Project

Please choose ONE project among the given five projects. The last three projects are programming projects. Choose any programming language you want. Note that you can also write programs for the first two projects also. The deadline for emailing me your project is Friday, July 29, 2016, at 11:59pm PST. Please email to sami.khuri@sjsu.edu your project as an attachment with subject: XXXX_Project.pdf where XXXX is your family name.

I - Position Weight Matrix

You are going to build a position weight matrix (PWM) for the 5' splice site (also known as donor splice site) and use it to score 9-mers. You are going to follow the steps of HandsOn_14 (HO_14), part A, and HandsOn_15, part A. Type your answers in this document and expand existing tables by adding rows. You can also submit excel spreadsheets.
You are going to build the PWM by considering the splice sites of a gene of your choice, as long as:

1) the gene has at least 8 exons, and
2) no one else has chosen that gene.

Hint: you can start your gene search at http://www.ncbi.nlm.nih.gov/gene/?term=cancer.
By clicking on the gene, you will see right away how many exons it has.

This problem has 3 parts:

A) Introduction of My Favorite Gene (MFG)
B) Building the PWM for YFG – Similar to Hands-On 14
C) Scoring sequences using the PWM of part B – Similar to Hands-On 15

A) Introduction to MFG

Start with the following information:
i)
Name of MFG: _____.
Accession Number: _____.
Number of exons: _____.
Size of 5' UTR: _____.
Size of 3' UTR: _____.
Write a paragraph explaining why you chose this gene and what is the function of the gene. What does the protein(s) MFG produce(s) do?

ii) Parse the sequence of MFG in exactly the same way as we did for HBB in Hands-On, and hand in a document entitled "MFG sequence" similar to "HBB sequence" of Hands-On 8.

B) Building the PWM

Read the instructions of HO_14.
1) Determine all the 9-mers and list them here:

$X_1 =$

$X_2 =$

$X_3 =$

X_4 =

X_5 =

X_6 =

X_7 =

X_8 =
…. Add as many lines as needed.

2) Copy & paste the 9-mers in http://weblogo.berkeley.edu/logo.cgi to create a logo. What can be said when comparing the logo you obtained to the following logo we studied in the lecture notes?

5′ splice site

Answer:

3) Fill in Table 1 that lists all 9-mers representing the 5'splice sites of _____.

Table 1: The 9-mers representing the 5' splice sites of _____.

	1	2	3	4	5	6	7	8	9
X_1									
X_2									
X_3									
X_4									
X_5									
X_6									
X_7									
X_8									
…									

4) Use Table 1 to fill Table 2 which represents the probability distribution of each base in each of the 9 positions. Note that this is the Position Weight Matrix representing the 9-mers.

Table 2: PWM of the 9-mers of the 5'splice sites of _____.

	1	2	3	4	5	6	7	8	9
A									
C									
G									
T									

5) Use Table 2 and Laplace rule for pseudocounts to build Table 3.

Table 3: PWM with pseudocounts using Laplace's rule.

	1	2	3	4	5	6	7	8	9
A									
C									
G									
T									

6) Use Table 3 and the fact that the genome-wide average G and C content is 44% to fill Table 4 which represents the log-odd scores of the 9-mers of the 5' splice sites of _____. Use log base 2.

Table 4: Log-odds of the PWM of the 9-mers from Table 3 where base = 2

	1	2	3	4	5	6	7	8	9
A									
C									
G									
T									

C) Scoring Sequences using the PWM

Read the instructions of HO_15.

In this problem we are going to use the PWM we built in part A to score sequences and to determine the cutoff value (threshold) for the PWM for the 5' splice site (donor splice site). In essence, you are going to answer all the questions of HO_15, part A.

1) Use Table 1 to score all the 9-mers for the 5' splice sites. Fill in the values of Table 5.

Table 5: Scores of the 5' splice sites of _____ with PWM of Table 1

Sequence	Score
X_1 =	
X_2 =	
X_3 =	
X_4 =	
X_5 =	
X_6 =	
X_7 =	
X_8 =	
...	

2) Recall that the gene you chose has _____ number of exons. You are going to randomly choose that many 9-mers from the sequence that are not 5' splice sites (but have the invariant GT in positions 4 and 5) and score them.

Table 6: Scores of randomly chosen 9-mers with GT (positions 4 and 5) with PWM of Table 1

Random Sequence	Start	End	Region	Score
.....				

