

Assignment 7

What is the main result of the article?

This article presents a statistical method, called the Expectation Maximization method, for finding protein binding sites (motifs) in unaligned DNA sequences. This method assumes each sequence contains at least one instance of the motif.

Briefly explain the Expectation Maximization method.

In the expectation maximization method we are given a set of unaligned sequences and a motif length. We make the assumption that a motif appears once in each unaligned sequence, however, where it appears is unknown. This method estimates the probability that a motif starts at some position j in a given sequence i given an initial guess at a description of the motif. These initial probability estimates are used to reestimate the probability that a letter is in the motif (for all letters in the alphabet – AGTC). This gives us a model of the motif and estimates the probability of each possible starting point of the motif of the sequences in the dataset.

Give an example of how the method works.

We are given a set of DNA sequences and the length of the motif. It is assumed that each sequence contains at least one instance of the motif. We would pass these to the following EM function receive a probabilistic model of what the motif could be.

```
1. EM (dataset, W) {
2.   choose starting point (p)
3.   do {
4.     reestimate z from p
5.     reestimate p from z
6.   } until (change in p < e)
7.   return
8. }
```

What is meant by "unaligned DNA fragments"

Unaligned DNA fragments is when the starting offsets of a motif are not known. Instead the methods described in this article estimate these offsets, which allow us to align the motifs in the DNA fragments.

Explain what the authors mean by their claim that the "reliability of the algorithm increases with the number of fragments but the computations increase only linearly".

Expectation Maximization takes $O(n)$ time where n is the number of subsequences in the data set. However, as you add more subsequences to the data set, the reliability of the probability estimates increases.