Group Meeting_Fangjie (Last: 2018-10)

> Recent progress

_ Arin's projects (1. nucleosome's preference on methylated DNA; 2. PCR bias)
_ SELEX with different temperatures and solvents
_ Quantification of TF's affinity

> SELEX with different temperatures and solvents

> Introduction

In the research of TF-DNA interaction, ultimately we hope to predict TFs' binding specificity from its structure. Such prediction can be represented by a model that takes protein surface features as input, and outputs the bound DNA sequences. However, such model would require much domain knowledge rather than being able to train completely de novo, because the amount of available protein structures is limited.

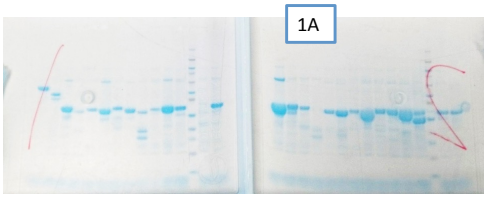
Further complication to the prediction is that, TFs can recognize distinctly different DNA sequences (E. Morgunova, et al., Elife 2018, 7) with different types of contacts. During the model training, it is benificial to separate specificities induced by direct contacts from those induced by indirect contacts, to reduce the complexity of the model.

The indirect TF-DNA contacts usually involve water-mediated hydrogen bounds, thus are less entropically favorable, and expected to be more sensitive to temperature and solvent changes. In light of such mechanism, we peformed SELEX under different temperatures and with different solvents, aiming to separate TFs' binding specificities into different categories.

> Experimental

TF expression

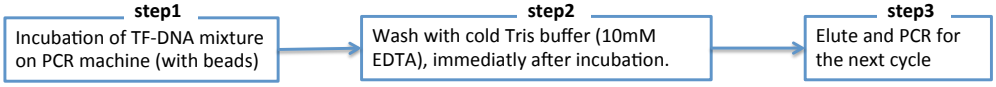
worked as usual (Fig 1A), also tested protein quantification with Bradford assay (Fig 1B). Most wells show visible color change compared to the background (buffer only, indicated with the red frame). It is possible to measure protein conc. In high-throughput



Preparation of c0 ligands

To obtain more counts for high-infomatin-content TF binding modes, with the expressed TFs, we performed 2 cycles of normal SELEX, and used the enriched ligands as cyc0 for this experiment. To also allow completely new binding modes to develop, the original cyc0 ligand (w/o TF signal) is also added.

SELEX with different temperatures and solvents



_ The aim of step 1 is to incubate under a constant temperature for long enough, to allow the reaction to reach equilibrium, thus the selectivity difference due to temperature and solvents can develop.
_ The aim of step 2 is to prevent dissociation as much as possible, because technically we cannot perform the washing step under different temperatures, and economically we can only use water as the washing solvent.

Conditions tested

_ Temperature effect

Temperature: 0°C, 10°C, 20°C, 30°C, 40°C, 50°C Solvent: H2O

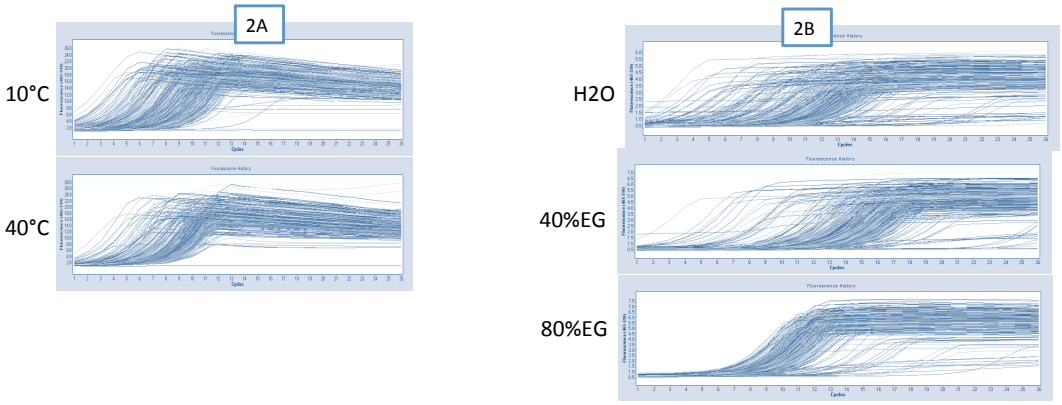
_ Solvent effect

Temperature: 25°C Solvent: H2O, 40%D2O, 80%D2O, 40%EG, 80%EG (EG: ethylene glycol)
Testing higher percentage of D2O or EG would be desirable, if the H2O is tightly bound at the TF-DNA interface. However, as the c0 DNA, TF protein, and 10x binding buffer are all prepared with H2O , it is much labour to further increase the solvent percentage.

