Ben Siranosian
CSCI1820 hw7

## 1.1 CPG ISLANDS

1) There should be a small but finite chance of transitioning from negative to positive states and vice-versa; the probabilities should be based on biological data. The information we need is the average length of an "ocean" and an "island" in our organism of choice. We want to model the HMM so the expected time spent in each state is equal to the average biological length of each feature.
The CpG island annotation of the 50kb region of chromosome 1 identified 7 islands occupying a total of 16kb. The average length of an island is about 2.29kb. The average length of an ocean is about 5.67kb. Therefore, the probability for transitioning from a given positive state to any negative state is 1/2290 = 0.000437, from a given negative state to any positive is 1/5670 = about 0.000176. These probabilities will be divided by 4 to be distributed among the four between group transition options a given state has.

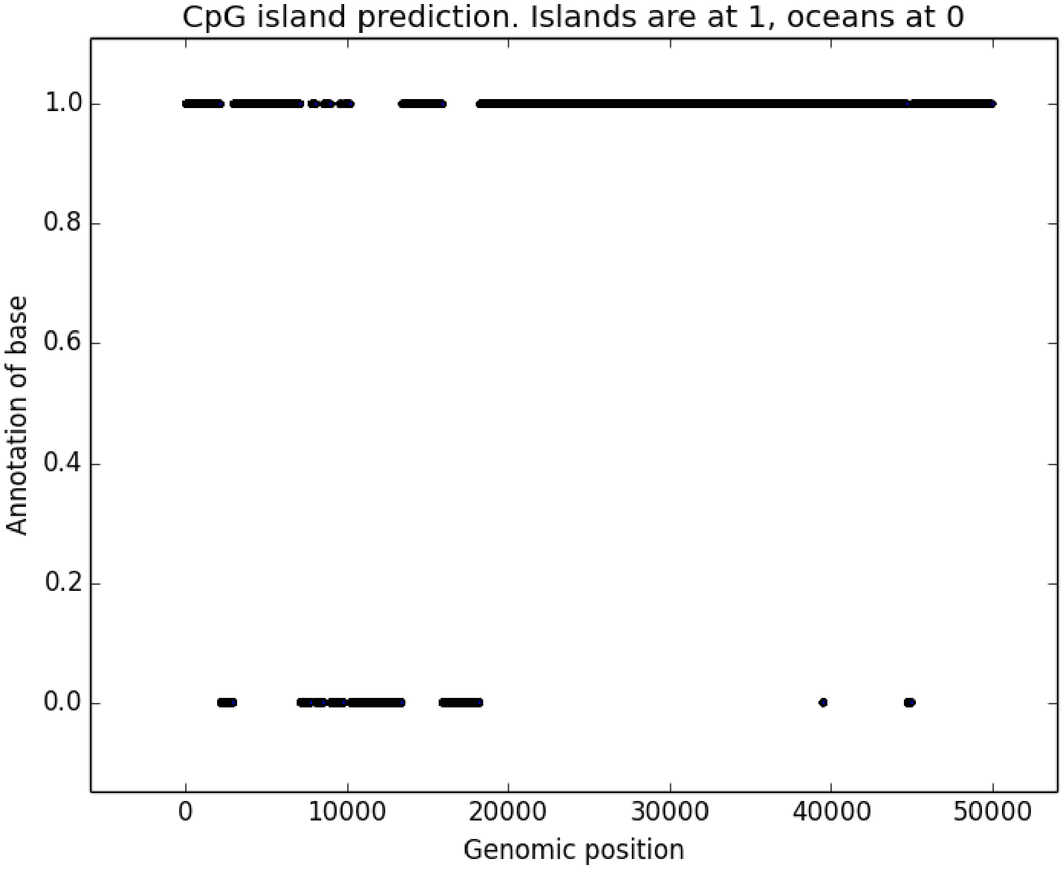
The initial probabilities should be based directly off the probability any given base is in that feature:

