HHS Public Access

Author manuscript

Methods. Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

Methods. 2020 October 01; 181-182: 24–34. doi:10.1016/j.ymeth.2019.08.008.

Computational Approaches for Inferring 3D Conformations of Chromatin from Chromosome Conformation Capture Data

Dario Meluzzi[1], **Gaurav Arya**[2,*]

[1]Department of Medicine, University of California San Diego, La Jolla, CA 92093

[2]Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708

Abstract

Chromosome conformation capture (3C) and its variants are powerful experimental techniques for probing intra- and inter-chromosomal interactions within cell nuclei at high resolution and in a high-throughput, quantitative manner. The contact maps derived from such experiments provide an avenue for inferring the 3D spatial organization of the genome. This review provides an overview of the various computational methods developed in the past decade for addressing the very important but challenging problem of deducing the detailed 3D structure or structure population of chromosomal domains, chromosomes, and even entire genomes from 3C contact maps.

INTRODUCTION

Eukaryotic chromosomes are made up of chromatin, a fibrous complex of DNA and histone proteins.[1] Understanding how the chromatin fiber is spatially organized inside the cell nucleus has become an increasingly important topic of study.[2-4] The reasons for the growing interest include: a greater appreciation for the biological roles of chromatin organization in all cellular and nuclear processes;[5] an increasing numbers of cancers,[6-8] developmental defects,[9, 10] and neurological disorders[11, 12] being linked to defects in chromatin organization; and recent developments in powerful microscopy and DNA sequencing technologies providing a wealth of new data.[13-16] The most obvious role of chromatin organization is in packaging the enormously long genomic DNA into the tiny confines of the cell nucleus, while enabling ready access to genes and regulatory elements on demand.[17, 18] Chromatin organization also plays critical roles in DNA transcription and recombination by bringing into proximity multiple functional DNA elements that are otherwise distant on the DNA sequence. Two striking illustrations of such long-range interactions, mediated by chromatin looping, include the classical case of promoter-enhancer interactions required for initiating transcription[19] and the fascinating rosette-like organization of the immunoglobulin heavy chain locus that appears in developing B-cells to facilitate V(D)J recombination.[20, 21]

*Corresponding author (gaurav.arya@duke.edu, Phone: 919-660-5435, Fax: 919-660-8963).

Lastly, chromatin organization enables physical segregation of genomic elements based on their function. For instance, mammalian genomes are organized into topologically associating domains (TADs) harboring common epigenetic marks at the 100 kb to 1 Mb scale,[22-24] into tissue-specific compartments of active and inactive chromatin at ~5 Mb scales,[25-27] and into the well-known chromosome territories at the largest scale.[28]

Biology has traditionally relied on light and electron microscopy for studying intracellular organelles and structures. However, these approaches are not ideal for visualizing the 3D organization of chromatin *in vivo*, owing to the low, diffraction-limited resolution of light microscopy (~200 nm) and to the highly invasive and non-specific nature of electron microscopy.[29] Typically, chromatin folding has been visualized by multicolor fluorescence *in situ* hybridization (FISH), which can simultaneously provide the 3D positions of multiple genomic loci separated by >100 kb along DNA sequence or >200 nm in space,[30] though discerning the folded configuration of chromatin remains challenging. While limitations in resolution can be overcome by super-resolution microscopy techniques,[31, 32] these lack the throughput necessary to study more than a few chromatin regions at once.

A recently developed set of experimental methods, known as chromosome conformation capture (3C), has enabled researchers to study chromatin interactions and conformations at an unprecedented resolution and throughput (Fig. 1).[33, 34] These methods begin by treating the chromatin inside cell nuclei with chemicals like formaldehyde. In this manner, DNA loci that are in close spatial proximity to each other become cross-linked, typically via intervening proteins. Next, the cross-linked DNA is digested using a restriction enzyme, yielding two kinds of DNA fragments: isolated and cross-linked fragments, with only the latter kind containing information about the interacting genomic loci. These loci are identified and quantified from the cross-linked fragments through a series of biochemical and bioinformatic steps. Then, by counting the number of times that each pair of loci is observed in cross-linked fragments collected from millions of cells, one can construct a 2D contact map that quantifies the frequency of interactions between all pairs of loci. While the original 3C method detected interactions between only a few preselected pairs of loci,[35] the latest and most advanced 3C variant, known as Hi-C, takes advantage of high-throughput sequencing to provide a comprehensive map of interactions across the entire genome at resolutions of up to ~1 kb.[36, 37] Importantly, because the frequencies of interactions between genomic loci must on average be related to their spatial proximity in some reciprocal manner, the contact maps also contain valuable information about the 3D conformation of the underlying chromatin fiber. Inferring such structural information from contact maps is however a challenging task, because many unknowns are associated with the underlying chromatin fiber, including its physical properties, its variability across populations of cells, and the uncertainties in the measured contact counts, and because of the high dimensionality of the configurational space. To tackle this multidimensional structure-determination problem, different strategies, assumptions, and approximations have been proposed to develop a variety of ingenious computational approaches, which have been described by many excellent reviews.[38-45]

In this review, we provide our perspective on these computational approaches, which can be generally categorized into three classes. The first class assumes a functional relationship

between spatial distance and interaction frequency, enabling the optimization of chromatin conformation directly in the distance space. The second class does not invoke any such functional relationship, but uses polymer models to directly predict chromatin interactions, thus enabling the optimization of structures in the interaction frequency space. The third class also uses a distance-frequency relationship, but describes the measured interaction frequencies in probabilistic terms, thus enabling the optimization of structures within the statistical parameter space. The three classes of approaches are not completely mutually exclusive, as some approaches aptly combine features of different classes. For instance, some aspects of polymer modeling may also be used in the first and third class of methods. Nevertheless, such classification serves to organize the several possible components of a computational protocol for obtaining 3D chromatin conformation from 3C data. This review article does not cover other interesting aspects of 3C experiments, such as the experimental protocols, the computational pipelines for generating contact maps, and the vast amount of chromosome biology and physics learnt from such maps using *de-novo* polymer models. For information on these topics we refer the reader to several pertinent reviews.[46-52]

PROBLEM DEFINITION

All computational methods considered in this review take as input a 2D contact map and generate as output one or more 3D conformations of chromatin. Before describing these methods, we define their input and output in mathematical terms.

Hi-C experiments can generate up to ~10^9 sequence reads.[53] However, even such large numbers of reads are insufficient to cover all possible inter-fragment interactions, whose number is much larger. For instance, the *MboI* enzyme generates ~10^6 fragments from the human genome, involving ~10^{12} possible inter-fragment interactions. Therefore, contact maps are typically constructed by dividing the genome into “bins” that are larger than the fragments[46] Interactions are then counted across genomic bins rather than fragments. Denoting the total number of bins by N, a Hi-C contact map is represented as a symmetric $N \times N$ matrix

$$\mathbf{C} \equiv \begin{bmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NN} \end{bmatrix} \tag{1}$$

where c_{ij} represents the count of fragment pairs with one fragment in bin i and the other in bin j. The contact maps are then corrected for various experimental biases such as those arising from sequence mappability, density of restriction sites, and GC content.[54] The contact maps of 3C and 5C experiments, which probe only selected pairs of interactions and can afford to probe them at much higher precision, are however best described in terms of the original inter-fragment interaction counts.[55] In 5C, which relies on distinct sets of forward and reverse oligonucleotide probes that may also differ in numbers, only interactions between fragments associated with oppositely oriented primers can be probed, and hence the contact matrix is no longer symmetric and is also not square in most cases. Nevertheless, the computational approaches developed for Hi-C maps may be adapted for maps obtained from 3C and 5C, and *vice versa*, as long as differences in the definition and

resolution of the interacting loci (fragments versus bins) between the two types of techniques are taken into account.[56] The contacts in Eq. 1 are often converted to interaction frequencies (IFs) or contact probabilities via $f_{ij} = c_{ij}/\sum_{i,j} c_{ij}$ and the resulting "normalized" matrix is denoted by **F.**

Given a contact map **C,** or **F,** as input, the desired computational output consists of one or more 3D conformations of the chromatin fiber that are consistent with that contact map. One such conformation, or structure, can be described by a $3 \times N$ matrix

$$\mathbf{R} \equiv [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N] \tag{2}$$

where $\mathbf{r}_i \equiv [x_i, y_i, z_i]^T$ represents the position vector of genomic bin $i = 1,\ldots,N$ in terms of Cartesian coordinates.

DISTANCE-OPTIMIZATION METHODS

The underlying idea in these methods is to convert the contact frequencies f_{ij} in the contact map into suitable *spatial*, or Euclidean, distances δ_{ij} between the interacting loci, and then determine the locus coordinates $\mathbf{r}_i$ of the chromatin conformation whose internal Euclidean distances $d_{ij} \equiv |\mathbf{r}_i - \mathbf{r}_j|$ best match the target distances δ_{ij} derived from the maps. The single solution for the chromatin conformation obtained in this manner is often referred to as the "consensus structure" (Fig. 2).

There are two key assumptions inherent in these methods. The first deals with the existence of a functional relationship for converting frequencies into distances. Certainly, such relationships are available for simple models of polymers. For instance, in *ideal chains*, where segments are connected by freely-rotatable joints, the frequency f_{ij} of overlap of two segments i and j decays as $f_{ij} \sim s_{ij}^{-3/2}$ with respect to their separation $s_{ij} \equiv |i - j|$ along the chain, while their spatial distance increases as $d_{ij} \sim s_{ij}^{1/2}$.[57, 58] Combining the two results yields the simple inverse relationship $d_{ij} \sim f_{ij}^{-1/3}$ between frequency and distance. Chromosomes are however considerably more complex than ideal chains due to energetic interactions and heterogeneity of the underlying chromatin fiber and due to confinement and molecular crowding effects. Furthermore, the interactions observed in 3C maps arise not only from random collisions between loci, but also from protein-mediated loops with unknown lifetimes. The d_{ij}-f_{ij} relationship in chromosomes is generally far too complex to be described by any tangible model. Nevertheless, studies have employed various functions to at least capture the reciprocal dependence of locus distances on frequency, often with adjustable parameters to obtain the best agreement with experimental data.

The second assumption is that the obtained consensus structure is a reasonable approximation of the "average" conformation of chromatin exhibited by the ensemble of cells from which the 3C data were derived. Chromosomes, on the other hand, are highly dynamic entities and the conformation of chromatin likely varies from cell to cell. This conclusion may in fact be directly deduced from the non-binary nature of the interaction frequencies.[49] Nonetheless, consensus structures may still provide valuable information on

the overall structure of chromosomes, and on differences in chromosome organization between different populations of cells examined by 3C.

Scoring Function and Constraints.

