Understanding Transcription Factor Binding Kinetics Using Statistical Learning Methods

UNIVERSITY OF CALIFORNIA
MERCED

Student: Ruoxi Zhao

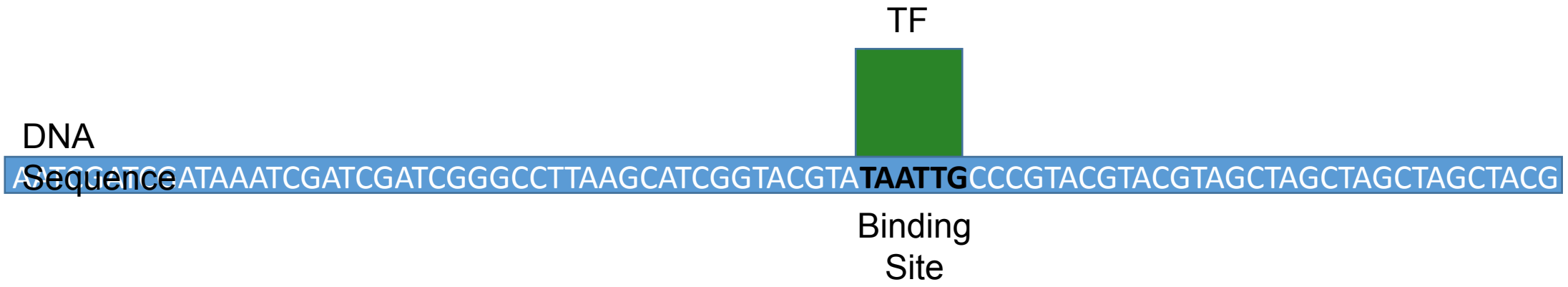
Faculty Mentor: Professor H. Tomas Rube

Department of Applied Math, School of Natural Sciences

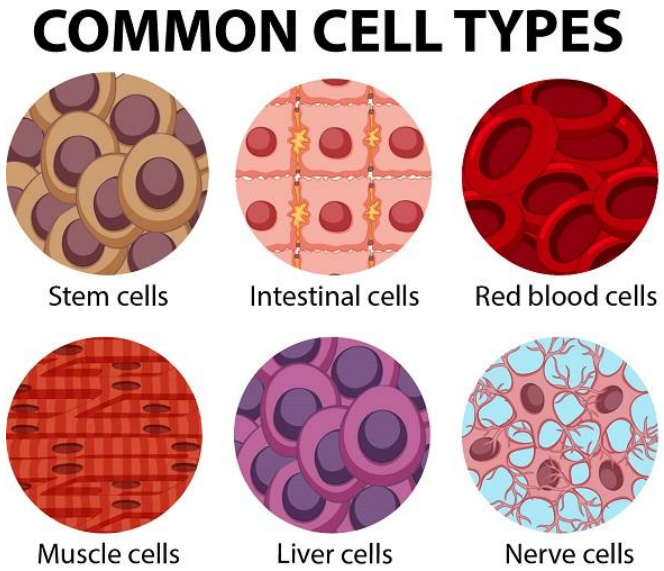
University of California, Merced

Transcription Factors

- Transcription factors (TFs):
  - Proteins that bind to **specific DNA sequence**
  - Regulate expression of nearby genes



Gene Regulation



Determines the shape and function of each cell

Adapt to changes in the environment

Image credit: shutterstock.com, wikipedia.org

How do TFs read sequence?

Goal: Learn mathematical model that predicts binding.

Binding Affinity and Equilibrium Binding

- The **dissociation constant** (k_D) quantifies the strength of biomolecular interaction

k_{on}:

k_{off}:

$$k_D = \frac{k_{off}}{k_{on}} = \frac{Dissociation\ rate\ constant}{Association\ rate\ constant}$$

Small k_D, strong binding affinity
Large k_D, weak binding affinity

- To predict equilibrium binding, we need to predict how k_D depends on sequence.

