DUE: Monday Apr 27 at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers. Your work must be legible -- if your handwriting isn't great, type it up and print it.

For full credit you must identify key assumptions and provide reasoning (or show work) behind answers. Whenever possible, partial credit will be given if adequate work is shown. Remember I encourage working together, but you MUST indicate all collaborations and/or assistance received or given.

Note that showing work means that if you utilize software for assistance (programs you write or stock software such as Excel), you should indicate as such and provide sufficient details so that I can judge the work. That may mean sending me (by email) source code or associated Excel files.

Questions 1-5 are required of all sections. Questions 6 (marked Advanced) is required for the graduate sections (5520 and 7000). Those in the undergraduate section may do the advanced questions for extra credit. Question 7 is extra credit for **ALL** sections.

Question #1 is a double question: worth 20 pt, all other problems have a maximum value of 10 points. Sub-problem values are marked when appropriate.

Questions

1. (20 pt: ** NOTE this question is worth twice normal value**) During an analysis of a promoter you identify six sites (shown below) that alter the expression of the promoter when they are deleted.
ACGGAG
ACGTGG
AGGAAG
AGGCAC
ACGCAC
AGGGAC

(a) (3pt) Since the number of sites is small, we will include pseudocounts distributed according to the background nucleotide frequencies in the genome. Assume you are working in a genome where %A=%T=20%, and %G=%C=30%. The sum of all pseudocounts per position should be 1. What are the pseudo-count values for each nucleotide?

(b) (4pt) Fill in the nucleotide count matrix, N(b,i) for this multiple alignment, including pseudocounts.

Count matrix (including pseudocounts):

	1	2	3	4	5	6
A						
C						

G						
T						

(c) (5pt) Now convert the above counts matrix into a probability matrix, recalling that

$$\mathrm{P(b,i)} = \mathrm{N(b,i)}/\sum_{\boldsymbol{k}=\mathbf{1}}^{\mathbf{4}} \boldsymbol{N}(\boldsymbol{k},\boldsymbol{i})$$

Probability Matrix:

	1	2	3	4	5	6
A						
C						
G						
T						

(d) (5pt) Now convert the above probabilitiy matrix into a scoring matrix, recalling that S(b,i)=log[P(b,i)/P(b)], where P(b) is defined by the genome's nucleotide frequencies.

Scoring Matrix

	1	2	3	4	5	6
A						
C						
G						
T						

(e) (3pt) Consider the following two new sequences:

Sequence 1: TCGGAG
Sequence 2: ACTGAG

Based on the scoring scheme determined in part D, which of these two sequences is a better fit to this motif model?

2. (10pt) You notice that only 60% of the peaks detected by ChIP-Seq have the known transcription factor motif nearby.

(a) (5pt) Give at least two explanations for the remaining 40%.

(b) (5pt) How would you test (experimentally or computationally) for the explanations you suggested in part A?

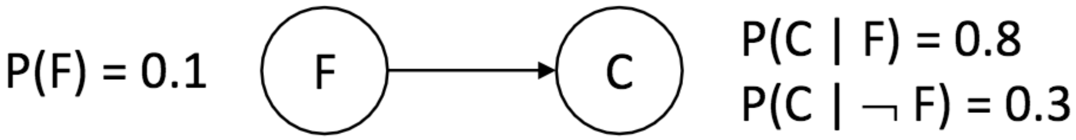
3. (10pt) Your colleague professor Stu Dent generated a genome-wide DNA methylation map for normal colon cells using MRE-seq (a restriction enzyme approach) and MeDIP-seq (an immunoprecipitation approach). In an intergenic region, he found an interesting locus. This locus is about 20kb. On one end of the locus, there is a 2kb CpG rich stretch that has both relatively low MRE-seq and MeDIP-seq signals. The rest 18kb has high level of MeDIP-seq signals.

(A) (3pt) Why might you suspect that this region encodes for a novel gene?

(B) (4pt) You decide to look at histone modification patterns across this region for more evidence. There are several genome-wide datasets available for this cell type: H3K4me1, H3K4me3, H3K27me3, H3K9me3, H3K36me3, and H3K9Ac. Which histone mark would you investigate for this locus and why?

(C) (3pt) Suggest at least one other source of data (NOT annotation or bisulfide sequencing) that may help you confirm or refute your suspicion, and why you think it may help?

4. (10pt) Consider the following simple Bayesian network diagram:



Where F is “having the flu” and C is “coughing” and both are binary (yes/no) variables.

(a) (3 pt)Given the diagram, what is the probability of both having the flu and a cough e.g. P(F,C)?

(b) (3 pt) What is the probability of having a cough e.g. P(C) from the diagram?

(c) (4 pt)Draw a Bayesian network to encode the following statement. Be sure to state all assumptions included in your diagram.

“A smell of sulphur (S) can be caused either by rotten eggs (E) or as a sign of the doom brought by the Mayan Apocalypse (M). The Mayan Apocalypse also causes the oceans to boil (B).”

5. Read “The what, where, how and why of gene ontology -- a primer for bioinformaticians” (PDF is available on Canvas as GOprimer.pdf) and answer the following questions:

a) (1pt) What are the three ontologies in GO?

b) (1pt) The relationships within GO form what kind of graph?

c) (1pt) What is the main difference between the full and filtered GO?

d) (1pt) Who does most of the current GO annotation?

e) (1pt) According to Figure 5, what is the most common type of evidence for information within GO?

f) (2pt) Give one pro and one con of the information content measure of GO terms.

g) (3pt) List 3 major criticism of GO.

6. (Advanced: 10pt) There are two competing methods for chromatin state annotation, Segway (Hoffman et. al. Nat. Methods 2012) and ChromHMM (Ernst & Kellis Nat. Biotech. 2010). Interestingly, the two groups came together to publish a paper (on Canvas) that compares the two approaches (Hoffman et. al. NAR 2012). Read the NAR paper and answer the following questions:

A. (3pt) Describe the difference between supervised and unsupervised methods.

B. (2pt) According to Figure 3, in what state do most phenotype-associated SNPs reside?

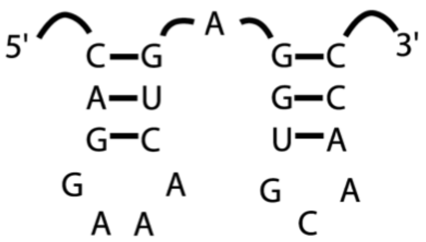
C. (2pt) The authors avoid (quite emphatically) declaring either method as superior. Based on the results presented, which method is better and why?

D. (3pt) What are the inherent tradeoffs in the number of states? [Here they use 25 states, but the original ChromHMM paper used 51 and in other papers they use 12, 16, 19, and 21.]

7. (Extra Credit) Consider the following (insanely) simple grammar for RNA secondary structure prediction:

$$S \rightarrow aSa'S \mid aS \mid \varepsilon$$

The first rule ($S \rightarrow aSa'S$) captures both bifurcation (splitting for multiple stem structures) and base pairing! Draw a parse tree indicating how this structure:



would be generated from this simple (but admittedly a bit odd) grammar. Recall: the parse tree must begin with S, each non-terminal (in this case “S”) is replaced at each step of tree with a single rule from the grammar, and all branches must end with ε. Hint: We will discuss secondary structure of RNAs starting on April 23rd.