CS1820 Midterm

Due: Friday, April 4 11:59pm

Directions. All students must answer Problem 0 (labeled required), computational students must answer problem 1 COMP, biology track students must answer problem 1 BIOL, and all students must answer three of the remaining five (labeled optional) for a total of five questions.

Genome Assembly

Problem 0 *Comparison of genome assembly projects (required - 20 points)*

The human genome was sequenced and assembled by two different methods. Read the two different papers posted on the website. Describe in detail the methods used by the public Human Genome Project, laid out in their Nature paper. Do the same for Craig Venter and Celera Genomics as laid out in their Science paper. What are some major differences and similarities between these methods (use the Istrail *et al* PNAS assembly comparison paper provided)?

Look at the findings from the analysis of the sequence of the human genome. What percent of the genome was found to be repetitive? What do these repeats consist of? What is the total number of genes found? Why was this number surprising?

Is every nucleotide of the human genome sequenced by these methods? What were some difficulties in the sequencing and assembly of the human genome? Why was Nobelist Ham Smith worried about windows in the wet lab vis a vis genome sequencing and assembly?

Problem 1 *BIOL Idury-Waterman (required - 35 points)*

Compute the genome assembly using the Idury-Waterman algorithm on the reads located in file sub_sequence.fasta.reads for $k = 5$ and $k = 6$. Each line in sub_sequence.fasta.reads is a read 6 bases long sampled from the forward (5' to 3') strand. Show all steps.

Problem 1 *COMP Finishing an assembly: mate-pairs and scaffolding (required - 35 points)*

In Homeworks 6A & 6B you were asked to implement fundamental components of a genomic assembler. In reality, what was built was a contig enumerator. Large pieces of contiguous DNA were assembled into contigs but the orientation and relative ordering of the contigs remained unknown. In this problem you will implement your own method for taking a set of contigs and producing one or more scaffolds.

Definitions. A *super-contig* or *scaffold* is a series of contigs that are ordered and oriented (usually from the use of mate pairs) but not necessarily connected.

We will provide you with two text files. The first (contigs.txt) will contain one contig per line and the second (mate_pairs.txt) will contain one mate pair per line. Each line in the mate pair file will consist of a read of some length, then an integer insert length distance (this distance includes the length of the first read), then another read (tab-delimited). The reads will be sampled from the forward (5' to 3') and reverse (3' to 5') strands. If the read was taken from the forward strand, the read listed first occurs before the read listed second (in absolute genomic index of the reads). If the read was taken from the reverse strand, then the read listed first occurs after the read listed second.

Given these files we ask you to **write a program that constructs the scaffolding for the corresponding genome**. You will be graded on how well you orient and order the contigs using the mate pairs. Keep in mind that the mate pair reads may contain sequencing errors (errors on a larger scale will be ignored for this exercise). In addition, the mate pair reads may also contain erroneous insert lengths (the estimated length between mates) or reversed orientation. Each type of error has an incidence rate of about 2%. One or both ends of the read may also fall into a region that was not assembled into a contig. The data we generated we reads of length 35 and a coverage of 5.

In addition please answer the following questions:

1. How many super-contigs were you able to assemble?
2. What parameters could you change to achieve a better assembly? Describe how modifying these parameters could help.
3. Assume now you receive mate pair reads from a diploid genome. What new problems would arise from attempting the assembly of the diploid genome?
4. Extra-credit (2 points): We should have warned you earlier, but the organism whose genome these contigs were generated from is extremely deadly. Use bioinformatics tools to identify this organism.

Please choose three of the following five problems to solve. None of these problems require any coding.

Bias and assumptions computational biology

Problem 2 *How NOT To Lie With Statistics (optional - 15 points)*

Over the course of the semester, we've learned how statistical biases can find their way into genomic analysis and how statistics can be contorted to yield a bias towards a given result. Think back to the readings from the book "How to Lie with Statistics." The following is a list of claims. Consider each claim and (a) provide an explanation as to how the claim might have a built-in statistical bias and (b) informally describe what you would need to do to in order to verify the claim and correct for bias.

1. **Alignment scores.** The best way to locally align two amino acid sequences from different species is to estimate the corresponding species evolutionary distance and choose the appropriate PAM or BLOSUM matrix.

