2D and 3D Genome Structure Modeling

Jianlin Cheng, PhD

Department of EECS
Informatics Institute
University of Missouri, Columbia
2019
Genome – Code of Life
Genome Sequencing (1D)
The Genomic Era

Collins, Venter, Human Genome, 2000

Personal Genome
Sequencing Revolution
$1000 Personal Genome!

http://www.futuretimeline.net/blog/2014/01/16.htm#.U3Vg-FhdWzs
>95% non-coding regions of a genome are not junk!!
Genome Conformation

Liberman-Aiden et al., 2009
3D Genome Structure is Important

- Spatial gene regulation
- Transcription efficiency
- Genome interpretation
- Function implication (ENCODer)
- Disease diagnosis & treatments
- Drug design
DNA is further packaged. **Nucleosomes** are arranged together into a **fiber** approximately 30 nanometers in diameter. The *precise structure of the chromatin fiber is not known*. Chromatin fiber is further organized into **chromatin loops**, and chromatin loops are further organized into higher-order structures. It has been suggested that **packaging plays a role in gene expression** (gene expression may require associated DNA to open up and acquire an unpackaged conformation). The fully condensed chromosome structure is only seen during **mitosis**.
The DNA is a remarkable molecule in many ways. It encodes our entire genome, and if stretched out in a thin thread would measure 1.8 m in length.
A complex of DNA and basic proteins (such as histone) in eukaryotic cells that is condensed into chromosomes in mitosis and meiosis.

There are two types. **Heterochromatic** is densely-coiled chromatin that appears as nodules in or along chromosomes and contains relatively few genes. **Euchromatic** is the less-coiled and genetically active portion of chromatin that is largely composed of genes.

http://www.daviddarling.info/encyclopedia/C/chromatin.html
Size of Elements

A) DNA

- Nucleosome
- 30 nm chromatin fiber?
- Mitotic chromosome
- Nucleus

B) Disordered loop

http://www.nig.ac.jp/section/maeshima/maeshima-e.html
Chromosomes

http://www.copernicusproject.ucr.edu/ssi/HSBiologyResources.htm
Two Compartments

Linear chromosome:
- Active genes: orange
- Inactive genes: blue

Chromatin looping between active genes and regulatory elements and clustering of genes at the site of active transcription facilitates formation of chromatin globules.

Active gene cluster associate with other expressed genes in active neighborhoods while inactive genes cluster in silent neighborhoods.

Active and silent neighborhoods associate in cis and in trans to form larger active and inactive compartments.

Nuclear organization reflects clustering of active and inactive loci in distinct compartments forming a fractal globule.

A. Sanyal et al., Current Opinion in Cell Biology, 2011
Chromosome Conformation Capturing Techniques
Chromosome Conformation Capturing (Hi-C)

A

1. Crosslink DNA
2. Cut with restriction enzyme
3. Fill ends and mark with biotin
4. Ligate
5. Purify and shear DNA; pull down biotin
6. Sequence using paired-ends
Hi-C Protocol

• Cells are cross-linked with formaldehyde
• DNA is digested with a restriction enzyme and ligated, resulting enriched with cross-linked elements with a biotin marked at the junction
• Shearing DNA and selecting biotin-containing fragments to create a Hi-C library
• Sequence the library to create a catalog of interacting fragments
Genomic Spatial Interaction (Contact) Data

Map reads to the human genome sequence

C[i,j] = # contacts

Intra- / inter-chromosome contact map

Wang et al., 2013
Hi-C Data Analysis Pipeline

- Paired-ends
  - SRR027956.1, end 1
  - SRR027956.1, end 2

A pair of reads were mapped to every chromosome independently.

Mq (-e 150)
- Chromosome 1
- Chromosome 2
- ... (other chromosomes)
- Chromosome Y

Only keep these four cases, all the others removed.

Contact Library

Proximity Matrices
2D Chromosome Contact Map

Chromosome Conformation Capturing
Construct 3D Shape of Genome

Images.google.com
Data Set I

• A normal human lymphoblastoid cell line (B-cell)
• 8.4 million read pairs uniquely mapped to the human genome reference sequence
• 6.7 million corresponded to long-range contacts between segments > 20 kb apart

Liberman-Aiden et al., 2009
Genome Wide Contact Map

• Divide genome into 1-Mb regions (loci)
• $M_{ij}$: number of contacts between loci $i$ and $j$
• The matrix reflects an ensemble average of the interactions in the original sample of cells
• Represented as a heat map
(B) Hi-C produces a genome-wide contact matrix. The submatrix shown here corresponds to **intrachromosomal interactions on chromosome 14**. (Chromosome 14 is acrocentric; the short arm is not shown.) Each pixel represents all interactions between a 1-Mb locus and another 1-Mb locus; intensity corresponds to the total number of reads (0 to 50). Tick marks appear every 10 Mb. (C) We compared the original experiment with results from a biological repeat using the same restriction enzyme [(C)], range from 0 to 50 reads]. Correlation is 0.99.
Relation between Euclidean Distance and Genomic Distance

• Average intrachromosomal contact probability $I_n(s)$ for pairs of loci separated by a genomic distance $s$ on chromosome $n$.
• $I_n(s)$ decreases monotonically on every chromosome
• Even at distances $> 200$ Mb, $I_n(s)$ is always much greater than the average contact probability between different chromosomes.
Probability of contact decreases as a function of genomic distance on chromosome 1, eventually reaching a plateau at ~90 Mb (blue). The level of interchromosomal contact (black dashes) differs for different pairs of chromosomes; loci on chromosome 1 are most likely to interact with loci on chromosome 10 (green dashes) and least likely to interact with loci on chromosome 21 (red dashes). Interchromosomal interactions are depleted relative to intrachromosomal interactions.
Observed/expected number of interchromosomal contacts between all pairs of chromosomes. Red indicates enrichment, and blue indicates depletion (range from 0.5 to 2). Small, gene-rich chromosomes tend to interact more with one another, suggesting that they cluster together in the nucleus.

**Formula:**

\[ E_{i,j} = R_i \times R_j \times N_{INTER}, \]

where \( R_i \) and \( R_j \) are the fractions of inter-chromosomal reads associated with \( i \) and \( j \), respectively, and \( N_{INTER} \) is the total number of inter-chromosomal reads for a cell sample. The actual observed number of inter-chromosomal contacts between chromosomes \( i \) and \( j \) divided by the expected number \( E_{i,j} \) indicates the enrichment or depletion of inter-chromosomal contacts between them.
Dividing each entry in the contact matrix (M) by the genome-wide average contact probability for loci at that genomic distance. The normalized matrix shows many large blocks of enriched and depleted interactions, generating a plaid pattern.
If two loci (here 1-Mb regions) are nearby in space, we reasoned that they will share neighbors and have correlated interaction profiles.

This process dramatically sharpened the plaid pattern; 71% of the resulting matrix entries represent statistically significant correlations (P ≤ 0.05).

