THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB Article

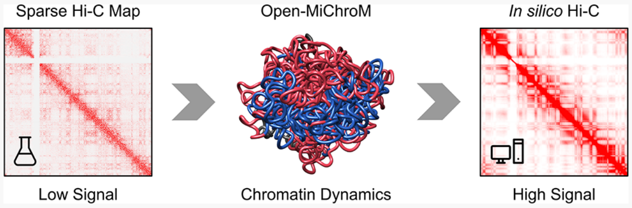
Chromosome Modeling on Downsampled Hi-C Maps Enhances the Compartmentalization Signal

Published as part of The Journal of Physical Chemistry virtual special issue "Yoshitaka Tanimura Festschrift".

Antonio B. Oliveira Junior, Cynthia Perez Estrada, Erez Lieberman Aiden, Vinícius G. Contessoto,* and José N. Onuchic*

Cite This: *J. Phys. Chem. B* 2021, 125, 8757–8767 Read Online

ACCESS | Metrics & More | Article Recommendations

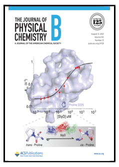


ABSTRACT: The human genome is organized within a nucleus where chromosomes fold into an ensemble of different conformations. Chromosome conformation capture techniques such as Hi-C provide information about the genome architecture by creating a 2D heat map. Initially, Hi-C map experiments were performed in human interphase cell lines. Recently, efforts were expanded to several different organisms, cell lines, tissues, and cell cycle phases where obtaining high-quality maps is challenging. Poor sampled Hi-C maps present high sparse matrices where compartments located far from the main diagonal are difficult to observe. Aided by recently developed models for chromatin folding and dynamics investigation, we introduce a framework to enhance the compartments' information far from the diagonal observed in experimental sparse matrices. The simulations were performed using the Open-MiChroM platform aided by new trained parameters in the minimal chromatin model (MiChroM) energy function. The simulations optimized on a downsampled experimental map (10% of the original data) allow the prediction of a contact frequency similar to that of the complete (100%) experimental Hi-C. The modeling results open a discussion on how simulations and modeling can increase the statistics and help fill in some Hi-C regions not captured by poor sampling experiments. Open-MiChroM simulations allow us to explore the 3D genome organization of different organisms, cell lines, and cell phases that often do not produce high-quality Hi-C maps.

■ INTRODUCTION

In eukaryotes, the genome is organized within a nucleus enclosed by a nuclear envelope formed by lipids and proteins.[1] The bare nuclear DNA interacts with several different proteins packing the genome as chromatin fibers that are arranged at different levels of structural organization.[2] In the past decade, experiments using DNA–DNA ligation performed in human cell lines were able to capture information about the genome architecture by creating a 2D heat map named Hi-C.[3] Hi-C data support the organization of chromosomes in territories and identified that the overall genome organization could be described by two main compartments termed A and B.[3] Compartmentalization is often related to epigenetic information and gene expression. Compartment A correlates with regions of the genome containing highly expressed genes. On the other hand, compartment B is associated with heterochromatin, where the chromatin fiber organization is densely packed.[3−5] Hi-C maps reveal self-interacting partitions in the order of megabases named topologically associating domains (TADs).[6] Additionally, the loci contact frequency decays as a function of the genomic separation presenting a less intense signal for interactions far from the main diagonal.[3,7] Physical modeling helps to understand the compartmentalization pattern formation and associates the compartments with phase separation of chromatin type A–B. Additionally, chromatin dynamics relate the decay function with the polymer compaction, chromosome territory formation, and loop extrusion mechanism.[5,8−15] Initially, Hi-C maps experiments were performed in human interphase cell lines.[3] Recently, those efforts were expanded to several different organisms, cell lines, tissues, and cell cycle phases.[4,7,16−18] In these cases, obtaining high-quality maps is challenging for several reasons, such as the number of sequencing reads, sequencing coverage, the number of cells, cell phase synchronization, and the reference genome for the alignment. The poor sampled Hi-C maps present high sparse matrices where compartments located far from the main diagonal are difficult to observe. In this work, we propose chromatin

Received: May 11, 2021
Revised: July 12, 2021
Published: July 28, 2021



https://doi.org/10.1021/acs.jpcb.1c04174

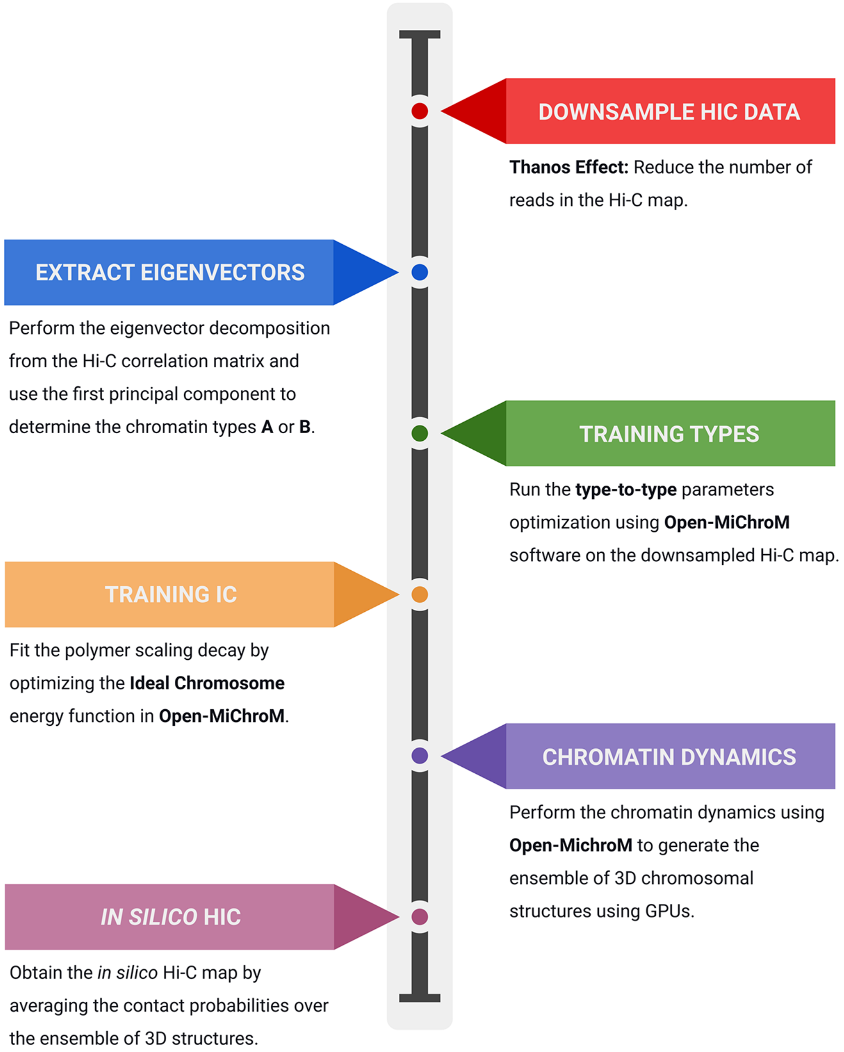

DOWNSAMPLE HIC DATA
Thanos Effect: Reduce the number of reads in the Hi-C map.
EXTRACT EIGENVECTORS
Perform the eigenvector decomposition from the Hi-C correlation matrix and use the first principal component to determine the chromatin types **A** or **B**.
TRAINING TYPES
Run the **type-to-type** parameters optimization using **Open-MiChroM** software on the downsampled Hi-C map.
TRAINING IC
Fit the polymer scaling decay by optimizing the **Ideal Chromosome** energy function in **Open-MiChroM**.
CHROMATIN DYNAMICS
Perform the chromatin dynamics using **Open-MichroM** to generate the ensemble of 3D chromosomal structures using GPUs.
IN SILICO HIC
Obtain the *in silico* Hi-C map by averaging the contact probabilities over the ensemble of 3D structures.

