Results:

Hox proteins are known to bind to TGAYNNAY motif, recently, Donaldson et al. have published ChIP-seq data for Hoxa2 in mouse [1]. We have used our minor groove width predictor in order to predict the minor groove width of each position in the ChIP-seq data. Next we have translated the 4 sequence letters of the ChIP-seq data to an 8 sequence+shape letters based on table1. We are interested in searching for the sequence and the sequence+shape binding motif of Hoxa2, to this end we have used DRIMUST and MEME in order to search for over represented motifs. DRIMUST identifies sequences that tend to appear at the top of a ranked list more often than in the rest of the list. This approach might help in eliminating sequence+shape motifs that are enriched even in random data as a result of shape that is highly dependent on sequence (such as the E-box) (I will elaborate on the dependency problem in the next paragraph). We have used DRIMUST for the sequence ChIP-seq data and were successfully able to find TGATGGAT motif (table2). We also used DRIMUST for the sequence+shape ChIP-seq data and again we were able to find TGATGGAT motif (table2) but this time we were able to add shape information to the sequence motif. Hoxa2 is the ortholog of Pb in drosophila (which is an anterior Hox protein), we have previously published the binding shape preferences of Hox proteins in drosophila and have found that Pb binds to sequences which have 2 narrow minor groove regions. Interestingly, using DRIMUST for ChIP-seq data in mouse we were able to find the same shape preferences for Hoxa2 (two narrow minor groove regions in position 3-4 and 8). We have also used MEME to search for over represented motifs (table3,4). MEME is a tool for discovering motifs in a group of related DNA or protein sequences. We have used MEME for the top 200 sequences in the ChIP-seq data and found general resemblances between the DRIMUST motifs and the MEME motif (table3). It is important to note that MEME discover more than one sequence+shape motif with low p-values, such as the E-box (which is enriched even in random data sets). This emphasizes the shape dependency problem. MEME is also able to perform discriminative motif discovery (to find motifs overrepresented in one set of sequences compared to in another set). To deal with the dependency problem and eliminate motifs that are also enriched in random data, we have also used MEME with the ChIP-seq top 200 sequences compare to the ChIP-seq bottom 200 sequences (table4). We have found that even with the discriminative motif discovery, the sequence and sequence+shape motifs are the same as the motif we found using DRIMUST. As this motif was enriched only on the top of the list, this suggests that the motif is specific to the binding of Hoxa2.

Shape dependency problem:
The minor groove width predictor is using 5-mer window, which means that any specific 5-mer enriched motif will always have the same minor groove width in position 3. We want to be able to distinguish between position that conserved in shape as a result of the protein binding preferences and position that are conserved in shape as a result of sequence conservation. In different words, we are interested in understanding how much information is added from the shape that we could not learn from using sequence alone. To do that, we looked at the shape of the motifs that were enriched at the top of the ChIP-seq list (and are assumed to have the binding site of the protein) compared to the shape of those motifs in all of the data (which are assumed to also have random, non-binding sequences) (for more information on the method see method paragraph on the bottom). As you can see from table5, it’s hard to see a difference between the shape preferences in the top of the list compared to the full list. This might be because it is hard to look at the shape differences with only 2 letters (lower case for narrow and upper case for wide). In order to future study the shape preferences, we have calculated the
average minor groove width in each position in the top of the list and in all of the list (figure 1). Interestingly, now we can see that positions 2, 3 and 4 tend to have significantly narrower minor groove width in the top of the list compare to the full list. Also, position 9 tend to have significantly wider minor groove width on the top of the list compare to the full list. Those results suggest that the Hoxa2 indeed tend to bind to motifs with district shape.

Questions to myself:
Does using the sequence+shape alphabet really help us in understanding shape preferences? Or it would be better to find the enriched sequence motifs and calculate the minor groove width in each position (which is sort of simulating the approach for SELEX as we done before)

Table 1: translating 4 letters sequence alphabet to 8 letters.

<table>
<thead>
<tr>
<th>Minor groove width</th>
<th>Narrow &lt; 4.8Å</th>
<th>Wide=&gt; 4.8Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ade</td>
<td>a</td>
<td>A</td>
</tr>
<tr>
<td>Cyt</td>
<td>c</td>
<td>C</td>
</tr>
<tr>
<td>Gua</td>
<td>g</td>
<td>G</td>
</tr>
<tr>
<td>Thy</td>
<td>t</td>
<td>T</td>
</tr>
</tbody>
</table>

Table 2: DRIMUST enriched motifs

<table>
<thead>
<tr>
<th>Seq</th>
<th>Seq + Shape</th>
</tr>
</thead>
</table>

Table 3: MEME enriched motifs

<table>
<thead>
<tr>
<th>Seq</th>
<th>Seq + Shape</th>
</tr>
</thead>
</table>
Table 4: MEME with discriminative motif search

<table>
<thead>
<tr>
<th>Seq</th>
<th>Seq + Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Memetic Pattern" /></td>
<td><img src="image2" alt="Memetic Pattern" /></td>
</tr>
</tbody>
</table>

Table 5: WebLogo for the enriched motifs (for more information see method).

<table>
<thead>
<tr>
<th>Full list</th>
<th>Top 100</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="WebLogo Pattern" /></td>
<td><img src="image4" alt="WebLogo Pattern" /></td>
</tr>
</tbody>
</table>

Figure 1: Average minor groove width in each position. Blue: when taking into account all of the ChIP-seq data. Green: when taking into account only the top 100 sequences in the ChIP-seq data. For more information see method.
Method:
In order to study the shape dependency, we took all of the sequence+shape motifs that DRIMUST found enriched in the ChIP-seq data (which constructed the PSSM in table2 seq+shape), and we translated them back to 4 letters sequence alphabet. We next search for all of those motifs in the ChIP-seq data (if from example DRIMUST found the sequence+shape motif tgAT we searched for the sequence TGAT without taking shape into account). Looking at the flanking regions of those motifs in the ChIP-seq data (2 positions from each size), we calculated the minor groove width for each position of the motif. In order to study shape dependency, we used this approach for the top 100 sequences in the ChIP-seq data, and compared it to the results of this approach for all of the sequences in the ChIP-seq data (see results in table5 and figure1).

For next week I am planning to:
- Adding p-values to the box plot.
- Doing the box plot for the sequence motifs (and not for the sequence+shape motifs) to see if using DRIMUST with 8 letters help in finding the shape preferences.
- Use MEME for the Harbison data.

References: