On Drosophila Chip Explore and Drosophila SELEX

By Michael Krakoff

Advisors: Professor Peter Bickel and Post-Doc Ben Brown

Introduction

For my summer research as provided by the VIGRE grant, I worked under the grand instruction of Professor Peter Bickel, but primarily under the guidance of post-doctoral student Ben Brown and Nathan Boley. The purpose of their research is to better understand the human genome, specifically with respect to DNA and RNA sequencing as well protein binding sites. The focus of my research concerned itself with dealing with a Chip-Explore and SELEX experiments, done on the Drosophila fruit fly.

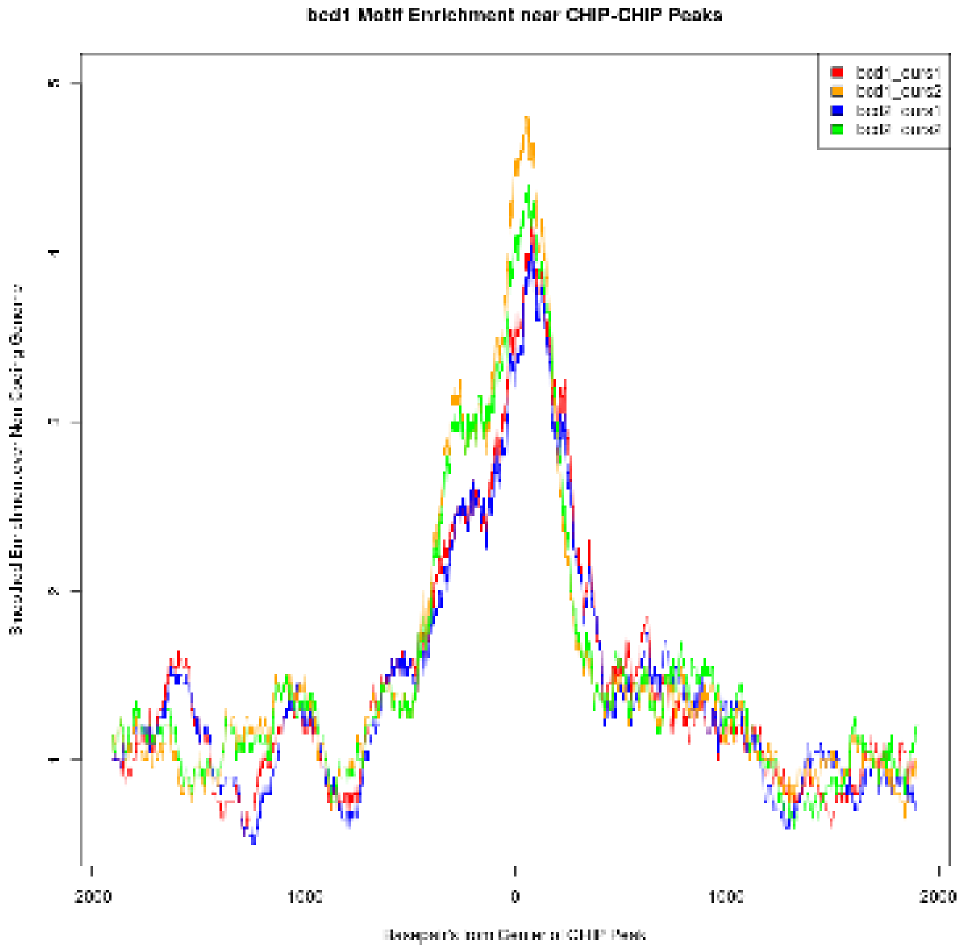
SELEX

While working with the SELEX experiment, I attempted through maximum likelihood estimation to attempt to better estimates for the entries in the energy matrix for each of the four base pairs: A, C, G, T. The protein in which I principally concerned myself, and on which the results presented below are based, is a bicoid. The likelihood function used is non-parametric in form, and features three independent variables and thus four dimensions. These variables to be controlled for are the: non-specific binding site, well-folded fraction, and lastly the additive energy. For the bicoid protein, a notable binding sequence is TAAT. Thus, I began with an energy matrix consisting of zeros corresponding to the position of the given base pair; that is to say that the first row had zero in the A entry, the second row has a zero in the T entry, etc. The idea being that one can add an additive energy to each non-zero base pair, thereby altering the energy matrix and helping to improve the maximum likelihood estimation. After setting a range of values for each of the three variables, and setting a step size, one can run as many k * j * i simulations where selected indices correspond to each the number of iterations for each of the specified variables. For the given base pair sequence and combination of variable values, once the value of the likelihood function appears to