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.

Problem set I

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 chrA.fasta).

2. Utilizing the dinucleotide frequencies from step 1, design and implement a first-order Markov model with four states (one for each nucleotide) where the transition probabilities correspond to the dinucleotide frequencies. Similarly, implement a first-order Markov model for sequence outside CpG islands. Your code should take as input a string S, begin and end coordinates (b, e) of a substring on S, and compute the CpG potential score.

$$\text{CpG potential} = \log\left(\frac{\Pr^{\text{CpG}} S[b,e]}{\Pr^{\text{non-CpG}} S[b,e]}\right)$$

   Run your code on the training data (chrA.fasta) provided and predict CpG islands based on the CpG potential being positive or negative. Provide statistics on how your answers differ from or match the original labeling of training data (chrA.islands), by publishing 'true positive,' 'false positive,' and 'false-negative' predictions of CpG islands.

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

Problem Set II

3. For each of the 5 data-sets (SNP matrices with n individuals/rows, and m sites/columns) provided, determine if a perfect phylogeny exists or not. You have to write code to do this.

4. Plot the running time of your code. The y-axis should be the time, and x-axis should be the number of entries in the matrix (size of the input). What function best satisfies the running time as a function of the input size?