CS 262 Tuesday, February 23, 2016
Computational Genomics Handout #4

Homework #4
Phylogenetics and Gene Prediction
Due at the beginning of class on Tuesday, March 8

Collaboration is allowed in groups of at most four students, but you must submit separate writeups. Please write the names of all your collaborators on your solutions. You are not allowed to copy group work. If you are working alone, we will drop the problem with the lowest score. If you submit your solutions after the submission deadline, you must write the date and time of submission on your writeup. Under no circumstances will a homework be accepted more than three days after its due date.

1. **Problem 1** (25 points) **Multiple Sequence Alignment**

(a) (12 points) Consensus multiple alignment versus sum-of-pairs multiple alignment. Definitions: (adapted from Gusfield, p. 352)

(1) Given a multiple alignment M of a set of strings S, the *consensus character* of column i of M is the character that maximizes the summed score between the character and all the characters in column i. (In case of ties, say by convention that we prefer A over C over G over T over 'gap'). The score of (gap, gap) is 0. Let $d(i)$ denote that maximum summed score in column i.

(2) The consensus string S_M derived from alignment M is the concatenation of the consensus characters for each column of M.

(3) The alignment score of S_M equals to the sum of column scores $d(S_M) = d(1) + ... + d(m)$, where m has m columns.

(4) The optimal consensus multiple alignment is a multiple alignment M for input string set S whose consensus string S_M has the largest alignment score over all possible multiple alignments of S.
Example:

$S = \{AGCC, ACC, TCC\}$, and match, mismatch, gap = +2, -2, -3. Consider the following alignments:
M_1: $\{AGCC, A-CC, T-CC\}$; S_{M_1} = A-CC, and $d(S_{M_1}) = (+2) + (-3) + (+6) + (+6) = 11$.
M_2: $\{AGCC, A-CC, -TCC\}$; S_{M_2} = AGCC, and $d(S_{M_2}) = (+1) + (-3) + (+6) + (+6) = 10$.

Show an example with three or more sequences where all optimal multiple alignments according to the above model are different from all optimal alignments according to the Sum-Of-Pairs model. In other words, since there may be several equally-scoring optimal alignments, the set of optimal alignments for the consensus model must be disjoint from the set of optimal alignments for the Sum-Of-Pairs model.

Assume either a match, mismatch, and gap penalty of $(+2, -2, -3)$ or $(+2, -1, -1)$ (you may find the second set of scoring parameters easier to prove). Let the alphabet be $\{A, C, G, T\}$.

(b) (13 points) Phylogenetic-treebased alignment.

Definitions:

(1) Given an input rooted binary tree T with a distinct string (from a set of strings S) written at each leaf, a *phylogenetic alignment* for T is an assignment of one string to each internal node of T, **subject to**

the additional constraint described below. Note that the strings assigned to internal nodes need not be distinct and need not be from the input strings in S.

(2) If strings s and s' are assigned to the endpoints of an edge (i, j), then that edge has edge score $D(s, s')$, which is simply the maximum pair-wise alignment score between the two strings s and s'. The score of a phylogenetic alignment is the total of all edge scores in the tree.

(3) The *phylogenetic alignment* problem for T is to find an assignment of strings to internal nodes of T (one string to each node) that maximizes the score of the alignment.

Additional constraint:

In the maximum pair-wise alignment between a leaf node x and an internal node s, a letter in x must match or mismatch some letter in s, and not be gapped. Also, the maximum pair-wise alignment between two adjacent internal nodes s and s' may not have any gaps. Then, constructing the multiple alignment of the input sequences given the phylogenetic alignment is straight-forward.

Note: it is also possible to define the problem where T is unrooted. If you prefer that definition, please go ahead and use it instead.

$S = \{x = ACC, y = AGCC, z = TCC\}$, and match, mismatch, gap $= +2, -2, -3$. Let T = [Nodes $= x, y, z, v_{yz}, v_{xyz}$; Edges $= (x, v_{xyz}), (v_{xyz}, v_{yz}), (v_{yz}, y), (v_{yz}, z)$, Root $= v_{xyz}$] , with leafs $\{x, y, z\}$ and root v_{xyz}.

