2)

**1. Protein Binding Microarray (PBM)**

PBMs use microarrays with double-stranded DNA probes to measure the fluorescence of alphaGST-tagged proteins bound to their sequence-specific binding sites on the probes. There are quite a few different PBM designs – with the UPBM, ME, and HK being the most popular. All three of these designs have roughly 44K DNA probe sequences that are engineered using deBruijn sequences to cover (nearly) all overlapping 10-mers at least once.

**Advantages:** Fast, relatively inexpensive, and provides real-valued measurements of binding.

**Disadvantages (2015):** Most array designs only cover all 10-mers, making it difficult to model binding sites larger than 10 base pairs. Contains a strong artifact of preferential binding at binding positions near the free-ends of the probes.

**2. SELEX-seq/HT-SELEX**

The SELEX-seq (also called HT-SELEX) method combines classical protein-DNA SELEX (Systematic Evolution of Ligands by EXponential enrichment) assays with massively parallel sequencing. Before the first round of SELEX, an oligonucleotide containing a randomized region that is flanked by defined primer docking sites is used to bind the protein complex of interest. DNA bound by the complex is then separated from unbound DNA and the bound DNA is then amplified by PCR and used for subsequent rounds of DNA binding and selection. SELEX-seq leverages the depth of next generation sequencing to characterize millions of selected DNA molecules at each round of selection. For each round, SELEX-seq produces integer-valued, poisson-distributed sequence read counts for all the selected reads.

**Advantages:** No limit to the size of the binding site. Can capture binding measurements of large protein complexes. Only the depth of sequencing limits quantity of binding data.

**Disadvantages (2015):** Relatively expensive. The initial pool always contains strong sequence biases and additional biases are introduced during the many rounds of PCR amplification. Provides integer-valued, poisson-distributed sequence read counts.

3. Chromatin immunoprecipitation (ChIP) is an important technique in the study of protein-gene interactions. Using ChIP, DNA-protein interactions are studied within the context of the cell. The basic steps in this technique are fixation, sonication, immunoprecipitation, and analysis of the immunoprecipitated DNA. Although ChIP is a very versatile tool, the procedure requires the optimization of reaction conditions.

**Advantage**:versatile tool. The first step of the technique is the cross-linking of DNA and proteins. Formaldehyde is the most commonly used cross-linking agent. The most important advantage of using formaldehyde is the ease of reversibility of the cross-links and its ability to form bonds that span a distance of approximately 2 angstrom.

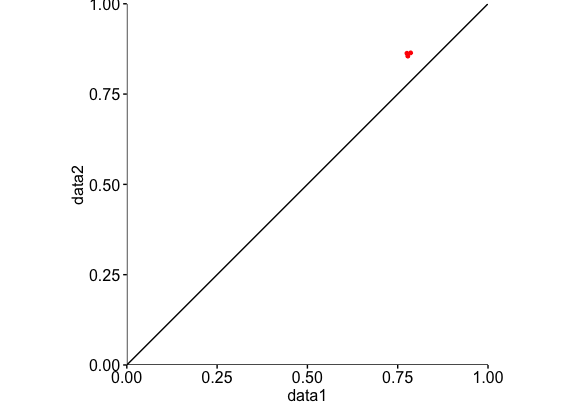
**Significant advantage** is that the genome coverage is not limited by the probe sequences fixed on the array. This is particularly important for analysis of repetitive regions of the genome, which are typically masked out on arrays.

**Disadvantage**: Although ChIP is a highly versatile procedure, it has several limitations and requires the optimization of conditions for a successful DNA extraction. The most vital step is probably the binding of antibody, and the quality of antibody is crucial for the recovery of DNA fragments.

4)

|  |  |  |
| --- | --- | --- |
|  | Shape and seq | Sequence |
| Myc | 0.85500 | 0.77794 |
| Max | 0.86412 | 0.78581 |
| Mad | 0.86296 | 0.77550 |

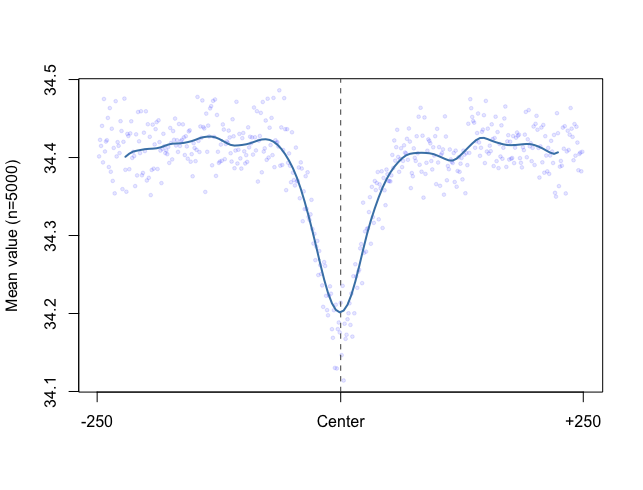
5)