not be changing, one records the values for the additive energy, non-specific binding site, and well-folded fraction and enters them into an optimizing algorithm called “Cobayla”. The Cobayla algorithm seeks to further optimize the values of the additive energy matrix which can further increase our estimate for the likelihood. In doing so, a new added to the additive energy matrix, and accordingly a new base pair is added on to the original base pair sequence.

Continuing to work on the bicoid protein, I repeated the above process, continuing to add on to the base pair sequence and thus the energy matrix, until I had created a ten base pair sequence. At each step, I continued to see that I was increasing the likelihood estimate, which was the goal all along. It was important to observe that none of the non zero entries in the energy matrix became too large in scale, for that could indicate possible divergence. My final energy matrix is shown below:

```
-1.071028    0.000000   -0.167093   -2.264730
-2.681211    0.000000   -3.944833   -0.559008
-6.238646   -8.104211   -10.705240   0.000000
0.000000   -12.243909  -7.507240   -9.175781
0.000000   -16.282147  -9.394255   -17.563493
-8.952273   -10.740814  -6.242276    0.000000
-8.526554    0.000000   -12.347747  -6.964473
-5.732735    0.000000   -5.831922   -4.795989
-1.052385   -0.513766    0.000000   -2.764154
-1.064713    0.000000   -1.010647   -2.576833
```

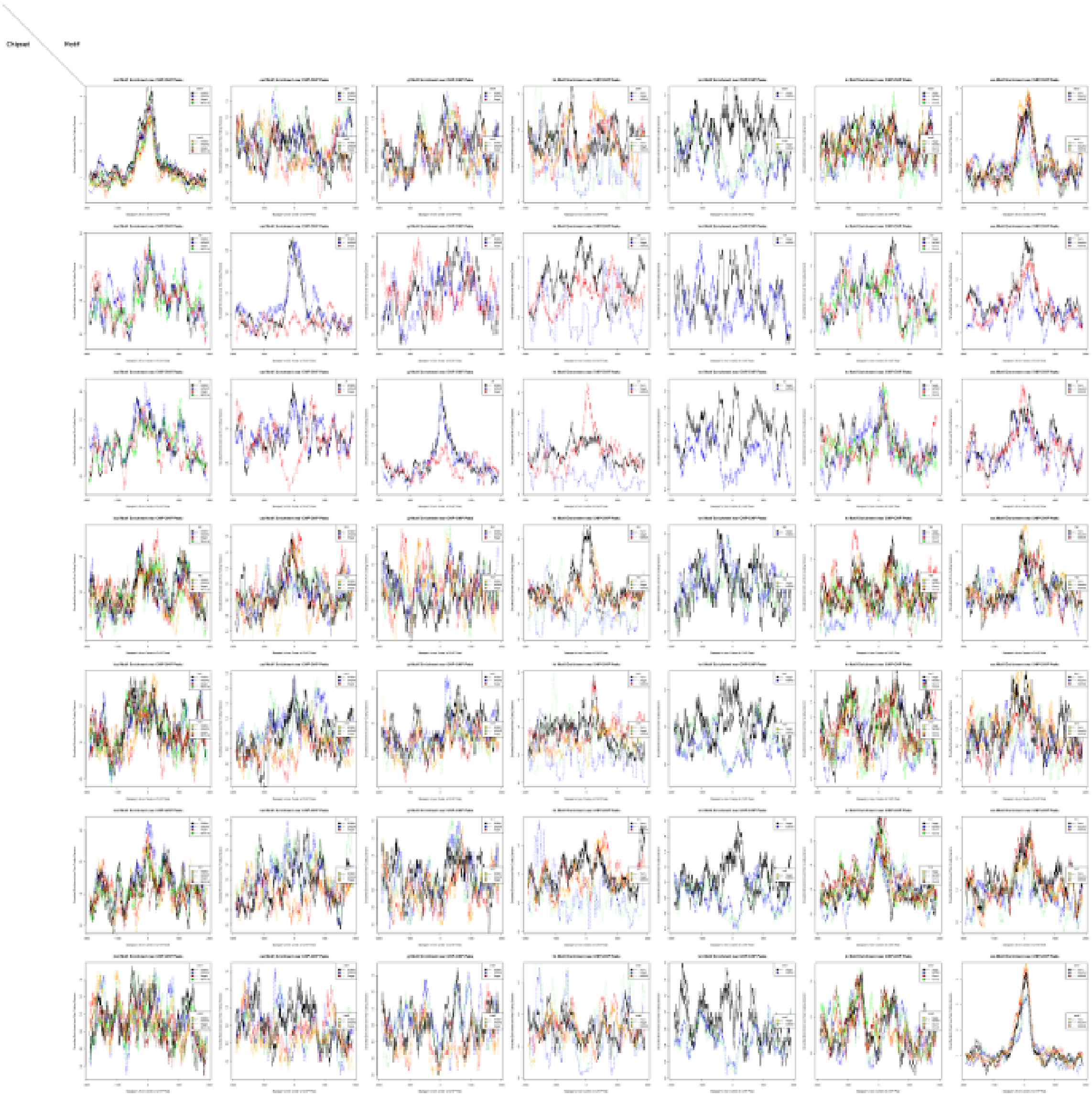
Note that the columns in corresponding order represent A, C, G, and T. The above matrix represents a motif which can be used in a Chip Explore experiment in an attempt to see if what is observed empirically agrees with what is to be expected under the assumptions of the model. To compare my motivated generated motif (above) to other motifs currently used, using a moving average kernel smoother, I plotted this ratio of observed over expected enrichment in the genome versus distance from the central peak (ideally enrichment should occur near or at a peak). This plot is shown below:



