Assignment 11

John Platig
Computational Biology II
Due on May 7, 2025

Introduction

In Assignment 4 you learned about DNA motif matching. Unsurprisingly, sequence motif matching is also commonly used for RNA binding proteins (RBPs). However, their secondary (and tertiary) structures also play an important role in determining an RBP-RNA binding event. In this assignment, you will revisit motif matching, but with a structural twist.

To begin, download the file "vienna_hox6_structure.txt" from the course website.

Question 1

(1 point) Describe how dot-bracket notation is used in RNA secondary structure.

Question 2.1

(2 point) Create a data frame using the "vienna_hox6_structure.txt". Use this to generate a new sequence+structure alphabet where a given base pair can be in one of eight states (ignore the 3'/5' distinction for paired bases). Using this data, what is the probability that a randomly chosen RNA nucleotide is paired? If you ignore paired/unpaired information, what is the probability of observing each of the four nucleotides if you randomly selected a nucleotide?

Question 2.2

(1 point) Use your data frame from Q2.1 to generate a probability that a randomly chosen base is in one of the eight states. Which states stand out as most frequent?

Question 3.1

(2 points)

Calculate the position frequency matrix (PFM) from the following sequences, where p = paired and u = unpaired. However, given the small number of sequences, let's add a pseudo count to the counts matrix, C_{ij}, where i indexes over the sequence position, j indexes over the alphabet, and n indicates the number of states/letters in the alphabet:

$$PFM = \frac{C_{ij} + \frac{1}{n}}{\sum_i C_{ij} + 1}$$

$$\vec{s}_1 = \begin{pmatrix} A.U & A.P & G.U & G.U & G.P \end{pmatrix} \tag{1}$$
$$\vec{s}_2 = \begin{pmatrix} C.P & C.U & C.U & G.U & G.P \end{pmatrix} \tag{2}$$
$$\vec{s}_3 = \begin{pmatrix} A.P & C.U & C.U & U.U & G.P \end{pmatrix} \tag{3}$$
$$\vec{s}_4 = \begin{pmatrix} C.P & G.P & C.U & G.U & G.P \end{pmatrix} \tag{4}$$
$$\vec{s}_5 = \begin{pmatrix} C.U & C.U & G.P & G.U & G.P \end{pmatrix} \tag{5}$$

Question 3.2

(1 point) Using the same sequences above, calculate the PFM, but only use the nucleotide sequence information.

Question 4.1

(2 points) Using the PFMs, calculate the **total** information content* for this new "motif" using the 8 letter alphabet and the 4 letter (nucleotide only) alphabet. Elaborate on the difference in the values. What does this tell us about including structure information in motif scanning? Does it seem useful?

* Start from Eqn. 2 in Stormo, 2000 (PMID: 10812473)

Question 4.2

(1 point) We had to sum over sequence positions in Q4.1 to get our answer. Look back at Stormo, 2000 (PMID: 10812473), and explain what is implied by our assumption to sum over positions.