Genome Organization

CS279
November 14, 2019
Adrian Sanborn
Questions for today

• What interesting biology can you learn from understanding and simulating DNA?
• What are similarities and differences between protein and genome simulations?
• How do you design simulations across different scales?
• What does current research in genome organization look like?
The Human Genome

• Your genome is your “source code.” But it is also a collection of very long polymers that have to be stored, maintained, and replicated.
• This is a lot of work for the cell!
One genome, different cell types

• All the cells in your body have (basically) the same DNA but radically different shapes and functions. How does this happen?

• Different genes are turned on ("expressed") or off in different cells. This generates different cell types

• This is a central problem in biology
Epigenetics

• *Epigenetics* is the study of how the same genome produces different functions.
  – DNA can be chemically marked
  – DNA can be left accessible or packaged away
  – DNA can form loops that bring together important sites
  – RNA is also regulated in many different ways
→ Gene regulation is closely tied to 3D genome organization.
Genome organization can be studied at many different scales.

1) DNA
   - Short region of DNA double-helix
   - 10 bp

2) Chromatin
   - "Beads-on-a-string" form of chromatin
   - 100 bp
   - 1 Kb
   - 10 Kb
   - 100 Kb

3) Loops and domains
   - Extended scaffold-associated form
   - 100 Kb
   - 1 Mb
   - 10 Mb

4) Whole chromosomes
   - Condensed scaffold-associated form
   - 100 Mb
   - 10 Mb
   - 100 Mb
1. DNA
DNA

- The structure of DNA is well-understood: the double helix!
- DNA, which stores information, has a much less dynamic structure than protein, which act as molecular machines.
- Simulations of DNA often look at protein binding to DNA.

E.g. CRISPR-Cas9 cleavage of DNA

Zuo and Liu, Scientific Reports 2016
DNA is relatively stiff

- DNA has repeating negative charge – it is a polyelectrolyte
- The negative charges repel, making DNA quite inflexible at the scale of 10s of base pairs
- Ions in solution affect DNA flexibility
  - Monovalent ions screen charge
  - Divalent cations (Mg2+) promote compaction
  - Important for simulation
- How do you compact DNA? By wrapping it around something with positive charges.
2. Chromatin
The nucleosome

- Around 150bp of DNA winds around an octamer of eight *histone proteins*
- Together, the DNA and the histone octamer are called a *nucleosome*
- Histones have many positively charged residues
- This wrapping compacts the genome linearly by about 7x
- Stringing multiple nucleosomes together gives *chromatin* – DNA and its associated proteins.
The chromatin fiber

- Repeating nucleosomes form the 10nm fiber, like beads on a string
- In a test tube, chromatin condenses to form a thick fiber about 30nm in diameter
Structure of the 30nm fiber

- Two main models: solenoid and zigzag
Structure of the 30nm fiber

- Two main models: solenoid and zigzag
- Use coarse-grained simulations to examine possible structures
  - Use metropolis Monte Carlo
  - Parameters: linker length, linker angle, twist angle, nucleosome thickness
  - Compare to experimental measurements: thickness, mass density, spring behavior

Wedemann and Langowski, J. Biophys. 2002
Structure of the 30nm fiber

• More recently, you can also run all-atom simulations
• All-atom challenges:
  – LOTS of atoms
  – many charged atoms can slow computation
  – solvent is very important

Coarse-grained simulations

- DNA base and backbone each represented by single beads.
- Challenges: models can be quite inaccurate, need to validate well

https://www.youtube.com/watch?v=4Z4KwuUfh0A
Nucleosomes and gene regulation

Nucleosomes affect gene expression in two important ways:

1. Presence of a nucleosome at the beginning of a gene inhibits its transcription

2. Chromatin can be *accessible* or *compacted*. Compaction inhibits transcription
   - This is determined by chemical modifications on histone tails
3. Chromatin compartments
Under the electron microscope, the accessible chromatin (*euchromatin*) appears light, while the compact chromatin (*heterochromatin*) appears dark.

http://medcell.med.yale.edu/histology/cell_lab/euchromatin_and_heterochromatin.php
Higher-order genome organization

• Under a microscope, all DNA sequences “look the same,” so you can’t match your image to genomic position
  – There are ways to mark specific DNA sequences (fluorescence in situ hybridization), but very low throughput
• We can use DNA sequencing to map self-contacts of the folded 3D genome
• This method is called *chromosome conformation capture* (3C)

Lieberman-Aiden, van Berkum et al. Science 2009
CONTACT MAPPING
Exploring structure via proximity
The genome is segregated into two major compartments
By zooming in on regions off the diagonal, you can identify additional sub-compartments

Rao & Huntley, et al., Cell 2014
3D genome reconstruction

Given a map of self-contacts, how can you reconstruct the structures that produced it?

