BIOL 437/GCB 536 Final Test

1. **Ground Rules:** You may use reference materials including class notes, slides, additional primary literature, the mid-term presentations, etc. But, you CANNOT ask anybody, CANNOT consult any forum, nor do a web page search specifically for answers. For example, do not google "how do I know a sequence is non-coding"? I will have already done all the permutations of such queries and I will know. Please sign below to acknowledge that you will follow the ground rules (you can digitally sign or sign on printed page and send a picture).

   Your Name:

   Your Signature

2. **Length: No more than 1 page single space, 11pt, answer per question.**

3. **Question 0 is mandatory. For Questions 1-5, choose 3 out of 5 to answer**.

4. **Tip:** Many of the questions are open-ended. In these cases, I am not looking for correct answer per se but whether you can state a **GOOD RATIONALE** for the proposed solutions. Your answers will be graded based on the criteria of: (1) reasonably correct; (2) creative and new; (3) good consideration of pros and cons of the approach.

5. Your answers in word file or pdf file must be **emailed back to me at Junhyong@sas.upenn.edu by 12/15/2018 11:59 PM**. Being late will cost you 5% per each 10 min.

6. **IMPORTANT:** Your answer file should have the following file name format: BIO437.[Your First Initial].[Your Last Name].doc/pdf. For example, BIO437.J.Kim.pdf

Remember your HONOR, MOTHER, and DIGNITY!

ready, fight!

Question 0 (Mandatory)—The following equations have been selected from various parts of your readings. Briefly describe their meaning or their use. NOTE, precision of your answer matters and I will be grading your answers based on how much your answer conveys your understanding of the equation.

Q0.1:

$$f(n) = O(q(n))$$

Q0.2:

$$\min(x\mathbf{W}) = N(x) + B[x, \min(\mathbf{W})\text{-}1] + 1$$
$$\max(x\mathbf{W}) = N(x) + B[x, \max(\mathbf{W})]$$

Q0.3:

$$V(i,j) = max\begin{cases} V(i-1,j-1) + C(S(i),P(j)) \\ V(i-1,j) + C(S(i),-) \\ V(i,j-1) + C(-,P(j)) \\ 0 \end{cases}$$

Q0.4:

$$P(S(t+1)=i) = \sum_{k=1}^{n} P(S(t+1)=i \mid P(S(t)=k)P(S(t)=k)$$

Q0.5:

$$f_0(0) = 1$$

and

$$f_k(i+1) = e_k(x_{i+1}) \sum_{j \in \Pi} f_j(i) \cdot t_{jk}$$

Q0.6:

$$d(w,x) + d(y,z) = d(w,z) + d(x,y) \geq d(w,y) + d(x,z)$$

Q0.7:

$$1 \cdot 3 \cdot 5 \cdot \cdot (2t-5) = \frac{(2t-5)!}{2(t-2)!} = \frac{\sqrt{2\pi(2t-5)}e^{2t-5}(2t-5)^{2t-5}}{2\sqrt{2\pi(t-2)}e^{t-2}(t-2)^{t-2}} \sim O(t^t)$$

Q0.8:

$$\Pr ob\{Accept\ state\ y\} = \min\{\frac{p(y \mid data)q(x,y)}{p(x \mid data)q(y,x)}, 1\}$$

Q0.9:

$$K(x,y) = K(x-y) = \exp(-\|x-y\|^p / d)$$

Q0.10:

$$R(\alpha_n^*) \le R_{emp}(\alpha_n^*) + \phi(\alpha, n),$$

For the next set of questions, answer 3 out of 5 questions.

Question 1: Cancer often arises out of translocation mutations where one part of one protein is fused to another protein creating a novel oncogenic fused protein. The Cancer Genome Atlas (TCGA) project has used nextgen sequencing to directly sequence many of the known cancer cell lines and patient primary tumors. We wish to develop an algorithm to identify novel fusion protein from samples. Design this algorithm, analyze its computational complexity, and discuss what might affect its sensitivity (avoiding false negatives) and precision (avoiding false positives). **Hint:** Assume that Nextgen sequence reads are single end sequences of 150 bp long. Suppose the fusion protein is from some part of protein A and some part of protein B: part(A)+part(B). Remember a read may only fall on part(A), only on part (B), or span the junction between part(A) and part(B). When it spans the junction, the read may be X bp on part(A) side and (150 -X) bp on part(B) side where X may be anything from 1 bp to 149 bp.