Figure 1. Workflow for performing the chromosome modeling based on poor-sampling Hi-C data. A high-quality Hi-C map is downsampled to test the model capability of using low-quality data. The compartments' annotations are obtained from the eigenvectors based on the downsampled maps. Using the PC1 annotations, the chromatin type A and B interactions are trained in the MiChroM energy function. Once the type-to-type interactions are trained, the IC parameters are optimized. Using the complete MiChroM model, the simulations are performed to generate the ensemble of chromosomal 3D structures. The contact probabilities of each loci pair along the trajectory are averaged to create the *in silico* 2D Hi-C map.

dynamics modeling to enhance the compartments' information far from the diagonal observed in sparse matrices. We used the experimental data from cell line GM12878.[7] We employed different techniques to partially downsample the Hi-C data to mimic poor sampling experiments. The simulations were performed using the Open-MiChroM platform, and new trained parameters were applied to the minimal chromatin model (MiChroM) energy function as described in Figure 1. The set of parameters for the chromatin type A or B interactions is extracted from the first component of the eigenvector of different levels of degradation of the experimental Hi-C. The simulations applying the trained parameters of the most downsampled experimental map (10% of the original data) present similar contact frequencies when compared with the complete experimental data set. The modeling results open a discussion on how simulation can increase the statistics and help fill in some Hi-C regions not captured by poor sampling experiments.

■ METHODS

Workflow for Enhancing the Compartmentalization Signal. Theoretical approaches to chromatin modeling have been successfully employed to generate the ensemble of chromosome structures that are consistent with Hi-C experiments.[19−21] To perform the 3D modeling, these models require high-quality experimental data to obtain information about the interaction of a loci pair. Methods using the maximum entropy approach such as MiChroM[8] (minimal chromatin model) use chromatin dynamics where the *in silico* map correlates $R = 0.96$ with the experimental Hi-C. The MiChroM energy function is built on two main assumptions: (1) the phase separation between chromatin types A and B and (2) the motor activity related to the polymer chain compaction, i.e., the ideal chromosome term.[8,9,22] (See the MiChroM model section for details.) Notwithstanding the accuracy of using the MiChroM energy function, the parameter optimization was trained using a high-quality Hi-C at 50 kb resolution. Here we propose to optimize the MiChroM energy

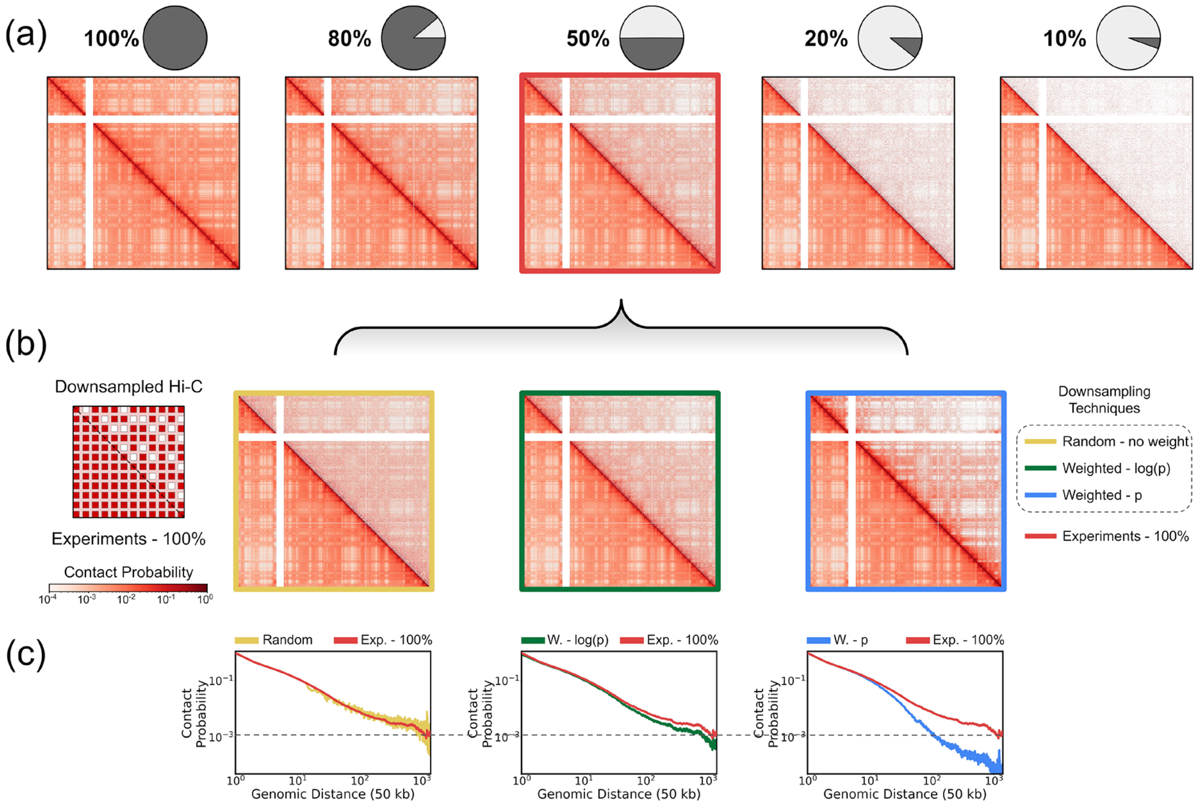


Figure 2. Downsampled Hi-C maps for chromosome 18 from GM12878 cell line. (a) Different levels of downsampling the Hi-C data (upper matrix) compared to the original Hi-C map (lower matrix). (b) Comparison of different methods for downsampling Hi-C maps; Random, Weighted-log(p) and Weighted-p, yellow, green, and blue borders, respectively. (c) Contact probability as a function of the genomic distance. The decay curve changes based on different downsampling techniques.

function in low-quality data to explore the limitation of this approach. Figure 1 presents the methodology workflow.

In the first step, to mimic low-quality data, the Hi-C maps[7] from the GM12878 cell line are downsampled. Three different algorithms were explored to reduce the number of counts in the Hi-C matrix. (See the Downsampling section.) In the sparsest matrix, Hi-C was reduced to 10% of the initial map. Next, the first principal component of the correlation matrices is extracted and associated with compartment annotations A and B with positive and negative values, respectively. Using chromatin types A and B, the optimization of the type-to-type energy parameters is performed using the Open-MiChroM software package. The optimization of the type-to-type parameters is related to the checkerboard patterns observed in the Hi-C. After the convergence of the type A—B interactions, training is performed in the Ideal Chromosome (IC) parameters. The IC energy function is associated with polymer compaction, where the function is calibrated to match the experimental scaling decay curve. The IC parameters optimization is the last step of the training process. Open-MiChroM uses the trained parameters to perform chromatin dynamics that generate an ensemble of 3D chromosomal structures over time. The average of the contact probability of each loci pair generates the *in silico* Hi-C map that is consistent with the experimental full Hi-C matrices.

Procedures for Downsampling the Hi-C Maps. The quality of a Hi-C map is associated with the number of counts/reads obtained from the high throughput sequencing used to generate the 2D heatmap after the alignment with the reference genome. For example, the Hi-C maps of the B-lymphoblastoid cell line (GM12878) obtained by Aiden Lab.[7] have around 15 billion reads. In this work, the experimental Hi-C map of chromosome 18 from the GM12878 was used for training the MiChroM parameters. Five different training of MiChroM energy function were performed in Hi-C maps with varying degrees of downsampling. Figure 2a shows the original (100%) Hi-C data compared to different levels of downsampling (80%, 50%, 20%, and 10%).

