Implementing Gibbs Sampling

Assignment 4

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

ABSTRACT

This work implements a Gibbs sampler in order to find motif sites in a set of given DNA sequences. The program implemented an algorithm for Gibbs sampling given in class. The motif and background models are reported, along with the motif sites. Bar graphs of the probabilities of each sequence position being sampled are also provided for each given DNA sequence. Final the results are compared to output from a Gibbs sampler provided online. Based on comparison to this output, the Gibbs sampler implemented here appears to be working correctly.

1. PROBLEM STATEMENT

The problem here is to implement a Gibbs sampler using the algorithm given in class. The Gibbs sampler is used to find motif sites in a set of DNA sequences.

2. METHODS

To solve the problem, I wrote a program to implement the Gibbs sampler. It takes in the test.fa file and reads the 10 DNA sequences from it. The program starts off by initializing an array a which has an entry for each DNA sequence given. Each element in a is randomized to some index in the sequence, not including the last 8 positions, as 8 is the size of the motifs being considered.

The program then goes into a loop for the burn in and sampling loops. As these loops are very similar, they are called using the same function. The outer loop is some number of iterations; 1000 for a burn in loop and 2000 for sampling loop. The inner loop goes through each sequence. The a value for the current sequence is set to 0. Then θ_{ij} is calculated for the motif and θ_{Bj} is calculated for the background.

$$\theta_{ij} = \frac{n_{ij} + \alpha_{ij}}{\sum_l (n_{il} + \alpha_{il})} \tag{1}$$

$$\theta_{Bj} = \frac{n_{Bj} + \alpha_{Bj}}{\sum_l (n_{Bl} + \alpha_B)} \tag{2}$$

Both θ values are calculated using the other 9 sequences.

After this, the probabilities of x_{ij} given either the motif or background models can be calculated. These are defined as

$$P(x_{ij}|\theta_m) = \prod_{k=1}^{w} \theta_{kb_j} \tag{3}$$

$$P(x_{ij}|\theta_B) = \prod_{k=1}^{w} \theta_{Bb_j} \tag{4}$$

Then we use these probabilities to calculated the values in the array r:

$$r_{ij} = \frac{P(x_{ij}|\theta_m)}{P(x_{ij}|\theta_B)} \tag{5}$$

Then the elements of r must be normalized.

Finally, the function ends with setting a[i] for sequence i to a random number based on the weights given by array r. If the loop is the sampling loop, the count for matrix c in the ith, a[i]th position is then incremented. On the last iteration of the sampling loop, the values from θ_{ij} and θ_{Bj} are collected, in order to aggregate their values for all 5 chains.

After all chains have finished running, the values for the aggregated θ_{ij} and θ_{Bj} values are normalized and printed. Then the counts in matrix c are converted into their proportions and any motif starting point greater than 0.5 is printed with its motif sequence, position number, and probability. Finally for each sequence, the probabilities for each position being sampled is graphed.

3. RESULTS

The motif model is shown in Table 1 and the background model is shown in Table 2. Looking at the largest probabilities in each row results in a motif sequence of ATAATTAT. This matches most of the motifs that are shown in Table 3. The motifs in table 3 are the motifs with the highest probability found in each sequence.

Figures 1 through 10 are bar plots that show the probability of each sequence position being sampled. In some sequences, such as sequence 1 the starting point of the motif is very certain. However for many of the sequences, there are multiple motif sites with high probabilities found.

