

Assignment 3: Sequence analysis

Submission: Send your answers as a PDF or TXT document to biomedin214-spr0506-submit@lists.stanford.edu (NOT to the staff list)

Question 1

Assume equal proportions of the four bases (ATGC) and independent random distribution of bases. What is the probability of observing a run of A's of length 6 in a sequence with 6 bases?

Question 2

Consider a gap-less alignment between two nucleotide sequences of length n :

$$\begin{matrix} X_1 X_2 \dots X_n \\ Y_1 Y_2 \dots Y_n \end{matrix}$$

Where each X_i and Y_i is the i th nucleotide base in the sequence X and Y respectively. Let S be a measure of similarity between the two sequences. As usual, we define S to be the sum over the pair-wise similarities:

$$S = \sum_{i=1}^n \sigma(X_i, Y_i)$$

where σ is a pair-wise similarity function between two nucleotides. Recall that the expected value of a function g of a discrete random variable T is:

$$E(g(T)) = \sum_{t \in T} P(T = t) \cdot g(t)$$

where the summation sums over all possible values of the random variable and $P(T=t)$ is the probability that T equals a particular value t . Let the pair-wise scoring scheme be the following, where we penalize transitions and transversions unequally:

	A	G	C	T
A	1	0	-1	-1
G	0	1	-1	-1
C	-1	-1	1	0
T	-1	-1	0	1

In the above matrix, each cell represents a particular value of σ . For example, according to the matrix,

$$\sigma(A,G) = \sigma(G,A) = 0.$$

Now suppose the two sequences, X and Y , are drawn randomly and independently from the yeast genome where the percentage of adenine is 31%, thymidine is 31%, cytosine is 19%, and guanine is 19%. Assume each position in the alignment is independent.

Compute an estimate for $E(\sigma)$, the expected value of σ .

Question 3

Which one or more of the following mutations are transitions (as opposed to transversions):

- a. A \leftrightarrow C
- b. A \leftrightarrow T
- c. A \leftrightarrow G
- d. C \leftrightarrow T
- e. C \leftrightarrow G
- f. T \leftrightarrow G
- g. purine \leftrightarrow purine
- h. purine \leftrightarrow pyrimidine
- i. pyrimidine \leftrightarrow pyrimidine

Question 4

If g is the penalty for opening a gap, and L is the penalty for lengthening the gap by 1, which of the following is most likely to be true and why?

- a. $g > L$
- b. $g < L$
- c. $g \sim L$

Question 5

Why might one choose to use correlation rather than Euclidean distance for clustering microarray data? (Three sentences or less.)

Question 6

Multiple Sequence Alignment and 1D Motifs

p53 proteins are involved in cell regulation and cancerous cells often have mutations in these genes. This family has a PROSITE profile that has 100% precision ($TP / (TP + FP)$) and 100% recall ($TP / (TP + FN)$).

Go to Prosite, find the p53 family, and give the pattern.

Question 7

Explain in words what that profile means. (One sentence.)

Question 8

What is the upper bound on the sum-of-pairs score for a multiple alignment?

Question 9

Gibbs Sampling

Over the next 4 questions, you will analyze 11 malaria DNA sequences for a conserved motif with Gibbs Sampling. The initial random alignment is below; one sequence has been left out.

```
sequence 1   C CAG A
sequence 2   G TTA A
sequence 3   G TAC C
sequence 4   A AGC T
sequence 5   C AGA T
sequence 6   T TTT G
sequence 7   A TAC T
sequence 8   C TAT G
sequence 9   A GCT C
sequence 10  G TAG A
```

Step one: Calculate a PSSM for the given alignment.

To calculate background frequencies, assume that in the malaria genome, 70% of positions have an A paired with a T. Use that information to calculate a log odds matrix for the central three positions. Use log base 2. To simplify things we will not use an alpha constant since in this example we don't have to worry about division by zero.

Write your answer in the following format:

[base] ; [col1 log odds] ; [col2 log odds] ; [col3 log odds].

