# Transferable model for chromosome architecture

Michele Di Pierro[a,1,2], Bin Zhang[a,1], Erez Lieberman Aiden[a,b], Peter G. Wolynes[a,c,d,e,2], and José N. Onuchic[a,c,d,e,2]

[a]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005; [b]Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030; [c]Department of Physics & Astronomy, Rice University, Houston, TX 77005; [d]Department of Chemistry, Rice University, Houston, TX 77005; and [e]Department of Biosciences, Rice University, Houston, TX 77005

In vivo, the human genome folds into a characteristic ensemble of 3D structures. The mechanism driving the folding process remains unknown. We report a theoretical model for chromatin (Minimal Chromatin Model) that explains the folding of interphase chromosomes and generates chromosome conformations consistent with experimental data. The energy landscape of the model was derived by using the maximum entropy principle and relies on two experimentally derived inputs: a classification of loci into chromatin types and a catalog of the positions of chromatin loops. First, we trained our energy function using the Hi-C contact map of chromosome 10 from human GM12878 lymphoblastoid cells. Then, we used the model to perform molecular dynamics simulations producing an ensemble of 3D structures for all GM12878 autosomes. Finally, we used these 3D structures to generate contact maps. We found that simulated contact maps closely agree with experimental results for all GM12878 autosomes. The ensemble of structures resulting from these simulations exhibited unknotted chromosomes, phase separation of chromatin types, and a tendency for open chromatin to lie at the periphery of chromosome territories.

human genome | genome architecture | maximum entropy | molecular dynamics | Hi-C

Chromatin comprises a highly flexible polymer composed of nucleosomes, DNA wrapped around histone proteins, connected to one another by a linker region of 20–50 bp. Hundreds of associated structural and regulatory proteins interact with the genetic material, coordinating the way chromatin folds to fit inside the nucleus of eukaryotic cells.

The resulting ensemble of partially organized structures brings sections of DNA separated by a great genomic distance into close spatial proximity, and plays an important role in controlling gene transcription (1, 2). Although some of the features of this ensemble can be explained using simple polymer physics (3–6), there is now ample evidence that specific biochemical interactions play a crucial role (7–10). Understanding the interplay between biochemistry, genome architecture, and transcriptional regulation is a major outstanding challenge.

For over two decades, molecular biology techniques that combine chromatin fragmentation and proximity ligation have given us quantitative information about how chromatin is organized in vivo (5, 11–13). In recent years, Hi-C experiments have made it possible to measure the frequency of contact between all pairs of genomic loci using a single experiment.

Here, we explore a physical model by which local interactions between genomic loci can lead to the conformations of human chromosomes in interphase. Specifically, we propose a theoretical energy landscape model for chromatin folding, designated the Minimal Chromatin Model (MiChroM), which uses the maximum entropy principle (14, 15) in combination with a minimal number of assumptions to model the structural consequences of the aforementioned biochemical interactions. Importantly, MiChroM can be used to model biochemical interactions even though the identity of the interacting biomolecules is unknown. MiChroM suggests a mechanism that is sufficient to explain chromatin organization and can be used to generate ensembles of 3D structures describing whole genomes. As we will show, contact maps generated in silico from these ensembles reproduce in detail the maps from Hi-C.

The first assumption made in MiChroM is that the genome is partitioned into intervals of a handful of types, such that each type of interval is marked by characteristic histone modifications and interacts with a characteristic combination of nuclear proteins. As a result, when two segments of chromatin come into contact, the effective free energy change due to this contact depends, to first order, on the chromatin type of each segment [also Jost et al. (16)].

This assumption is supported by both biochemical and structural data. For instance, five distinct types of chromatin have been found in Drosophila cells based on the binding patterns of nuclear proteins (17). Further, analysis of original Hi-C maps (5) suggested that human chromatin is partitioned into two compartments, A and B, each associated with distinct long-range contact patterns. More recently, Rao et al. (9) used kilobase-resolution Hi-C experiments to show that the human genome can be further partitioned into six subcompartments (A1 and A2 and B1, B2, B3, and B4), each correlated with particular histone marks and associated with a particular pattern of long-range contacts. A similar partitioning of the genome was observed also in the mouse (9, 18) and Drosophila (19, 20). Both the boundaries of these genomic intervals and their chromatin types may change along with changes in cell state (9). The close association between interval types and long-range contact patterns suggests that intervals of the same type segregate together in the nucleus.

