**Question 2**

**a.** SELEX-seq (Systematic Evolution of Ligands by EXponential Enrichment) and PBM) are both in vitro experiments. SELEX-Seq is now the optimal high-throughput technique for characterizing DNA-binding specificities of transcription factors. PBM (Protein binding microarray) characterize TFs' sequence specificities in a high-throughput manner. The PBM technology allows for determination of the DNA binding site specificity of proteins in a single day. The chip consists of a support surface such as a glass slide, nitrocellulose membrane, bead, or [microtitre plate](https://en.wikipedia.org/wiki/Microtitre_plate" \o "Microtitre plate), to which an array of capture proteins is bound.

**b.** ChIP-seq is used to analyze protein/DNA interactions. It combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

**c.** The SELEX-seq platform is flexible and can be used to determine the relative affinities to any DNA sequence for any transcription factor or multiprotein complex. PBM and SELEX-seq are both quantitative so they can describe the relative affinity of the protein-DNA interactions. Also they both can work for large amounts of protein.

Disadvantages of these methods are that can be hard to find a surface that can keep the proteins from denaturing and limiting their interactions with other molecules.

An advantage of the ChIP-seq method is the large quantities of data that it generates. A disadvantage is that it only gives qualitative binding data, and doesn’t give as much specifics on binding strength.

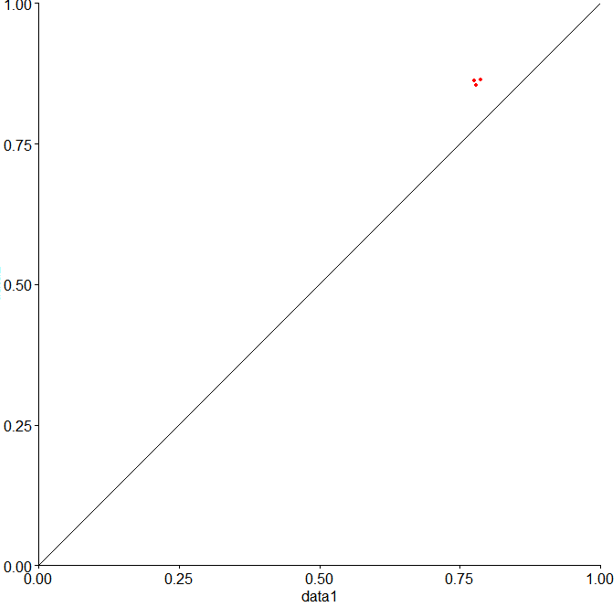
Compare and discuss the advantage and disadvantage of these methods

**Question 4**

|  |  |  |
| --- | --- | --- |
| Filename | 1-mer | 1-mer plus shape |
| Mad.txt | 0.775607 | 0.862834 |
| Max.txt | 0.786137 | 0.864298 |
| Myc.txt | 0.778085 | 0.854663 |

**Question 5**

**a.**

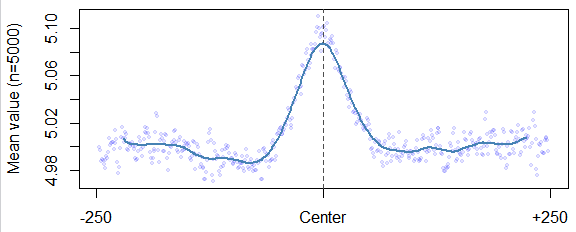


**b. Data 1** represents the 1-mer R2 values on the x-axis, and **Data 2** represents the 1-mer + shape R2 values on the y-axis. The three data points are all above the y=x line, meaning that the R2 values for the 1-mer + shape are higher than the R2 values for the 1-mer across the board. Sometimes, adding factors that are not relevant to the model will not improve the R2 values. This is not the case here; adding the shape parameter to the model improved the R2 values and thus the efficacy of the model, and the close clustering of the data points suggests that this is true for all of the values.

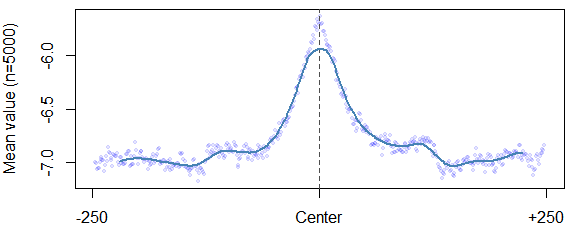
**Question 7**

**a.**

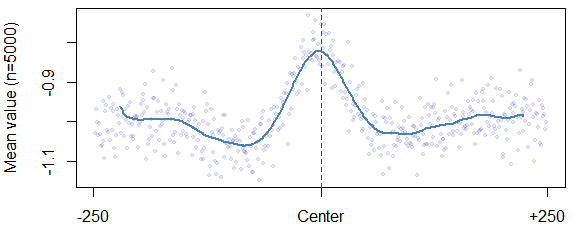
**plotShape(pred$MGW)**

****

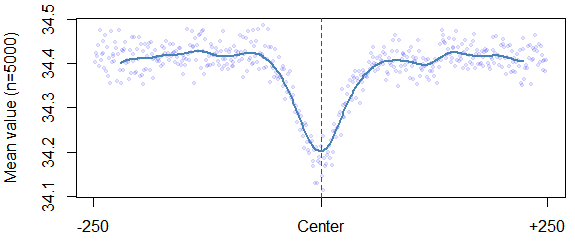
**plotShape(pred$ProT)**

****

**plotShape(pred$Roll)**

****

**plotShape(pred$HelT)**

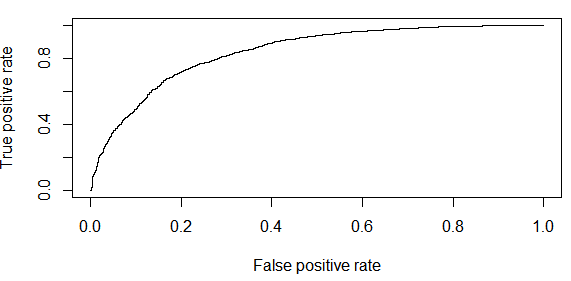
****

**b.** The bound used Chip-Seq and unbound was randomly chosen from the genome. The middle part of the graph with a peak is where the minor groove width is larger and this is where it is recognized by the transcription factors. This is also true for propeller twist and roll. However, for helical twist, the center is where it is the smallest, and this is where the transcription factors recognize it.

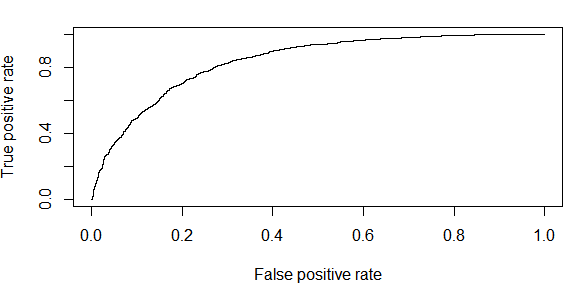
**Question 8**

**a.**

**1-mer + shape----AUC: 0.8390385**



**1-mer----AUC: 0.8380748**



**b.** The AUC for 1-mer + shape is slightly larger than the AUC for the 1-mer but in fact the values are very similar. The difference in these values may be due to random factors, and the numbers could even change after a few runs. This means that adding the shape as a parameter doesn't really improve the accuracy of the model, but it doesn’t worsen it either. It has little effect.