P(initial island) = 16/50 = 0.32 P(initial ocean) = 34/50 = 0.68

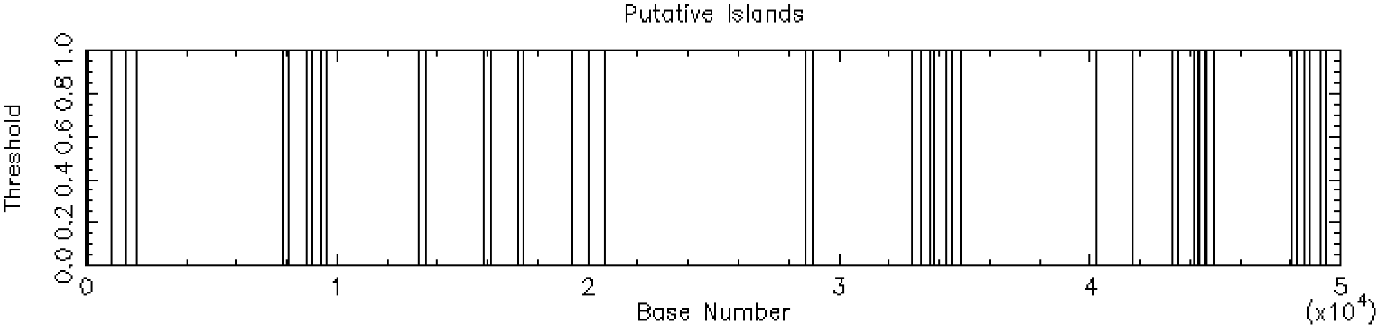
The between state transition probabilities can be added to form an 8x8 state transition matrix. The matrix should ideally be normalized so that the row sums are 1, but that changes the numbers so slightly I chose to keep them the same.

| | **A+** | **C+** | **G+** | **T+** | **A-** | **C-** | **G-** | **T-** |
|---|---|---|---|---|---|---|---|---|
| **A+** | 0.180 | 0.274 | 0.426 | 0.120 | 0.00011 | 0.00011 | 0.00011 | 0.00011 |
| **C+** | 0.171 | 0.368 | 0.274 | 0.188 | 0.00011 | 0.00011 | 0.00011 | 0.00011 |
| **G+** | 0.161 | 0.339 | 0.375 | 0.125 | 0.00011 | 0.00011 | 0.00011 | 0.00011 |
| **T+** | 0.079 | 0.355 | 0.384 | 0.182 | 0.00011 | 0.00011 | 0.00011 | 0.00011 |
| **A-** | 0.000044 | 0.000044 | 0.000044 | 0.000044 | 0.300 | 0.205 | 0.285 | 0.210 |
| **C-** | 0.000044 | 0.000044 | 0.000044 | 0.000044 | 0.322 | 0.298 | 0.078 | 0.302 |
| **G-** | 0.000044 | 0.000044 | 0.000044 | 0.000044 | 0.248 | 0.246 | 0.298 | 0.208 |
| **T-** | 0.000044 | 0.000044 | 0.000044 | 0.000044 | 0.177 | 0.239 | 0.292 | 0.292 |

2) Code for the HMM model I implemented can be found in Viterbi.py
My algorithm performs well on test cases I gave it. For example, in a sequence of a hundred bases with obvious island/ocean transitions, it clearly picks the proper state. It also correctly identifies that the sequence starts in an island state. However, when applied to the entire 50kb segment, it makes transitions more frequently than expected and annotates much of the segment as an island. This is kind of annoying because Viterbi is such a clear cut and well defined algorithm, and I can't find any errors in my code (surprise, surprise). An annotation of the 50kb segment is reproduced below.



3) My algorithm gave much different results when compared with the GenBank annotation. However, other bioinformatics programs also differed from the GenBank data. For



example, EMBOSS Cpgplot run with default parameters identified many islands, typically smaller than those in the GenBank annotation. See the figure below.

## 1.2 HMM MODELING

1) A first order HMM could be used to differentiate between helices and strands. The model would perform best if the amino acid frequencies or transitions between amino acids are different in the stand helix state vs the strand state. It would also be useful if helices and strands had predictable lengths.
   A HMM for this would be similar to the model for CpG islands. 20 states would be used for being in the helix and 20 for being in the strand. Transition probabilities within the group could be calculated from empirical protein data. Between group transition probabilities and initial probabilities should be based on the lengths and frequencies of each feature, just like CpG islands.

   A sheet has many strands stacked next to each other. However, these strands don't always occur next to each other in the amino acid sequence. According to Wikipedia, the most reliable source on the planet, "individual strands can also be linked in more elaborate ways with long loops that may contain alpha helices or even entire protein domains." Therefore, a first order HMM could not be used to infer structure including sheets. A higher order HMM could be used to infer structure about the simple sheet motifs, like the beta hairpin (Wikipedia again), "in which two antiparallel strands are linked by a short loop of two to five residues, of which one is frequently a glycine or a proline." It would be likely that an amino acid is part of a beta hairpin if it is part of a strand and the residue 3, 4, or 5 positions before it is part of a strand. Higher order HMMs may not be able to identify the more complicated structural motifs.