2. **Convergence or divergence.** There might be particular patterns of amino acids that are repeatedly selected for use in secondary structure motifs like alpha sheets and beta helices. Alternatively, there might be some combinations of amino acids that do not occur because of structural instability or steric problems. A further problem is that the 20 amino acids do not occur with equal frequencies. The three most frequent occurring amino acids are glycine, alanine, and leucine - account for a quarter of all residues. They occur four times as often as the least frequent amino acids - tryptophan, histidine, cysteine, and methionine.

3. **Protein Evolution.** The exon shuffling hypothesis claims it is likely that the overwhelming majority of proteins have evolved from a very small number of archetypal proteins.

4. **Identities.** The significance of percentage identity in an alignment between two proteins is very much a function of the lengths of the sequences being compared. So percentage identity by itself is not meaningful.

5. **Protein Data Bank.** For an analysis of random proteins amino acid content one may use all the proteins stored in the Protein Data Bank. The protein data bank stores the crystallized structures of proteins derived from researcher's experiments.

Sequence Alignment

Problem 3 *Circular chromosomes (optional - 15 points)*

Prokaryotes have circular chromosomes rather than packed linear chromosomes as we eukaryotes have. Describe a local alignment algorithm for circular chromosomes (objective function, initialization, recursion, overview of traceback). You will be graded on the following, in order of most points to least: optimality, termination, runtime.

Problem 4 *DNA vs. Protein alignments (optional - 15 points)*

If we look at identifying homology between two genes, should we compare DNA coding regions or protein sequences of the genes? Let us define the question in term of PAM similarity matrices. We used in class PAM matrices for aligning protein sequences. And we presented the statistical theory behind general similarity matrices such as PAM. The key concept was that of substitution between one amino acid to another and collecting the statistics from alignments that have a lot of conservation between proteins sequences. Remember 1 PAM is on average one amino acid change per 100 amino acids in a protein sequence.

We can define PAM matrices for DNA sequences as well. You need to think what could be meaningful substitutions there. Again, 1 PAM is one substitution in 100 DNA bases on average. As such substitution in coding regions could be synonymous and non-synonymous 1 percent change means different things at the DNA coding region level compared to the protein sequence level. One amino acid corresponds to three DNA bases in the corresponding codon. Evolutionary data shows that about on average 1.5 synonymous DNA substitutions happen for every one non-synonymous DNA substitution.

Based on this data provide an informal argument based on information theory that at evolutionary distances, such as 120 PAMs, significant amounts of information available in the protein comparison are lost when comparing corresponding DNA sequences

Information Theory and Similarities Matrices

Problem 5 *Relative Entropy (optional - 15 points)*

Given a gapless alignment of sequences, we can define a consensus sequence by a plurality vote of the bases. The way that we score information of the consensus is to evaluate the following function, where i is a position in the sequence from 1 to n, the length of the sequence.

$$\sum_{i=1}^{n} [p_{x_i} \log_2(p_{x_i}/.25)]$$

where x_i is the consensus base at position i and p_{x_i} is the proportion of sequences where x_i is found at position i (ie. a match with the consensus). But let's focus on a single base x_i from the consensus sequence:

$$I(p_{x_i}) = p_{x_i} \log_2(p_{x_i}/.25)$$

Answer each of the following. "Explain the consequences" means to describe what the given value of I implies for the relative frequency of x_i (the consensus base) to the three other bases (which $\neq x_i$).

a) What is $\min_{p_{x_i}} I(p_{x_i})$? What is $argmin_{p_{x_i}} I(p_{x_i})$? Why? Explain the consequences.

b) What is $\max_{p_{x_i}} I(p_{x_i})$? What is $argmax_{p_{x_i}} I(p_{x_i})$? Why? Explain the consequences.

c) For what p_{x_i} does $I(p_{x_i}) = 1$? Explain the consequences.