The second assumption made in MiChroM is that certain pairs of genomic "anchor" loci tend to form loops. This tendency is encoded in the model as a change in the effective free energy of a chromatin configuration when the two anchor loci are in contact.

## Significance

Chromatin consists of DNA and hundreds of proteins that interact with the genetic material. In vivo, chromatin folds into non-random structures. The physical mechanism leading to these characteristic conformations, however, remains poorly understood. Here, we introduce a model that generates chromosome conformations by using the idea that chromatin can be subdivided into types based on its biochemical interactions. Chromatin types, which are distinct from DNA sequence, are partially epigenetically controlled and change during cell differentiation, thus constituting a link between epigenetics, chromosomal organization, and cell development. The degree of accuracy achieved by this model supports the viability of the proposed physical mechanism of chromatin folding and makes the computational model a powerful tool for future investigations.
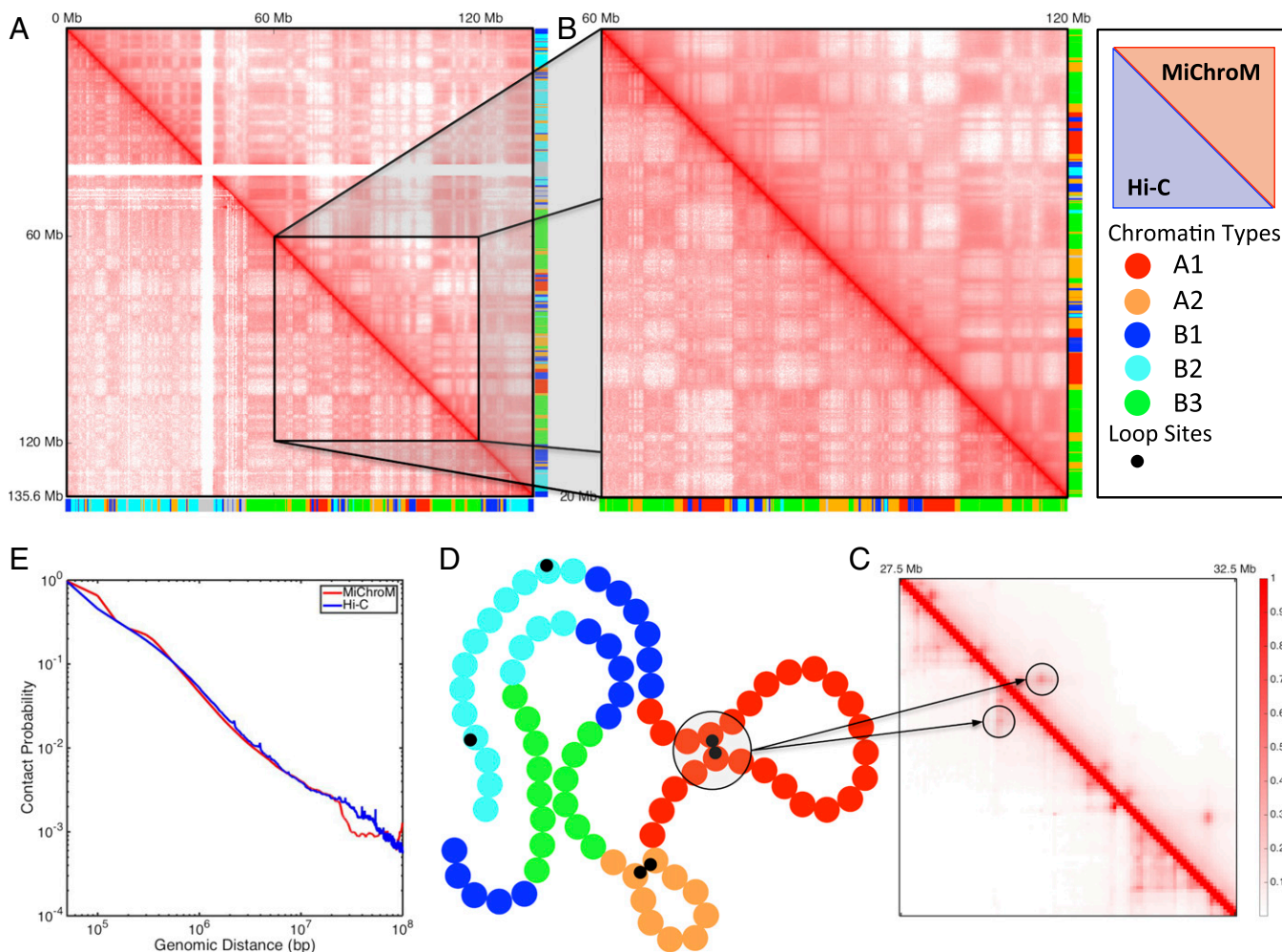
**Fig. 1.** Contact maps obtained by computational modeling using MiChroM closely reproduce experimental maps obtained using the Hi-C protocol. (*A–C*) Contact probability map of chromosome 10 of B-lymphoblastoid cells (GM12878). The map obtained from MiChroM is shown in the upper diagonal section of each map, whereas the lower diagonal region shows the maps reported by Rao et al. (9). A symmetrical figure indicates that the experimental data are reproduced well. Color bars on axes show the chromatin-type sequences of chromosome 10. (*A*) Complete contact map of chromosome 10 (log scale). (*B*) Magnification of the 60- to 120-Mb region of chromosome 10 (log scale). At this magnification, the relationship between chromatin types and spatial proximity is clearly visible; for example, it is easy to see that contacts between B-type loci are more frequent than contacts between B-type and A-type loci. MiChroM accurately reproduces the pattern of contacts measured by Hi-C. (*C*) Further magnification of the 27.5- to 32.5-Mb region (linear scale). The peaks in contact probability characterizing the loops are clearly visible in both experimental and computational maps (more details are provided in *SI Appendix*). (*D*) Schematic representation of MiChroM. Each bead represents 50 kb of chromatin belonging to one single chromatin type represented by its specific color as in the color bars of *A* and *B*. A smaller black bead marks the location of loop sites. (*E*) Probability of contacts as a function of genomic distance from experiment and in silico.

