Agreement form for Bioinformatics take-home midterm

I understand that the take-home midterm in Bioinformatics is intended to test my own personal knowledge of and ability to apply concepts, methods, and factual information from the course. I therefore agree to do **ALL** of the following:

1. Complete all problems **on my own**, without discussing them with any other person (other than Drs. Weisstein and Beck)

2. Use **only the following resources**:
    - course PowerPoint slides and your own lecture notes
    - *Learning Python, 5th edition* (linked on the course Blackboard site)
    - any code and documentation that you or a member of your own team produced for an earlier assignment in this course
    - any other resource that has been approved *in writing* by Dr. Weisstein

3. **Report any evidence** that other students are violating or seeking to violate these terms.

I understand that failure to uphold these terms constitutes a serious violation of Truman's Academic Integrity policy and will be treated accordingly.

Signature: _______________________________________ Date: ______________

1a. Write and thoroughly document a Python program that (i) imports RNA sequences from FASTA format, (ii) reports any sequences that do not represent valid pre-mRNA transcripts, and (iii) for each of the remaining sequences, determines the corresponding mRNA sequences (after RNA processing). You may assume that each transcript contains **AT MOST** one intron and has had its poly-A tail removed; additionally, you do **NOT** need to identify start/stop codons or valid reading frames.
Tip: Review the online lecture notes on gene expression, especially the sections on RNA processing and on the nature of a "consensus sequence."

b. Use your program from part (a) to analyze the sequences contained in the file *MidBio1.txt*, and print the resulting output.

2. We have defined D as the proportion of sites that differ between two DNA sequences. Let us now define $v = V / D$ as the percentage of those differences that are ***transversions***. For example, consider two sequences in which 97% of sequences are identical, 2% differ by a transition, and 1% differ by a transversion. For this pair of sequences, $D = 0.03$ and $V = 0.01$, so $v = 0.01/0.03 = 0.33$.

a. Calculate and graph the Jukes-Cantor and Kimura 2-parameter distances as functions of v. Assume that the two sequences being compared differ at 20% of their positions ($D = 0.20$). Your graph should have an interval size **no greater than 0.01**; that is, it should plot d_{JC} and d_{K2P} for $v = 0.00$, $v = 0.01$, $v = 0.02$, and so on, up to the maximum possible value of v.

b. Explain why the Jukes-Cantor distance does not depend on v, but the Kimura 2-parameter distance does.

