Robert Hyun

HW#3

2. SELEX-seq and PBM are examples of in vitro, high-throughput binding assays. Both methods select desired oligonucleotides from a pool via selection using target protein. Proteins that are bound through selection then are amplified and the amplified oligonucleotides are then sequenced.

For PBM, the process uses a dsDNA microarray and epitome-tagged TF is bound to the dsDNA microarray. The epitope is labeled with a fluorophore-tagged antibody making the desired dsDNA visible by scanning.

For SELEX-seq, desired oligonucleotides are selected by the protein ExdHox and are amplified by PCR. This process is continued multiple times in order to better isolate the wanted oligonucleotide.

ChIP-seq is an in vivo experiment that also utilizes high-throughput binding assay that analyzes protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. From the cells themselves the DNA is cross link and sonicated. Specific antibody is added to select for desired DNA and then separated through immuno-precipitation. The crosslinks are then reversed, the DNA labeled then put through sequencing.

The advantage of PBM is that it is fast, relatively inexpensive and provides real-valued measurements of binding but the disadvantage is that it is difficult for it to model binding sites larger then 10 base pairs and has preferential binding at a binding positions near the free-ends of the probes.

The advantage of SELEX-seq is that there is no limit of the size of the binding site and can capture binding measurements of large protein complexes. The disadvantage is that it is relatively expensive and many rounds of PCR amplification introduce additional biases to already sequence biases to the initial pool of oligonucleotides.

ChIP-seq has genome coverage that is not limited by the probe sequences fixed on the array but it is, like SELEX-seq, expensive. Also when there is not a sufficient amount of readings, here is a significant loss of sensitivity or specificity. For example, ChIP-seq has a bias towards GC-rich content in fragment selection.

For the coding, using DNAshapeR package in order to generate a feature vector of “1-mer” sequence model and a feature vector for “1-mer+shape” model for the proteins illuminated what was calculated for each of these models. The 1-mer model lists the nucleotide sequence using groupings of four numbers in binary code. The 1-shape model gives data to quantitatively depict the structure of the protein’s DNA. In order to get the values for each of the proteins for both of these models, the code was kept the same but the data retrieval for the protein and the featureType code was changed indicting whether the model would be 1-mer or 1-mer+shape. The two feature vectors then were used to build a L2-regularized MLR model for the 1-mer and 1-mer+shape features using 10-fold cross validation. Prior to this, prediction without L2-regularied MLR model was done in order to get the R^2 value when the prediction with the L2-Regularized MLR model was conducted. These predictions were made possible by the caret package. The folds are selected at random but the average R^2 value was obtained for both of these features from the three proteins: *Mad, Max, Myc*.

5. From the L2-regularized MLR model it was shown that the high-throughput in vitro data analysis was more accurate in deciphering the DNA strucure when given access to the 1-mer+shape data as opposed to just the 1-mer data (sequence). The 1-mer vs 1-mer+shape pdf has three points and all of them are above m=1. As the x axis of this graph is 1-mer and the y-axis is 1-mer+shape, it is shown quantitatively that the 1-mer+shape feature vector was more accurate. This indicates that the process of determining DNA structure for these three proteins becomes more accurate when sequence and data of shape is provided. The R^2 values were put into Data preparation with data1 being 1-mer and the x axis and data2 being the 1-mer+shape data and on the y-axis.

For the coding, the different values for the DNA shape parameters: minor groove, propeller twist, Roll, and helix twist were obtained by substituting the appropriate code for the DNAshapeR prediction.

Result Analysis

7. Plot Shape analysis of in vivo data

1. Minor Groove Width [MGW]- the data shows that the mean value for width is not symmetrical about the y-axis where the center is. Also of note is that the minor groove has the most width in the center and width of the minor groove is not necessarily proportional to distance away from the center. The minor groove in this case appears to have a shape that is not symmetrical from the center.
2. Propeller Twist- appears to be most significant at the center and tapers off as distance from the center increases at a certain point. The further away the point from the center is, the more variability is observed as the data making up the average is most wide spread at the values furthest away from the center in the x-axis.
3. Roll- For Roll, the data at every x-axis shows a lot of variability, even x-values that are close to the center. This indicates that proximity to center does not have as much effect in accuracy for Roll as observed in the data for Propeller Twist. We learn that Propeller Twist values are the highest when obtained from the center.
4. Helix Twist- There is less variability for values for helix twist when observed near the center. However unlike the other DNA shape parameters, the helix twist has the lowest value at the center. Like the data for Minor groove width and Propeller Twist there is more variability for values when the point of the DNA analyzed is far from the center.

For all of these tables, the data was analyzed via plotShape. The dark blue line indicates the average value of all the data in one point of the x-axis. These values were obtained with a sample size of 500 while the next procedure used a sample size of 30.

8. For the in vitro data, we built a logistic regression model for 1-mer and 1-mer+shape. Of note is there high AUC score which indicates that the model does well in representing the DNA sequence. The DNA sequence is known and the model for both 1-mer and 1-mer+shape features does well in predicting the sequence. From the results of the prediction model for the in vitro data, we look at also the ROC curve to notice that whether the feature was 1-mer or 1-mer+shape there was no significant difference shown in the ROC curve. This indicates that the DNA shape is not necessary for the protein in question to determine the structure of the DNA. It appears that sequence is enough for the protein to determine the DNA structure satisfactorily and the addition information provided 1-shape does not make a difference in accuracy.

For the coding for the ROC curves we used both unbound and bound data to determine how accurate the prediction model is by looking at the generated AUC value and ROC curve. Through the code we merged the two data sets bound and unbound in order to be able to generate the feature vectors for 1-mer and 1-mer+shape, a process similar for the high throughput analysis of *Mad, Max and Myc*.