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.

All problems have a maximum value of 10 points. Subproblem values are marked when appropriate. We did not discuss rounding conventions in class, so I did not take off for what I detected as rounding errors.

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:

3pts (a) P(s = “CRICK”)
= .0156 * .0532 * .0591 * .0156 * .0593 = 4.5x10^{-8}
3pts (b) P(s = “WATSON”)
= the letter “O” is an invalid character in the amino acid alphabet, therefore it is not in the “world” of possibilities and has (by definition) a probability of zero. = 0

4pts (c) P(s = “charged” or “aromatic”) [For definition of charged and aromatic, see slide #2 of Jan20ORFfinding lecture.]
=> charged = D, E, R, K, or H =
.0530 + .0660 + .0532 + .0593 + .0228 = 0.2543
aromatic = F, Y, W, or H =
0.0401 + .0307 + .0115 + .0228 = 0.1051
P(s = “charged” or “aromatic”) =
P(charged) + P(aromatic) – P(charged and aromatic) =
0.2543 + 0.1051 – .0228 = 0.3366

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)).

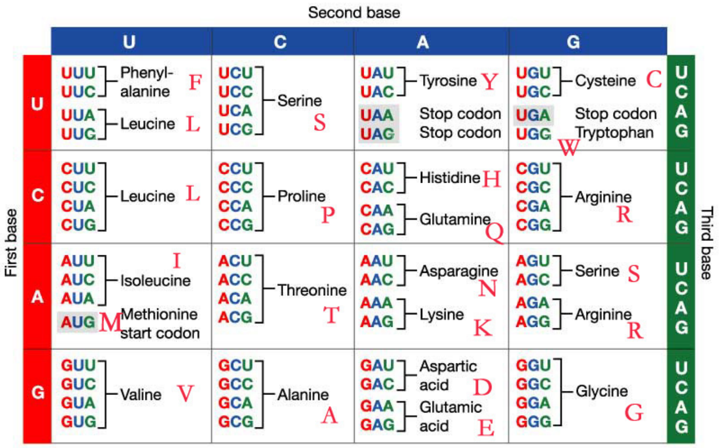
P(s(i+1)|s(i)) = P(s(i+1)*s(i))/P(s(i))
P(s(i) = A) = .1202+.0505+.0483+.0912 = .3102
P(s(i) = C) = .0665 + .0372 + .0396 + .0484 = .1917
P(s(i) = G) = .0514 + .0522 + .0363 + .0499 = .1898
P(s(i) = T) = .0721 + .0518 + .0656 + .1189 = .3084

	A	C	G	T
A	0.1202/.3102 = 0.3875	0.0505/.3102 = 0.1628	0.0483/.3102 = 0.1557	0.0912/.3102 = .2940
C	0.0665/.1917 = 0.3469	0.0372/.1917 =0.1941	0.0396/.1917 = 0.2066	0.0484/.1917 = 0.2525
G	0.0514/.1898 = 0.2708	0.0522/.1898 = 0.2750	0.0363/.1898 = 0.1913	0.0499/.1898 = 0.2629
T	0.0721/.3084 = 0.2338	0.0518/.3084 = 0.1680	0.0656/.3084 = 0.2127	0.1189/.3084 = .3855

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?

(1pt probabilities of individual nucleotides)

GC content = 35% therefore P(G) = P(C) = 0.35/2 = 0.175
AT content = 1-.35 = 0.65 => P(A) = P(T) = 0.325



