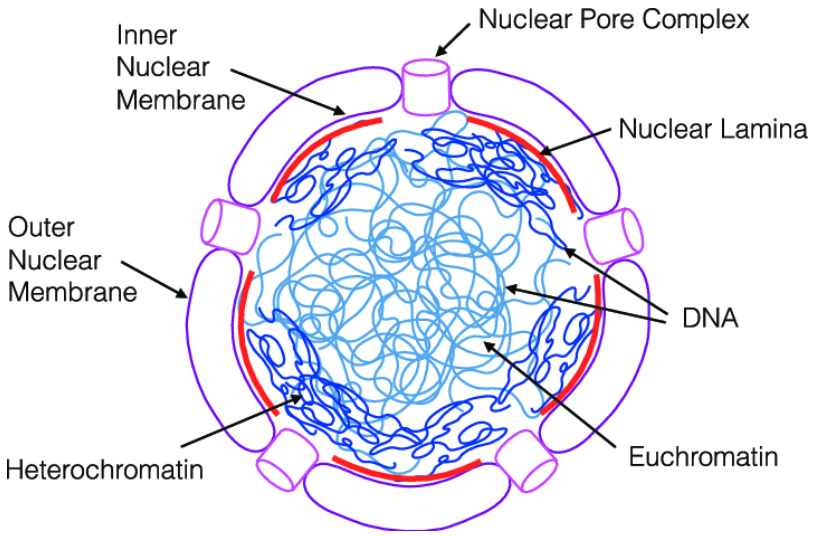
The Role of Protein Occupancy in DNA Compartmentalization