Additionally, it was employed three different methods for reducing the Hi-C data. Figure 2b presents the 50% downsampled Hi-C map generated by these three procedures. In the first approach, named "Random - no weight," all Hi-C contacts have the same probability of being removed (Figure 2b - yellow border). Weaker interactions located far from the main diagonal have the same chance of being removed as stronger interactions close to the diagonal. Figure 2c presents the contact probability decay as a function of the genomic distance. By employing the "Random" approach (yellow curve), the curve decays similarly to the 100% data (red curve). However, there is a significant increase in the noise generated by the lack of statistics in each genomic segment.

In the second approach called "weighted - p," the probability $p_{i,j}$ of maintaining the loci pair (i, j) interaction is selected based on the number of counts $C_{i,j}$ of that contact, and it is defined as follows:

$$p_{i,j} = \frac{C_{i,j}}{\sum_{i,j} C_{i,j}} \tag{1}$$

https://doi.org/10.1021/acs.jpcb.1c04174

Figure 2b (blue border) presents the downsampled Hi-C map employing this second technique. There is a significant decrease in the number of reads in regions far from the main diagonal. Figure 2c (right panel) show a significant deviation in the scaling curve when compared the data from the full Hi-C (red curve) with the downsampled data (blue curve). The third approach is called ″weighted - log(p),″ where the probability $p_{i,j}$ of maintaining the loci pair (i, j) interaction is defined as

$$p_{i,j} = \frac{\log(C_{i,j})}{\sum_{i,j} \log(C_{i,j})} \tag{2}$$

Figure 2b (green border) presents the downsampled Hi-C map generated by the third method. Figure 2c (middle panel) shows the contact probability as a function of genomic distance for the original map (red curve) and the downsampled data (green curve). There is a slight deviation in the scaling for regions far from the main diagonal, i.e., interactions between segments that are far in the genomic separation. This work will employ the ″weighted - log(p)″ approach for training the MiChroM energy function. The choice is based since the third approach present less noise than the first (″Random″), and minor deviation of the scaling decays in comparison with the second (″weighted - p″)

MiChroM Energy Function Based on the Maximum Entropy Approach. MiChroM physical potential considers the chromosome chain as a homopolymer with beads connected by springs. In order to create a polymer model to reproduce the experimental Hi-C contact frequencies, we assume a homopolymer potential with an additional term associated with a observable function $(\phi(r))$. A simulation in the canonical (NVT) ensemble gives the following value

$$\phi_{\mathrm{HP}} = \langle \phi(\vec{r}) \rangle_{U_{\mathrm{HP}}} = \frac{\int \phi(\vec{r}) \mathrm{e}^{-\beta U_{\mathrm{HP}}(\vec{r})} \,\mathrm{d}r}{\int e^{-\beta U_{\mathrm{HP}}(\vec{r})} \,\mathrm{d}r} = \int \phi(\vec{r}) \pi^{\mathrm{HP}}(\vec{r}) \,\mathrm{d}r \tag{3}$$

where $\vec{r}$ is the vector of positions in Cartesian space of all the loci in the chromosome, $\pi^{HP}(\vec{r})$ is the probability density for the homopolymer model and $\beta = 1/k_B T$. Employing the Maximum Entropy approach introduced by Jaynes,[23] we consider the probability density $\pi^{ME}(\vec{r})$ that reproduces the experimental values of a set of n observables $\phi_i(\vec{r})$. eq 4 show constraints that define the probability density:

$$\begin{aligned}
c_0 &= \int \pi^{\mathrm{ME}}(\vec{r}) \,\mathrm{d}r - 1 \\
c_1 &= \int U^{\mathrm{HP}}(\vec{r}) \,\pi^{\mathrm{ME}}(\vec{r}) - \frac{3}{2} N k_{\mathrm{B}} T \\
c_i^{\mathrm{data}} &= \int \phi(\vec{r}) \,\pi^{\mathrm{ME}}(\vec{r}) \,\mathrm{d}r - \phi_i^{\mathrm{exp}} \quad i = 1, ..., n
\end{aligned} \tag{4}$$

Each of the constraint equations must be equal to zero. The first constraint c_0 ensures that $\pi^{ME}(\vec{r})$ is normalized, i.e., the summation of the probabilities must be equal to one. The second constraint c_1 determines the average potential energy to be equal to the thermal energy $3Nk_B T/2$. The last set of constraints c_i^{DATA} where i can be related to n observables, and the equations ensure that the expectation values and the experimental values coincide. To determine the probability density $\pi^{ME}(\vec{r})$ we maximize the information entropy

$$S = -\int \pi^{\mathrm{ME}}(\vec{r}) \ln(\pi^{\mathrm{ME}}(\vec{r})) \,\mathrm{d}r \tag{5}$$

subject to the constraints. This approach is based on the fact that a constrained maximization of the entropy is equivalent to minimizing the amount of additional information built into the distribution other than the one contained in the constraints themselves. Using Lagrange multipliers we obtain the following condition

$$\frac{\partial S}{\partial \pi^{\mathrm{ME}}} - \lambda_0 \frac{\partial c_0}{\partial \pi^{\mathrm{ME}}} - \lambda_1 \frac{\partial c_1}{\partial \pi^{\mathrm{ME}}} - \sum_{i=1}^{n} \lambda_i{}^D \frac{\partial c_i^{\mathrm{data}}}{\partial \pi^{\mathrm{ME}}} = 0 \tag{6}$$

which leads to the probability distribution

$$\pi^{\mathrm{ME}}(\vec{r}) = \frac{\mathrm{e}^{-\lambda_1 U_{\mathrm{HP}}(\vec{r}) - \sum_{i=1}^{n} \lambda_i{}^D \phi_i(\vec{r})}}{\int \mathrm{d}r \quad \mathrm{e}^{-\lambda_1 U_{\mathrm{HP}}(\vec{r}) - \sum_{i=1}^{n} \lambda_i{}^D \phi_i(\vec{r})}} \tag{7}$$

Recognizing that λ_1 coincides with β, we can think of the maximum entropy probability distribution $\pi^{ME}(\vec{r})$ as the distribution sampled from the maximum entropy potential energy:

$$U_{\mathrm{ME}}(\vec{r}) = U_{\mathrm{HP}}(\vec{r}) + \frac{1}{\beta} \sum_{i=1}^{n} \lambda_i{}^D \phi_i(\vec{r}) \tag{8}$$

MiChroM energy function employed in this work has two assumptions. First, chromosome loci are classified into chromatin types A and B; each chromatin type contains specific interaction patterns with the other types. The type-to-type potential is related to the chromatin phase separation associated with the compartmentalization observed in the Hi-C maps. The second assumption is relative to the ideal chromosome which there is a gain/loss of $\gamma(d)$ effective free energy every time a pair of loci come into contact, and this potential depends on the genomic distance d. These 2 assumptions are replaced in c^{data} from eq 4. The set of constraint generated by all observables gives the following expressions:

$$\begin{aligned}
c_0 &= \int \pi^{\mathrm{MiChroM}}(\vec{r}) \,\mathrm{d}r - 1 \\
c_1 &= \int U^{\mathrm{HP}}(\vec{r}) \pi^{\mathrm{MiChroM}}(\vec{r}) - \frac{3}{2} N k_{\mathrm{B}} T \\
c_T^{kl} &= \int T_{kl}(\vec{r}) \pi^{\mathrm{MiChroM}}(\vec{r}) \,\mathrm{d}r - T_{kl}^{\mathrm{exp}} \\
&\quad \{ \ \forall \ k, l \in \text{types}: l \geq k \} \\
c_G^{\mathrm{d}} &= \int G_{\mathrm{d}}(\vec{r}) \pi^{\mathrm{MiChroM}}(\vec{r}) \,\mathrm{d}r - G_{\mathrm{d}}^{\mathrm{exp}} \quad 3 \leq d \leq d_{\mathrm{cutoff}}
\end{aligned} \tag{9}$$

