

THE UNIVERSITY *of York*

MRes Examinations Examinations ,2008

COMPUTATIONAL BIOLOGY  
Part BI

INTRODUCTION TO MACHINE LEARNING (IML)  
Open Examination

Issued at:

**Monday 23rd March 2009**

Submission due:

**Monday 6th April 2009**

Your attention is drawn to the Guidelines on Mutual Assistance and Collaboration in the Student's Handbook.

All queries on this assessment should be addressed to  
**James Cussens, Daniel Kudenko & Simon O'Keefe.**

**Your examination number must be written on the front of your submission.  
You must not identify yourself in any other way.**

---

## **1 Profile hidden Markov models (40 marks)**

### **1.1 Alignments and profile HMMs (20 marks)**

Consider the fictional protein multiple alignment given in Figure 1. Suppose a profile HMM is built to model this multiple alignment, where the columns marked \* correspond to five match states. The profile HMM will have the architecture of a main model under the HMMER Plan 7 architecture. An example of a Plan 7 HMM can be found on page 31 of the HMMER manual found at <ftp://selab.janelia.org/pub/software/hmmer/CURRENT/Userguide.pdf>. The main model consists of the states  $B$ ,  $E$  and all states between  $B$  and  $E$ . The main model includes a delete state just after state  $B$  and a delete state just before  $E$ .

1. Draw the profile HMM, labelling the states. (1 mark)
2. For each of the 4 sequences in Figure 1 give, where possible, the state path through the profile HMM that corresponds to the alignment of that sequence. (8 marks)

```

S1 KQ.....I
S2 TSKDGLD.
S3 KHI.K.TK
S4 KASAEATA
    ***    **

```

Figure 1: Fictional protein multiple alignment

3. Estimate, where possible, the parameters of the HMM using
  - maximum likelihood estimation (3 marks) and
  - a Bayesian approach with a prior of your choice (3 marks).
4. Explain the relationship between pseudo-counts and Dirichlet distributions. (3 marks)
5. Contrast the Bayesian and maximum likelihood approaches, giving at least one advantage of each approach. (2 marks)

## 1.2 Using HMMER (20 marks)

1. Use HMMER (with default settings) to construct a profile HMM from the alignment in file `rrm.slx` which is available on the Biology Linux system at `/biol/programs/hmmer/src/hmmer-2.2g/tutorial/rrm.slx`. Ensure that the resulting HMM is in a file called `rrm.hmm`. State which command you used. (1 mark)
2. Calibrate `rrm.hmm`. State which command you used to do this calibration. (1 mark)
3. Why do we calibrate? (3 marks)
4. Give the highest scoring amino acid for the first ten positions according to `rrm.hmm`. (3 marks)
5. Why might the highest probability amino acid not be the highest scoring amino acid for some position? (4 marks)
6. Which of the sequences in `rrm.slx` (if any) pass through a delete state when aligned to `rrm.hmm`? (3 marks)
7. What is the probability of an *L* being emitted from match state 1? (Hint: You need to refer to the HMMER manual to answer this.) (5 marks)

## 2 Distinguishing Human DNA from Phage DNA (40 marks)

The problems ask you to perform inductive learning experiments on a dataset containing human and phage DNA sequences. The goal is to compute a hypothesis that is able to distinguish human sequences from phage sequences.

The files containing the sequences can be found on the IML web page. There are two sequence files: `hum-pos` (containing human sequences) and `hum-neg` (containing phage sequences). Each line in these files corresponds to one sequence. If you have problems accessing these files, please send an email to `kudenko@cs.york.ac.uk`.

Please hand in all Python scripts used in the solutions of the problems on a CD. Please provide a brief documentation for each Python script (i.e., what is the input, what is the output, and how it is executed). Also, please hand in a description of your experiments, and a description of the files on the CD (each file's purpose should be clearly identified in this description).

### 2.1 C4.5 (20 marks)

1. Transform the sequences into positional representation for C4.5. What is the error rate when applying 10-fold cross-validation? (4 marks)
2. Transform the sequences into n-gram subsequence frequency representation, for  $n = 3$  and 4. What is the best representation given the results of 10-fold cross-validation on C4.5? (7 marks)
3. Try the union of the subsequence frequency representations from the above problem (i.e. 4-gram frequency union 3-gram frequency). What is the 10-fold cross-validation error rate? (3 marks)
4. Consider the 5-gram subsequence frequency representation. This representation results in a large number of features and therefore in a slow-down of C4.5. Try a *simple* feature selection method to reduce the number of features. What is the 10-fold cross-validation error using this reduced set? (6 marks)

### 2.2 Nearest Neighbour (20 marks)

1. This problem asks you to apply nearest-neighbour learning to the sequence dataset. This will require you to implement a Python program for nearest neighbour learning. Try both 1-NN and 3-NN learning in the questions below.
1. Use positional mismatch distance (i.e., simply counting the number of character mismatches at each sequence position, as presented in the lectures). What is the 10-fold cross-validation error rate? (10 marks)

2. Use the dynamic programming algorithm presented in the lectures to compute the minimum edit-distance of sequences. The costs of insertion, deletion, and replacement of non-equal characters should all be set to 1. What is the 10-fold cross-validation error rate? (10 marks)

### 3 Protein Secondary Structure Prediction (20 marks)

Protein secondary structures are formed from the folding of the linear sequence of amino acids into a three-dimensional structure. The three classes of secondary structures are the  $\alpha$ -helix, the  $\beta$ -sheet and the random coil. There is much interest in predicting the secondary structure of proteins from the primary structure (amino acid sequence).

Numerous papers have been published on the use of neural network models to attack the problem.

1. Explain the model used by Kakumani *et al.* [1]. Your answer should discuss the data used, the neural model and how it is trained, and how the effectiveness of the architecture was evaluated. [10 marks]
2. Briefly discuss *two* examples from the recent literature of neural networks (no earlier than 2000) applied to this problem. [10 marks]

[1] Kakumani, R., Devabhaktuni, V. and Ahmad, M. O., "A Two-Stage Neural Network Based Technique for Protein Secondary Structure Prediction", 30th Annual IEEE EMBS Conference, Vancouver, BC, August 20-24, 2008.