The plaid pattern suggests that each chromosome can be decomposed into two sets of loci (arbitrarily labeled A and B) such that contacts within each set are enriched and contacts between sets are depleted.
Key dates

• 4/25/2019: Reading assignment 4 (3D genome) due
• 5/1/2019: Report of Project 3 due
• 5/8/2019: Final presentation of the class
• 5/13/2019: Final report of the class due
The first two principal components (PC) clearly corresponded to the plaid pattern (positive values defining one set, negative values the other).

The entire genome can be partitioned into two spatial compartments such that greater interaction occurs within each compartment rather than across compartments.
FISH Validation of Two Compartments

L1 – L3: Compartment A
L2 – L4: Compartment B
A strong correlation was observed between the number of Hi-C reads $m_{ij}$ and the 3D distance between locus $i$ and locus $j$ as measured by FISH [Spearman’s $r = -0.916$, $P = 0.00003$], suggesting that Hi-C read count may serve as a proxy for distance.

Pairs of loci in compartment B showed a consistently higher interaction frequency at a given genomic distance than pairs of loci in compartment A. This suggests that compartment B is more densely packed.
Open More Accessible, Gene Rich Compartment VS less Accessible

H3K27/H3K36: activating/repressive chromatin Marks
DNAsel: deoxyribonuclease I, sensitivity, measure chromatin accessibility
Contact probability as a function of genomic distance averaged across the genome (blue) shows a power law scaling between 500 kb and 7 Mb (shaded region) with a slope of $-1.08$ (fit shown in cyan).

When plotted on log-log axes, $I(s)$ exhibits a prominent power law scaling between $\sim500$ kb and $\sim7$ Mb, where contact probability scales as $s^{-1}$. This range corresponds to the known size of open and closed chromatin domains.
• **Power-law** dependencies can arise from polymer like behavior.

• **Equilibrium globule**: a compact, densely knotted configuration originally used to describe a polymer in a poor solvent at equilibrium.

• **Fractal Model**: This highly compact state is formed by an unentangled polymer when it crumples into a series of small globules in a “beads-on-a-string” configuration. These beads serve as monomers in subsequent rounds of spontaneous crumpling until only a single globule of globules-of-globules remains. (hierarchical organization)
Fractal Model

• Lack knots and would facilitate unfolding and refolding, for example, during gene activation, gene repression, or the cell cycle.

• In a fractal globule, contiguous regions of the genome tend to form spatial sectors whose size corresponds to the length of the original region.

• In contrast, an equilibrium globule is highly knotted and lacks such sectors; instead, linear and spatial positions are largely decorrelated.
Comparison of Two Models

**FOLDED POLYMER**
- Equilibrium globule
- Cross-section view

**Fractal globule**
- Cross-section view
The equilibrium globule model predicts that contact probability will scale as $s^{-3/2}$, which we do not observe in our data.

We analytically derived the contact probability for a fractal globule and found that it decays as $s^{-1}$; this corresponds closely with the prominent scaling we observed ($s^{-1.08}$).

3D distance between pairs of loci: $s^{1/2}$ for an equilibrium globule, $s^{1/3}$ for a fractal globule. Although 3D distance is not directly measured by Hi-C, we note that a recent paper using 3D-FISH reported an $s^{1/3}$ scaling for genomic distances between 500 kb and 2 Mb.
Monte Carlo simulations to construct ensembles of fractal globules and equilibrium globules (500 each).

Contact probability (for fractal globules, $s^{-1}$, and for equilibrium globules, $s^{-3/2}$)

3D distance (for fractal globules $s^{1/3}$, for equilibrium globules $s^{1/2}$)

Lack of entanglements and the formation of spatial sectors within a fractal globule.
Multi-Scale Fractal Model
Hi-C Data Analysis II

Data Sets

• Primary human acute lymphoblastic leukemia (B-ALL) B-cell
• The MHH-CALL-4 B-ALL cell line (CALL4)
• The follicular lymphoma cell-line (RL)
• Sequenced by Illumina HiSeq 2000
### Number of Reads of the samples

<table>
<thead>
<tr>
<th>Samples</th>
<th>Total number of reads</th>
<th>Utilized for analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal B cell</td>
<td>12,887,282</td>
<td>YES</td>
</tr>
<tr>
<td>RL1</td>
<td>60,272,006</td>
<td>NO</td>
</tr>
<tr>
<td>RL2</td>
<td>61,043,078</td>
<td>NO</td>
</tr>
<tr>
<td>RL3</td>
<td>65,579,872</td>
<td>NO</td>
</tr>
<tr>
<td>RL4</td>
<td>125,256,746</td>
<td>YES</td>
</tr>
<tr>
<td>Call4_1</td>
<td>62,741,712</td>
<td>NO</td>
</tr>
<tr>
<td>Call4_2</td>
<td>62,607,906</td>
<td>NO</td>
</tr>
<tr>
<td>Call4_3</td>
<td>133,542,778</td>
<td>YES</td>
</tr>
<tr>
<td>ALL B-Cell</td>
<td>77,888,742</td>
<td>YES</td>
</tr>
</tbody>
</table>

### Read coverage

<table>
<thead>
<tr>
<th></th>
<th>Read coverage of gene region</th>
<th>Read coverage of non-gene region</th>
<th>Read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call4 cell line</td>
<td>2.81129121903121</td>
<td>2.35780895</td>
<td>100</td>
</tr>
<tr>
<td>RL cell line</td>
<td>1.47413416423764</td>
<td>1.14648699</td>
<td>100</td>
</tr>
<tr>
<td>Normal B-cell</td>
<td>0.186290512630489</td>
<td>0.1725446</td>
<td>76</td>
</tr>
<tr>
<td>ALL B-cell</td>
<td>1.788587532589</td>
<td>1.49037874</td>
<td>120</td>
</tr>
</tbody>
</table>
The sequencing quality score at a position is calculated as $Q_{\text{solexa}} = -10 \log_{10} \frac{p}{1-p}$, where $p$ is the probability of a sequencing error at the position. A score 30 means the probability of a sequencing error at the position is $\sim 0.001$. A score 20 or above may be considered acceptable.
Mapping Reads to Human Genome

UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

Position/search chr7:27,104,502-27,212,501

Scale
chr7:1 27120000 27130000 27140000 27150000 27160000 27170000 27180000 27190000 27200000 27210000 27220000 27230000

UCSC Genes Based on RefSeq, UniProt, GenBank, CGD and Comparative Genomics

Human ESTS That Have Been Spliced

ENCOD Enhancer- and Promoter-Associated Histone Mark (H3K4me1) on 8 Cell Lines

ENCOD Promoter-Associated Histone Mark (H3K4me3) on 9 Cell Lines

ENCOD Digital Dnase I Hypersensitivity Clusters

ENCOD Transcription Factor ChIP-seq

Mammal Cons

Multiz Alignments of 44 Vertebrates

Simple Nucleotide Polymorphisms (dbSNP build 138)
Un-normalized intra-chromosomal heat maps for primary ALL B-Cell
chromosomal heat maps for primary ALL B-Cell

Sequential Component Normalization:

\[ \frac{M[i,j]}{|M[i]|} \]

\[ \frac{M[i,j]}{|M[j]|} \]

Repeat until symmetric
Correlate intra-contact numbers of 23 pairs of chromosomes
Inter-Chromosome Contacts and Chromosome Translocation

A Inter-chromosomal 11-14

100Kb res.

