Final Project

Josh You

Identifying CpG islands (CGIs) is an important problem within computational biology. CpG islands are sections of a genome that have an unusually high number of CG pairs, the "p" in CpG referring to the phosphodiester bond between C and G. Normally, CG is by far the least frequent nucleotide pair because the C in CG pairs tends to become methylated, causing it turn to a T. However, in certain areas methylation is suppressed, leading to a high proportion of CG pairs relative to the rest of the genome. In mammalian genomes, CGIs are associated with the 5' end of genes, making them an important marker for finding genes.[1] This makes finding CGIs an important sub-problem in gene prediction.

Gardiner-Garden and Frommer (1987) defined a CpG island as a region of at least 200 base pairs with a GC proportion of at least 0.5 and a CG count that is at least 0.6 of what would be expected by random chance (meaning that CG counts can still be lower in CGIs than a gene sequence where all pairs are equally likely).[2] Given this definition, a straightforward way to find CGIs would be to search a genome using a 200 base-pair window (i.e. search the first 200 bases, then the first through 201st base…) and note which of these sequences meet the above criteria. However, this approach can inadequate since the definition relies on strict cutoffs that can be arbitrary. Researchers have identified genome data containing differentially methylated regions, or parts of a genome with different methylation patterns than the rest of the genome, that are

[1] Han, Leng et al. "CpG Island Density and Its Correlations with Genomic Features in Mammalian Genomes." *Genome Biology* 9.5 (2008): R79. *PMC*.

[2] Gardiner-Garden, M., and M. Frommer. "CpG islands in vertebrate genomes." *Journal of molecular biology* 196.2 (1987): 261-282.

important identifiers but fail to meet the 0.6 threshold[3]. On the flip side, the statistical definition can lead to false positives by identifying regions with small clusters of CpGs and large stretches that are CpG-depleted as CGIs.

An alternative method is searching for CGIs is through the use of hidden Markov models (HMM). An HMM is a probabilistic model defined by the following: a set of outputs, a set of states that are hidden from the observer, a probability distribution for each state that indicates the probability that the state produces each output, and a state transition matrix with the probability that each state will transition into every other state (including itself) at each point. In addition, an initial probability distribution is defined that indicates the probability that the HMM will begin in each state. In my implementation, this matrix was independent from the state transition matrix.

I implemented an HMM with eight states: {A+, C+, G+, T+, A-, C-, G-, T-}, with "+" indicating that the state is part of a CGI, and "- " indicate that it is not. Each of these states outputs their corresponding base pair with probability 1. I used state transition values from Durbin, Eddy, Krogh, and Mitchison (1998)[4], who calculated the transition probabilities empirically from previously identified CGIs in the human genome. For the initial probability distribution, I used data from Han et. al. (2008)[5] who estimated that CGIs comprised around 1.4% of the human genome, and I assumed that each base within each category (+ or -) was equally likely. The next step is to identify the most likely series of states the model went through

[3] Wu, Hao et al. "Redefining CpG Islands Using Hidden Markov Models."*Biostatistics (Oxford, England)* 11.3 (2010): 499–514. *PMC*.

[4]Durbin, Richard, Eddy, Sean, Krogh, Anders, Mitchison, Graeme. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998.

[5] Han, Leng, et al. "CpG island density and its correlations with genomic features in mammalian genomes." *Genome Biol* 9.5 (2008): R79.

to generate a given set of observations. This is done through a dynamic programming algorithm called the Viterbi algorithm (look at my code for details).

Finally, I tested my program on some sample genome data, and I compared results with an online CpG island finder that uses the simple approach of searching a given sequence for areas that meet the statistical definition of a CpG.[6] One of the sequences was a human CGI sequence acquired from the NCBI[7], and the others were randomly generated using the UC Riverside online tool. The second randomly generated sequence used a GC content parameter of 0.4, the others were uniform. All the sequences were approximately 1000 bases long. The table below indicates the start and end indices of any CGI islands identified by the two programs:

	HMM	CPG Island Finder
NCBI genome sequence	(491, 1033)	(306, 1005)
Random sequence 1	(461, 633)	(12, 987)
Random sequence 2	None	None
Random sequence 3	None	(1, 977)

The overall trend is that the HMM was stricter in identifying CGIs than the online tool. This effect was much more pronounced for the randomly generated data: my guess is that since the HMM parameters are more tailor to how CGIs actually look like in real genome data, the HMM is less likely to identify portions of random data that have a high CG count as a CGI, whereas the naive statistical procedure is probably not sensitive to any differences between

[6] "DataBase of CpG Islands & Analytical Tools." From Research Center for Medical Excellence, National Taiwan University. http://dbcat.cgm.ntu.edu.tw/
[7] http://www.ncbi.nlm.nih.gov/nuccore/4220551?report=fasta

randomly generated DNA sequences and actual DNA sequences. I wanted to find a sequence where the HMM identified CGIs that the simple technique rules out, but I had no such luck.