The SELEX-seq Experiment

DNA with binding site

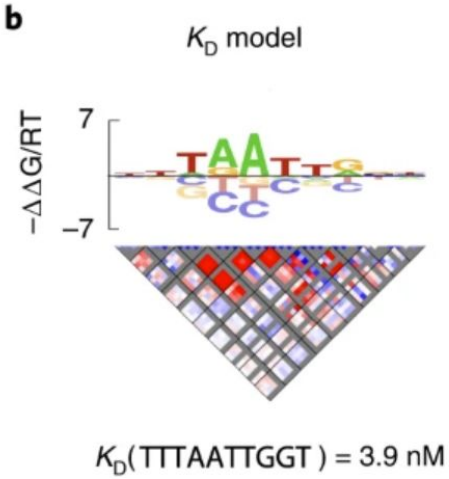
Randomized DNA Sequences

Adding transcription factors

Equilibrium

Extract the binding sequences

Sequence the DNA data



t_0

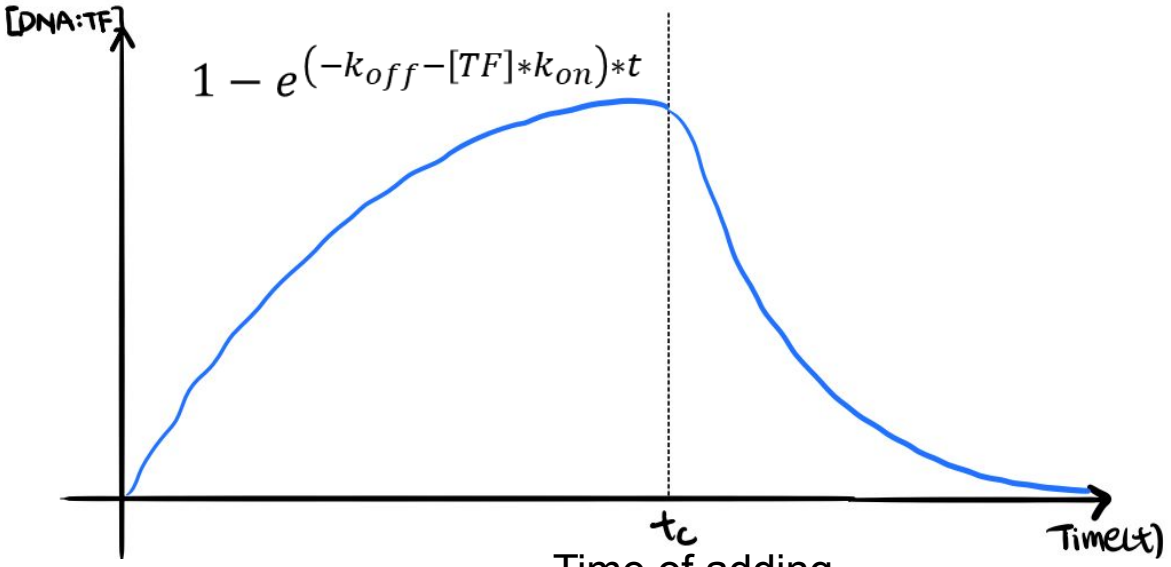
As time progresses

t_f

Model of Initial DNA Binding

Probability of bound at equilibrium:

$$P(bound) = \frac{[DNA{:}TF]}{[DNA]_{total}} = \frac{[TF]}{[TF]+k_D}$$



Time of adding competitors

Differential equations describing single sequence binding kinetics:

$$\frac{d}{dt}[DNA{:}TF](t) = [DNA](t) * [TF] * k_{on} - [DNA{:}TF](t) * k_{off}$$

After solving this, we got:

$$\frac{[DNA{:}TF](t)}{[DNA]_{total}} = \left(1 - e^{(-k_{off}-[TF]*k_{on})*t}\right) * \left(\frac{[TF]}{[TF]+k_D}\right)$$

Current Binding Models Blind to Kinetics

- The current binding model predicts k_D
- The k_D ***does not tell you***:
  - How long are TF molecules bound?
  - How quickly is the binding equilibrium reached?
- To answer such questions, we need:

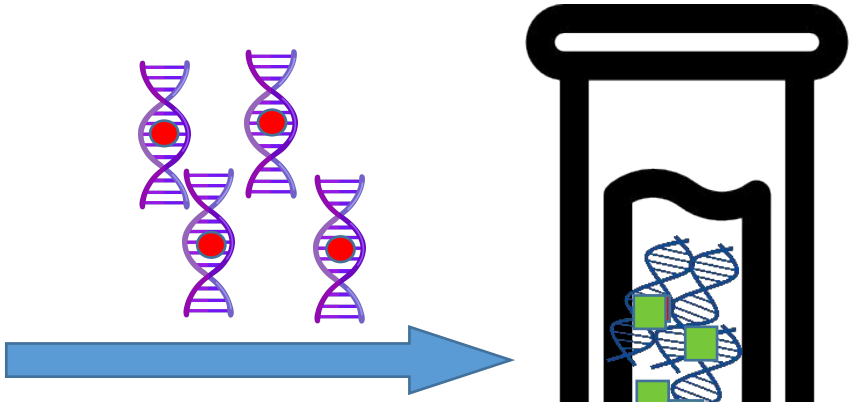
$$k_D = \frac{k_{off}}{k_{on}}$$

k_{on}:

k_{off}:

- Current binding models predict binding strength, blind to kinetics
- Goal: Learn how binding kinetics depend on sequence