10X enlarge

10Kb res.

121.48Mb

105.42Mb

105.47Mb

100Kb res.

10X enlarge

Chromosome 11

Chromosome 14
Cancer Causing Chromosome Translocation

- Reconstruction of translocated chromosomes
- Two cancer related genes
Comparison of Inter-Chromosome Contact Profiles of Different Cells

The expected number of contacts between chromosome $i$ and $j$ was calculated by:

$$E_{i,j} = R_i \times R_j \times N_{\text{INTER}}$$
Gene-Gene Contact Map of HoxA Gene Cluster:

HoxA gene region (27,104,502 – 27,212,501) on chromosome 7, Transcription factor controlling embryonic development
Hi-C Databases

- 4DN data portal: https://data.4d nucleome.org

- ENCODE database: https://wwwencodeproject.org
3D Genome Structure Modeling

Trieu, Cheng, 2014
Spatial Representation of Chromosome or Genome

- Divided into $N$ equal-size (e.g. 1MB) consecutive units sequentially
- The center of a unit is denoted by a point and its coordinate $(x, y, z)$
Contact Driven Structure Modeling

Random Structure

Normalized Contact Map

Permutation & Optimization

Trieu, Cheng, 2014
Opportunities

• Chromosome contact data can be generated easily and cheaply
• Chromosome contact data is rather reliable
• A 3D model of a genome is very valuable in studying spatial regulation of gene expression and methylation
Challenges

- Genome and chromosome is very large (3 billion nucleotide of human genome)
- Genome structure is very dynamic
- No known experimental genome structure other than some point distance data generated by FISH
- Relationship between contact and distance is not deterministic
- Hi-C data is noisy
Maximum Likelihood or Distance-Based Approaches

• Convert interaction frequency (contact number) into distance
  \[ d \propto \frac{1}{IF^a} \]

  or

  \[ IF \propto \frac{1}{d^a} \]

• Translate distance into \((x, y, z)\) coordinates
A MCMC Approach

MCMC5C

• Formulate a probabilistic model linking 5C/Hi-C data to physical distances

• **Markov chain Monte Carlo (MCMC)** approach called MCMC5C to generate a representative sample from the posterior distribution over structures from IF data.
MCMC5C

- Structural properties (base looping, condensation, and local density) were defined in the models.

- Applied these methods to a biological model of human myelomonocyte cellular differentiation and identified distinct chromatin conformation signatures corresponding to each of the cellular states.

- Run on Hi-C data and produce a model of human chromosome 14 at 1Mb resolution that is consistent with previously observed structural properties as measured by 3D-FISH.
Other Existing Methods

• 5C3D (Fraser et al.): translates IF values into physical distance estimates and then uses a gradient descent approach to find the 3D conformations.
• Bau et al. Interactions are modeled with springs whose equilibrium length depends on the observed IF values, subject to certain constraints based on the structure of the 30-nm fiber, optimized by Integrative Modeling Platform.
Other Existing Methods

• Duan et al., convert interaction frequencies to Euclidean distances and then seek conformation minimizing the misfit, with addition of a set of clash avoidance constraints and a few prior known knowledge about the yeast genome organization. The constrained optimization problem is solved to find the best structure.
Possible Drawbacks of Existing Methods

• Objective function (sum of square difference between predicted and derived distance) is debatable.

• Assume each IF is equally reliable.

• The absence of an underlying probabilistic model, preventing the calculation of confidence intervals on specific structural properties (e.g. distance between two genomic sites)
Probalistic Model of Chromatin Conformation

• A chromosome is modeled as a continuous piece-wise linear curve in 3D.

• Theoretical interaction frequency between fragment \( i \) an \( j \), denoted \( IF(i, j) \), is inversely correlated with the distance between two fragments in 3D conformation: \( IF(i,j) = f(D_s(i,j)) \), where \( D_s(i,j) \) is the Euclidean distance between sites \( i \) and \( j \) and \( f \) is an appropriately chosen function.

\[
f(D_s(i,j)) \propto 1/D_s(i,j)^\alpha
\]
\[ \hat{IF}(i, j) | IF(i, j), \sigma(i, j) \] = N(\hat{IF}(i, j); IF(i, j), \sigma(i, j)^2) \\

\text{re } N(x; \mu, \sigma^2) \text{ is the normal density function}
Observed IF and Theoretical IF (Hi-C)

\[ \Pr[\hat{r}(i, j) | r(i, j)] = N(\hat{r}(i, j); r(i, j), r(i, j) + \kappa). \quad (2) \]

The role of \( \kappa \), which we set to 10, is to avoid having small read counts being assigned too low a variance.
Posterior Probability (Structure | IF data)

The observed data \( \hat{IF} \) defines a posterior distribution over the set of possible conformations of the chromatin: 
\[ \Pr[S|\hat{IF}] = \frac{\Pr[\hat{IF}|S] \cdot \Pr[S]}{\Pr[\hat{IF}]} \].
Since there are no constraints imposed on the structure space and the probability of the observed data (\( \hat{IF} \)) is constant with respect to \( S \), we get 
\[ \Pr[S|\hat{IF}] = \zeta \cdot \Pr[\hat{IF}|S] \],
for some constant \( \zeta \), and thus

\[ \Pr[S|\hat{IF}] = \zeta \cdot \prod_{i,j} \Pr[\hat{IF}(i,j)|IF(i,j) = f(D_s(i,j), \sigma(i,j))]. \]
Sampling Conformations from Posterior Distribution to maximize probability

• A random structure \( R_0 \) is initially chosen to seed the process \((t=0)\), where each point is placed randomly in a cube of side length \( 10 \times \text{avg}(f(\text{IF})) \).

• Repeat: The current structure \( R_t \) is randomly perturbed to obtain a new structure \( R'_t \). If \( \Pr[R'_t | \text{IF}] > \Pr[R_t | \text{IF}] \), the perturbation is obtained and we set \( R_{t+1} = R'_t \). Otherwise, we set \( R_{t+1} = R_t \).

• For values of \( t \) sufficiently large, find a structure with high probability: \( \Pr[S | \text{IF}] \), let \( R_t = S \).
Random Structure Perturbation

• Randomly choose one point $S(i)$ along the structure and moving it by a vector $v$ randomly choosing within a sphere of radius $r$ (e.g. $r = 0.25$ nm)

• The likelihood of the resulting structure is then quickly obtained from that of the old by updating the terms corresponding to the pairs of points involving $i$. 
Assessing Mixing

• $R_1, \ldots, R_k$ of early iterations are highly dependent on $R_0$.

• Determine at what point $m$, the Markov process has mixed, i.e., $R_m$ is independent of $R_0$.

