BISC 481

HW 3

2.

A. SELEX-Seq is a method that determines the binding site of proteins. First, DNA templates need to be produced to make the entire RNA pool from which the protein binding will be tested. Then, the pool has to be marked with a label, such that if the protein binds to a sequence, it will be tagged so that it can be identified. Protein is added to the RNA pool, and the bound RNA is identified and used to make more RNA to narrow down the pool. The process is repeated until most of the pool can bind to the protein.

Protein Binding Microarray is a method similar to SELEX Seq, in that it determines the binding site of proteins. It uses a chip of nitrocellulose membrane. An array of DNA sequences are stuck to the chip so that the fluorescently tagged proteins (fluorescence is only activated when proteins are bound) can bind to the DNA sequences. The affinity of the protein to the DNA is shown by the intensity of the fluorescence which is scanned to determine protein/transcription binding signal per sequence. It also provides information on the relative quantitative amounts of protein bound to each DNA sequence.

B. CHIP Seq identifies binding sites of DNA associated proteins inside cells and provides information on transcription factor binding and gene expression. The library of DNA sequences comes from DNA found in vivo. Cross linked DNA-protein complexes are enriched using antibodies. The crosslinks are then precipitated out of the cell content to be isolated and identified.

C. Advantages: The in vitro methods do not limit the DNA library to what is found in test cells, so it can have a broader range and better test for mutated sequences’ binding affinity. In vivo methods, however, account for DNA as it is in cells, like gene silencing and chromatin remodeling.

Disadvantages: In vitro methods do not account for gene regulation. Also the use of fluorescence in PBM only provides relative binding strength. CHIP Seq uses antibodies which are not always available for every protein and are expensive.

5. The plot shows data points from the R squared values of three proteins calculated from the two methods 1-mer and 1-mer+shape. The data points lie above the line with the slope of 1, which shoes that the R squared values of 1-mer+shape are higher than those of 1-mer. This makes sense because 1-mer+shape provides additional information so one would expect it to be a better model.

7. Minor Groove Width: The plot shows that the width of the minor groove

widens at its center (of the sequence) and narrows out. The minor groove width

shows a slight asymmetry that could contribute to increased binding specificity as

transcription factors generally have an “irregular shape.”

Propeller Twist: The degree of twist refers to the intrabase relationship. The

propeller twist stays in the range of -7 to -6.5 degrees till the center of the sequence

where it heads towards -6 degrees. This can cause a noticeable change in shape in

the local region that can contribute to binding specificity as it can change a protein’s

access to base pairs.



Helix Twist: Helix twist provides insight into the winding of DNA. At the

center of the sequence the helix twist is slightly lower potentially making the DNA

unwind a little locally which could influence binding specificity and accessibility.



Roll: Roll is a interbase pair measure, which opens up space between base

pairs. It becomes less negative at the center, which differentially provides access,

changing shape locally.



8. The accuracy of each model is measured by how well the model differentiates

between “true positives” and “false positives” shown by the ROC curve. The AUC

score is determined by calculating the area under the curve. If the model was

perfect, it would have a score of 1. The 1-mer+shape model had a AUC score of

0.8377604 while the 1-mer model had a AUC score of 0.8421153 suggesting it is

more accurate.

1-mer+shape:



1-mer only