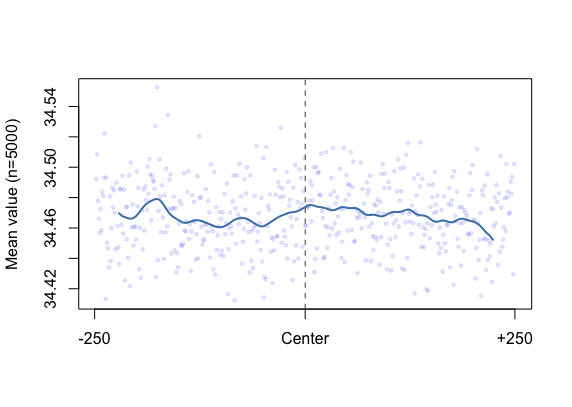
data1 <- c(0.77794, 0.785808, 0.7755)-X axis shows sq

data2 <- c(0.85500, 0.864119, 0.86296)-Y axis shows sq+shape

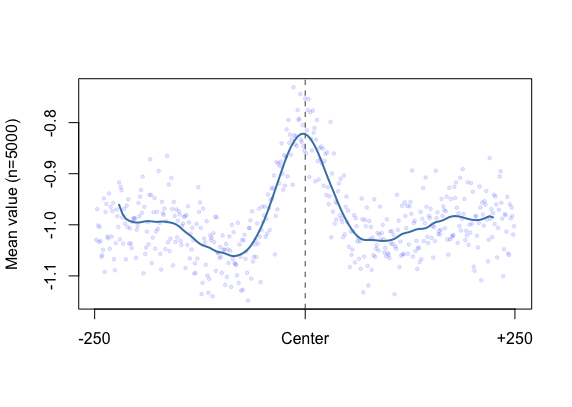
As table in question 3 and graph in question 4 shows our R^2 values for shape and sq. are closer to one, which means we have more accurate data. Our data set 1 and data set 2 are different by about 0.1, this tell us that shape is important because it gives us closer Regression value to 1.