• After mixing, i.e. for $k \geq m$, any sample $R_k$ is representative of the target distribution. For $d$ sufficiently large, $R_k$ and $R_{k+d}$ are independent.
Convergence Determination

- Run two independent chains $R$ and $R'$ in parallel, from independently chosen initial conformations $R_0$ and $R_0'$.
- Mixing is achieved if the samples $\{R_{k/2}, \ldots, R_k\}$ and $\{R'_{k/2}, \ldots, R'_k\}$ cannot be distinguished from each other. Specifically, the average pairwise structural distances among $R_k$ is compared to that between $R_k$ and $R'_k$.
- After mixing is achieved, collect samples every $d=k/20$ iterations.
Clustering of Structure Ensembles

• Distance metric: N*N intra-structure distance matrix Ds.

• The distance \((S,T)\) between two structures \(S\) and \(T\) is:

\[
\text{dist}(S, T) = \sqrt{\sum_{i,j} (D_S(i, j) - D_T(i, j))^2}
\]
Structure Clustering

- Hierarchical clustering
- Visualization: tree dendrogram
- Visual inspection is performed to determine the tree height cutoff and number of subfamilies
- Choose maximum likelihood structure from each cluster as representative and assigning it a weight proportional to the number of structures in its cluster.
Hierarchical Clustering

```
f  l  c  e  h  a  k  b  i  d  m  g  j``
The most accurate model is the one that is best able to predict unseen pairwise interaction frequencies. For each of a set of possible a leave-one-out cross-validation was performed. Find a $a$ to minimize

$$MSE(\alpha) = \frac{1}{n} \sum_{(i,j)} (DS_{(i,j);\alpha} (i, j)^{-\alpha} - IF(i,j))^2.$$
Figure 3 Leave-one-out cross-validation. Value of the mean-squared-errors as a function of $\alpha$, obtained for a leave-one-out cross-validation on the HB-1119 dataset. The minimum error is found for an exponent of 2.0, although values of $\alpha$ between 1 and 3 do not produce significantly worse errors.
C (Distance Scaling Factor) Calibration

Without physical measurement of the distance between pairs of points along the sequence, it is difficult to accurately estimate the value of C. However, based on the average IF value of pairs of fragments located less than 5kb apart along the sequence and following Bystricky et al. [51] that packed chromatin has a physical length of 1 nm for every 110-150bp, C was estimated as approximately 50 nm.
Assessment of Mixing

**Figure 4** Mixing of parallel MCMC5C runs (HB-1119 dataset).
Distance between consecutive structures (sampled every $10^6$ iterations) from within one of two parallel MCMC5C runs (blue and red curves) or across the two runs (green curve), on the HB-1119 5C dataset. The runs converge to the same distribution very rapidly (in less than 250 seconds) and the cross-run distance (green) drops to within the same range as the within-run distances (blue and red curves) after $350 \times 10^5$ iterations.
Figure 4 shows that mixing is achieved after approximately $350 \times 10^5$ iterations, which requires less than 250 seconds of running time. Passed this point, structures sampled every $10^6$ steps from the two parallel runs are undistinguishable from each other and sample structures from the same distribution.

250 structures were sampled after burn-in from each of the two runs. The two ensembles of structures were then combined and the 500 structures were clustered based on their structural similarity.

Analysis of the two THP-1 5C datasets produced similar results, and runs of a larger number of parallel MCMC chains confirm that they all sample similar structures.
Gold structure: a computationally constructed 3D structure used to generated IF data.

Simulated structure: models constructed from the IF data of the gold structure.
Verification by Sampling from Simulated True Structures

Figure 6 HB-1119 Structures from simulated data aligned to gold standard structure. The "gold standard" structure is used as a reference structure to which structures from four different parallel MCMC5C runs on simulated data generated from the gold standard structure are aligned. The gold standard structure is shown highlighted with a white glow and the transcription start sites for the HoxA genes are annotated. The structures found from the simulated data are shown in superimposition to the gold standard structure and show a high degree of alignment.
Structure Clustering and Sub-Structure Families. Sub-structure families may correspond to chromatin structures of cells in different stages.

Figure 5: Mixing and subclustering of HB-1119 structures. Mixing and hierarchical clustering (Ward's method) of structure similarity. The five-hundred structures come from two parallel MCMC runs on the HB-1119 dataset (pools of 250 structures from each run were used). The colors along the top indicate which run each structure originated from (run one = blue, run two = red) and demonstrates that the sampling process has successfully mixed. The blocks in the heatmap and the dendrogram indicate the presence of sub-clusters of structures (numbered in the dendrogram). The two clusters (numbered 1 and 2) both contain structures from the two parallel runs (blue and red vertical bars), indicating that the structures are conserved across runs and are not an artifact of the burn-in process.
Conformations of HoxA in Undifferentiated and Differentiated Conditions

Figure 7 Models of HoxA cluster before and after differentiation. Maximum likelihood structures found by MCMC5C from the undifferentiated and differentiated THP-1 datasets (A and B, respectively). The HoxA gene transcription start sites are annotated on each of the structures.
Figure 8 THP-1 clustering of undifferentiated and differentiated structures. Hierarchical clustering (Ward’s method) of one-thousand structures from four parallel MCMC runs, two on the undifferentiated THP-1 dataset and two on the differentiated THP-1 dataset (250 structures each). The colors along the top indicate which state each structure originated from (undifferentiated run one = blue, run two = red; differentiated run one = pink, run two = orange) and demonstrate a clear distinction between the two states, indicating that the undifferentiated and differentiated cell states specify different structure signatures.
