BMI 776

HOMEWORK 1 (0 LATE DAYS)

Noah Cohen Kalafut
Computer Science Doctoral Student
University of Wisconsin-Madison
nkalafut@wisc.edu

February 14, 2021

1 Part 1

Done! (learn_motif.py)

Some optimization measures have been taken for the initial sampling. example_2 matches near-perfectly. example_1 matches the given solution for most positions.

2 Part 2

2.1 Question A

By following the definitions

$$p_{c,k} = \frac{n_{c,k} + d_c}{N - 1 + d_b} \quad p_{c,0} = \frac{n_{c,0} + d_c}{(N-1)(L-W) + d_b}$$

for the given data, using pseudocounts of 1, we can first simplify to[1]

$$p_{c,k} = \frac{n_{c,k} + 1}{10 - 1 + 4} = \frac{n_{c,k} + 1}{13} \quad p_{c,0} = \frac{n_{c,0} + 1}{(10-1)(8-4) + 4} = \frac{n_{c,0} + 1}{40}$$

We will sort rows alphabetically by base (ACGT). For our counts excluding sequence 5, we have[2]

$$n = \begin{bmatrix} 12 & 1 & 5 & 0 & 1 \\ 5 & 5 & 1 & 1 & 3 \\ 12 & 2 & 1 & 7 & 2 \\ 7 & 1 & 2 & 1 & 3 \end{bmatrix}$$

Then,

$$p = \begin{bmatrix} 13/40 & 2/13 & 6/13 & 1/13 & 2/13 \\ 6/40 & 6/13 & 2/13 & 2/13 & 4/13 \\ 13/40 & 3/13 & 2/13 & 8/13 & 3/13 \\ 8/40 & 2/13 & 3/13 & 2/13 & 4/13 \end{bmatrix} \tag{1}$$

[1] Notice that $N = 10$ not 9. This is because the $N - 1$ portion of the equation already accounts for taking out one sample. $d_b = 4$ since d_b is meant to represent the sum of all pseudocounts for the column.

[2] $n_{c,k}$ is the number of times base c appears at position k in the motif. $n_{c,0}$ counts the occurrences of c outside of a motif

2.2 Question B

Using 1 above, we can calculate the likelihood ratios[3]

$$LR(j) = \frac{\prod_{k=j}^{j+W+1} p_{c_k,k-j+1}}{\prod_{k=j}^{j+W+1} p_{c_k,0}}$$

for all possible j.

j	1	2	3	4	5
$LR(j)$	$\frac{\frac{2*2*2*4}{13^4}}{\frac{13*6*6*8}{40^4}} \approx 0.77$	$\frac{\frac{6*2*2*2}{13^4}}{\frac{6*6*8*13}{40^4}} \approx 1.15$	$\frac{\frac{6*3*1*4}{13^4}}{\frac{6*8*13*8}{40^4}} \approx 1.29$	$\frac{\frac{2*6*2*4}{13^4}}{\frac{8*13*8*6}{40^4}} \approx 1.72$	$\frac{\frac{2*3*2*3}{13^4}}{\frac{13*8*6*13}{40^4}} \approx 0.40$

After normalizing over all j, the probability of the motif starting in the second position is approximately 21.58%

2.3 Question C

For a palindromic constraint, we redefine $p_{c,k}$: [4]

$$p_{c,k} = \frac{n_{c,k} + n_{t,W-k+1} + d_c + d_t}{2N - 2 + d_b + d_u} = \frac{n_{c,k} + n_{t,W-k+1} + 2}{26}$$

Then,[5]

$$p = \begin{bmatrix} 13/40 & 6/26 & 8/26 & 4/26 & 4/26 \\ 6/40 & 9/26 & 10/26 & 4/26 & 7/26 \\ 13/40 & 7/26 & 4/26 & 10/26 & 9/26 \\ 8/40 & 4/26 & 4/26 & 8/26 & 6/26 \end{bmatrix}$$

3 Part 3

3.1 Question A

Suppose we extend our previous PWM to $W + w$ width, where $w \in \mathbb{Z}^+$. What then happens to the optimal log likelihood?

