Using Markov Chains to Identify Sequences as CpG or non CpG Islands

Assignment 2

Caitlin Ross
Computer Science Department, Rensselaer Polytechnic Institute
rossc3@rpi.edu

ABSTRACT

This work looks at the use of Markov chains to model CpG island regions. Previously identified CpG islands from human chromosome 12 and random non-CpG island sequences from the same chromosome are used to determine the transition matrices for the Markov model. The model is tested using test sets collected from the same data sets. To test the model, each test sequence has its log odds ratio calculated. Based on a histogram plot of the log odds ratios, the model appears to model the sequences as expected, with the CpG test sequences having a positive log odds ratio, while the non-CpG test sequences tend to have negative log odds ratios. Finally the log odds ratios of 10 unknown sequences are calculated in order to determine whether it is likely for the sequence to be a CpG island or not.

1. PROBLEM STATEMENT

The problem studied here is to use Markov chains to model CpG island regions. To do this, we need a set of previously identified CpG islands and a full chromosome sequence. The CpG sequences will form a positive training set for our model, while randomly chosen sequences from the chromosome sequence will form the negative training set for the model. The two training sets are used to find their respective group's transition matrix. At this point the positive and negative sets' transition matrices can be used to calculate the log odds ratio of an unknown sequence in order to determine whether this sequence is likely to be a CpG island or not.

2. METHODS

This section first discusses the relevant mathematics to this work, followed by a description of the program developed.

2.1 Relevant Mathematics

A Markov chain consists of states that undergo transitions from one state to another. Each transition is assigned some probability, which depends only on the current state and not any previous states. In this model, there are 16 transitions for each pair of bases. From textbook 'Biological Sequence Analysis' by Durbin et al., the transition probability is defined as

$$a_{st} = P(x_i = t | x_{i-1} = s) \tag{1}$$

(i.e. a_{st} is the probability that t follows s in the sequence).

The probability of a given sequence x, is

$$P(x) = P(x_1) \prod_{i=2}^{L} a_{x_{i-1},x_i} \tag{2}$$

where L is the length of the sequence and x_i are letters in the sequence.

To find the transition probabilities for the positive training model, the following equation was used:

$$a^{+}_{st} = \frac{c^{+}_{st}}{\sum_{t'} c^{+}_{st'}} \tag{3}$$

where c^{+}_{st} is the number of times that t follows s in the sequence and t' represents any letter in the alphabet used. Equation (3) is also used similarly to find a^{-}_{st} for the negative training set.

To compute the log odds ratio of a given sequence once the transition matrices are found, the following equation is used:

$$\sum_{i=2}^{L} \log \frac{a^{+}_{x_{i-1},x_i}}{a^{-}_{x_{i-1},x_i}} \tag{4}$$

where $x_i \in A, C, T, G$.

2.2 Program

The program uses the files `CpG` and `chr12.fa`. The file `CpG` was downloaded from http://genome.ucsc.edu and contains 1211 FASTA sequences. These sequences are from CpG islands found in the human genome on chromosome 12 using the December 2013 assembly. 'chr12.fa' is a FASTA file downloaded from http://hgdownload.cse.ucsc.edu that contains the full DNA sequence for human chromosome 12. These two files are used to build the Markov chain model and then to test that the model is correct. The program also uses the file `test_sequences.fasta`, downloaded from the course website. This file is another FASTA file, that contains 10 sequences from different mammalian genomes. The Markov model is used to identify the sequences in this file as likely or unlikely to be CpG islands.

The program uses Biopython to read the FASTA files. The function `pos_samp()` reads the `CpG` file to get the sequences. It then randomly chooses 200 sequences to return as the positive training set, and the other 1011 sequences as the positive test set.

The program then creates the negative training and test sets from the chromosome 12 DNA sequence from `chr12.fa`. The function `random_seq()` takes in the chromosome DNA sequence, and the length of the sequence the function should return. Since the CpG sequences are a small portion of the chromosome 12 sequence, choosing sequences from random in the chromosome will be likely to return non-CpG sequences. For every CpG sequence in the positive sets, the length is used to choose the size of the sequences in the negative sets.