The consensus structure is obtained by minimizing the deviation of its internal distances d_{ij} from the corresponding distances $\delta_{ij} = \mathcal{F}(f_{ij})$ inferred from the 3C contact map via a suitable function $\mathcal{F}$ that converts frequencies to distances. In its simplest formulation, this deviation is quantified as a weighted sum of squares of the individual deviations of the internal distances from the distance restraints

$$S(\mathbf{R}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} w_{ij}(d_{ij} - \delta_{ij})^2 \quad (3)$$

where w_{ij} are weights and $S(\mathbf{R})$ is the objective or scoring function that needs to be minimized to attain the consensus structure denoted by $\mathbf{R}_0$. Without w_{ij}, the scoring function would be dominated by pairs of loci with large δ_{ij}. Since the largest distances are inferred from the smallest f_{ij} values, the largest δ_{ij} represent the least reliable inter-locus distances. Thus, to ensure that the scoring function is appropriately weighted based on the reliability of each restraint, w_{ij} should ideally decrease with δ_{ij} (or increase with f_{ij}). A popular weighting function is $w_{ij} = 1 / \delta_{ij}^2$,[59-61] though $w_{ij} = 1/\delta_{ij}$,[62] $w_{ij} = 1$,[63-65] and $w_{ij} \propto |\text{Zscore}(f_{ij})|^p$, with $p = 0.5$ or 2,[66, 67] have also been used. Interestingly, the latter approach assigns larger weights to frequencies that deviate from its average value in the map, irrespective of whether f_{ij} is smaller or larger than the average. Instead of using a functional form for $w_{ij}(\delta_{ij})$, one approach used two different values of w_{ij}, a large value for all δ_{ij} smaller than some cutoff distance and a small one for larger distances.[68] In some cases, the nature of the weighting function is restricted by the optimization approach. For instance, the multidimensional scaling approach discussed below generally requires $w_{ij} = 1$,[45, 69, 70] though using log-transformed distances may allow $w_{ij} = 1 / \delta_{ij}^2$ to be used.[71] The quadratic form of the above scoring function implies that even small deviations of the structure from the targeted distances δ_{ij} get penalized. Hence, any uncertainties in the function $\mathcal{F}$ used to derive the δ_{ij}'s propagate to the obtained consensus structure. To alleviate some of the inherent bias introduced through the uncertain function $\mathcal{F}$, several approaches have attempted to use "flat-bottomed" restraints that do not penalize deviations from δ_{ij} until they become larger than a cutoff.[72-74] Another approach devised a bell-shaped Lorentzian scoring function that assigns larger weights to consistent distance restraints whose values are not affected by the violation of inconsistent restraints.[75]

The scoring functions discussed above constrains the path of the chromatin fiber based purely on the distance restraints obtained from 3C data. However, because of the two assumptions and the data reliability issues discussed above, the consensus structures obtained after optimization may not be physically or biologically realistic. In addition, as this scoring function lacks an associated length scale, all structures proportional to $\mathbf{R}_0$ are equally valid solutions. Hence, the scoring function is often accompanied by additional structural constraints based on physical properties of chromatin and biological data. Chain connectivity and excluded volume represent two commonly implemented *physical*

constraints. The first constraint, which accounts for the physical connectivity of adjacent loci i and $i + 1$, is usually implemented as an upper-bound limit d_{max} on the spatial distance between the two loci given their lengths and the linear density of chromatin.[61, 63, 72] The second constraint, accounting for the excluded volume of each locus, is typically implemented as a lower-bound limit d_{min} on the spatial distance between all locus pairs i and j, where d_{min} is either the diameter of the chromatin fiber for high-resolution structures, or a suitable larger value for coarser models.[63, 72, 73, 76] Two common *biological* constraints include confinement and position restraints. The confinement constraint ensures that every locus in the genome is located inside the cell nucleus, usually approximated as a sphere of size consistent with that obtained from microscopy.[61, 63] Confinement constraints may also be imposed on individual chromosomes, as they are known to occupy specific territories within the nucleus.[28] In this case, the size of the confinement domain may be obtained from whole-chromosome FISH painting or estimated by scaling down the size of the nucleus with the ratio of the chromosome to genome length.[60] Microscopy also provides information on the positioning of specific chromosomal regions, such as centromeres, telomeres, and lamin-associated domains, thus allowing related constraints to be implemented.[61, 63, 76]

Frequency-Distance Conversion.

The relationship used for converting frequencies into distances plays a key role in these methods. The simple relationship $\delta_{ij} = \gamma f_{ij}^{-1/3}$ obtained from the polymer physics of ideal chain and fractal-globule models provides one possible option.[61, 73] A softer version of this inverse relationship, $\delta_{ij} = \gamma / f_{ij}$, has also been used, especially in some of the earlier methods.[59, 63] The unknown prefactor γ that sets the scale of the system is generally chosen or optimized so that the strongest frequencies yields an appropriate contact distance, e.g., thickness of chromatin fiber, or that the resulting structure exhibits a more or less uniform density in the nucleus.[61] Recognizing that chromosomes are complex and likely not to follow any single, tangible relationship between frequency and distance, many recent methods have begun to use a more open relationship $\delta_{ij} = \gamma f_{ij}^{-\alpha}$, where parameter α is used as an unknown parameter that also needs to be optimized.[62, 65, 75, 77-79] Values of α obtained in this manner have been found to range between 0.5 and 1.4, highlighting the variable and complex nature of chromatin. Extending the concept of variability even further, one study assigned locus-dependent α's to account for the tendency of some loci to co-cluster.[80] Several approaches have also considered using linear relationships, e.g., $\delta_{ij} = -mf_{ij} + c$ (with $m > 0$), to only assign $\mathcal{F}$ within a select range of frequencies deemed to be sufficiently reliable.[60, 66]

A more robust alternative to assuming relationships for $\mathcal{F}$ is to derive such a function directly from a combination of 3D-FISH and 3C measurements, which respectively provide inter-locus distances and corresponding interaction frequencies. Such a strategy has been used by several researchers,[64, 67, 70, 81] yielding various different kinds of calibration curves. In some cases, the δ_{ij} vs. f_{ij} data could be well described by the power-law function $\delta_{ij} \sim f_{ij}^{-\alpha}$ with $\alpha = 0.25$[70] and $\alpha = 0.39–0.49$,[81] while in other cases it was better described by a decaying exponential function[64] and a 5th-degree polynomial.[67, 82] An alternative approach is to derive the embedding function itself during optimization via nonlinear dimensionality

reduction.[69, 83] One of these studies[69] recovered a reciprocal relationship between δ_{ij} and f_{ij} that could in fact be described by the power-law $\delta_{ij} \sim f_{ij}^{-\alpha}$. However, the recovered exponent a was found to vary across different Hi-C data sets, attesting to the large variability in chromatin behavior or in Hi-C experiments, and to the inaccuracy inherent in using a single function $\mathcal{F}$.

Structure Determination.

The last aspect of these approaches involves optimizing chromatin conformation by minimizing the scoring function, often subject to the additional constraints mentioned earlier. A variety of optimization approaches have been employed, which may be roughly classified into two categories: (i) methods that use classical numerical optimization algorithms, which smartly traverse the multidimensional vector space **R** to locate the global minimum $\mathbf{R}_0$, and (ii) methods that use the established mathematical formalism of multidimensional scaling to obtain $\mathbf{R}_0$ through matrix algebra operations.

Classical numerical optimization offer flexibility, easy integration of a wide variety of weights and constraints, and the opportunity to observe intermediate structural solutions. While gradient-descent (or gradient-ascent) optimization provide a simple and efficient approach for optimizing scoring functions containing a few distance restraints and no constraints,[59, 74, 75] more sophisticated optimization methods are required when analyzing contact maps involving hundreds of loci (bins) with an even larger number of constraints. A number of studies have used the open-source IPOPT software for such optimization.[63, 64, 68, 80] IPOPT is an interior-point gradient-based algorithm that is especially adept at tackling high-dimensional, nonlinear constrained optimization problems.[84] Another commonly used optimization approach is simulated annealing.[85] Here, the loci are represented by particles that interact with each other via harmonic potentials representing each of the distance restraints, *i.e.*, particles i and j interact with a harmonic spring of equilibrium length δ_{ij} and spring constant $2w_{ij}$, whereupon the scoring function simply becomes the total energy of this system of interacting particles. The particle positions are sampled stochastically or deterministically using Monte Carlo (MC)[66, 82, 86] or molecular dynamics (MD)[65, 72, 73] simulations at a fictitious temperature. This temperature is set to a high value at the beginning of the simulation and is gradually lowered in steps during the simulation, enabling both rapid exploration of the configuration space and increased chance of trapping a configuration at or at least near the global minimum. In addition, multiple copies of simulations are performed to collect many possible candidates for the global minimum. Such candidates can then be clustered based on structural similarity to identify the consensus structure.[66] While external constraints are easily handled in MC methods, they typically need to be converted into stiff potentials and moved into the scoring function in MD simulations.

Multidimensional scaling (MDS)[87] provides a more elegant and efficient alternative to determining the consensus structure, though this approach works best with quadratic scoring functions without constraints and with equal weights. Given a complete set of pairwise distances between points in some vector space of high dimension K, MDS allows one to infer the vector coordinates of those points in a space of dimension $k \leq K$ so that the

cumulative deviation in the inferred distances from the given distances is minimized. This approach has found applications in sensor network localization[88] and in determining molecular structures based on interatomic distances obtained from nuclear magnetic resonance.[89] More recently, MDS has been use to determine chromatin conformations that best satisfy distance restraints. While the most general forms of MDS scoring, or stress, functions require numerical algorithms known as stress majorization to obtain optimal structures,[90] the simplest form of MDS (with $w_{ij} = 1$) can be solved algebraically. This involves formulating a suitably centered $N \times N$ "Gram" matrix from distances δ_{ij} inferred from the contact map, and then determining the largest three eigenvalues and corresponding eigenvectors of the Gram matrix. Simple rescaling of normalized eigenvectors by the square-root of their corresponding eigenvalues then yields the desired consensus structure.[45] Such an approach was first applied by the developers of the original 3C technique on the limited number of interactions they obtained.[35] Since then, the approach has been applied to the more exhaustive Hi-C contact maps, which pose additional challenges. One of the main challenges in dealing with Hi-C data, especially from single cells, is that the contact maps are often quite sparse, i.e., contain many zero c_{ij}'s. Furthermore, due to low reliability of some interactions, pairs of inferred distances may not satisfy the triangle inequality ($\delta_{ij} + \delta_{jk} \geq \delta_{ik}$ for any loci i, j, and k). Hence, significant efforts have been devoted to assigning more reliable distances to missing or low frequency data using shortest-path approaches (e.g., the Floyd-Warshall algorithm) adopted from graph theory,[45, 78, 91] recurrence plots adopted from non-linear time-series analysis,[92] and regularization terms that account for missing data.[62] Another strategy for dealing with map sparseness is multi-staged implementation of MDS, where MDS is first applied at high resolution to strongly interacting domains with large c_{ij}, e.g., TADs, and then applied at a lower resolution to weaker interactions across these domains.[70] Researchers have also begun to consider optimization on distance manifolds to reduce the importance of distances derived from low frequencies[86] or to entirely eliminate the use of the function $\mathcal{F}$.[69, 83]

POLYMER PHYSICS-BASED METHODS

Approaches based on polymer physics involve optimizing the parameters of a polymer-chain model of the chromatin fiber, whose conformational ensemble best recapitulates the experimentally-derived 3C contact maps. As shown in Fig. 3, such "training" of the model is usually accomplished via an iterative process involving three components: a polymer model, a sampling algorithm, and a parameter optimizer. The polymer model captures the physical properties of chromatin fiber, using a few known and unknown parameters. The sampling algorithm generates a thermodynamic ensemble of chromatin conformations consistent with the parameters of the model. These conformations are used for generating a "predicted" version of the contact map. The parameter optimizer compares the predicted map against the experimental map to suggest parametric refinements to the model in order to improve the agreement between the two maps. This process is repeated until the best possible agreement is achieved. Since all information about inter-locus interactions is embedded in the polymer model, which is assumed to accurately capture the conformational behavior of the chromatin fiber, these methods do not need to invoke any specific functional relationship between spatial distance d_{ij} and interaction frequency f_{ij}. Furthermore, it is not necessary to generate

conformations that are individually consistent with the contact map; only the conformation ensemble as a whole should be consistent. Therefore, these approaches by construction incorporate variability in chromatin structures and interactions across members of the conformational ensemble, mimicking the population nature of the interactions derived from 3C experiments.