This assumption is well-supported by historical literature (8), and has been further confirmed by recent high-resolution Hi-C maps of the human genome, where loops are visible as peaks in the contact probability map (9). Most loops are associated with convergent pairs of CCCTC-binding factor (CTCF)–binding motifs, which have been proposed to help orchestrate loop formation via extrusion (21). MiChroM, however, makes no assumption about the particular mechanism of loop formation. Loops associated with the presence of CTCFs typically enclose a few hundred kilobases of DNA, and there is evidence that such structures are involved in diverse regulatory functions, including activation, repression, and insulation (8).

Finally, MiChroM assumes that every time a pair of loci comes into contact, there is a gain/loss of effective free energy, $\gamma(d)$, that depends only on the genomic distance, $d$. This "ideal chromosome" term models the local structure of chromatin in the absence of compartmentalization or looping (15), and is sequence translational invariant by construction. The form of the ideal chromosome potential is supported by the widespread evidence that chromatin can behave like a liquid crystal (22–24), and is consistent

with the popular notion of the existence of a higher order fiber in chromatin (25–27), although remaining more general.

To build a physical model for chromatin, we use the maximum entropy principle to convert the above three assumptions into an information theoretical energy function. The effective energy that maximizes the information theoretical entropy takes the following form (*SI Appendix*):

$$U_{MiChroM}(\vec{r}) = U_{HP}(\vec{r}) + \sum_{\substack{k \geq l \\ k,l \in \text{Types}}} \alpha_{kl} \sum_{\substack{i \in \{\text{Loci of Type } k\} \\ j \in \{\text{Loci of Type } l\}}} f(r_{ij})$$

$$+ \chi \sum_{(i,j) \in \{\text{Loop Sites}\}} f(r_{ij}) + \sum_{d=3}^{d_{cutoff}} \gamma(d) \sum_{i} f(r_{i,i+d}),$$

