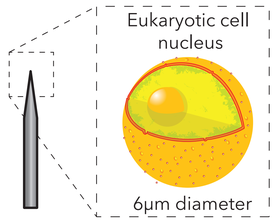
A multi-scale ensemble model of chromatin conformation
~ or ~ DNA folds like a bundle of yarn in every cell

Ben Siranosian
Computational Biology
Undergraduate Honors Thesis

Stretched end-to-end, the DNA in every human cell would reach over

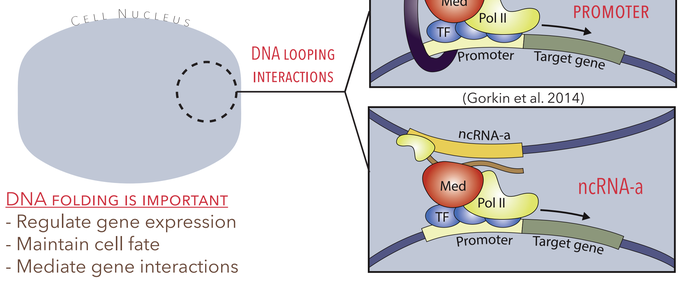
2 METERS LONG!

To fit in a cell nucleus, DNA folds up into an area 1/100,000 the size of the head of a pin.



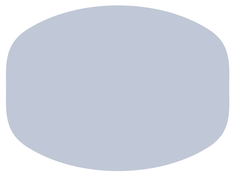
Questions
How is it organized?
Why so tightly folded?
How can we study it?

DNA forms a complex 3D structure inside each cell. Think of it like a piece of yarn crumpled up.



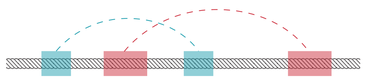
DNA folding is important
- Regulate gene expression
- Maintain cell fate
- Mediate gene interactions

The Hi-C experiment



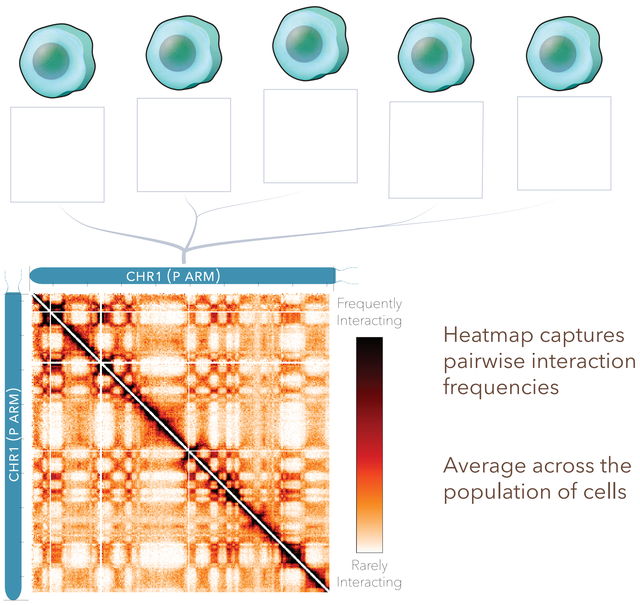
CELLULAR CHROMATIN

Hi-C reveals which sections fold together in the cell



PAIRWISE INTERACTION FREQUENCY

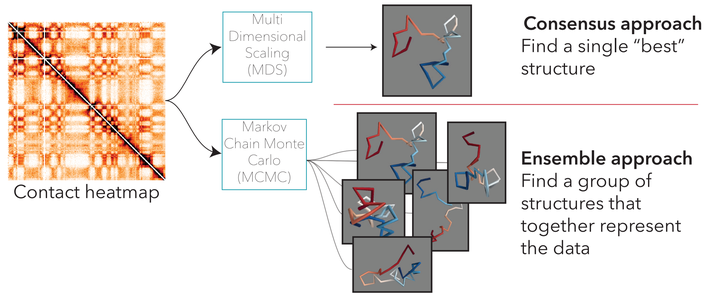
MILLIONS OF CELLS



Heatmap captures pairwise interaction frequencies

Average across the population of cells

THE 3D RECONSTRUCTION PROBLEM



Consensus approach
Find a single "best" structure

Ensemble approach
Find a group of structures that together represent the data

An ensemble approach is better
- Matches experimental design
- Hi-C done on a population of cells
- Average properties across many structures

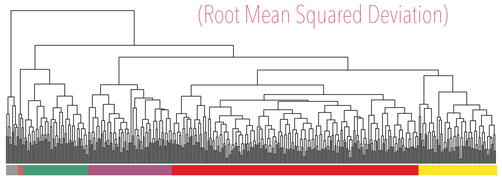
A 2-STEP ENSEMBLE APPROACH

My solution to the 3D reconstruction problem

Step 1

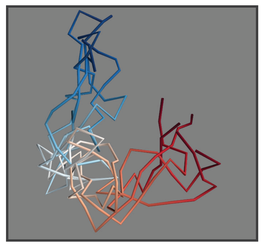
- Break the genome into smaller pieces.
- Randomly fold each piece millions of times, trying to get closer to the "true structure."
- Save structures every 100,000 folds. Generate an ensemble of 5000 structures.
- Compare to the experimental data.
- Analyze properties of the structures.

Clustering on RMSD
(Root Mean Squared Deviation)



Cluster	Characteristic	%
	elongated center, folded left end	0.03%
	elongated structure	0.01%
	folded at both ends	13%
	folded in center and at right end	17%
	very folded, typical loop domain	50%
	ends always close	16%

3D alignment

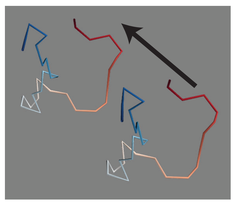


Step 2

Step 1 only works for small pieces of DNA. Step 2 uses a similar Markov Chain Monte Carlo (MCMC) sampling procedure, but for the whole chromosome.

- Start with a member of the ensemble for each DNA section
- Change the organization of the structures randomly.
- Attempt to better reconstruct the experimental data.

THE MCMC MOVES

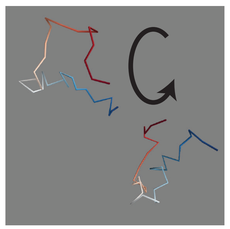


Translate

Move an element linearly in 3D space

If C is a nx3 vector containing the points in 3D space, the translation is:

$$C' = C+v$$

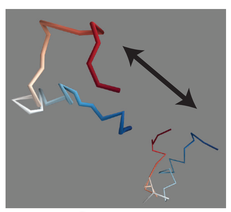


Rotate

Rotate a structure by a random amount around x, y and z axis

R is a rotation matrix, then:

$$C'' = RC'$$

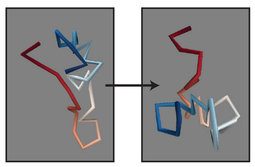


Scale

Scale a structure by a random factor

S is a scaling matrix, then:

$$C''' = SC''$$



Swap

Change for a different member of the ensemble

References

Giorgetti, L. et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell 157, 950-63 (2014).
Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell 14, 762-775 (2014).
Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289-93 (2009).

Acknowledgments

Nicola Neretti and Sorin Istrail for providing guidance through the entire project.
Members of the Neretti lab, especially Steven Criscione and Alan Hwang.
Brown CCMB
Anyone who supported me on this thesis journey, thank you!

CCMB