_ Combined

Temperature: -20°C Solvent: 40% EG

From qPCR, it is visible that TFs bind less DNA at higher temperature (2A). D2O do not have an obvious effect. EG significantly reduced the amount of bound DNA (2B), likely due to the precipitation of proteins in EG.



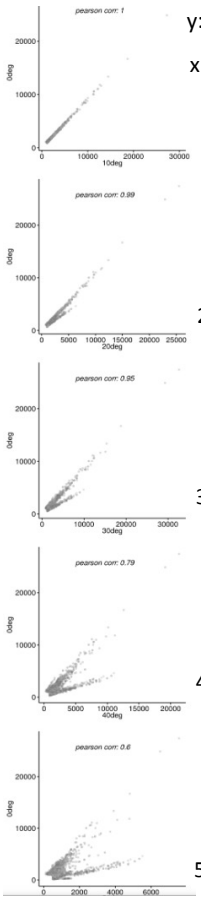
> Quality control

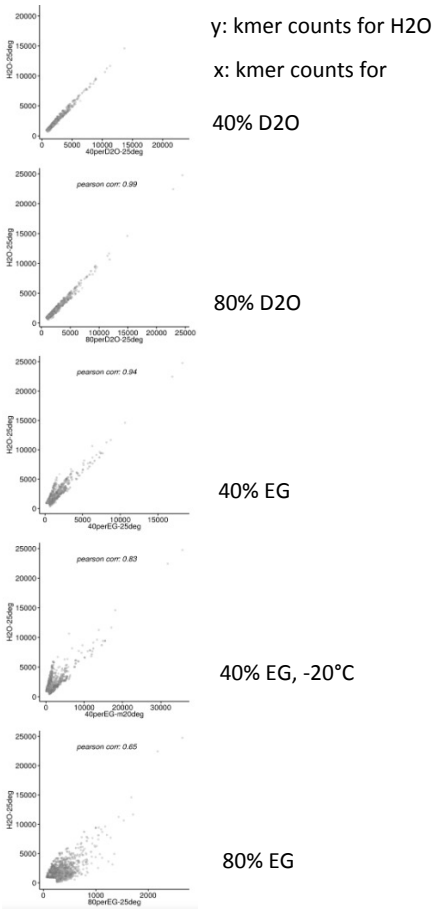
From the priliminary analyses, it seems our strategy of using 10 mM EDTA in washing indeed supressed the dissociation of TF-DNA complex, so that the selectivity difference during the binding step is visible for many TFs (ELK3 as example below):

_ For temperature change, all results are compared to 0 deg result (left), the selectivity difference gradually increases with the difference in temperature

_ For solvent change, all results are compared to H2O result (right), the difference between D2O and H2O results is negligible, and more visible with Ethylene glycol (EG)

Such behaviors are in accord with the expectation.

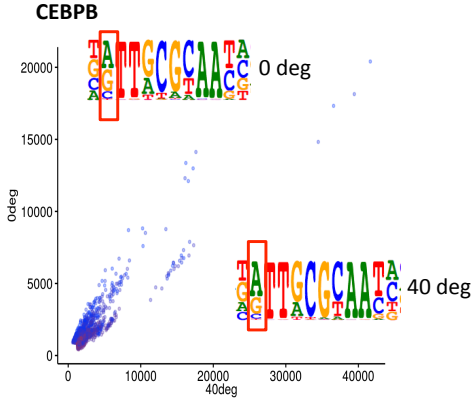


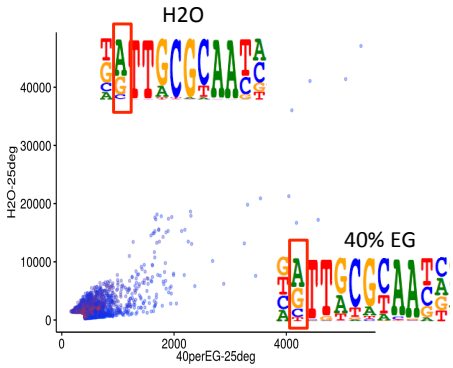


We noted that
_ most TFs have little signal under 50°C, likely due to the heat instability of protein.
_ As suggested by qPCR, most TFs failed to enrich signals with 80% EG
_ For many TFs, more or less specificity change has accompanied the temperature or solvent change. From such changeWe hope to discriminate between the enthalpy-driven mode and the entropy-driven mode. However, **most of the variance is accounted for by the stringency difference between conditions, and complicated by the dimer modes.**