Competitor DNA: Suppresses Rebinding

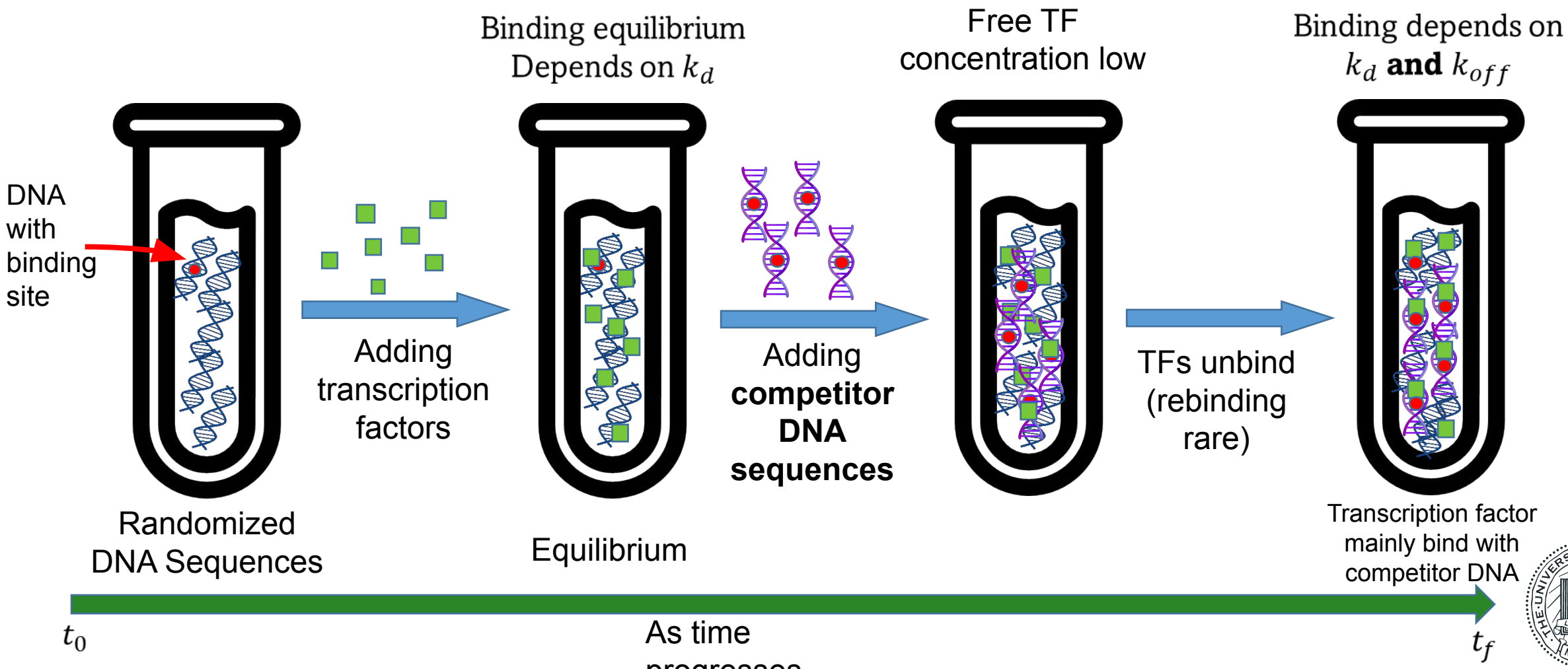


Adding Competitor DNA sequence

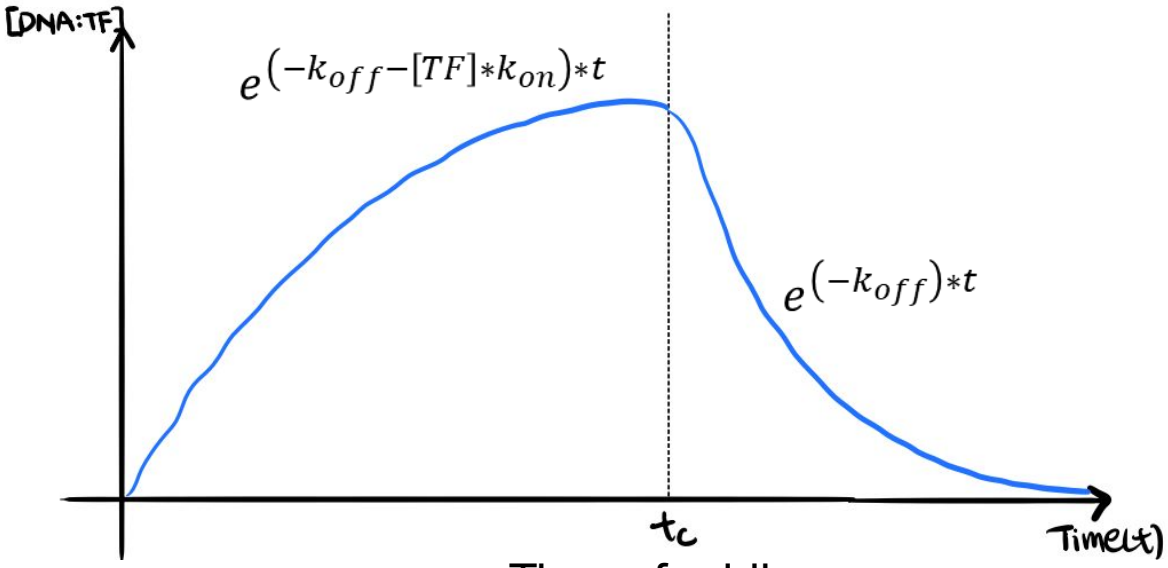
After reaching the equilibrium

- Add competitor with high-affinity binding site
- Free TFs absorbed
- Rebinding to randomized DNA suppressed

Competition SELEX Experiment



Modeling Binding Kinetics with Competitor



Time of adding competitors

Differential equations describing binding kinetics:

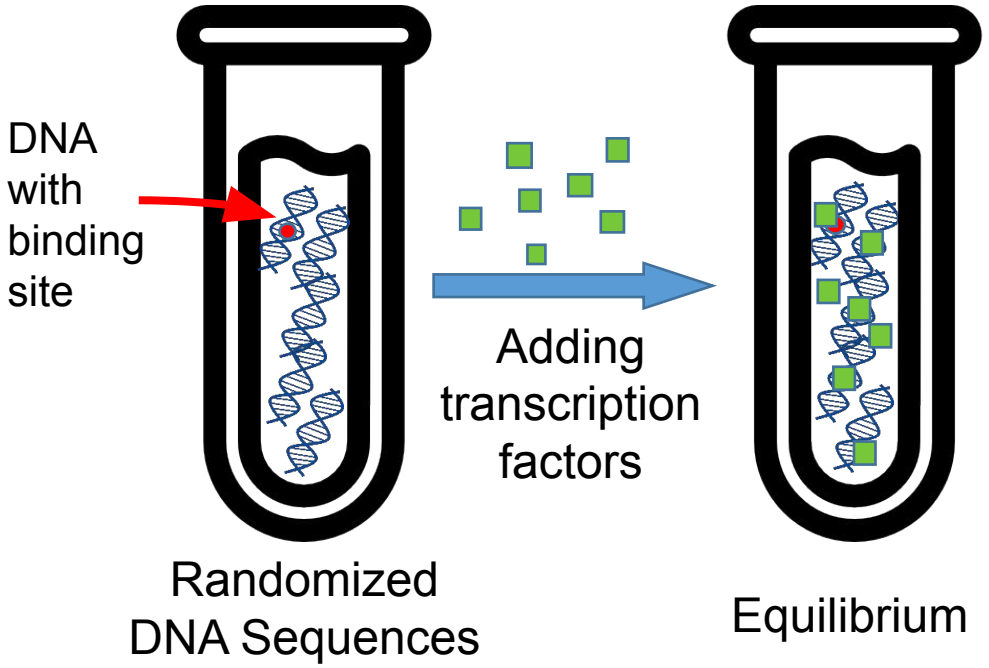
Equals 0 because [TF] = 0

$$\frac{d}{dt}[DNA:TF](t) = \cancel{[DNA](t) * [TF] * k_{on}} - [DNA:TF](t) * k_{off}$$

Solution :