For example, if at the first position we have 3/10 of the sequences having an A, the second position has 10/10 sequences with an A, and the third has 1/10 sequences with an A, then the output would be for base A would be:

A ; -.222 ; 1.515 ; -1.807

Do the same for all 4 bases.

Question 10

Step Two: Align the left-out sequence.

AGTCG is the left-out sequence. Find the log odds score for each of three possible positions in the left-out sequence.

Write your answers in the following format:

[site position] ; [subsequence] ; [log odds score].

For example:

If the subsequence you are matching to the matrix is GCT and the log odds of having a G in the first position is 1.233, the log odds of having a C in the second position is -.034, and the log odds of having a T in the third position is -.854, then the output would be:

1 ; GCT ; .345

Question 11

Step Two continued:

Calculate the probability of a match at each position in the left-out sequence. (Odds score = $2^{(\log \text{ odds score})}$.)

Write your answer in the following format:

P([subsequence])=[score].

Continuing the above example:

The odds score of GCT is 1.27. If the odds score of the second position is .221 and the odds score of the third position (in the left-out sequence) is .923, then the probability of position one is .526, written as:

$$P(\text{GCT}) = .526$$

Question 12

Step Two continued:

Which subsequence is the most likely location for the motif in the left-out sequence? (Just take the subsequence with the highest probability in the last question.)

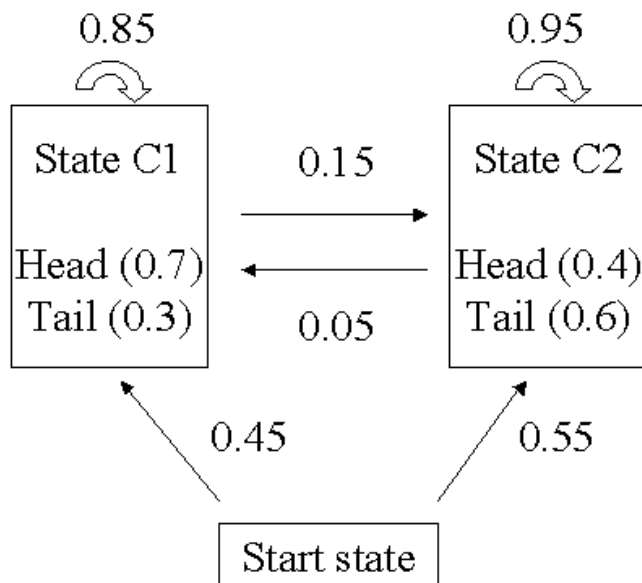
Question 13

Does Gibbs sampling always take the most probable subsequence to align the motif? How does it choose the position?

Question 14

This question should give you a feeling of how HMM works. The following questions are based on a greatly simplified HMM structure, and we will provide some analogy to protein secondary structure prediction with HMMs.

Consider an HMM that models the process of repetitively tossing two biased coins, C1 and C2.



The above HMM has two output symbols head (H) and tail (T) and two states (the two coins, C1 and C2). Parameters without "()" indicate the likelihood of the model moving from one state to another state. Parameters with "()" indicates the likelihood of tossing H or T within a particular state.

Compute the probability of the output sequence "tthhh" being generated by the above HMM using the forward algorithm given here.

Forward algorithm implements the following recurrence relation (Pr = probability of):

$$\alpha_i(\text{state } s) = \begin{cases} \text{Pr}(\text{from start to state } s) * \text{Pr}(\text{output symbol at state } s) & \text{for } i = 1 \\ \sum_{\text{all possible } r} [\alpha_{i-1}(\text{previous state } r) * \text{Pr}(\text{transition from } r \text{ to } s) * \text{Pr}(\text{output symbol at state } s)] & \text{for } i > 1 \end{cases}$$

You may work out the following by hand or by code.