> Stringency difference between conditions

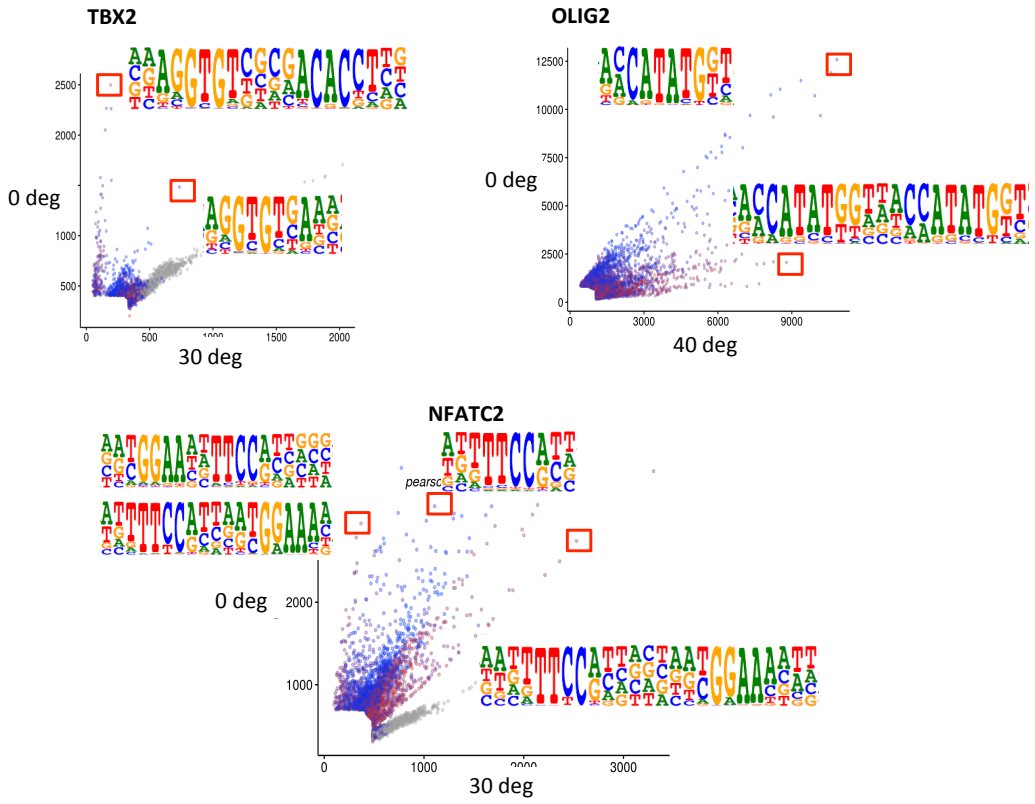
With CEBPB as an example, we show the general tendency for the selection stringency:
1. higher stringency is observed at a higher temperature (left)
2. higher stringency is with the solvent of H2O (right), compared to EG.





> Selectivity change involving dimer modes: are they meaningful?

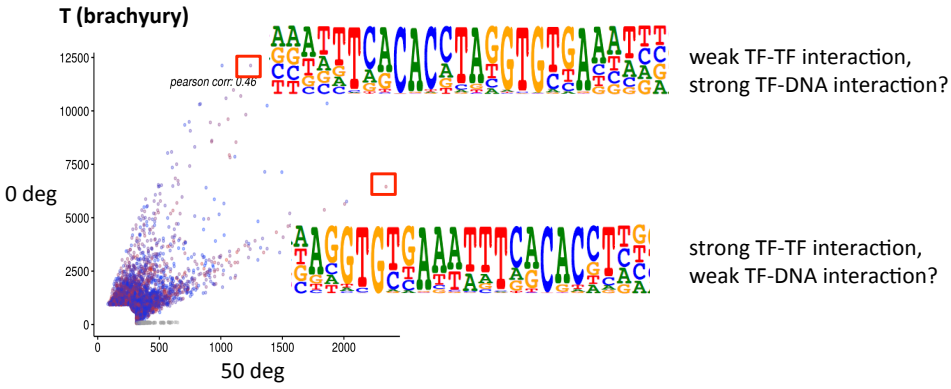
In addition, much of the selectivity change involves dimer modes. Dimer modes can be favored at low temperature (e.g. TBX2, also for MYBL2, VENTX, PKNOX1), at high temperature (e.g. OLIG2, also for HOXA2, HOXD12, HOXB13, EVX1, PDX1, CUX2, ELK3, MLX, ALX3, RHOXF1, CLOCK...), or different dimer modes favored at different temperatures (e.g. NFATC2)