Prof. Leonid Mirny Lab:
Vishnu Emani, Kevin Zhao

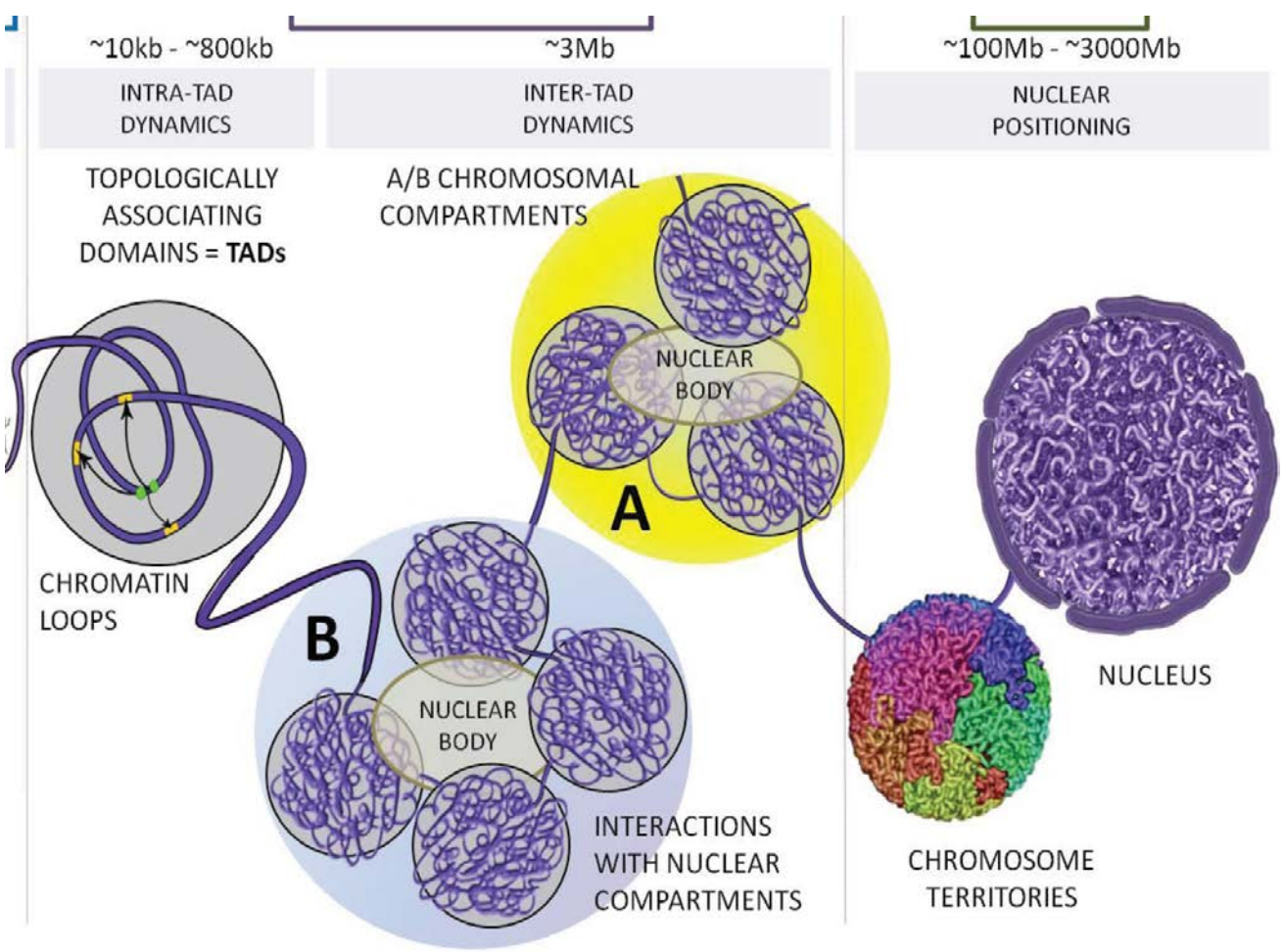
Background/ Overview

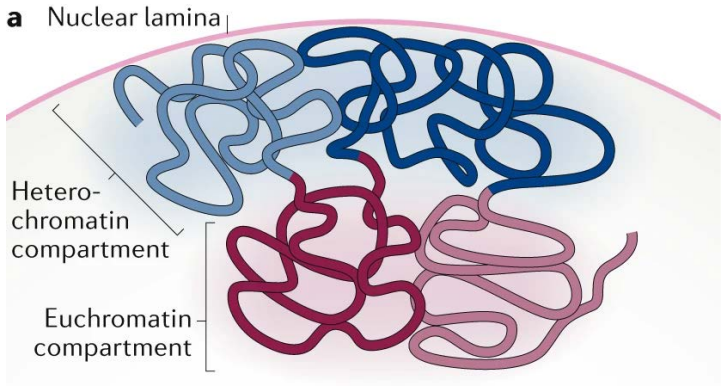
- Most basic: Every cell has a nucleus, with DNA
- Q: “How is the DNA organized in the nucleus of a cell and what structures does it take?”



Compartments

- Many layers of organization
- Compartments = large scale
  - About once every million base pairs