$$\begin{aligned} \alpha_1(s=C1) &= \\ \alpha_2(s=C1) &= \\ \alpha_3(s=C1) &= \\ \alpha_4(s=C1) &= \\ \alpha_5(s=C1) &= \end{aligned}$$

$$\begin{aligned} \alpha_1(s=C2) &= \\ \alpha_2(s=C2) &= \\ \alpha_3(s=C2) &= \\ \alpha_4(s=C2) &= \\ \alpha_5(s=C2) &= \end{aligned}$$

$$\text{Pr}(\text{output sequence}="tthhh" \mid \text{HMM}) =$$

Question 15

Compute the state sequence that was most likely to have generated the observed sequence "tthhh" using the Viterbi algorithm (the algorithm that was discussed in class). The Viterbi algorithm realizes the following recurrence relation:

$$\xi_i(\text{state } s) = \begin{cases} \text{Pr}(\text{from start to state } s) * \text{Pr}(\text{output symbol at state } s) & \text{for } i = 1 \\ \text{maximum}_{\text{all possible } r} [\xi_{i-1}(\text{previous state } r) * \text{Pr}(\text{transition from } r \text{ to } s) * \text{Pr}(\text{output symbol at state } s)] & \text{for } i > 1 \end{cases}$$

Again, you may compute the following by hand or code.

$$\begin{aligned} \delta_1(s=C1) &= \\ \delta_2(s=C1) &= \\ \delta_3(s=C1) &= \\ \delta_4(s=C1) &= \\ \delta_5(s=C1) &= \end{aligned}$$

$$\begin{aligned} \delta_1(s=C2) &= \\ \delta_2(s=C2) &= \\ \delta_3(s=C2) &= \\ \delta_4(s=C2) &= \\ \delta_5(s=C2) &= \end{aligned}$$

Most probable sequence of states, q =

Question 16

Instead of heads/tails, how many output symbols are there for a protein secondary structure prediction HMM model (where we predict secondary structure from protein sequence)?

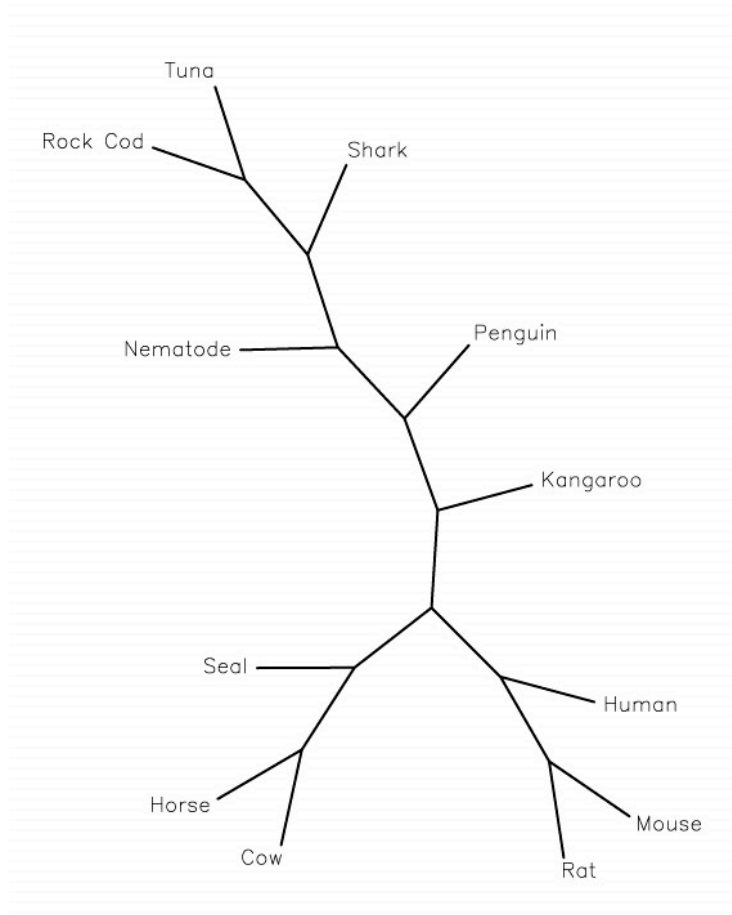
Question 17

Instead of C1/C2, how many states are there for a protein secondary structure prediction HMM model?

