DUE: Friday Jan 29th 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), 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. Given the following codon usage table (as percentages) from H. influenza:

Ala (A) 7.81	Gln (Q) 3.94	Leu (L) 9.62	Ser (S) 6.88
Arg (R) 5.32	Glu (E) 6.60	Lys (K) 5.93	Thr (T) 5.45
Asn (N) 4.20	Gly (G) 6.93	Met (M) 2.37	Trp (W) 1.15
Asp (D) 5.30	His (H) 2.28	Phe (F) 4.01	Tyr (Y) 3.07
Cys (C) 1.56	Ile (I) 5.91	Pro (P) 4.84	Val (V) 6.71

Calculate the following:

(a) P(s = “CRICK”)
(b) P(s = “WATSON”)
(c) P(s = “charged” or “aromatic”) [For definition of charged and aromatic, see slide #2 of Jan20ORFfinding lecture.]

2. We observe the following empirical frequencies for 2-mers in H. influenza:

	A	C	G	T
A	0.1202	0.0505	0.0483	0.0912
C	0.0665	0.0372	0.0396	0.0484
G	0.0514	0.0522	0.0363	0.0499
T	0.0721	0.0518	0.0656	0.1189

Where the first nucleotide s(i) is the row and the second nucleotide s(i+1) is given in the columns, hence the frequency of AC is 0.0505. Convert the above frequency matrix into a transition matrix for the Markov model of di-nucleotide sequences discussed in class. Note that each entry of the matrix is the conditional probability: P(s(i+1)| s(i)).

3. Consider a random i.i.d. model of nucleotide sequence where the GC content is 35%. Assume that the frequency of G = C and A = T. What are the expected probabilities of each of the 20 amino acids?

4. Given the i.i.d model described in Q#3, what is the p-value associated with an open reading frame prediction of 50 amino acids in length?

5. Consider Ravenhall et. al. Inferring Horizontal Gene Transfer. PLoS Comp Biol 11(5): e1004095 (2015). doi:10.1371/journal.pcbi.1004095
(This article is available on D2L as RavenhallPLoS2015.pdf).

(a) Briefly describe the difference between parametric and phylogenetic approaches to detecting horizontal gene transfer.

(b) What are the pros and cons of the parametric approaches?

6. (Advanced) Due to redundancy in the genetic code, a sequence of amino acids could be encoded by several DNA sequences. For a ten amino acid long protein fragment, what is the lower and upper bound for the number of possible DNA sequences that can encode this protein sequence?

7. (Advanced) Describe a method for finding, within a collection of protein sequences, the longest English language word. The English word may be a subsequence within any protein sequence in the set. Identify the assumptions of your method.