Assignment 2

BISC577A, Unit 3

Jae Cha

ChIP-seq, which is in vivo experiment, is method that analyzes protein-DNA interactions, measuring specific biological modifications along genome such as detecting DNA-binding proteins, e.g. TF (transcription factor), binding sites. The experimental procedure is almost the same as ChIP-chip except the last step where sequencing is used to replace microarray. Unlike in vitro experiments like SELEX-seq and PBM, ChIp-seq isolate chromatin, use antibody to immunoprecipitate target factor in chromatin, and DNA bound to this factor gets then sequenced, to produce target DNA site library bound to target protein in vivo, eventually revealing which genomic regions were bound by the factor at chromatin isolation. Also, unlike ChIP-chip, the enriched DNA segments is rather sequenced directly than dependent on hybridization in tiling arrays. It doesn't require much starting material and has quite high spatial resolution and good signal to noise ratio as well as dynamic range. But, it's quite costly and needs a lot of tissue, which might be problematic for some types of samples. SELEX-seq is in vitro high-throughput technique that procudes oligonucleotides of ssDNA or RNA that target ligand specifically bind to so as to determine TF-DNA sequence specificity by using classical protein-DNA SELEX assays with massive parallel sequencing. Its latent specificity could be precedent for distinguishment of similar TF specificities between in vivo and in vitro. Also, it could determine relative affinities to DNA sequence for any TF. But, the product should be purified after PCR, such as from dsDNA or other left-overs. Also, its range over readout (to be analyzed) is known to be limited, and prior sequence-based alignment of different oligomers are needed. PBM (protein binding microarray) also can characterize TF sequence specificity in a high-throughput fashion. It can provide whole comprehensive table with preference of particular TF for any sequence variant though it required prior background knowledge of the protein of interest and affinity reagents.

Two different sequence models “1-mer” and “1-mer+shape” are built as prediction models for three different in vitro data of Mad, Max, and Myc.

```
7534 samples
 144 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 6782, 6780, 6780, 6779, 6780, 6781, ...
Resampling results across tuning parameters:

  lambda        RMSE       Rsquared
  3.051758e-05  0.3809422  0.7748270
  6.103516e-05  0.3809422  0.7748270
  1.220703e-04  0.3809422  0.7748270
  2.441406e-04  0.3809422  0.7748270
  4.882812e-04  0.3809422  0.7748270
  9.765625e-04  0.3809422  0.7748270
  1.953125e-03  0.3809422  0.7748270
  3.906250e-03  0.3809422  0.7748270
  7.812500e-03  0.3809422  0.7748270
  1.562500e-02  0.3809422  0.7748270
  3.125000e-02  0.3811609  0.7747944
  6.250000e-02  0.3842917  0.7743338
  1.250000e-01  0.3946565  0.7727535
  2.500000e-01  0.4222592  0.7679309
  5.000000e-01  0.4771609  0.7555835
  1.000000e+00  0.5544751  0.7312023
  2.000000e+00  0.6344850  0.6950211
  4.000000e+00  0.6991414  0.6551491
  8.000000e+00  0.7432415  0.6208768
  1.600000e+01  0.7698562  0.5968907
  3.200000e+01  0.7846970  0.5823424
  6.400000e+01  0.7925893  0.5742313
  1.280000e+02  0.7966512  0.5699719
  2.560000e+02  0.7988001  0.5677692
  5.120000e+02  0.8008122        NaN
  1.024000e+03  0.8008122        NaN
  2.048000e+03  0.8008122        NaN
  4.096000e+03  0.8008122        NaN
  8.192000e+03  0.8008122        NaN
  1.638400e+04  0.8008122        NaN
  3.276800e+04  0.8008122        NaN

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using  the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.015625.
> average_Rsquared5 <- mean(na.omit(model5$results$Rsquared))
> head(average_Rsquared5)
[1] 0.71613
```

Figure 1. L2-regularized MLR model for “1-mer” on Mad

```
7534 samples
 274 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 6780, 6781, 6781, 6779, 6782, 6782, ...
Resampling results across tuning parameters:

  lambda        RMSE       Rsquared
  3.051758e-05  0.3014281  0.8634341
  6.103516e-05  0.3014281  0.8634341
  1.220703e-04  0.3014281  0.8634341
  2.441406e-04  0.3014281  0.8634341
  4.882812e-04  0.3014281  0.8634341
  9.765625e-04  0.3014281  0.8634341
  1.953125e-03  0.3014281  0.8634341
  3.906250e-03  0.3014281  0.8634341
  7.812500e-03  0.3014281  0.8634341
  1.562500e-02  0.3014281  0.8634341
  3.125000e-02  0.3014281  0.8634341
  6.250000e-02  0.3161308  0.8499889
  1.250000e-01  0.3507648  0.8171492
  2.500000e-01  0.3837904  0.7871111
  5.000000e-01  0.4209726  0.7595674
  1.000000e+00  0.4715008  0.7293100
  2.000000e+00  0.5382431  0.6882220
  4.000000e+00  0.6111136  0.6317770
  8.000000e+00  0.6760116  0.5670342
  1.600000e+01  0.7249594  0.5099457
  3.200000e+01  0.7578188  0.4692471
  6.400000e+01  0.7776414  0.4448170
  1.280000e+02  0.7888304  0.4311913
  2.560000e+02  0.7947790  0.4240596
  5.120000e+02  0.8010289        NaN
  1.024000e+03  0.8010289        NaN
  2.048000e+03  0.8010289        NaN
  4.096000e+03  0.8010289        NaN
  8.192000e+03  0.8010289        NaN
  1.638400e+04  0.8010289        NaN
  3.276800e+04  0.8010289        NaN

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using  the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.03125.
> average_Rsquared6 <- mean(na.omit(model6$results$Rsquared))
> head(average_Rsquared6)
[1] 0.7336331
```

Figure 2. L2-regularized MLR model for "1-mer+shape" on Mad

```
8568 samples
 144 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 7712, 7711, 7711, 7711, 7712, 7712, ...
Resampling results across tuning parameters:

  lambda        RMSE        Rsquared
  3.051758e-05  0.09942387  0.7858999
  6.103516e-05  0.09942387  0.7858999
  1.220703e-04  0.09942387  0.7858999
  2.441406e-04  0.09942387  0.7858999
  4.882812e-04  0.09942387  0.7858999
  9.765625e-04  0.09942387  0.7858999
  1.953125e-03  0.09942387  0.7858999
  3.906250e-03  0.09942387  0.7858999
  7.812500e-03  0.09943377  0.7858939
  1.562500e-02  0.10015205  0.7854702
  3.125000e-02  0.10255425  0.7840143
  6.250000e-02  0.10915894  0.7794362
  1.250000e-01  0.12290319  0.7669988
  2.500000e-01  0.14323516  0.7399677
  5.000000e-01  0.16517483  0.6947949
  1.000000e+00  0.18351620  0.6392346
  2.000000e+00  0.19645176  0.5877872
  4.000000e+00  0.20450912  0.5505702
  8.000000e+00  0.20913087  0.5276691
  1.600000e+01  0.21163308  0.5148803
  3.200000e+01  0.21293695  0.5081473
  6.400000e+01  0.21360578  0.5046590
  1.280000e+02  0.21428316        NaN
  2.560000e+02  0.21428316        NaN
  5.120000e+02  0.21428316        NaN
  1.024000e+03  0.21428316        NaN
  2.048000e+03  0.21428316        NaN
  4.096000e+03  0.21428316        NaN
  8.192000e+03  0.21428316        NaN
  1.638400e+04  0.21428316        NaN
  3.276800e+04  0.21428316        NaN

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using  the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.00390625.
> average_Rsquared3 <- mean(na.omit(model3$results$Rsquared))
> head(average_Rsquared3)
[1] 0.7025783
```

Figure 3. L2-regularized MLR model for “1-mer” on Max

```
8568 samples
 274 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 7712, 7710, 7710, 7712, 7712, 7711, ...
Resampling results across tuning parameters:

  lambda        RMSE        Rsquared
  3.051758e-05  0.07983256  0.8641997
  6.103516e-05  0.07983256  0.8641997
  1.220703e-04  0.07983256  0.8641997
  2.441406e-04  0.07983256  0.8641997
  4.882812e-04  0.07983256  0.8641997
  9.765625e-04  0.07983256  0.8641997
  1.953125e-03  0.07983256  0.8641997
  3.906250e-03  0.07983256  0.8641997
  7.812500e-03  0.07983256  0.8641997
  1.562500e-02  0.08293992  0.8539958
  3.125000e-02  0.08989418  0.8312003
  6.250000e-02  0.09770819  0.8074273
  1.250000e-01  0.10777506  0.7813824
  2.500000e-01  0.12195142  0.7483146
  5.000000e-01  0.14044348  0.6999057
  1.000000e+00  0.16037194  0.6323146
  2.000000e+00  0.17804236  0.5566054
  4.000000e+00  0.19170558  0.4908092
  8.000000e+00  0.20111950  0.4454490
  1.600000e+01  0.20705466  0.4181150
  3.200000e+01  0.21046769  0.4032364
  6.400000e+01  0.21232627  0.3953695
  1.280000e+02  0.21429740         NaN
  2.560000e+02  0.21429740         NaN
  5.120000e+02  0.21429740         NaN
  1.024000e+03  0.21429740         NaN
  2.048000e+03  0.21429740         NaN
  4.096000e+03  0.21429740         NaN
  8.192000e+03  0.21429740         NaN
  1.638400e+04  0.21429740         NaN
  3.276800e+04  0.21429740         NaN

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using  the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.0078125.
> average_Rsquared4 <- mean(na.omit(model4$results$Rsquared))
> head(average_Rsquared4)
[1] 0.7200874
```