Polymer Model.

The interaction frequencies probed by the 3C method arise from a combination of *specific* interactions, usually attributed to protein-mediated looping of chromatin, and *non-specific* interactions, associated with random collisions between loci. A polymer model representing chromatin should capture both kinds of interactions in its ensemble of conformations.

The frequencies of *non-specific* interactions between loci are dictated by the separation distance between the loci *along* the chromatin fiber, by the physical properties of the fiber, and by its confinement. Each of these effects can be captured in a bead-chain model, where each bead represents a segment of the chromatin fiber. Usually the beads represent segments of fixed length, which can be as short as ~3 kb[93-95] to as long as ~500 kb,[96] depending on the desired resolution. Adjacent beads along the fiber are connected by rigid bonds,[94, 95] harmonic springs,[93, 97] or finite extensible nonlinear elastic springs[98-100] to account for the stretching resistance of the fibers. Such “connectivity” restraints are already sufficient to produce polymer-like conformations. Chromatin fibers, however, are also resistant to bending, and their bending rigidity is often treated using harmonic[93] or cosine potentials[99] in the bending angle subtended by three adjacent beads along the fiber. The excluded volume of the beads prevents them from penetrating each other, and can be represented by hard-sphere[101] or short-range repulsive potentials[93, 99] between non-adjacent beads. The above polymer representation is often referred to as the self-avoiding wormlike chain (SA-WLC). Confinement effects arising from nuclear boundaries can also be incorporated using excluded-volume potentials.[95, 96, 98, 99] When modeling individual chromosomes or chromosomal domains, these confinement effects may arise from other chromosomes or domains whose boundaries are not known. Such unknown effects can be implicitly included in the models for specific interactions.

The choice of interaction potentials depends on the resolution of the model. At high resolutions, where beads represent the actual thickness of the chromatin fiber, i.e., ~30 nm, stiff bonded and excluded-volume potentials are used. Since the bead sizes are shorter than the persistence length of chromatin, a bending potential is also required to obtain realistic fiber conformations.[93] Most parameters in these potentials can be fixed based on experimental estimates of chromatin thickness, elastic modulus, and persistence length.[93] At lower resolutions, where beads represent folded-up balls of the fiber, much softer bonded and excluded-volume potentials need to be used, and the bending potential becomes redundant.[97]

The *specific* interactions arising from chromatin looping and the confinement effects described earlier are the primary unknowns of the model. These interactions can be approximated by looping restraints between pairs of beads. Possible implementations of the restraints include harmonic spring potentials,[93, 101] a square-well potential,[94] or its smoother

differentiable form.[98] Then, the unknown parameters, to be determined by the optimization procedure, are the locations of the looping restraints along the fiber, together with the stiffness and equilibrium lengths of the harmonic restraints, or the depth and width of the square-well restraints. These restraints can be established across all pairs of beads, thus counting on the optimization algorithm to identify the true looping partners, while converging the strengths of unnecessary restraints to zero values.[94, 98] Alternatively, researchers have applied restraints only across those locus pairs exhibiting strong peaks in the contact maps.[93, 95] To further reduce the number of adjustable parameters, the equilibrium lengths of the harmonic restraints are set to zero,[93] as excluded-volume interactions can enforce a minimum distance even between strongly looped segments, and the widths of the square-well restraints are often set to a physically reasonable value.[99]

Conformational Sampling.

A key step in these methods involves generating at each optimization step an ensemble of bead-chain conformations consistent with its parameters at that step. In theory, a molecular system *at equilibrium* should exhibit conformations consistent with its thermodynamic state, or, more technically, its statistical-mechanical ensemble. In the canonical ensemble with fixed number of molecules, volume, and temperature, the probability ρ of observing a specific conformation $\mathbf{R}$ should follow the Boltzmann distribution $\rho(\mathbf{R}) \propto \exp(-U(\mathbf{R})/k_B T)$, where $U(\mathbf{R})$ is the total potential energy of the conformation, k_B is the Boltzmann constant, and T is the temperature. Whether chromosomes obey such a distribution or even represent an equilibrium system is debatable, but in the absence of any concrete alternatives, this distribution at least provides a useful starting point for obtaining physically-realistic conformations.[99] Ideally, to obtain the most realistic ensembles, U and T should be *quantitatively* described. This may however be unfeasible in the case of low-resolution models that use more ad-hoc energy potentials and temperature.[96, 97]

Various sampling methods have been used in the context of generating Boltzmann-distributed conformations of chromatin. Monte Carlo (MC) simulations provide a computationally efficient and flexible strategy for sampling conformations.[94, 95] In the most common implementation, the Metropolis-Hastings algorithm,[102] conformations are sampled through simple trial “moves”, such as translation of randomly chosen beads, which are accepted with a probability $P_{acc} = \min[1, \exp(-\Delta U/k_B T)]$, where ΔU is the change in the potential energy associated with each move. Such an approach, repeated over millions of trial moves, eventually yields the desired ensemble of conformations. However, this approach becomes inefficient for sampling the conformations of long polymer chains, especially those strongly confined and possessing looping restraints. More efficient sampling of chains has been achieved through biased regrowth of chains via geometric sequential importance sampling,[95] or through use of quaternions, instead of Euler angles, for implementing rotational moves on rigid clusters of beads within chains.[103] Molecular dynamics (MD) simulations provide an easily implementable approach for sampling conformations.[96] A number of freely-available MD simulation software such as LAMMPS[104] and GROMACS[105] work well with user-supplied interaction potentials. Other sampling methods such as Langevin dynamics[99] and Brownian dynamics simulations[93, 97] provide a good compromise between efficiency and usability.

The sampled conformations are used to compute the *predicted* interaction frequencies denoted by $\hat{f}_{ij}$. Though the spatial distance at which two segments of chromatin get crosslinked in 3C experiments is not known, it is reasonable to assume that crosslinking occurs at small distances on the order of the fiber diameter. By setting a reasonable distance threshold ξ for contact, typically 30 nm or smaller, $\hat{f}_{ij}$ calculation simply boils down to counting the fraction of conformations in the ensemble with $d_{ij,k} \leq \xi$, where $d_{ij,k}$ is the shortest of the distances between pairs of beads representing loci i and j in structure k. In other words, $\hat{f}_{ij} = \frac{1}{n}\sum_{k=1}^{n} \Theta(\xi - d_{ij,k})$, where Θ is the Heavyside step function and n is the size of the ensemble.[93, 94] Some approaches use the smoother sigmoidal function for counting contacts.[99] Two particularly attractive approaches for obtaining $\hat{f}_{ij}$ involves estimating them from distributions of inter-bead distances,[106] or their mean-square values,[97] which can both be more reliably measured from the ensemble than the direct counting method for loci exhibiting very small $\hat{f}_{ij}$.

Parameter Optimization.

The last component in the optimization loop adjusts the parameters $P^{(m)}$ of the polymer model (where m is the iteration number) based on differences in $\hat{f}_{ij}^{(m)}$ predicted from the simulated conformations and the 3C map f_{ij} to obtain a new set of parameters $P^{(m+1)}$ that should achieve better agreement between the two maps in the next iteration. The optimization stops when the difference between the maps becomes smaller than some specified tolerance, yielding the desired ensemble of conformations.

The original 3C or ChIP-chip techniques can probe interactions between only a handful of loci. If the interactions are sufficiently independent of each other, i.e., changing one interaction does not influence other interactions, then the imposed looping restraints can be optimized independently via simple rules that increase or decrease a restraint strength when its predicted interaction frequency is weaker or stronger than the experimental counterpart. For instance, one study optimized the stiffnesses k_{ij} of harmonic restraints via $k_{ij}^{(m+1)} = k_{ij}^{(m)}\left[\varepsilon(f_{ij}) / \varepsilon\left(\hat{f}_{ij}^{(m)}\right)\right]^{a}$, where ε is an inverse error function and a is a parameter that governs the accuracy and speed of convergence[101]. The chosen function not only ensured that k_{ij}'s were correctly up- or down-scaled based on the relative magnitudes of the predicted and experimental frequencies, but that strong restraints were adjusted more aggressively than the weaker ones.

More formal approaches are required for carrying out optimization of systems involving the large numbers of interactions probed by 5C or Hi-C experiments, where interactions are also more likely to be correlated. This requires defining an error function $S(\mathbf{k}) = \frac{1}{2}\|\widehat{\mathbf{F}}(\mathbf{k}) - \mathbf{F}\|^2$ that quantifies the difference between the predicted and experimental maps, where $\mathbf{k}$ is a vector of all imposed restraint stiffnesses k_{ij}. The objective is then to minimize this error under the condition that all $k_{ij} > 0$. This is most naturally achieved through a gradient-descent algorithm wherein the stiffnesses are updated via $\mathbf{k}^{(m+1)} = \mathbf{k}^{(m)} - h\nabla S^{(m)}/\|\nabla S^{(m)}\|$, where h is a step size and $\nabla S^{(m)}$ is the gradient of the scoring function with respect to each restraint

stiffness.[97] Alternatively, a reweighting scheme may be used to update restraint strengths, as done recently for optimizing the depths B_{ij} of square-well restraints.[94] Here, the *existing* ensemble of conformations is used for predicting how the existing set of interactions might change with small perturbations in B_{ij} via $\hat{f}_{ij}^{(m)} = \sum_{k=1}^{n} \Theta(\xi - d_{ij,k}^{(m)}) \exp(-\Delta E_k / k_B T) / \sum_{k=1}^{n} \exp(-\Delta E_k / k_B T)$, where ΔE_k is the change in the energy of all restraints in the k^{th} conformation in the ensemble due to the change in B_{ij} and Θ is the previously defined function for counting contacts. A Monte Carlo search in B_{ij}'s is carried out to minimize the deviation between the reweighted and experimental frequencies. These locally optimized parameters are then used in the next iteration to generate a new ensemble of conformations, and the process is repeated until convergence.