These constraints define the MiChroM probability distribution, and the energy function is given by

$$\begin{aligned}
U_{\mathrm{MiChroM}}(\vec{r}) = U_{\mathrm{HP}}(\vec{r}) &+ \sum_{k \geq lk, l \in \text{types}} \alpha_{kl} \sum_{i \in \{\text{loci of type } k\} j \in \{\text{loci of type } l\}} f(r_{ij}) \\
&+ \sum_{d=3}^{d_{\mathrm{cutoff}}} \gamma(d) \sum_i f(r_{i,i+d})
\end{aligned} \tag{10}$$

The first term of $U_{\mathrm{MiChroM}}(\vec{r})$ is related to the generic homopolymer potentials. The second term represents the type-to-type interaction resulting from the constraint c_T^{kl}, and the last potential is called Ideal Chromosome that comes from the last constraint c^d_G. The function $f(r_{i,j})$ is the probability of cross-link[8,9] and can be written as

$$f(r_{i,j}) = \frac{1}{2}(1 + \tanh[\mu(r_c - r_{i,j}])) \tag{11}$$

where μ ad r_c are determined based on the experimental Hi-C maps. The function $f(r_{i,j})$ must return 1 when two beads are in contact (distance between the center of two beads is equal to 1, in reduced units σ), e.i., $f(1) = 1$. $f(r_{i,j})$ also must decreases monotonically with the distance and the minimum of the experimental probabilities must match with the next nearest neighbor, e.g, $f(2)$=min$\{P^{exp}_{i,i+2}\}$. The parameters adjusted for the Hi-C maps of GM12878 cell line[7] are μ = 3.22 and r_c = 1.78.[8] The Lagrange multipliers α and γ remain to be determined. The optimization procedure for training the parameters based on the experimental Hi-C maps is discussed below.

Parameters optimization. The optimized values of the Lagrange multipliers (in this work, α's and γ's) can be obtained via the minimization of the objective $\Gamma(\lambda)$ defined as

$$\Gamma(\lambda) = \ln\left(\frac{z(\lambda)}{z_0}\right) + \beta\sum_{i=1}^{n} \lambda_i \phi_i^{\text{exp}} \tag{12}$$

where $\beta = 1/k_BT$, with k_B as the Boltzmann constant and T as the temperature. $z(\lambda)$ and z_0 are the partition functions for the homopolymer with and without the maximum entropy correction, respectively. The exact solution of this function is a difficult statistical problem for an arbitrary value of λ, however, a possible solution is to use an iterative method to find an ensemble that satisfies the constraints. For this, the ratio between the partition function of the $\Gamma(\lambda)$ is simplified as a cumulant expansion:

$$\begin{aligned}\frac{z(\lambda)}{z_0} &= \frac{\int \mathrm{e}^{-\beta U_{\text{ME}}(\vec{r})}\,\mathrm{d}r}{\int \mathrm{e}^{-\beta U(\vec{r})}\,\mathrm{d}r} \\ &= \frac{\int \mathrm{e}^{-\beta U(\vec{r})}\mathrm{e}^{-\beta\sum_{i=1}^{n}\lambda_i\phi_i^{\text{exp}}}\,\mathrm{d}r}{\int \mathrm{e}^{-\beta U(\vec{r})}\,\mathrm{d}r} = \left\langle \mathrm{e}^{-\beta\sum_{i=1}^{n}\lambda_i\phi_i^{\text{exp}}}\right\rangle \\ &= \mathrm{e}^{\sum_{n=1}^{\infty}\frac{(-\beta)^n}{n!}\langle\langle(\sum_{i=1}\lambda_i\phi_i^{\text{exp}})^n\rangle\rangle}\end{aligned} \tag{13}$$

The first two terms (n = 1 and n = 2) from the cumulant expansion can be described as

$$\begin{aligned} c1 &= -\beta\sum_{i=1} \lambda_i\langle\phi_i^{\text{exp}}(\vec{r})\rangle \\ c2 &= \frac{\beta^2}{2}\sum_i\sum_j \lambda_i\lambda_j[\langle f_i f_j\rangle - \langle f_i\rangle\langle f_j\rangle]\end{aligned} \tag{14}$$

If we keep only these two terms of the expansion and insert the eq 13 in eq 12, we have an approximate expression

$$\Gamma(\lambda) = \frac{\beta^2}{2}\lambda^{\text{T}}B\lambda - \beta[\langle f_i\rangle - f_i^{\text{exp}}]^{\text{T}}\lambda \tag{15}$$

where B is a Hermitian matrix with elements B_{ij} = $\langle f_i f_j\rangle - \langle f_i\rangle\langle f_j\rangle$.

Following the approaches described in previous works,[24,25] eq 15 has its extreme value of $\Gamma(\lambda)$ at

$$\lambda = \frac{1}{\beta}B^{-1}[\langle f_i\rangle - f_i^{\text{exp}}]^{\text{T}} \tag{16}$$

By employing the cumulant approximation only to a generic homopolymeric model, there is no guarantee that λ solutions would reproduce the experimental data. However, eq 16 is only an approximate solution of λ. If $\langle f_i\rangle$ is not equal $\langle f_i^{exp}\rangle$, we use an iterative procedure to find more accurate values that eventually would converge to reproduce experimental measurements.[26,27] The iteration algorithm is as follow:[9]

I - Perform simulations with the potential energy $U_{ME}(\vec{r})$ to estimate the ensemble averages $\langle f_i\rangle$ and the matrix B.

II - Check convergence of the iteration by calculating the percentage of error defined as $\sum_i |\langle f_i\rangle - \langle f_i^{\text{exp}}\rangle| / \sum_i f_i^{\text{exp}}$, where $|\cdot|$ correspond to absolute values.

III - If the error is less than a tolerance value *tol*, the iteration has converged and we stop the simulations. Otherwise, we update λ using the expression $\lambda^{l+1} = \lambda^l + \delta(\frac{1}{\beta}B^{-1}(\lambda^l))([\langle f_i(\lambda^l)\rangle - f_i^{\text{exp}}]^{\text{T}})$, where λ^l are the Lagrangian multipliers used in step I and $\delta \in (0,1)$ is dampening parameter. Now with the updated λ^{l+1} values, we go back to step I and restart the iteration.

Simulations Details. Chromatin dynamics simulations were performed using Open-MiChroM software package that uses OpenMM API.[28] The chromosome 18 model consists of 1561 beads at 50 Kb resolution. The input information for training Open-MiChroM simulations uses chromatin types annotations for determining the polymer sequence of beads A or B. The types annotation was obtained by the first principal component of the correlation matrix extracted from the Hi-C map. Open-MiChroM receives the parameters α for Types and γ for IC in a configuration text file. A collapse run of 10^6 steps is performed at a high temperature to randomize the initial configuration of the sampling simulation. At this collapse run, a harmonic potential term is used to accelerate the process. The harmonic potential for collapsing the polymer is removed for the trainig and long sampling simulations. The training parameters simulations were carried out for 20 replicas over 1×10^7 steps for each iteration with a trajectory snapshot saved at every 10^3 steps. Trajectories are stored in a binary format .cndb (Compact Nucleome Data Bank file). The training parameters are calculated based on all stored 3D structures. Once the parameters optimization is converged, more extended production simulations were performed for 30 replicas running for 1×10^8 steps and with a 3D structure frame saved every 10^3 steps, leading to a total of 3×10^6 structures. Contact probabilities of each loci pair are averaged over the ensemble of 3D structures to generate the *in silico* Hi-C map. Parameters for the homopolymer potential U_{HP} follow the same values described in Open-MiChroM study.[28] Hi-C maps are plotted using the juicebox software tool.[29] The 3D structure representation of the chromosome was made using Chimera software.[30] Open-MiChroM package, trajectory data, analysis scripts, simulation tutorials, .cndb file converter to .pdb or .gro files are available at the Nucleome Data Bank (NDB) server[20] (https://ndb.rice.edu).