Figure 4. L2-regularized MLR model for "1-mer+shape" on Max

```
6926 samples
 144 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 6234, 6234, 6234, 6234, 6233, 6232, ...
Resampling results across tuning parameters:

  lambda        RMSE       Rsquared
  3.051758e-05  0.3741327  0.7782157
  6.103516e-05  0.3741327  0.7782157
  1.220703e-04  0.3741327  0.7782157
  2.441406e-04  0.3741327  0.7782157
  4.882812e-04  0.3741327  0.7782157
  9.765625e-04  0.3741327  0.7782157
  1.953125e-03  0.3741327  0.7782157
  3.906250e-03  0.3741327  0.7782157
  7.812500e-03  0.3741327  0.7782157
  1.562500e-02  0.3741327  0.7782157
  3.125000e-02  0.3744965  0.7781523
  6.250000e-02  0.3775201  0.7776340
  1.250000e-01  0.3875459  0.7758580
  2.500000e-01  0.4143209  0.7704810
  5.000000e-01  0.4676630  0.7570520
  1.000000e+00  0.5434125  0.7315101
  2.000000e+00  0.6228073  0.6956054
  4.000000e+00  0.6879572  0.6582451
  8.000000e+00  0.7330583  0.6273191
  1.600000e+01  0.7604372  0.6061893
  3.200000e+01  0.7758067  0.5934025
  6.400000e+01  0.7839728  0.5863221
  1.280000e+02  0.7881927  0.5825823
  2.560000e+02  0.7924668  0.5340876
  5.120000e+02  0.7925094        NaN
  1.024000e+03  0.7925094        NaN
  2.048000e+03  0.7925094        NaN
  4.096000e+03  0.7925094        NaN
  8.192000e+03  0.7925094        NaN
  1.638400e+04  0.7925094        NaN
  3.276800e+04  0.7925094        NaN

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using  the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.015625.
> average_Rsquared <- mean(na.omit(model2$results$Rsquared))
> head(average_Rsquared)
[1] 0.7190249
```

Figure 5. L2-regularized MLR model for “1-mer” on Myc

```
6926 samples
 274 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 6233, 6234, 6234, 6234, 6234, 6232, ...
Resampling results across tuning parameters:

  lambda        RMSE       Rsquared
  3.051758e-05  0.3044091  0.8551475
  6.103516e-05  0.3044091  0.8551475
  1.220703e-04  0.3044091  0.8551475
  2.441406e-04  0.3044091  0.8551475
  4.882812e-04  0.3044091  0.8551475
  9.765625e-04  0.3044091  0.8551475
  1.953125e-03  0.3044091  0.8551475
  3.906250e-03  0.3044091  0.8551475
  7.812500e-03  0.3044091  0.8551475
  1.562500e-02  0.3044091  0.8551475
  3.125000e-02  0.3044091  0.8551475
  6.250000e-02  0.3237244  0.8372757
  1.250000e-01  0.3475603  0.8157906
  2.500000e-01  0.3750757  0.7942302
  5.000000e-01  0.4129851  0.7705247
  1.000000e+00  0.4686106  0.7390194
  2.000000e+00  0.5398607  0.6931013
  4.000000e+00  0.6138105  0.6317843
  8.000000e+00  0.6768448  0.5657804
  1.600000e+01  0.7231997  0.5094154
  3.200000e+01  0.7535558  0.4698781
  6.400000e+01  0.7716196  0.4459729
  1.280000e+02  0.7816803  0.4326430
  2.560000e+02  0.7869946  0.4256434
  5.120000e+02  0.7925509        NaN
  1.024000e+03  0.7925509        NaN
  2.048000e+03  0.7925509        NaN
  4.096000e+03  0.7925509        NaN
  8.192000e+03  0.7925509        NaN
  1.638400e+04  0.7925509        NaN
  3.276800e+04  0.7925509        NaN

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using  the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.03125.
> average_Rsquared <- mean(na.omit(model2$results$Rsquared))
> head(average_Rsquared)
[1] 0.7307367
```