On the plot above, the red and blue curves represent enrichment plots generated using the motif generated above, while the orange and green represent plots based on a motif generated using similar methods. There are two plots per motif as we are plotting this motif against two different types of bicoid chipsets. The plots generated from the “ours2” motif seems to be a better fit than the plots generated by the “ours1” (my own) motif, as evidenced by the higher enrichment value and that this enrichment occurred at a point closer to zero. Nevertheless, it appears that the motif I generated appears to have a fair amount of validity as the enrichment occurs near the expected distance of zero, and for the most part, the enrichment value remains high for each chipset used.

Chip Explore

While working on the Chip Explore and subsequently Chip Enrichment projects, I attempted to explore if a motif generated in a SELEX experiment for a specific protein (recall the motif generated for bicoid previously) would cause enrichment when used against a foreign chipset unrelated to the given protein. The seven proteins used are: bicoid (bcd), caudal (cad), giant (gt), hunchback (hb), knirps (kni), kruppel (kr), and snail (sna). For each of the seven proteins, I created an enrichment plot using the protein's own chip set as well its own set of motifs, and then continued to use the protein's own chip set, yet now use each of the other six motifs created seven plots for a given a chipset. I thus created a seven by seven plot matrix shown on the next page (due to size limitations), where the rows correspond to specific chip set, while the columns correspond to the motif specified for a chipset.



Though difficult to make out, the rows and columns are ordered in the following way: bcd, cad, gt, hb, kni, kr, and sna. Recall that rows represent a given chipset, while the columns a different motif. The diagonal of the plot matrix accordingly represents each motif being used for its intended chipset. This should be clear as there appears to be significant enrichment around zero on each of those plots as evidenced by the clear and defined peaks. For the most part, it appears that when using a motif designated for its own specific chipset on a foreign chipset, it only produces noise and that there is no significant enrichment. Perhaps most interesting to note, is that the motifs generated for the snail protein appear to induce significant enrichment when run on a foreign chipset. This is most evident for the bicoid and caudal proteins, as shown in entries (1,7) and (2,7) of the plot matrix.