Computational methods for predicting 3D genomic organization from high-resolution chromosome conformation capture data

Kimberly MacKay\textsuperscript{1*} and Anthony Kusalik\textsuperscript{1}

\textsuperscript{1} Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

\textsuperscript{*} Corresponding Author

Corresponding Author Contact Information:

telephone number: +1-306-966-4886

fax number: +1-306-966-4884

e-mail: kimberly.mackay@usask.ca
Abstract

The advent of high-resolution chromosome conformation capture assays (like 5C, Hi-C and Pore-C) has allowed for unprecedented sequence-level investigations into the structure-function relationship of the genome. In order to comprehensively understand this relationship, computational tools are required that utilize data generated from these assays to predict 3D genome organization (the 3D genome reconstruction problem). Many computational tools have been developed that answer this need but a comprehensive comparison of their underlying algorithmic approaches has not been conducted. This manuscript provides a comprehensive review of the existing computational tools (from November 2006 to September 2019, inclusive) that can be used to predict 3D genome organizations from high-resolution chromosome conformation capture data. Overall, existing tools were found to use a relatively small set of algorithms from one or more of the following categories: dimensionality reduction, graph/network theory, maximum likelihood estimation and statistical modelling. Solutions in each category are far from maturity and the breadth and depth of various algorithmic categories have not been fully explored. While the tools for predicting 3D structure for a genomic region or single chromosome are diverse, there is a general lack of algorithmic diversity among computational tools for predicting the complete 3D genome organization from high-resolution chromosome conformation capture data.

Key Phrases

Genome Organization, 3D Genome Prediction, 3D Genome Reconstruction Problem, High-Resolution Chromosome Conformation Capture Data, Hi-C, 5C
1 Introduction

This manuscript provides a survey of the existing computational tools that can be used for predicting 3D genomic organization from high-resolution chromosome conformation capture data. Relevant biological and computational background is provided in Section 2. Section 3 describes the 3D genome reconstruction problem (3D-GRP) formalism. Section 4 provides an overview of existing tools for solving the 3D-GRP. Two of these existing tools (one consensus and one ensemble) are described in more detail in Section 5. Similarly, Section 6 provides an overview of existing tools for solving the related, but simpler, problem of predicting 3D organization for a single chromosome or genomic region. An exemplar consensus and ensemble tool for solving this simpler problem are discussed in more detail in Section 7. Finally, a discussion of the shortcomings of the existing approaches and future research directions can be found in Section 8.

2 Background

Like many areas of biology, the relationship between genomic structure and function is closely linked [1]. Alterations in the 3D organization of chromosomes have been demonstrated in a wide variety of nuclear and cellular processes, including DNA translocation [1], differentiation [2], serum response [3], therapeutic response [4] and response to DNA damage [5]. The unique spatial organization of the genome that is seen under these different cellular conditions is hypothesized to be a crucial mechanism driving various nuclear and cellular functions. It has been theorized that this dynamic organization of the genome may be driven by global regulation of gene expression (or vice-versa) [6–10] since 3D genome organization has been shown to facilitate interactions between genes and their regulatory elements [11, 12].

Traditionally, microscopy techniques have been utilized to visualize the spatial organization of chromosomes within the nucleus. While informative, they do not provide sequence-
level information about the observed organizations [13]. Therefore other biological techniques must be used (either in combination or standalone) to allow for the sequence-level inference of 3D genomic organization. Many such biological techniques have been developed to assay the 3D genome organization at various sequence-level resolutions [1, 14–16]. In general, these techniques are able to determine whether a single (or multiple) pair(s) of genomic regions are in close 3D physical proximity. Genomic regions in close proximity are more commonly referred to as "interacting".

Table 1: Biological techniques that can be used to assay 3D genome organization. Techniques are categorized based on the number of genomic regions they assay.

See table1.xlsx

Table 1

| Category              | Description                                                                 |
|-----------------------|-----------------------------------------------------------------------------|
| one-by-one            | (used to detect an interaction between a single pair of genomic regions)     |
| one-by-all            | (used to detect all the interactions between one genomic region and the rest of the genome) |
| many-by-many          | (used to detect interactions between many genomic regions and many other loci, where many is the number of loci on a chip or microarray) |
| many-by-all           | (used to detect interactions between many genomic regions and the rest of the genome) |
| all-by-all            | (used to detect all the interactions occurring between mappable regions of the genome) |