and includes, respectively, the potential energy, $U_{HP}(\vec{r})$, characterizing a generic homopolymer; the interactions between chromatin types (assumption 1); the interactions between loop anchors
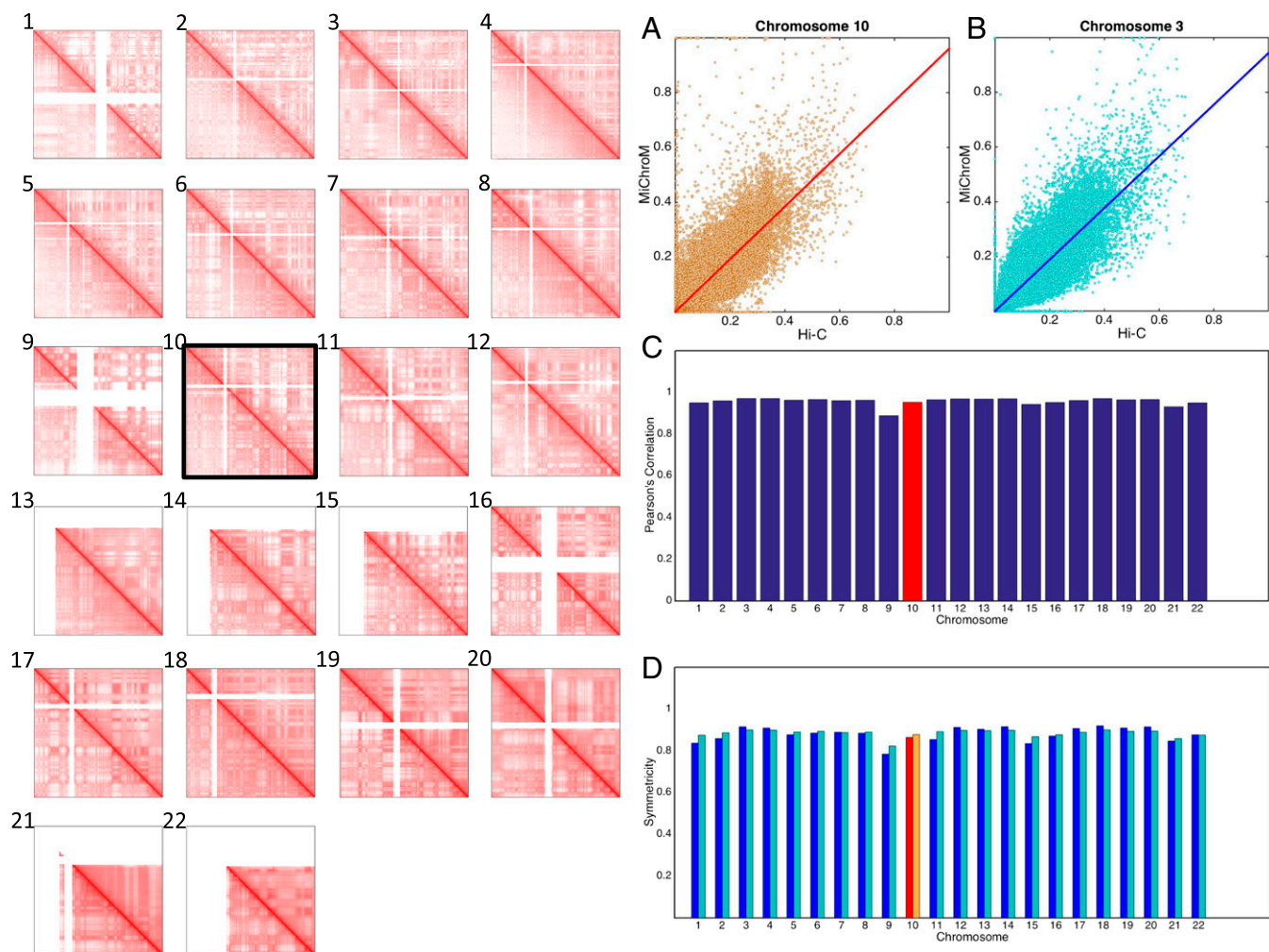
**Fig. 2.** MiChroM generates 3D structures for the chromosomes of B-lymphoblastoid cells (GM12878) that are in agreement with experimental data. Panels 1–22 show contact maps of chromosomes 1–22 represented in log scale; the upper diagonal regions show results from computational modeling, and the lower diagonal regions show maps obtained using the Hi-C protocol (9). Chromosome 10 was the only chromosome used for calibration. The contact maps for the remaining 21 chromosomes were generated de novo using the chromatin-type sequence and location of loops as the only inputs. The quality of the generated contact maps is consistently high, as shown by the symmetry of the maps. (*A*–*D*) Different measures of the accuracy of the maps generated by our model. (*A* and *B*) Scatter plots of Hi-C vs. MiChroM contact probability for all of the possible contacts, together with a linear fit of the data obtained by using the least squares method. (*A*) Data from chromosome 10; the slope of the fitting function is 0.96, and the intercept is 0.0003. (*B*) Data from chromosome 3; the slope in this case is 0.94, and the intercept is 0.0003. The experimental and computational datasets show a strong linear relationship confirming their similarity. The quality of such a linear relationship remains substantially unaltered whether we look at chromosome 10, which was used for calibration, or at chromosome 3, which was independently generated. (*C*) Pearson's correlation between the contact probabilities generated by MiChroM and as measured by Hi-C for all of the 22 chromosomes. The average correlation is 0.956. For chromosome 10, which was used for calibration, the correlation is 0.952 (also *SI Appendix*, Figs. S2 and S3). (*D*) Symmetry score (definition is provided in *SI Appendix*, Fig. S1) for the combined Hi-C/MiChroM maps for the 22 chromosomes. The score is bounded by 0 and 1, with 1 indicating a perfectly symmetrical matrix (i.e., perfect agreement between model and experiments). The average symmetry score is 0.88 using both the two-norm (blue bars) and the Frobenius norm (cyan).