Question 18

Phylogenetics

One of the outputs from the multiple sequence alignment program, ClustalW, is a dendrogram. A dendrogram shows the evolutionary relationships between proteins/species. The distance between species on the tree represents the alignment score for the sequences. Below is shown the dendrogram as drawn by WebPhylip after aligning several different organisms' myoglobins.



This tree is unrooted. What is a way to determine where the root would be relative to the other sequences?

Question 19

Can you suggest an alternative protein besides myoglobin that would give better discrimination for very close species (eg chimp, gorilla and human)? Why? If you can't think of such a protein, at least give a property that it must have.

Question 20

Briefly compare maximum parsimony and maximum likelihood methods for building dendrograms. Show that you understand the meaning of each and the difference between them. (3 sentences or less.)

Question 21

Gene Finding.

One task after completion of a genome project is predicting the locations of genes. The following questions will go through some of the issues involved.

A common way of identifying gene sequences is to locate "open reading frames," which are regions from a start codon (ATG) until the next stop codon (TAA, TAG, or TGA), within the same reading frame. A problem with this approach is false positives, or gene sequences that do indeed have a start and stop codon, but are not actually not a gene sequence. The probability of such false positives depend on the sequence length. Assuming that each nucleotide has the same probability ($1/4$) of being at any position in a DNA sequence and that we have identified a start codon, what is the probability of not having a stop codon until 150 nucleotides (i.e., 50 codons) following the start codon? I.e., the sequence would be a start codon followed by 150 nucleotides followed by a stop codon, but you are only calculating the probability of the middle 150 nucleotides.

List your calculation and results below and explain your calculations briefly. Keep in mind that there are three stop codons that you should not count.

Question 22

Again assuming that each of the 4 nucleotides (A, T, G, C) shows up at each position with the same probability ($1/4$), how long does a gene sequence have to be for the probability of a random sequence as defined in the question above to be less than or equal to 0.01? List your calculation and results below. Explain your calculations briefly. (Again, do not include the stop codons.)

Question 23

Comparative genomics allows one to predict coding exons with great confidence in the absence of anything but sequence information. Very briefly, how can a multiple genome alignment be used to locate coding exons in a sequence?

Question 24

You are now faced with the task of finding coding regions (exons) within a long stretch of human DNA sequence. The sequence is at

<http://www-helix.stanford.edu/bmi214/assignment3/data/AB049169.txt>

A stretch of DNA has 6 possible open reading frames (subsequences of the DNA that code for proteins): three on each strand, since codons are three bases long. Shifting the open reading frame yields completely different protein sequences. For example, the sequence:

ATGGTCTT

has the following 3 ways of being split into codons:

ATG GTC TT
A TGG TCT T
AT GGT CTT

For the reverse complement of the sequence:

AAGACCAT

you get the following codons:

AAG ACC AT
A AGA CCA T
AA GAC CAT

Gene prediction algorithms often look at all 6 possible open reading frames for the sequence that has the longest open reading frame (longest possible protein sequence before hitting a stop codon). These algorithms also must find intron-exon boundaries.

One of these programs for gene finding is GenScan. Go to <http://genes.mit.edu/GENSCAN.html> Use GenScan to identify exons in this sequence. You can also view the output graphically via the link to a pdf.

How many exons does GenScan find?

Question 25

The name of the gene is AB049169. Go to the Santa Cruz human genome browser at <http://genome.ucsc.edu/>. Click on "Genome Browser" on the upper left of the screen. Type in the gene name in the "position or search term" field and submit.

You should see the gene in the browser. Below the gene are optional tracks to display. You can play around with these, but make sure to turn on "Known genes", "Genscan", and "Human mRNAs" if they aren't already. Using the "dense" option will give you more detailed information. Make sure to hit "refresh" after changing track options.

GenScan didn't appear to get the gene totally correct, compared to the known information and the conservation. What did GenScan do wrong? (Although we can't know for sure that it's incorrect, just compared to the knowledge on the display.)

- a. Prediction of alternative promoters (multiple biologically-correct transcription start sites for the same gene) where there was none
- b. Prediction of alternative splicing (multiple biologically-correct ways to splice the same gene) where there was none
- c. Used the wrong open reading frame for the protein.
- d. Predicted an extra exon