ECE-S690-503 Genomic Signal Processing: Phylogeny and Information Theory Homework Assignment 4

Due on:

June 3rd, 2009

Introduction

In class, we briefly covered that DNA sequences can be analyzed with entropy measures and that these can lead to exciting applications such as sequence logos. Also, Briefly how to construct phylogenetic trees. **You must turn in your Matlab code in your solutions, or you will receive a 0. Start Now!**

1 Introduction to Phylogeny

```
X: TACCCGAT
Y: TAAACGAT
Z: AAAACGCG
W: AAAACGAT
```

Define the distance $S(u, v)$ between two sequences u and v to be simply the number of letter substitutions (Hamming distance). Is this distance function $S(u, v)$ ultrametric on these four sequences? Regardless of ultrametricity, build the average linkage (a.k.a., UPGMA) tree TAL for these four sequences.

Unfortunately, we do not have enough data and we wish to assess the confidence of the branches of this tree. **Bootstrap the above data using 5 replicates and compute the confidence of each branch.**

2 Information Theory and DNA

Entropy of a sequence is defined as:

$$H = -\sum_{i=1}^{4} p_i log_2(p_i) \tag{1}$$

where p_i is the probability of base $i \in \{A, C, T, G\}$.

In this exercise, we will look at the entropy of a) whole sequences, b) of windowed sequences, of c) particular base positions, and d) even look at how the entropy may differ from a more "random sequence".

From the Pinho and Neves paper on the three-state model in compression, take the sequence GI:42759850, Saccharomyces Cerevisiae Chromosome III. **What is the length of the sequence? What is the probability, p_i, of each base?**
What is the entropy of the sequence?

Before we have explored *featuresparse* in Matlab. Use this command to extract all the coding sequences (look at the *Indices*). You will notice that some of the indices are backwards. **Why is this?** Use *seqrcomplement* to get the correct coding sequences from these parts. Paste together all the coding sequences into one sequence (do a loop so they are in sequence).
What is the length of the coding sequence?

What is the probability, p_i, of each base of the coding sequence?
What is the entropy of the coding sequence?
How does this entropy differ from the total sequence entropy? Why do you think it differs like this?

Now, calculate the entropies of each Codon Position (CP), so you'll be calculating for 3 base positions, on the contiguous coding sequence.
Can you explain why some CPs may have a higher entropy than others? Take the mean of the three CP entropies. What is the mean of the CP entropies? How does the mean of the CP entropies of the coding sequence compare to the entropy of the coding sequence (without CP sub-computations)?
How does the mean of the coding sequence CPs differ from the total coding sequence entropy?

Now, take the coding sequence and convert it to amino acids. Taking this sequence, convert it back to nucleotides using *aa2nt*. (Since *aa2nt* is a many to one mapping, it picks random nucleotides where there are degrees of freedom. This will result in a more randomized sequence that yields the same amino acid structure). Using this more random coding sequence(MRCS), **calculate the total entropy of the sequence.**
Calculate the entropies of each CP on the contiguous coding sequence(MRCS). What is the mean of the CP entropies(on MRCS)? How does the mean of the more random coding sequence(MRCS) CPs differ from the total MRCS entropy?

How do the CP entropies of the MRCS differ from the CP entropies of the real coding sequence? How does the total entropy of this MRCS differ from the total entropy of the real coding sequence? Why do you think this is?

Now I would like you to make a function to take a window length, N, of a sequence and to compute the entropy of that window. **Plot the entropy of the full sequence and coding sequence (separate plots) for $N = 50$. List the 50-length base sequences for windows that fall under 1 bit of entropy.**
Also do these plots again for another arbitrary N.

3

In class, we learned of a way to represent the information content of binding sites (or any sites that are highly conserved but still have differences/mutations). The only problem is that we need to have many realizations from the "random process". I will have you test the sequence logo equations in this problem, but a nice Web generator for sequence logo content can be found at `http://weblogo.berkeley.edu/`.

Go to a database of E. Coli binding sites: `http://arep.med.harvard.edu/ecoli_matrices/`. Click on the *lexA* link. Then, the *alignment* link contains the list of binding sites. Unfortunately between each sequence, there are notes about the sequence (you will need to strip these when importing the sequence into Matlab).

From class:

$$R_{sequence}(l) = 2 - (H(l) + e(n)) \tag{2}$$

where $H(l) = -\sum p(b,l) log_2(p(b,l))$ and $b \in A, C, G, T$ and l is the position of the sequence. So, you may ask "What is $e(n)$?" Well, we are going to set $e(n) = 0$ and see how our magnitude of our sequence logo differs from the Berkeley Weblogo for the same sequence.

Find the $R_{sequence}(l)$ where l is each position in the sequence. Plot this $R_{sequence}$ vs. l. Also, make a $b \times l$ table of the base information, b_i per l position. Print this table out. (Remember the $b_i = p(b,l)R_{sequence}(l)$).

(Find Berkeley's plot of the sequencelogo for this sequence here:
`http://weblogo.berkeley.edu/img/lexA.png`).
How does your $R_{sequence}$ vs. l plot compare to the contour of this binding site at Berkeley's Weblogo? How do you think the $e(n)$ compensates the Information content calculation?