(assumption 2); and the translational invariant compaction term (assumption 3).

This potential function contains 27 parameters that must be provided to specify the model fully. Once the potential function is fully specified, it is possible to perform molecular dynamics simulations of chromatin using as input the classification of loci into chromatin types and the location of loops. This procedure is directly analogous to the simulation of protein folding using amino acid sequence and disulfide bond positions as the only input.

Determining the optimal value for these 27 parameters requires a training dataset. In this case, we iteratively adjusted the parameter set to reproduce data extracted from a Hi-C contact map of chromosome 10 generated using GM12878 cells (9). To do so, we modeled human chromosome 10, which is 136 Mbp long, as a polymer containing 2,712 monomers, each representing 50 kb of DNA. We used the annotations generated by Rao

et al. (9) to assign each monomer a chromatin type, as well as to specify the positions of loops between pairs of monomers. In each iteration, we combined these polymer specifications with the current parameter set to generate an ensemble of structures. We then used this ensemble to generate a simulated map of pairwise intermonomer contact frequencies, and compared this contact map with the one obtained by Rao et al. (9) experimentally to choose the next set of parameters (*SI Appendix*).

The simulated contact maps obtained using the final set of parameters correspond closely to the experimental contact maps obtained for chromosome 10 (Pearson's $r = 0.95$). This correspondence goes beyond the visually obvious "checkerboard" pattern in the simulated contact map (Fig. 1). In general, all features larger than 300–400 kb in the experimental contact map (i.e., features that are about an order of magnitude larger than the size of an individual monomer in our simulations) appear to
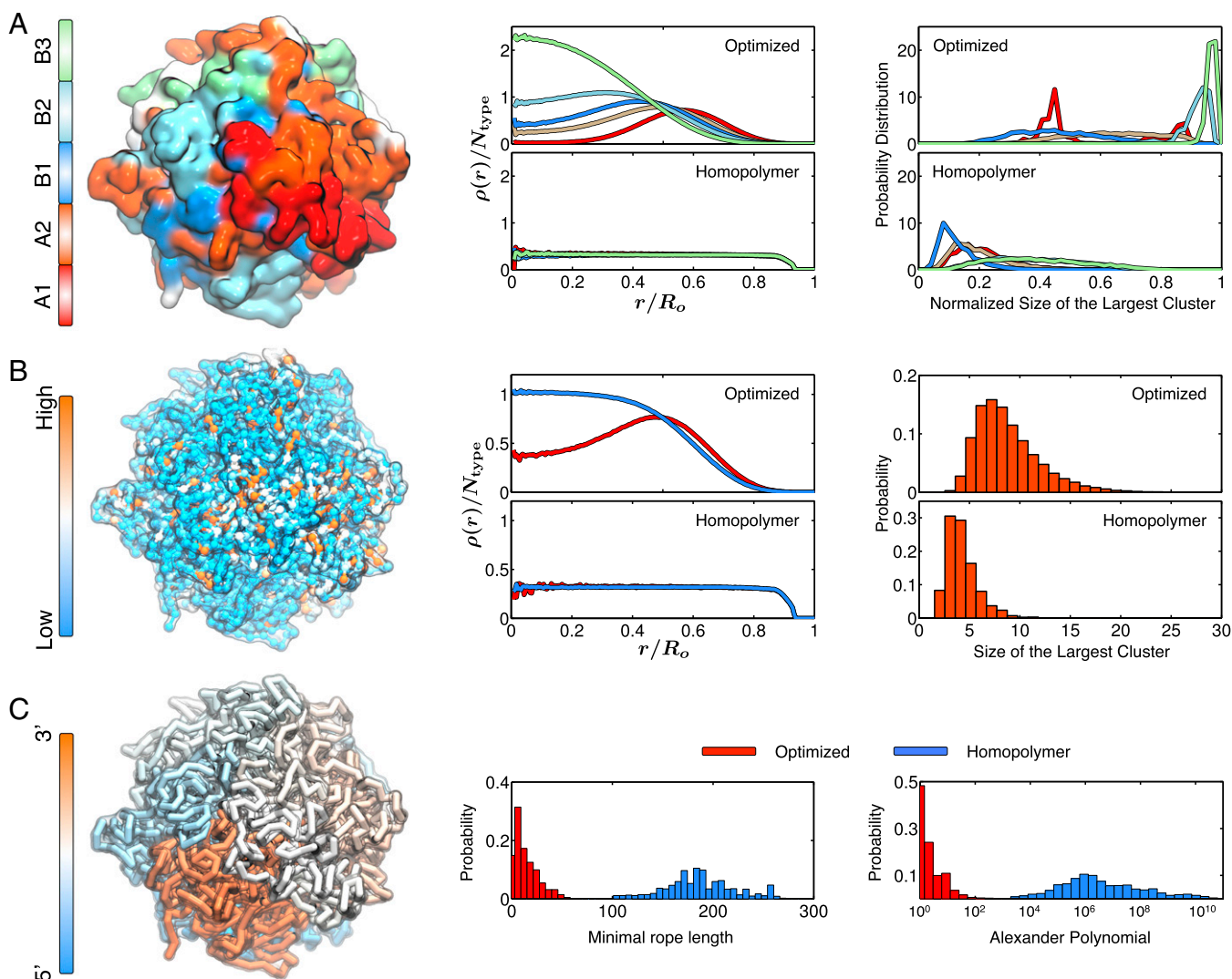