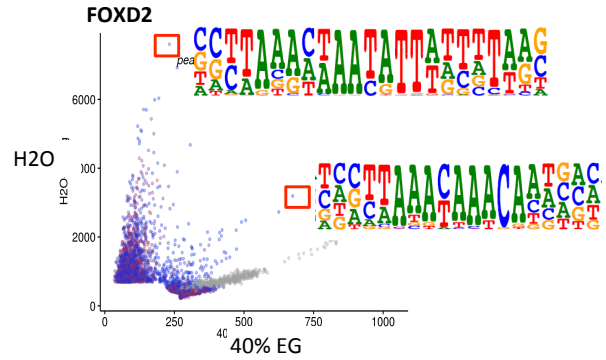
It is difficult to evaluate whether such change is due to stringency, or due to the different nature of the underlying TF-DNA interactions (that we are interested in). Because for dimers:
1. All different enrichment patterns of dimers can be explained only considering stringency, e.g., dimers enriched under low temperature may feature weak protein-protein interaction but strong protein-DNA interaction, whereas strong protein-protein interaction TF dimers will enrich more under high temperatures.
2. When comparing dimer modes with monomer modes, the observed selectivity differece can also arise from the fact that dimer modes usually has higher infomation content, thus enriches faster with conditions of a higher stringency.

To tell whether the changes involving dimers are meaningful, running SELEX with a gradient of TF concentration might be meaningful. If the binding specificity changes similarily between the dilution process and temperature increase, then such change is likely explained by stringency.

We also noticed specificity changes with dimer orientation (e.g. T(brachyury)), which can also be interpreted by stringency if without further information.



More complexed changes are associated with FOXD2. The change involves trimerization (?), there is probably too many ways to explain only with stringency.



> How to understend the effect of temperature and solvent change?

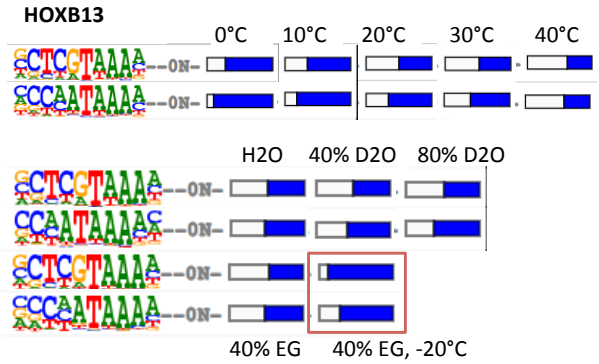
Another important question is how the conditions used affect the underlying interactions.

Intuitively, we expect both low temperature and EG to enhance all interactions except the hydrophobic ones, most of tested TFs do follow such rules. As mentioned above, TF motifs usually have lower stringencies under such conditions. Additionally, the overall direction of change on the kmer spectrum is also similar for both conditions.

While both EG and low temperature are conceived to increase the strength of ionic interaction and direct H-bonds, intuitively, the EG effect differs from low temperature in two aspects:
1. EG decreases more of the strength of hydrophobic interaction
2. Whereas low temperature increases solvent-mediated H-bond, EG decreases its strength (as the dihedral angle formed by -OH is not fixed for EG),

HOB13 was reported (E. Morgunova, et al., Elife 2018, 7) to have two major binding modes: the CAA (enthalpic) mode, which features well-defined water-mediated H-bond, and the TCG (entropic) mode that is with less fixed waters. We expect lower temperature favors the CAA (enthalpic) mode, in water, lowering the temperature does increase a bit the relative strength of CAA to TCG mode. However the effect is minor, probably because the CAA-mode has an hydrophobic interaction between Ile-262 and T, the strength of which will decrease when lowering the temperature. The effect of D2O and EG on the balance of the two modes are hard to explain, the CAA mode slightly increase its relative strength at higher conc of D2O, and remained the same for EG.

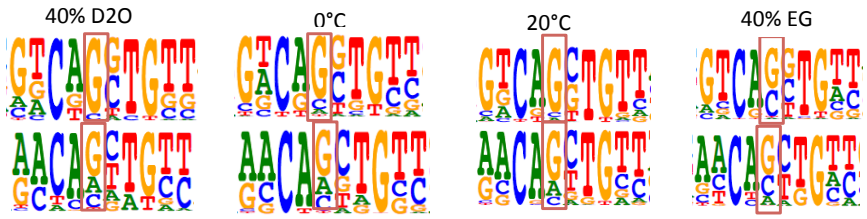
The most prominent change occurs at -20°C with EG, the relative strength of these two modes reversed. I haven't yet come out with a straightforward explanation. Experiments with more conditions might be required to rationalize the underlying mechanisms.



MYF6 was also reported to have 2 modes from Yimeng's list. Binding of the 1st mode involves water-mediated H-bond, while binding of the 2nd mode (left half) does not. The result does seem to suggest that the 1st mode (with H2O H-bond) decreased a bit in EG. The real ratio change could be larger because the signal of MYF6 is weak in this batch, likely dues to not enough protein concentration and dues to a high background of wash using cold EDTA buffer (cf. the qPCR profiles). In addition, the base "C" became more preferred in the center of the motif in EG (red boxed).

