UNIVERSITY OF CAMBRIDGE

## M. PHIL. IN COMPUTATIONAL BIOLOGY

Friday, 13 May, 2022   2:00 pm to 4:00 pm

# COMPUTATIONAL BIOLOGY

*Attempt **ALL** questions.*

*There are **THREE** questions in total.*

*The questions carry equal weight.*

***STATIONERY REQUIREMENTS***
*Cover sheet*
*Treasury Tag*
*Script paper*

***SPECIAL REQUIREMENTS***
*Calculator - students are permitted to bring an approved calculator.*

**You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.**

## 1 Deep Learning

### Hopfield Network

1. Explain the meaning of each term in the following equation for a Hopfield network. How would you repeatedly use it to update the state of the network?

$$v_i = f(\sum_{j=1}^{N} w_{ij} v_j)$$

[25%]

2. How would you determine the weights $w_{ij}$ in the network? [10%]

3. A useful quantity to calculate is:

$$E = -\frac{1}{2} \sum_{ij} w_{ij} v_i v_j$$

Explain what this measures and how it should be interpreted. [10%]

4. Define the Traveling Salesman Problem (TSP). How would you set up the Hopfield Network to solve the TSP? [15%]

### Perceptron

5. A perceptron with $N$ inputs is defined by

$$y = f(\sum_{j=1}^{N} w_j x_j)$$

Explain each of the terms in this equation. [20%]

6. Given the expression:

$$F = \frac{1}{2}(t - y)^2 + \beta \sum_{j=1}^{N} w_j^2$$

where $\beta > 0$, what is $t$? What does $F$ measure? Derive a rule for updating weights to minimise $F$. [20%]

## 2 Genomics II

1. In the context of the following experiment, where we want to estimate the expression of a given gene under the values of ER (level of oestrogen receptor with two possible values, positive + and negative −), and the dose (different doses of radiation):

| Sample | ER | Dose |
|---|---|---|
| Sample 1 | + | 37 |
| Sample 2 | − | 52 |
| Sample 3 | + | 65 |
| Sample 4 | − | 89 |
| Sample 5 | + | 24 |
| Sample 6 | − | 19 |
| Sample 7 | + | 54 |
| Sample 8 | − | 67 |

   (a) Write down the design matrix for a model with no interaction between ER and Dose.

   (b) Write down the design matrix for a model with interaction between ER and Dose.

   (c) Give two advantages of using RNA-SEQ instead of microarrays to run this experiment. [25%]

2. In the context of pathway analysis:

   (a) Explain what are the goals and the steps needed to run a gene ontology analysis and a gene set enrichment analysis.

   (b) When we are trying to quantify the involvement of a pathway in our data set, is there any difference in using just the list of genes or split by overexpression and underexpression? [25%]

3. Describe a typical ChIP-SEQ analysis pipeline and describe the typical files generated in the process (fastq, bam, bed) until obtaining a list of differentially binding sites. [25%]

4. In the context of single cell RNA-SEQ analysis, explain:

   (a) What are unique molecular identifiers (UMI's) and what are their advantages?

   (b) What are doublets and how can we detect them?

   (c) Explain briefly the goals of differential abundance analysis and pseudotime analysis [25%]

## 3 Genome Sequence Analysis

1. Give a formal definition of a Hidden Markov Model (HMM) and its parameters. [30%]

2. A scientist has collected data on the binding of a certain protein to DNA in a sample of cells. The data show, for each position in the genome, the proportion of cells in the experiment in which the protein was bound at that position. She would like to identify regions of the genome that are particularly rich in binding sites and also those that are depleted, relative to 'typical' regions.

   Discuss how a HMM might be used to address this question, indicating how the data might be processed, what the HMM would look like and what implementation issues there might be to consider. [40%]

3. Define the unscaled forward variable in a HMM, and show that it can be calculated iteratively given a set of parameter values and observed data. [30%]

## END OF PAPER