DUE: Friday Feb 9th 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 and 7 (marked Advanced) are required for the graduate sections (5520 and 7000). All questions are weighted equally in the overall grade. Those in the undergraduate section may do the advanced questions for extra credit.

All problems have a maximum value of 10 points. Sub-problem values are marked when appropriate.

Questions

1. Given the following codon usage table (as percentages):

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) (3pts) P(s = “MENDEL”)
(b) (3pts) P(s = “ROSALIND”)
(c) (4pts) P(s = “charged” or “aromatic”) [For definition of charged and aromatic, see slide #12 of Jan 24 lecture.]

2. We observe the following empirical frequencies for dinucleotides:

	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 P(GC) = 0.0522. 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 transition matrix is the conditional probability: P(s(i+1)| s(i)).

3. The sum of all amino acids should be 1. The sum of all codons is 1. However, STOP is not a valid amino acid. Yet we should still be able to calculate the probability of any amino acid from the nt frequencies, by proper *normalization*. In this case, we normalize by the total probability that leads to codons. So: P(amino acid) = P(all codons for the amino acid) / P(all codons that are valid amino acids). Assume P(A) = 0.3, P(T) = 0.3, P(C) = 0.2, and P(G) = 0.2

(4pts) What is the P(all codons that are valid amino acids)?

Using proper normalization, what is the probability of the following amino acids?
(3pts) Ile (I)
(3pts) Trp (W)

4. Consider the following read returned from the sequencing facility:

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
GCATGTGGTGAGGTGGTAGTGATGGTGATATAGAGTGGTAGTATAAGTGT
+
IIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIAIIGIICI
```

Recommendation: Refer to the Wikipedia page on FASTQ format for the encoding schemes discussed below.

(a) (2pts) Assume the quality scores are encoded using the Sanger offset (Phred+33). Is this sequence of generally good quality?

(b) (5pts) Under this encoding, what base is the lowest quality? (You may circle it in the above) What is the probability of this position being correct?

(c) (3pts) You realize that you were mistaken in the encoding and it is given in the Illumina 1.3+ (Phred+64) format. Under this encoding scheme, is this sequence of generally good quality? Is the worst position still the one you circled in question b?

5. You begin to sequence the genome of Tamatoa, gathering 5,000 reads that were each 600 base pairs long. You have hypothesized that the Tamatoa genome is about 2 million bases long.

(a) (5pts) At what coverage have you sequenced the genome thusfar?

(b) (5pts) If the coverage of the Tamatoa genome were 6X, what is the probability that a base will be unsequenced?

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? (5 pts per bound)

7. (Advanced) 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 Canvas as RavenhallPLoS2015.pdf).

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

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