3) Check out all the scores you obtained in Tables 5 and 6 and decide on a good cutoff value that can be used as threshold. Threshold Value: _____.

4) i) Number of True Positives? _____. Explain.

ii) Number of False Positives? _____. Explain.

iii) Number of True Negatives? _____. Explain.

iv) Number of False Negative? _____. Explain.

II – Hidden Markov Models

Problem 1

We have three different coins, one fair and two biased. We build a hidden Markov model, λ, with the following parameters:
Three states: X, Y, and Z. Alphabet = {H,T}.

The transition probability matrix is A, and π gives the initial probabilities:

$$A = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \qquad \pi = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

The probabilities of emission are:

$b_X(H) = b_X(T) = 1/2$ $b_Y(H) = 3/4$ $b_Y(T) = 1/4$ $b_Z(H) = 1/4$ $b_Z(T) = 3/4$

Use the Viterbi algorithm to find the sequence Q that most likely generated HHHHTHTTTT.

Problem 2

A recent study focused on the relationship between birth weights of California women and the birth weights of their daughters. The weights were split into three categories: low (below 6 pounds), average (between 6 and 8), and high (above 8 pounds).

Among women whose own birth weights were low:

- 50 percent of the daughters had low birth weights,
- 45 percent had average weights, and
- 5 percent had high weights.

Women whose own birth weights were average had:

- Daughters with average weights half of the time, while
- The other half was split evenly between low and high categories.

Women whose own birth weights were high had female babies with:

- High weights 40 percent of the time,
- Low and average weights each occurring 30 percent of the time.

Suppose that the initial generation of mothers surveyed contained 25 percent low birth women, 60 percent average weight, and 15 percent high weight.

1) Use L for low, V for average, and H for high.
   a) Write out the transition probabilities of this Markov Model representing California female birth weights.
   b) Write out the transition probability matrix: A, of this Markov Model where the rows and columns are in the following order: L, V, and H.
   c) Write out the initial probability matrix: Π, of this Markov Model in the following order: L, V, and H.

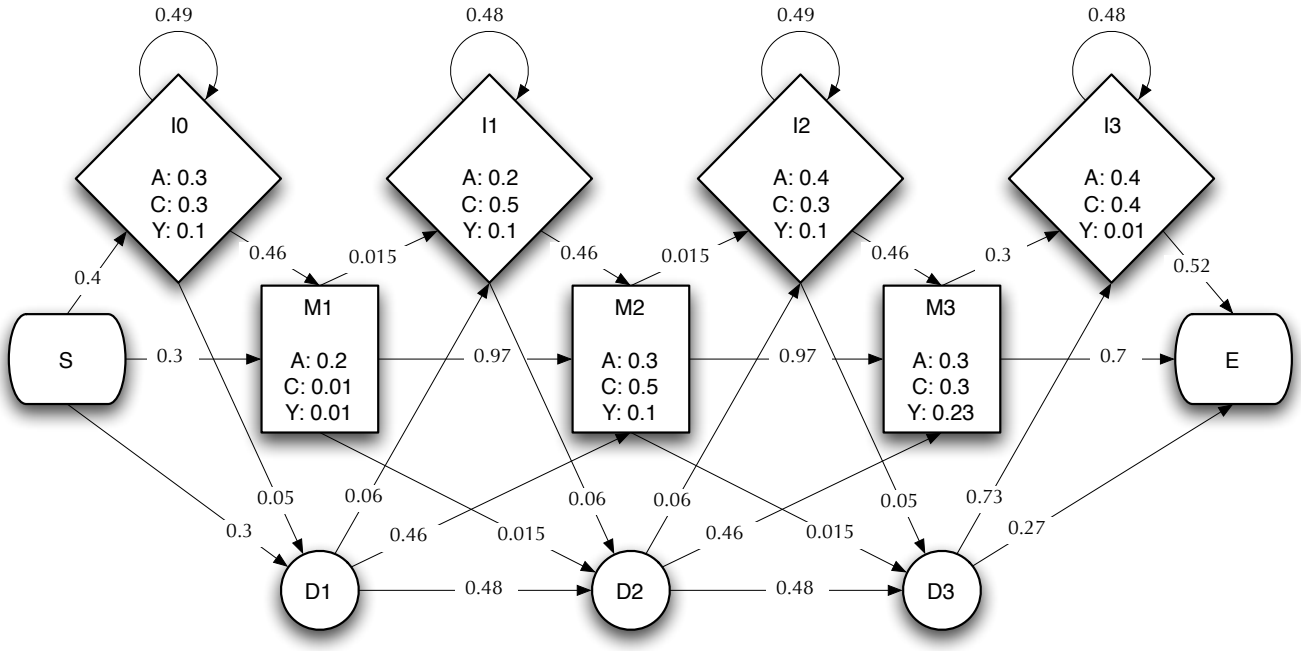
2) Draw the state diagram of the California female birth weights.