3.1.1 The Majority Case

Theorem: Increasing W, where possible, when every motif occurrence can be extended in the same direction, will improve/increase or not affect the optimal log likelihood of a PWM **for responsibly chosen pseudocounts**.

We would expect that adding an additional column to p would improve or replicate the log likelihood of the previous PWM – simply adding detail to the model. This is correct in most cases. However, there is a complication in proving this: The background parameters.

If we can prove that there exists some PWM state for which the log likelihood of the new PWM exceeds or equals that of the old, such would be sufficient to prove our theorem. Further, if we can provide proof for $W + 1$, it follows by induction that the theorem holds for $W + w$.

Suppose that the motif positions are fixed from the previous PWM and that $W < L$, making the new PWM possible. Note that this also makes the background parameters unoptimizable.

The bases now part of the motifs are excluded from the new background calculation. We will refer to these bases as N. Similarly, we will refer to sequences as X_i, the previous motif bases (not including N) as M, and the previous background as B. $m, k \in M_i$ will refer to base m at position k in the original motif M for sequence X_i.

Let's refer to the old weights as p and the new as $\mathbf{p}$.

[3] This can be interpreted as the odds of the subset being a motif versus the product of random chance.

[4] In a real implementation, the algorithm would likely try palindromic and non-palindromic constraints before deciding on one to keep using LR. Also note that c, t are a base pair.

[5] Our constraint is only on the motif, so the background is unaffected.

Let's additionally assume that all motif occurrences are extended in the same direction. Let new bases be in position s of the motif, where either $s = W + 1$ or $s = 0$.

As the log likelihood increases with the likelihood, if likelihood improves, so will the log likelihood. So, let's examine the differences between the likelihood functions of the two PWMs.

$$\text{Old PWM} \rightarrow \prod_{i=0}^{S-1} \left(\prod_{m,k \in M_i} p_{m,k} \prod_{b \in B_i} p_{b,0} \right)$$

$$\text{New PWM} \rightarrow \prod_{i=0}^{S-1} \left(\prod_{m,k \in M_i} \mathbf{p}_{m,k} \prod_{n \in N_i} \mathbf{p}_{n,s} \prod_{b \in B_i - N_i} \mathbf{p}_{b,0} \right)$$

Then, suppose we fix $p_{c,k}$ for all M, shifting k if necessary. We can also set our new weights (either $p_{c,W+1}$ or $p_{c,0}$) equivalent to our old background. Applying these steps sequentially, we are left with

$$\prod_{i=0}^{S-1} \left(\prod_{m,k \in M_i} p_{m,k} \prod_{b \in B_i} p_{b,0} \right) \stackrel{?}{=} \prod_{i=0}^{S-1} \left(\prod_{m,k \in M_i} \mathbf{p}_{m,k} \prod_{n \in N_i} \mathbf{p}_{n,s} \prod_{b \in B_i - N_i} \mathbf{p}_{b,0} \right)$$

$$\prod_{i=0}^{S-1} \left(\prod_{b \in B_i} p_{b,0} \right) \stackrel{?}{=} \prod_{i=0}^{S-1} \left(\prod_{n \in N_i} \mathbf{p}_{n,s} \prod_{b \in B_i - N_i} \mathbf{p}_{b,0} \right)$$

$$\prod_{i=0}^{S-1} \left(\prod_{b \in B_i - N_i} p_{b,0} \right) \stackrel{?}{=} \prod_{i=0}^{S-1} \left(\prod_{b \in B_i - N_i} \mathbf{p}_{b,0} \right)$$

The question becomes: Which background weights better describe the new background? Since the new background weights are calculated using this background, for responsibly chosen pseudocounts, the right-hand side of the last equation above, corresponding to the new PWM, should be greater than or equal to the left. However, this method has caveats. This will be true with the use of no pseudocounts or with pseudocounts that are small relative to the sample size.