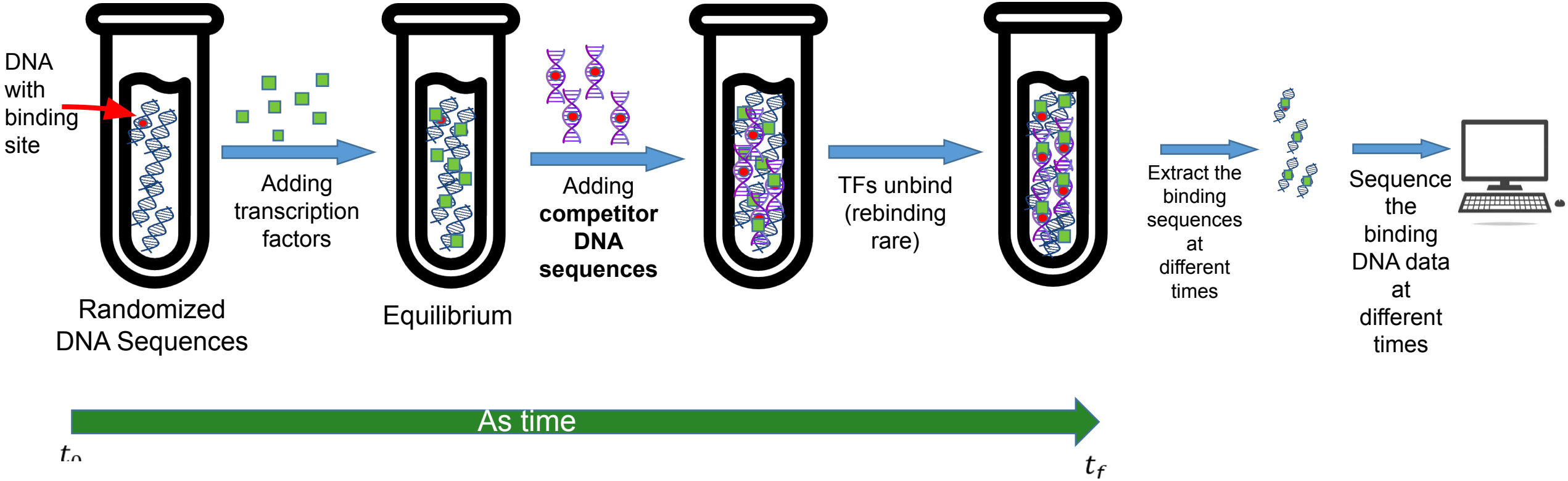
$$[DNA:TF](t) = \left(e^{-k_{off}*t}\right) * [DNA:TF](0)$$

The SELEX-seq Experiment



t_0 → As time progresses → t_f

Proof-of-Concept Experiment



TF:	Dll (Distal-less) TF from fruit fly
Competition time:	0, 1, 5, 30 min
Sequenced:	Input DNA and bound sequences at each timepoint.
My project:	Analyzed the data to see if experiment worked.