Instead of directly optimizing restraints, an alternative strategy involves optimizing the elements of a transformation matrix **W** that relates the interaction frequencies of the restrained pairs of beads to the strengths of their restraints via the linear relationship $\mathbf{k} = \mathbf{W}\mathbf{f}$.[93] The approach boils down to determining the elements of **W** that yield a set of restraint stiffnesses, such that the simulated ensemble of conformations using these stiffnesses reproduces the experimental interaction frequencies of the restrained loci. Within the optimization algorithm, this requires using the current elements of the matrix $\mathbf{W}^{(m)}$ to predict *two* sets of stiffnesses: one set is obtained from the experimental frequencies, via $\mathbf{k}^{(m)} = \mathbf{W}^{(m)}\mathbf{f}$, and the other set is obtained from the frequencies of the current ensemble, via $\hat{\mathbf{k}}^{(m)} = \mathbf{W}^{(m)}\hat{\mathbf{f}}^{(m)}$. The difference between the two stiffnesses $\boldsymbol{\varepsilon}^{(m)} = \mathbf{k}^{(m)} - \hat{\mathbf{k}}^{(m)}$ is then used to update the matrix via the least mean-square algorithm: $\mathbf{W}^{(m+1)} = \mathbf{W}^{(m)} + 2\mu_m \boldsymbol{\varepsilon}^{(m)} \cdot \hat{\mathbf{f}}^{(m)}$, where μ_m is a "gain factor" that governs the stability and speed of convergence. The above iteration is repeated until the matrix predicts similar stiffnesses from the predicted and experimental frequencies, implying that both maps are also similar. A key advantage of this approach is that the converged matrix provides an explicit relationship between restraints and frequencies, revealing how *each* restraint affects the interactions between *all* pairs of restrained loci. This allows one to distinguish true looping partners from "secondary" interactions that yield peaks in the contact map due to the looping of neighboring loci.

Researchers have also attempted to optimize restraint strengths using the maximum entropy principle.[98, 99] The underlying concept here is to impose the fewest or softest possible restraints to a "baseline" polymer model that still allows the restrained system to reproduce the experimental contact maps. If the potential energy of the baseline model is denoted by $U_0(\mathbf{R})$, then the potential energy of the optimally restrained system is given by $U(\mathbf{R}) = U_0(\mathbf{R}) + \sum_{i,j} \alpha_{ij} f_{\text{SW}}(r_{ij})$, where $f_{\text{SW}}(r)$ is a smooth square-well potential of a prescribed width defining contact distance and unit depth that simultaneously defines the potential energy of the restraint and counts interactions between loci i and j since $\hat{f}_{ij}$ is simply the ensemble average $\langle f_{\text{SW}}(r_{ij}) \rangle$. The parameters α_{ij} represent Lagrange multipliers that define the strengths of the restraints, whose values are obtained by maximizing the free energy associated with imposing the constraint that the predicted contact map must match the experimental map.[98] This is typically achieved using an optimization scheme involving cumulant expansion of the free energy. The maximum-entropy approach can be extended

further to include other kinds of experimental data into the model, including the local structure of chromatin and the association tendency between similar types of chromatin, i.e., loci exhibiting similar histone modifications.[99]

A few methods using polymer models do not strictly follow every aspect of the optimization scheme described in Fig. 3. For instance, in one approach, the simulated ensemble of conformations were required to satisfy only one-sided constraints arising from excluded volume interactions and nuclear confinement, without following any specific thermodynamic distribution.[96] Another approach used simulations of a confined polymer model without any looping restraints to model the Hi-C "background" contacts arising from non-specific interactions, i.e., random collisions between loci and confining obstacles.[95] The remaining contacts that were not accounted for by the unrestrained polymer model were assigned to specific looping interactions, which were then added to the model as distance constraints in the second stage of the optimization. A third approach used a polymer model with essentially no constraints, and generated Boltzmann-distributed conformations using a fictitious energy given by $U(\mathbf{R}) = \sum_{i,j} f_{ij} d_{ij}$, rather than the characteristic potential energy of polymer chains.[103]

MAXIMUM LIKELIHOOD AND BAYESIAN METHODS

The statistical methods known as maximum likelihood (ML) and Bayesian inference can be used to estimate the unknown parameters of a system by processing some data **D** with an appropriate statistical model. Whereas ML yields point estimates of the desired parameters, Bayesian inference yields their probability distribution and can also take advantage of prior knowledge or beliefs about the system under study. Both methods have been applied to the problem of inferring 3D conformations of chromatin from contact maps. In this context, the data **D** are obtained from 3C experiments, and the unknown parameters include the chromatin conformation **R** and the parameters $\boldsymbol{\theta}$ of a statistical model that describes the production of the data **D.** Common to both methods is the construction of a likelihood function, denoted by $P(\mathbf{D}|\mathbf{R},\boldsymbol{\theta})$, which describes the probability of observing the data given the parameters. ML involves maximizing $P(\mathbf{D}|\mathbf{R},\boldsymbol{\theta})$ to obtain the best estimates of **R** and $\boldsymbol{\theta}$, thus yielding a "consensus" structure $\mathbf{R_0}$ consistent with the available data **D.** On the other hand, Bayesian inference also requires defining the prior distribution of the parameters, $P(\mathbf{R},\boldsymbol{\theta})$, and sampling the posterior distribution, $P(\mathbf{R},\boldsymbol{\theta}|\mathbf{D})$, thus yielding an ensemble of structures **R** consistent with the underlying population (Fig. 4).

Likelihood function.

Formulating the likelihood function $P(\mathbf{D}|\mathbf{R},\boldsymbol{\theta})$ requires choosing (i) a representation for the system, (ii) the form of the observed data, and (iii) a statistical model describing the behavior of the system. To represent the 3D conformation of a chromosome, various formulations have been adopted. A simple representation involves points or beads corresponding to the restriction fragments generated by 5C or Hi-C experiments.[56,107] Another natural choice is a chain of points or beads matching the genomic segments used to define the bins of the contact matrix. Such segments are typically of equal length,[71, 108-112] but can also be defined by hierarchical clustering of the contact matrix[113] or by matching the

extents of TADs.[108, 114] The volume of each bead may be proportional to the length of the genomic segment represented by the bead.[113] Actual bead diameters can be determined from values of chromatin density in the nucleus.[111] Also, each bead may have both a hard-core radius, to enforce excluded volume, and a larger soft-core radius, to detect contacts between beads.[113] Although chromosomes are suitably represented as continuous chains of beads, contact matrices often include gaps due to poor mappability of reads. Beads that fall in those gaps are omitted from some optimization procedures.[71, 109, 110]

The second requirement of the likelihood function is an appropriate form for the data **D**. A possible choice is the matrix **C** of contact counts c_{ij}.[56, 71, 108-111, 115] In some cases, matrices from multiple experiments are used simultaneously, allowing integration of data from different restriction enzymes.[115] Using contact counts from Hi-C experiments requires some care in regard to experimental biases.[54] These can be ignored for simplicity,[56] or can be explicitly modeled by including appropriate covariates in the likelihood function.[108-110, 115] Alternatively,[71, 112, 113] the contact matrix may be corrected using published procedures[54, 116-118] before computing the likelihood function. Further conversion of contact counts c_{ij} to frequencies f_{ij} may be required for some likelihood formulations.[56, 113] The likelihood function can also be defined using data **D** in the form of distances δ_{ij} between pairs of loci i and j.[107, 112] These distances may be obtained from c_{ij} by assuming $\delta_{ij} = \gamma c_{ij}^{-\alpha}$. Here, γ can be set arbitrarily[112] or estimated from published data, such as the average spatial distance between genomic loci,[107] and α can be treated as an unknown parameter.[107] In one study, α was varied within a specified range and, to choose the best value, the inferred structures were compared to a known structure using the Spearman correlation coefficient of the internal distances.[112] The likelihood function may also rely on data from other types of experiments. For example, FISH measurements have been used to constrain the radius of gyration of the inferred chromosome structure.[111]

The last requirement is a statistical model that yields the probability of observing each data point. For example, each contact count c_{ij}, may be assumed to obey a binomial distribution, which in turn is approximated by a normal distribution with an unknown mean μ_{ij} and standard deviation σ_{ij} equal to the mean plus a small constant.[56] On the other hand, the discrete nature of c_{ij} suggests they be modeled as independent Poisson random variables, thus yielding

$$P(\mathbf{D} \mid \mathbf{R}, \theta) = \prod_{1 \le i < j \le N} \frac{e^{-\mu_{ij}} \mu_{ij}^{c_{ij}}}{c_{ij}!} \tag{4}$$

where each mean contact count μ_{ij} is in turn related to the spatial distance d_{ij} between the interacting loci i and j. The relation is an inverse power law, $\log \mu_{ij} = \alpha_0 + \alpha_1 \log d_{ij}$, where $\alpha_1 < 0$, and $\alpha_0 > 0$ determines the scale of the 3D structure.[56, 71, 108] Because such scale cannot be derived from contact counts alone, the value of α_0 may be set arbitrarily,[56, 109, 110] deduced by assuming $d_{1,n} = 1$,[108] or inferred using non-metric MDS.[71] The definition of mean contact counts μ_{ij} can also be refined to include experimental biases inherent in the Hi-C data. In this case $\log \mu_{ij} = \alpha_0 + \alpha_1 \log d_{ij} + \mathbf{v}_{ij}^T \beta$, where $\mathbf{v}_{ij}$ is a vector of covariates quantifying the biases, and β is a vector of corresponding coefficients.[108, 109, 115]

Although convenient for modeling contact counts, the Poisson distribution may not be adequate for contact maps containing many zero entries. To overcome this issue, the use of a zero-truncated Poisson model was proposed,[109, 110] where the log likelihood is computed by excluding the terms associated with zero c_{ij}'s. Another potential problem is that the c_{ij}'s may not be truly independent, because neighboring loci along a chromosome may form similar contacts with farther loci. Moreover, the actual variance of the c_{ij}'s may be larger than allowed by the Poisson distribution. These problems can be addressed by including additional random variables in log μ_{ij} that account for variance over-dispersion and interdependency of contact counts.[110]

