6.047/6.878/HSPH IMI.231/HST.507 Fall 2018
Problem Set 3: Motifs and RNA Structures

Due Monday, October 22 at 11:59pm

Please submit a zip file of your solution file via Stellar, as in previous psets.

1 Gibbs sampling for motif discovery

In this problem, you will implement a Gibbs sampler to discover sequence motifs. We have provided a Python skeleton `gibbs.py`. Submit all code you write.

(a) Recall the Gibbs sampling algorithm for this problem: Initialize the motif position in each sequence. Until convergence: re-estimate the position weight matrix (PWM) from all the motifs except one, score every position in the excluded sequence, and sample a k-mer from the excluded sequence with probability proportional to the score.

We have intentionally not specified many of the implementation details. Describe and justify the design decisions you made in your implementation. For example, how do you choose the sequence to exclude when recomputing the position weight matrix?

(b) We have provided four test cases. `data1` is a synthetic data set where the motif is identical across the sequences. `data2` is a synthetic data set with a degenerate motif. `data3` and `data4` are yeast transcription factor binding sites of *ACE2* and *MBP1*, respectively.

Run your Gibbs sampler on the test data to discover motifs of length 10. You will need to repeat this procedure several times on each data set due to the stochastic nature of Gibbs sampling.

Submit plain text files containing the most consistently found PWM for each sequence. Use Weblogo[1] to create a *sequence logo* from each PWM and include them in your writeup.

2 Evolutionary signatures of motifs

In this problem, you will search for enriched (over-represented) k-mers in regions conserved across the yeast clade *Saccharomyces*. Submit all code you write.

(a) We have provided the sequence of all intergenic regions in *S. cerevisiae* in the file `allinter`. We have also provided an annotation of conservation in the file `allintercons`. Each position marked with $*$ corresponds to a conserved nucleotide. For simplicity, we will look for motifs which are non-degenerate, exact matches.

Compute the frequency and conservation of all 6-mers. Submit a plain text file with the 50 most frequently occurring and 50 most conserved motifs (those with the highest proportion of conserved instances).

(b) Compare frequently occurring motifs to highly conserved motifs. Are there biases in the sequence properties of either class? If so, where does this bias come from?

Which of the two lists should we use to direct further inquiry into yeast transcription factor binding sites? We have provided an annotation of known yeast motifs `yeast_motifs.txt`. Which known motifs does your scan of 6-mers find?

[1] `http://weblogo.threeplusone.com/create.cgi`

3 RNA secondary structure

In this problem we will explore the output of the Nussinov algorithm on random RNA sequences. Submit all code you write.

(a) Implement the Nussinov algorithm, scoring A–U, G–U, and C–G pairs as -1 and all other pairs as 0.

(b) Generate 1000 RNA sequences of length 100 where each base is drawn uniformly at random. What is the average score for these sequences?

(c) How does the score vary as a function of length? (You will need to repeat (b) for various lengths.)

(d) How does the score vary as a function of GC content? Is this function symmetric around GC content equal to 0.5? Why or why not? (You will need to repeat (b) for different distributions from which you draw bases.)

(e) Given an RNA transcript of interest, how should you interpret the score output by the Nussinov algorithm with respect to your observations about its dependence on length and sequence composition? Is there a better way to estimate the effect of these biases on the score?

4 Probabilistic model for transcription factor binding sites (6.878 only)

In this problem we will derive the probabilistic model underlying position weight matrices (PWMs) and use it to study CCCTC-binding factor (*CTCF*) binding sites. CTCF is a conserved zinc-finger protein which binds to thousands of locations in the human genome and acts as an insulator/repressor.

The data provided for this problem comes from Kim *et al*. "Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome." *Cell*. 2007 Mar 23;128(6):1231–45. Submit all code you write.

(a) Explain how to estimate the motif model $\mathrm{M} = [\mathrm{m_{ij}}]$ where $\mathrm{m_{ij}} = \mathrm{P}(\mathrm{position\ i = nucleotide\ j})$.

(b) Describe and justify an algorithm to estimate the background model B. What assumptions does your model make? What are some of its weaknesses?

(c) Recall that a PWM gives the log odds of observing a particular nucleotide at a particular position in the motif model against in the background distribution.

Use your algorithm from (a) to estimate M from `ctcf_binding_site_sequences.txt` and your algorithm from (b) to estimate B from `chr11_region.fa`. Include these distributions in your writeup.

Estimate a PWM for CTCF using M and B and include it in your writeup.

(d) An alternative visual representation of transcription factor binding sites is a *sequence logo* which gives the *information content* at every position (intuitively, how important each position is for protein binding affinity).

Use WebLogo[2] to generate a sequence logo for `ctcf_binding_site_sequences.txt`. Include it in your writeup.

(e) Compare your sequence logo to the published logo `ctcf_motif.jpg`. What could account for any differences?

(f) Discuss the limitations of PWMs as a representation of transcription factor binding sites. What assumptions are made? Do they hold in general?

(g) Because the entries of a PWM are log odds scores, we can score a k-mer by simply adding up the appropriate entries of the PWM.

Convert the published Position Frequency Matrix (PFM) `ctcf_pwm.txt` to a PWM and use it to scan for CTCF binding sites in `chr11_region.fa`. This region flanks the gene insulin-like growth factor 2 (*IGF2*). Plot the scores at every position for each strand and include the plots in your writeup.

(h) Recall that short k-mers frequently occur by chance throughout the genome. Estimate the probability distribution of scores by randomly sampling 1 million 20-mers from chromosome 11 and scoring them using the published PWM. You may want to use the full sequence of chromosome 11 rather than the region we have provided[3]. Plot a histogram of this distribution.

A simple way to use this null distribution to filter out hits that occurred by chance is to only keep hits with $\mathrm{P}(\mathrm{score} > \mathrm{threshold}) < 10^{-5}$. Based on the distribution you estimated, what is the threshold?

In a plain text file, report the location, score, and sequence (be sure to account for strand orientation) of each 20-mer which meets the threshold.

(i) Discuss some limitations of filtering PWM matches in this manner. Does the method by which we sample 20-mers matter? How will the sequence properties of randomly chosen genomic regions affect the answer? Can we choose regions in a more principled way to account for sequence properties? Is it possible to estimate the probability of a PWM match occurring by chance without sampling?

[2] `http://weblogo.threeplusone.com/create.cgi`
[3] `ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/chr11.fa.gz`