The subset of fragments that are the most conserved across the ensemble of structures are found to lie within the central core region of the structures. These fragments are spatially close to each other and may be involved in looping contacts that are important for the maintenance of the chromatin structure and are therefore highly conserved.
Model the long arm of human chromosome 14 (88.4 Mb region) from Hi-C data published by Lieberman-Aiden et al. [18] at a 1Mb resolution (89 fragments in total). Lieberman-Aiden et al. [18] proposed the existence of two physically disjoint compartments, whereby compartment A was found to correlate with open and actively transcribed chromatin, while compartment B was found to be more densely packed and repressed.
MCMC5C Availability

• MCMC5C modeling movie:
  https://www.youtube.com/watch?v=2alQvUDIGNc
Distance-based approach: convert IF to distance

• Duan et al., the resulting conversion approximately follows $d \propto 1/\text{IF}$.

• Mateos-Langerak et al. [50] also suggest a relationship of the form $d \propto 1 / \text{IF}^a$.

• Bau et al. [28] convert their IF via a linear transformation of the IF’s z-score.
Spatial Representation of A Genome Region at a Scale

- A genome region (e.g. a chromosome) is divided into $N$ equal / variable size units sequentially.
- The spatial position of the center of a unit is denoted by a point and its coordinate $(x, y, z)$ and constraints on the size of unit (radius of sphere).
- The consecutive points are joined into fragments forming the folding trace of the region.

Structure Construction at One Scale

Initial Structure Representation of Units

Contact Map Between Units

Sampling & Optimization

A Data Driven Optimization Approach

- Convert observed interaction frequencies into distance map

\[ d_{ij} = \frac{1}{IF_{ij}^\alpha} \]

**Objective**: Adjust positions \((x,y,z)\) coordinates of beads in 3D space to satisfy converted expected distances as well as possible

**Challenges**: approximated distances, noise, conflicts

Tuan and Cheng, NAR, 2017
Problem of Traditional Objective Functions

Square Error

\[ \sum_{i<n; j<n} (x_{ij} - d_{ij})^2 \]

- Very sensitive to noisy / conflicting distances
- Cannot build 3D models of large genomes involving noisy inter-chromosomal contacts
Deriving Soft Constraints by Lorentzian Function

- Use Lorentzian function to measure the satisfaction of a distance restraint

\[
\frac{c \times c}{c \times c + (x - d)^2}
\]

[0, 1]

- Tolerate noisy restraints and maximize the satisfaction of feasible restraints

Tuan and Cheng, NAR, 2017
Objective Function

- Objective function
  - Sum of all distance constraints
  - Adjacent beads have maximum weight to enforce their proximity

\[
fn = \sum_{|i-j|=1} \frac{c \cdot c \cdot IF_{\text{max}}}{c \cdot c + (x_{ij} - d_{ij})^2} + \sum_{|i-j| \neq 1} \frac{c \cdot c \cdot IF_{ij}}{c \cdot c + (x_{ij} - d_{ij})^2}
\]

- \(c\) is set to average of \(d_{ij}\) to control gradient vanishing during optimization

Tuan and Cheng, NAR, 2017
Large-Scale Adaptive Step-Size Gradient Ascent
Correlation Between Reconstructed Distances and True Distances

Yeast datasets  (Duan et al. Nature 2010)
A Open Source Tool - LorDG

- Github: [https://github.com/BDM-Lab/LorDG](https://github.com/BDM-Lab/LorDG)
Contact-Driven Modeling at Chromosome Scale

- **Input:** initial representation of chromosome, contact map, and physical distance restraints
- **Objective:** find 3D chromosome structures that satisfy the contact map and physical contact restraints as much as possible.
- **Scoring Function**
- **Optimization**
- **Output:** an ensemble of 3D shapes

Data Preparation

- **Data Sets**: Normal B-Cell, ALL B-Cell
- **Unit Size**: 1Mb
- **Unit Number**: Chr. 1, 248 – Chr. 22, 50
- **Contact map normalization**: \( C_{ij}' = \frac{C_{ij}}{\text{expected IF}} \)
- **Remove noisy contacts with low interaction frequency**

### Normalized Interaction Frequencies (IF) on the normal B-Cell

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Max IF</th>
<th>Min IF</th>
<th>Average IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86.34</td>
<td>0.035</td>
<td>1.35</td>
</tr>
<tr>
<td>2</td>
<td>83.04</td>
<td>0.037</td>
<td>1.53</td>
</tr>
<tr>
<td>3</td>
<td>59.81</td>
<td>0.037</td>
<td>1.74</td>
</tr>
<tr>
<td>4</td>
<td>70.76</td>
<td>0.035</td>
<td>1.95</td>
</tr>
<tr>
<td>5</td>
<td>171.92</td>
<td>0.040</td>
<td>1.84</td>
</tr>
<tr>
<td>6</td>
<td>59.68</td>
<td>0.031</td>
<td>1.86</td>
</tr>
<tr>
<td>7</td>
<td>130.00</td>
<td>0.035</td>
<td>1.99</td>
</tr>
<tr>
<td>8</td>
<td>122.22</td>
<td>0.056</td>
<td>2.29</td>
</tr>
<tr>
<td>9</td>
<td>190.85</td>
<td>0.043</td>
<td>2.86</td>
</tr>
<tr>
<td>10</td>
<td>76.12</td>
<td>0.085</td>
<td>2.31</td>
</tr>
<tr>
<td>11</td>
<td>58.51</td>
<td>0.040</td>
<td>2.19</td>
</tr>
<tr>
<td>12</td>
<td>85.86</td>
<td>0.042</td>
<td>2.23</td>
</tr>
<tr>
<td>13</td>
<td>151.09</td>
<td>0.212</td>
<td>3.62</td>
</tr>
<tr>
<td>14</td>
<td>112.48</td>
<td>0.116</td>
<td>3.42</td>
</tr>
<tr>
<td>15</td>
<td>103.02</td>
<td>0.110</td>
<td>3.14</td>
</tr>
<tr>
<td>16</td>
<td>100.09</td>
<td>0.046</td>
<td>3.31</td>
</tr>
<tr>
<td>17</td>
<td>86.08</td>
<td>0.081</td>
<td>3.12</td>
</tr>
<tr>
<td>18</td>
<td>108.33</td>
<td>0.173</td>
<td>3.80</td>
</tr>
<tr>
<td>19</td>
<td>96.56</td>
<td>0.061</td>
<td>4.82</td>
</tr>
<tr>
<td>20</td>
<td>106.89</td>
<td>0.089</td>
<td>4.05</td>
</tr>
<tr>
<td>21</td>
<td>63.01</td>
<td>0.174</td>
<td>5.98</td>
</tr>
<tr>
<td>22</td>
<td>79.05</td>
<td>0.057</td>
<td>5.83</td>
</tr>
</tbody>
</table>
Scoring Function for Optimization

\[ S = \sum_{|i-j| \neq 1} \left( \tanh(d_c - d(i,j)) \cdot \frac{IF_{ij}}{T} + W1 \cdot \frac{\tanh(d(i,j) - d_{min})}{T} \right) \]

\[ + \sum_{|i-j| \neq 1} \left( W2 \cdot \frac{\tanh(d_{max} - d(i,j))}{T} + W3 \cdot \frac{\tanh(d(i,j) - d_c)}{T} \right) \]

