

A Model for SELEX Data

In Spring of 2009, I worked with Peter Bickel, Juli Atherton, and Ben Brown to analyze SELEX data collected at the Berkeley Drosophila Transcription Network Project (BDNTP). We developed a new binding model, software to fit the model, and fit models for several Drosophila transcription factors. We also developed a method and corresponding software to assess the quality of our model when compared to Chip-Chip data.

A typical SELEX experiment begins in round 0 with a random pool of oligonucleotides in solution with a target. In the BDNTP data, the oligos were 16 bp double stranded DNA sequences and the target was a transcription factor found in Drosophila. Eventually, a dynamic equilibrium is reached in the solution such that the concentration of unbound transcription factor is constant. Then, the bound oligos are separated from the solution, the transcription factor is separated from the DNA, and the DNA is amplified via PCR. After amplification a sample is taken from the amplified DNA and sequenced. The remaining DNA is added into round 1 and the process is repeated.

Noting that the bound oligos are separated when the solution is at equilibrium, one can write down the probability that an oligo is bound given its binding affinity and the concentration of unbound transcription factor. We assume that each 16 oligo sequence contains a shorter subsequence that the target binds to. Furthermore, we assume that each bp type at each position in this subsequence contributes an additive term to the total binding affinity that is independent of its neighboring basepairs. We call this set of energies the energy matrix. One of the notable points of our method is that it determines (aligns) the subsequences while it fits the model.

My contribution was primarily in developing the software to fit the energy matrix. I used the method of maximum likelihood to fit the model. Since the likelihood surfaces that typical SELEX data produce are not smooth and because they have multiple, deep, local minimums finding the maximum is not easy. I wrote the code that evaluates the lhd given an energy matrix and set of transcription factor concentrations in C. Then, I wrote a python interface to the C code and used several of the methods in the scipy optimization package to fit the model. Because the lhd surfaces are generally so nasty, most of the techniques are not particularly effective. However, given a decent starting location, the constrained Nelder-Mead algorithm (COBYLA) works reasonably well. In fact, when compared to chip-chip data, several of the fit models were better than any model that has been published up until now.

I also developed software to compare the fit SELEX model to chip-chip data. The approach is as follows. We take the center of the top 1000 called peaks from a chip-chip experiment on the same transcription factor. Then, for every contiguous subsequence in the 500 bp flanking regions of every peak, we calculate the empirical distribution of scores (in this case binding energies). We then take the score at the 0.1 percentile, and plot every point that is greater than this cut-off score. Finally, we smooth the plot with a 200 bp moving average. For the majority of factors, the resulting plot's clearly show a correspondence between the high energy binding motifs as determined from SELEX and the regions called through the chip-chip experiments.