This is an update on the Fkh project in which I have been working for the past couple weeks. In this project, we wanted to analyze covariation between protein sequence and DNA sequence/shape using PBM data from Bulyk’s lab. Starting with data for various Fkh proteins, we started with full probe sequences with normalized signal-intensity. First analysis, I did not get any correlation between DNA sequence/shape and protein sequence. The following is what was done and a few things that could be done:

1) FIMO was used to get the sequences, but with this approach I get a lot of probes that had low signal intensity from PBM, and, at the same time, I do not get a lot of the top scoring ones depending on the cutoff that is used. I used the PWM given in the paper, and as a possible alternative, I can either try different cutoffs or, as suggested by Iris, try to use MEME and find the motif enriched on the top high intensity probes.

2) For the Fkh case, if I try to align the DNA motifs from different Fkh proteins to get a pairwise similarity score, I end up having some of them not aligned as the motifs can vary a lot (as an example RYAAAYA vs AYGC). I can't really think of any other way to analyze them all together - only grouping them according to the pwm type, but any suggestions are welcome.

3) If I try getting the correlation between protein and DNA similarity (either shape or sequence), I get no correlation at all, which I hope is not the case. I tried calculating the similarity using the blosum algorithm, and I tried both in clustalw and R, but the scores vary a lot depending on how the alignment is done. I schedule to talk to Iris about this on Tuesday, and hopefully then I will rerun the analysis and have better results this time.