\[ + \sum_{|i-j| = 1} \left( IF_{max} \cdot \frac{\tanh(da_{max} - d(i,j))}{T} + W1 \cdot \frac{\tanh(d(i,j) - d_{min})}{T} \right) \]

*\text{tanh: hyperbolic tangent function}*

*\text{IF}_{ij}: interaction frequency between units i and j}*

*T: total interaction frequencies*

*d_c: contact distance threshold*

*d_{min}: minimum distance between two units*

*d_{max}: maximum distance between two units*

*da_{max}: maximum distance between two adjacent units*

*W1, W2, W3, W4: weight parameters in order to maximize % satisfied contacts + % satisfied non-contacts*
**Table S1** The weight parameters and the percent of contacts after removing nose for 23 pairs of chromosomes at 1MB resolution.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Percentage of contact pairs</th>
<th>$W_1$</th>
<th>$W_2$</th>
<th>$W_3$</th>
<th>$W_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49%</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>55%</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>66%</td>
<td>1.2</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>65%</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>64%</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>6</td>
<td>70%</td>
<td>1.0</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>7</td>
<td>74%</td>
<td>1.0</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>8</td>
<td>77%</td>
<td>1.0</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>9</td>
<td>75%</td>
<td>1.0</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>10</td>
<td>78%</td>
<td>1.0</td>
<td>1.8</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>11</td>
<td>77%</td>
<td>1.0</td>
<td>1.8</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>12</td>
<td>75%</td>
<td>1.0</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>13</td>
<td>94%</td>
<td>1.0</td>
<td>4.6</td>
<td>4.6</td>
<td>4.6</td>
</tr>
<tr>
<td>14</td>
<td>86%</td>
<td>1.0</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>15</td>
<td>89%</td>
<td>1.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>16</td>
<td>86%</td>
<td>1.0</td>
<td>3.9</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>17</td>
<td>85%</td>
<td>1.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>18</td>
<td>97%</td>
<td>1.0</td>
<td>9.5</td>
<td>9.5</td>
<td>9.5</td>
</tr>
<tr>
<td>19</td>
<td>92%</td>
<td>1.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>20</td>
<td>93%</td>
<td>1.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>21</td>
<td>94%</td>
<td>1.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>22</td>
<td>96%</td>
<td>1.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>23</td>
<td>83%</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Estimating Parameters

• FISH data (Mateos-Langerak et al., 2009) for physical distances of Chr. 1 and 11 at various genomic distances
• $d_{\text{min}}, d_{\text{max}}$: min and max distance between pairs at 1Mb away. (0.2 um , 1.8 um)
• $d_{\text{max}}$: max distance between all pairs
• $d_c$: a threshold resulting in the same percent of contacts in our data
• $d_c$ and $d_{\text{max}}$ are chromosome length dependent (1.73 – 2.24 um, 2.45 – 3.32 um)
Steepest Gradient Ascent with Backtracking Line Search

• **Random Initialization**: \((x_1^0, y_1^0, z_1^0), (x_2^0, y_2^0, z_2^0), \ldots, (x_N^0, y_N^0, z_N^0)\) in \([-0.5, 0.5]\)

• **Update**:

\[
\begin{align*}
X_1^{t+1} &= X_1^t + \eta^* \Delta X \\
Y_1^{t+1} &= Y_1^t + \eta^* \Delta Y \\
Z_1^{t+1} &= Z_1^t + \eta^* \Delta Z \\
X_2^{t+1} &= X_2^t + \eta^* \Delta X \\
Y_2^{t+1} &= Y_2^t + \eta^* \Delta Y \\
Z_2^{t+1} &= Z_2^t + \eta^* \Delta Z \\
\vdots & \quad \vdots \\
X_N^{t+1} &= X_N^t + \eta^* \Delta X \\
Y_N^{t+1} &= Y_N^t + \eta^* \Delta Y \\
Z_N^{t+1} &= Z_N^t + \eta^* \Delta Z
\end{align*}
\]

*Step size \((\eta)\) is adjusted dynamically during iterations to avoid too big moves.*
Structure Modeling Movie

At YouTube without music: http://www.youtube.com/watch?v=C03R7A9kYc8

At NSF CAREER project web site with music:
http://people.cs.missouri.edu/~chengji/genome_modeling_movie.mp4

Two Compartment Validation on Normal B-Cell

Purple and green denote regions in two different components using principle component analysis on contact correlation map.

Model Selection

- Use TM-Score (Zhang & Skolnick, 2004) to superpose every pair of models in an ensemble of models
- Calculate GDT-HA score: percent of unit pairs within specific distance thresholds
- Choose centroid model as representative
## Satisfaction of Contacts in Representative Models of Normal B-Cell

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Satisfied contact pairs (%)</th>
<th>Satisfied Non-contact pairs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>81</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>82</td>
<td>84</td>
</tr>
<tr>
<td>4</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>83</td>
<td>86</td>
</tr>
<tr>
<td>6</td>
<td>83</td>
<td>83</td>
</tr>
<tr>
<td>7</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>8</td>
<td>89</td>
<td>83</td>
</tr>
<tr>
<td>9</td>
<td>88</td>
<td>90</td>
</tr>
<tr>
<td>10</td>
<td>87</td>
<td>88</td>
</tr>
<tr>
<td>11</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>12</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>13</td>
<td>90</td>
<td>89</td>
</tr>
<tr>
<td>14</td>
<td>93</td>
<td>84</td>
</tr>
<tr>
<td>15</td>
<td>91</td>
<td>90</td>
</tr>
<tr>
<td>16</td>
<td>89</td>
<td>88</td>
</tr>
<tr>
<td>17</td>
<td>90</td>
<td>92</td>
</tr>
<tr>
<td>18</td>
<td>89</td>
<td>94</td>
</tr>
<tr>
<td>19</td>
<td>92</td>
<td>96</td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>21</td>
<td>92</td>
<td>99</td>
</tr>
<tr>
<td>22</td>
<td>91</td>
<td>92</td>
</tr>
</tbody>
</table>
## Violations of Contacts on Normal B-Cell

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Average distance (unsatisfied contact pairs)</th>
<th>Average IF (unsatisfied contact pairs)</th>
<th>Average distance (unsatisfied non-contact pairs)</th>
<th>$d_c$</th>
<th>Average IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.98</td>
<td>0.44</td>
<td>1.43</td>
<td>1.73</td>
<td>1.6</td>
</tr>
<tr>
