NOTE TO COMP 462/561 – Fall 2016 students: These are questions taken from the 2014 final exam. It does not include all questions on that exam, because some of those questions covered material we did not cover this year. Consequently, you should expect that your exam will be a bit longer than what you have here.

Question 1. (16 points, 4 points each)

Indicate whether the following statements are true or false. **Give a two line explanation for each**. *Credits will be given only if the justification is clear and correct.*

a) True or False? *Justify.* Indels that occur in the coding region of a gene will always cause a frame shift.

False. Indels that are of size that is a multiple of 3 will not cause a frame shift.

b) True or False? *Justify.* An RNA-seq experiment can be used to measure the abundance of each of the 20,000 human proteins in a given sample.

False. An RNA-seq experiments measures the abundance of mRNA transcripts, not of the proteins they encode. The amount of protein copies present at time t in a cell is a function of the transcriptional level of the gene in the past, the splicing and translational efficiency, and the protein's and mRNA's degradation (among others).

c) Give *two* reasons why the secondary structure inferred for a given RNA sequence using the Nussinov algorithm may not always reflect the true structure this sequence adopts in cells.

1) The energy model used by the algorithm is very coarse; it does not capture all the fine details of the structure, e.g. pseudoknots, non-canonical interactions, etc.
2) The structure of an RNA sequences is dynamic. At a given time t, it may not be folded in its most stable structure.

d) (4 points) Gene expression can be assessed using a RNA-seq experiment. In that case, the expression level of gene *g* is obtained using the FPKM measure (Fragment Per KiloBase Per Million reads).

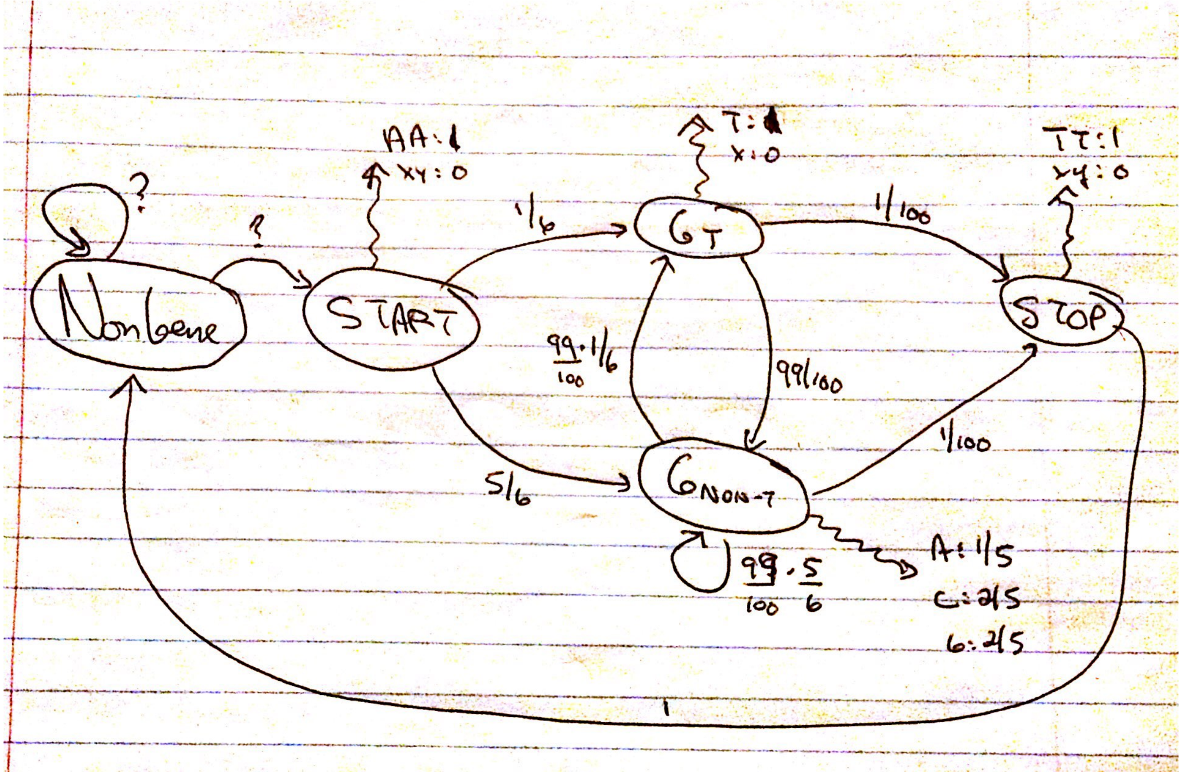
Explain the two types of normalization this entails, and why they are necessary.
"per KiloBase":
Normalizes for the length of the gene. Because the number of reads observed for a given gene is proportional to the length of gene, one needs to factor this out in order to be able to compare genes of different lengths.
"per Million Reads":
Normalizes for the amount of sequencing done. Allows comparing two datasets for which the coverage were different.