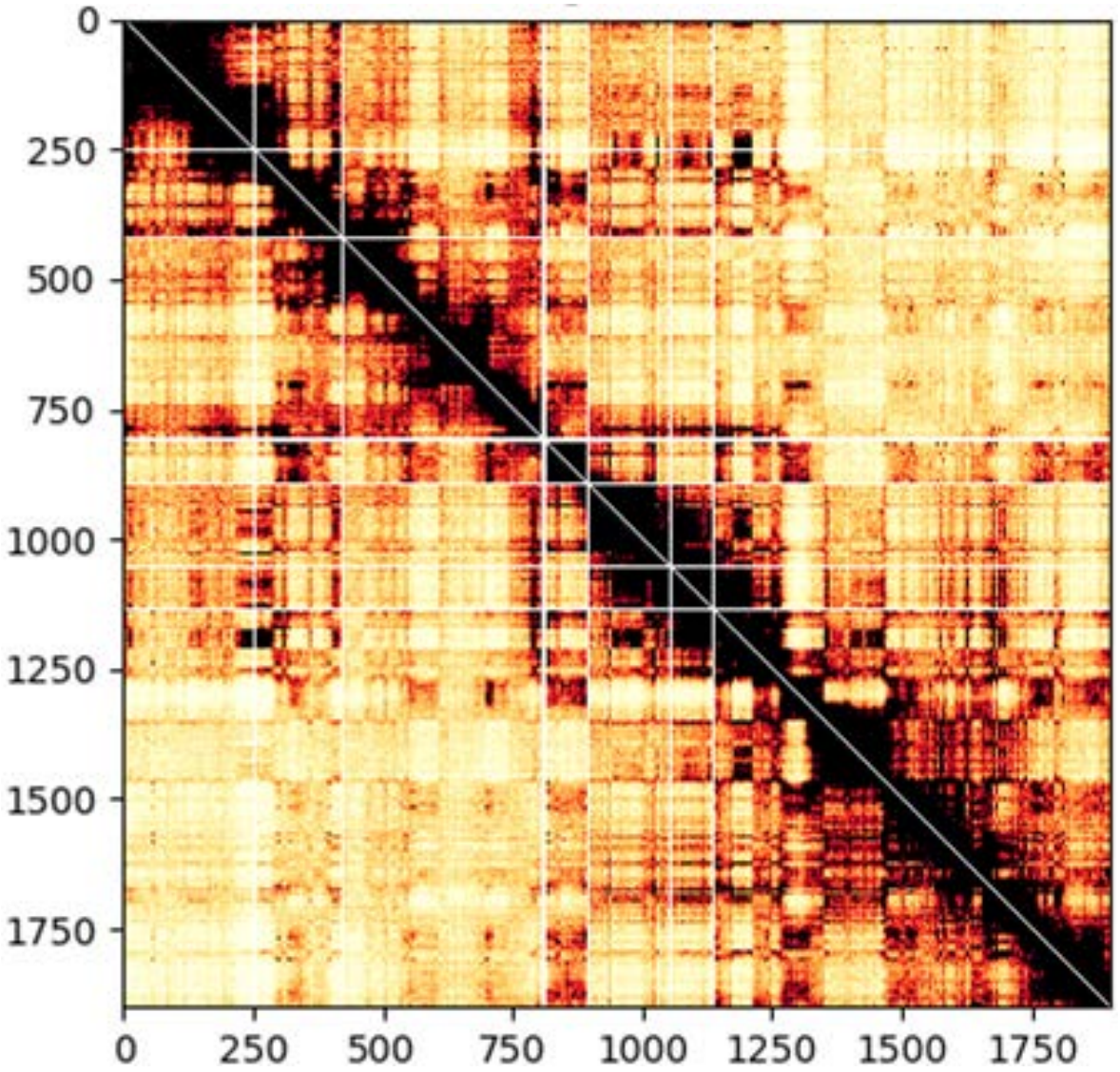




Experimental Methods

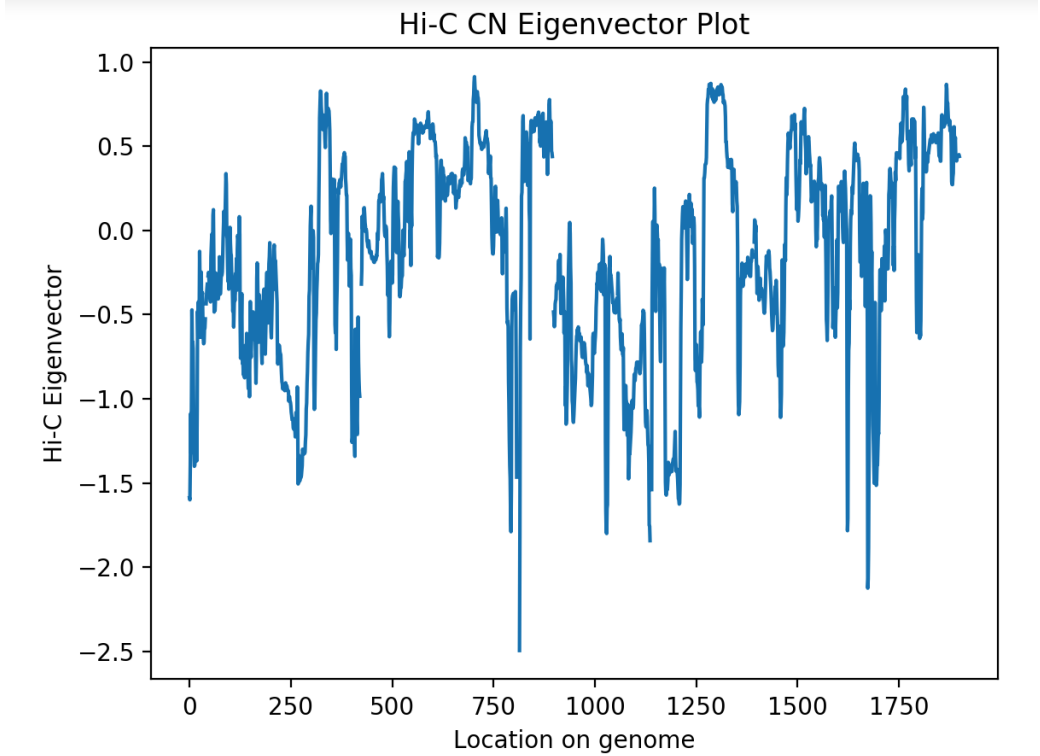
Hi-C Chromatin Capturing

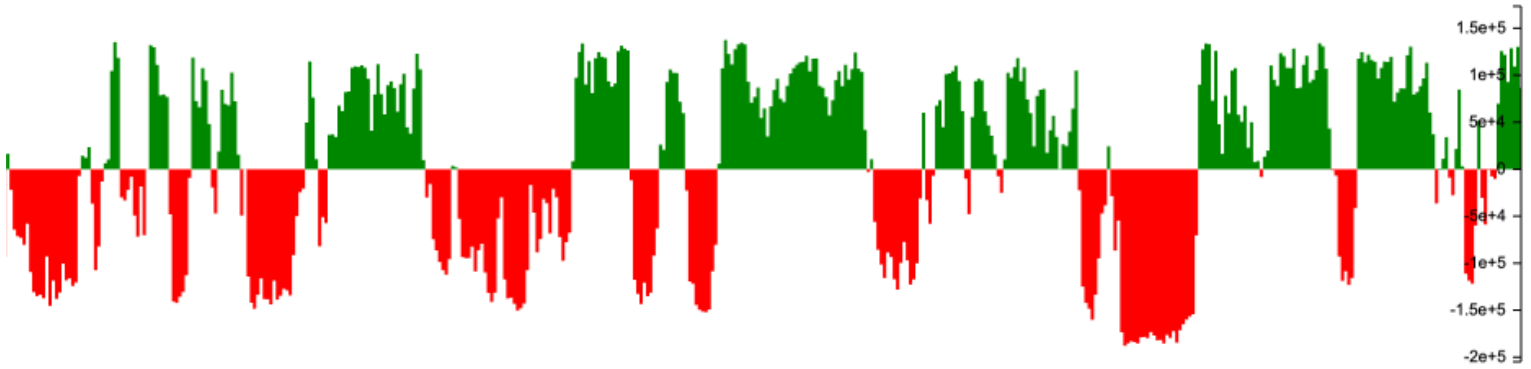
- Many methods to analyze how DNA is positioned
  - Convenient method is to examine contact within genome