Figure 6. L2-regularized MLR model for "1-mer+shape" on Myc

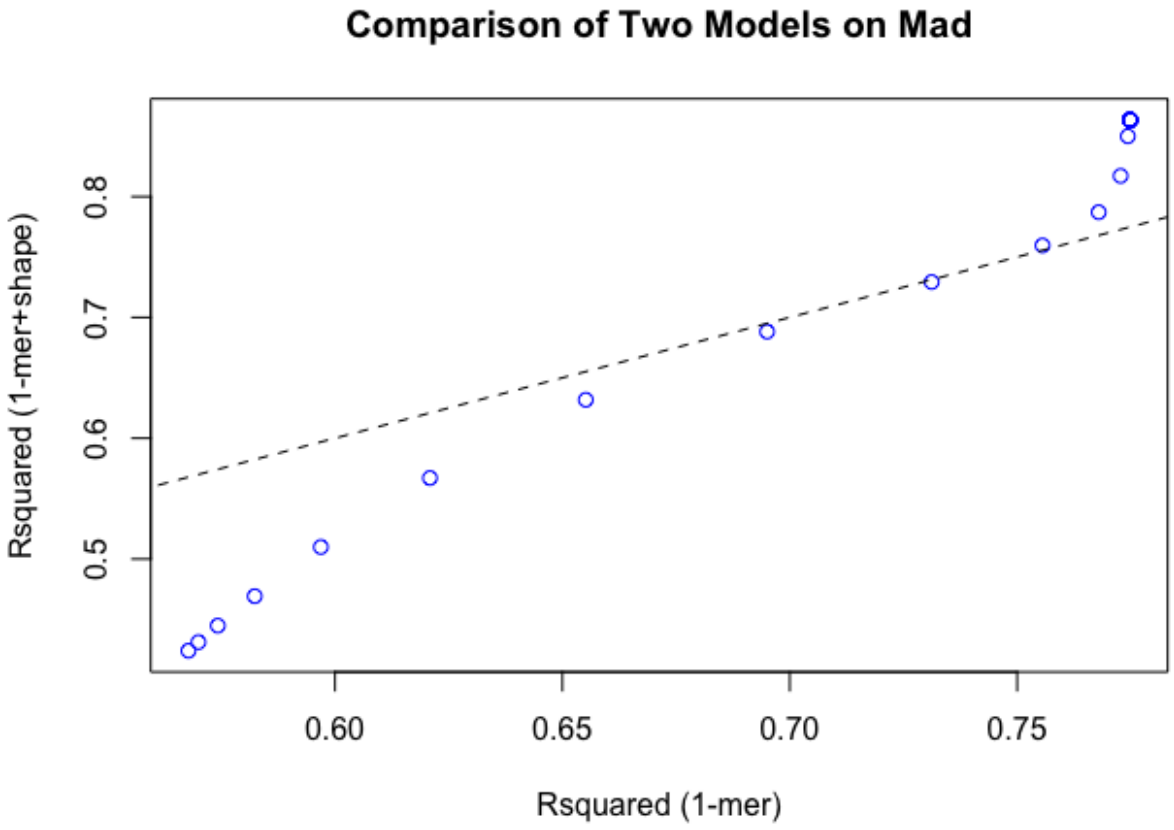

Figure 7. Comparison "1-mer" vs. "1-mer+shape" on Mad

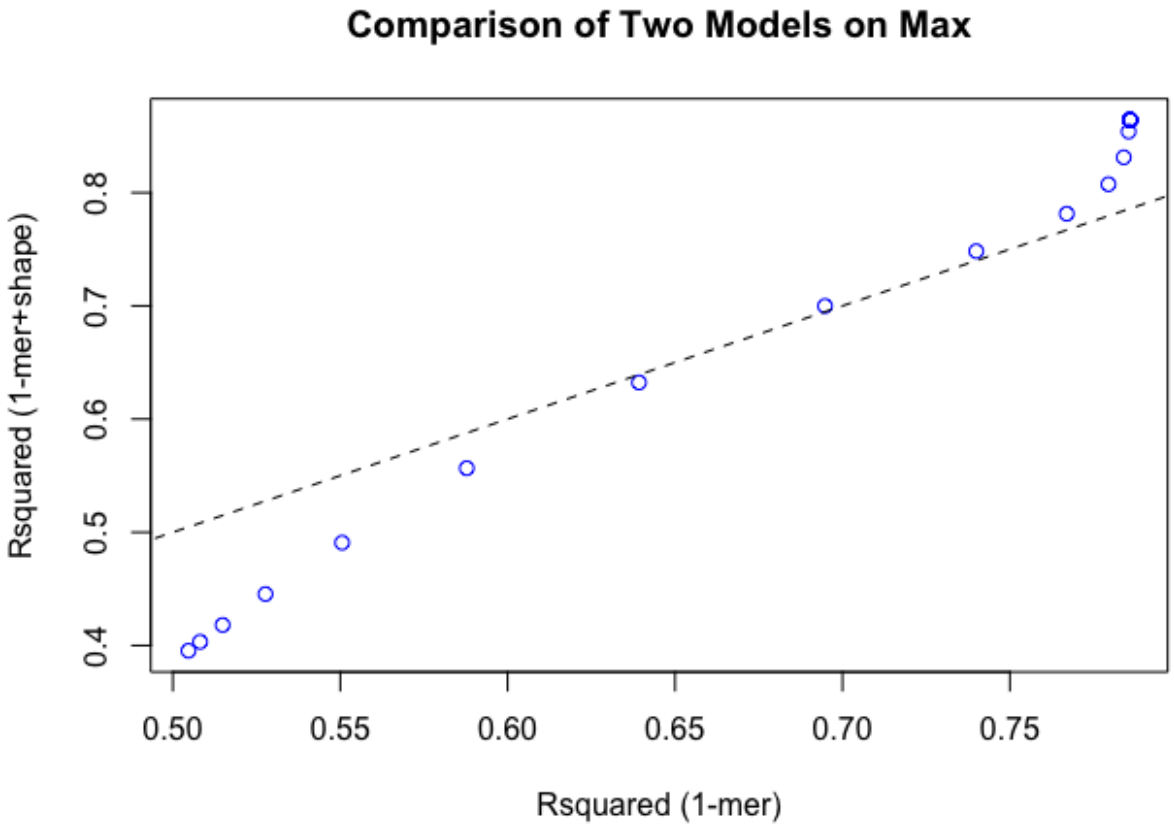

Figure 8. Comparison "1-mer" vs. "1-mer+shape" on Max

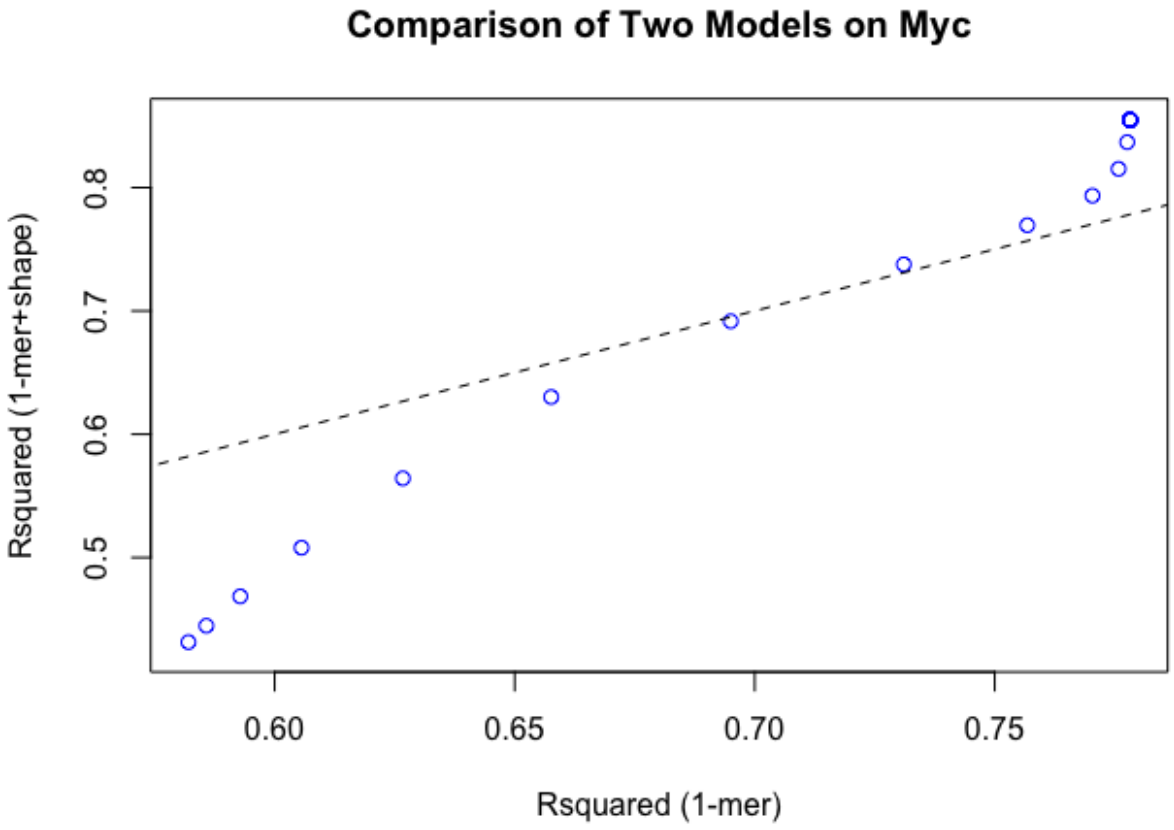

Figure 9. Comparison "1-mer" vs. "1-mer+shape" on Myc

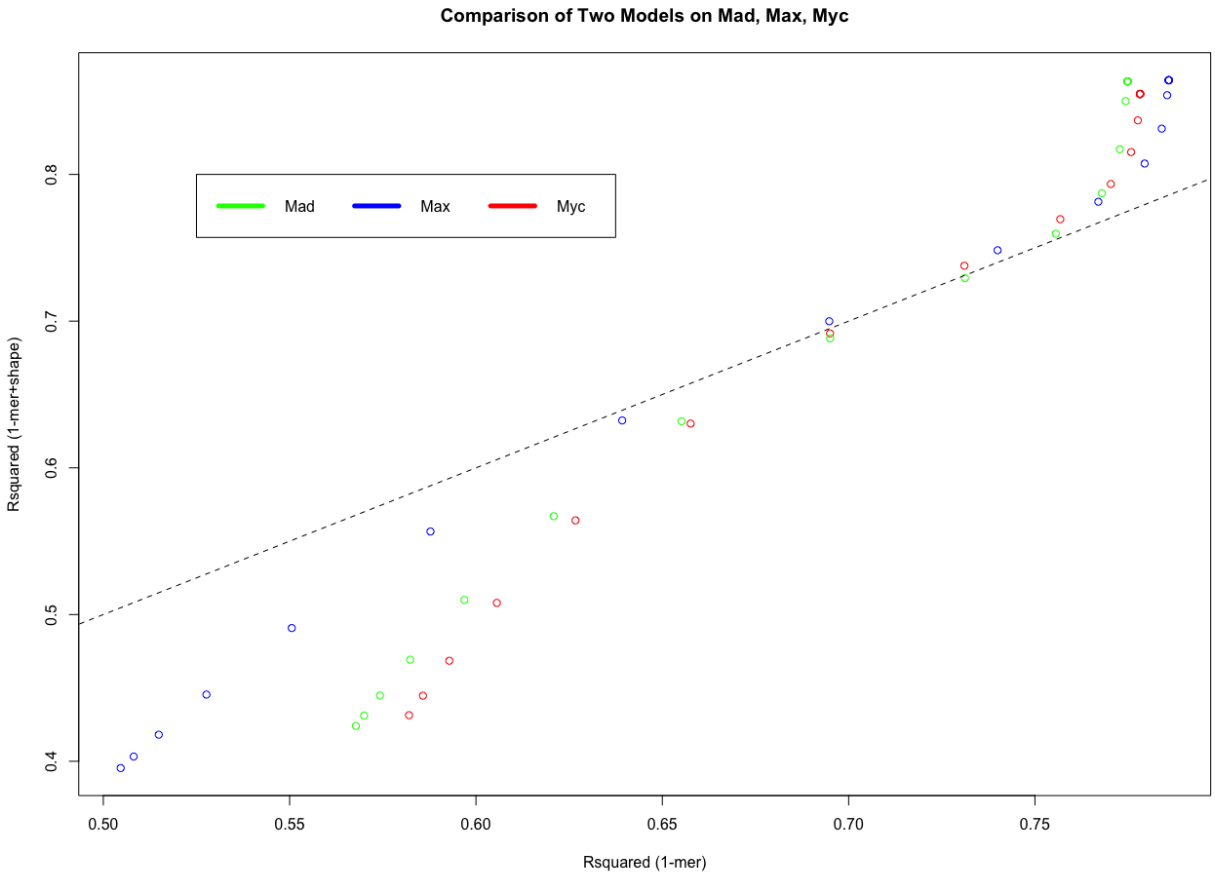

Figure 10. Comparison "1-mer" vs. "1-mer+shape" on Mad, Max, Myc

To see if the shape-augmented model outperforms significantly enough, we compute p-value with null hypothesis that two models perform the same (no outperformance). Note that R-squared is measure of how close the data are to fitted regression, i.e. how well the variation in response is explained by the fitted model. The improvement looks quite modest here, e.g. 0.716 vs. 0.733 of average R-squared between 1-mer and 1-mer+shape on Mad, but we want to know if this is statistically significant. One idea could be utilizing one-tail or two-tail t-test. For example, on data set Mad, we can set up the null hypothesis that there is no difference between two R-squared means of 1-mer and 1-mer+shape, and run t-test on list of generated R-squared of 1-mer and 1-mer+shape since if there's no difference between two average R-squared, then neither model outperforms the other. Then, we can obtain, in this case, p-value = 0.6475, which is way greater than 0.05. So, we can accept the null, i.e., there's not much statistically significant difference between two. This can be also confirmed by running one-tail test on 1-mer+shape with fixed mean = 0.716, which will yield p-value=0.6122. Hence, again, we can't reject the null hypothesis that average R-squared of 1-mer+shape is statistically significantly greater than average of R-squared of 1-mer. The results are similar on the remaining data sets Max and Myc. Even running t-test on combined data sets, i.e., three R-squared means from each data for 1-mer and the same for 1-mer+shape, yielded the similar result.

But, another approach by using Fisher's exact test yielded different result. In general, if we have contingency table, say 2 X 2 matrix with rows = student A, B and columns = Pass MATH577, fail MATH577, the null hypothesis states that the proportions/ratios of passing MATH577 are not different for two students. In our case, if we construct contingency table with rows = 1-mer, 1-mer+shape, and columns = averaged R-squared of Mad, Max, and Myc, the null hypothesis states that the proportions of mean R-squared on Mad (or Max, Myc) are not different between 1-mer and 1-mer+shape. Since the table should be integer-valued for computation, Fisher's test can be run by updating values multiplied by power of 10. The resulting p-value was 0.0005, so null hypothesis can be rejected, i.e. there's significant difference in R-squared on Mad/Max/Myc between two models. (But, this approach might not be entirely correct because unlike t-test, proportion of R-squared on one data over the all doesn't necessarily represent the performance)