Question 2. (16 points)

Suppose that the Philae mission to comet 67P was equipped with a DNA sequencer. It finds DNA on the comet and researchers determine that on that comet, genes have a highly simplified structure:

- Proteins are made of only 4 types of amino acids. Each amino acid is encoded by a single nucleotide.
- Genes always start with ‘ A A ’
- Genes always end with ‘ T T ’
- The body of genes is made of any number of ‘A’, ‘C’, or ‘G’, ‘T’, but never contain two consecutive T’s, as those would be interpreted as a STOP signal. In genes, C’s and G’s are each twice as frequent as A’s and T’s.
- Genes contain no introns
- Genes are on average 100 bp long.

a) (9 points) Draw an HMM that could be used to make gene predictions in this type of DNA. Include all the transition and emission probabilities. If you think that there’s some information you are missing in order to choose some of these probabilities, mark them as “?”.



b) (4 points) Indicate what information you would need to be able to choose the value of probabilities you’ve marked as “?”.

We would need to know the average length of intergenic regions.

c) (3 points) If no one is able to give you the information you specified in (b), name the algorithm that you could use to estimate them?

Baum-Welch algorithm, or Viterbi training.

Question 3. (15 points)

When using Sanger sequencing to read a short piece of DNA, the probability of sequencing error, i.e. calling nucleotide x whereas the correct nucleotide is y, increases as a function of the position within the read. Suppose that a read has length 1000 and that Pr[error at position p] = p / 1000. Assume that insertions (i.e. the inclusion in a read of a nucleotide that was not present in the DNA sequence) never happen. Deletions (the omission of a nucleotide) are possible but very rare, so we will assume that a read contains at most one deletion of one nucleotide.

a) (10 points) Give a modified version of the Needleman-Wunsch algorithm that could be used to align two reads, under the assumption given below. Pay particular attention to the substitution and indel scoring schemes.

Suppose we are aligning reads S and T, and we are given a substitution matrix M and gap penalty c.

We get the following recurrence:

$$\begin{aligned} X(i,j) = \max\ & X(i-1,j-1) + M'(S[i],i, T[j],j) \\ & X(i-1,j) + c \\ & X(i,j-1) + c \end{aligned}$$

where $M'(S[i],i, T[j],j)$ is the expected score of the match between the unknown nucleotide at position i in S and that at position j in T. That is:

$$\begin{aligned} M'(a, i, b, j) = & M(a,b)*(1-i/1000)*(1-j/1000) + \\ & Sum_{b'\neq b}\, M(a,b')*(1-i/1000)*(j/1000)+ \\ & Sum_{a'\neq a}\, M(a',b)*(i/1000)*(1-j/1000) \\ & Sum_{a'\neq a}\, Sum_{b'\neq b}\, M(a',b')*(i/1000)*(j/1000) \end{aligned}$$

We initialize $X(0,0) = 0$, $X(1,0)=c$, $X(0,1)=c$. We then calculate only entries of the form $X(i-1,i)$, $X(i,i)$, and $X(i+1,i)$, because only those correspond to scenarios with at most one deletion per read.

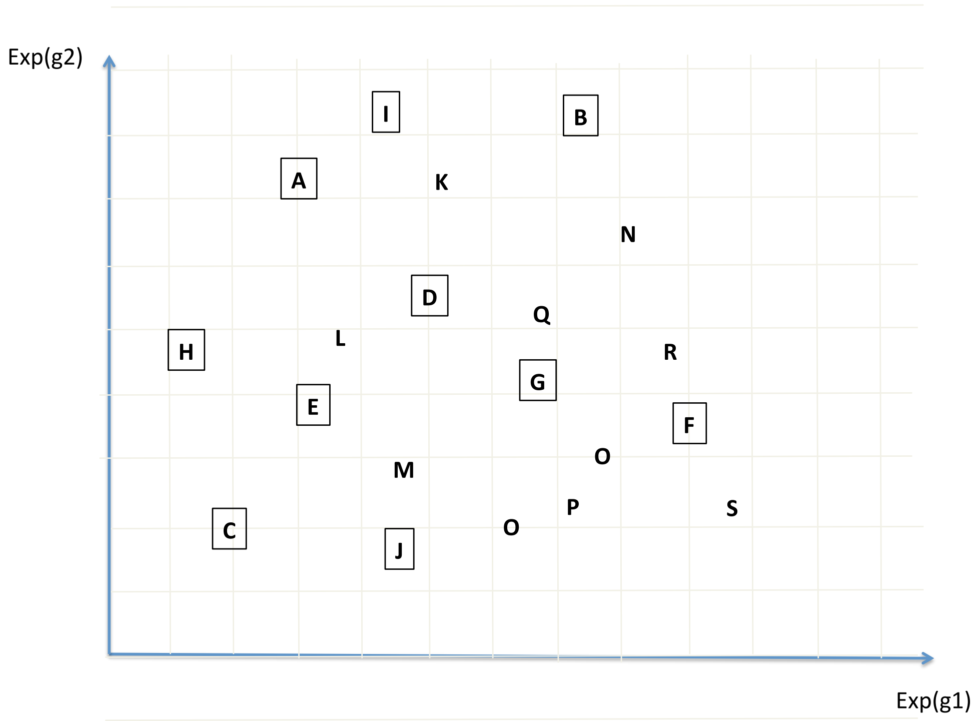
Thus the algorithm runs in time $O(n)$, where n is the length of the reads.

