DUE: Monday Apr 11 at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers, show all work (as necessary), identify key assumptions, and indicate all collaborations and 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 or R), 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.

Your work must be legible -- if your handwriting isn't great, type it up and print it.

Questions

1. Consider a new algorithm for predicting whether a particular RNA binding protein binds to an exon. 10,000 exons are evaluated by the prediction method and a cutoff of 2 was selected. Everything scoring above a 2 was considered positive for the RNA binding protein whereas below this level was classified as negative. These results were then compared to a gold standard method of determining whether the RNA binding protein associates with the exon. The results are shown in the following table:

	"Gold Standard" Outcome		
Prediction Method	Positive	Negative	Total
Positive	125	25	150
Negative	375	9475	9850
Total	500	9500	10,000

Calculate:
a) Sensitivity
b) Specificity
c) Positive predictive value

2. Your colleague, professor Stu Dent, generated a genome-wide DNA methylation map for normal colon cells using MRE-seq and MeDIP-seq. 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 intermediate MRE-seq and MeDIP-seq signals. The rest 18kb has high level of MeDIP-seq signals.

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

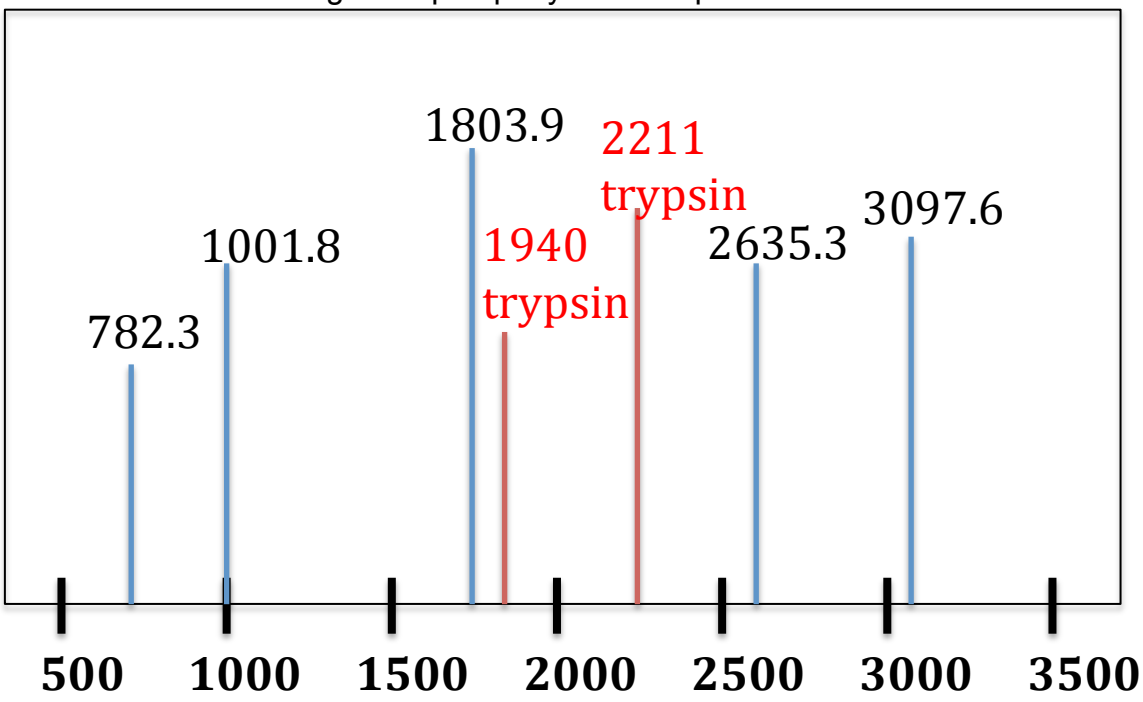
(B) You decide to look at histone modification patterns across this region for more evidence. There are several genome-wide datasets available for this cell type: H3K4me1, H3K4me3, H3K27me3, H3K9me3, H3K36me3, and H3K9Ac. Which histone mark would you investigate for this locus and why?

(C) Suggest at least one other source of data that may help you, and why you think it may help?

3. a) Explain how these two comparisons are the similar.
   I. GRO-seq vs. RNA-seq
   II. Ribo-seq vs. Proteomics

b) Describe at least one Ribo-seq scoring method (FLOSS, ORFscore) and how that method tells you that you have a real ribosome protected fragment. Extra credit: Explain both methods and show some make believe example math for both scouring methods (Show math for both a gene that you believe is ribosome enriched and a gene you don't believe).

4. Consider the following example query MALDI spectrum:



It belongs to one of these three proteins:
Protein A>
MQNSANHGRGFAMWEVPPRRKLRKGCPVWESTLDVVNSLSDRIRQACGCA

Protein B>
MQLHQVFPRSISLARNVCPNPKTDRLSGNTIMMREPVWLTNSWKGLTLIR

Protein C>
MLNLYPAGEVAPLPPQTAIPPSMRGVLHKPLVSWRHPTRNEIAKSIEFMR

You fragment with a trypsin which we will assume cuts after every K. (Which is not 100% accurate but we will assume for this homework.)

A. Build a "database" of all the masses of expected fragments from the three proteins above using the masses below. Use this website: http://db.systemsbiology.net:8080/proteomicsToolkit/FragIonServlet.html to calculate the fragment sizes using the **Mono (M+H)$^+$** mass.

Name of fragment	Fragment sequence	Mass Mono (M+H)$^+$

B. Which protein does the MALDI spectrum (above) represent?

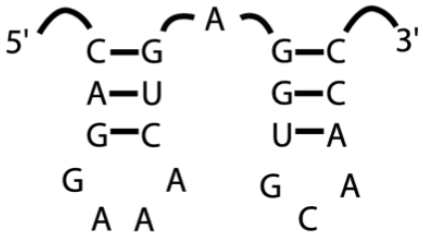
C. What protein does the 1001.8 fragment come from? How did this fragment affect your answer to 3B?

5. Consider the following (insanely) simple grammar:

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

The first rule ($S \rightarrow aSa'S$) captures both bifurcation (splitting for multiple stem structures) and base pairing. The second rule ($S \rightarrow aS$) is for single stranded regions. The last rule $S \rightarrow \varepsilon$ ends the description of this part of the structure. Recall that at each step you replace a single Nonterminal (the capital S) by the application of a single rule. Also recall that each 'a' is replaced by a character in your string and the aSa' nomenclature assumes two nucleotides that are base paired.

Draw a parse tree indicating how this structure:



would be generated from this simple (but admittedly a bit odd) grammar.

Hint #1: Slide #16 of the Mar30 lecture (bottom of page 8) shows an example

parse tree (in blue and black, right side of slide) using a different grammar (on the slide) for this same structure!

6. 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 D2L) that compares the two approaches (Hoffman et. al. NAR 2012). Read the paper and answer the following questions:

A. Describe the difference between supervised and unsupervised methods.

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

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

D. 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.]