Or, though not actually tested here, but like Zhou et al's PNAS paper, we can assess Spearman's rank correlation, take the differences in this values between models to get null distribution of rank correlation differences, then calculate empirical p-values.

```
> t.test(na.omit(model5$results$Rsquared),na.omit(model6$results$Rsquared), var.equal=TRUE, paired=FALSE)

        Two Sample t-test

data:  na.omit(model5$results$Rsquared) and na.omit(model6$results$Rsquared)
t = -0.46033, df = 46, p-value = 0.6475
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09403996  0.05903378
sample estimates:
mean of x mean of y
0.7161300 0.7336331

> t.test(na.omit(model6$results$Rsquared),mu=0.7161300)

        One Sample t-test

data:  na.omit(model6$results$Rsquared)
t = 0.51394, df = 23, p-value = 0.6122
alternative hypothesis: true mean is not equal to 0.71613
95 percent confidence interval:
 0.6631813 0.8040850
sample estimates:
mean of x
0.7336331

> t.test(rsq[1,],rsq[2,], var.equal=TRUE, paired=FALSE)

        Two Sample t-test

data:  rsq[1, ] and rsq[2, ]
t = -2.3846, df = 4, p-value = 0.07561
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.033708500  0.002559166
sample estimates:
mean of x mean of y
0.7125777 0.7281523

> t.test (rsq[2,], mu=0.7125777)

        One Sample t-test

data:  rsq[2, ]
t = 3.7817, df = 2, p-value = 0.06335
alternative hypothesis: true mean is not equal to 0.7125777
95 percent confidence interval:
 0.7104322 0.7458724
sample estimates:
mean of x
0.7281523

> r_contingency
         [,1]     [,2]     [,3]
[1,] 0.716130 0.702578 0.719025
[2,] 0.733633 0.720087 0.730737
> r_fisher <- fisher.test(1000000*r_contingency, simulate.p.value=TRUE)
> r_fisher$p.value
[1] 0.0004997501
```

CTCF transcription factor is actually transcriptional repressor, which is known as 11-zinc finger protein. It's also often referred as CCCTC-binding factor because it binds as transcription factor to repeats of the core sequence CCCTC. This CFCT binds to the consensus sequence that is defined by its zinc finger motifs. This zinc motif in them typically consists of alpha helix pair linked by zinc coordination. One helix binds in major groove at each half-site, while the second helix helps maintaining the overall structure. And, even though some local distortions to roll or propeller twist could be observed, DNA itself retains B-DNA form with mean helical twist of ~34°, usually quite low propeller twist in range between -16° and -20° in average, as well as deep, narrow minor groove. Indeed, average 5Å MGW is quite close to 5.2 Å of 1bna, recalling the previous assignment. Even if the mean propeller twist is not really within the typical bound, if we only consider dinucleotide values for helical parameter, for example, ≈-8° for CC, it seems to make sense (Or, some twist might have happened when binding). Helical twist is very close to that of typical B-DNA.

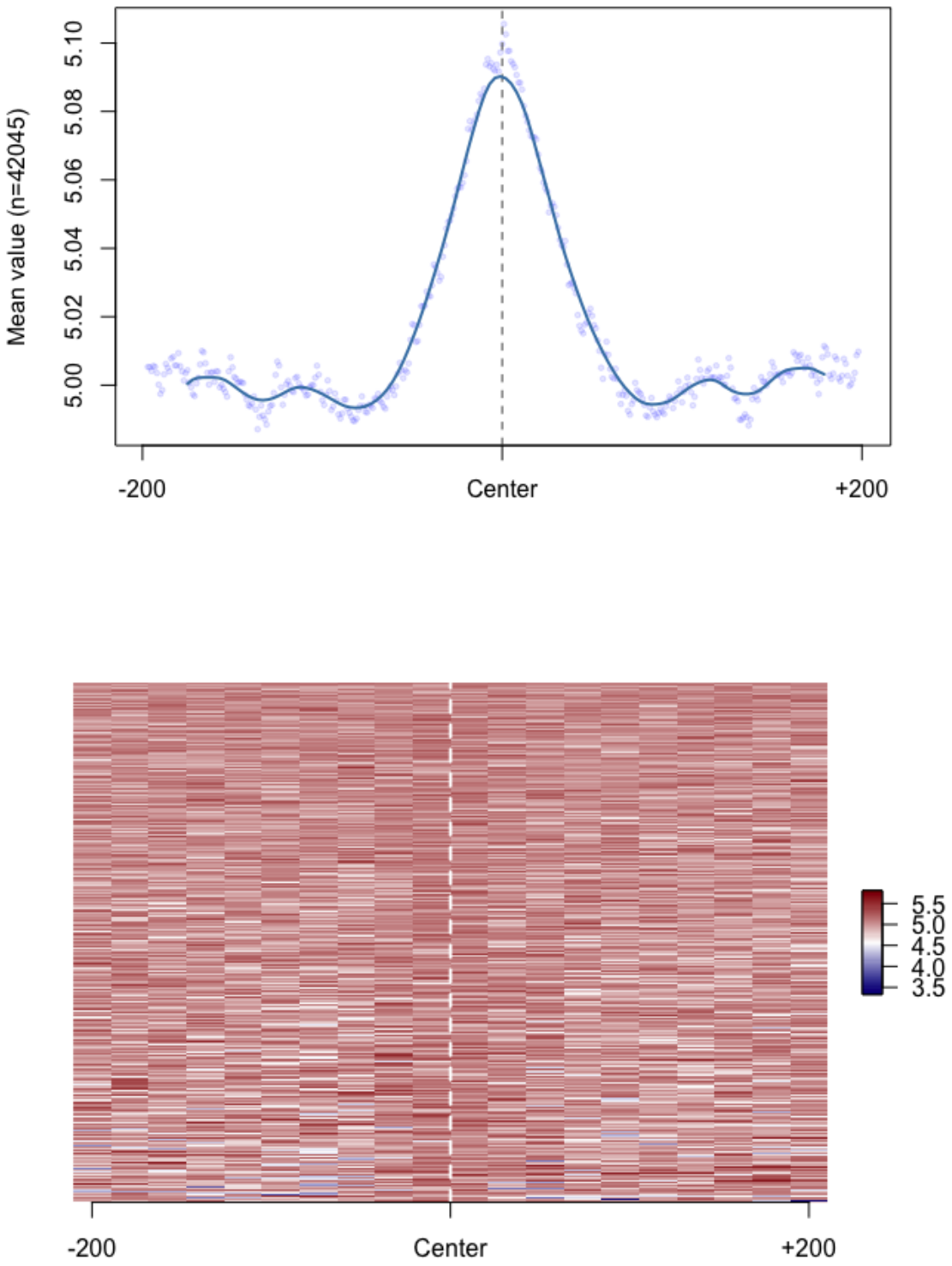

Figure 11. Ensemble plot for minor groove width (MGW)