**Fig. 3.** Structural characterization of the simulated conformations for chromosome 10. (*A*) Chromatin of different types phase-separate, with A types localizing at the surface and B types in the interior. (*Left*) Surface plot for the chromosome colored by chromatin types, with the coloring scheme shown at the side. (*Center*) Radial density profiles for different chromatin types calculated from the ensemble of simulated chromosome (*Top*) and homopolymer (*Bottom*) conformations. $\rho(r)/N_{type}$, normalized radial density; $r/R_o$, normalized radius. (*Right*) Probability distributions for the size of the largest cluster found for each chromatin type from the ensemble of simulated chromosome (*Top*) and homopolymer (*Bottom*) conformations. The cluster sizes for different types are normalized by the total number of genomic loci for that type. (*B*) Genomic loci with high gene expression spatially colocalize at the exterior of the chromosome. (*Left*) Chromosome structure colored by gene expression, with red and blue representing high and low expression, respectively. (*Center*) Radial density profiles for genomic loci with high and low gene expression calculated from the ensemble of simulated chromosome (*Top*) and homopolymer (*Bottom*) conformations. (*Right*) Probabilities for finding the largest cluster of size N for highly expressed genomic loci from the ensemble of simulated chromosome (*Top*) and homopolymer (*Bottom*) conformations. (*C*) Simulated chromosomes adopt knot-free conformations. (*Left*) Chromosome structure colored by genomic distances, with one end of the chromosome shown in blue and the other end shown in red. (*Center*) Probability distributions of the knot invariant measured as the minimal rope length for the ensemble of simulated chromosome (red) and homopolymer (blue) conformations. (*Right*) Probability distributions of the knot invariant measured as the Alexander polynomial for the ensemble of simulated chromosome (red) and homopolymer (blue) conformations.

be accurately recapitulated by the MiChroM model. Notably, the power law scaling relationship between the probability of forming contacts and genomic distance, often used to justify the nonequilibrium fractal globule model, is also reproduced with great accuracy by this equilibrium model (Fig. 1*E*).