Question 2: I am working to identify novel proteins that might be involved in axon guidance of olfactory sensory neurons. That is, how the axons of olfactory neurons in the nose find the right anatomical region to connect to in the brain. The general hypothesis is that it must involve some transmembrane protein that senses some signaling molecule released by the target region. We carry out RNAseq from carefully isolated regions of the nose that have sensory neurons and obtained 3,000 possible mRNAs from three different sets of neurons, those that project to proximal regions, those that project to medial regions, and those that project to distal regions. Describe how you would analyze this data to come up with candidate proteins. **Hint**: consider how to map the transcripts to genes to identify full length transcripts. Consider what kinds of amino-acid composition transmembrane proteins might have. Consider how to construct the appropriate feature space and approach learning method to predict these guidance proteins.

Question 3: Sequence motifs are short patterns in promoter regions of genes, which are usually binding sites for transcription factors. ChIP-seq is a way of isolating short sequences of DNA that is physically associated with some protein. I am studying the transcription factor, GATA4 and my goal is to characterize the DNA sequence motif for GATA4 binding. ChIP-seq experiments were carried out on mouse embryonic heart cells and the data were pre-processed for various controls and statistical significance and we were given 2,000 sequences of intervals of the genome ranging from 100 bp to 500 bp where GATA4 might have bound. (Note that transcription factor binding sites are never a unique pattern but a family of patterns that might be summarized with a position specific weight matrix.)

Q3.1: Design a method to process the 2,000 sequences and identify the GATA4 binding motif. Describe your approach, the computational complexity of your approach, and the output.

Q3.2: I have the hypothesis that GATA4 binding motif changes dependent on cellular context. Therefore, we did another experiment using kidney cells and obtained another 1,200 sequences. Discuss how you would use this data in combination with the original data to test my hypothesis.

Question 4: Tracing cell lineages can tell us a lot about the kinds of events that led to particular tissue traits or diseases like cancer. A tumor is typically a heterogenous collection of various mutant clones that arose during carcinogenesis. Furthermore, when a tumor becomes metastatic and establishes new tumor mass in various parts of the body, those might derive from some special subset of the original tumor. If we can trace the cell lineage of the heterogeneous clones, we may be able to reconstruct driver mutations that led to different tumor phenotypes. With single cell RNA sequencing, we can also use the resulting sequence data to infer DNA sequence variation—for example, when the RNA sequence has different base-pairs or has insertions or deletions from baseline, we might infer that those are due to DNA mutations. To reconstruct the cell lineage of human small cell lung carcinoma, we carried out single cell RNA sequencing of the lung tumor mass and two other metastatic sites in lymph nodes. We also carried out tissue RNAseq from normal cells to establish baseline sequences. I received sequence data from

100 individual cells from each location for a total of 300 cell sequence data and one tissue sequence data. Each sequence data is already annotated for base-substitution and indel mutations.

Q4.1: Design a method to use this data to infer cell lineage trees. Make sure you describe how you would use the each kind of mutation and the assumption involved in the method. Describe possible computational problems.

Q4.2: I wish to infer a suite of mutations that are hypothesized to be responsible for metastasis. Describe how you would use this data to test that hypothesis.

Question 5: I learned about various dimension reduction and clustering methods in class. I recently received a dataset of 10,000 human kidney single cell RNA sequence data. I try PCA, tSNE, and non-metric multi-dimensional scaling and the resulting pictures do not seem to show distinct clusters of cells I expected. From previous pathology information, I expected there to be at least six distinct cell types and likely at least another six more. I ask you to develop a novel custom dimension reduction and clustering algorithm. Describe what kind of approach you would take and why. Describe the pros and cons of your method compared to tSNE.

Bonus Question

Answer the following for extra 10%!

Recall the maximum likelihood phylogeny method discussed in class and also the Metropolis Hastings MCMC algorithm. Suggest a possible construction of a proposal function for Metropolis-Hastings algorithm to compute a Bayesian estimate of a phylogeny.