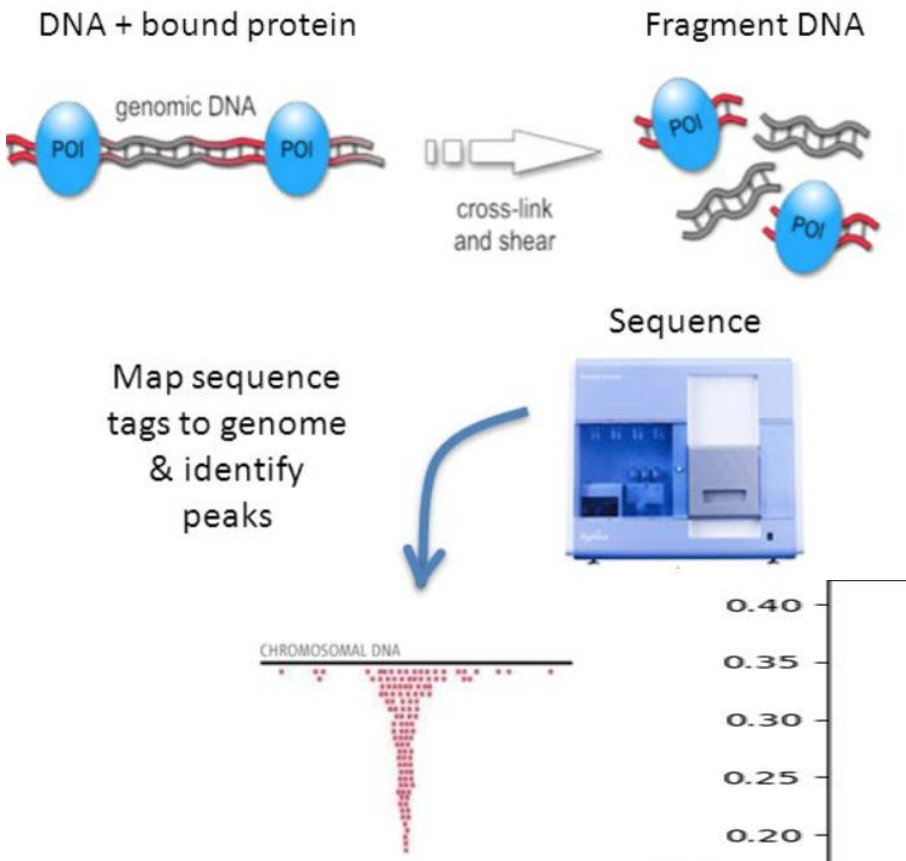
Analyzing Hi-C

- Hi-C too much data
- Eigenvectors:
  - Decomposition of the matrix into vectors that summarize the behavior of the matrix
- Eigenvectors indicate compartments