First step: Parse Data

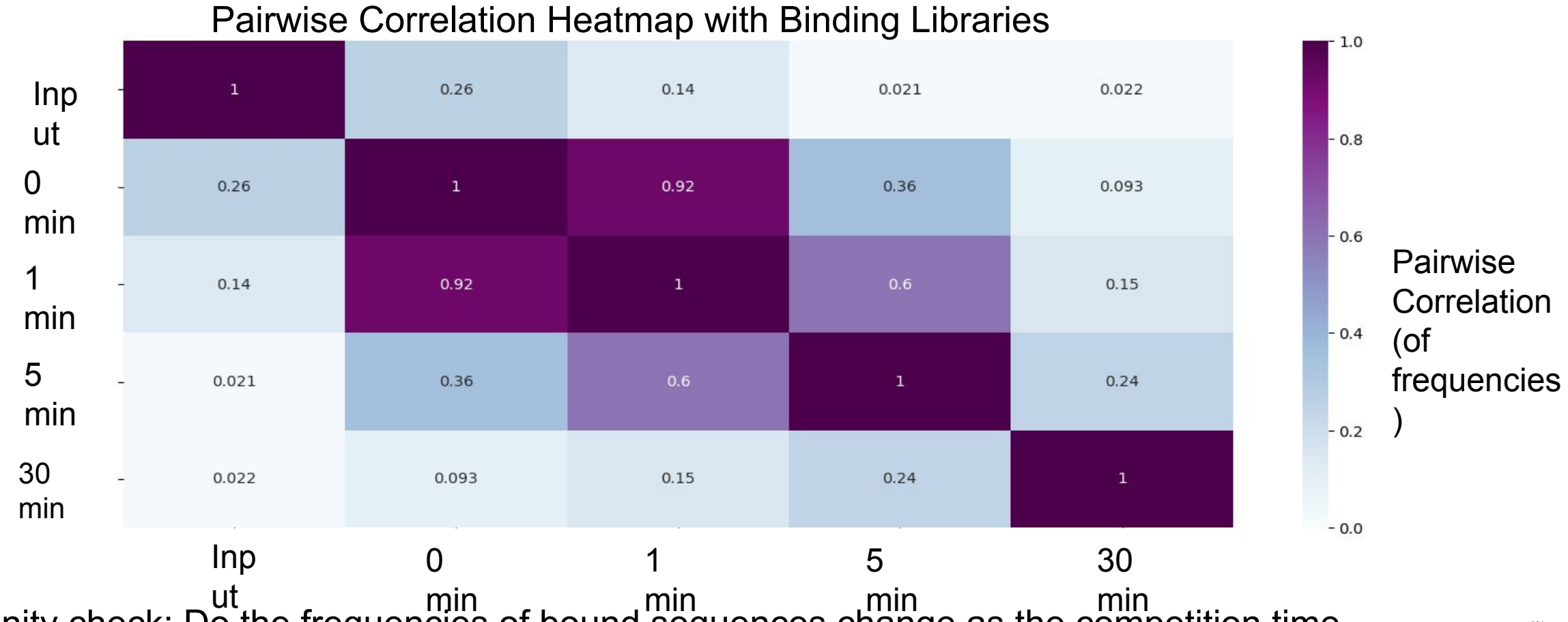
Enrichment in 0 minute

Randomized DNA sequence

	index	freq0B	freq1B	freq5B	freq30B	freqR0
0	ATCTAATTAA	1.618241e-04	3.253455e-04	2.551404e-04	0.0	5.430562e-06
1	TAATTGCTGT	1.315473e-05	2.409324e-05	4.849469e-05	0.0	3.443771e-07
	CGATATTACT	2.088053e-06	3.466654e-07	3.145292e-07	0.0	2.092753e-06
3	TCCCCATCAC	6.264158e-07	1.733327e-07	4.289035e-07	0.0	3.284828e-06
4	GCTATGACTA	0.000000e+00	0.000000e+00	5.775900e-06	0.0	1.271546e-06
...	...	...	...	...	...	...
1017172	CGACGCCAGG	0.000000e+00	0.000000e+00	0.000000e+00	0.0	7.947164e-08
1017173	CAAGAGAAGG	0.000000e+00	1.733327e-07	0.000000e+00	0.0	5.298110e-08
1017174	CGCAAGACGT	0.000000e+00	1.733327e-07	0.000000e+00	0.0	1.589433e-07
1017175	GGCTGCCTGG	0.000000e+00	0.000000e+00	0.000000e+00	0.0	5.298110e-08
1017176	GAGACTGGCG	0.000000e+00	0.000000e+00	0.000000e+00	0.0	1.854338e-07

1017177 rows × 6 columns

Bound sequence libraries changed after competition

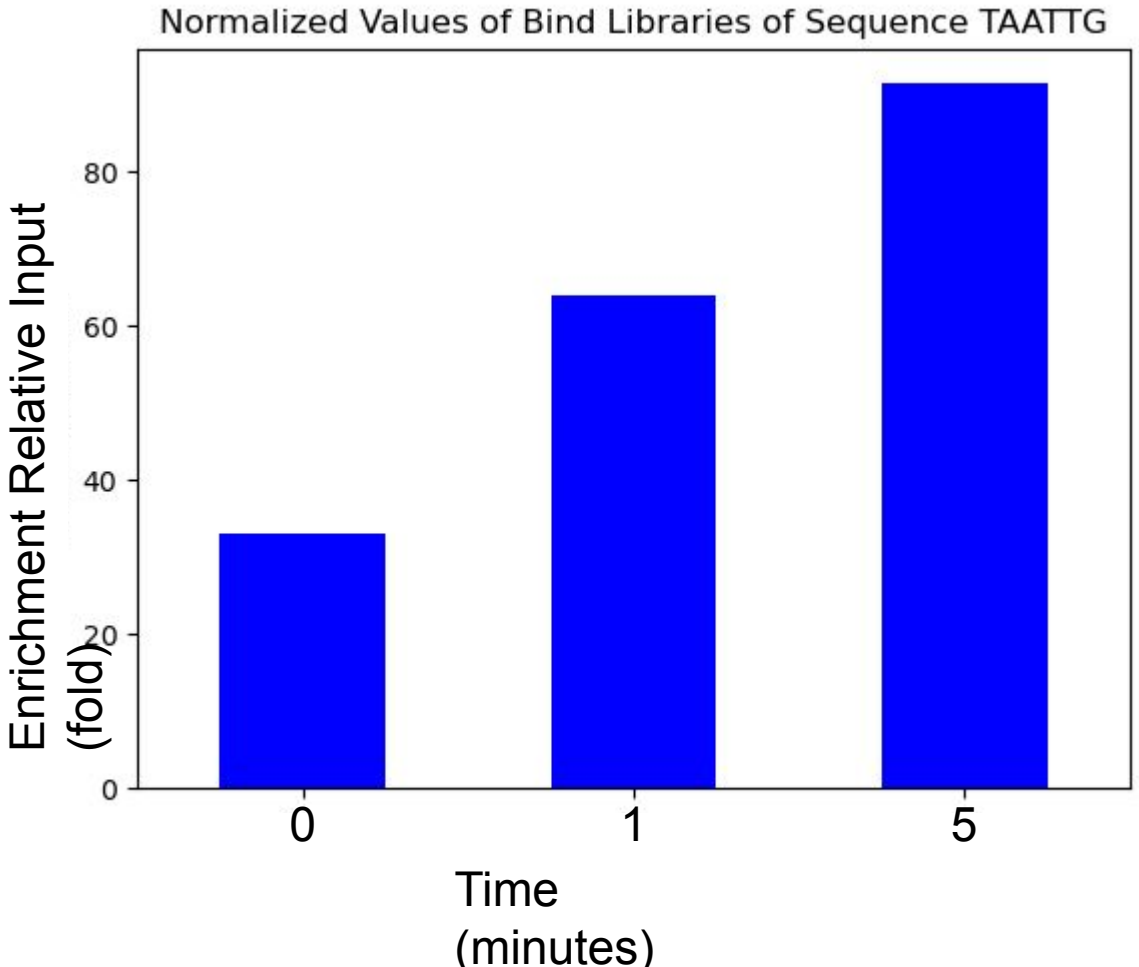


- Sanity check: Do the frequencies of bound sequences change as the competition time increases?
- Take-away:
  - As expected, libraries become more dissimilar as time goes on