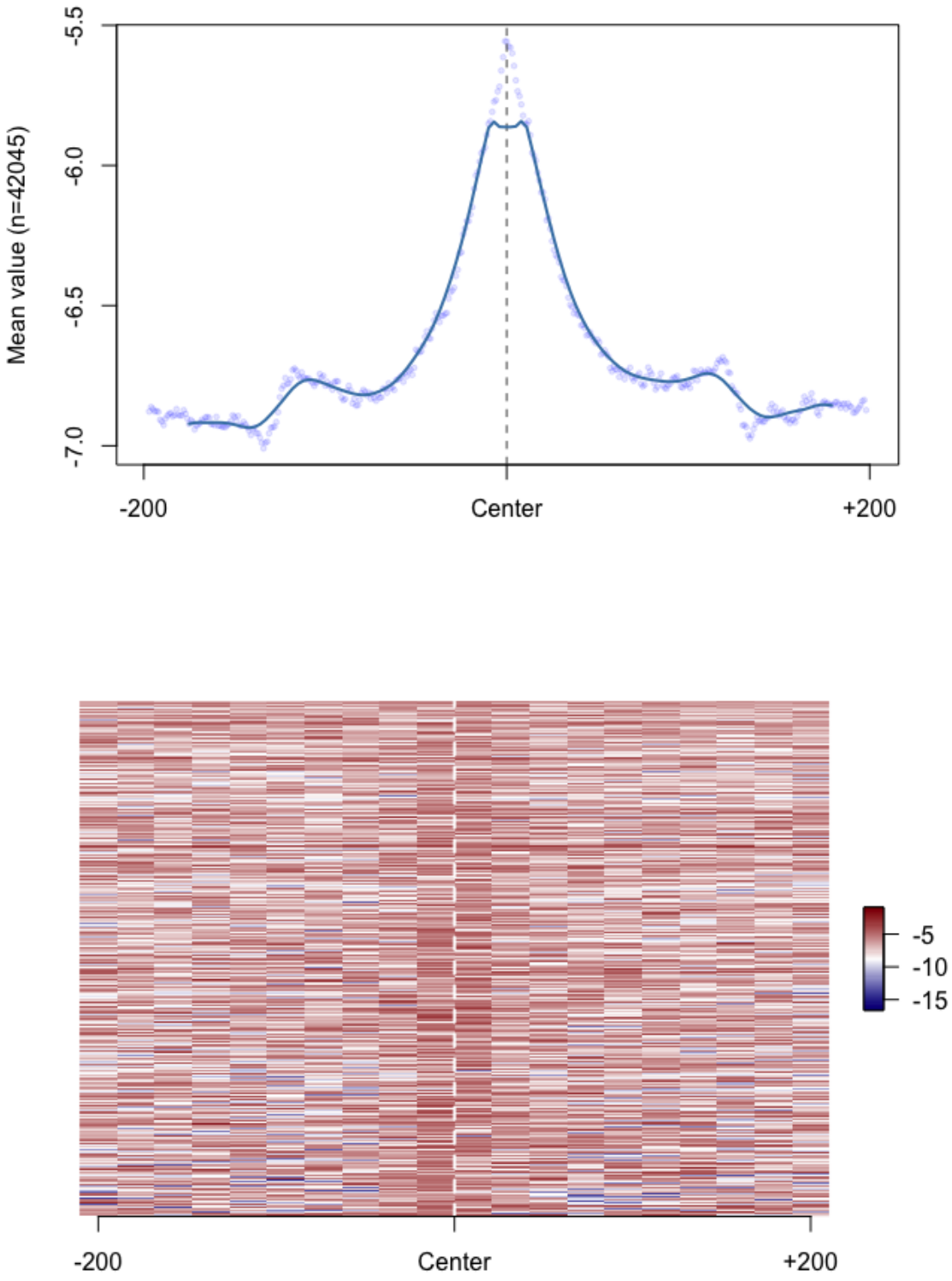

Figure 12. Ensemble plot for propeller twist (ProT)

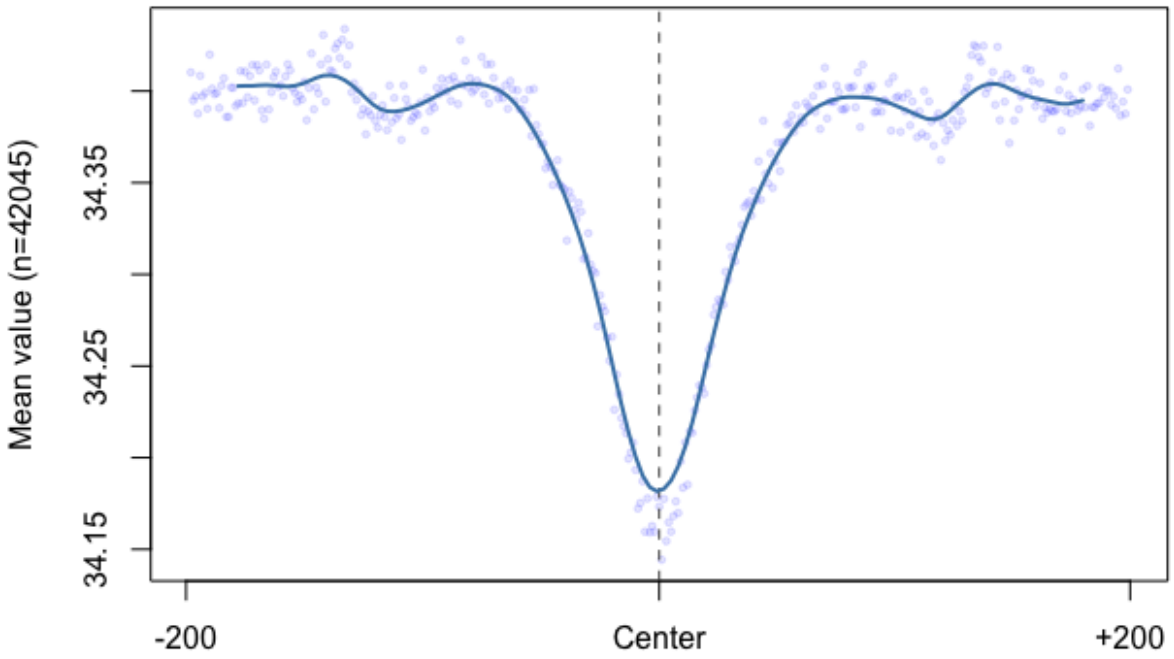

Figure 13. Ensemble plot for helical twist (HelT)

Two different logistic models for “1-mer” and “1-mer+shape” are built as prediction models for three different in vivo ChIP-seq data of CTCF TF of Mus musculus. Comparison can be done via receiver operating characteristic curve (ROC), which is a plot of true positive rate against the false positive rate with different possible cutpoints. The closer the ROC curve follows left borderline and then top border, the more accuracy, i.e., how well the test separates the tested groups into two, we attain. On the other hand, diagonal-like curve will imply less accuracy. This accuracy can be measured by area under ROC curve (AUC) where area 1 represents indeed a perfect test whereas area 0.5 represents meaningless. AS we have 0.827 for “1-mer” and 0.813 for “1-mer+shape”, the models are considered to be “good”/”accurate.”