■ RESULTS AND DISCUSSION

Training MiChroM Types Parameters on Downsampled Hi-C maps. MiChroM energy function has two main terms: the type-to-type and Ideal Chromosome

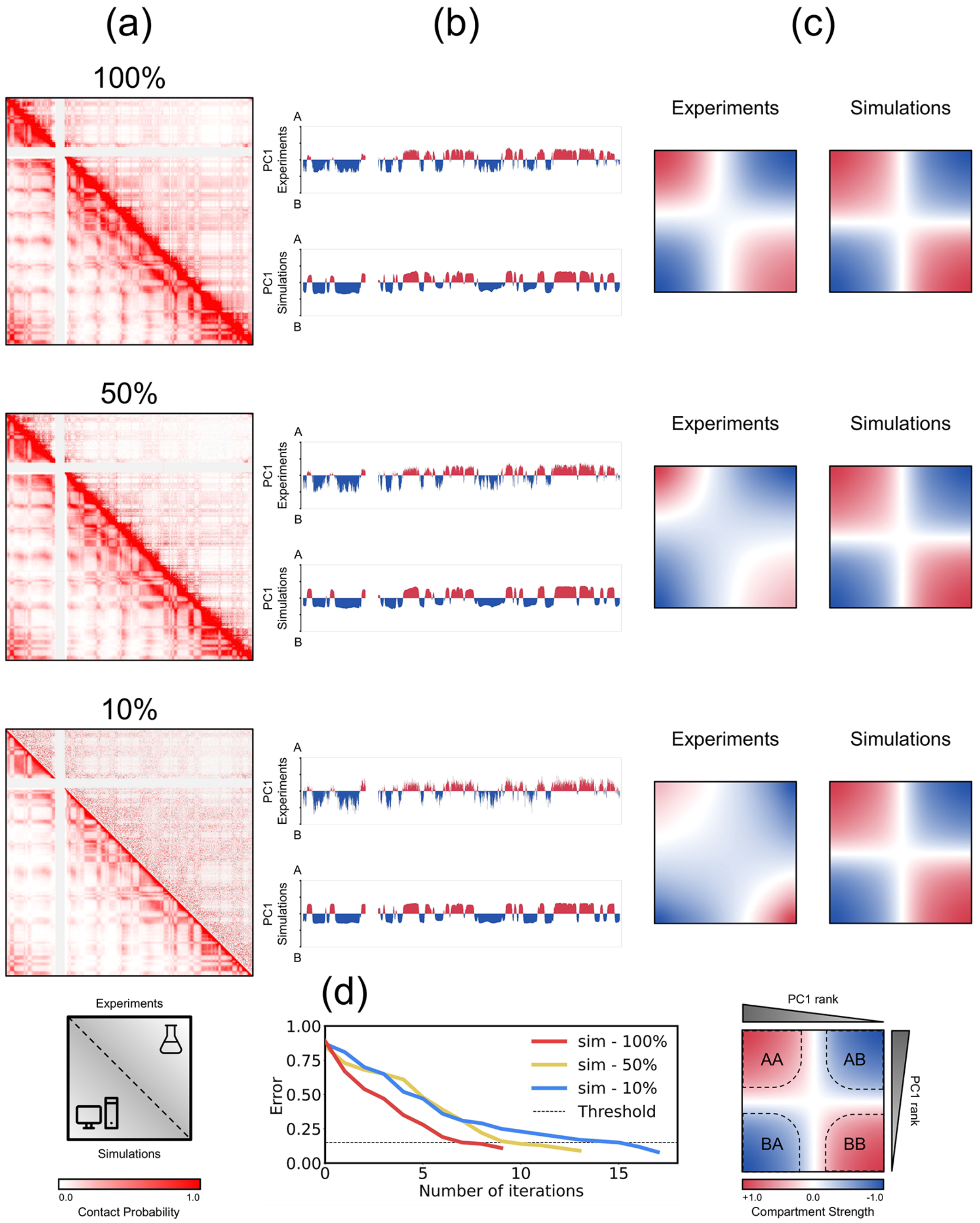


Figure 3. (a) *In silico* Hi-C maps generated using only the type-to-type trained parameters based on the downsampled Hi-C maps. (b) The first principal component obtained from the correlation matrix from the *in silico* Hi-C and the experimental downsampled maps. (c) Saddle plots comparing the compartmentalization signal from simulations and downsampled experiments. (d) Error calculation during the optimization process in different iteration steps for different levels of downsampling.

interactions. These potentials have parameters to be optimized based on experimental Hi-C maps (see Methods for details). As mentioned in the simulation workflow, the first step explored is the optimization of the type-to-type interactions. The minimization procedure is performed to find optimum parameters for the Lagrange multipliers α from the eq 10. The parameter α is related to the phase separation between the chromatin types A and B. Here, we employed a model with three different chromatin beads: Type A, Type B, and NA (nonspecific type). The sequence of compartmental types for each chromosome is extracted from the first principal component (PC1) of the eigenvector decomposition from the correlation matrix.[3] By convention, positive values of the PC1 denominated to the chromatin type A and negative values to type B. The same procedure for obtaining the chromatin sequences is applied for all downsampled Hi-C maps and used as input for simulations. The ”Types” energy function training includes the homopolymer chain potential and the type-to-type interactions defined as

$$U_{\mathrm{MiChroM}}^{\mathrm{types}}(\vec{r}) = U_{\mathrm{HP}}(\vec{r}) + \sum_{\substack{k \geq l \\ k,l \in \mathrm{types}}} \alpha_{kl} \sum_{\substack{i \in \{\text{loci of type } k\} \\ j \in \{\text{loci of type } l\}}} f(r_{ij}) \quad (17)$$

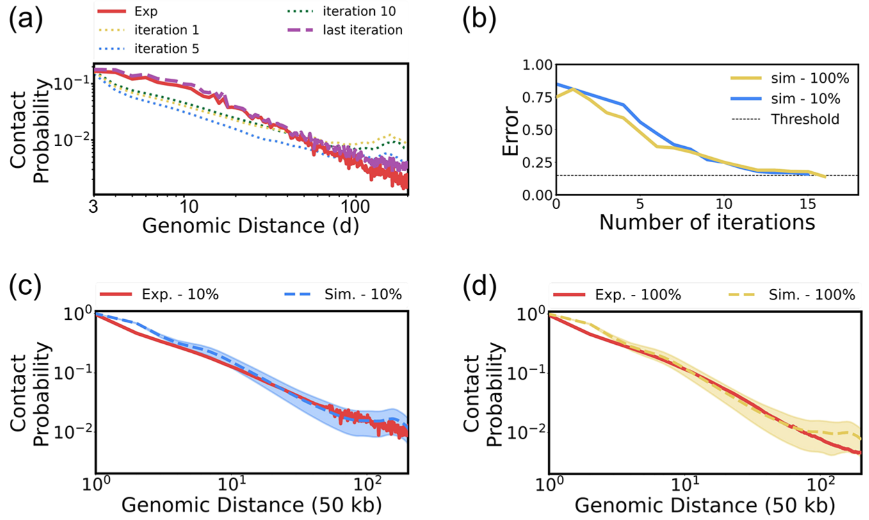


Figure 4. (a) Contact probability as a function of the genomic distance for different iteration steps of the ideal chromosome parameters training. (b) Error value as a function of the number of iterations of the IC parameter optimization. (c and d) Contact probability as a function of the genomic distance for different degrees of downsampling the experimental Hi-C maps, 10 and 100%, respectively. Each shaded area presents the standard deviation for a genomic distance.

where the $\alpha_{k,l}$ represents interactions between the chromatin type k and l, i.e., A−A, A−B, A−NA, B−NA, and NA−NA. Multiple iterations were performed until the error (described in step II of the optimization section) drops below 10%. At this point, we consider that the $\alpha_{k,l}$ values are converged. Figure 3a presents the *in silico* Hi-C maps after the optimization of the type-to-type parameters. Even though the simulations were performed using only the "types" potential, there is compartmentalization in regions far from the main diagonal in the *in silico* Hi-C maps.