The biological techniques used for detecting 3D genomic organization can be broadly classified into the following categories based on the number of genomic regions they assay: one-by-one (used to detect an interaction between a single pair of genomic regions); one-by-all (used to detect all the interactions between one genomic region and the rest of the genome); many-by-many (used to detect interactions between many genomic regions and many other loci, where many is the number of loci on a chip or microarray); many-by-all (used to detect interactions between many genomic regions and the rest of the genome); and all-by-all (used to detect all the interactions occurring between mappable regions of the genome). Table 1 provides the specific names and citations for some of the biological techniques in each of these categories. Briefly, these techniques all follow five general steps (with slight modifications): (1) chemical cross-linking, (2) fragmentation, (3) ligation, (4) reverse cross-linking, and (5) technique-specific detection. A visual overview of the general workflow for each technique can be found in the review by Denker and de Laat [14]. Additional information regarding the biological background for these techniques can be found in the review recently published by Han et al. [33]. For the purpose of this manuscript, "high-resolution chromosome conformation capture" (HR-3C) will refer to the many-by-many,
many-by-all and all-by-all techniques.

Algorithms for predicting 3D genome structure utilize a set of pairwise interactions and associated frequencies as input. Typically, this data is extracted from the results of a many-by-many, many-by-all or all-by-all (HR-3C) assay. The one-by-one and one-by-all techniques do not generate enough pairwise data points to allow for an accurate prediction of 3D genomic structure on their own. It is possible that the data from a one-by-one or one-by-all assay could be combined with data from a HR-3C assay and used as input to a 3D prediction algorithm, but this is not common practise in the field. The following paragraphs present a brief overview of how a set of pairwise interactions and associated frequencies can be extracted from a HR-3C assay’s raw data (sequencing reads).

In general, HR-3C techniques utilize next-generation sequencing technologies to identify the sequences of interacting regions of the genome. Once these sequencing reads are generated, they are typically processed through a read mapping and filtering pipeline like HiCUP [34]. Briefly, this process involves quality control, read-splitting and independent read-mapping. Mapping sequence reads to a reference genome results in the generation of a matrix called a whole-genome contact map. A whole-genome contact map is a \( N \times N \) matrix, where \( N \) is the number of genomic "bins" where each bin represents a contiguous sequence of linear DNA [35–37]. In general, the size of the whole-genome contact map (the number of genomic bins) is approximately equal to the total genome size divided by the assay’s experimental resolution. Each cell \((A_{i,j})\) of a whole-genome contact map \((A)\) indicates the count of how many times genomic bin \(i\) has been found to interact with genomic bin \(j\). These counts are symmetric along the diagonal (i.e. \(A_{i,j} = A_{j,i}\)) and are often referred to as the frequency of the interaction between \(A_i\) and \(A_j\) (or interaction frequency).

After the whole-genome contact map is generated, interaction frequencies are normalized to correct for some of the inherent biases resulting from HR-3C experiments. These biases include (but are not limited to) discrepancies in DNA compaction or "visibility" [38], GC content [39, 40] and copy number variation [41]. Various computational methods have
been developed to dampen these biases through normalization [40, 42–45]. Most commonly, an iterative correction and eigenvector (ICE) decomposition [38] or Knight-Ruiz normalization [42, 43] is applied resulting in fractional interaction frequencies. ICE decomposition aims to achieve equal visibility across all genomic regions and results in relative interaction frequencies. Knight-Ruiz normalization performs matrix balancing resulting in fractional interaction frequencies where the rows and columns sum to 1. A comprehensive comparison of the normalization methods for HR-3C data has been recently published by Lyu et al. [46].

Downstream analysis of normalized whole-genome contact maps has uncovered unique genome-level patterns including distance-dependent interaction frequencies and more interactions between genomic regions on the same chromosome (cis-chromosomal interactions) than between regions on different chromosomes (trans-chromosomal interactions) [29, 35]. Further computational analysis of whole-genome contact maps has revealed the presence of various "hallmarks" of 3D genome organization. For instance, statistical analysis of whole-genome contact maps has revealed the presence of structural subunits called topologically associating domains (TADs) [47]. TADs are linear regions of DNA where interactions occur more frequently within the domain instead of between domains [47]. Originally, TADs were hypothesized to be structural building blocks for 3D genome organization but it has been determined that they serve no structural importance [48, 49].