Next, we applied the MiChroM model to the remaining GM12878 autosomes by combining the potential function with the experimentally derived monomer type and loop annotations. When each chromosome is simulated separately, the resulting intrachromosomal contact map closely corresponds to the experimental contact map in every case. Notably, the correspondence for autosomes that were not used to train the potential function was typically as close (Pearson's r = 0.95) as the correspondence

for chromosome 10 (Fig. 2 and *SI Appendix, Supplementary Text and Figs. S2–S47*).

When we examined the ensemble of 3D structures for each individual chromosome, we observed that each chromosome formed a compact chromosome territory. We also observed the phase separation of chromatin types within this territory, leading to subvolumes comprising only a single type of genomic interval (Fig. 3*A*). Usually, only a single subvolume formed for each subcompartment, although we observed multiple subvolumes of a single type in some cases. Similarly, we see that highly expressed genes [as measured by RNA sequencing (28)] tend to occupy spatial subvolumes, which is expected, given that highly expressed genes lie predominantly in the A compartment. Overall, these

findings are consistent with the notion that different types of intervals colocalize in distinct spatial compartments. Interestingly, the A compartment tends to be less densely packed and to lie at the periphery of the chromosome territory. These observations are consistent with the findings of prior studies using both microscopy and Hi-C (9, 29, 30). Notably, a control model composed of a simple self-avoiding homopolymer chain failed to exhibit any of these results and, instead, recapitulated the expected properties for an equilibrium globule (Fig. 3 *A* and *B* and *SI Appendix*, Fig. S3).

It is commonly assumed that one essential feature of chromosomes is the absence of knots, because one might suppose that a highly knotted structure could create obstacles to the transcription process. We studied the extent of knotting in the ensemble of chromosome structures sampled from the optimized energy landscape and from the homopolymer potential. To quantify knotting in a particular conformation of the chromosome, we used two different knot invariants: the Alexander polynomial and the minimal rope length required to generate a topologically equivalent knot (15, 31). Both measures show that the configurations produced by MiChroM are largely devoid of knots. In contrast, the homopolymer control system tended to form extraordinarily complex knots (Fig. 3*C*). This topological feature is a direct result of inferring the energy landscape from the three physical assumptions explained above. Remarkably, the simple equilibrium mechanism underlying MiChroM produces ensembles of structures that are devoid of knots.

Finally, we used MiChroM to simulate chromosomes 17 and 18 jointly (*SI Appendix*, Fig. S1). This simulation allowed us to explore whether the MiChroM potential function, which was trained using a single intrachromosomal contact map for chromosome 10, could successfully reproduce genome architecture at a larger scale. The resulting intrachromosomal contact maps are essentially the same as those intrachromosomal contact maps simulated in isolation (Pearson's *r* = 0.96). The phenomenon of phase separation of chromatin types now extends to both chromosomes, creating larger regions of space occupied by one single type. Spatial confinement introduces artifacts in the frequency of interchromosomal contacts; therefore, the interchromosomal contact map from simulation shows somewhat increased probabilities with respect to Hi-C. Even with the biased intensity, the two-chromosome map shows a correct pattern of interchromosomal interactions.

When we examined the 3D ensemble, we found that, despite the extensive contacts between the chromosomes, the chromosomes were not entangled with one another (*SI Appendix*, Fig. S1*B*); instead, we observed the formation of nonoverlapping chromosome territories. This last result highlights the fact that MiChroM can successfully recapitulate features of the nucleus as a whole.

The MiChroM assumes that chromosomes fold under the action of a cloud of proteins that bind with different selectivity to different sections of chromatin, and offers a simple strategy for recapitulating the energy landscape created by such interactions. This energy landscape brings about transient contacts rather than permanent ones, which is consistent with the fact that most of the experimentally observed contacts between two genetic loci only occur in a small fraction of cells at a given time (5, 32). Contacts associated with loop formation tend to be more frequent; accordingly, our optimization algorithm assigns them a larger free energy gain upon formation. In humans, we find that six types of chromatin are sufficient to reproduce the arrangement of interphase DNA in vivo. The fact that our model can be reliably transferred from one chromosome to the rest suggests the plausibility of the proposed energetic mechanism, even if the underlying biochemical details remain unclear at the present time.

As shown, MiChroM is able to explain and reproduce the results of DNA proximity ligation experiments. Nevertheless, caution must be applied in the interpretation of these results. Hi-C experiments are performed using millions of cells at once, and report only a population average. We know little about what happens in individual cells at specific moments in time. For instance, a typical cell population interrogated by Hi-C may contain entirely separate subpopulations, as well as fluctuating or even oscillating configurations. These subpopulations and configurations would be lost in MiChroM.