	0°C	10°C	20°C	30°C	40% D2O	80% D2O	40%EG	40%EG -20°C
1. AACANNTGTT	2612	2678	1981	2147	2979	2285	1730	1420
2. GTCAGSTGTT	1192	1303	945	1088	1298	999	973	775
Ratio (1:2)	2.191275	2.055257	2.096296	1.973346	2.295069	2.287287	1.778006	1.832258

MYF6

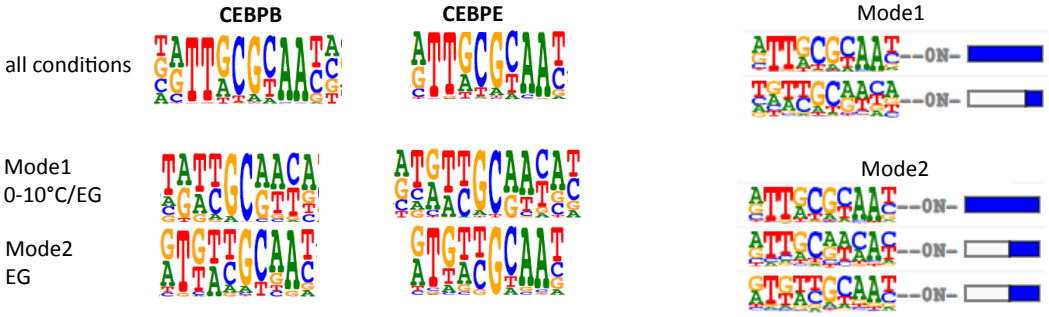


> New modes and motif changes induced by temperature/solvent change

However, we can still pick out changes that is less likely due to the stringency effect. When we observe
1. obvious changes within motif of monomer or strict dimer TFs (development of new specificity, or reversed order of kmer ranks)
2. ratio change between modes with a similar IC

More binding modes can be found at lower temperatures or with EG, because interactions are overall strenghened, and signals from weaker interactions but lower IC can be retained.

The following show the hand-picked examples:
Many changes with relatively large effect sizes are with the strict dimer TFs. For example, CEBPs developed two new modes under low temperature or in EG (left panel), while mode1 is found under low-temp/EG, mode2 is only found in EG. Comparing their strength to the normal mode (right panel) suggest that mode 1 is relatively weak, and mode 2 is approx. half the strength of the normal mode. However, SELEX is an exponential process, both of these modes might be stronger than seen here.



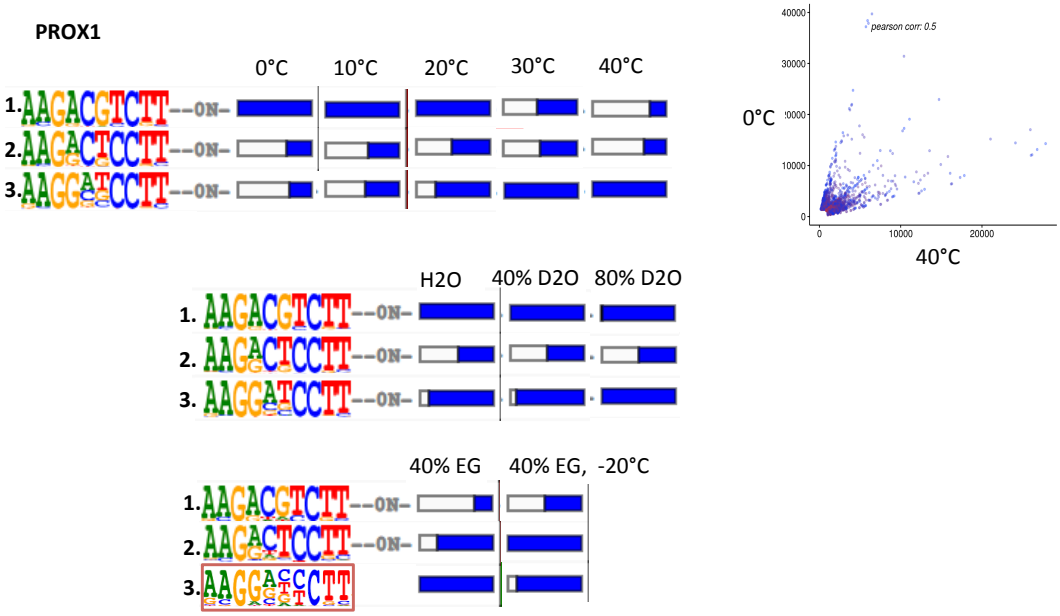
BHLHB8 has two modes for the central dinucleotide of its motif, the result seems to suggest that the "TA" mode is more preferred under high temperature (comparing 40°C to 0°C, but this can still be due to the stringency effect). However, in EG, the order is reversed, "GC" gets preferred over "TA"