The program then uses the positive and negative training sets to find the transition matrix for each set for the Markov model using equation (3), as explained in Section 2.1. The program uses the function `calc_trans_matrix()`, which takes in one of the training sets to calculate its transition matrix.

Now that both transition matrices are found, they can be used to find the log odds ratio in equation (4) of an unknown sequence. First we use the positive and negative training sets to test the correctness of the transition matrices. Each sequence in these sets is passed to the function `calc_log_odds_ratio()`, which calculates the log odds ratio of the sequence and normalizes the value by dividing by the length of the sequence. Then a histogram of the counts of the log odds ratio is plotted.

After this, the program reads in the sequences contained in the `test_sequences.fasta` file. The normalized log odds ratio of these sequences are found. If the normalized log odds ratio is positive, it is considered to most likely be a CpG island, otherwise it's considered unlikely to be a CpG island.

3. RESULTS

Tables 1 and 2 show the transition matrices for the positive and negative sets, respectively. In the positive set's matrix, the C and G nucleotides columns show the highest probabilities. Since this matrix is created from the set of CpG island sequences, it is expected that C and G are more likely to follow another nucleotide than A and T. This trend doesn't hold in the negative training set's transition matrix. One thing to note is that the probability of a G to follow a C is only approximately 5%. So CpGs are rarely found in the random sequences from Chromosome 12.

Figure 1 shows the histogram of the log odds ratios calculated from the positive and negative tests sets. Based on the histogram, the Markov model appears to be correct as the CpG sequences mostly have positive log odds ratios, while most of the ratios for the random chromosome 12 sequences are negative.

Table 1: Positive Training Set Transition Matrix

+	A	C	G	T
A	0.19704	0.26645	0.41745	0.11905
C	0.15636	0.36629	0.28588	0.19147
G	0.16457	0.35469	0.35954	0.12119
T	0.08902	0.36991	0.34713	0.19394

Table 2: Negative Training Set Transition Matrix

-	A	C	G	T
A	0.32785	0.16925	0.23915	0.26374
C	0.35308	0.25461	0.05246	0.33985
G	0.28183	0.21274	0.25446	0.25096
T	0.21691	0.19940	0.24718	0.33652

Table 3: Negative Training Set Transition Matrix

Sequence ID	Log Odds Ratio
0	0.30956
1	-0.13199
2	0.23125
3	0.14662
4	0.17286
5	-0.00964
6	-0.06452
7	-0.39547
8	0.36713
9	0.15661

Table 3 shows the log odds ratios for the 10 sequences contained in `test_sequences.fasta`. Sequences 0, 2, 3, 4, 8, and 9 appear to be CpG islands, based on their log odds ratios. Sequences 1 and 7 appear to be non-CpG islands based on their log odds ratio as well. Sequences 5 and 6 have negative log odds ratios as well, however their ratios are close to 0, so it's not as certain that they are non-CpG islands.

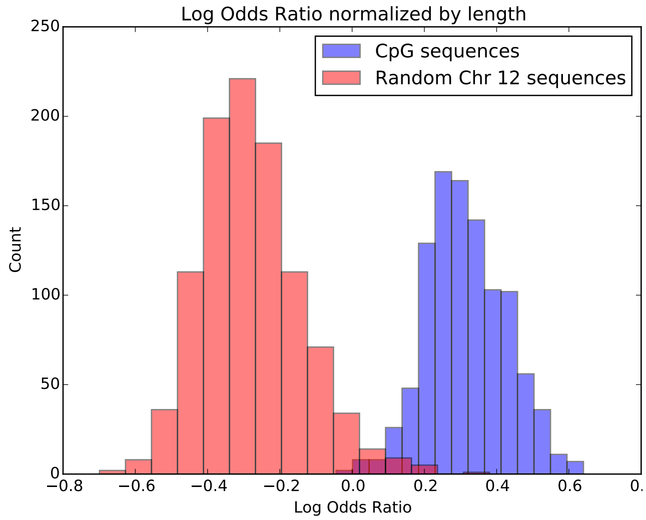

Figure 1: Histogram of log odds ratios of the positive and negative test sets normalized for the length of the sequences