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9476  -0.6794   0.0243   0.7368   3.4815

Coefficients: (29 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.566068 270.260537  -0.058 0.954070
X1            8.159449 270.261038   0.030 0.975915
X2            8.324427 270.261047   0.031 0.975428
X3            8.215030 270.261059   0.030 0.975751
X4            7.978090 270.261021   0.030 0.976450
X5            0.077150   0.171087   0.451 0.652033
X6            0.299716   0.176777   1.695 0.089991 .
X7            0.324966   0.178338   1.822 0.068426 .
X8                  NA         NA      NA       NA
X9            0.022376   0.174384   0.128 0.897901
X10           0.616403   0.177934   3.464 0.000532 ***
X11           0.401906   0.175001   2.297 0.021642 *
X12                 NA         NA      NA       NA
X13          -0.270636   0.172028  -1.573 0.115671
X14           0.512927   0.175926   2.916 0.003550 **
X15           0.337207   0.176926   1.906 0.056661 .
X16                 NA         NA      NA       NA
X17           0.001139   0.176162   0.006 0.994843
X18           0.485858   0.177939   2.730 0.006324 **
X19           0.800471   0.181178   4.418 9.95e-06 ***
X20                 NA         NA      NA       NA
X21           0.170525   0.173769   0.981 0.326428
X22           0.410991   0.172061   2.389 0.016911 *
X23           0.516247   0.179185   2.881 0.003963 **
X24                 NA         NA      NA       NA
X25           0.001819   0.172591   0.011 0.991593
X26           0.609752   0.179119   3.404 0.000664 ***
X27           0.129938   0.180620   0.719 0.471896
X28                 NA         NA      NA       NA
X29           0.032637   0.177020   0.184 0.853724
X30           0.516526   0.176469   2.927 0.003422 **
X31           0.774910   0.177214   4.373 1.23e-05 ***
X32                 NA         NA      NA       NA
X33          -0.426793   0.172019  -2.481 0.013099 *
X34           0.228955   0.177160   1.292 0.196232
X35           0.420648   0.174006   2.417 0.015630 *
X36                 NA         NA      NA       NA
X37           0.175118   0.177251   0.988 0.323168
X38           0.614410   0.183990   3.339 0.000840 ***
X39           0.691514   0.176304   3.922 8.77e-05 ***
X40                 NA         NA      NA       NA
```

```
X41          0.143905   0.177280   0.812 0.416941
X42          0.763439   0.179883   4.244 2.19e-05 ***
X43          0.510925   0.173215   2.950 0.003181 **
X44                NA         NA      NA       NA
X45         -0.063388   0.176826  -0.358 0.719989
X46          0.427036   0.172069   2.482 0.013073 *
X47          0.603167   0.173456   3.477 0.000506 ***
X48                NA         NA      NA       NA
X49          0.107923   0.172627   0.625 0.531852
X50          0.580576   0.177900   3.263 0.001100 **
X51          0.524668   0.177986   2.948 0.003200 **
X52                NA         NA      NA       NA
X53          0.215246   0.173581   1.240 0.214966
X54          0.776228   0.177732   4.367 1.26e-05 ***
X55          0.736248   0.178589   4.123 3.75e-05 ***
X56                NA         NA      NA       NA
X57         -0.455561   0.179616  -2.536 0.011203 *
X58          0.490985   0.177825   2.761 0.005762 **
X59          0.871557   0.182687   4.771 1.84e-06 ***
X60                NA         NA      NA       NA
X61         -0.352121   0.174742  -2.015 0.043895 *
X62          0.343350   0.175344   1.958 0.050212 .
X63          0.310103   0.181268   1.711 0.087129 .
X64                NA         NA      NA       NA
X65          0.194903   0.176912   1.102 0.270596
X66          0.688738   0.181809   3.788 0.000152 ***
X67          0.741696   0.182024   4.075 4.61e-05 ***
X68                NA         NA      NA       NA
X69          0.089602   0.173799   0.516 0.606170
X70          0.511750   0.173679   2.947 0.003214 **
X71          0.643368   0.175808   3.659 0.000253 ***
X72                NA         NA      NA       NA
X73          0.033207   0.170084   0.195 0.845206
X74          0.448757   0.176714   2.539 0.011103 *
X75          0.238920   0.181575   1.316 0.188235
X76                NA         NA      NA       NA
X77          0.318275   0.173792   1.831 0.067048 .
X78          0.304644   0.178536   1.706 0.087945 .
X79          0.704604   0.180937   3.894 9.85e-05 ***
X80                NA         NA      NA       NA
X81         -0.128903   0.174073  -0.741 0.458991
X82          0.372127   0.178930   2.080 0.037549 *
X83          0.180419   0.177163   1.018 0.308498
X84                NA         NA      NA       NA
X85         -0.190459   0.169043  -1.127 0.259873
X86          0.563253   0.181736   3.099 0.001940 **
X87          0.399426   0.181048   2.206 0.027371 *
```

```
X88               NA         NA      NA       NA
X89         0.500252   0.176687   2.831 0.004636 **
X90         0.759318   0.176174   4.310 1.63e-05 ***
X91         0.877815   0.177609   4.942 7.72e-07 ***
X92               NA         NA      NA       NA
X93         0.110890   0.173606   0.639 0.522989
X94         0.519107   0.177136   2.931 0.003384 **
X95         0.560911   0.180685   3.104 0.001907 **
X96               NA         NA      NA       NA
X97         0.059087   0.169690   0.348 0.727685
X98         0.362830   0.176932   2.051 0.040299 *
X99         0.429126   0.176359   2.433 0.014964 *
X100              NA         NA      NA       NA
X101        -0.268958   0.172029  -1.563 0.117949
X102        0.003304   0.181251   0.018 0.985457
X103        0.421132   0.181142   2.325 0.020079 *
X104              NA         NA      NA       NA
X105        -0.051988   0.170083  -0.306 0.759862
X106        0.634228   0.180501   3.514 0.000442 ***
X107        0.364533   0.178254   2.045 0.040853 *
X108              NA         NA      NA       NA
X109        0.116783   0.173790   0.672 0.501598
X110        0.290683   0.179422   1.620 0.105209
X111        0.406056   0.181603   2.236 0.025355 *
X112              NA         NA      NA       NA
X113        0.071766   0.171743   0.418 0.676043
X114        0.547531   0.183024   2.992 0.002775 **
X115        0.475032   0.179792   2.642 0.008239 **
X116              NA         NA      NA       NA
X117        0.418484   0.168056   2.490 0.012769 *
X118        0.555931   0.178077   3.122 0.001797 **
X119        0.784943   0.174093   4.509 6.52e-06 ***
X120              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2772.6  on 1999  degrees of freedom
Residual deviance: 1767.1  on 1908  degrees of freedom
AIC: 1951.1

