Turn in all plots and code (to

gailr@ece.drexel.edu

Finding ORFs empirically.

Mycoplasmas are members of the class Mollicutes and comprise a large group of bacteria which lack a cell wall, have small genomes and a characteristically low G+C content. Mycoplasmas are of interest because they are believed to represent a minimal life form, having yielded to selective pressure to reduce genome size. The species with the smallest genome size in this class is Mycoplasma genitalium (580 kb).

1. Obtain the dna sequence of M. genitalium can be obtained from the EntrezSite with the accession number NC_000908.
2. In a nucleotide sequence an obvious thing to look for is if there are any open reading frames. The function **seqorfs** can be used to determine the ORFs in a sequence. Use Seq orfs on M. Genitalium.
3. The variable orf is a structure with information about the start/stop positions of each ORFs, its length and which reading frame it is in. Here the minimum number of codons for an ORF to be considered valid is 10 (default value). The minimum number of codons for an ORF to be considered valid can be also set. The original genome paper gave the number of genes as about 470. Set the seqorf minimum length threshold to 90 and 100 respectively, what do you get?
4. What are the total number of ORFs (e.g. set the minimum length to 1).

The classic approach to decide whether an ORF is a good candidate as a gene is to calculate the probability of seeing an ORF of a certain length L in a random sequence. To test the significance of ORFs a single-nucleotide permutation test can be used.

1. Permute the sequence using genitalium_seq(randperm(length(genitalium_seq))).
2. Run seqorfs on it.
3. Make a histogram of the original ORFs and the random ORFs. If the distributions are too long-tailed, Feel free to only plot a histogram of ORFs up to lengths 300->500 aa (e.g.

ORFLength(ORFLength<500)). Adjust the size of the bins (try to make a histogram that looks nice comparing them).

4. Find the maximum random orf length. How many Genitalium lengths are greater than this random maximum length? What are they?
5. Set an empirical threshold to empirical_threshold=prctile(ORFLength_random,95). This is a more tolerant threshold in order to keep all ORFs of length equal to or greater than the top 5% of random ORFs. What is it? How many of the Genitalium orfs are above it?

C+G Content Segmentation with an HMM.

Phages are viruses that infect bacteria, and Bacteriophage lambda infects the bacterium Escherichia coli, a very well studied model system. Bacteriophage lambda was the one of the first viral genomes to be completely sequenced (1982). It contains about 48502 bases. The Genome repository at the NCBI contains more interesting information about it.

1. Get the Blambda sequence NC_001416.
2. Plot the ntdensity with a window size of 2000, 3000, and 4000 respectively. Turn in the plots.

The analysis of the plots reveals that the phage genome is composed of two halves with completely different GC content: the first GC rich, the second AT rich. This is an example of change point in a genome.

You can use an HMM to segment the Lambda Phage genome into blocks of these two states. You can start generating random transition and emission matrices as input to the Expectation Maximization (EM) algorithm that better estimates those parameters.

1. Make random transition and emission matrices.
   T=rand(2,2);
   E=rand(2,4);

   % Normalize matrices

```
T(1,:) = T(1,:) ./ (norm(T(1,:),1));
T(2,:) = T(2,:) ./ (norm(T(2,:),1));
E(1,:) = E(1,:) ./ (norm(E(1,:),1));
E(2,:) = E(2,:) ./ (norm(E(2,:),1));
```

2. Encode the BLambda into integers.

```
seq=nt2int(BLambda);
```

3. Train the estimation and the transition states for the HMM using the initial transition and emission states, and the numerical Blambda sequence. (e.g. hmmtrain)
4. Estimate the segmentation states with the Viterbi algorithm and the matrices previously calculated. (e.g. hmmviterbi)
5. Make a plot of both the ntdensity and these estimated states.

```
ntdensity(BLambda);

hold on

plot(estimatedStates-1,'k--') % for visualization the states are coded as -1/1

hold off
```

6. Now, estimate the segmentation states with the viterbi algorithm using initial guesses: T and E. Do the above plot in 5., but now with the random states. Compare this plot to plot 5 – how does it differ?