BHLHB8

seed: AMCAKMTGKT

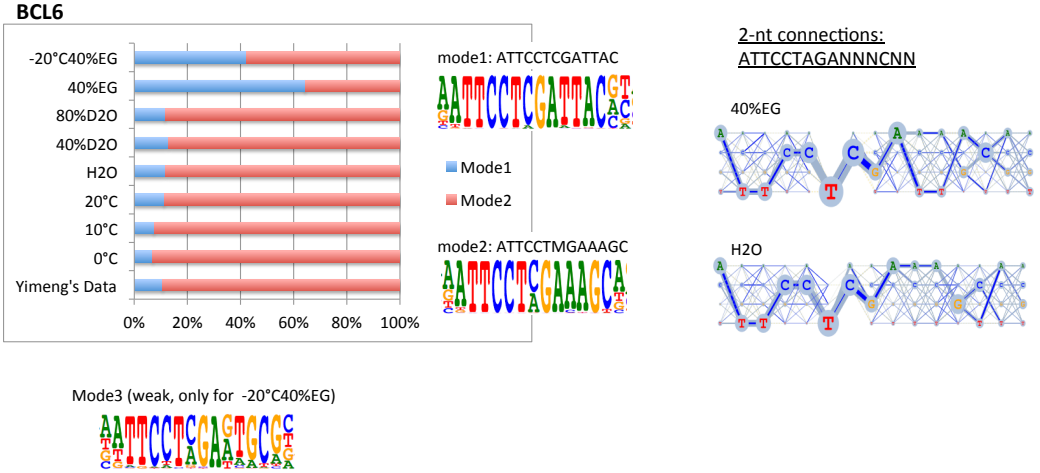
PROX1 is also a strict dimer and has 3 major modes, the balance between these modes has a pretty large effect size (as seen in the xyplot of 40°C vs 0°C). When temperature increases, the strength of mode 3 increases whereas that of mode 1 decreases. D2O do not have a significant effect. Replacing H2O with EG decreased mode 1, increased mode 2, and changed the motif of mode 3 (red box). It is hard to figure out a straightforward explanation for all the observed changes, maybe mode 1 involves the largest amount of H2O mediated H-bond so that it decreases with temperature, and also decreases when replacing H2O with EG.

PROX1



Balance of the two dimer modes also varied a lot for the C2H2 TF BCL6, between the solvents of H2O and EG (left). Such change is visible with the 2-nt connection matrix, where mode 1 is almost invisible in H2O but preferred over mode 2 in EG. In EG, another weak mode (mode 3) has also developed.

BCL6



There are also quite drastic changes with GLI2 and GLI3, but with a small effect size. The new mode developed in EG is only 10% the strength of the normal one.

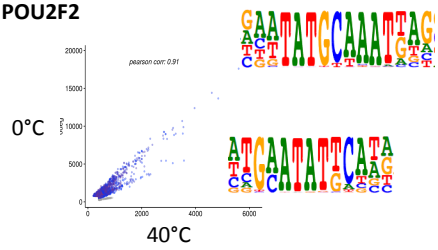
GLI2,GLI3


all conditions


weak
mode in 40% EG

When selectivity originates from stringency, the same direction of change is usually observed for low temperature and EG. However, for POU2F2, high temperature and EG prefers the same mode. However still cannot exclude the protein-level equlibrium of different TF oligomer species.