Here is a phylogenetic alignment, which is just a labeling of the internal nodes: Label v_{yz} with "AGCC", and v_{xyz} with "AGCC". Then, the alignment score is $D(x, v_{xyz})+D(v_{xyz}, v_{yz})+D(v_{yz}, y)+D(v_{yz}, z) = (+3) + (+8) + (+8) + (-1) = 18$.

Show an example with three or more sequences where all their optimal phylogenetic alignments differ from either the set of optimal consensus multiple alignments or the set of optimal Sum-Of-Pairs multiple alignments.

Similar to the previous problem, assume either a match, mismatch, and gap penalty of $(+2, -2, -3)$ or $(+2, -1, -1)$. Let the alphabet be $\{A, C, G, T\}$.

2. **Problem 2** (15 points) **Phylogenetic Trees**

a. In each step of building the tree by UPGMA, we compute a table of distances and choose two clusters i and j with the smallest distance d_{ij}, and create a new node at height $h_{ij} = d_{ij}/2$. Then, we remove the d_{ij} from a set of all pairwise distances and replace all distances d_{il}, d_{jl} where l is any other cluster with

$$d_{uv} = \frac{1}{|C_u||C_v|} \sum_{p \in C_u, q \in C_v} d_{pq}. \tag{1}$$

and create a new table of distances. (We ignore the diagonal terms in the tables, i.e., when $u = v$).

i. (3 points) Show that any distance $d_{i'j'}$ in this new table is no less than d_{ij}, i.e., $d_{i'j'} \geq d_{ij}$ for all i', j'.

ii. (3 points) Show that if $C_r = C_i \cup C_j$ and if C_s is any other cluster, then equation (1) implies

$$d_{rs} = \frac{d_{is}|C_i| + d_{js}|C_j|}{|C_i| + |C_j|}.$$

b. Given the following sequences, aligned with no gaps:

```
X: GCGGGCTG
Y: GCCGTCTG
Z: ACCGTCGG
W: ACCGTCTG
```

Define the distance d_{uv} between two sequences u and v to be simply the number of letter substitutions (Hamming distance).

i. (3 points) Build the average linkage (a.k.a., UPGMA) tree T_{AL} for these four sequences.

ii. (2 points) Is this distance function d_{uv} ultrametric on these four sequences?

c. (4 points) Show that the neighboring leaves (1, 3) and (2, 4) in the tree given below have the smallest distances according to the Neighbor-Joining distance metric D_{ij}.

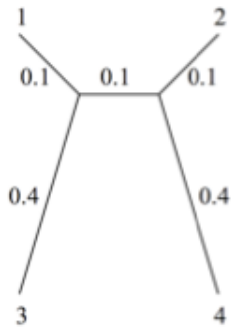

Figure 1: 2 (c)

3. **Problem 3** (25 points) **Jukes-Cantor model of evolution**

Before solving this problem, make sure you review lecture 13 notes on the Jukes-Cantor model.

(a)

i- (2 points) Assume that the genomic sequences of species X and Y are 95% identical, so they differ at $p_{XY} = 5\%$ of sites. What is the evolutionary distance d_{XY} between the two sequences in terms of substitutions per site assuming the Jukes-Cantor model of molecular evolution?

ii- (1 point) Now assume that species Y and Z differ at $p_{YZ} = 20\%$ of sites. What is the evolutionary distance d_{YZ} between them?

iii- (3 points) Is d_{YZ} simply 4 times larger than d_{XY}? Is it more than 4 times larger? Less? Briefly explain why this makes sense, based on how p and d are defined.

(b) (8 points) Let t_{XY} be the evolutionary time between sequences X and Y and t_{YZ} be the time between sequences Y and Z. Let $t_{XZ} = t_{XY} + t_{YZ}$ be the time between X and Z. According to the Jukes-Cantor

model, the probability that a site is identical between X and Y is $r(t_{XY}) = 1 - p_{XY} = (1 + 3e^{-\mu t_{XY}})/4$, while the probability that a site transitions to one of the other letters is $s(t_{XY}) = (1 - e^{-\mu t_{XY}})/4$. Show that $r(t_{XY} + t_{YZ}) = r(t_{XY})r(t_{YZ}) + 3s(t_{XY})s(t_{YZ})$.

