COMP0082 Bioinformatics

Introduction to the course

Contact details

Prof. David Jones

Email: d.t.jones@ucl.ac.uk

http://www.cs.ucl.ac.uk/staff/D.Jones

See Moodle for info on office hours.

The Course- motivation for biological material

- Modern molecular biology and especially genomics has led to vast quantities of data: DNA/protein sequence, gene expression.
- This mainly consists of long strings or large matrices which in their raw form are not very interesting or informative.
- What’s needed is integration and mining of data for patterns.
- Machine (deep) learning techniques are proving to be very good for extracting useful patterns in these types of data.

From our always reliable friend, Wikipedia...

Bioinformatics is the application of information technology and computer science to the field of **molecular biology**. ... Its primary use since at least the late 1980s has been in **genomics and genetics**, particularly in those areas of genomics involving large-scale DNA sequencing. Bioinformatics now entails the creation and advancement of **databases, algorithms, computational and statistical techniques, and theory** to solve formal and practical problems arising from the **management and analysis of biological data**.

Also from our always reliable friend, Wikipedia...

Computational biology is an interdisciplinary field that applies the techniques of **computer science, applied mathematics and statistics** to address **biological problems**. The main focus lies on developing **mathematical modeling and computational simulation techniques**. By these means it addresses scientific research topics with their theoretical and experimental questions without a laboratory.

You can waste a lot of energy trying to distinguish these terms. Or simply decide it doesn’t really matter and use whichever term you prefer!

Motivation

- In order to extract useful information from biological data, it ***is*** necessary to understand biological principles involved.
- In this course we will introduce some basic molecular biology/ genomics and look at ways in which computers can be used to analyse it (bioinformatics), with an emphasis on machine learning applications (though not exclusively).

Things you should be aware of

- This module is designed as a supplement to a machine learning course
  - I will assume you know all about machine learning methods and how to apply them e.g. using Matlab, SK-Learn, PyTorch or similar tools
  - I will assume that you know little or nothing about biology
- It doesn’t cover the whole field of bioinformatics
  - e.g. there is very limited coverage of bioinformatics algorithms
- There is a bit of biological jargon to learn
  - To work in the field effectively, you have to be able to converse with biologists and even read some biology papers
- Background reading on biological topics is **NOT** optional
  - The molecular biology lecture material is not sufficient on its own to get a top exam mark
  - You are expected to do some self-study and read around the highlighted topics using e.g. the Stryer book
- The emphasis is on **biology** not computer science
  - The course is designed to help a machine learning specialist work more effectively in biology. Nothing more, nothing less.

Course lecture content

- I will give ~7 weeks of lectures at the start of the course.
- Dr. Daniel Buchan will then lecture on high throughput ‘omics methods.
- REMEMBER to regularly check Moodle for last minute announcements or changes.

Coursework & Homework

- Coursework:
  - One “mini-project” worth 40% of total marks, starting after reading week
  - Submission in the form of approximately publication-ready bioinformatics papers
  - START PLANNING TIME FOR THIS NOW!!!
    - Coding
    - Running code & collating results
    - Writing paper
- Homework:
  - Doing own reading on basic biology
  - Reading specified research papers
    - Not graded
    - But both will be important for the exam

Exam

- Written exam – 2 hours (in person)

- 60% of total mark
  - Need to answer 3 questions
  - One in section A, two (from three) in section B.
  - Section A (compulsory question) will be entirely about cell/molecular biology – no CS at all!
  - Section B will be about applications

Moodle

- *All* communication concerning this course (including lecture changes/cancellations – if any) will be done via Moodle.
- Coursework submission will be via Moodle (Turnitin)
- Make sure you are familiar with Turnitin submission before deadline day!
- Make sure you check Moodle just before every lecture in case there are urgent announcements

Slides and Handouts

Everything’s on MOODLE

Potentially Useful Books

- **Good starting point:**
  - **Kratz – Molecular and Cell Biology for Dummies (good starting point)**
- **Good reference books with excellent diagrams:**
  - **Stryer- Biochemistry**
  - **Alberts et al- Molecular Biology of the Cell**
- More specialized books on bioinformatics:
  - Baldi and Brunak – Bioinformatics – a machine learning approach (2nd ed)
  - Lesk- Introduction to bioinformatics
  - Orengo, Jones and Thornton – Bioinformatics (available as e-book via library)
  - Durbin, Eddy, Krogh and Mitchison – Biological sequence analysis (for protein HMMs)