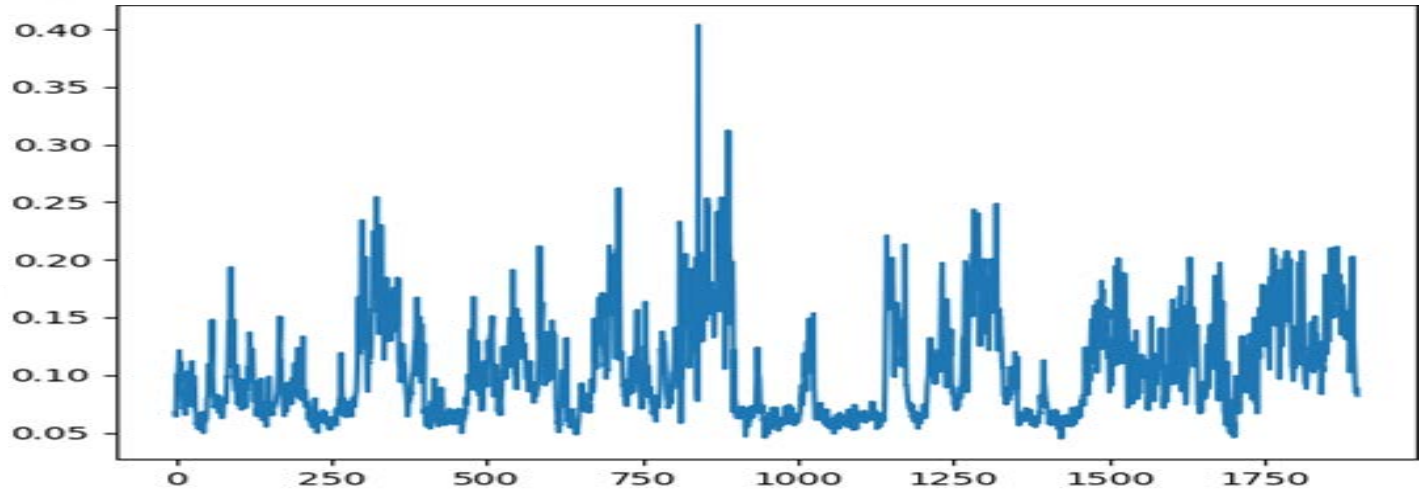




ChIP-Seq



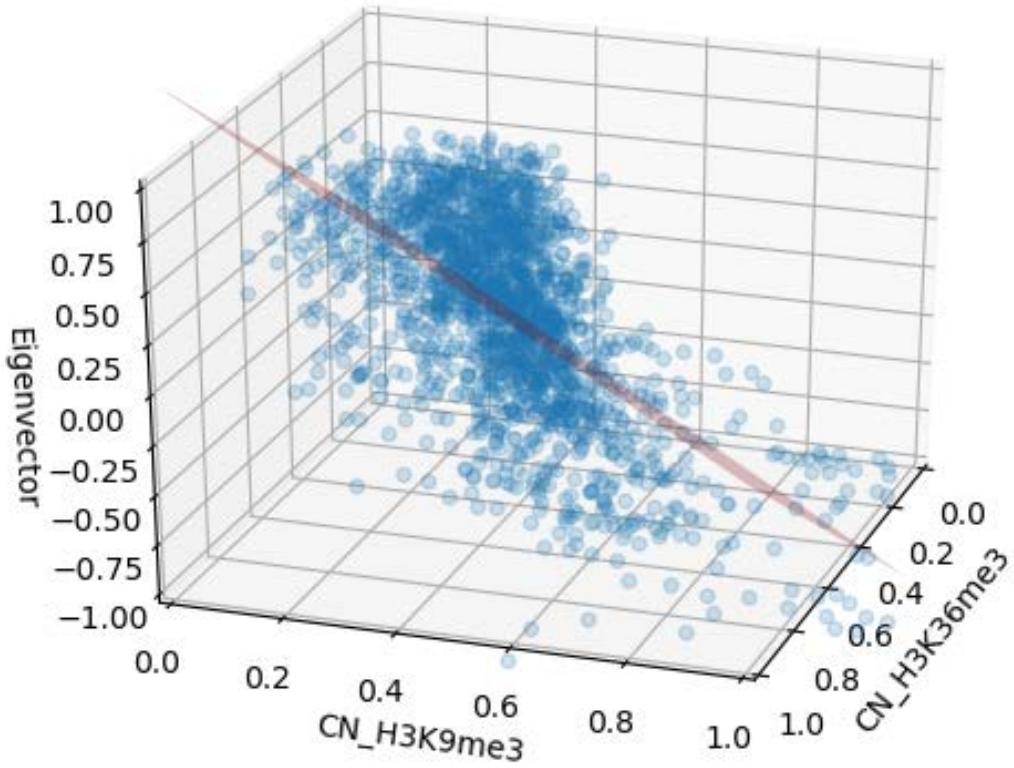
- ChIP-Seq measures the protein occupancy at every point in the genome
- Different regions have different protein contacts



Results

Regression

- Different proteins were used as the independent variables of a regression on the eigenvector to evaluate the influence of each protein