(c) (6 points) Use the result from (b) to derive a relationship between p_{XZ}, p_{XY} , and p_{YZ} .

(d) (5 points) Show that the distances computed by the Jukes-Cantor model satisfy $d_{XZ} = d_{XY} + d_{YZ}$.

4. **Problem 4** (25 points) **Gene Prediction**

A simple strategy for locating genes in compact genomes not containing introns is to look for long open reading frames (ORFs). An ORF is defined as a sequence of DNA beginning with a start codon (ATG) and containing no in-frame stop codons (TAA, TAG, or TGA). ORF scanning works because genes contain long open reading frames which are unlikely to occur by chance.

(a) Suppose we select a random ORF from a section of noncoding DNA in which all positions are independent and each base is equally likely.

i- (6 points) What is the probability distribution for the length of such an ORF, where the length includes the start but not the stop codons?

ii- (2 points) What is the probability that the length is at least 100 codons (300 bp)?

iii- (4 points) What is the probability that the length is at least 100 codons if the base distribution is P(A) = 0.2, P(C) = 0.3, P(G) = 0.3, and P(T) = 0.2?

(b) Suppose we would like to find genes in Saccharomyces cerevisiae (bakers yeast) by ORF scanning. We estimate that only about 5% of yeast coding regions are less than 100 codons long based on our experience with other organisms. Therefore, we will predict any ORF of at least 100 codons as a gene. Assume for simplicity that the yeast genome has a uniform distribution of base pairs P(A) = P(C) = P(G) = P(T) = 0.25.

i- (7 points) For a noncoding region of length L, show that the probability of predicting at least one false positive gene in the region is no more than:

$$1 - \left(1 - \frac{1}{64}\left(\frac{61}{64}\right)^{99}\right)^{L-299}$$

Explain why this is an upper bound but not the exact probability.

ii- (3 points) Assume for simplicity that all the *S.cerevisiae* noncoding regions longer than 300 bp have length 500 bp and that there are 2,000 such regions. Using the upper bound from part i., and ignoring the possibility of multiple false positives in a single region, provide an estimate for the total number of false positive genes we will predict. Given that the yeast genome has about 6,000 real genes (95% of which we will correctly predict), what is the sensitivity and specificity of this approach?

Definitions:

Sensitivity = true genes predicted correctly / total true genes
Specificity = true genes predicted correctly / total genes predicted

iii- (3 points) Now suppose we would like to try ORF scanning in the human genome. Since many human exons are short, we set our threshold at 50 codons (150 bp). Inspired by part i., give an upper bound on the probability that we will predict a false positive coding region in a typical human intron of length 2,000 bp. Give a similar bound for a typical human intergenic region of length 50,000 bp.

5. **Problem 5** (15 points) **Basic Understanding of Population Genetics**

(a) (3 points) If beak length in Robins is controlled by a single pari of alleles with long beak associated with the presence of a dominant allele, what are the frequencies of dominant and recessive alleles if a population consists of 64 long beaked birds and 36 short beaked birds? Estimate the percentage of the population that is homozygous for the dominant (long beak) allele.

(b) (3 points) You have sampled a population in which you know that the percentage of the homozygous recessive genotype (aa) is 36%. Using that 36%, calculate the frequencies of the genotypes "AA" and "Aa".

(c) (3 points) A very large population of randomly-mating laboratory mice contains 35% white mice. White coloring is caused by the double recessive genotype, "aa". Calculate allelic and genotypic frequencies for this population.

(d) (3 points) Cystic fibrosis is caused by an autosomal, recessive mutation. Now assume that, among a certain population the frequency of the cystic fibrosis allele is 1/2500. What percentage of marriages between the population, who have NO disease, could possibly produce diseased children? Assume the cystic fibrosis locus is at equilibrium.

(e) (3 points) Do you think the following statement is true? "The allele frequencies changes faster with selection against dominant phenotypes than recessive phenotypes". (Explain your logic.)