Second base

First base	U	C	A	G	Third base
U	UUU UUC] Phenyl-alanine F UUA UUG] Leucine L	UCU UCC UCA UCG] Serine S	UAU UAC] Tyrosine Y UAA Stop codon UAG Stop codon	UGU UGC] Cysteine C UGA Stop codon UGG Tryptophan W	U C A G
C	CUU CUC CUA CUG] Leucine L	CCU CCC CCA CCG] Proline P	CAU CAC] Histidine H CAA CAG] Glutamine Q	CGU CGC CGA CGG] Arginine R	U C A G
A	AUU AUC AUA] Isoleucine I AUG M Methionine start codon	ACU ACC ACA ACG] Threonine T	AAU AAC] Asparagine N AAA AAG] Lysine K	AGU AGC] Serine S AGA AGG] Arginine R	U C A G
G	GUU GUC GUA GUG] Valine V	GCU GCC GCA GCG] Alanine A	GAU GAC] Aspartic acid D GAA GAG] Glutamic acid E	GGU GGC GGA GGG] Glycine G	U C A G

(7 pts) Basic principles: From this we can calculate the probability of each three letter code. Codons are either:

$0.175^3 = 0.005359$
$0.175^2 * 0.325 = 0.009953$
$0.175 * 0.325^2 = 0.01848$
$0.325^3 = 0.03433$
Some amino acids are from a single codon whereas others are from two, four or six codons. (g = c = .175; a = u = .325) The

Leu (L) = UUA + UUG + CUU + CUC + CUG + CUA = .03433 + .01848 + .01848 + .009953 + .009953 + .01848 = 0.109676

Ile (I) = AUU + AUC + AUA = 0.03433 + 0.01848 + 0.03433 = 0.08714

Ser (S) = UCU + UCC + UCA + UCG + AGU + AGC = .01848 + .009953 + .01848 + .009953 + .01848 + .009953 = 0.085299

Arg (R) = CGU + CGC + CGA + CGG + AGA + AGG = .009953 + .005359 + .009953 + .005359 + .01848 + .009953 = 0.059057
Thr (T) = ACU + ACC + ACA + ACG = .01848 + .009953 + .01848 + .009953 = 0.056866
Val (V) = GUU + GUC + GUA + GUG = 0.01848 + 0.009953 + 0.01848 + 0.009953 = 0.056866

Lys (K) = AAA + AAG = .03433 + .01848 = 0.05281
Asn (N) = AAU + AAC = 0.03433 + .01848 = 0.05281
Phe (F) = UUU + UUC = 0.03433 + 0.01848 = 0.05281
Tyr (Y) = UAU + UAC = 0.03433 + 0.01848 = 0.05281

Ala (A) = GCU + GCC + GCA + GCG = .009953 + .005359 + .009953 + .005359 = .030624

Gly (G) = GGU + GGC + GGA + GGG = .009953 + .005359 + .009953 + .005359 = .030624
Pro (P) = CCU + CCC + CCA + CCG = 0.009953 +0.005359 + 0.009953 + 0.005359 = .030624

Gln (Q) = CAA + CAG = .01848 + .009953 = 0.028433
Glu (E) = GAA + GAG = .01848 + .009953 = 0.028433
Asp (D) = GAU + GAC = 0.01848 + 0.009953 = 0.028433
His (H) = CAU + CAC = 0.01848 + 0.009953 = 0.028433
Cys (C) = UGU + UGC = 0.01848 + 0.009953 = 0.028433

Met (M) = AUG = .01848

Trp (W) = UGG = .0099553

(2pts: Normalization) There are three codons that do NOT encode amino acids. The sum total of all amino acid probabilities should sum to 1, and must therefore be normalized by the total sum of codons that encode an amino acid.
P(notstop) = 1-0.0713 = 0.9287

Leu (L) = 0.109676/0.9287 = 0.1181

Ile (I) = 0.08714/0.9287 = 0.09383

Ser (S) = 0.085299/0.9287 = 0.09185

Arg (R) = 0.059057/0.9287 = 0.06359
Thr (T) = 0.056866/0.9287 = 0.06123
Val (V) = 0.056866/0.9287 = 0.06123

Lys (K) = 0.05281/0.9287 = 0.05686
Asn (N) = 0.05281/0.9287 = 0.05686
Phe (F) = 0.05281/0.9287 = 0.05686
Tyr (Y) = 0.05281/0.9287 = 0.05686