The classification of loci into chromatin types and the position of chromatin loops, which are inputs of our model, are strongly associated with epigenetic features (histone modifications and bound CTCF motifs in convergent orientation) that can be directly and inexpensively assayed by ChIP sequencing. Exploiting these associations along with MiChroM opens up the possibility of predicting in silico the 3D structure of whole genomes starting from 1D genomics data, which are often already publicly available.

1. Cremer T, Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2(4):292–301.
2. Bickmore WA (2013) The spatial organization of the human genome. *Annu Rev Genomics Hum Genet* 14:67–84.
3. Grosberg AY, Nechaev SK, Shakhnovich EI (1988) The role of topological constraints in the kinetics of collapse of macromolecules. *J Phys (Paris)* 49(12):2095–2100.
4. Grosberg A, Rabin Y, Havlin S, Neer A (1993) Crumpled globule model of the 3-dimensional structure of DNA. *Europhys Lett* 23(5):373–378.
5. Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293.
6. Gürsoy G, Xu Y, Kenter AL, Liang J (2014) Spatial confinement is a major determinant of the folding landscape of human chromosomes. *Nucleic Acids Res* 42(13):8223–8230.
7. van Steensel B (2011) Chromatin: Constructing the big picture. *EMBO J* 30(10):1885–1895.
8. Phillips JE, Corces VG (2009) CTCF: Master weaver of the genome. *Cell* 137(7):1194–1211.
9. Rao SSP, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–1680.
10. Collepardo-Guevara R, et al. (2015) Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: A multiscale computational study. *J Am Chem Soc* 137(32):10205–10215.
11. Cullen KE, Kladde MP, Seyfred MA (1993) Interaction between transcription regulatory regions of prolactin chromatin. *Science* 261(5118):203–206.
12. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295(5558):1306–1311.
13. Dostie J, et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16(10):1299–1309.
14. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106(4): 620–630.
15. Zhang B, Wolynes PG (2015) Topology, structures, and energy landscapes of human chromosomes. *Proc Natl Acad Sci USA* 112(19):6062–6067.
16. Jost D, Carrivain P, Cavalli G, Vaillant C (2014) Modeling epigenome folding: Formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res* 42(15):9553–9561.
17. Filion GJ, et al. (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* 143(2):212–224.
18. Dixon JR, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380.
19. Eagen KP, Hartl TA, Kornberg RD (2015) Stable chromosome condensation revealed by chromosome conformation capture. *Cell* 163(4):934–946.
20. Sexton T, et al. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 148(3):458–472.
21. Sanborn AL, et al. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA* 112(47):E6456–E6465.
22. Boy de la Tour E, Laemmli UK (1988) The metaphase scaffold is helically folded: Sister chromatids have predominantly opposite helical handedness. *Cell* 55(6):937–944.

23. Naumova N, et al. (2013) Organization of the mitotic chromosome. *Science* 342(6161): 948–953.

24. Zhang B, Wolynes PG (2016) Shape transitions and chiral symmetry breaking in the energy landscape of the mitotic chromosome. *Phys Rev Lett* 116(24): 248101.

25. Bascom GD, Sanbonmatsu KY, Schlick T (2016) Mesoscale modeling reveals hierarchical looping of chromatin fibers near gene regulatory elements. *J Phys Chem B* 120(33):8642–8653.

26. Grigoryev SA, et al. (2016) Hierarchical looping of zigzag nucleosome chains in metaphase chromosomes. *Proc Natl Acad Sci USA* 113(5):1238–1243.

27. Maeshima K, Hihara S, Eltsov M (2010) Chromatin structure: Does the 30-nm fibre exist in vivo? *Curr Opin Cell Biol* 22(3):291–297.

28. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.

29. Boettiger AN, et al. (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* 529(7586):418–422.

30. Hübner MR, Spector DL (2010) Chromatin dynamics. *Annu Rev Biophys* 39:471–489.

31. Stasiak A, Katritch V, Kauffman LH (1998) *Ideal Knots* (World Scientific, River Edge, NJ).

32. Bantignies F, et al. (2011) Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell* 144(2):214–226.

PHYSICS

BIOPHYSICS AND
COMPUTATIONAL BIOLOGY