<td>2</td>
<td>1.97</td>
<td>0.52</td>
<td>1.48</td>
<td>1.73</td>
<td>1.84</td>
</tr>
<tr>
<td>3</td>
<td>1.94</td>
<td>0.60</td>
<td>1.51</td>
<td>1.73</td>
<td>2.09</td>
</tr>
<tr>
<td>4</td>
<td>2.05</td>
<td>0.57</td>
<td>1.31</td>
<td>1.73</td>
<td>2.36</td>
</tr>
<tr>
<td>5</td>
<td>2.03</td>
<td>0.55</td>
<td>1.39</td>
<td>1.73</td>
<td>2.22</td>
</tr>
<tr>
<td>6</td>
<td>1.97</td>
<td>0.60</td>
<td>1.42</td>
<td>1.73</td>
<td>2.23</td>
</tr>
<tr>
<td>7</td>
<td>2.37</td>
<td>0.62</td>
<td>1.68</td>
<td>2.00</td>
<td>2.39</td>
</tr>
<tr>
<td>8</td>
<td>2.34</td>
<td>0.69</td>
<td>1.60</td>
<td>2.00</td>
<td>2.76</td>
</tr>
<tr>
<td>9</td>
<td>2.42</td>
<td>0.72</td>
<td>1.72</td>
<td>2.12</td>
<td>3.49</td>
</tr>
<tr>
<td>10</td>
<td>2.58</td>
<td>0.71</td>
<td>1.92</td>
<td>2.23</td>
<td>2.78</td>
</tr>
<tr>
<td>11</td>
<td>2.56</td>
<td>0.71</td>
<td>1.80</td>
<td>2.23</td>
<td>2.6</td>
</tr>
<tr>
<td>12</td>
<td>2.54</td>
<td>0.67</td>
<td>1.89</td>
<td>2.23</td>
<td>2.69</td>
</tr>
<tr>
<td>13</td>
<td>2.55</td>
<td>1.32</td>
<td>1.75</td>
<td>2.23</td>
<td>4.34</td>
</tr>
<tr>
<td>14</td>
<td>2.47</td>
<td>1.15</td>
<td>1.69</td>
<td>2.23</td>
<td>4.12</td>
</tr>
<tr>
<td>15</td>
<td>2.46</td>
<td>1.02</td>
<td>1.85</td>
<td>2.23</td>
<td>3.78</td>
</tr>
<tr>
<td>16</td>
<td>2.58</td>
<td>0.91</td>
<td>1.79</td>
<td>2.34</td>
<td>4.02</td>
</tr>
<tr>
<td>17</td>
<td>2.56</td>
<td>0.93</td>
<td>2.06</td>
<td>2.34</td>
<td>3.78</td>
</tr>
<tr>
<td>18</td>
<td>2.66</td>
<td>1.36</td>
<td>1.72</td>
<td>2.34</td>
<td>4.54</td>
</tr>
<tr>
<td>19</td>
<td>2.57</td>
<td>1.40</td>
<td>1.84</td>
<td>2.34</td>
<td>5.85</td>
</tr>
<tr>
<td>20</td>
<td>2.56</td>
<td>1.35</td>
<td>1.78</td>
<td>2.34</td>
<td>4.88</td>
</tr>
<tr>
<td>21</td>
<td>2.62</td>
<td>1.60</td>
<td>1.93</td>
<td>2.34</td>
<td>7.28</td>
</tr>
<tr>
<td>22</td>
<td>2.58</td>
<td>1.78</td>
<td>1.64</td>
<td>2.34</td>
<td>7.05</td>
</tr>
</tbody>
</table>
3D Models for 22 Chromosomes of Normal B-Cell

Models: 200KB resolution
Models of Chr. 2 for Normal and Malignant Cells

Primary Leukemia B-Cell

Normal B-Cell

The similarity scores measured as average GDT-HA scores. Y-axis denotes the similarity scores and X-axis the indices of chromosomes. ‘Blue bars’ represent the average GDT-HA scores of models within the same model ensemble constructed from the whole normalized data sets of the normal B-cell for each chromosome, ‘red bars’ the average GDT-HA scores between models constructed from sampled data sets with those constructed from the whole data sets and ‘green bars’ the average GDT-HA scores between models of the leukemia B-cell and those of the normal B-cell.
The percentage of recovered contacts in all chromosomes in validation

The superimposition of the plots of IFs of all missing contacts and the plots of IFs of recovered contacts for chromosome 1 (left) and chromosome 11 (right). Y-axis denotes interaction frequencies and X-axis the indices of contacts. The green tails visualized the contacts that were not recovered and had lower IFs.
Average of spatial distances and IFs of region pairs within and across (between) compartments in chromosomes 1 and 11.


© The Author(s) 2014. Published by Oxford University Press.
The contact scores and non-contact scores of the chromosomal models of the leukemia B-cell.

The models of chromosome 1 at resolution of 200 K (A) and at resolution of 1 MB (B).

Pairwise comparison:
1Mb versus 200Kb
Scoring Function for Genome Optimization

\[ Fn = \sum_{\text{contacts} \ (i,j): |i-j| \neq 1} \left( W_1[\text{chr1}, \text{chr2}] \cdot \tanh(d_c^2 - d_{ij}^2) \cdot \frac{F_{ij}}{\text{totalIF}} + W_2[\text{chr1}, \text{chr2}] \cdot \frac{\tanh(d_{ij}^2 - d_{\min}^2)}{\text{totalIF}} \right) + \]

\[ \sum_{\text{(i,j): |i-j|=1 & chr1=chr2}} \left( W_1[\text{chr1}, \text{chr2}] \cdot IF_{\text{max}} \cdot \frac{\tanh(d_{\max}^2 - d_{ij}^2)}{\text{totalIF}} + W_2[\text{chr1}, \text{chr2}] \cdot \frac{\tanh(d_{ij}^2 - d_{\min}^2)}{\text{totalIF}} \right) + \]

\[ \sum_{\text{non-contacts} \ (i,j): |i-j| \neq 1, \text{chr1}=\text{chr2}} \left( W_3[\text{chr1}, \text{chr1}] \cdot \frac{\tanh(d_{\max}^2 - d_{ij}^2)}{\text{totalIF}} + W_4[\text{chr1}, \text{chr1}] \cdot \frac{\tanh(d_{ij}^2 - d_{\max}^2)}{\text{totalIF}} \right) + \]

\[ \sum_{\text{non-contacts} \ (i,j): |i-j| \neq 1, \text{chr1} \neq \text{chr2}} \left( W_3[\text{chr1}, \text{chr2}] \cdot \frac{\tanh(d_{\max}^2 - d_{ij}^2)}{\text{totalIF}} + W_4[\text{chr1}, \text{chr2}] \cdot \frac{\tanh(d_{ij}^2 - d_{\max}^2)}{\text{totalIF}} \right) + \]

\[ d_{\min}^2 = 0.2 \mu m^2, \ d_{\max}^2 = 1.8(\mu m^2) \]
\[ d_c^2 = 6\mu m^2, \ d_{\max}^\text{intra} = d_{\max}^2 = 20\mu m^2 \]
\[ d_{\max}^\text{inter} = 13\mu m \]

Contact Pairs
Adjacent Pairs
Non-contacts on same chromosome
Non-contacts on different chromosomes

Trieu, Cheng, Bioinformatics, 2016
Modeling of 3D Genome
Test on Simulated Data of the Yeast Genome Structure

Known Structure (Duan et al., 2010)  Reconstructed Model

Correlation vs. Noise level

Trieu, Cheng, Bioinformatics, 2016
A 3D Model of Genome of Human B-Cell in an Ensemble

Trieu, Cheng, Bioinformatics, 2016
Largely Conserved Chromosomal Structure, Dynamic Genome Structure

Chromosome 11

Pairwise structural similarity between chromosomal models: 0.71
Pairwise structural similarity between genome models: 0.16

Trieu, Cheng, Bioinformatics, 2016
Center – Periphery Architecture and Dynamic Inter-chromosomal Interactions

- Small chromosomes in center
- Large chromosomes in periphery
- Genome structure ensemble (dynamics)
- Correlation of inter-chromosomal interactions (0.48), p-value < e-16

Trieu, Cheng, Bioinformatics, 2016
Enriched Telomere and Centromere Inter-Chromosomal Interactions

- Telomere-telomere interactions > telomere/centromere interactions > other interactions

Trieu, Cheng, Bioinformatics, 2016
MOGEN: Model of Genome (Tool)

1. Reconstruction process video: http://calla.rnet.missouri.edu/mogen/video/3D Genome_Movie.mp4

2. Genome structures of healthy cells: http://calla.rnet.missouri.edu/mogen/normal_cell/