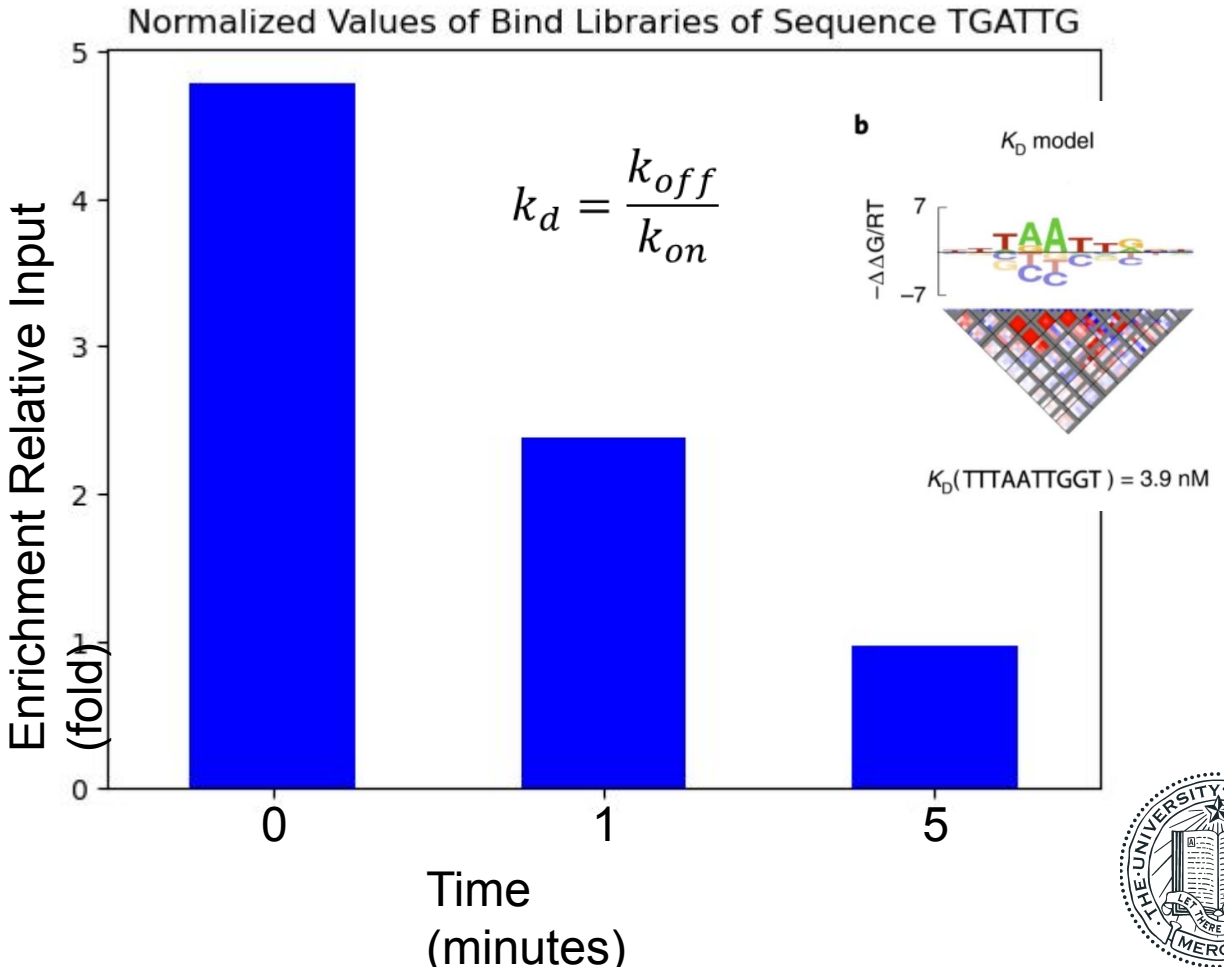
Low-affinity sequences unbind faster

CCBM
NSF-CREST
CENTER FOR CELLULAR AND BIOMOLECULAR MACHINES

High Binding Affinity sequence TAATTG



Low Binding Affinity sequence TGATTG



Conclusion

- Combining SELEX-seq with competition can probe binding kinetics with high throughput.
- Proof-of-concept experiment works as expected:
  - Bound sequence libraries changed after competition
  - Low-affinity sequences unbind faster
- Future Step: Build machine-learning method for inferring kinetic binding models

Acknowledgement/References

Center for Cellular and Biomolecular Machines at UC Merced (NSF-HRD-1547848 and NSF-HRD-2112675)

CCBM-CREST, UC Merced

Tomas Rube

Roberto C. Andresen Eguiluz

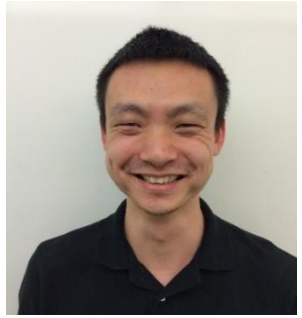
Petia Gueorguieva

Jose Zamora Alvarado,
All the mentors and fellow students

Experiment Dr. Rube performed while visiting Mann Lab, Columbia University

Richard Mann

Siqian Feng



William Glassford

https://www.degruyter.com/document/doi/10.1515/medgen-2021-2073/html
https://medlineplus.gov/genetics/understanding/howgeneswork/geneonoff/#:~:text=Gene%20regulation%20is%20an%20important,to%20changes%20in%20their%20environments.
https://www.nature.com/articles/s41587-022-01307-0
https://www.nature.com/scitable/definition/transcription-factor-167/
https://app.clickup.com/36071687/v/b/s/66168757

Binding Equilibrium

Binding Equilibrium:
As time progresses, the number of transcription factors binding with the DNA sequence will reach an equilibrium, with an equal rate of binding and unbinding.

For a **single sequence** at equilibrium:

$$[DNA] * [TF] * k_{on} = [DNA{:}TF] * k_{off}$$

In a long run, the bind on and bind off rates are the same on average.

Binding Affinity

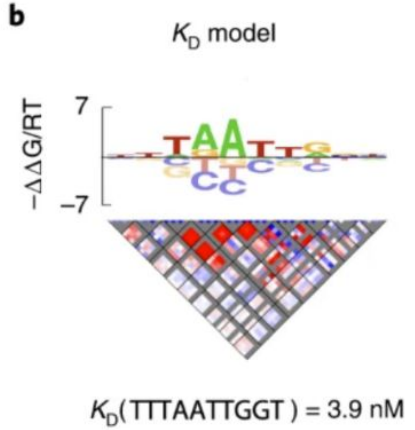
Binding Affinity:
The strength of the binding interaction between a single biomolecule (protein, DNA) to its ligand/binding partner.