Instead of using contacts c_{ij} in the likelihood function, one can use distances $\delta_{ij} = \gamma c_{ij}^{-\alpha}$, which can be assumed to follow a normal distribution with mean d_{ij} and variance σ[2107, 112]:

$$P(\mathbf{D} \mid \mathbf{R}, \theta) = \prod_{1 \le i < j \le N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\delta_{ij} - d_{ij})^2\right\} \quad (5)$$

The parameters α and σ can be inferred together with the distances d_{ij}.[107] Another option is to keep α constant and to eliminate the variance from Eq. 5 by assuming that $\sigma^2 \propto \sum(\delta_{ij} - d_{ij})^2$.[112] Although the above examples of likelihood functions are generally applicable to data from bulk Hi-C experiments, analysis of data from single-cell Hi-C experiments requires different statistical models. One approach is to express the log-likelihood as a sum of logistic functions of d_{ij}, thus introducing adjustable parameters for the contact distance and the steepness of the step.[111]

The likelihood function can be pushed even further. For example, the need for an explicit relation between c_{ij} and d_{ij} can be avoided, and a population of conformations for an entire diploid genome can be inferred at once via a single likelihood maximization. These ambitious requirements were met by expressing the likelihood function in terms of two large matrices: a $2N \times 2N \times M$ matrix **R** containing the coordinates of the beads for all diploid genome structures in the population, and a $2N \times 2N \times M$ binary matrix **W** assigning contacts to pairs of beads that overlap within each estimated structure, where N is the number of beads per haploid genome and M is the population size.[113] Estimating M genomic structures simultaneously also allows estimation of interaction frequencies, which are compared directly to the contact map **F**. The resulting likelihood function is thus $P(\mathbf{F}|\mathbf{R}) = P(\mathbf{F}|\mathbf{W})P(\mathbf{W}|\mathbf{R})$, i.e., a product of the probability of observing the assigned contacts **W** given the estimated structures **R**, and the probability of observing the Hi-C interaction frequencies **F** given the contact assignments **W**.

Likelihood maximization.

Having defined the likelihood function, one can proceed to determine a consensus structure $\mathbf{R_0}$ that recapitulates the observed data **D**. Assuming such structure to be the most probable given the data, then the task is to find **R** and $\boldsymbol{\theta}$ that maximize $P(\mathbf{D}|\mathbf{R},\boldsymbol{\theta})$. To achieve this goal, several optimization techniques can be employed with various tradeoffs in computational efficiency and reliability. A sufficiently simple likelihood function that depends only on **R** can be maximized using the gradient ascent method or the adaptive gradient algorithm.[112]

More complex scenarios can be handled with appropriate iteration schemes. For instance, optimization of both **R** and the parameters a_0 and a_1 can be performed through a coordinate-descent algorithm that randomly initializes **R** and then alternates between (i) maximizing the likelihood with respect to a_0 and a_1 with a fixed conformation **R**, and (ii) maximizing the likelihood with respect to **R** while keeping a_0 and a_1 fixed.[71] The individual optimization problems can be solved by using the interior point filter algorithm implemented in the IPOPT code.[84]

Even more challenging functions require iteration at multiple levels. For example, maximizing the likelihood function $P(\mathbf{F}|\mathbf{R}) = P(\mathbf{F}|\mathbf{W})P(\mathbf{W}|\mathbf{R})$, used in Ref. [113], entails optimizing a structure matrix **R** and a contact assignment matrix **W** that are both very large. Here, one iteration level involves two steps: (i) updating the contact assignments in **W** by maximizing $P(\mathbf{F}|\mathbf{W})$, and (ii) estimating **R** by maximizing $P(\mathbf{W}|\mathbf{R})$.[113] Specifically, $P(\mathbf{F}|\mathbf{W})$ is maximized by comparing inter-bead distances to appropriate thresholds determined from **R**, while $P(\mathbf{W}|\mathbf{R})$ is maximized through simulated annealing and conjugate gradient algorithms in IMP.[119] Another iteration level takes advantage of the idea that enforcing frequent contacts before infrequent ones can efficiently guide the search for an optimal structure. Thus, the above procedure is repeated by incrementally populating the matrix **F** with sets of contact probabilities arranged from largest to smallest.[113]

One advantage of using ML is that additional constraints can be introduced without complicating the likelihood function. For example, the excluded volume of beads, the confinement of beads within the nuclear volume, and the distance of certain beads from the nuclear periphery based on FISH data can all be enforced by using constrained optimization methods.[113]

Bayesian inference.

An alternative to ML is Bayesian inference, which explicitly recognizes the existence of probability distributions for chromatin structures and auxiliary parameters. This approach, named inferential structure determination, was previously shown to be effective in obtaining the structure of macromolecules from nuclear magnetic resonance data.[120] The general scheme is derived from Bayes' theorem to yield the posterior distribution of conformation **R** and parameters $\boldsymbol{\theta}$ given the observed data **D**, i.e., $P(\mathbf{R},\boldsymbol{\theta}|\mathbf{D}) = P(\mathbf{D}|\mathbf{R},\boldsymbol{\theta})P(\mathbf{R},\boldsymbol{\theta})/P(\mathbf{D})$, where the now familiar likelihood function $P(\mathbf{D}|\mathbf{R},\boldsymbol{\theta})$ is multiplied by $P(\mathbf{R},\boldsymbol{\theta})$, which is the prior distribution of **R** and $\boldsymbol{\theta}$ based on some assumed behavior of **R**. The normalizing constant $P(\mathbf{D})$ ensures that $0 \leq P(\mathbf{R},\boldsymbol{\theta}|\mathbf{D}) \leq 1$, but is usually not needed to estimate **R** and $\boldsymbol{\theta}$ from $P(\mathbf{R},\boldsymbol{\theta}|\mathbf{D})$. Thus, to evaluate the posterior distribution $P(\mathbf{R},\boldsymbol{\theta}|\mathbf{D})$, it is only necessary to define an appropriate prior distribution $P(\boldsymbol{R},\boldsymbol{\theta})$. This requirement can be sidestepped by assuming a uniform distribution, i.e., a non-informative prior, so that $P(\mathbf{R},\boldsymbol{\theta}|\mathbf{D}) \propto P(\mathbf{D}|\mathbf{R},\boldsymbol{\theta})$,[56, 108-110] or one can take advantage of the prior to obtain conformations **R** that are more physically realistic. For example, the dependency between the 3D positions of neighboring genomic loci can be modeled with a normal distribution, thus yielding the prior $P(\mathbf{R},\theta) \propto \prod_{i \geq 2} \exp\left\{-\lambda(l_i - l_{i-1})d^2_{l_i,l_{i-1}}\right\}$, where the l_i's are integer indices of the genomic loci for which contact counts are available, and the tunable parameter λ determines the smoothness of the chain.[115] Another option is to assume that the potential energy U of a

conformation obeys the Boltzmann distribution,[107, 111] yielding $P(\mathbf{R},\boldsymbol{\theta}) \propto \exp(-U(\mathbf{R},\boldsymbol{\theta})/k_B T)$. Such a formulation affords great flexibility in choosing relevant physical properties of chromatin. For instance, U may include potentials for stretching, bending, and excluded volume, using physical parameters $\boldsymbol{\theta}$ similar to those used in polymer models.[107, 111]

Once the likelihood function and priors are defined, the posterior distribution $P(\mathbf{R},\boldsymbol{\theta}|\boldsymbol{D})$ can be sampled to infer conformations **R** and model parameters $\boldsymbol{\theta}$ from the observed data **D.** The purpose of sampling may be to find a single consensus structure that maximizes the posterior,[108-110, 115] or to obtain an ensemble of conformations that can be further studied, e.g., via clustering methods.[56,107] A general approach for sampling the posterior distribution is to employ Markov chain Monte Carlo (MCMC) with the Metropolis-Hastings algorithm, [102] which draws conformations from the posterior without requiring evaluation of $P(\mathbf{D})$. Specific refinements of this strategy may be required based on the complexity of the posterior. When the latter lacks adjustable model parameters, MCMC may be adequate by itself to produce a conformation ensemble.[56] In this case, starting with a random initial structure, each MC step generates a proposal structure by randomly displacing a randomly chosen point in the chain.

The presence of unknown model parameters requires more elaborate schemes. A possible solution is to use Gibbs sampling,[121] which alternates between chromosome structures and model parameters.[108-111] In regards to chromosome structures, the Hamiltonian, or Hybrid, MC method[122] can be used to efficiently sample the posterior distribution of **R** while keeping $\boldsymbol{\theta}$ fixed.[108-111] In this case, each MC step produces a proposal conformation by performing a short MD simulation, where the 3D coordinates are updated by numerical integration.[109, 110] Complicated posteriors may lead the Gibbs sampler to become trapped in local peaks. A smart solution to this problem is to combine he Gibbs sampler with replica exchange MC,[123] where a fictitious inverse temperature determines the weight of the likelihood function on the posterior distribution.[111, 124] Additional improvements are possible: the model parameters can be initialized using Poisson regression and then refined using adaptive rejection sampling.[108] Moreover, the initial conformations **R** can be obtained using sequential importance sampling with a rejection control technique that improves efficiency.[108]

Besides the Gibbs sampler, other schemes have been proposed. For example, a posterior that includes a distance penalty to enforce the connectivity of the chain[115] can be maximized by iterating over two steps: (i) fitting a generalized linear model (GLM) obtained from the log likelihood function by omitting the terms for α_1 and the distance penalty, and (ii) minimizing the distance penalty by adjusting groups of sequential coordinates to obtain an initial structure, followed by updating the GLM coefficients through simulated annealing with Hamiltonian dynamics.[115] Another iterative scheme uses the expectation maximization algorithm[125] to estimate the model parameters while also generating an ensemble of conformations.[107] After initializing the model parameters and generating an initial ensemble of structures through Brownian dynamics, the computation alternates between two steps. In the expectation step, a gradient ascent algorithm is used to refine each structure by maximizing its likelihood, which is calculated from the posterior using the current estimates of model parameters. In the maximization step, a grid search is performed to estimate the

model parameters that maximize the likelihood of the ensemble of conformations refined in the previous step.[107]

CONCLUSION

In this review, we have provided an overview of available methods for converting 2D contact data from Hi-C experiments into 3D chromatin conformations that are consistent with such data. We have found it expedient to group such methods into three classes. The first class includes methods that covert contact frequencies into internal distances and feed the latter to a scoring function whose optimization yields a single consensus structure. The second class includes methods that avoid the conversion from contact frequencies to internal distances and instead use polymer models to obtain conformation ensembles that recapitulate the experimental contact frequencies. The third class includes methods that relate the contact frequencies to internal distances through a statistical model, whose parameters are then optimized to agree with the contact frequencies.