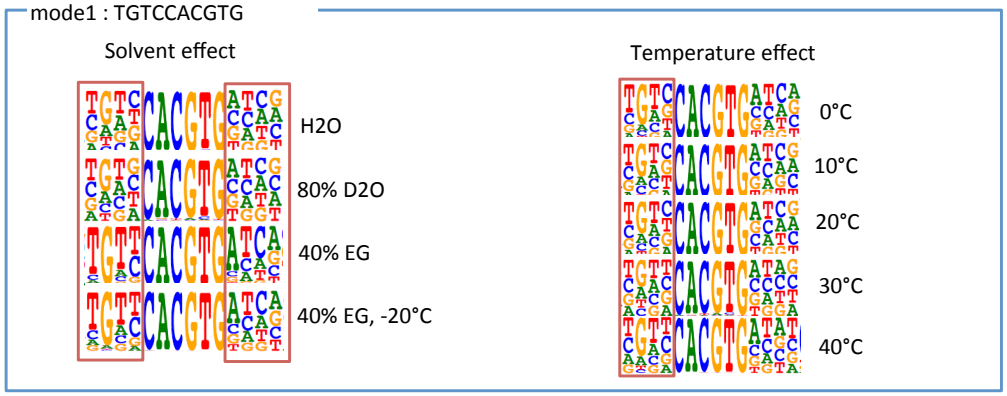
POU2F2



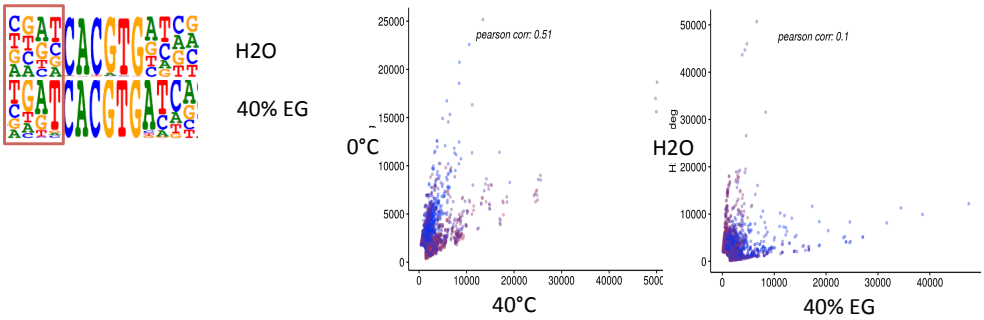
The stringency on the flankings of MLX motifs also changes conter-intuitively. Most TFs has higher stringency at high temperature and in H2O (compared to EG), such as the afore-mentioned CEBP, but flankings of MLX has higher stringency under lower temperature and in EG. Interestingly, with EG, high temperature do give a higher stringency (40%EG vs 40%EG-20°C). Similar tendency is also observed for another mode (mode 2). We confirmed that such conter-intuitive stringency change is limited only to the flankings, the overall behavior of MLX also accords to most TFs (cf. the xyplots, with more dimers in high temp, and more dimers in H2O compared to EG)

MLX

mode1 : TGTCCACGTG



mode2: ATCACGTGAT



> Perspective and future work

_ The effect size by changing solvent and temperature is pretty weak if excluding those caused by stringency change, and types of interactions for each bp of the motif can be complex, therefore it is necessary to gather more data with a variety of different conditions, to gain statistic power against the noise, figure out the rules guiding the specificity change.

_ Run SELEX with serial dilutions of TF proteins, to help separate stringency effects from specificity changes that characterize different types of interactions

_ Try more solvents/additives,

Polar aprotic: THF, DMF, MeCN, DMSO, nitromethane, Dichloromethane

Polar protic: formate, MeOH, EtOH, acetate, 1- and 2- propanol

Kosmotropic-Chaotropic series, e.g., trehalose, glycerol, TMAO, thiourea, urea, MgCl2, guanidinium thiocyanate

Polyamines that binds to DNA minor groove or compacts DNA, e.g. Spermidine, spermine

DNA minor groove binder, intercalaters, distorters

Side chain analogs of amino acids

_ From all possible interactions, intuitively the deconvolution of solvent-mediated H-bond, or hydrophobic effects from other interactions would be relatively easy, maybe we can put some focus there when choosing additional conditions to test, rather than having a pervasive goal.

_ This report showed some of the hand-picked examples of relatively large effect size (still quite small though), this is quite time-consuming and becomes tricky for weak effects. I assume some kmer-based approach should be developed to systematically analyze the whole dataset. E.g., still needs to think of the details but probably we can use kmer ranks as features to perform dimension reduction, find out kmer clusters that increase/decrease together, then further figuer out the underlying mechanisms to explain those kmers with the same behavior.

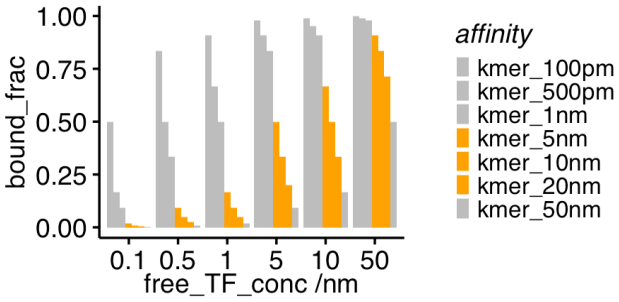
_ Binding signals not described by PWM model could be more sensitive to solvent/temp changes, try also look into MI signals and DNA shape signals.

_ Check ChIP-seq data between warm-blooded and cold-blooded animals to find in-vivo significance for this project

> Absolute TF binding affinity by fitting saturation effect

Sequencing has been utilized to characterize the relative affinities of a TF to different DNA sequences (G. D. Stormo, Z. Zuo, Y. K. Chang, Brief Funct Genomics 2015, 14, 30). To move this further, here we hope to use the sequencer to measure the absolute binding affinity of a TF. In our current SELEX protocol, quantifying the absolute amount of bound and unbound is difficult, it is difficult even for collecting the unbound fraction. However, the sequencing approach accurately measures the relative abundance of different DNA sequences within the bound library. Leveraging such infomation, we can figure out the absolute affinity of a TF, by titrating a fixed amount of DNA with a series of TF concentrations.

