FEBS Letters xxx (2015) xxx–xxx

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Modeling chromosomes: Beyond pretty pictures

Maxim V. Imakaev [a,1], Geoffrey Fudenberg [b,c,1], Leonid A. Mirny [a,b,c,*]

[a] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Graduate Program in Biophysics, Harvard University, Cambridge, MA 02138, USA
[c] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 1 September 2015
Accepted 3 September 2015
Available online xxxx

Edited by Wilhelm Just

Keywords:
Chromatin
Hi-C
Polymer
Model
Simulation

ABSTRACT

Recently, Chromosome Conformation Capture (3C) based experiments have highlighted the importance of computational models for the study of chromosome organization. In this review, we propose that current computational models can be grouped into roughly four classes, with two classes of *data-driven* models: consensus structures and data-driven ensembles, and two classes of *de novo* models: structural ensembles and mechanistic ensembles. Finally, we highlight specific questions mechanistic ensembles can address.

1. Review of Hi-C

Chromosome Conformation Capture (3C [1]) based methods provide high-resolution and genome-wide maps of contact frequencies between genomic positions. 3C methods convert spatial contacts between pairs of genomic loci into molecular products that can be assayed using high-throughput sequencing. To obtain these molecular products, the 3C protocol involves: crosslinking chromatin to freeze contacts in place, digesting chromatin with restriction enzymes to break full chromosomes into fragments, and capturing interactions between spatially contacting fragments using proximity ligation. Depending on the specific approach, contacts between fragments are either read out: 1-by-1 (3C [1]), 1-by-all (4C [2,3]), many-by-many (5C [4]), or all-by-all (Hi-C [5], TCC [6], and 3C-seq [7]). 3C-based methods are usually performed on large populations of cells, producing population-average maps of chromosomal contact frequencies, though single-cell approaches have also been developed [8].

3C-based methods display many layers of chromosomal organization in higher eukaryotes, and computational models can aid understanding at each level. Following [9], mammalian chromosome display roughly five levels of organization: (1) chromosome territoriality (cis/trans ratio [5,6]); (2) distance-dependent contact frequency, $P(s)$ [5,10]; (3) genomic compartments (eigenvectors) [5,11]; (4) topological domains (TADs) [12,13]; (5) point interactions [14]. *Drosophila* chromosomes display similar levels of organization [7,15].

Interestingly, yeast and bacterial chromosomes appear to be organized on different principles, and each requires independent modeling efforts. Importantly, they are not simply scaled-down human chromosomes; for example, neither displays alternating compartments. In yeast, chromosome organization is dominated by strong centromere-centromere clustering, consistent with a Rabl-type conformation both in *Cerevisiae* [1,16] and *Pombe* [17,18]. In *Caulobacter* [19,20] and *Subtilis* [21,22], a major feature is co-alignment of two chromosomal arms.

2. Challenges for models

One major challenge for developing spatial models of chromosomes is that Hi-C maps generally do not represent information from single in vivo conformations. This is underscored by comparing conventional Hi-C maps with maps from single-cell Hi-C experiments [8]. In a conventional Hi-C experiment, hundreds of millions of cells are pooled together, creating a population-average map of contact probability. A striking feature of conventional Hi-C maps is that there are almost no regions of zero contact probability; any genomic locus may be found in contact with any

[*] Corresponding author at: Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.febslet.2015.09.004
0014-5793/© 2015 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.

other locus in some, potentially very small, fraction of cells. Consistently, single-cell Hi-C experiments show that contact maps of individual cells are highly variable [8]; each individual cell only realizes a subset of possible contacts, which are different in every cell. A similar difference was observed between single-cell and population-average contact maps in polymer simulations [10]; contacts in individual realizations of the polymer model were highly variable, while the contact map averaged over many realizations was homogeneous. Since a single structure produces a very sparse contact map, a diverse set, or an *ensemble of conformations,* is needed to reproduce a population-average Hi-C map [23].

Another challenge for models is the complicated relationship between Hi-C contact probability and spatial distance, as measured by imaging. While average contact probability and average spatial distance of two loci are often highly anti-correlated [24,25], Hi-C probes a particular part of the pairwise distance distribution and focuses on small distances (contacts), which can be very different from the mean or median distance. In particular, Hi-C contact frequency may increase despite two loci becoming further apart on average. Interestingly, we found this situation occurs in published data [14] (peak-4-loop has roughly 4-fold higher contact probability despite being further away on average than peak-3-control; nevertheless, the small distance behavior of the CDF is in agreement with Hi-C). This illustrates that Hi-C contact probabilities cannot be simply translated into spatial distances. Reconciling microscopy measurements of chromosomal organization with Hi-C is an important challenge [26,27], yet will require very high resolution [28,29] and high-throughput [30] imaging experiments to probe the infrequently sampled small-distance regime of the spatial distance distribution.