Number of Fisher Scoring iterations: 14
```

Figure 14. Logistic regression model on bound and unbound data (1-mer)

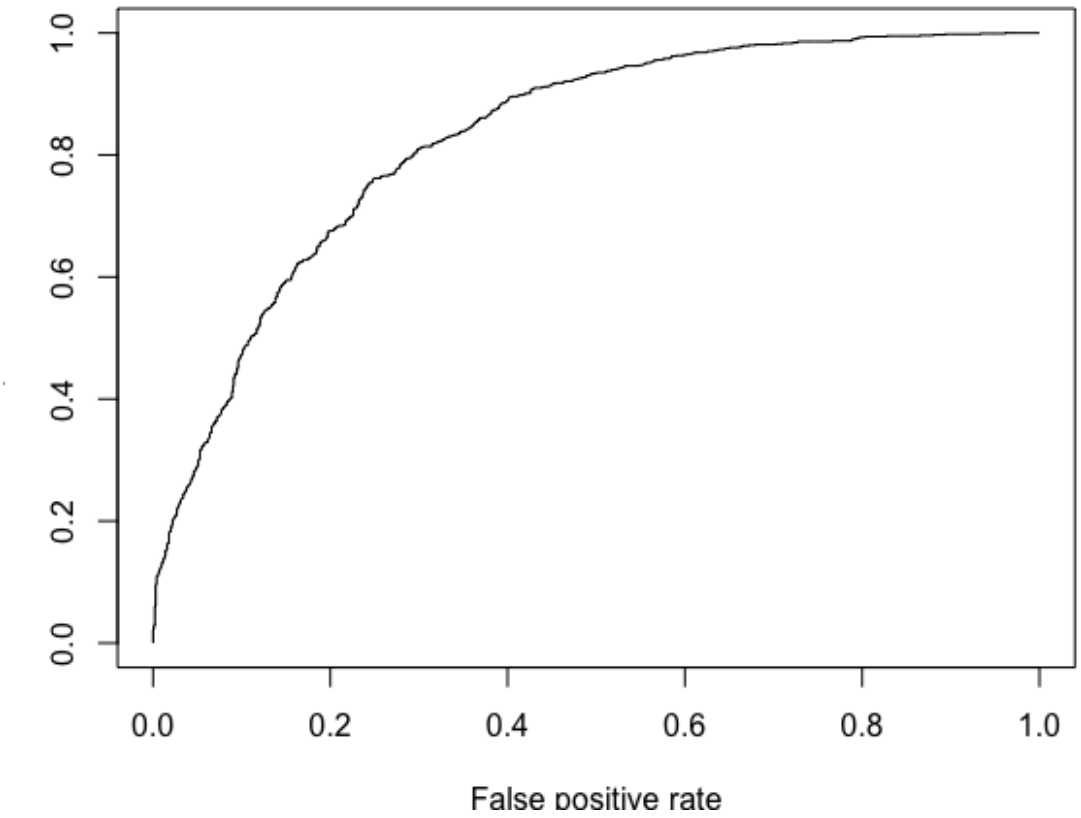

Figure 15. ROC plot of logistic regression model on bound and unbound data (1-mer). AUC = 0.827536

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9728  -0.5996   0.0183   0.6266   3.8727

Coefficients: (29 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.266e+02  4.464e+02   0.284 0.776729
X1          -1.726e+02  4.467e+02  -0.386 0.699167
X2          -1.727e+02  4.467e+02  -0.387 0.699118
X3          -1.727e+02  4.467e+02  -0.387 0.699115
X4          -1.730e+02  4.467e+02  -0.387 0.698531
X5           4.446e-01  3.859e-01   1.152 0.249190
X6           2.499e-01  3.068e-01   0.814 0.415360
X7           3.023e-01  2.644e-01   1.143 0.252858
X8                  NA         NA      NA       NA
X9           8.235e-01  4.925e-01   1.672 0.094510 .
X10          1.466e+00  5.210e-01   2.813 0.004904 **
X11          2.010e-01  5.272e-01   0.381 0.702980
X12                 NA         NA      NA       NA
X13         -5.002e-01  4.986e-01  -1.003 0.315731
X14         -5.318e-01  5.183e-01  -1.026 0.304845
X15         -2.686e-01  5.110e-01  -0.526 0.599099
X16                 NA         NA      NA       NA
X17         -5.371e-01  5.176e-01  -1.038 0.299397
X18         -7.005e-02  5.378e-01  -0.130 0.896374
X19         -1.600e-01  5.161e-01  -0.310 0.756514
X20                 NA         NA      NA       NA
X21          3.064e-01  5.258e-01   0.583 0.560102
X22         -1.094e-01  5.275e-01  -0.207 0.835743
X23          8.676e-01  5.384e-01   1.611 0.107081
X24                 NA         NA      NA       NA
X25          4.545e-01  5.426e-01   0.838 0.402254
X26          1.177e+00  5.611e-01   2.098 0.035928 *
X27          5.683e-01  5.308e-01   1.071 0.284368
X28                 NA         NA      NA       NA
X29         -5.251e-01  4.911e-01  -1.069 0.284963
X30          3.492e-02  5.116e-01   0.068 0.945582
X31          3.979e-01  4.938e-01   0.806 0.420357
X32                 NA         NA      NA       NA
X33         -6.963e-01  5.193e-01  -1.341 0.179934
X34          2.051e-01  5.368e-01   0.382 0.702419
X35          1.713e-01  5.173e-01   0.331 0.740629
X36                 NA         NA      NA       NA
X37          7.768e-01  5.451e-01   1.425 0.154108
X38          4.929e-01  5.476e-01   0.900 0.368052
X39          7.552e-01  5.278e-01   1.431 0.152515
X40                 NA         NA      NA       NA
```

```
X41         -3.824e-01  5.214e-01  -0.733 0.463356
X42          6.242e-01  5.616e-01   1.111 0.266371
X43         -5.193e-01  5.313e-01  -0.977 0.328332
X44                 NA         NA      NA       NA
X45          1.449e+00  5.155e-01   2.812 0.004930 **
X46          9.992e-02  5.367e-01   0.186 0.852317
X47          6.200e-01  5.223e-01   1.187 0.235245
X48                 NA         NA      NA       NA
X49         -3.656e-01  5.330e-01  -0.686 0.492776
X50         -1.981e-01  5.499e-01  -0.360 0.718704
X51          4.166e-01  5.415e-01   0.769 0.441658
X52                 NA         NA      NA       NA
X53         -3.844e-01  5.175e-01  -0.743 0.457594
X54          7.991e-01  5.443e-01   1.468 0.142094
X55         -1.113e-01  5.254e-01  -0.212 0.832161
X56                 NA         NA      NA       NA
X57          2.383e-01  5.134e-01   0.464 0.642503
X58          1.409e-01  5.414e-01   0.260 0.794642
X59          6.142e-01  5.093e-01   1.206 0.227829
X60                 NA         NA      NA       NA
X61         -4.188e-03  5.221e-01  -0.008 0.993600
X62          1.500e-01  5.310e-01   0.282 0.777595
X63          5.790e-03  5.228e-01   0.011 0.991162
X64                 NA         NA      NA       NA
X65         -5.231e-01  5.248e-01  -0.997 0.318904
X66         -4.069e-02  5.550e-01  -0.073 0.941552
X67          1.729e-01  5.275e-01   0.328 0.743028
X68                 NA         NA      NA       NA
X69         -1.075e+00  5.155e-01  -2.086 0.036954 *
X70         -6.830e-01  5.467e-01  -1.249 0.211549
X71          1.435e-02  5.138e-01   0.028 0.977724
X72                 NA         NA      NA       NA
X73          3.362e-01  5.115e-01   0.657 0.510980
X74          5.766e-01  5.099e-01   1.131 0.258169
X75          7.637e-01  5.228e-01   1.461 0.144095
X76                 NA         NA      NA       NA
X77          1.350e+00  5.163e-01   2.615 0.008926 **
X78          1.214e+00  5.512e-01   2.203 0.027583 *
X79          1.260e+00  5.406e-01   2.330 0.019792 *
X80                 NA         NA      NA       NA
X81         -9.743e-01  5.222e-01  -1.866 0.062084 .
X82          1.613e-01  5.382e-01   0.300 0.764460
X83         -3.621e-01  5.211e-01  -0.695 0.487085
X84                 NA         NA      NA       NA
X85          9.043e-01  5.062e-01   1.787 0.074003 .
X86          7.261e-01  5.293e-01   1.372 0.170157
X87          1.059e+00  5.097e-01   2.078 0.037721 *
```