1. Knowledge-based: 3C or other data can be used as constraints
2. Mechanism-based: propose a mechanism that organizes the genome and simulate it

Mechanisms that organize the genome are still quite mysterious. Need computation to tackle this problem.
“Minimal Chromatin Model” (MiChroM)

- Knowledge-based: use 1D compartment info
- Beads in the same compartment prefer to interact
- Train energy function on one chromosome and apply it to the whole genome

Di Pierro et al., PNAS 2016
Mechanism-based method: binding of multivalent proteins
• Have beads (proteins) that prefers to bind to one “type” of chromatin, and binds to multiple sites
• This causes clustering that generates compartments

Brackley et al., NAR 2016
Population-based modeling – use contacts as constraints

- Assignment step: decide which contacts are most important for each individual structure
- Modeling step: optimize the structure based on those constraints

Tjong et al., PNAS 2016
Single-cell modeling

- With single-cell experiments, data is much more sparse, but corresponds to a unique structure

Stevens et al., Nature 2017
4. Loops and domains
The genome is full of “spooky action at a distance”
Loops appear as off-diagonal peaks in the contact map.
Domains tile the genome. Loops often form at their boundaries

*Domain* – a contiguous region of DNA that interacts much more with itself than with its neighbors. Appears as a square on the contact map diagonal.

*Loop* – an interaction between two specific DNA sites. Appears as a peak off the diagonal.
Loop formation correlates with nearby gene activation
How do loops and domains form?

We can propose mechanisms, simulate them, and compare to the contact maps.
Coarse-grained simulations of chromatin

- DNA is a polymer of identical repeating units
  - Each unit represents 1kb (1000 base pairs)
- Local forces drive compaction

\[
U_n = \sum_{i=n-1,n+1} K_{\text{bond}} (R_{\text{bond}} - d_{n,i})^2 \\
+ 4\epsilon_{\text{LJ}} \sum_{i \neq n} \left[ \left( \frac{\sigma_{\text{LJ}}}{d_{n,i}} \right)^{12} - \left( \frac{\sigma_{\text{LJ}}}{d_{n,i}} \right)^6 \right] \\
d_{n,i} = \|x_n - x_i\|
\]

Polymer backbone

Lennard-Jones forces
Coarse-grained simulations of chromatin

- DNA is a polymer of identical repeating units
  - Each unit represents 1kb (1000 base pairs)
- Local forces drive compaction
A simple model: loop anchors are attracted to each other

Add a weak, long-range attractive force between loop anchors. Loops form, but domains are faint.

Simulation Chromatin compaction only
Simulation Long-range looping
Experiment
A clue for a better mechanism: loops are anchored at convergent CTCF sites
Looping via diffusion should not produce convergent sites
Looping via diffusion should not produce convergent sites.
Model: loops and domain form via "extrusion"
A DAY IN THE LIFE OF AN EXTRUSION COMPLEX

Najeeb Tarazi, Adrian Sanborn
Equations for simulating extrusion

\[ U_n = \sum_{i=n-1,n+1} K_{\text{bond}} (R_{\text{bond}} - d_{n,i})^2 \]

\[ + 4\epsilon_{\text{LJ}} \sum_{i \neq n} \left[ \left( \frac{\sigma_{\text{LJ}}}{d_{n,i}} \right)^{12} - \left( \frac{\sigma_{\text{LJ}}}{d_{n,i}} \right)^6 \right] \]

\[ + \sum_{i \in L(n)} K_{\text{loop}} (R_{\text{loop}} - d_{n,i})^2 \]

\[ d_{n,i} = \|x_n - x_i\| \]

Coarse-grained simulations:
- DNA is a polymer of identical repeating units
- Local forces drive compaction
- Certain pairs of sites are bonded in a loop
- Extrusion causes the looped sites to move over time
Extrusion simulations can recapitulate contact maps using just CTCF binding sites

Sanborn, Rao et al., PNAS 2015
LOOP EXTRUSION FORMS
SPATIALLY SEGREGATED DOMAINS
Extrusion simulations predict how genome organization will change when you edit CTCF sites

Sanborn, Rao et al., PNAS 2015
Extrusion simulations predict how genome organization will change when you edit CTCF sites
DNA extrusion by condensin complex was demonstrated in 2018
Common themes: protein simulation vs genome simulation

- A wide variety of data-driven and mechanism-driven approaches are used
- Simulations extend the experiments
- Simulations are used to understand the relationship between structure and function
- Active and expanding areas of research!