One final step done was to compare the program output

Figure 1: Probabilities for Sequence 1

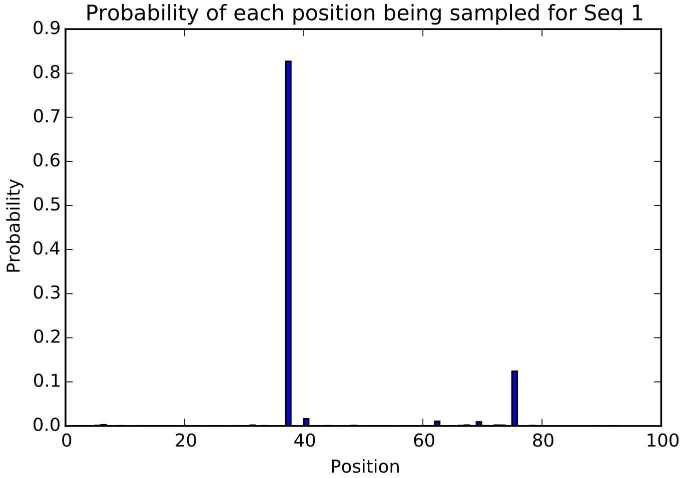
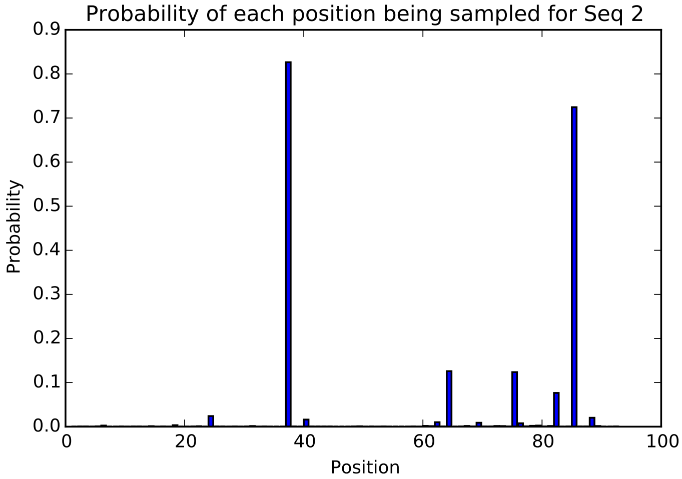


Figure 2: Probabilities for Sequence 2



with the output from http://ccmbweb.ccv.brown.edu/cgi-bin/gibbs.12.pl?data_type=DNA. The output from the website is shown in Table 4. As can be seen when comparing Tables 3 and 4, the results match exactly.

Table 1: Motif Model

A	C	G	T
0.7692	0.0769	0.0769	0.0769
0.0769	0.0769	0.0769	0.7692
0.6769	0.07692	0.1692	0.0769
0.7184	0.1169	0.0876	0.0769
0.0892	0.1323	0.0769	0.7015
0.0769	0.0769	0.1199	0.7261
0.7584	0.0769	0.0876	0.0769
0.0769	0.0907	0.1015	0.7307

Table 2: Background Model

A	0.2938
C	0.2201
G	0.2420
T	0.2440

Table 3: Motif Sites

Starting Position	Motif	Probability
37	ATAACTAT	0.8266
85	ATAATGAT	0.7246
90	ATAATTAT	0.9876
66	ATGATTAT	0.8546
16	ATAATTAT	0.9834
10	ATAATTAT	0.9789
28	ATAATTAT	0.9915
64	ATAATTAT	0.9938
36	ATAATTAT	0.9524
85	ATAATTAT	0.974

Table 4: Motif Sites

	Starting Position	Motif
seq 1	37	ATAACTAT
seq 2	85	ATAATGAT
seq 3	90	ATAATTAT
seq 4	66	ATGATTAT
seq 5	16	ATAATTAT
seq 6	10	ATAATTAT
seq 7	28	ATAATTAT
seq 8	64	ATAATTAT
seq 9	36	ATAATTAT
seq 10	85	ATAATTAT