Ala (A) =.030624/0.9287 = 0.03298
Gly (G) = 0.030624/0.9287 = 0.03298
Pro (P) = 0.030624/0.9287 = 0.03298

Gln (Q) = 0.028433/0.9287 = 0.03062
Glu (E) = 0.028433/0.9287 = 0.03062
Asp (D) = 0.028433/ 0.9287 = 0.03062
His (H) = 0.028433/0.9287 = 0.03062
Cys (C) = 0.028433/0.9287 = 0.03062

Met (M) = 0.01848 / 0.9287 = 0.01990

Trp (W) = 0.0099553 / 0.9287 = 0.01072

If all of your math is correct, the sum of the amino acid probabilities should be roughly 1 (only differing by rounding error).

I know this was tedious - but a few things to get out of this exercise. First, computers are useful for tedious calculations and coding up an answer here (in Excel or in your favorite language) would prevent simple errors. Second, the world of possible probabilities, in this case amino acids, must sum to 1. Finally, given this underlying model these are now my expected frequencies and therefore can be used to ask questions such as whether particular amino acids are overused than expected given their background frequencies.

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?

Note that unlike question #3 (which was amino acid probabilities), this question is concerned with the codon probabilities explicitly.

(2 pts Calculating Start/Stop probabilities)
P(AUG) = 0.325 * 0.325 * 0.175 = 0.0185
P(stop) = P(UAA) + P(UAG) + P(UGA) =
(0.325 * 0.325 * 0.325) +
(0.325 * 0.325 * 0.175) +
(0.325 * 0.175 * 0.325) = 0.0343 + .0185 + .0185 = .0713
P(notstop) = 1-0.0713 = 0.9287

(8pts) We talked about the precise probability of a particular open reading frame as: P(MET) * P(not stop)$^{k-1}$ * P(stop)
and that we could approximate this probability by:
P(not stop)^{k} * P(stop)
I will accept either method of calculating the probability of a particular open reading frame. Below I give the equation for the approximation method.

The p-value is the probability of getting a sequence of this length or more extreme by our null model. This would be calculated as:

$$\text{p-value} = \sum_{x=50}^{\infty}(0.9287)^x(0.0713) = 1 - \sum_{x-1}^{49}P(0.9287)^x(0.0713) = 1\text{-}0.9039 = 0.0961$$

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).

(5pts) (a) Briefly describe the difference between parametric and phylogenetic approaches to detecting horizontal gene transfer.

Parametric approaches rely only on the genome under study whereas phylogenetic methods are comparative, requiring many sequenced genomes. Parametric approaches assume uniformity in the characteristics of the genome and look for anomalies in this uniformity. Phylogenetic approaches either assume an underlying model or rely on a reference species tree and tend to be applied to genes as the underlying unit of interest.

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

PROs:

- Good when only one or a few genomes available.

CONs:

- Not accounting for the host's intragenomic variability will result in overpredictions.
- transferred segments need to exhibit the donor's signature and to be significantly different from the recipient's
- difference between the two (donor and recipient) tends to vanish over time

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?

(5pts) Lower bound is when the amino acid sequence is all MET which is encoded by a single codon, therefore only one possible sequence encodes for poly-MET:
AUGAUGAUGAUGAUGAUGAUGAUGAUGAUG

(5pts) The upper bound is when the peptide is encoded using only amino acids that have six possible codons, namely Leucine (L), Arginine (A), or Serine (S). In this case there can be as many as 6^{10}= 60,466,176 possible different nucleotide sequences that encode the 10 amino acid peptide 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.

There are many possible answers for this question. However, here is one:

Assume I have a dictionary of valid English language words. There are letters in the English language that cannot appear in an amino acid sequence, so remove all words containing any of these letters. Sort the remaining words by length. For each word, starting with the longest, search the protein sequence set for the occurrence of this word. There are many ways to do this search but what you want is a glocal search – meaning you want **all** of the English word to match **some** part of the protein sequence. Once a word match is found, you can stop as you have found the **longest** English language word in the protein sequence set.