Each class of methods includes a variety of techniques and refinements that enable structural recovery at different scales, ranging from domains, to chromosomes, to whole diploid genomes. The large number of proposed methods creates opportunities for further research and improvement. For example, choosing the method most appropriate for a given biological question will benefit from efforts to assess objectively the strengths and weaknesses of the available methods.[126] Also, careful evaluation and comparison of present and future methods will benefit from the availability of standardized test cases where the solution structure or ensemble is known in advance. This practice is well established in the Critical Assessment of Structure Prediction (CASP) experiments, which are periodically performed to track the progress of computational methods for predicting protein structure from amino acid sequence.[127] However, CASP relies on experimental data that are currently unavailable to provide the ground truth for chromatin conformation inference, which is therefore more difficult to validate than protein structure prediction. The variety of available methods for 3D genome structure determination and their likely complementary strengths and weaknesses also suggest the possibility to apply such methods simultaneously in order to obtain a consensus solution based on suitable criteria. Similar strategies have been proposed, for example, to improve the reliability of protein-ligand predictions,[128, 129] protein structure alignments,[130] protein structure comparison,[131] and protein secondary structure prediction.[132-134]

There are also opportunities for further improvement of the current methods and their specific implementations. For example, a major area of concern is computational efficiency, especially when attempting to reconstruct the structure of whole diploid genomes at high resolution. Execution speed and complexity of the reconstructed structures may both increase by exploiting high-performance hardware, such as large computer clusters and GPUs, which have already been used for 3D genome visualization.[77] Efficiency could also be improved through the use of multi-resolution models, where increasingly refined models of chromatin are threaded through structural solutions obtained from lower-resolution models and then locally optimized at higher resolution. Such approaches have found applications in building high-resolution models of bacterial genomes,[82, 135, 136] and similar

ideas could be applied to obtaining refined structures of eukaryotic genomes using available high-resolution models of chromatin.[137-140] Also, the accuracy of the reconstructed structures may be improved by deriving constraints from additional experimental techniques such as super-resolution microscopy and soft X-ray tomography.[141] This trend has already begun with the integration of epigenetic data from ChIP-seq and DNase-seq experiments toward the prediction of chromatin conformation.[142-144] Lastly, it remains to be explored whether the reconstruction of 3D genomic structure from experimental data can be improved by taking advantage of increasingly popular, data-greedy machine learning algorithms. For example, deep neural networks[145] have been applied to predict transcription factors binding sites,[146] protein secondary structure,[147] and protein-protein interactions.[148] Indeed, as the quantity of experimental data continues to grow, such strategies are already finding their way into predicting 3D chromatin architecture.[149]

ACKNOWLEDGMENTS

D.M. was supported by grants from the National Institutes of Health (5F32DK112682 and 1K01DK119687).

REFERENCES

1. Alberts B et al. Molecular Biology of the Cell, Sixth Edition Molecular Biology of the Cell, Sixth *Edition*, 1–1342 (2015).
2. Parmar JJ, Woringer M & Zimmer C How the Genome Folds: The Biophysics of Four-Dimensional Chromatin Organization. Annual Review of Biophysics 48, null (2019).
3. Szalaj P & Plewczynski D Three-dimensional organization and dynamics of the genome. Cell Biology and Toxicology 34, 381–404 (2018). [PubMed: 29568981]
4. Bickmore WA in Annual Review of Genomics and Human Genetics, Vol 14, Vol. 14. (eds. Chakravarti A & Green E) 67–84 (Annual Reviews, Palo Alto; 2013).
5. Dekker J et al. The 4D nucleome project. Nature 549, 219–226 (2017). [PubMed: 28905911]
6. Zhang Y et al. Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations. Cell 148, 908–921 (2012). [PubMed: 22341456]
7. Akdemir KC, Chin L, Futreal A & Grp IPSA Spatial organization of the genome and genomic alterations in human cancers. Human Genomics 10, 1 (2016). [PubMed: 26744305]
8. Smith KS, Liu LL, Ganesan S, Michor F & De S Nuclear topology modulates the mutational landscapes of cancer genomes. Nature Structural & Molecular Biology 24, 1000-+ (2017).
9. Zhang M, Wang FC, Kou ZH, Zhang Y & Gao SR Defective Chromatin Structure in Somatic Cell Cloned Mouse Embryos. Journal of Biological Chemistry 284, 24981–24987 (2009).
10. Cuartero S & Merkenschlager M Three-dimensional genome organization in normal and malignant haematopoiesis. Current Opinion in Hematology 25, 323–328 (2018). [PubMed: 29702522]
11. Ausio J, de Paz AM & Esteller M MeCP2: the long trip from a chromatin protein to neurological disorders. Trends in Molecular Medicine 20, 487–498 (2014). [PubMed: 24766768]
12. Iwase S & Martin DM Chromatin in nervous system development and disease. Molecular and Cellular Neuroscience 87, 1–3 (2018). [PubMed: 29248671]
13. Elisa Z et al. Technical implementations of light sheet microscopy. Microscopy Research and Technique 81, 941–958 (2018). [PubMed: 29322581]
14. Girkin JM & Carvalho MT The light-sheet microscopy revolution. Journal of Optics 20, 20 (2018).
15. Hauser M et al. Correlative Super-Resolution Microscopy: New Dimensions and New Opportunities. Chemical Reviews 117, 7428–7456 (2017). [PubMed: 28045508]
16. Reuter JA, Spacek DV & Snyder MP High-Throughput Sequencing Technologies. Molecular Cell 58, 586–597 (2015). [PubMed: 26000844]
17. Dogan ES & Liu C Three-dimensional chromatin packing and positioning of plant genomes. Nature Plants 4, 521–529 (2018). [PubMed: 30061747]

18. Fazary AE, Ju YH & Abd-Rabboh HSM How does chromatin package DNA within nucleus and regulate gene expression? International Journal of Biological Macromolecules 101, 862–881 (2017). [PubMed: 28366861]
19. Maston GA, Evans SK & Green MR in Annual Review of Genomics and Human Genetics, Vol. 7 29–59 (Annual Reviews, Palo Alto; 2006).
20. Jhunjhunwala S, van Zelm MC, Peak MM & Murre C Chromatin Architecture and the Generation of Antigen Receptor Diversity. Cell 138, 435–448 (2009). [PubMed: 19665968]
21. Ebert A, Hill L & Busslinger M in Molecular Mechanisms That Orchestrate the Assembly of Antigen Receptor Loci, Vol. 128 (ed. Murre C) 93–121 (Elsevier Academic Press Inc, San Diego; 2015).
22. Dixon JR, Gorkin DU & Ren B Chromatin Domains: The Unit of Chromosome Organization. Molecular Cell 62, 668–680 (2016). [PubMed: 27259200]
23. Gonzalez-Sandoval A & Gasser SM On TADs and LADs: Spatial Control Over Gene Expression. Trends in Genetics 32, 485–495 (2016). [PubMed: 27312344]
24. Dekker J & Heard E Structural and functional diversity of Topologically Associating Domains. Febs Letters 589, 2877–2884 (2015). [PubMed: 26348399]
25. Lieberman-Aiden E et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009). [PubMed: 19815776]
26. Wang JY & Jia ST New Insights into the Regulation of Heterochromatin. Trends in Genetics 32, 284–294 (2016). [PubMed: 27005444]
27. Dixon JR et al. Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–336 (2015). [PubMed: 25693564]
28. Cremer T & Cremer M Chromosome Territories. Cold Spring Harbor Perspectives in Biology 2 (2010).
29. Rouquette J, Cremer C, Cremer T & Fakan S in International Review of Cell and Molecular Biology, Vol 282, Vol. 282. (ed. Jeon KW) 1–90 (Elsevier Academic Press Inc, San Diego; 2010). [PubMed: 20630466]
30. Wang SY et al. Spatial organization of chromatin domains and compartments in single chromosomes. Science 353, 598–602 (2016). [PubMed: 27445307]
31. Cremer C, Szczurek A, Schock F, Gourram A & Birk U Super-resolution microscopy approaches to nuclear nanostructure imaging. Methods 123, 11–32 (2017). [PubMed: 28390838]
32. Boettiger AN et al. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature 529, 418-+ (2016). [PubMed: 26760202]
33. Sati S & Cavalli G Chromosome conformation capture technologies and their impact in understanding genome function. Chromosoma 126, 33–44 (2017). [PubMed: 27130552]
34. Barutcu AR et al. C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. Journal of Cellular Physiology 231, 31–35 (2016). [PubMed: 26059817]
35. Dekker J, Rippe K, Dekker M & Kleckner N Capturing chromosome conformation. science 295, 1306–1311 (2002). [PubMed: 11847345]
36. Belton JM et al. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods 58, 268–276 (2012). [PubMed: 22652625]
37. Belaghzal H, Dekker J & Gibcus JH Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods 123, 56–65 (2017). [PubMed: 28435001]
38. Lin DJ, Bonora G, Yardimci GG & Noble WS Computational methods for analyzing and modeling genome structure and organization. Wiley Interdisciplinary Reviews-Systems Biology and Medicine 11, 14 (2019).
39. Bianco S, Chiariello AM, Annunziatella C, Esposito A & Nicodemi M Predicting chromatin architecture from models of polymer physics. Chromosome Research 25, 25–34 (2017). [PubMed: 28070687]
40. Zhang B & Wolynes PG Genomic Energy Landscapes. Biophysical Journal 112, 427–433 (2017). [PubMed: 27692923]