Normalized whole-genome contact maps can also be used to infer a 3D structure of the genome (or a single genomic region). The process of predicting a model of the 3D genomic organization from a contact map is known as the 3D genome reconstruction problem (3D-GRP) [50] (described in more detail below). Many computational methods have been developed that utilize the data from HR-3C experiments to predict 3D genomic organization. Classically, existing programs have been broadly classified based on the number of genome models the method produces. Ensemble tools generate a collection of structures which represent the different genome organizations that may be present within a population of cells while consensus tools generate one structure which represents the population-averaged
genome organization [35]. This manuscript provides a comprehensive review of the existing
tools published from November 2006 (the year 5C was first described [22]) to September
2019 that use data extracted from HR-3C techniques to predict a 3D structure of complete
genomes or a genomic region (Sections 4 and 6 respectively). A brief overview of the main
chromosome models used by these existing tools is provided in Section 3 of this manuscript.
The subsequent sections assume that the reader has some familiarity with the following
concepts: multi-dimensional scaling [51, 52], shortest path algorithms [53–56], expectation
maximization [57], genetic algorithms [58], gradient descent (or ascent) [59], simulated an-
nealing [60, 61], and Markov chain Monte Carlo sampling [62].

3 Problem Formalism

As mentioned above, the process of predicting a 3D genomic organization from HR-3C
data is known as the 3D genome reconstruction problem (3D-GRP) [50]. It should be noted
that the 3D-GRP has also been referred to as the 3D chromatin structure modelling problem
[63] and that these two phrases can be used interchangeably. More formally, the 3D-GRP
can be formulated as an optimization problem that tries to optimize the combined distance
between multiple pairs of genomic regions. Informally, this is represented as a geometry
problem [64] where the genomic bins are encoded as points and the goal is to find each
point’s (x, y, z) coordinates such that the pairwise distances between points best capture
the corresponding interaction frequencies. It is assumed that, on average, a pair of genomic
regions with a small interaction frequency will be further away in 3D space than a pair of
genomic regions with a higher interaction frequency [6, 65–72]. This relationship is often
modelled through the following inverse function for a given pair of genomic regions (i and
j): dist_{i,j} = \frac{1}{A_{i,j}^{\alpha}} \text{ where dist is the distance between the two genomic regions, } A_{i,j} \text{ is the}
corresponding normalized interaction frequency (a value between 0 and 1) from the whole-
genome contact map, and } \alpha \text{ is an exponential factor with a value typically between 0.1 and}
Most existing methods focus on finding the optimal value (or a set of values) for \( \alpha \) and each point’s \((x, y, z)\) coordinates so that the computed distances closely recapitulate the original normalized frequencies from the whole-genome contact map [50]. Formally, the 3D-GRP can be defined in the following way when Euclidean distance is used:

\[
\text{Given a whole-genome contact map } A \text{ with bins from } 1..N, \text{ determine } \alpha \text{ and each point’s } (x, y, z) \text{ coordinates such that}
\]

\[
\text{for } i = 1..N \text{ and } j = 1..N \quad \text{dist}_{i,j} = \frac{1}{A_{i,j}^\alpha} \quad (1)
\]

\[
\text{and the sum}
\]

\[
\sum_{i=1,j=1}^{N} \left| \text{dist}_{i,j} - \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \right| 
\]

\[
\text{is minimized and}
\]

\[
(x_i, y_i, z_i) \neq (x_j, y_j, z_j) \text{ where } 1 \leq i \leq N, 1 \leq j \leq N, i \neq j \quad (3)
\]

In order to predict a 3D organization for a complete genome or genomic region, individual chromosomes need to be modelled as a set of points that can be assigned 3D coordinates. In general, existing methods use one of the following chromosome models. (1) Beads: each individual chromosome is represented as a collection of \( M \) beads where \( M \) is the number of genomic bins that constitute the linear extent of a chromosome. (2) Beads-on-a-String: again, each individual chromosome is represented as a collection of \( M \) beads. Unlike the beads model, "strings" of a fixed length are used to connect each pair of adjacent beads. Typically, these represent beads that are linearly adjacent on a chromosome. (3) Beads-on-a-Spring: this representation is similar to (2) but beads on an individual chromosome are connected with "springs" to represent the linear extent of a chromosome. Springs typically
have a variable length that is based on attractive and repulsive forces of the connected beads. 

(4) Graph/Network: each bin from the whole-genome contact map is represented as a node in a network. Edges between nodes represent interactions from the contact map. Often, edges between bins on the same chromosome that are linearly adjacent are not included. (5) Polymer: each chromosome is represented as a line which is composed of consecutive line segments. Each line segment encodes a genomic bin or a genomic region that is delimited by two endonuclease restriction sites. (6) Piecewise curve: this is a mathematical formulation where each chromosome is represented as a set of connected 3D curves. Each curve represents an individual genomic bin or region.

4 Existing Tools for solving the 3D-GRP

A comprehensive list of the existing computational tools for predicting 3D genomic organization from HR-3C data is available in Table 2. This table represents the majority of tools in the existing literature at the time of manuscript submission. Additional information regarding how these manuscripts were selected can be found in Section 13.1.