For example, simulation (the following plot) shows that the ratio between kmers with different affinities will vary with TF concentration, especially when TF concentration approach / surpass their Kds (note the relative ratio change for colored kmers from TF conc 1nM to 50nM). Such saturation effect contains infomation regarding TF's affinity to each kmer.



To test the possibility of fitting, I generated the following hypothetical data set (assuming a homogeneous background kmer counts). The relative enrichment of each kmer to a ref kmer (10nM kmer as ref here) is calculated based on kmer's affinity and the TF concentration. In experiment, such relative enrichment ratios can be obtained by sequencing the cyc1 library followed by counting the kmers.

abs affinity	20	10	5	2.5	abs TF conc /nM
kmer_100pm	1.493	1.98	2.941	4.808	kmer / kmer_10nm
kmer_500pm	1.463	1.905	2.727	4.167	
kmer_1nm	1.429	1.818	2.5	3.571	
kmer_5nm	1.2	1.333	1.5	1.667	
kmer_10nm	1	1	1	1	
kmer_20nm	0.75	0.667	0.6	0.556	
kmer_50nm	0.429	0.333	0.273	0.238	

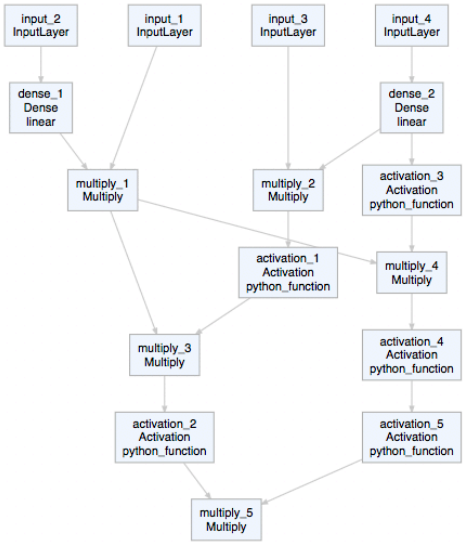
However, when performing such measurement, we only know the relative affinities of the kmers but not their absolute affinities. Therefore, the available data we really have after the measurement is inside the blue border of the table

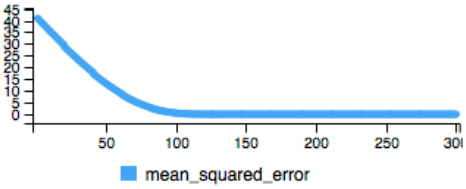
	rel affinity	20	10	5	2.5	abs TF conc /nM
kmer_100pm	0.01	1.493	1.98	2.941	4.808	kmer / kmer_10nm
kmer_500pm	0.05	1.463	1.905	2.727	4.167	
kmer_1nm	0.1	1.429	1.818	2.5	3.571	
kmer_5nm	0.5	1.2	1.333	1.5	1.667	
kmer_10nm	1	1	1	1	1	
kmer_20nm	2	0.75	0.667	0.6	0.556	
kmer_50nm	5	0.429	0.333	0.273	0.238	

I tried if using the infomation can fit for the abs affinity of the ref kmer (10nm)

$$\frac{kmer}{kmer_{ref}} = \frac{1-\frac{1}{(rel\_affinity)K_d^{ref}+[TF]}}{1-\frac{1}{K_d^{ref}+[TF]}}$$

It is possible to trick Keras, the software for NN, to do some custom optimization work. The following illustration (left) shows structure of the NN to optimize according to the formular above. The Keras model minimizes the mean squared error between the predicted ratios and the real ratios between kmers, the fitting worked as suggested by the decrease of loss function (right). The true Kd of the reference kmer here is 10nM. I provided an initial value 100nM, after fitting, it converges to 9.999931nM.





Because in the formular, it is the ratio between K_d^{ref} and [TF] that dictates the kmer ratios, therefore it is impossible to fit for Kd without knowing the absolute value of [TF]. However, as shown above we can measure the absolute value of [TF] in high-throughput. Although no guarentee that 100% of the recombinant proteins are active, we should still be able to get Kds for all TFs in the correct order.

Alternatively, we can also label the TFs with a known-affinity tag, then attach the catcher of the tag to a reference DNA ligands with defined sequences (these ligands will show up in sequencing). Then after sequencing by comparing the enrichemnt of target sequnce to the reference DNA ligand, we will also be able to deduce the absolute affinity.

For conceivable application with the measured Kds, we would be able to infer the TF concentration in cells by comparing the ChIP-seq peak area of the TF at sites with different affinities. In addition, previously I have been struggling for a long time to find a non-linear optimizer for the dissociation data, using Keras might be a choice that worth another try.