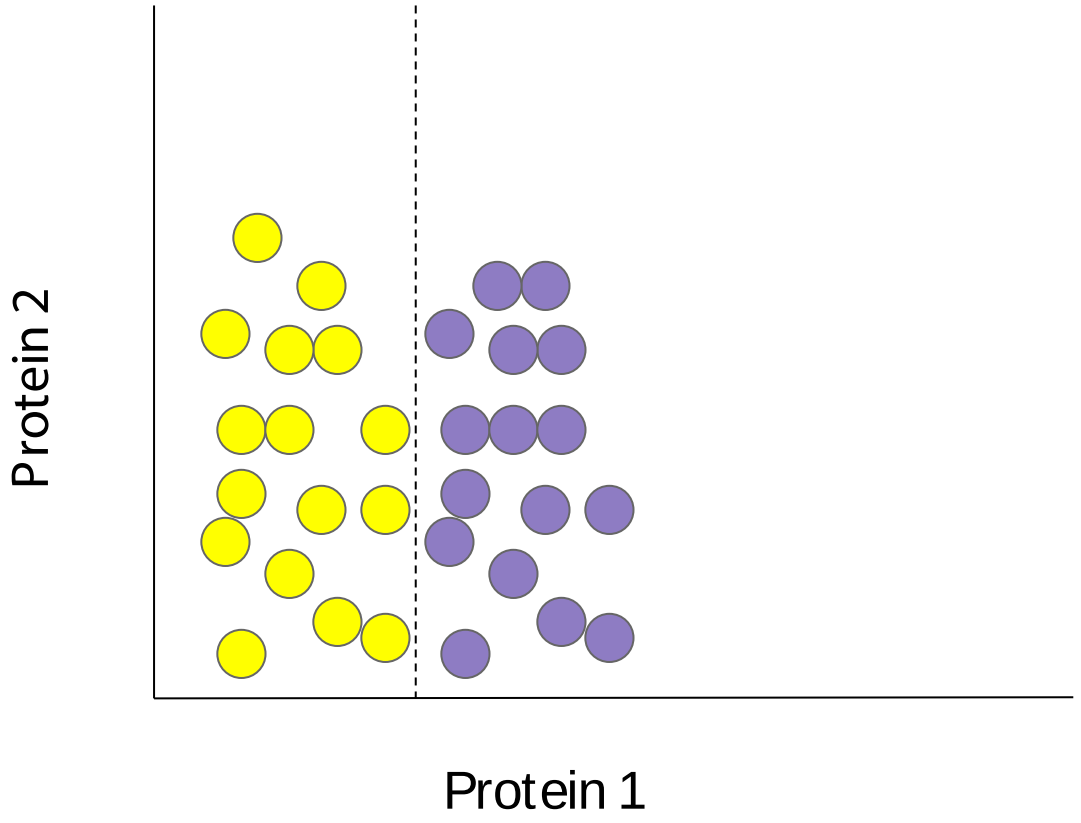
Regression

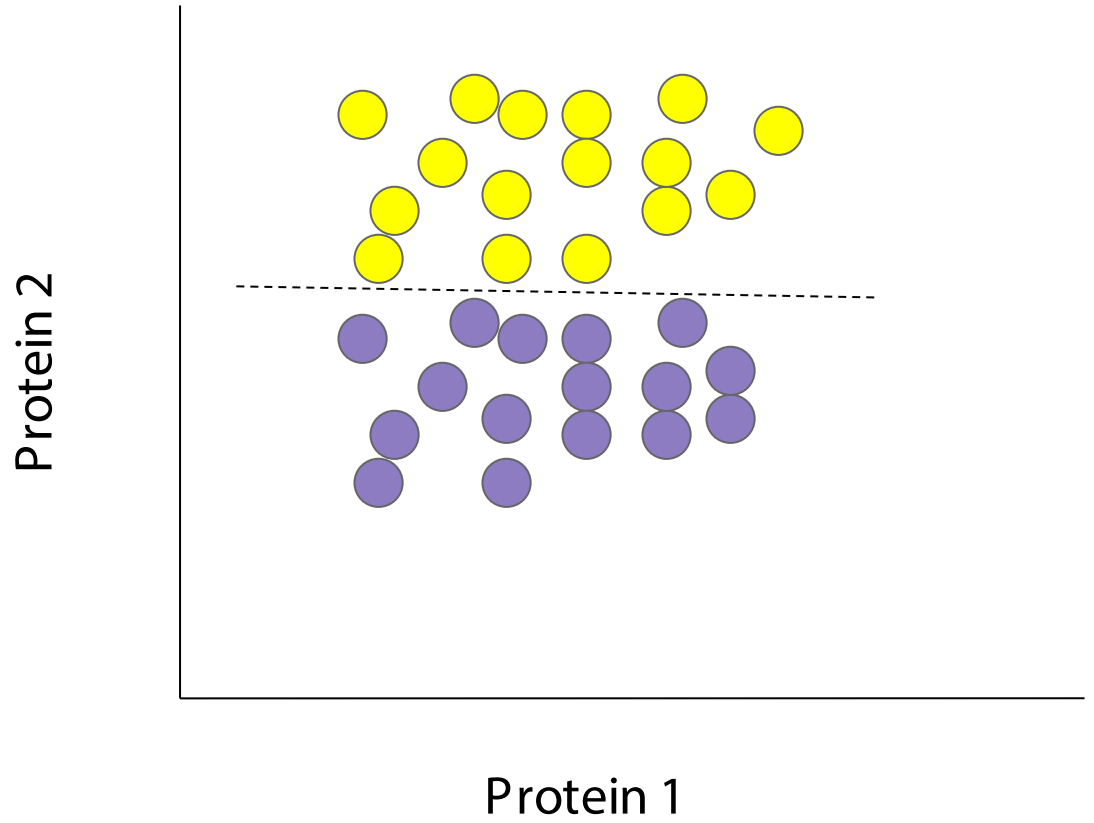
- First, proteins that are highly correlated with others are removed
- The regression is performed with the remaining proteins and the protein with the smallest (absolute value) coefficient is removed
- This process is repeated with the remaining proteins

Regression Coefficients for CN Proteins							
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	-0.147		0.457	1.019	0.865	-0.300	-1.317
2							
3							
4							

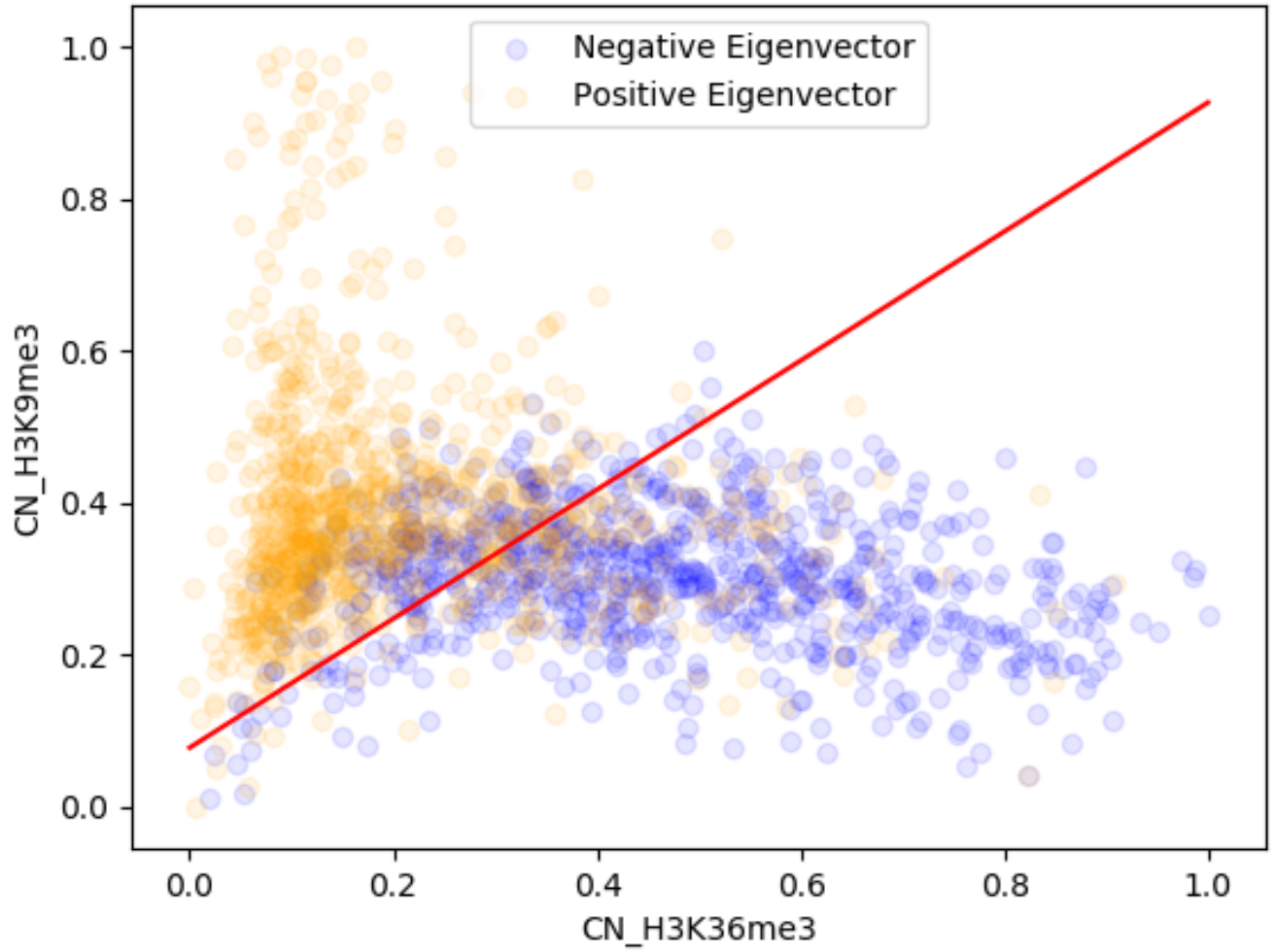
Regression Coefficients for NPC Proteins							
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	-0.096		0.302	0.338	0.324	2.008	-1.112
2							
3							
4							