- The **dissociation constant** (k_d) measures the equilibrium, which it quantifies the strength of biomolecular interaction

k_{on}:

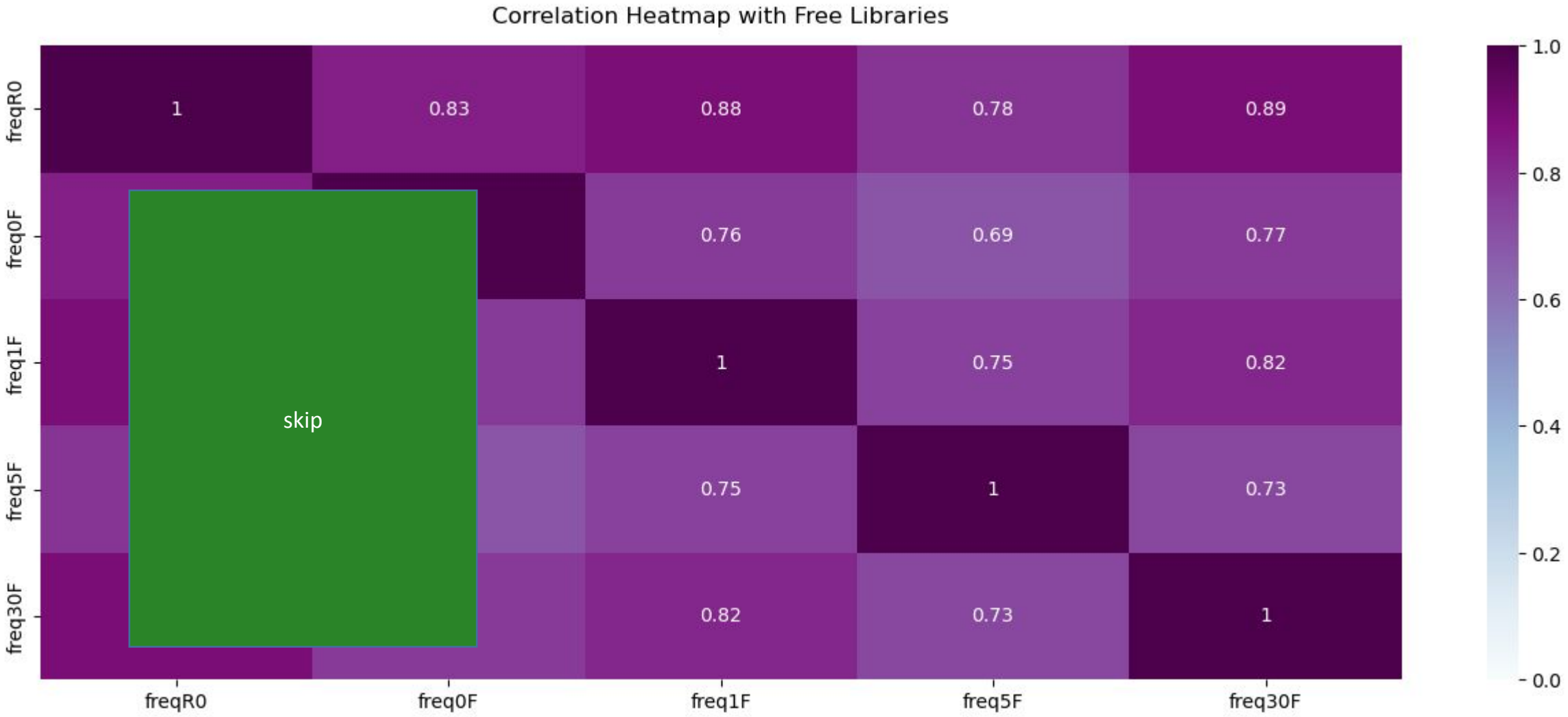
k_{off}:

$$k_d = \frac{k_{off}}{k_{on}}$$

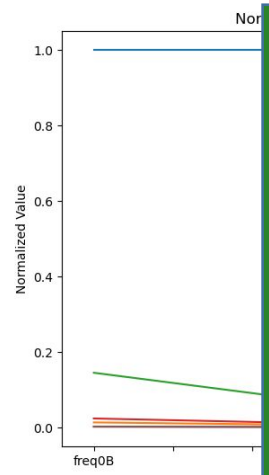


The binding kinetics can be used to predict the binding affinity of the transcription factor for different DNA sequences, as well as the kinetics of binding in the presence of competing DNA sequences.

