Biological Data Analysis (CSE 182) : Assignment 4

Background

CpG dinucleotides (C followed by a G) occur with a much lower frequency in the sequence of vertebrate genomes compared to what is expected. The frequency of CG dinucleotides in the human genome, which has a 42% GC content, is 0.01 which is significantly lower than the expected frequency of CGs (0.0441). Regions of the human genome with elevated frequency of CG dinucleotides are referred to as CpG islands and are typically found in the promoter regions of human genes.

Problems

1. Given a DNA sequence (fasta file), write a program to calculate the frequency of each dinucleotide. Compare the observed frequency of each dinucleotide to its expected frequency (based on the frequencies of A, C, G and T nucleotides). Identify the dinucleotides for which the observed frequency is significantly different than the expected frequency (show the results on input file chrB.fasta). [10 pts]

2. Devise a Hidden Markov Model (Q, Σ, π, e, T) to detect CpG islands in a DNA sequence. For CpG islands, we can use a first-order Markov model with four states (one for each nucleotide) where the transition probabilities correspond to the dinucleotide frequencies. Similarly, we can construct a first-order Markov model for sequence outside CpG islands. Therefore, the overall HMM has a total of 8 states (4 each for a CpG island and non-CpG island). You can assume that $\pi = 0$ for CpG island states. Assume that in CpG islands, the frequency of CG is 0.06 and all other dinucleotides are equally likely. Similarly, outside of CpG islands, the frequency of CG is 0.01 and all other dinucleotides are equally likely. Also assume that the probability of transitioning from a CpG island to non-CpG island is small (0.005) and similarly, the probability of transitioning from a non-CpG island to a CpG island is 0.001. [15 pts]

3. Implement the Viterbi algorithm for finding the most likely sequence of hidden states for a DNA sequence using the above HMM. Analyze the input file (chrB.fasta) using the Viterbi algorithm to identify CpG islands. Output the locations of the predicted CpG islands to a text file with the start and end coordinates of each CpG island (one per line). [30 pts]

4. The transition and emission probabilities of an HMM can be estimated using training data, i.e. data for which the hidden states or labels are known. Consider a DNA sequence S for which we have labeled each nucleotide as I (CpG island) or O (non-CpG).

   ```
   S =    GGGACTACCACTCACGCAGAGCCAATCAGAACTCGCGGTGGGGGCTGCTGGTTCTTCCAG
   L =    OOOOOOOOOOOOOOIIIIIIIIIIIIIIIIIOOOOOOOOOOOOIIIIIIIOOOOOOOOOO
   ```

   Given the labels, we can calculate the transition probability $T[I, I]$ (of staying in a CpG island) as

   $$\frac{n_{II}}{n_{II} + n_{IO}}$$

   where n_{II} is the number of positions i such that $L[i] = I$ and $L[i+1] = I$. Similarly, we can calculate $T[O, O]$ and other transition and emission probabilities. Use the location of the CpG islands for the DNA sequence provided in the fasta file "chrA.fasta" to estimate the transition and emission probabilities for the HMM. [25 pts]

5. Re-run your HMM with the parameters estimated in (4) on the input file (chrB.fasta) to identify the locations of the CpG islands using the Viterbi algorithm. Do the results differ compared to the labels obtained in (3) ? Report the number of CpG islands that are shared between the two outputs and the number that differ. [20 pts]

Input files

- chrA.fasta: DNA sequence for which the locations of the CpG islands are known
- chrA.islands: locations (start & end) of the CpG islands in chrA.fasta
- chrB.fasta: DNA sequence for which we want to predict the CpG islands