Support Vector Machine (SVMs)





ChIP-Seq Separation (SVM) Plots



Multidimensional SVM Results

- Suggests that H3K36me3, H3K9me3 and H3K4me3 are the most influential proteins for CN

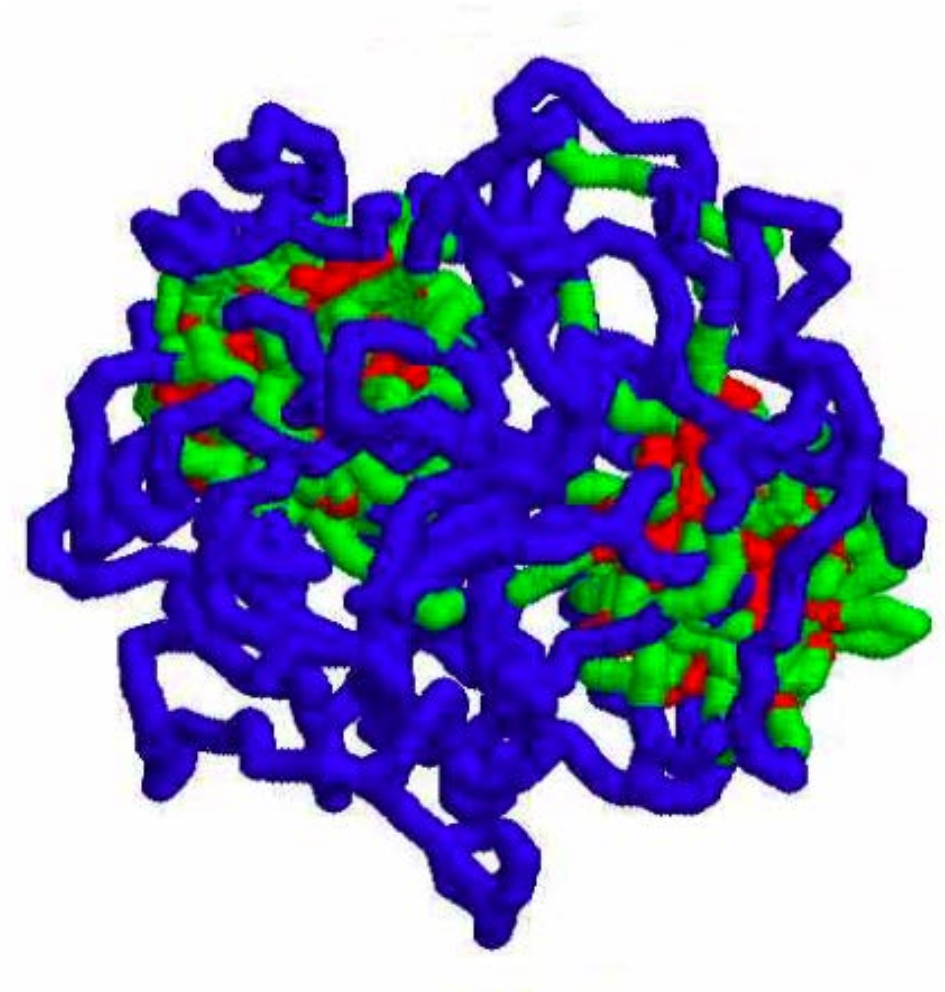
Classification Coefficients for CN Proteins							
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	-2484.477		4353.858	17296.220	18691.338	-10434.279	-22160.805
2			8711.859	15103.567	19609.969	-11174.201	-29771.394
3				10831.438	17194.062	-6627.479	-18027.950
4				6949.258	10710.975		-11220.823

Classification Coefficients for NPC Proteins							
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	1463.089		6560.088	6151.266	28590.458	54607.701	-32237.576
2			302.190	7279.376	24280.016	51135.497	-20749.016
3				5961.584	21359.202	39107.273	-16058.077
4					18949.180	28035.934	-12969.610