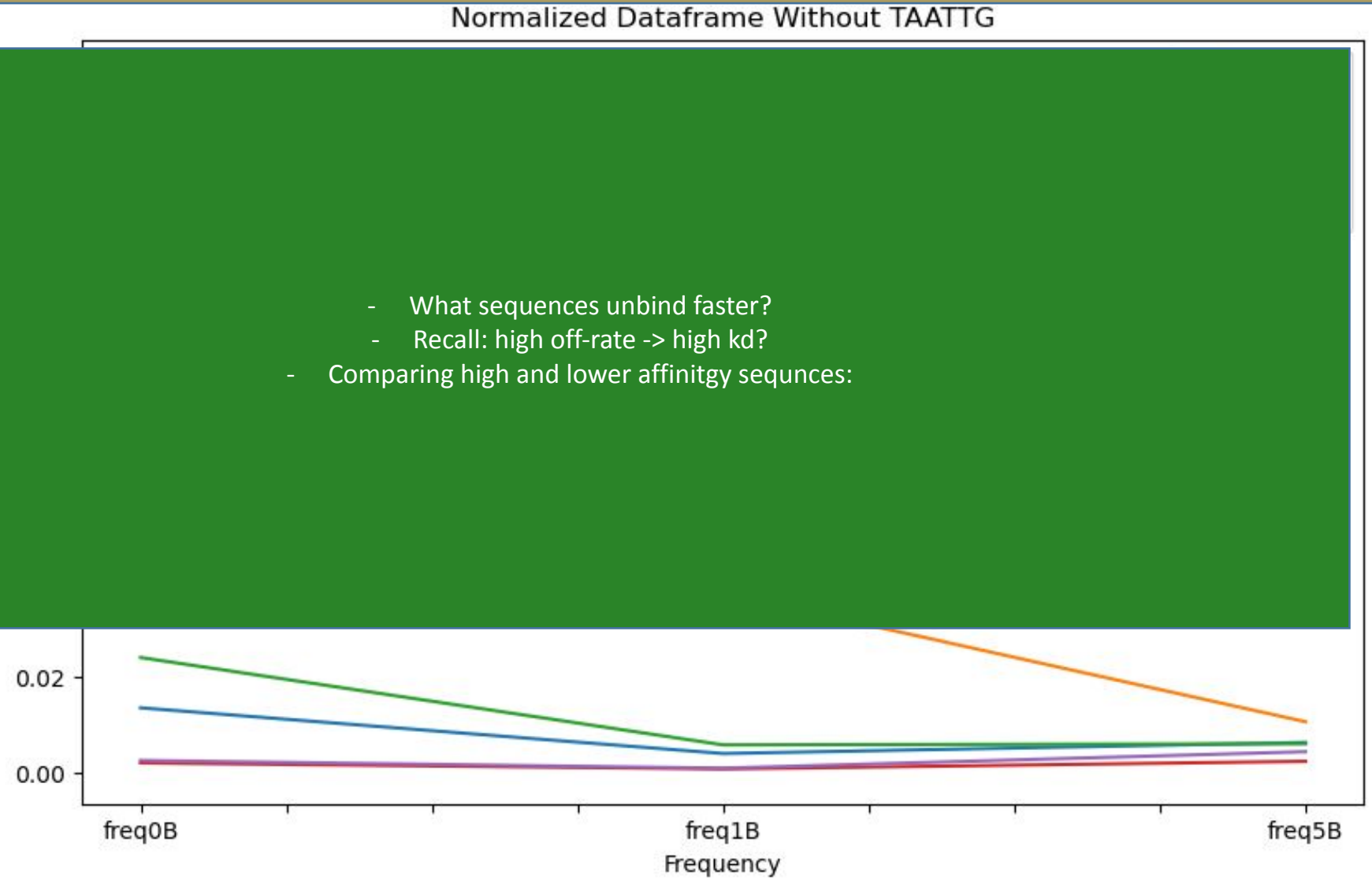
Results


Correlation Heatmap with Free Libraries
freqR0
freq0F
freq1F
freq5F
freq30F
1
0.83
0.88
0.78
0.89
0.76
0.69
0.77
1
0.75
0.82
0.75
1
0.73
0.82
0.73
1
skip
1.0
0.8
0.6
0.4
0.2
0.0

Results

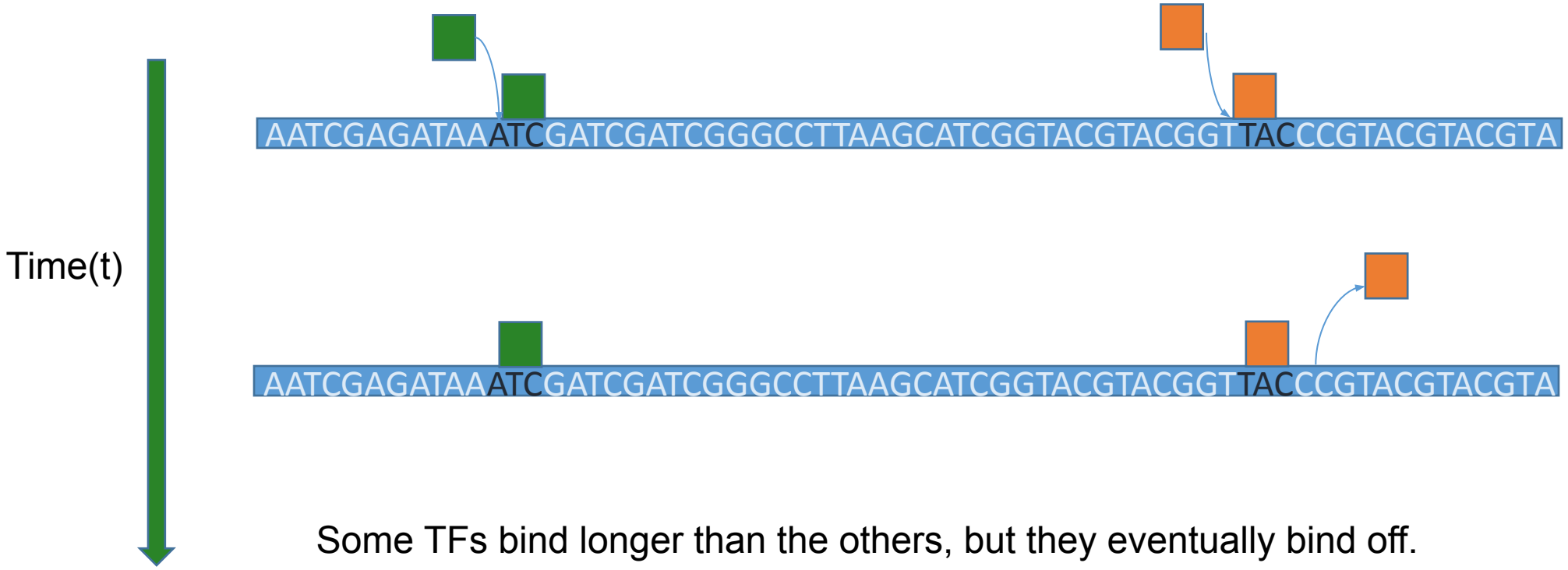


- What sequences unbind faster?
- Recall: high off-rate -> high kd?
- Comparing high and lower affinitgy sequnces:



Binding Kinetics

- Describes the dynamics binding interaction between two molecules, expressed as k_{on} (rate of association) and k_{off} (rate of disassociation).
- Studying the transcription factor binding kinetics can help us better understand the dynamics of **gene regulation**.



Some TFs bind longer than the others, but they eventually bind off.

Fruit Fly Distal-less(Dll) TFs

FlyTF is a database of computationally predicted and/or experimentally verified site-specific transcription factors (TFs) in the fruit fly Drosophila melanogaster. It covers the DNA-binding characteristics of the proteins and a more fine-grained annotation of both DNA binding and regulatory properties.

Competition SELEX Experiment