41. Tiana G & Giorgetti L Integrating experiment, theory and simulation to determine the structure and dynamics of mammalian chromosomes. Current Opinion in Structural Biology 49, 11–17 (2018). [PubMed: 29128709]
42. Le Dily F, Serra F & Marti-Renom MA 3D modeling of chromatin structure: is there a way to integrate and reconcile single cell and population experimental data? Wiley Interdiscip. Rev.-Comput. Mol. Sci 7, 13 (2017).
43. Serra F et al. Restraint-based three-dimensional modeling of genomes and genomic domains. FEBS Lett. 589, 2987–2995 (2015). [PubMed: 25980604]
44. Rosa A & Zimmer C in New Models of the Cell Nucleus: Crowding, Entropic Forces, Phase Separation, and Fractals, Vol. 307 (eds. Hancock R & Jeon KW) 275–349 (Elsevier Academic Press Inc, San Diego; 2014).
45. Lesne A, Riposo J, Roger P, Cournac A & Mozziconacci J 3D genome reconstruction from chromosomal contacts. Nat. Methods 11, 1141–1143 (2014). [PubMed: 25240436]
46. Lajoie BR, Dekker J & Kaplan N The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods 72, 65–75 (2015). [PubMed: 25448293]
47. Forcato M et al. Comparison of computational methods for Hi-C data analysis. Nature Methods 14, 679-+ (2017). [PubMed: 28604721]
48. Sajan SA & Hawkins RD in Annual Review of Genomics and Human Genetics, Vol 13, Vol. 13. (eds. Chakravarti A & Green E) 59–82 (Annual Reviews, Palo Alto; 2012).
49. Dekker J, Marti-Renom MA & Mirny LA Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. 14, 390–403 (2013). [PubMed: 23657480]
50. Schmitt AD, Hu M & Ren B Genome-wide mapping and analysis of chromosome architecture. Nature Reviews Molecular Cell Biology 17, 743–755 (2016). [PubMed: 27580841]
51. Nicoletti C, Forcato M & Bicciato S Computational methods for analyzing genome-wide chromosome conformation capture data. Current Opinion in Biotechnology 54, 98–105 (2018). [PubMed: 29550705]
52. Xu C & Corces VG Towards a predictive model of chromatin 3D organization. Seminars in Cell & Developmental Biology 57, 24–30 (2016). [PubMed: 26658098]
53. Rao SSP et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014). [PubMed: 25497547]
54. Yaffe E & Tanay A Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics 43, 1059–U1040 (2011). [PubMed: 22001755]
55. Ferraiuolo MA, Sanyal A, Naumova N, Dekker J & Dostie J From cells to chromatin: Capturing snapshots of genome organization with 5C technology. Methods 58, 255–267 (2012). [PubMed: 23137922]
56. Rousseau M, Fraser J, Ferraiuolo MA, Dostie J & Blanchette M Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinformatics 12, 16 (2011). [PubMed: 21226914]
57. De Gennes P-G & Gennes P-G Scaling concepts in polymer physics. (Cornell university press, 1979).
58. Rubinstein M & Colby RH Polymer physics, Vol. 23 (Oxford university press New York, 2003).
59. Fraser J et al. Chromatin conformation signatures of cellular differentiation. Genome Biol. 10 (2009).
60. Peng C et al. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. Nucleic Acids Res. 41, 11 (2013).
61. Ay F et al. Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. 24, 974–988 (2014). [PubMed: 24671853]
62. Zhang Z, Li G, Toh K-C & Sung W-K 3D chromosome modeling with semi-definite programming and Hi-C data. J. Comput. Biol. 20, 831–846 (2013). [PubMed: 24195706]

63. Duan Z et al. A three-dimensional model of the yeast genome. Nature 465, 363–367 (2010). [PubMed: 20436457]
64. Tanizawa H et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res. 38, 8164–8177 (2010). [PubMed: 21030438]
65. Adhikari B, Trieu T & Cheng JL Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. BMC Genomics 17, 9 (2016). [PubMed: 26819243]
66. Bau D et al. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nature Structural & Molecular Biology 18, 107-+ (2011).
67. Umbarger MA et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. Molecular cell 44, 252–264 (2011). [PubMed: 22017872]
68. Xie WJ et al. Structural modeling of chromatin integrates genome features and reveals chromosome folding principle. Sci Rep 7, 2818 (2017). [PubMed: 28588240]
69. Zhu GX et al. Reconstructing spatial organizations of chromosomes through manifold learning. Nucleic Acids Res. 46, 15 (2018).
70. Rieber L & Mahony S miniMDS: 3D structural inference from high-resolution Hi-C data. Bioinformatics 33, I261–I266 (2017). [PubMed: 28882003]
71. Varoquaux N, Ay F, Noble WS & Vert JP A statistical approach for inferring the 3D structure of the genome. Bioinformatics 30, 26–33 (2014).
72. Nagano T et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59 (2013). [PubMed: 24067610]
73. Stevens TJ et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59-+ (2017). [PubMed: 28289288]
74. Trieu T & Cheng JL MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. Bioinformatics 32, 1286–1292 (2016). [PubMed: 26722115]
75. Trieu T & Cheng JL 3D genome structure modeling by Lorentzian objective function. Nucleic Acids Res. 45, 1049–1058 (2017). [PubMed: 28180292]
76. Paulsen J et al. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. Genome Biol. 18, 15 (2017). [PubMed: 28118844]
77. Szalaj P et al. An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization. Genome Res. 26, 1697–1709 (2016). [PubMed: 27789526]
78. Li J, Zhang W & Li X 3D genome reconstruction with ShRec3D+ and Hi-C data. IEEE/ACM transactions on computational biology and bioinformatics 15, 460–468 (2018). [PubMed: 26955049]
79. Segal MR & Bengtsson HL Reconstruction of 3D genome architecture via a two-stage algorithm. BMC Bioinformatics 16, 10 (2015). [PubMed: 25592313]
80. Liu T & Wang Z Reconstructing high-resolution chromosome three-dimensional structures by hi-C complex networks. BMC Bioinformatics 19, 496 (2018). [PubMed: 30591009]
81. Shavit Y, Hamey FK & Lio P FisHiCal: an R package for iterative FISH-based calibration of Hi-C data. Bioinformatics 30, 3120–3122 (2014). [PubMed: 25061071]
82. Yildirim A & Feig M High-resolution 3D models of Caulobacter crescentus chromosome reveal genome structural variability and organization. Nucleic Acids Res. 46, 3937–3952 (2018). [PubMed: 29529244]
83. Ben-Elazar S, Yakhini Z & Yanai I Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. Nucleic Acids Res. 41, 2191–2201 (2013). [PubMed: 23303780]
84. Wächter A & Biegler LT On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Mathematical Programming 106, 25–57 (2006).
85. Ingber L Simulated annealing: Practice versus theory. Mathematical and computer modelling 18, 29–57 (1993).
86. Paulsen J, Gramstad O & Collas P Manifold based optimization for single-cell 3D genome reconstruction. PLoS Comput. Biol. 11, e1004396 (2015). [PubMed: 26262780]

87. Borg I & Groenen P Modern multidimensional scaling: Theory and applications. Journal of Educational Measurement 40, 277–280 (2003).
88. Ji X & Zha H in IEEE INFOCOM 2004, Vol. 4 2652–2661 (IEEE, 2004).
89. Glunt W, Hayden TL & Raydan M Molecular conformations from distance matrices. Journal of Computational Chemistry 14, 114–120 (1993).
90. De Leeuw J & Mair P Multidimensional scaling using majorization: SMACOF in R. (2011).
91. Szalaj P et al. 3D-GNOME: an integrated web service for structural modeling of the 3D genome. Nucleic Acids Res. 44, W288–W293 (2016). [PubMed: 27185892]
92. Hirata Y, Oda A, Ohta K & Aihara K Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. Sci Rep 6, 9 (2016). [PubMed: 28442706]
93. Meluzzi D & Arya G Recovering ensembles of chromatin conformations from contact probabilities. Nucleic Acids Res. 41, 63–75 (2012). [PubMed: 23143266]
94. Giorgetti L et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell 157, 950–963 (2014). [PubMed: 24813616]
95. Gursoy G, Xu Y, Kenter AL & Liang J Computational construction of 3D chromatin ensembles and prediction of functional interactions of alpha-globin locus from 5C data. Nucleic Acids Res. 45, 11547–11558 (2017). [PubMed: 28981716]
96. Kalhor R, Tjong H, Jayathilaka N, Alber F & Chen L Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nature biotechnology 30, 90 (2012).
97. Le Treut G, Kepes F & Orland H A Polymer Model for the Quantitative Reconstruction of Chromosome Architecture from HiC and GAM Data. Biophys. J. 115, 2286–2294 (2018). [PubMed: 30527448]
98. Di Pierro M, Cheng RR, Aiden EL, Wolynes PG & Onuchic JN De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. Proceedings of the National Academy of Sciences 114, 12126–12131 (2017).
99. Zhang B & Wolynes PG Topology, structures, and energy landscapes of human chromosomes. Proceedings of the National Academy of Sciences 112, 6062–6067 (2015).
100. Di Pierro M, Zhang B, Aiden EL, Wolynes PG & Onuchic JN Transferable model for chromosome architecture. Proceedings of the National Academy of Sciences 113, 12168–12173 (2016).
101. Junier I, Dale RK, Hou C, Képès F & Dean A CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the β-globin locus. Nucleic Acids Res. 40, 7718–7727 (2012). [PubMed: 22705794]
102. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 21, 1087–1092 (1953).
103. Caudai C, Salerno E, Zoppe M & Tonazzini A Inferring 3D chromatin structure using a multiscale approach based on quaternions. BMC Bioinformatics 16, 11 (2015). [PubMed: 25592753]
104. Plimpton S Fast parallel algorithms for short-range molecular dynamics. Journal of computational physics 117, 1–19 (1995).
105. Van Der Spoel D et al. GROMACS: fast, flexible, and free. Journal of computational chemistry 26, 1701–1718 (2005). [PubMed: 16211538]
106. Meluzzi D & Arya G Efficient estimation of contact probabilities from inter-bead distance distributions in simulated polymer chains. J. Phys.-Condes. Matter 27, 12 (2015).
107. Wang SY, Xu JB & Zeng JY Inferential modeling of 3D chromatin structure. Nucleic Acids Res. 43, 12 (2015).
108. Hu M et al. Bayesian Inference of Spatial Organizations of Chromosomes. PLoS Comput. Biol. 9, 14 (2013).
109. Park J & Lin S 245–261 (Springer International Publishing, Cham; 2015).
110. Park J & Lin SL Impact of data resolution on three-dimensional structure inference methods. BMC Bioinformatics 17, 13 (2016). [PubMed: 26823083]