Thus, in these cases there exists some possible PWM configuration $\mathbf{p}$ such that the likelihood of the PWM increases or remains the same for $W + 1$. So, the theorem is proven for all $W + w$, $w \in \mathbb{Z}^+$.

Please note that this does not mean that larger W is better. Along with reducing readability and applicability, larger W can run into far more local minima, meaning that the practical likelihood achieved could be larger than smaller W.

3.1.2 The Exception

Now for an example of irresponsibly chosen pseudocounts. Consider

$$\begin{array}{c} \underline{AC}GA \\ \underline{CG}AC \\ \underline{GA}CG \end{array}$$

with $d_c = 1$.

Then, we have

$$p = \begin{bmatrix} 3/10 & 2/7 & 2/7 \\ 3/10 & 2/7 & 2/7 \\ 3/10 & 2/7 & 2/7 \\ 1/10 & 1/7 & 1/7 \end{bmatrix}$$

Choosing $W + 1$, we then have

$$p = \begin{bmatrix} 2/7 & 2/7 & 2/7 & 2/7 \\ 2/7 & 2/7 & 2/7 & 2/7 \\ 2/7 & 2/7 & 2/7 & 2/7 \\ 1/7 & 1/7 & 1/7 & 1/7 \end{bmatrix}$$

Now, consider the likelihood with fixed motif positions

$$\mathbb{P}(D|p_W) = \left(\frac{3}{10}\right)^6 \left(\frac{2}{7}\right)^6 > \left(\frac{2}{7}\right)^1 2 = \mathbb{P}(D|p_{W+1})$$

The irresponsibility here comes from the fact that the pseudocount is too high compared to the sample size. Because of this, fewer samples for the new background data means that the pseudocount affects the data more than is necessary.

3.2 Question B

To use this prior, we calculate the matrix d with the following definition

$$d_{c,k} = \sum_j \mathbb{P}(\alpha^{(j)}|n_k)\alpha_c^{(j)}$$

Using the purines and pyrimidines as rows, we calculate

$$\mathbb{P}(\alpha^{(j)}|n_k) = \begin{bmatrix} 24/36 & 3/9 & 6/9 & 7/9 & 3/9 \\ 12/36 & 6/9 & 3/9 & 2/9 & 6/9 \end{bmatrix}$$

Combining this with the Dirichlet mixture prior,

$$d = \begin{bmatrix} 3 & 2 & 3 & 10/3 & 2 \\ 2 & 3 & 2 & 5/3 & 3 \\ 3 & 2 & 3 & 10/3 & 2 \\ 2 & 3 & 2 & 5/3 & 3 \end{bmatrix}$$

This provides us with

$$p = \begin{bmatrix} 15/46 & 3/19 & 8/19 & 10/57 & 3/19 \\ 7/46 & 8/19 & 3/19 & 8/57 & 6/19 \\ 15/46 & 4/19 & 4/19 & 31/57 & 4/19 \\ 9/46 & 4/19 & 4/19 & 8/57 & 6/19 \end{bmatrix}$$

4 Part 4

4.1 Question A

We are given the following p-matrix

$$p = \begin{bmatrix} 0.256 & 0.104 & 0.032 & 0.027 & 0.057 & 0.053 & 0.213 & 0.076 & 0.204 & 0.968 & 0.964 & 0.457 \\ 0.243 & 0.447 & 0.019 & 0.034 & 0.852 & 0.396 & 0.012 & 0.010 & 0.010 & 0.013 & 0.011 & 0.039 \\ 0.257 & 0.135 & 0.026 & 0.338 & 0.020 & 0.010 & 0.365 & 0.898 & 0.775 & 0.010 & 0.010 & 0.267 \\ 0.244 & 0.314 & 0.922 & 0.601 & 0.071 & 0.541 & 0.410 & 0.016 & 0.012 & 0.010 & 0.015 & 0.237 \end{bmatrix}$$

