Stella Wroblewski

14 May 2020

Biological vs. Algorithmic Efficiency in the Motif Finding Problem

My project is aiming to explore the alignments and divergence between biological efficiency and effectiveness and algorithmic runtime. To do so, I decided to implement multiple of the motif finding algorithms described in the textbook including the greedy motif finder (model solution from class), randomized motif search, and Gibbs sampler. I then timed each algorithm on the COVID-19 DNA strand, using timeit, 100 times.

We implemented the greedy motif search in a previous homework. The pseudocode can be seen below.

```
GREEDYMOTIFSEARCH(Dna, k, t)
    BestMotifs ← motif matrix formed by first k-mers in each string
                 from Dna
    for each k-mer Motif in the first string from Dna
        Motif1 ← Motif
        for i = 2 to t
            form Profile from motifs Motif1, …, Motifi - 1
            Motifi ← Profile-most probable k-mer in the i-th string
                     in Dna
        Motifs ← (Motif1, …, Motift)
        if Score(Motifs) < Score(BestMotifs)
            BestMotifs ← Motifs
    return BestMotifs
```

I implemented the randomized motif search from the pseudocode described in the textbook. The pseudocode can be seen below.

```
RANDOMIZEDMOTIFSEARCH(Dna, k, t)
    randomly select k-mers Motifs = (Motif1, …, Motift) in each string
```

```
    from Dna
BestMotifs ← Motifs
while forever
    Profile ← Profile(Motifs)
    Motifs ← Motifs(Profile, Dna)
    if Score(Motifs) < Score(BestMotifs)
        BestMotifs ← Motifs
    else
        return BestMotifs
```

I implemented the Gibbs sampler from the pseudocode described in the textbook. The pseudocode can be seen below.

```
GIBBSSAMPLER(Dna, k, t, N)
    randomly select k-mers Motifs = (Motif1, …, Motift) in each string
        from Dna
    BestMotifs ← Motifs
    for j ← 1 to N
        i ← Random(t)
        Profile ← profile matrix constructed from all strings in Motifs
                except for Motifi
        Motifi ← Profile-randomly generated k-mer in the i-th sequence
        if Score(Motifs) < Score(BestMotifs)
            BestMotifs ← Motifs
    return BestMotifs
```

Project link: https://repl.it/join/clpuwgqp-wroblews

The results were:

Greedy Motif Search time: 867.6189 s

Randomized Motif Search time: 10.3905 s

Gibbs Sampling time: 0.8221 s

These results clearly show that the Gibbs sampler was implemented the most efficiently, while Greedy has significantly the longest runtime of all three.

I am ok having my report posted on the webpage!

All pseudocode taken from:

Compeau, Phillip, and Pavel Pevzner. *Bioinformatics Algorithms: an Active Learning Approach*. 2nd ed., vol. 1, Active Learning Publishers, 2015.