Simulations

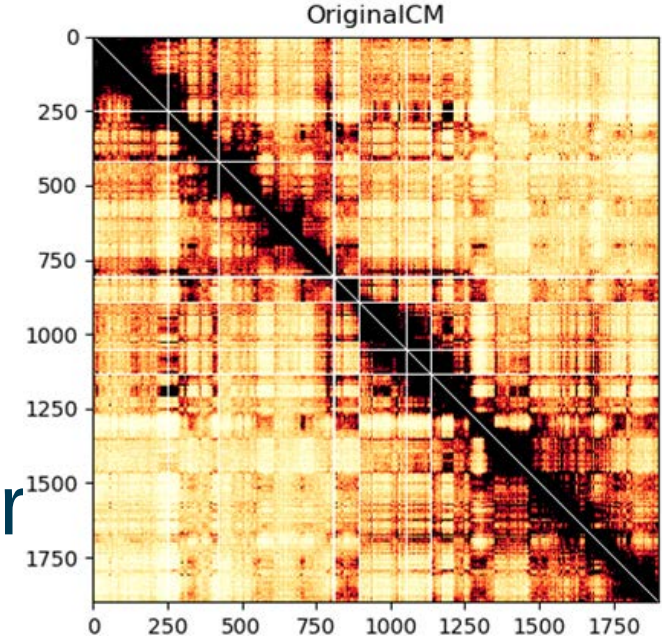
Simulation Overview

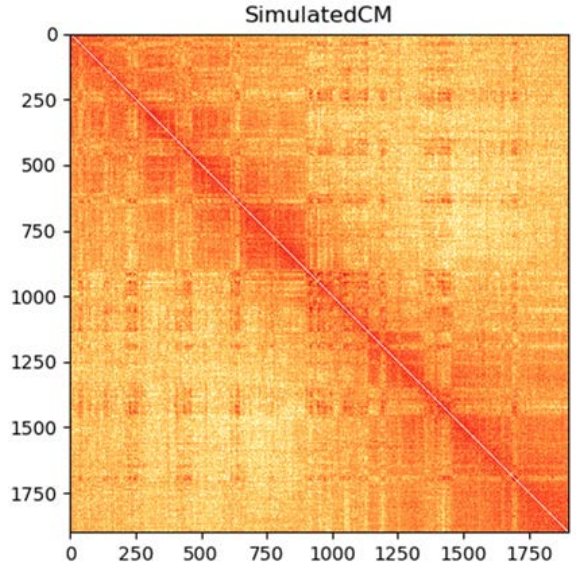
- OpenMM was used to model forces exerted on a polymer
  - Random thermal force
  - Attractive & repulsive forces between monomers



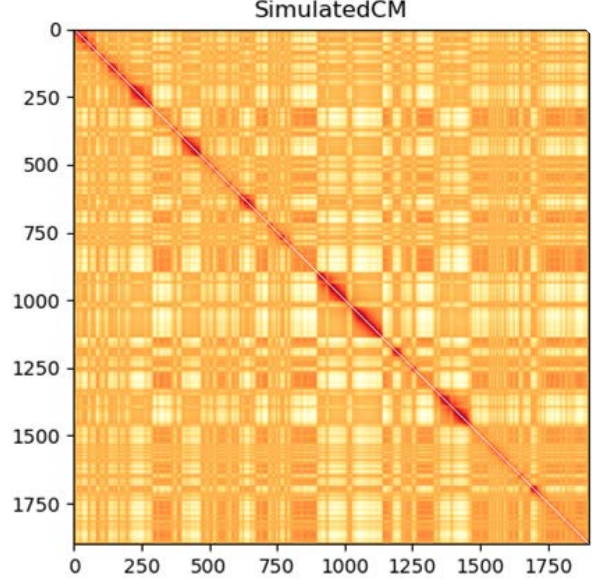
Simulation Data Analysis

- The simulated contact map and eigenvector is calculated and compared with the experimental data
- A high correlation between eigenvectors indicates that the simulation is realistic in modeling the nucleus

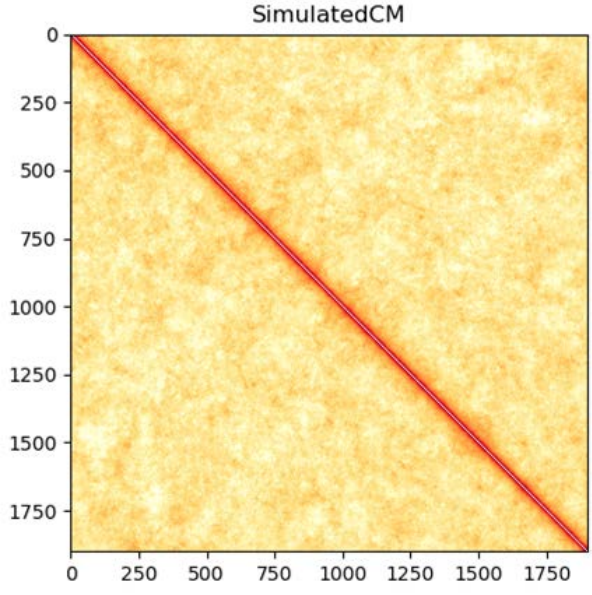




Simulation Models



- Variable “stickiness”
  - The stickiness of monomers was determined by the ChIP-Seq (transformed by various functions)
- Stochastic “stickiness”
  - Binary stickiness was assigned to monomers with a random probability based on the value of the corresponding ChIP-Seq track



Future Work

- Simulations with multiple ChIP-Seq tracks
- Generate more data for stochastic models
- More rigorous methods to find the most influential proteins from ChIP-Seq

Acknowledgements

Thanks to:

- Our mentors Martin Falk and Sameer Abraham for their guidance
- Dr. Gerovitch and the PRIMES program for giving us this opportunity
- Our parents