X88	NA	NA	NA	NA	
X89	2.883e-01	5.072e-01	0.568	0.569796	
X90	9.804e-01	5.246e-01	1.869	0.061651	.
X91	5.441e-01	5.132e-01	1.060	0.289084	
X92	NA	NA	NA	NA	
X93	-1.432e-01	5.057e-01	-0.283	0.777065	
X94	5.291e-02	5.146e-01	0.103	0.918109	
X95	2.173e-01	5.090e-01	0.427	0.669469	
X96	NA	NA	NA	NA	
X97	4.378e-01	5.131e-01	0.853	0.393486	
X98	8.407e-02	5.406e-01	0.156	0.876409	
X99	2.933e-01	5.287e-01	0.555	0.579057	
X100	NA	NA	NA	NA	
X101	-7.678e-01	5.254e-01	-1.461	0.143913	
X102	1.422e-01	5.554e-01	0.256	0.797930	
X103	6.235e-01	5.231e-01	1.192	0.233288	
X104	NA	NA	NA	NA	
X105	3.502e-01	5.011e-01	0.699	0.484647	
X106	1.645e+00	5.299e-01	3.105	0.001905	**
X107	7.813e-01	4.921e-01	1.588	0.112390	
X108	NA	NA	NA	NA	
X109	4.334e-01	4.839e-01	0.896	0.370491	
X110	-4.023e-01	5.251e-01	-0.766	0.443578	
X111	2.668e-01	4.816e-01	0.554	0.579645	
X112	NA	NA	NA	NA	
X113	-9.224e-01	3.887e-01	-2.373	0.017643	*
X114	8.250e-02	3.320e-01	0.249	0.803743	
X115	-1.809e-01	3.023e-01	-0.598	0.549545	
X116	NA	NA	NA	NA	
X117	4.771e-01	2.530e-01	1.886	0.059306	.
X118	7.559e-01	2.522e-01	2.997	0.002729	**
X119	1.068e+00	2.625e-01	4.068	4.75e-05	***
X120	NA	NA	NA	NA	
X121	1.392e+00	1.531e+00	0.909	0.363397	
X122	4.643e+00	1.662e+00	2.794	0.005203	**
X123	-2.261e-02	1.645e+00	-0.014	0.989034	
X124	2.915e+00	1.682e+00	1.733	0.083059	.
X125	1.399e+00	1.709e+00	0.818	0.413093	
X126	2.255e-01	1.703e+00	0.132	0.894666	
X127	7.565e-01	1.686e+00	0.449	0.653590	
X128	7.542e-02	1.769e+00	0.043	0.965992	
X129	-6.275e-01	1.760e+00	-0.357	0.721411	
X130	-7.346e-01	1.697e+00	-0.433	0.665108	
X131	3.897e+00	1.697e+00	2.297	0.021637	*
X132	-1.363e+00	1.741e+00	-0.783	0.433531	
X133	8.729e-01	1.702e+00	0.513	0.608058	
X134	6.421e-01	1.747e+00	0.368	0.713231	

```
X135         2.625e+00  1.766e+00   1.487 0.137078
X136         7.926e-02  1.660e+00   0.048 0.961912
X137        -2.494e+00  1.708e+00  -1.461 0.144102
X138         1.331e+00  1.762e+00   0.755 0.450100
X139        -2.142e-01  1.751e+00  -0.122 0.902618
X140         1.227e+00  1.722e+00   0.713 0.475896
X141         2.249e+00  1.665e+00   1.351 0.176761
X142        -4.386e-01  1.740e+00  -0.252 0.800977
X143        -5.009e-01  1.779e+00  -0.282 0.778220
X144         2.451e+00  1.653e+00   1.482 0.138259
X145         1.026e+00  1.652e+00   0.621 0.534556
X146        -1.804e-01  1.606e+00  -0.112 0.910555
X147         3.322e-01  1.086e+00   0.306 0.759785
X148         2.386e+00  1.095e+00   2.178 0.029375 *
X149         9.826e-01  1.093e+00   0.899 0.368849
X150         1.002e+00  1.123e+00   0.892 0.372146
X151        -9.601e-01  1.107e+00  -0.867 0.385906
X152         7.380e-01  1.059e+00   0.697 0.485927
X153         6.106e-01  1.077e+00   0.567 0.570817
X154         1.224e+00  1.071e+00   1.144 0.252830
X155        -5.522e-02  1.086e+00  -0.051 0.959452
X156         3.907e+00  1.126e+00   3.470 0.000520 ***
X157         2.552e-01  1.118e+00   0.228 0.819467
X158         8.416e-01  1.146e+00   0.734 0.462671
X159         1.986e+00  1.097e+00   1.811 0.070174 .
X160         1.521e+00  1.080e+00   1.408 0.159084
X161         1.098e+00  1.113e+00   0.986 0.323929
X162         1.547e+00  1.066e+00   1.452 0.146421
X163        -6.634e-01  1.083e+00  -0.612 0.540325
X164        -1.089e+00  1.118e+00  -0.975 0.329736
X165         1.626e-01  1.120e+00   0.145 0.884614
X166         9.596e-01  1.090e+00   0.881 0.378486
X167         1.543e-01  1.092e+00   0.141 0.887630
X168         7.783e-01  1.053e+00   0.739 0.459844
X169         8.224e-01  1.093e+00   0.753 0.451709
X170        -6.880e-01  1.158e+00  -0.594 0.552325
X171        -8.968e-01  1.073e+00  -0.835 0.403501
X172         1.725e+00  1.091e+00   1.581 0.113971
X173        -1.418e-01  1.224e+00  -0.116 0.907781
X174        -2.184e+00  1.435e+00  -1.522 0.128123
X175        -8.784e-01  1.448e+00  -0.607 0.544105
X176        -3.279e+00  1.586e+00  -2.068 0.038671 *
X177        -1.043e+00  1.553e+00  -0.671 0.502045
X178         1.346e-01  1.529e+00   0.088 0.929824
X179        -1.149e-02  1.551e+00  -0.007 0.994092
X180        -6.474e-01  1.596e+00  -0.406 0.685050
X181         2.554e+00  1.627e+00   1.570 0.116387
```

```
X182        -2.062e-01  1.578e+00  -0.131 0.895999
X183         1.394e+00  1.565e+00   0.891 0.372993
X184         1.255e+00  1.577e+00   0.796 0.426132
X185        -9.422e-01  1.579e+00  -0.597 0.550676
X186         1.032e+00  1.604e+00   0.643 0.519987
X187         6.324e-01  1.567e+00   0.404 0.686575
X188        -1.387e-01  1.560e+00  -0.089 0.929120
X189        -6.548e-01  1.563e+00  -0.419 0.675224
X190         6.980e-01  1.574e+00   0.443 0.657519
X191         1.260e+00  1.586e+00   0.795 0.426800
X192        -2.413e+00  1.566e+00  -1.541 0.123395
X193         7.788e-01  1.540e+00   0.506 0.612967
X194         9.956e-02  1.491e+00   0.067 0.946774
X195         8.369e-01  1.571e+00   0.533 0.594244
X196         5.572e-01  1.572e+00   0.354 0.723042
X197        -1.248e+00  1.559e+00  -0.800 0.423428
X198        -2.590e-01  1.519e+00  -0.171 0.864615
X199         2.830e+00  1.320e+00   2.144 0.032071 *
X200        -3.267e-02  8.381e-01  -0.039 0.968902
X201         3.283e+00  9.959e-01   3.297 0.000978 ***
X202         2.051e+00  1.003e+00   2.046 0.040800 *
X203         2.016e+00  1.000e+00   2.015 0.043889 *
X204         4.282e-01  9.924e-01   0.431 0.666152
X205         1.335e+00  1.005e+00   1.328 0.184085
X206         8.355e-01  9.958e-01   0.839 0.401442
X207         9.288e-01  9.752e-01   0.952 0.340900
X208         1.367e+00  1.026e+00   1.332 0.182931
X209         2.078e+00  9.859e-01   2.108 0.035047 *
X210         3.356e+00  9.981e-01   3.363 0.000772 ***
X211         4.990e-01  9.761e-01   0.511 0.609231
X212         1.823e+00  9.982e-01   1.826 0.067864 .
X213         3.085e+00  1.022e+00   3.017 0.002549 **
X214         3.000e+00  9.950e-01   3.016 0.002565 **
X215         1.920e+00  9.903e-01   1.939 0.052549 .
X216        -7.898e-01  9.597e-01  -0.823 0.410530
X217        -1.170e+00  9.755e-01  -1.200 0.230263
X218         7.398e-01  1.002e+00   0.739 0.460201
X219         9.754e-02  9.842e-01   0.099 0.921051
X220         6.804e-01  9.941e-01   0.684 0.493678
X221         1.195e+00  9.850e-01   1.213 0.225030
X222         9.760e-01  1.001e+00   0.975 0.329624
X223         7.381e-01  9.754e-01   0.757 0.449221
X224        -3.562e-01  9.625e-01  -0.370 0.711340
X225         1.502e+00  9.605e-01   1.564 0.117734
X226         1.461e+00  8.403e-01   1.739 0.082104 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2772.6  on 1999  degrees of freedom
Residual deviance: 1587.4  on 1802  degrees of freedom
AIC: 1983.4

Number of Fisher Scoring iterations: 15
```

Figure 16. Logistic regression model on bound and unbound data (1-mer+shape)

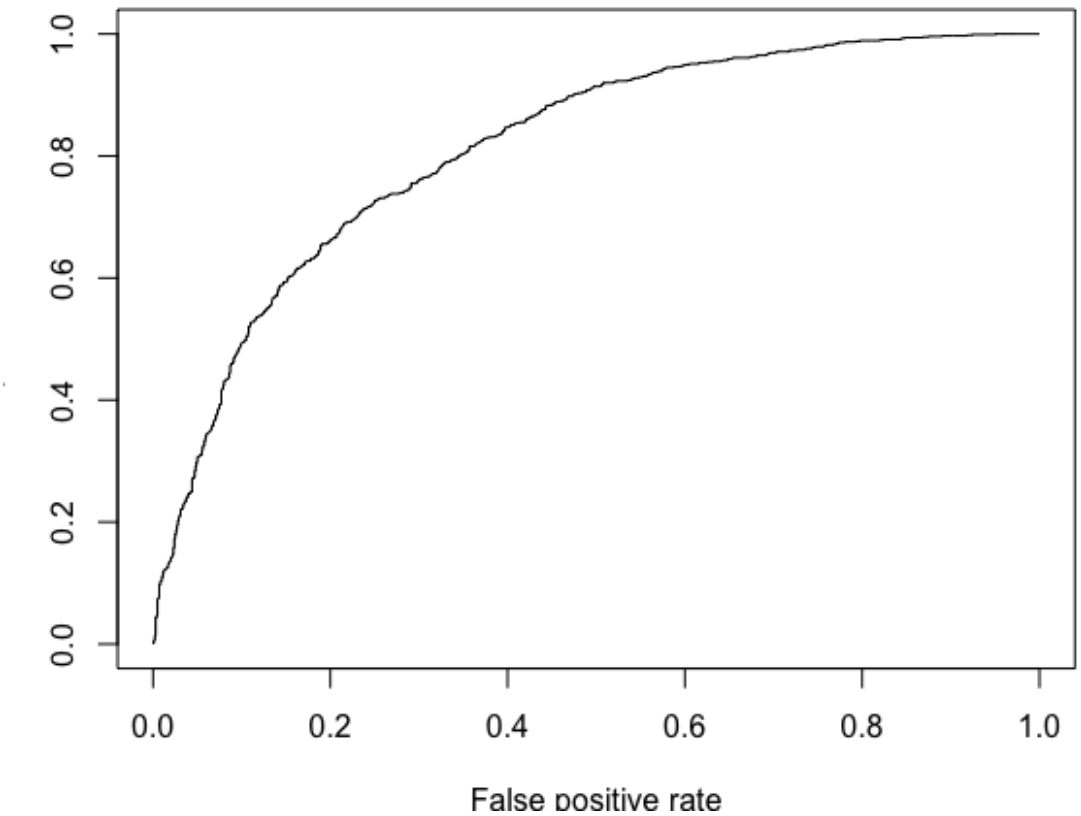

Figure 17. ROC plot of logistic regression model on bound and unbound data (1-mer+shape). AUC = 0.813343