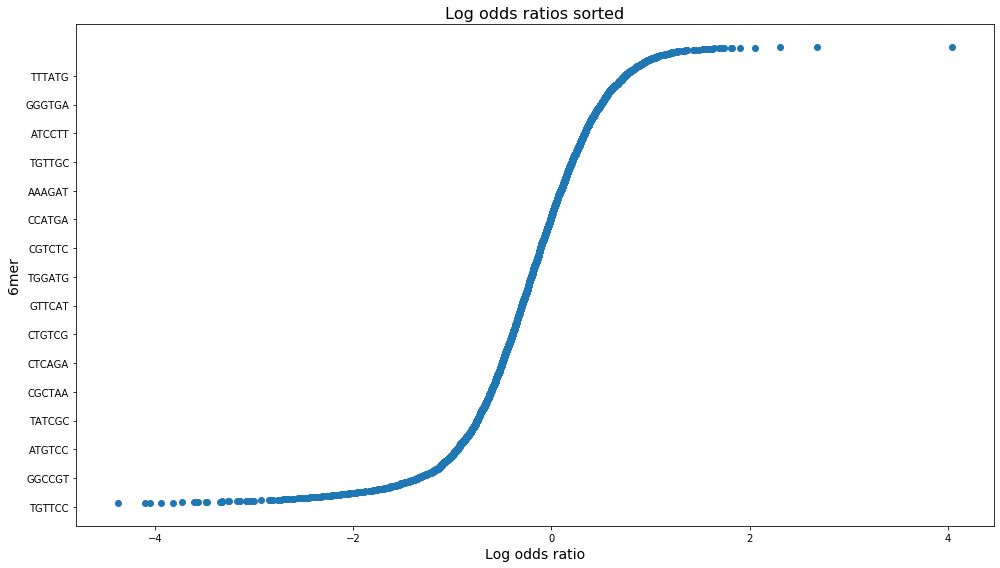
CSE520 Homework 4

Tony Adkins ([anadkins@uw.edu](mailto:anadkins@uw.edu))

# Genomics

## Sorted log odds ratios of 6mers



* Maximum log odds ratio: TCTTTC, 4.03856
* Minimum log odds ratio: CGACCT, -4.37356

The minimum above was calculated excluding the numerous 6mers which had zero probability, resulting in a negative infinity log odds ratio. The minimum of around -4 means the 6mer CGACCT has less than a 2:100 odds of splice usage. The maximum 5mer TCTTTC’s log odds ratio of around 4 means something closer to 55:1, a nearly 3000-fold increase in odds likelihood.

## KL-divergence Gradient Descent

### ∂L/∂wk derivation

Where yi is a constant from the data and is defined as

So, we can take the derivative as the sum of the derivatives of the two parts of the sum, starting with the first half:

Which can be further split into two, allowing us to find the derivatives of each separately because

So, now to find

We must apply the chain rule F’(w) = f’(g(w))g’(w) where g(w) =

Which requires us to find the derivative of g(w) = :

-

Then, to find f’, we find

So, we combine these results to find

Similarly, we can find the derivative for

by applying the chain rule F(w) = f’(g(w))\*g’(w), where g(w) = , which we can rewrite for simplicity as

So,

Then,

Which we then bring back to our original equation to solve

This just leaves us the last part of the original equation to derive, which is simple

Now, to put that all together, we say

So, finally:

### ∂L/∂w0 derivation

With the loss function below, we will now similarly find the derative with regards to

Again, we will start by taking the derivative as the sum of the derivatives of the two parts of the sum, starting with the first half:

Which can be further split into two, allowing us to find the derivatives of each separately because

So, now to find

We must apply the chain rule F’(w) = f’(g(w))g’(w) where g(w) =

Which requires us to find the derivative of g(w) = :

-

Then, to find f’, we find

So, we combine these results to find

Similarly, we can find the derivative for

by applying the chain rule F(w) = f’(g(w))\*g’(w), where g(w) = , which we can rewrite for simplicity as

So,

Then,

Which we then bring back to our original equation to solve

This just leaves us the last part of the original equation to derive, which is even simpler

Now, to put that all together, we say

So, finally:

## Gradient Descent Implementation

Results from the gradient descent implementation

A close up of a logo

Description automatically generatedA close up of a logo

Description automatically generated

The resulting error did not ever grow between iterations before reaching convergence within the λ = 0.0001 limitation, it was always converging.

### 10 most enhancing 6mers, with weights

1. GGGGGG = 0.4529777389211724
2. GGGAGG = 0.332653489646616
3. GGAGGG = 0.28457701160538845
4. GGGGGA = 0.26016371655006026
5. AGGGGG = 0.2593333591585411
6. GGGGGT = 0.24563617235171237
7. TGGGGG = 0.24290842901762238
8. GGGGAG = 0.23059204646574685
9. GGGGAA = 0.20745649695682045
10. AGGGGA = 0.20296590405196308

### 10 most repressive 6mers, with weights

1. CTTGGA = -0.6054054956108216
2. CGGTCG = -0.5133396767128425
3. AGGTCG = -0.5117146171699465
4. CTTGGC = -0.4109720738731291
5. CTTGGT = -0.3935653618870085
6. TGGTCG = -0.35594982435656036
7. ACGGTC = -0.233725673298892
8. AAGGTC = -0.20328247017022158
9. TTGGAG = -0.19159093667071778
10. TTGGTA = -0.18280566786238048

Here we can see a pattern that the top 6mers to increase the odds ratio, or most enhancing motifs, are those with the most guanine (G), with a secondary preference for adenine (A). All are 4 or more G mixed with 1 or 2 A or T.

On the other hand, for most repressive 6mers, there is a pattern that they contain a 2mer of GG often paired with a 2mer of TT or AA.