b) (5 points) Suppose that you are interested in reading as accurately as possible the sequence of a portion of a genome and that you have multiple reads originate from that region. Describe informally in 4-5 lines how could you combine the information contained in the reads in order to obtain the most accurate prediction about the exact sequence of that genomic region?

Suppose you have n reads covering a particular position of the genome we want to sequence (assuming homozygosity). The simplest approach would be to infer the majority base (i.e. the base observed the most often in the reads aligned to this position) as the most likely base. However variable error rates should also be factored in. One way would be to give more weight to bases that are close to the beginning of their read. A better way would be to use Bayes theorem to calculate Pr[true nucleotide | R1, R2, .. Rn], factoring in position-dependent error rates.

Question 5. (16 points)

Consider the following result of a microarray experiment, where in each case we measured the expression of only two genes, g1 and g2. Assume that samples A, B, …, J (shown with boxes in the figure below) come from patients with a specific disease, while samples K, L, …, S (shown without boxes) come from healthy patients.



a) (5 points) Draw the linear classifier that obtains the smallest number of classification errors (False-positives + False-negatives) on this data set.

*I can't easily draw it now, but the line separates points **I, B, A**, K**, D, H**, L, **E, C** from points N, Q, R, **G, F**, M, O, P, O', S, **J.***

Note: The sample O is present in the figure twice by mistake. Assume it is O and O'.

b) (3 points) What is the sensitivity of your classifier on this training data?

7/10.

c) (3 points) What is the specificity of your classifier on this training data?

8/10

d) (5 points) If one wants a sensitivity of at least 90%, where should the boundary of the linear classifier be moved in order to maintain a specificity that remains as high as possible? Draw you answer on the figure above, using a dashed line.

The classifier would separate R, O, O', P, F, S from the rest.

Question 7. (15 Points)

You are given two RNA sequences, $S = s_1 \ldots s_m$ and $T = t_1 \ldots t_n$, that you want to align. Sequence S has a known nested secondary structure (without pseudoknots), given to you in the form of a list of pairs of positions in S that form base pairs:
Struct = $\{ (l_1,r_1), (l_2,r_2), \ldots, (l_k,r_k) \}$, where $1 \leq l_i < r_i \leq m$ for all i ε {1…k}.
The secondary structure of sequence T is not known but is believed to be related to that of S.

Give an alignment algorithm to align sequences S and T using a given substitution matrix M and linear gap penalty c, subject to the following constraint:
If nucleotides s_i and s_j form a base pair in S (i.e. (i, j) ε Struct), and t_a is aligned to s_i and t_b is aligned to s_j, then t_a and t_b must be complementary nucleotides (i.e. A-U, C-G, G-C, or U-A pairs).

This question turned out to be harder than expected. Any solid attempt at it was marked generously. Here is my solution:

Here, we have to use a combination of Needleman-Wunsch and Nussinov algorithms.
Let X(i,j,k,l) be the score of the optimal structure-preserving alignment for si…sj against tk…tl.
We get

$$\begin{array}{ll} X(i,j,k,l) = \max \{ X(i+1,j-1,k+1,l-1) + M(Si,Tk) + M(Sj,Tl) & \text{if: (1) (i,j) not in Struct} \\ & \text{Or} \\ & \text{(2) Tk and Tl are complemenary} \\ \quad X(i+1,j,k,l) + c & \\ \quad X(i,j,k+1,l) + c & \\ \quad X(i,j-1,k,l) + c & \\ \quad X(i,j,k,l-1) + c & \\ & \\ \quad \mathrm{Max}_{i < i' < j} \, \mathrm{Max}_{k < k' < l} \{ X(i,i',k,k') + X(i'+1,j, k'+1,k) \} & \\ \quad \} & \end{array}$$

Page left blank intentionally. Use it if you need extra space.

Page left blank intentionally. Use it if you need extra space.