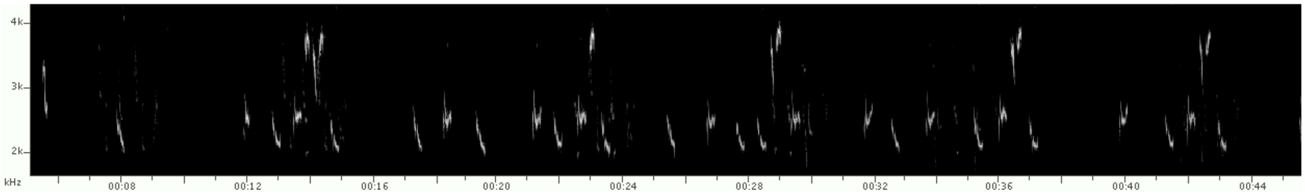

Figure 1: Sonogram of a male Baltimore Oriole.

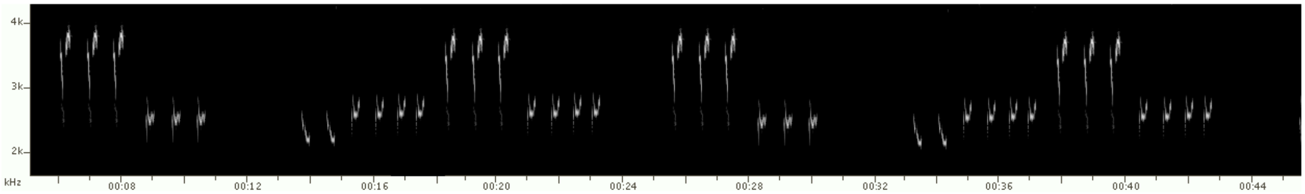

Figure 2: Sonogram of a male Baltimore Oriole alerting neighboring birds of danger.

Hidden Markov Models

Problem 6 *Bird songs (optional - 15 points)*

Songbirds can generate sequences of sounds roughly corresponding to syllables of a complex language. Songbirds have been previously used as model organisms for the study of language evolution [1] and, recently, the genetic bases of autism spectrum disorder [2]. Songbirds present an interesting model organism to study the vocal learning component of language that cannot be found in many model systems. Both songbirds and humans learn language during development phases, rely on social interactions to build language faculty, and require specialized brain function to produce new sounds.

A sonogram can be used to visualize the sound waves of bird songs. The x-axis of a sonogram is time and the y-axis is the frequency of the sound. A normal male Oriole song can be seen in Figure 2. The sound can be heard at http://macaulaylibrary.org/audio/113501/raven-viewer and the sonogram can be visualized with Quicktime at http://macaulaylibrary.org/audio/113501/raven-viewer. Consider this the song of a normal Oriole.

An Oriole attempting to alert others of a possible threat has a different sequence of sounds (this data is simulated).

Your task is to describe the properties of a hidden Markov model that can differentiate normal Oriole songs from alert songs. You should **describe the state space, the observations, the state transition probabilities, and the emission probabilities**. Sonograms are continuous measurements but your HMM should operate in discrete space so you will have to discretize the sonograms. **Given a new sonogram of an Oriole, what algorithm would you use to find the most probable sequence of hidden states?**

Extra-credit

Problem 7 *Learning to sing (optional - 5pts)*

Typically, it is believed that songbirds learn how to sing from their parents and social interactions with other birds. But Professor Istrail's bird Coco is not a typical bird. Coco was born in Albuquerque, New Mexico to a pair of renegade cockatiels. Coco's parents were on the run from the law so they begrudgingly had to leave Coco with a woman in Albuquerque that took care of birds. Shortly thereafter, Professor Istrail adopted Coco and brought him back to wonderful Rhode Island. Coco has a cushy life in his new home, watching television, listening to music (mostly the Beatles), and flying around the house.

After living in Rhode Island for some time, Coco began to sing to the surprise of Professor Istrail (coco_singing.mov). Your task is to **describe a model and algorithm to determine what inspired Coco to sing.** Your algorithm should be able to determine if Coco's songs were learned from television, music, other birds outside, humans, his parents in the brief time they were together, or some combination. Hauser *et al.*[1] is a suggested reading which can be found on the course web.

References

[1] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[2] S Carmen Panaitof. A songbird animal model for dissecting the genetic bases of autism spectrum disorder. *Disease Markers*, 33(5):241–249, 2012.