ELL796: Minor Test I

February 23, 2018

Maximum Marks: 8

Name: _____

Entry Number: _____

1. The Minister of State for HRD recently argued against Darwin's theory of evolution, saying it should be removed from school and college curricula in India. His argument was quoted as follows:

   > *Since man is seen on Earth, he has always been a man. Nobody, including our ancestors, in written or oral, said they ever saw an ape turning into a human being.*

   Explain what is wrong with this argument. (NB: You are not being asked to give a general defence of evolution, but rather a specific response to this argument, based on evolutionary theory.) **[2]**

   (1) Apes did not turn into humans; the most recent common ancestor was equally distant from both.

   (2) There is plenty of fossil evidence of intermediate forms/species between that ancestor and modern humans.

   (3) Language & writing evolved late in human cultural evolution!

2. Consider the below DNA sequence.

   `ACCATGTTGCCACATGTGAACCGGTACCGTAG`

   Suppose we seek to model this using a simple Markov model (without hidden states).

   (a) Give the maximum-likelihood (i.e., frequency-based) estimates of the parameters of the distribution $p(x_i|x_{i-1} = \texttt{A})$, the distribution over nucelotides where the preceding nucelotide is `A`. **[2]**

   $p(A|A) = \frac{1}{8}$

   $p(C|A) = \frac{1}{2}$

   $p(T|A) = \frac{1}{4}$

   $p(G|A) = \frac{1}{8}$

   (b) Let H_A denote the Shannon entropy of the above distribution. Compute the value of H_A. (The general

formula for Shannon entropy is $H = -\sum_i p_i \log_2 p_i$.) [1]

$$-\left(-\frac{3}{8} - \frac{3}{8} - \frac{2}{4} - \frac{1}{2}\right) = \frac{14}{8}$$

(c) What is the maximum possible value that H_A could have for any DNA sequence? [0.5]

$$4\left(-\frac{1}{4}\log_2\frac{1}{4}\right) = 2$$

(d) In general, do you expect H_A to be lower for coding regions, or for non-coding regions? Why? [0.5]

For coding regions, due to expected non-uniformity in codon distribution.

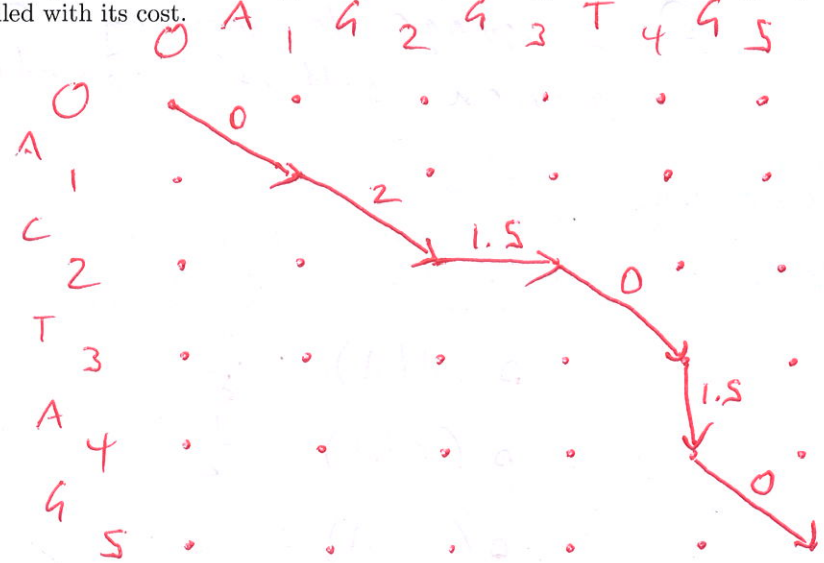
3. Consider the below alignment of two DNA sequences.

A	C	–	T	A	G
A	G	G	T	–	G

Suppose that we have the following cost matrix δ:

	A	C	T	G	–
A	0	2	2	1	1.5
C	2	0	1	2	1.5
T	2	1	0	2	1.5
G	1	2	2	0	1.5
–	1.5	1.5	1.5	1.5	

(a) Draw the alignment grid for the above alignment, clearly showing the corresponding alignment path, with each edge labelled with its cost. [1.5]



(b) Obtain the total cost of the above alignment. [0.5]

$$2 + 1.5 + 1.5 = 5$$