We may ignore the background column. We can first calculate the entropy of each position

$$\mathbb{H}(C) = -\sum_c \mathbb{P}(c)\log_2 \mathbb{P}(c)$$

$$= [\ 1.774 \quad 0.512 \quad 1.277 \quad 0.816 \quad 1.300 \quad 1.610 \quad 0.584 \quad 0.896 \quad 0.260 \quad 0.280 \quad 1.700\ ]$$

Notice that the maximal entropy for four bases (expected number of bits needed to optimally encode a position) is

$$\mathbb{H}_{\text{max}} = \log_2 4 = 2$$

We can then calculate the decrease in entropy for each position

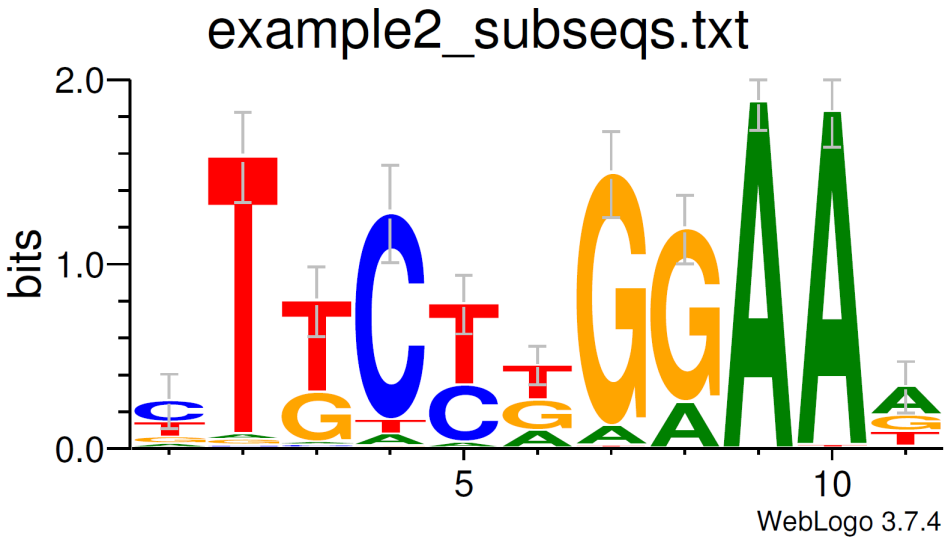
$$\mathbb{H}_{\text{max}} - \mathbb{H}(C) = [\ 0.226 \quad 1.488 \quad 0.723 \quad 1.184 \quad 0.700 \quad 0.390 \quad 1.416 \quad 1.104 \quad 1.740 \quad 1.720 \quad 0.300\ ]$$

Finally, we determine the height for the bases in each position. This is done by simply dividing up the above by frequency, which is conveniently given in the columns of p

	Height										
A	0.024	0.048	0.020	0.067	0.037	0.083	0.108	0.225	1.684	1.658	0.137
C	0.101	0.028	0.025	1.009	0.277	0.005	0.014	0.011	0.023	0.019	0.012
G	0.031	0.039	0.244	0.024	0.007	0.142	1.272	0.856	0.017	0.017	0.080
T	0.071	1.372	0.435	0.084	0.379	0.160	0.023	0.013	0.017	0.026	0.071

4.2 Question B

Done! (Logo.pdf - also shown below)



4.3 Question C

An information content logo can quickly convey the significance of the different positions of the sequence found. A taller stack will tell us that we expect fewer bits to be needed to encode that position. In other words, a large amount of the time, a small number of bases is seen. This small number of bases is shown by the size of characters in the stack. A large number in a large stack will occur by far most frequently in that position. A small stack will occur when the base probabilities are more spread out. These types of stack will not display a clear majority probability and will generally be of less use for identifying the motif.

Using our example above, we can see that the first and last positions are mainly chance. There are trends but they are somewhat insignificant when compared to the rest of the sub sequence. We are fairly certain that the motif will generally obey the pattern _T(T|G)C(T|C)_GGAA_. We could be dealing with two sequential motifs. Perhaps the motif we're looking for is only 9 width. In any case, there is a clear commonality in the sequences given.