Quiz 3&4: Motifs and RNA Structures
შეფასება: 16 ქულა

მოცემულია 3 დავალება, თითოეული ფასდება 6 ქულით + 4 ბონუს ქულა (ჯამურად 22 ქულა). ქვიზის მაქსიმალური შეფასებაა 16 ქულა (რაც გულისხმობს 16 + 6 ბონუს ქულა). სურვილისამებრ შეგიძლიათ, ამოარჩიოთ და შეასრულოთ თქვენთვის სასურველი დავალებები.
ზოგიერთ სავარჯიშოში საჭიროა შეკითხვებზე სიტყვიერი პასუხის გაცემა, რაც სასურველია ატვირთოთ word/pdf ფაილის სახით.
საბოლოოდ ატვირთეთ ყველა სამუშაო ფაილი დაზიპული ფაილის სახით.

1 (6 ქულა + 2 ბონუს ქულა) Gibbs sampling for motif discovery *(Discussed in Lecture 10)*

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) (3 ქულა) 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 ev- ery 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) (3 ქულა) 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.

(c) (**+2 ბონუს ქულა** - არასავალდებულო დავალება) Use Weblogo http://weblogo.threeplusone.com/create.cgi to create a *sequence logo* from each PWM and include them in your writeup.

2 (6 ქულა) Evolutionary signatures of motifs *(Discussed in Lecture 10)*

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) (3 ქულა) 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 frequentlyoccurring and 50 most conserved motifs (those with the highest proportion of conserved instances).

(b) (3 ქულა) Compare frequently occurring motifs to highly conserved motifs. Are there biases in the sequence propertiesof 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?

3 (6 ქულა + 2 ბონუს ქულა) RNA secondary structure (Lecture 7)

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

(a) (2 ქულა) Implement the Nussinov algorithm, scoring A–U, G–U, and C–G pairs as −1 and all other pairs as 0.

(b) (2 ქულა) Generate 1000 RNA sequences of length 100 where each base is drawn uniformly at random. What is the average score for these sequences?

(c) (1 ქულა) How does the score vary as a function of length? (You will need to repeat (b) for various lengths.)

(d) (1 ქულა) 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 drawbases.)

(e) (+2 ბონუს ქულა) 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?