3. Genome structures of malignant cells: http://calla.rnet.missouri.edu/mogen/cancer_cell/

4. Experimental data on synthetic datasets: https://missouri.box.com/s/5cnwys9e1qgt8o20pdh8544meb5myay0

5. MOGEN: Executable program - https://missouri.box.com/s/a8nynixow4aixj6fd0k2xf6n4wof772

http://calla.rnet.missouri.edu/mogen/
MOGEN at GitHub

- [https://github.com/BDM-Lab/MOGEN](https://github.com/BDM-Lab/MOGEN)
3D Structure of the Entire Genome
Smaller Chromosomes are Closer to the Center
Description

GMOL is an application designed to visualize genome structure in 3D. It allows users to view the genome structure at multiple scales, including: global, chromosome, loci, fiber, nucleosome, and nucleotide. This software was built upon the pre-existing Jmol package.

GMOL Web Site

Categories

Bio-Informatics, Visualization

License

GNU Library or Lesser General Public License version 3.0 (LGPLv3)

Features

- Interactively visualize genome structures in 3D
- Supports multiple scales/resolutions: global, chromosome, loci, fiber, nucleosome, nucleotide
- Measure distances and angles between points in the structure
- Rotate and scale models to analyze them even more
- Select parts of the structure based on index, scale, or sequence information
- Get DNA sequence for selected structures or portions of structures
- Includes existing Jmol functions
GMOL - Multi-Scale Visualization of 3D Genome Structure
Hierarchical High-Resolution 3D Genome Modeling
A Hierarchical Modeling Algorithm – A Divide-Conquer Strategy
Two-Scale Hierarchical Modeling

Trieu et al., 2018
Ratio of Scaling Global Low-Resolution Model to Match Local High-Resolution Models

\[ d_{xy} = \min_{i,j} (d_{ij} + d_{xi} + d_{yj}) \]
Model Refinement by Inter-Domain Distance Restraints

• Intra-domain distance restraints have been satisfied
• Inter-domain orientations are adjusted according to distance between the centers of domains
• The distances between adjacent domains are enforced
A High-Resolution Model at 5KB
Correlation between reconstructed distances and wish distances
Consistency between 5KB and 1MB Models

Similar topology between 5kb-resolution model (left) and 1mb-resolution model (right)

Correlation between models at 1MB and 5KB resolution
Validation with FISH Data

<table>
<thead>
<tr>
<th>Chr</th>
<th>L1 (position, MB)</th>
<th>L2 (position, MB)</th>
<th>L3 (position, MB)</th>
<th>L1-L2</th>
<th>L2-L3</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>130.72-130.75</td>
<td>130.29-130.32</td>
<td>129.86-129.89</td>
<td>0.8</td>
<td>7.1</td>
</tr>
<tr>
<td>13</td>
<td>86.37-86.40</td>
<td>85.46-85.49</td>
<td>84.55-84.58</td>
<td>1.9</td>
<td>13.9</td>
</tr>
<tr>
<td>14</td>
<td>71.60-71.63</td>
<td>72.20-72.23</td>
<td>72.80-72.83</td>
<td>1.2</td>
<td>7.8</td>
</tr>
</tbody>
</table>
Comparison with miniMDS

Blue: Hierarchical3DGenome
Red: miniMDS

Models of human Chromosome 19 of 5KB resolution produced by Hierarchical3DGenome (A) and miniMDS (B) respectively. The boundaries between TADs in miniMDS models are not well joined.
IV. 3D Genome Modeling Tools
GenomeFlow

- **1D function**: map Hi-C reads to a reference genome
- **2D function**: generate contact maps from mapped reads, normalize them, and identify TADs
- **3D function**: reconstruct, visualize, and analyze 3D chromosome/genome models and integrate other genomics data (e.g., gene expression, methylation, ChIP-Seq, histone modification, chromatin accessibility, enhancer, promoter)
- **A user-friendly graphical user interface (GUI)**

Trieu et al., 2018
(A) Visualization of a contact matrix and its TADs

(B) Reconstruction of a genome model in real time

(C) Chromatin loops are identified and highlighted

(D) A model is annotated by two groups of genes (in red and green)
Open-Source Genome Modeling Tools

- LorDG: [https://github.com/BDM-Lab/LorDG](https://github.com/BDM-Lab/LorDG)
- Hierarchical3DGenome: [https://github.com/BDM-Lab/Hierarchical3DGenome](https://github.com/BDM-Lab/Hierarchical3DGenome)
- MOGEN: [https://github.com/BDM-Lab/MOGEN](https://github.com/BDM-Lab/MOGEN)
- Chromosome3D: [https://github.com/multicom-toolbox/Chromosome3D](https://github.com/multicom-toolbox/Chromosome3D)
- Gen3D: [https://sourceforge.net/projects/gen3d/](https://sourceforge.net/projects/gen3d/)
- GMOL: [https://sourceforge.net/projects/gmol/](https://sourceforge.net/projects/gmol/)
- 3DMax: [https://github.com/BDM-Lab/3DMax](https://github.com/BDM-Lab/3DMax)
Project 4

• Apply a gradient descent method LorDG / MOGEN to build 3D models for human chromosome 7.


• How to assess the models (check the paper)?

• Visualization of models

• Reference models (constructed from somewhat different contact data of Chr. 7: [http://calla.rnet.missouri.edu/mogen/](http://calla.rnet.missouri.edu/mogen/)

• LorDG at GitHub: [https://github.com/BDM-Lab/LorDG](https://github.com/BDM-Lab/LorDG)

• MOGEN at GitHub: [https://github.com/BDM-Lab/MOGEN](https://github.com/BDM-Lab/MOGEN)
Data II: Hi-C data of Chromosome 7

HWI-EAS313_0025:7:78:7863:19096|7 40786044 7 45502707|case 0
HWI-EAS313_0025:7:63:14188:6924|7 128562332 7 128562745|case 0
HWI-EAS313_0025:7:94:3739:17610|7 154672365 7 154672125|case 0
HWI-EAS313_0025:7:78:17921:8167|7 90166635 7 90166267|case 0
HWI-EAS313_0025:7:98:5753:2860|7 146851262 7 82792917|case 0
WI-EAS313_0025:7:104:3993:13804|7 101472788 7 101472557|case 0
HWI-EAS313_0025:7:4:13123:19420|7 63615352 7 62522230|case 0
HWI-EAS313_0025:7:54:7610:3364|7 24688078 7 24687498|case 0
HWI-EAS313_0025:7:6:8245:1169|7 47788402 7 47788122|case 0

[http://sysbio.rnet.missouri.edu/3dgenome/contact_ALL_chrom7](http://sysbio.rnet.missouri.edu/3dgenome/contact_ALL_chrom7)

The LorDG /MOGEN software package includes some sample data and examples of how to run the program.
Timeline

• April 30: discussion of plan
• May 2: presentation of the plan
• May 9: Presentation of your results
Acknowledgements

• Tuan Trieu, Oluwatosin Oluwadare, Chenfeng He, Sharif Ahmed, Lingfei Xu
• Zheng Wang, Renzhi Cao, Avery Wells
• Charles Caldwell, Kristen Taylor, Aaron Briley