Figure 3b shows the PC1 signal for each locus along the whole chromosome chain for both experiments and simulations and different downsampling levels. The eigenvectors extracted from the *in silico* maps present a stronger signal in comparison with the experimental data. Interestingly, the simulations used the experimental PC1 as input for determining chromatin beads A and B. Even if the experiments provide a weak signal for PC1 that is strong enough to distinguish between chromatin types A and B, the results suggest that chromatin dynamics employing the MiChroM energy function amplify the PC1 signal and, consequently, the compartmentalization. Figure 3c presents the saddle plots for both experiments and simulation for different degrees of downsampling. The saddle plot is a heat map that highlights the A/B compartments' interaction strength. The strong interactions of the same chromatin types AA and BB are located in the matrix main diagonal corners. On the other hand, weaker intercompartmental-type interactions AB are situated on the corner of the antidiagonal.[31,32] The data generated from simulations do not present significant differences in the saddle plots. On the other hand, the data extracted from the downsampled Hi-C maps show a significant deviation from the original data. This observation suggests that downsampling the Hi-C data leads to a weaker compartmentalization signal, i.e., the eigenvector components still distinguish between positive and negative values but the amplitude decreases. The long production simulations can sample a more extensive variety of chromosomal structures, which results in a better sampled Hi-C map when averaging the contact probabilities over multiple different structures. PC1 shows a more robust signal even when using only 10% of the original data for the training. However, to perform parameter optimization on poorly sampled data, more iterations are needed to converge the parameters. Figure 3d shows the error value used in training as a function of the number of iterations. The error threshold is set at 10% (dashed gray line). The training simulation using the original Hi-C 100% (red curve) reaches the threshold with less iterations than using the 50% and 10% downsampled maps, yellow and blue curves, respectively.

Training Ideal Chromosomes on Downsampled Hi-C Maps. For parameter α on which the type-to-type interactions are trained, the MiChroM potential is still not fully optimized. As mentioned before, α is associated with the compartmentalization and phase separation of chromatin types A and B. Although there is agreement between simulations and experimental maps, the polymer scaling presents deviations. The contact probability curve decay is associated with the polymer chain compaction due to different motor activities in the chromatin, which is considered to be an effective potential named ideal chromosome (IC). The IC parameters γ in the MiChroM energy function need to be calibrated. The procedure employed in γ minimization is the same applied for training α parameters but using the full MiChroM potential described in eq 10, where α values are already optimized. The IC parameters were set to $\gamma(d) = 0$ in the first iteration for all genomic distances d. The experimental values $(f(d)^{\exp})$ are the average value over all probabilities for a given genomic distance d in the experimental Hi-C maps. In the case of $d = 3$, $f(3)^{\exp} = \frac{1}{N}\sum_{i=3}^{N-3} P_{(i,i+3)}$, where $P_{(i,i+3)}$ is the contact probability between the loci i and $i + 3$ and N is number of beads in the chromosome chain. Figure 4a presents the contact probability as a function of the genomic distance for different training iterations. Iteration 1 (dotted yellow curve) shows a more significant deviation of the decay curve than the experimental data extract from the original Hi-C (solid red curve). The method goes over different iteration along the minimization process following the optimization protocol presented in the Methods section. In the IC training, there

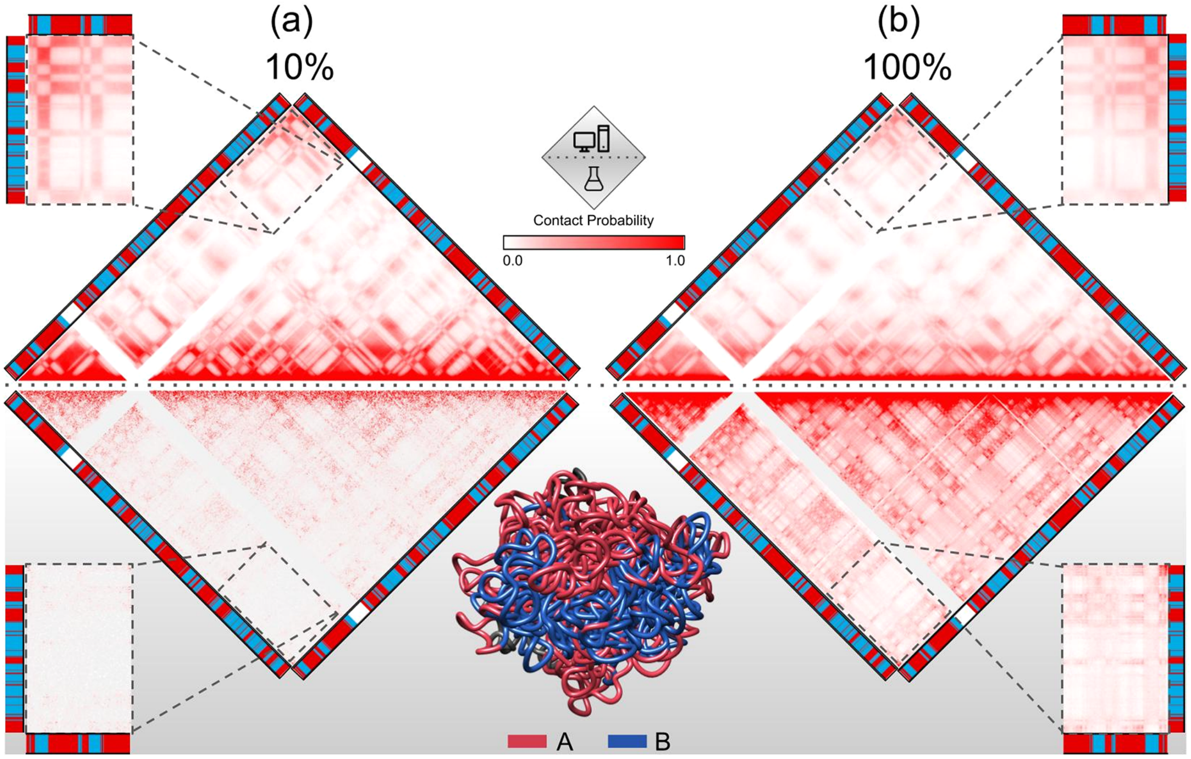


Figure 5. *In silico* Hi-C maps generated by employing the complete trained MiChroM energy function. (a and b) Comparison between *in silico* (top) Hi-C with the downsampled maps (bottom), 10 and 100%, respectively. The corner highlights regions of the Hi-C maps far from the main diagonal. The *in silico* maps present compartmentalization signals even when using only 10% of the original data to optimize the MiChroM parameters. At the bottom is presented a representative structure colored by chromatin type annotations A and B, red and blue, respectively.

