Name: _____Student ID:_____

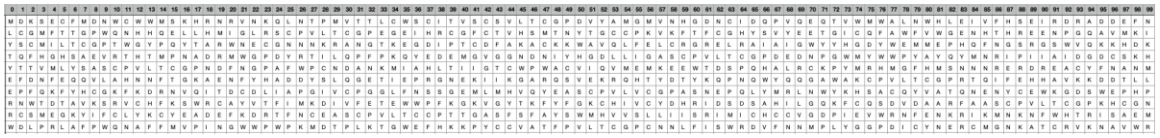
11300A Bioinformatics

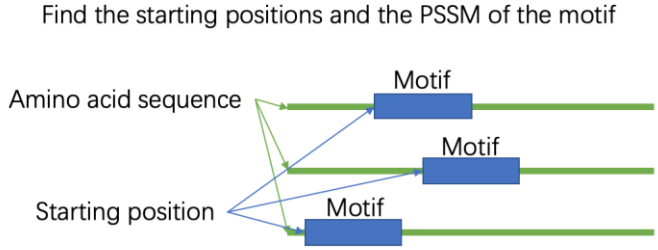
Homework 4

Jan. 10, 2023

1. Identifying protein motifs using Expectation Maximization (EM) algorithm.

Suppose we have 200 protein sequences of length 100, in the dataset. Each character represents an amino acid. The length of the motif is 10. The figure below shows the first 10 samples in the dataset.





Write a program to identify the starting position of the motif in each protein sequence. Write a report which includes: a) your understanding on EM algorithm, b) the core algorithm, c) the final result (starting positions and PSSM of the motif), d) the difficulties you met.

Hint:

1. In short, we need to find the common (or similar) part shared by this group of proteins.
2. How to represent a motif?
   a) We first assume the motifs have a fixed length, and each protein has one motif.
   b) Note that those similar parts are not necessarily to be exactly the same. We need a statistical model, such as the PSSM to describe them. (see PPT Multiple Sequence Alignment Page29-36 and PPT Motifs and domains Page34-36)
3. How to apply EM?
   a) Randomly initialize the starting positions for all the sequences.
   b) Given the positions, you can extract the motif from each protein.

c) Estimate the probability (or score) of each position to be the starting point given the motifs you extracted. (see PPT Motifs and domains Page36)
d) Choose the positions that has the highest probability as the updated starting position.
e) Go to b).