Extra credit problems

3) Find the probability that a woman whose birth weight was average has a granddaughter with an average birth weight.
4) What would the distribution look like for the generation of the great-great-granddaughters?
5) Find the distribution of birth weights after one generation if the initial probability distribution is (0.4, 0.3, 0.3).
6) Suppose the distribution of birth weights of a generation of daughters is (0.31, 0.45, 0.24). Can you find the distribution of birth weights of the mothers?

Problem 3

Consider the following profile HMM that can be found in Rachel Karchin introductory article on "Hidden Markov Models and Protein Sequence Analysis" at: http://www.cse.ucsc.edu/research/compbio/ismb99.handouts/KK185FP.html
As the caption of the article mentions, the figure highlights one of several possible paths for obtaining ACCY.
The insertion states are labeled from left to right by: I0, I1, I2, and I3. The matching states are labeled from left to right by: M1, M2, and M3.



The figure gives all the transition probabilities but not all the probabilities of emissions. Assume that the probabilities of emissions of the three amino acids, A, C, and Y are given by the following table:

	M1	M2	M3	I0	I1	I2	I3
A	0.2	0.3	0.3	0.3	0.2	0.4	0.4
C	0.01	0.5	0.3	0.3	0.5	0.3	0.4
Y	0.01	0.1	0.23	0.1	0.1	0.1	0.01

Note that each column does not add up to one since each state can emit more than the 3 amino acids shown in the table. Each state can emit any of the 20 amino acids.

Use the Viterbi algorithm to find the most likely path through the model that can produce ACCY. You may want to use the following table where the deletion states are not shown since they do not emit any amino acids.

	M1	M2	M3	I0	I1	I2	I3
A							
C							
C							
Y							
End							

III - Phylogenetic Trees

The reconstructing of phylogenetic trees is a general problem in biology. As seen in class, it is used in molecular biology to help understand the evolutionary relationships among proteins, for example.

This project consists in choosing one of the four algorithms mentioned below, choosing the appropriate referenced article(s), reading, understanding and implementing it as described in the article and comparing it to an existing package.

- Phylogenetic Trees Based on Pairwise Distances [FD96]
- Phylogenetic Trees Based on Neighbor Joining [SN87]
- Phylogenetic Trees Based on Maximum Parsimony [Fel96]
- Phylogenetic Trees Based on Maximum Likelihood Estimation [BT86], [Fel81].

[BT86] Bishop, M. and Thompson, E. Maximum likelihood alignment of DNA sequences. Journal of Molecular Biology. 190:159-165; 1986.
[FD96] Feng, D. and Doolittle, R. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. Methods Enzymol. 266; 1996.
[Fel81] Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 17:368-376; 1981.
[Fel96] Felsenstein, J. Inferring phylogeny from protein sequences by parsimony, distance and likelihood methods. Methods Enzymol. 266; 1996.
[SN87] Saitou, N. and Nei, M. The neighbor joining method: a new method for reconstructing phylogenetic trees. Molecular Biology Evolution; 4:406-425; 1987.

IV - Gene Prediction

Gene prediction consists in identifying regions of genomic DNA that encode proteins.

Some of the existing models that identify and distinguish coding regions from non-coding regions are based on:

- Hidden Markov Model,
- Neural Network,
- Probabilistic model,
- Linear discrimination analysis,
- Decision tree classification,
- Quadratic discriminant analysis,
- Stochastic context free grammars.

This project consists in choosing one of the above techniques and implementing the prediction (search) algorithm, which will be able to search a given database for genes that do code for proteins.

Your algorithm should be compared to an existing package.

V – Profile HMM

This project consists in choosing at least 20 proteins (of at least 100 amino acids each) that belong to the same family (orthologs) and building a profile HMM. Compare your profile HMM to existing packages.

Critique your profile HMM by giving its strong and weak characteristics.