are hundreds of Lagrange multipliers to be optimized. The range of *d* goes from 3 to 500. In principle, the range limit of *d* is the polymer chain length. Here, we used the same numbers presented in the first MiChroM training. In the last iteration, the simulation curve presents a smaller value of the error calculation. Figure 4b shows the error value for different iteration steps. The training process shows similar learning rates to reach the threshold value (error below 10%) for using the full Hi-C map (yellow curve) and the downsampled 10% Hi-C (blue curve). Figure 4c,d shows the decay curve of the simulations after the training compared to experiments 10 and 100%, respectively. There is agreement between simulation and experiments, which suggests that the ensemble of 3D structures generated by the MiChroM energy function has the polymer compaction close to what is expended from experiments. Figure 5 shows the converged *in silico* Hi-C maps trained on the basis of downsampled experimental maps. Simulations were performed by employing the complete trained MiChroM energy function that includes the type-to-type and IC potentials. The chromatin dynamics simulations performed with Open-MiChroM[28] generate an ensemble of chromosomal 3D structures. The loci contacts formed in the 3D structures are averaged over the simulated trajectory and mapped into the 2D *in silico* Hi-C. A total of 30 simulation replicas were carried out with different initial loci positions and velocities for sampling different configurations of the energy landscape. There are compartments observed in the simulated maps that are not seen in the 10% downsampled map, especially for regions far from the main diagonal. This suggests that the ensemble of 3D structures generated by the MiChroM energy function employing only two terms related to the chromatin phase separation and motor activity can enhance the compartmentalization. The enrichment of the compartments is related to the better statistics obtained from the chromatin dynamics simulations even when the parameter optimization comes from the downsampled Hi-C data. The zoom-in region of the Hi-C maps highlights these signal differences. Figure 5 shows a representative structure of chromosome 18 of the GM12878 cell line obtained from the trajectories generated by Open-MiChroM based on the full experimental Hi-C map. It is possible to observe a phase separation between chromatin types A and B. Figure 6a shows the knots calculation comparing simulations using the full MiChroM potential with a homopolymer and MiChroM applying only the type-to-type interactions.

Simulations using the complete training MiChroM energy function show fewer knots along the trajectory compared to the homopolymer simulation. As reported in previous studies,[8,28] the motor activity associated with the IC potential leads the polymer to equilibrate by favoring short local interactions. It is worth mentioning that the polymer physical model allows chain crossing to consider topoisomerase II activity. Interestingly, simulations employing only the type-to-type interactions (blue curve) also present fewer knots than the homopolymer. It suggests that the phase separation between chromatin types A and B also allows the polymer to become less entangled. Figure 6b presents Pearson's correlation as a function of the genomic distance for different simulations with their respective 3D structure. The optimization of the "types" potential significantly improves the compartment sampling compared to the homopolymer simulation baseline. The optimization of the type-to-type parameters reduces the polymer energetic frustration where

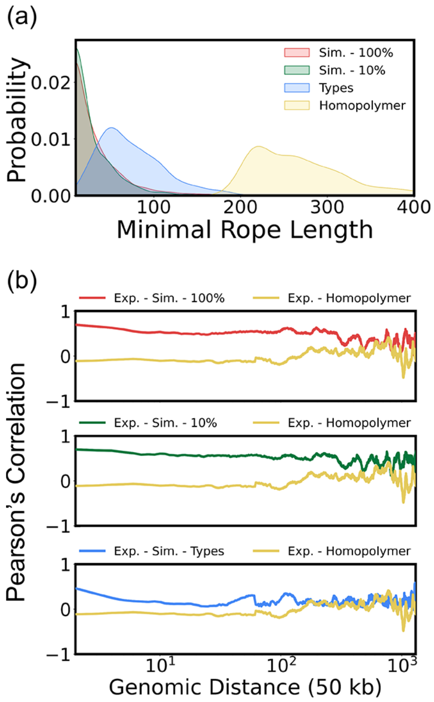


Figure 6. (a) Probability distributions of the minimal rope length ratio that is a topological invariant. The calculation was performed over a trajectory for different energy functions. The red and green curves are the distributions using the complete training MiChroM parameters with 100 and 10% of the original data, respectively. The simulation using only the type-to-type interactions is presented in blue, and the homopolymer results are presented in the yellow distribution. (b) Pearson's correlation as a function of the genomic distance for different training sets. The red and green curves are generated using the complete MiChroM training on 100 and 10% of the original Hi-C map, respectively. The blue curve shows the results using only the type-to-type interactions. The yellow curve presents the simulations of a homopolymer that is used as a baseline for comparison.

there is an increase in the number of favorable interactions when phase separation occurs.

CONCLUSIONS

Several computational studies perform chromosome modeling employing physical models with parameters determined utilizing experimental data extracted from Hi-C maps.[33] These approaches apply different metrics for converting the number of counts presented in the Hi-C to distances in the structures.[34−37] The 2D Hi-C heat maps are generated on the basis of an ensemble of cells, and the number of reads or the intensity of the signal in these maps is associated with a unique value for the whole population of cells. Once generating the 3D modeling, most of these approaches lack a consideration of the variability of chromosomal structures. Additionally, the multidimensional reduction techniques applied in the distance matrices lack more details about the physical properties of the polymeric chain. On the other hand, physical modeling approaches based on chromatin dynamics consider the heterogeneity of the structures and shed light on the underlying mechanisms of the loci phase separation and motor activity such as extruders.[4] Notwithstanding the good agreement between physical models of Hi-C maps and experiments, a well-sampled experimental Hi-C map is required to optimize the loci interaction parameters. Here we explore chromatin dynamics and modeling using poorly sampled Hi-C maps. By training the MiChroM energy function parameters using only the minimal information on locus compartments A and B, the simulations enhanced the compartmentalization signal of regions in the Hi-C map far from the main diagonal, i.e., spatial contacts between regions far from each other in sequence. The information extracted from the first-principles component of the correlation matrices from downsampled Hi-C seems to be enough for a physical polymer model to characterize the loci interactions related to phase separation and compartment formation. The prediction capability of these models allows for optimizing parameters even for downsampled Hi-C maps using only 10% of the original data, and MiChroM chromatin dynamics generates an ensemble of 3D structures consistent with the complete experimental Hi-C map. This prediction power of the MiChroM energy function provides the needed tool for modeling chromosomal 3D structures based on poorly sampled Hi-C data. MiChroM modeling allows us to explore the 3D genome organization of different organisms, cell lines, and cell phases that often do not produce high-quality Hi-C maps.

AUTHOR INFORMATION

Corresponding Authors

Vinícius G. Contessoto − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77251, United States; Instituto de Biociências, Letras e Ciências Exatas, UNESP - Univ. Estadual Paulista, Departamento de Física, São José do Rio Preto, SP, Brazil;* orcid.org/0000-0002-1891-9563; Email: vinicius.contessoto@rice.edu

José N. Onuchic − *Center for Theoretical Biological Physics, Department of Physics & Astronomy, Department of Chemistry, and Department of Biosciences, Rice University, Houston, Texas 77251, United States;* orcid.org/0000-0002-9448-0388; Email: jonuchic@rice.edu

Authors

Antonio B. Oliveira Junior − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77251, United States*

Cynthia Perez Estrada − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77251, United States; The Center for Genome Architecture, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States*

Erez Lieberman Aiden − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77251, United States; The Center for Genome Architecture, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c04174

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Michele Di Pierro, Sumitabha Brahmachari, Ryan Cheng, and Peter Wolynes for important discussions and suggestions during the development of this work. This research was supported by the Center for Theoretical Biological Physics sponsored by the NSF (grants PHY-2019745 and CHE-1614101) and by the Welch Foundation (grant C-1792). J.N.O. is a Cancer Prevention and Research Institute of Texas (CPRIT) Scholar in Cancer Research. A.B.O.J. acknowledges the Robert A. Welch Postdoctoral Fellow Program. V.G.C. is also funded by FAPESP (São Paulo State Research Foundation and Higher Education Personnel) and CAPES (Higher Education Personnel Improvement Coordination) grants 2016/13998-8 and 2017/09662-7. In addition to CTBP support, E.L.A. was also funded by the Welch Foundation (Q-1866), a USDA Agriculture and Food Research Initiative Grant (2017-05741), and an NIH Encyclopedia of DNA Elements Mapping Center Award (UM1HG009375). We also thank AMD for some of the computational resources used in this research.