A final challenge for developing spatial models of chromosomes is determining how to compare them with Hi-C data. Simple correlation between Hi-C maps can be misleading due to the very strong dependence of contact probability on distance in all maps. For example, a Hi-C map for mouse chr1 (CH12 cells, 100 kb resolution [14]) correlates with a same-size region of a human chr3 (GM12878 cells, 100 kb resolution [14]) with Pearson's $r = 0.41$, and Spearman's $r = 0.83$, while there obviously is no underlying relationship between the two maps. For this reason, we favor comparisons that consider a range of features (e.g. $P(s)$, TADs, compartments, specific interactions, see [9]) rather than simply relying on the correlation between two Hi-C maps.

3. Four classes of spatial models

An increasingly-common research goal has been to develop spatial models of chromosomes that can reproduce essential features of various experimentally-obtained contact maps (often either from Hi-C or 5C experiments; for convenience, we use the term Hi-C in what follows). However, the approaches to this problem have differed greatly in their assumptions and implementation. Moreover, different approaches can be used to address different questions. We believe that modeling approaches can be divided into roughly four categories, where the first two are *data-driven* approaches, and the latter two are *de novo* modeling approaches.

4. Data-driven models

A compelling approach is to directly use Hi-C data to produce a spatial model of a chromosome. This has led to a variety of methods that range from reproducing a single structure (consensus structure models) to reproducing an ensemble of structures (data-driven ensembles) (Fig. 1).

4.1. Consensus structure models

Consensus structure approaches aim at reconstructing a single chromosome structure from Hi-C maps [16,31–35]. These methods usually assume some relationship between the contact probability and the spatial distance between loci. Based on this relationship, these models impose a set of constraints, and generate a consensus structure. However, as discussed above, the structures produced by these approaches are inconsistent with Hi-C maps, as a Hi-C map has to be described by a highly variable ensemble of structures. In a sense, looking for the consensus structure of a chromosome is analogous to searching for the consensus structure of an unfolded protein. The conceptual misunderstanding underlying consensus models is after interpreting average contact frequencies as average distances, assuming that there are only small fluctuations around the average distance. This assumption is clearly violated in imaging experiments, which show that the variability in spatial distance between two loci is often similar to their average separation [25]. While consensus structure approaches can be thought of as methods for visualizing Hi-C data, transformations made by these approaches (contact frequencies to distances, distances to 3D structure) can lead to information loss and distortion.

We note that reconstruction of a single chromosome conformation from a single-cell Hi-C map [8] is actually a very different problem, and in this case it is well-justified to derive a consensus structure attempting to satisfy the observed contacts. Also, the authors carefully considered only the structures of the single-copy X-chromosome to avoid ambiguities arising from mapping Hi-C interactions onto homologous chromosomes. However, it is not yet clear whether the resolution of currently-available single-cell Hi-C data is sufficient to faithfully reconstruct the structure of a chromosome, as current experiments report roughly one contact per 100 kb region. As follows, further computational tests would be useful for these single-cell modeling approaches; for example, whether reconstructed structures have a similar $P(s)$ as the single-cell Hi-C data.

4.2. Data-driven ensembles

A second set of data-driven modeling approaches aim at simulating an *ensemble of structures* to reconstruct experimental Hi-C maps [6,25,36,37]. Since the variability needed to reconstruct experimental maps cannot be achieved by setting rigid distance constraints between different genomic regions, these methods usually employ a very flexible set of constraints. Interestingly, successful models either explicitly or implicitly make use of a polymer description of chromatin. Regardless of the nomenclature, a simulated region of chromatin fiber is described by a series of monomers (referred to alternately as 'points', 'beads', or 'particles') that interact via a number of forces. The first essential interaction is linear connectivity, imposed by harmonic bonds between adjacent monomers. The second essential interaction is that of excluded volume interactions between each pair of monomers, where monomers either interact as hard or soft spheres upon collisions. Additional fiber stiffness is often imposed as a function of the angle between each triplet of linearly connected monomers. Finally, a set of pairwise interactions between monomers, inferred by fitting to the Hi-C contact map, is usually added on top of the basic polymer interactions. Data-driven ensemble approaches then use Monte Carlo or Molecular Dynamics to sample the space of possible spatial structures given these interactions and generate a set of conformations that is variable enough to reproduce a Hi-C contact map.

We note that the boundary distinguishing data-driven ensembles and consensus structure approaches is not a strict division, and that specific approaches give different degrees of variability.