7) Helical twist-bound

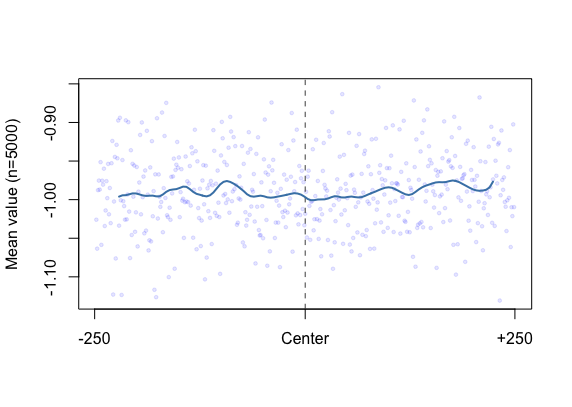
Helical twist unbound



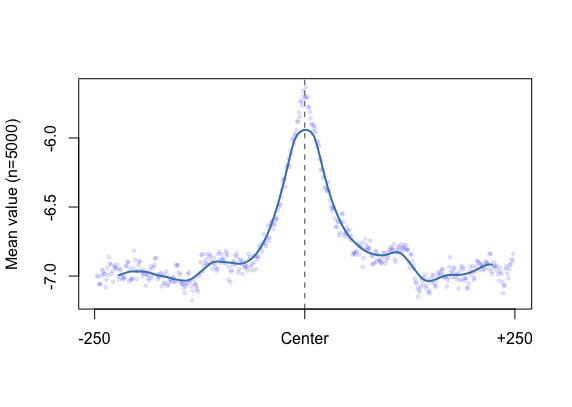
Roll-bound



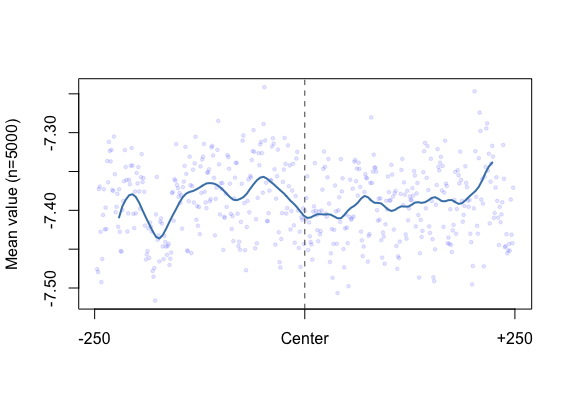
Roll Unbound



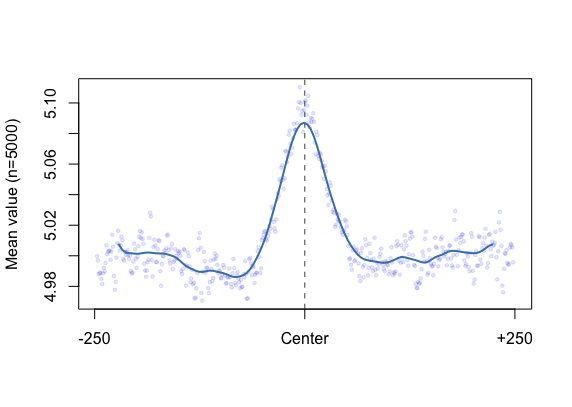
Prot-bound

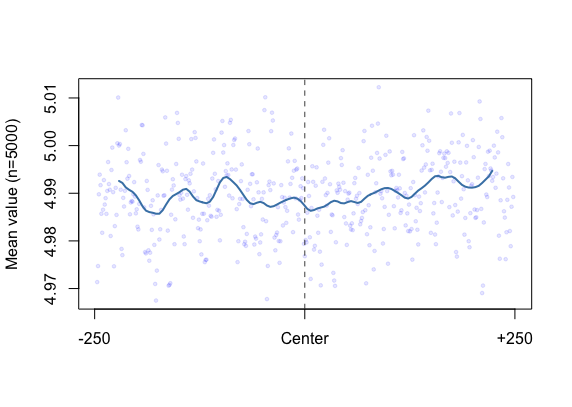


Prot Unbound



minor gorve-width-bound

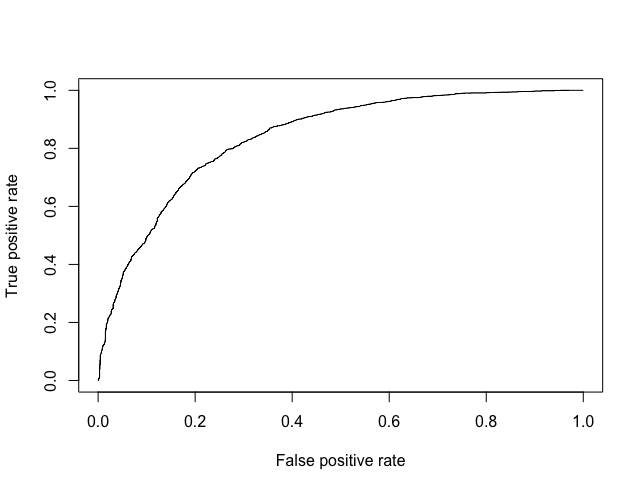


MGW-U

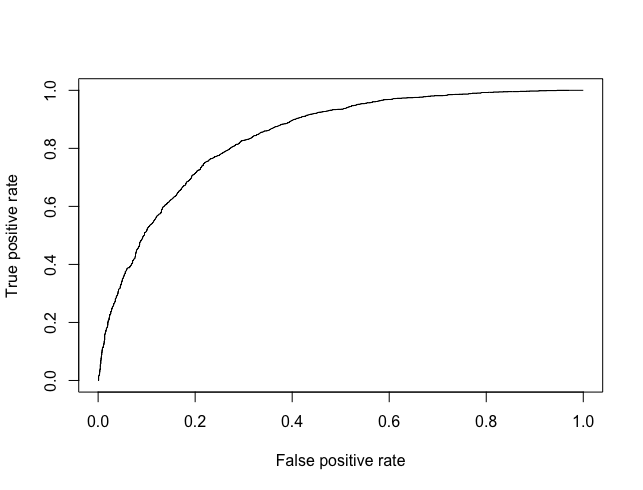
All the bound graphs for MGW, Prot , Roll, Helt follow the same shape except HElt has a decrease in the center instead of increase. All the unbound graphs also follow the similar shape and pattern of line. Our Y values for each graph are changing but the x-axis shows the same range of -250-250.

Our unbound graphs shows approximately linear graph where as our bound graphs shows hyperbola graphs. Our unbound sequences are random sequence because we took the average values. Our unbound graphs also are over fitting, rather than showing the general view of our model it shows so many details about our model. In regression analysis, over fitting a model is a real problem. An overfit model can cause the regression coefficients, p-values, and R-squared to be misleading.

8)seq+shape



seq



We have similar trend for sq+shape and sq only graph. We also do not have any value change; therefore we can conclude the shape is not important in these models. Our value for shape+sq=0.839 and for sq only was = 0.841, very close values again confirm our conclusion.