■ REFERENCES

(1) Van Driel, R.; Fransz, P. F.; Verschure, P. J. The eukaryotic genome: a system regulated at different hierarchical levels. *J. Cell Sci.* **2003**, *116*, 4067−4075.

(2) Ridgway, P.; Almouzni, G. Chromatin assembly and organization. *J. Cell Sci.* **2001**, *114*, 2711−2712.

(3) Lieberman-Aiden, E.; van Berkum, N. L.; Williams, L.; Imakaev, M.; Ragoczy, T.; Telling, A.; Amit, I.; Lajoie, B. R.; Sabo, P. J.; Dorschner, M. O.; Sandstrom, R.; Bernstein, B.; Bender, M. A.; Groudine, M.; Gnirke, A.; Stamatoyannopoulos, J.; Mirny, L. A.; Lander, E. S.; Dekker, J.; et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **2009**, *326*, 289−293.

(4) Cheng, R. R.; Contessoto, V. G.; Lieberman Aiden, E.; Wolynes, P. G.; Di Pierro, M.; Onuchic, J. N. Exploring chromosomal structural heterogeneity across multiple cell lines. *eLife* **2020**, *9*, No. e60312.

(5) Di Pierro, M.; Cheng, R. R.; Aiden, E. L.; Wolynes, P. G.; Onuchic, J. N. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 12126−12131.

(6) Dixon, J. R.; Selvaraj, S.; Yue, F.; Kim, A.; Li, Y.; Shen, Y.; Hu, M.; Liu, J. S.; Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **2012**, *485*, 376−380.

(7) Rao, S. S.; Huntley, M. H.; Durand, N. C.; Stamenova, E. K.; Bochkov, I. D.; Robinson, J. T.; Sanborn, A. L.; Machol, I.; Omer, A. D.; Lander, E. S.; et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **2014**, *159*, 1665−1680.

(8) Di Pierro, M.; Zhang, B.; Aiden, E. L.; Wolynes, P. G.; Onuchic, J. N. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 12168−12173.

(9) Zhang, B.; Wolynes, P. G. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6062−6067.

(10) Brackley, C. A.; Johnson, J.; Michieletto, D.; Morozov, A. N.; Nicodemi, M.; Cook, P. R.; Marenduzzo, D. Nonequilibrium chromosome looping via molecular slip links. *Phys. Rev. Lett.* **2017**, *119*, 138101.

(11) MacPherson, Q.; Beltran, B.; Spakowitz, A. J. Bottom−up modeling of chromatin segregation due to epigenetic modifications. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 12739−12744.

(12) Mirny, L. A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **2011**, *19*, 37−51.

(13) Fudenberg, G.; Imakaev, M.; Lu, C.; Goloborodko, A.; Abdennur, N.; Mirny, L. A. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **2016**, *15*, 2038−2049.

(14) Krepel, D.; Cheng, R. R.; Di Pierro, M.; Onuchic, J. N. Deciphering the structure of the condensin protein complex. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 11911−11916.

(15) Krepel, D.; Davtyan, A.; Schafer, N. P.; Wolynes, P. G.; Onuchic, J. N. Braiding topology and the energy landscape of chromosome organization proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1468−1477.

(16) Gibcus, J. H.; Samejima, K.; Goloborodko, A.; Samejima, I.; Naumova, N.; Nuebler, J.; Kanemaki, M. T.; Xie, L.; Paulson, J. R.; Earnshaw, W. C. A pathway for mitotic chromosome formation. *Science* **2018**, *359*, eaao6135.

(17) Dudchenko, O.; Batra, S. S.; Omer, A. D.; Nyquist, S. K.; Hoeger, M.; Durand, N. C.; Shamim, M. S.; Machol, I.; Lander, E. S.; Aiden, A. P.; et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **2017**, *356*, 92−95.

(18) Hoencamp, C.; Dudchenko, O.; Elbatsh, A. M.; Brahmachari, S.; Raaijmakers, J. A.; van Schaik, T.; Cacciatore, Á. S.; Contessoto, V. G.; van Heesbeen, R. G.; van den Broek, B.; et al. 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science* **2021**, *372*, 984−989.

(19) Zhang, B.; Wolynes, P. G. Genomic Energy Landscapes. *Biophys. J.* **2017**, *112*, 427−433.

(20) Contessoto, V. G.; Cheng, R. R.; Hajitaheri, A.; Dodero-Rojas, E.; Mello, M. F.; Lieberman-Aiden, E.; Wolynes, P. G.; Di Pierro, M.; Onuchic, J. N. The Nucleome Data Bank: web-based resources to simulate and analyze the three-dimensional genome. *Nucleic Acids Res.* **2021**, *49*, D172.

(21) Di Pierro, M.; Potoyan, D. A.; Wolynes, P. G.; Onuchic, J. N. Anomalous diffusion, spatial coherence, and viscoelasticity from the energy landscape of human chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 7753−7758.

(22) Zhang, B.; Wolynes, P. G. Shape Transitions and Chiral Symmetry Breaking in the Energy Landscape of the Mitotic Chromosome. *Phys. Rev. Lett.* **2016**, *116*, 248101.

(23) Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620−630.

(24) Agmon, N.; Alhassid, Y.; Levine, R. D. An algorithm for finding the distribution of maximal entropy. *J. Comput. Phys.* **1979**, *30*, 250−258.

(25) Mead, L. R.; Papanicolaou, N. Maximum entropy in the problem of moments. *J. Math. Phys.* **1984**, *25*, 2404−2417.

(26) Pitera, J. W.; Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445−3451.

(27) Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **2013**, *138*, 084107.

(28) Oliveira, A. B.; Contessoto, V. G.; Mello, M. F.; Onuchic, J. N. A Scalable Computational Approach for Simulating Complexes of Multiple Chromosomes. *J. Mol. Biol.* **2021**, *433*, 166700.

(29) Durand, N. C.; Robinson, J. T.; Shamim, M. S.; Machol, I.; Mesirov, J. P.; Lander, E. S.; Aiden, E. L. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **2016**, *3*, 99−101.

(30) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(31) Flyamer, I. M.; Gassler, J.; Imakaev, M.; Brandão, H. B.; Ulianov, S. V.; Abdennur, N.; Razin, S. V.; Mirny, L. A.; Tachibana-Konwalski, K. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **2017**, *544*, 110−114.

(32) Abramo, K.; Valton, A.-L.; Venev, S. V.; Ozadam, H.; Fox, A. N.; Dekker, J. A chromosome folding intermediate at the condensin-to-cohesin transition during telophase. *Nat. Cell Biol.* **2019**, *21*, 1393−1402.

(33) Jerkovic, I.; Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 511.

(34) Zhang, Z.; Li, G.; Toh, K.-C.; Sung, W.-K. 3D chromosome modeling with semi-definite programming and Hi-C data. *J. Comput. Biol.* **2013**, *20*, 831−846.

(35) Lesne, A.; Riposo, J.; Roger, P.; Cournac, A.; Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nat. Methods* **2014**, *11*, 1141.

(36) Zhu, G.; Deng, W.; Hu, H.; Ma, R.; Zhang, S.; Yang, J.; Peng, J.; Kaplan, T.; Zeng, J. Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic Acids Res.* **2018**, *46*, No. e50.

(37) Oluwadare, O.; Zhang, Y.; Cheng, J. A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. *BMC Genomics* **2018**, *19*, 161.