Figure 3: Probabilities for Sequence 3

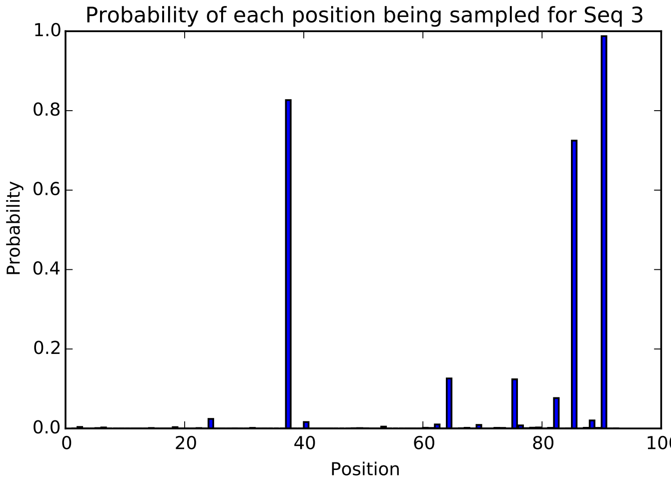


Figure 4: Probabilities for Sequence 4

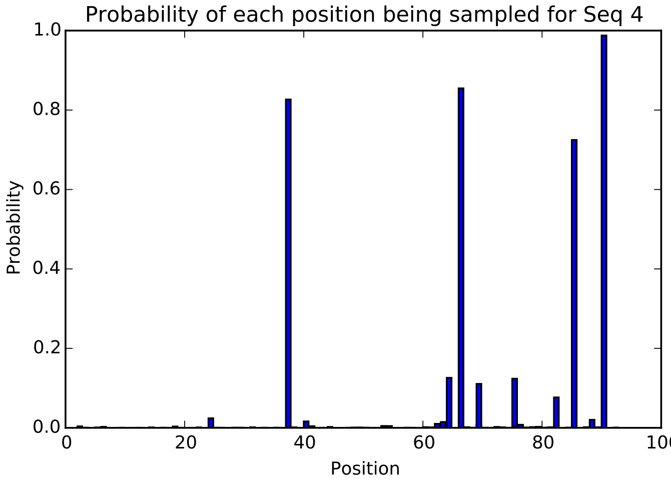


Figure 5: Probabilities for Sequence 5

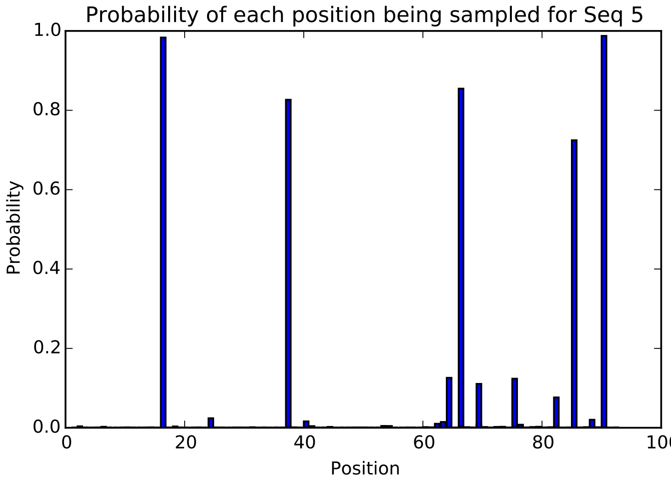


Figure 6: Probabilities for Sequence 6

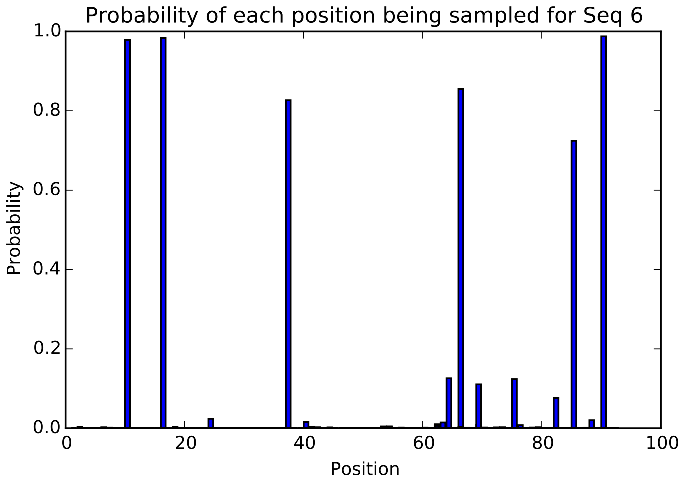


Figure 7: Probabilities for Sequence 7