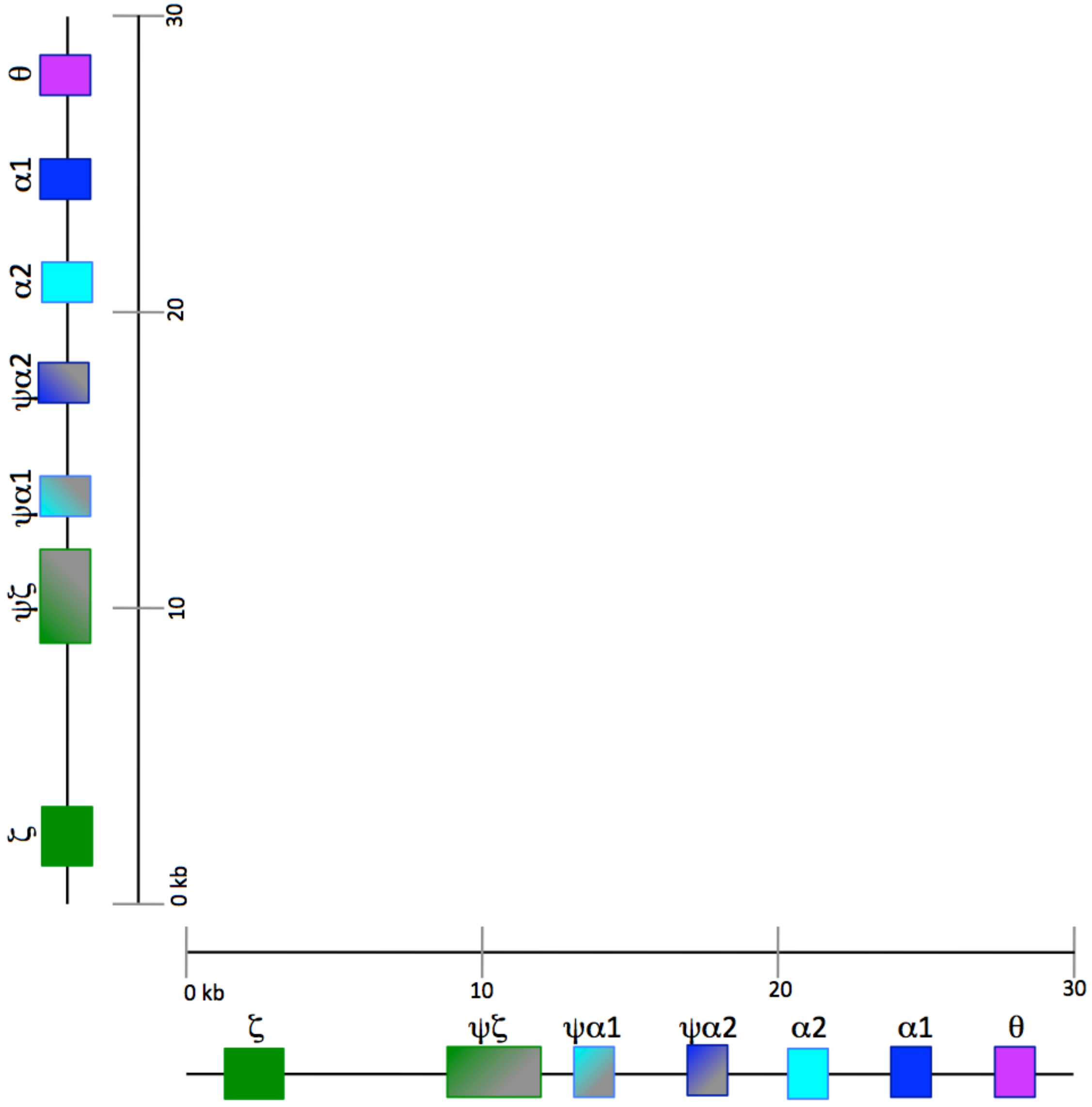
c. For what **exact** value of v does the Jukes-Cantor distance most closely approximate the more precise Kimura 2-parameter distance? **Clearly explain why.**

3. The file ***MidBio3.docx*** contains an alignment of 13 DNA sequences from the same genomic region. What model of molecular evolution (Jukes-Cantor, Kimura 2-parameter, HKY85, general time-reversible, or fully general) would be most appropriate for analyzing these sequences? **Clearly justify your decision**, showing any relevant calculations.

4. Following a gene duplication event, the two copies of the gene may diverge in sequence and function. In some cases, one copy of the gene loses most or all of its function as a result of accumulating mutations: this copy is called a ***pseudogene***.

a. Describe the <u>type of evolutionary pattern</u> you would expect to see in: (i) a functional hemoglobin gene, (ii) a pseudogene, (iii) an intergenic region. <u>Clearly explain your reasoning</u>. (Evolutionary patterns include balancing selection, directional selection, disruptive selection, purifying selection, and neutral evolution.)

b. The figure below shows the structure of the *α* globin gene clusters in mammals. <u>Sketch a dot plot of this entire cluster with itself</u>. Your dot plot should show **ALL** regions of substantial genetic similarity, according to the patterns you predicted in part (a). <u>Clearly explain your reasoning.</u> Note: the prefix "ψ" denotes a pseudogene of the corresponding gene, while the suffixes "1" and "2" denotes copies produced by gene duplication and subsequent divergence.
Tip: Review the online lecture notes on sequence alignment, particularly those dealing with large genomic regions (e.g., Lecture 5a, Slides 14–17).


θ
α1
α2
ψα2
ψα1
ψζ
ζ
30
20
10
0 kb
0 kb
10
20
30
ζ
ψζ
ψα1
ψα2
α2
α1
θ