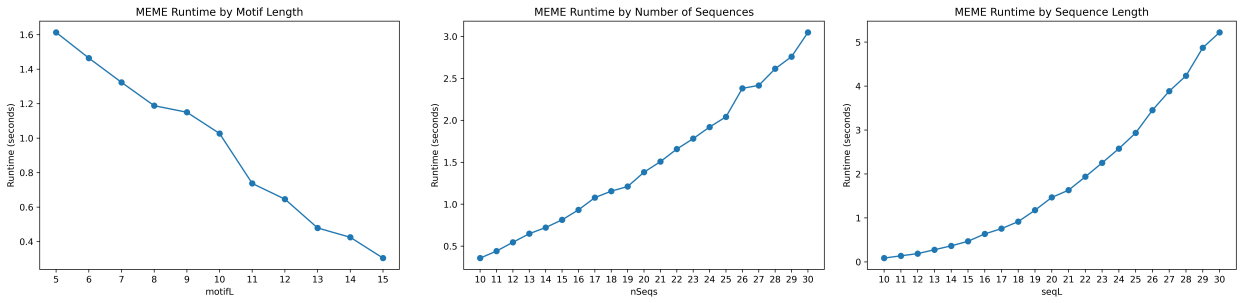
Jeff Dong - kqm018
Project Milestone 2
CS 4223

Project Milestone 2

1. I am still working on the same project as proposed in the first milestone submission. To simply state it again, I will be implementing and experimenting with the different probabilistic modeling approaches for motif finding done in python.

2. Although we did talk about the general idea and implementations of a couple different probabilistic algorithms (more specifically MEME and Gibb's Sampling), it was only at a high level. I am not implementing variations of these well-established algorithms, but to find more exact formulas and details on how to calculate probabilities I have been looking through various other resources on the internet. These resources primarily being lectures on this topic from other universities.

3. As of right now, I have completed the main functions needed to run a basic EM algorithm including functions that can perform the basic E-step and M-step. Additionally, I have implemented functions to get the initial starting point that MEME performs in OOPS mode. Going forward, I intend to implement a ZOOPS mode for MEME as well as the Gibb's Sampling approach. Additionally, creating (or finding existing functions) to visualize the PWM matrix is also on my list of things to do for this project to better visualize the results from my algorithms. Although my current implementation is far from optimized and another goal on my list to do, here are some preliminary charts of the runtime of my MEME algorithm with default values of nSeqs=20, seqL=20, motifL=7, nMuts=1 unless otherwise specified:



4. As currently shown in my figures above, my MEME algorithm is very unoptimized in it's implementation. Although the asymptotic complexity looks about right, the constant factors and implementation details bring the runtime up significantly compared to my PDMF/SDMF implementations in MATLAB. Regardless, it's good that I've been able to at the very least properly implement one of the probabilistic methods for motif finding. Given that I'm currently doing an exhaustive search when initializing the PWM for MEME which comprises most of the runtime, I imagine that the Gibb's Sampling approach will have a much lower runtime relatively speaking.