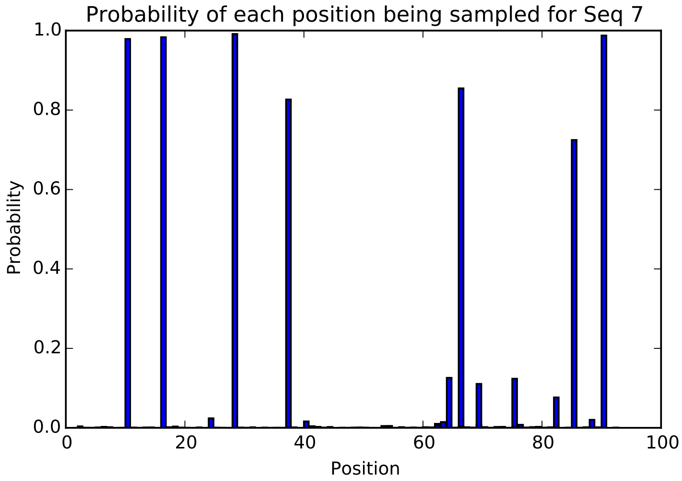


Figure 8: Probabilities for Sequence 8

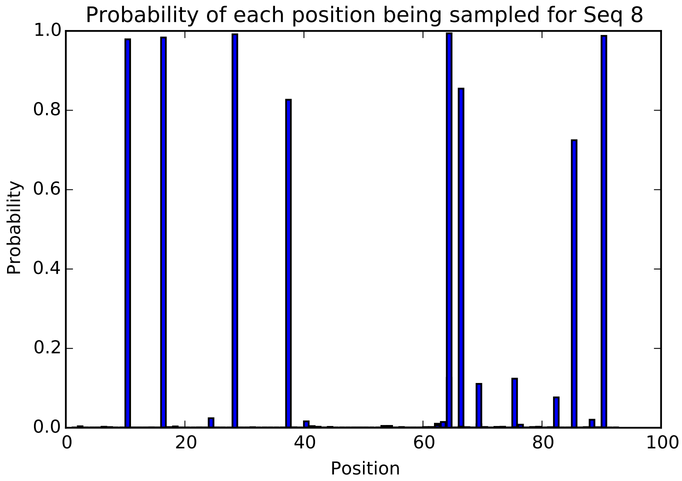


Figure 9: Probabilities for Sequence 9

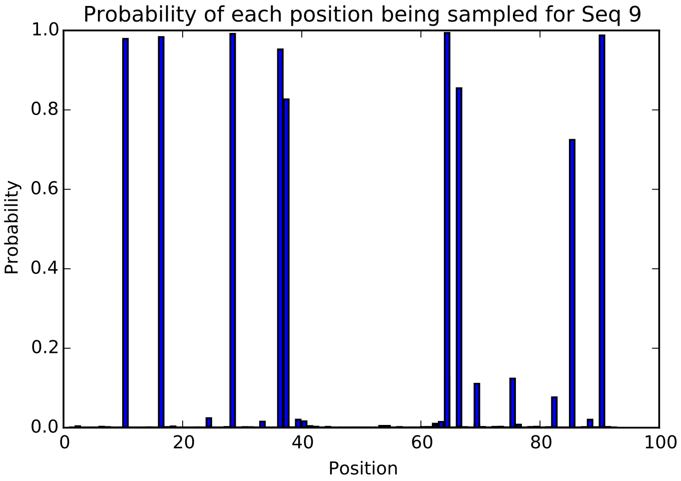
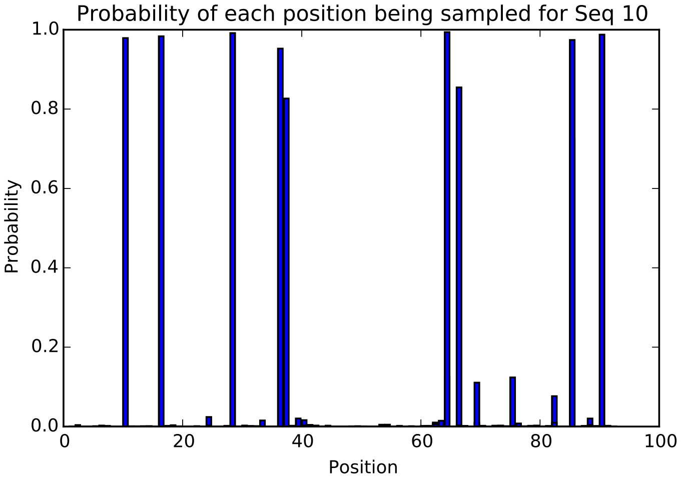


Figure 10: Probabilities for Sequence 10


Probability of each position being sampled for Seq 10
Probability
Position