals throughout the course. Students will have access to this until the Bioinformatics Exam. As such there will be no need to install software in advance, unless specifically requested. **Important: Please make sure you save files from the online training environment to your own computer to avoid data loss as retainment of data on the online training environment is not guaranteed.** Alternatively, you can install software used in practicals on your computer. Installation instructions on the software used in each practical are provided in the respective practical section on this site.

IT support will be provided if students have problems installing software on their computer. Please contact IT Manager Paul Judge (pj237@cam.ac.uk) if you need IT support assistance.

Topics

Foundations - Unix

Dr Bajuna Salehe

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 & 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 & 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 (GWAS) that covers the key concepts on genetic variation, the use of single nucleotide polymorphisms (SNPs) and methods used to identify genetic variations associated with complex traits and diseases. We will go through step-by-step exercises on study design, popular genotyping platforms, quality control steps and the statistical methods for association testing, managing population stratification, and addressing multiple comparisons. This will be performed using PLINK (https://zzz.bwh.harvard.edu/plink/) and R Studio. We will also cover how to interpret GWAS findings, linking genetic variations to biological pathways and additional discussions on how GWAS findings are applied in prediction of disease risk using polygenic risk scores and pharmacogenomics.

Contacts

Key Course Contacts

Course Organiser: Dr Matt Castle (mdc31@cam.ac.uk)

Course Administrator: Cathy Hemmings (cgh32@cam.ac.uk)

Course IT Systems Administrator: Paul Judge (pj237@cam.ac.uk)

General

BBS Student Liaison and General SBS Admin: FacBiol@admin.cam.ac.uk

PDN Part II Teaching Admin: part2@pdn.cam.ac.uk

Zoology Teaching Admin: teaching@zoo.cam.ac.uk

Plant Sciences Part II Teaching Admin: ugadmin@plantsci.cam.ac.uk

Lecturers

Dr Bajuna Salehe (bs579@cam.ac.uk)

Dr Vicki Hodgson (vjh33@cam.ac.uk)

Dr Katy Brown (kab84@cam.ac.uk)

Dr Soumya Banerjee (sb2333@cam.ac.uk)

Dr Sergio Martinez-Cuesta (sermarcue@gmail.com)

Ruhina Laskar (rl757@medschl.cam.ac.uk)

Useful links

BBS Website

BBS website for students: https://www.biology.cam.ac.uk/undergrads/nst/bbs

Policy for Recording

The University's policy for recording:
https://www.educationalpolicy.admin.cam.ac.uk/policy-index/recording

Permissible Subject Combinations

It is essential for BBS students, studying in NST Part II BBS, to be able to attend ALL lectures offered by both their Major and Minor Subjects. For this reason, only combinations of subjects which have compatible timetables are permitted.

For information on permissible combinations please refer to the following link:
https://www.biology.cam.ac.uk/undergrads/nst/bbs/subject-combinations

Corresponding Cohort Paper Codes

(NST2BBS) NST Part II BBS minor (Bioinformatics) (Paper code = 102)

(NST2BBS) NST Part II BBS (Zoology) (Paper code = 11_18)

(NST2ZO) NST Part II Zoology (Bioinformatics module) (Paper code= 102)

(NST2BBS) NST Part II BBS (PDN) (P5 Bioinformatics) (Paper code = 2_16)

(NST2PDN) NST Part II BBS PDN (Bioinformatics module) (Paper code = 102)

(NST2BBS) NST Part II BBS (Plant Sciences) (Paper code = 9_11)

(NST2PL) NST Part II Plant Sciences (Bioinformatics module) (Paper code = 102)

Key:
NST = Natural Sciences Tripos
BBS = Biological and Biomedical Sciences
PDN = Physiology, Development and Neuroscience