111. Carstens S, Nilges M & Habeck M Inferential Structure Determination of Chromosomes from Single-Cell Hi-C Data. PLoS Comput. Biol. 12, 33 (2016).
112. Oluwadare O, Zhang YX & Cheng JL. A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. BMC Genomics 19, 17 (2018). [PubMed: 29301490]
113. Tjong H et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc. Natl. Acad. Sci. U. S. A. 113, E1663–E1672 (2016). [PubMed: 26951677]
114. Hua N et al. Producing genome structure populations with the dynamic and automated PGS software. Nat. Protoc. 13, 915–926 (2018). [PubMed: 29622804]
115. Zou CC, Zhang YP & Ouyang ZQ HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. Genome Biol. 17, 14 (2016). [PubMed: 26821746]
116. Imakaev M et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 9, 999-+ (2012). [PubMed: 22941365]
117. Cournac A, Marie-Nelly H, Marbouty M, Koszul R & Mozziconacci J Normalization of a chromosomal contact map. Bmc Genomics 13, 13 (2012). [PubMed: 22233093]
118. Hu M et al. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 28, 3131–3133 (2012). [PubMed: 23023982]
119. Russel D et al. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. Plos Biology 10, 5 (2012).
120. Rieping W, Habeck M & Nilges M Inferential Structure Determination. Science 309, 303–306 (2005). [PubMed: 16002620]
121. Geman S & Geman D STOCHASTIC RELAXATION, GIBBS DISTRIBUTIONS, AND THE BAYESIAN RESTORATION OF IMAGES. Ieee Transactions on Pattern Analysis and Machine Intelligence 6, 721–741 (1984). [PubMed: 22499653]
122. Duane S, Kennedy AD, Pendleton BJ & Roweth D HYBRID MONTE-CARLO. Physics Letters B 195, 216–222 (1987).
123. Swendsen RH & Wang JS REPLICA MONTE-CARLO SIMULATION OF SPIN-GLASSES. Physical Review Letters 57, 2607–2609 (1986). [PubMed: 10033814]
124. Habeck M, Nilges M & Rieping W Replica-exchange Monte Carlo scheme for Bayesian data analysis. Physical Review Letters 94, 4 (2005). [PubMed: 15918241]
125. Dempster AP, Laird NM & Rubin DB MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA EM ALGORITHM. J. R. Stat. Soc. Ser. B-Methodol. 39, 1–38 (1977).
126. Trussart M et al. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res. 43, 3465–3477 (2015). [PubMed: 25800747]
127. Moult J, Fidelis K, Kryshtafovych A, Schwede T & Tramontano A Critical assessment of methods of protein structure prediction (CASP)Round XII. Proteins-Structure Function and Bioinformatics 86, 7–15 (2018).
128. Plewczynski D, Lazniewski M, Von Grotthuss M, Rychlewski L & Ginalski K VoteDock: Consensus Docking Method for Prediction of Protein-Ligand Interactions. Journal of Computational Chemistry 32, 568–581 (2011). [PubMed: 20812324]
129. Ren XD et al. Novel Consensus Docking Strategy to Improve Ligand Pose Prediction. Journal of Chemical Information and Modeling 58, 1662–1668 (2018). [PubMed: 30044626]
130. Stamm M & Forrest LR Structure alignment of membrane proteins: Accuracy of available tools and a consensus strategy. Proteins-Structure Function and Bioinformatics 83, 1720–1732 (2015).
131. Sharma A & Manolakos ES Multi-criteria protein structure comparison and structural similarities analysis using pyMCPSC. Plos One 13, 15 (2018).
132. Wei Y, Thompson J & Floudas CA CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. Proceedings of the Royal Society a-Mathematical Physical and Engineering Sciences 468, 831–850 (2012).
133. Kieslich CA, Smadbeck J, Khoury GA & Floudas CA conSSert: Consensus SVM Model for Accurate Prediction of Ordered Secondary Structure. Journal of Chemical Information and Modeling 56, 455–461 (2016). [PubMed: 26928531]

134. Kandoi G, Leelananda SP, Jernigan RL & Sen TZ in Prediction of Protein Secondary Structure, Vol. 1484 (eds. Zhou Y, Kloczkowski A, Faraggi E & Yang Y) 35–44 (Humana Press Inc, Totowa; 2017).

135. Hacker WC, Li S & Elcock AH Features of genomic organization in a nucleotide-resolution molecular model of the Escherichia coli chromosome. Nucleic Acids Res. 45, 7541–7554 (2017). [PubMed: 28645155]

136. Le TB, Imakaev MV, Mirny LA & Laub MT High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342, 731–734 (2013). [PubMed: 24158908]

137. Grigoryev SA, Arya G, Correll S, Woodcock CL & Schlick T Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. Proceedings of the National Academy of Sciences 106, 13317–13322 (2009).

138. Arya G, Zhang Q & Schlick T Flexible histone tails in a new mesoscopic oligonucleosome model. Biophys. J. 91, 133–150 (2006). [PubMed: 16603492]

139. Arya G & Schlick T Role of histone tails in chromatin folding revealed by a mesoscopic oligonucleosome model. Proceedings of the National Academy of Sciences 103, 16236–16241 (2006).

140. Nam G-M & Arya G Torsional behavior of chromatin is modulated by rotational phasing of nucleosomes. Nucleic Acids Res. 42, 9691–9699 (2014). [PubMed: 25100871]

141. Smith EA et al. Quantitatively Imaging Chromosomes by Correlated Cryo-Fluorescence and Soft X-Ray Tomographies. Biophysical Journal 107, 1988–1996 (2014). [PubMed: 25418180]

142. Di Pierro M, Cheng RR, Aiden EL, Wolynes PG & Onuchic JN De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. Proceedings of the National Academy of Sciences of the United States of America 114, 12126–12131 (2017). [PubMed: 29087948]

143. MacPherson Q, Beltran B & Spakowitz AJ Bottom-up modeling of chromatin segregation due to epigenetic modifications. Proceedings of the National Academy of Sciences of the United States of America 115, 12739–12744 (2018). [PubMed: 30478042]

144. Brackley CA et al. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome Biology 17, 16 (2016). [PubMed: 26831908]

145. Pouyanfar S et al. A Survey on Deep Learning: Algorithms, Techniques, and Applications. Acm Computing Surveys 51, 36 (2019).

146. Alipanahi B, Delong A, Weirauch MT & Frey BJ Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology 33, 831-+ (2015).

147. Guo YB, Wang BY, Li WH & Yang B Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. Journal of Bioinformatics and Computational Biology 16, 19 (2018).

148. Zhang L, Yu GX, Xia DW & Wang J Protein-protein interactions prediction based on ensemble deep neural networks. Neurocomputing 324, 10–19 (2019).

149. Schreiber J, Libbrecht M, Bilmes J & Noble WS Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv, 103614 (2017).

HIGHLIGHTS

- Computational methods can infer 3D genome structure from 3C contact maps.
- Distance-optimization methods convert contact maps to internal distances.
- Polymer physics methods recapitulate contact maps from polymer model simulations.
- Maximum-likelihood and Bayesian methods infer parameters of statistical models.

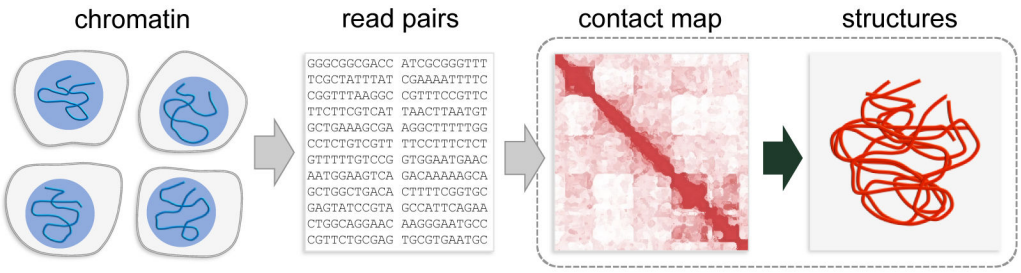

Figure 1:

Schematic of the overall pipeline for studying 3D organization of chromatin using 3C technology. The primary topic of this review article on computational methods for recovering 3D structure and structure populations of chromatin from 2D contact maps is highlighted by the dashed box. The maps themselves are generated from the sequences of read pairs, which are in turn collected from crosslinked chromatin in a large ensemble of cells.

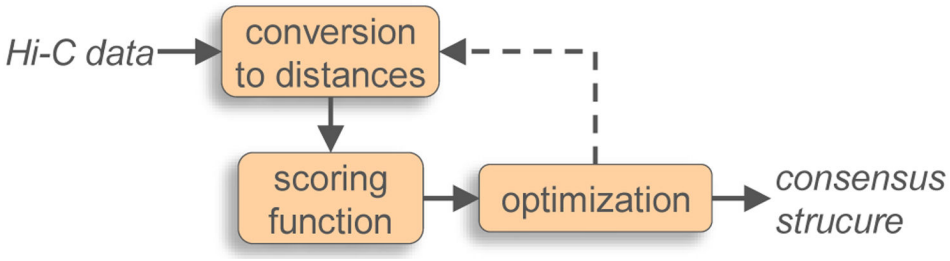

Figure 2:

General scheme for computational methods that rely on distance-optimization to generate a consensus structure that is consistent with an experimental contact map. Blocks represent the main components common to these methods. Arrows represent the main flow of information. Not all methods attempt to optimize the parameters of the relation used to convert contact frequencies into inter-locus distances.

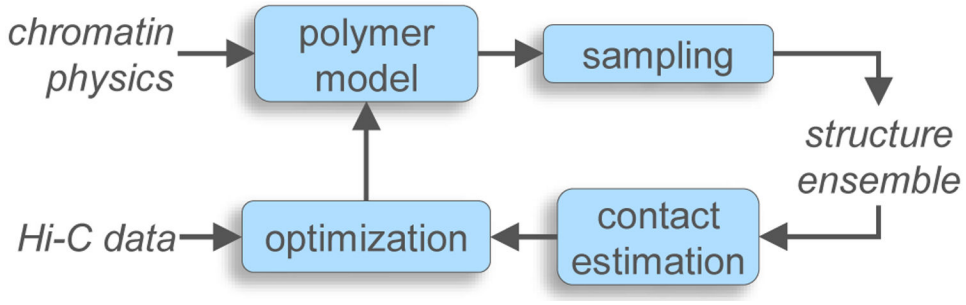

Figure 3:
General scheme for computational methods that use polymer physics to generate an ensemble of structures that is consistent with an experimental contact map. Blocks represent the main components common to these methods. Arrows represent the main flow of information.

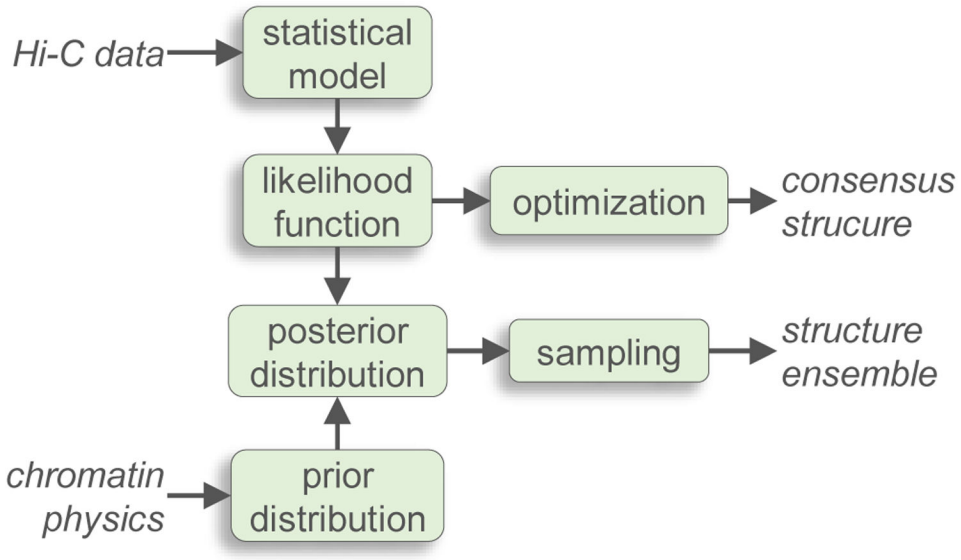

Figure 4:
General scheme for computational methods that use maximum likelihood or Bayesian inference to generate a consensus structure or an ensemble of structures, respectively, that is consistent with an experimental contact map. Blocks represent the main components common to these methods. Arrows represent the main flow of information.