Table 2: Existing computational tools for predicting 3D genome organization from HR-3C data. Tools are categorized as either consensus or ensemble and then listed in alphabetical order. Tools marked with an asterisk (*) did not appear to be actively maintained at the time of manuscript submission. Column headings are as follows: Name, the tool’s name or abbreviated reference (Panel labels from Figure 1 are provided in parentheses); Technique, the general algorithmic strategy employed; CHR Model, a description of the chromosome model utilized; Additional Data, any additional biological datasets required; a priori Constraints, a descriptor denoting whether a priori information is required and/or assumed; Language, the programming language used to implement the tool; Availability Mode, a description of how the tool was deployed; Website, a link to the tool. Abbreviations are as follows: MDS, Multi-Dimensional Scaling; MLE, Maximum Likelihood Estimation; LAD, lamin-associated domains; DamID, DNA adenine methyltransferase identification. IPOPT is a software library used for nonlinear optimization. A dash indicates "not applicable", "not available", or "none", as appropriate.

See table2.xlsx
See Figure 1 (attached)

**Figure 1:** An example of a predicted 3D genome organization from each of the existing consensus (A-G) and ensemble (H-M) tools. Tool name or abbreviated reference can be found at the top of each panel and the organisms (and cell type, when applicable) are listed at the bottom of the panel. The abbreviation ESCs stands for embryonic stem cells. Permission was obtained to reprint these images where required.

In Table 2, the existing tools are categorized based on the number of predicted genome organizations they produce (i.e. ensemble vs consensus). Tools marked with an asterisk (*) did not appear to be actively maintained (DAAM) at the time of manuscript submission. This designation was given if the software presented in the original manuscript(s) could no longer be accessed. Typically, this was due to obsolete or nonfunctional website uniform resource locators (URLs). An example of the output produced by each tool can be found in Figure 1. In each case, the images were extracted from the corresponding original publication. Permission was obtained to reprint these images where required. All of the existing tools utilize either heuristics or approximations in their solution.

Five of the seven consensus methods and three of the six ensemble methods listed in Table 2 provide access to the source code or a web interface. As mentioned in Table 2, the method developed by Stevens *et al.* only works with single-cell interaction data while all other methods accept interaction data from a population of cells. Currently, none of the available, actively maintained ensemble methods are usable for solving the 3D-GRP in the general case. This is because they rely on hypothesized "hallmarks" of genome organization like TADs, the presence of binding motifs for proteins often found at TAD boundaries (like CTCF) or require diploid, un-phased datasets to make their predictions. This is problematic since these genomic "hallmarks" have been shown to not exist in some organisms like *Arabidopsis thaliana* [84, 85]. This could pose a major barrier going forward as investigations into the 3D genomic organization of non-model organisms continues.

---

1Permission to reprint was obtained for the following panels: B (license number: 4751560452148), C (license number 4703250178326), G (license number: 4751560316055), I (license number 4703251132971). All other panels contain images that are allowed to be reprinted under a Creative Commons License. Additional information pertaining to reprint permissions for each image can be found in Appendix A.
5 Exemplar 3D-GRP Tools

The following section provides a more detailed discussion of an exemplar consensus and an exemplar ensemble method for solving the 3D-GRP. These methods were chosen since they were the most recent additions to the set of tools presented in Table 2 that have been used by the community to predict 3D genome structure based on real (rather than simulated) HR-3C datasets.

5.1 Consensus: miniMDS

miniMDS [77] is a consensus method that combines metric multi-dimensional scaling (MDS) with a divide-and-conquer approach to solve the 3D-GRP. Briefly and in general terms, the local structure of each chromosome is solved and then fitted to a low-resolution global genome prediction. First, a hidden Markov model is used to locally partition each chromosome into a set of subproblems. This hidden Markov model is derived from the TAD-finding algorithm developed by Dixon et al. [47] to identify local regions of a chromosome where edges of the region preferentially interact with the opposite side of the region. Each subproblem is then converted to a distance matrix based on equation 4 and metric MDS is used to solve a high-resolution local structure. It should be noted that the zero-distances (typically unmappable genomic regions) are ignored by MDS. This step is then repeated for each complete chromosome at a lower-resolution. High-resolution local structures are fitted to these lower-resolution chromosome structures using the Kabsch algorithm [86]. Finally, this fitting is repeated at an even lower resolution using the whole dataset to generate a low resolution global 3D structure. This global structure is then used as the final guide to position the chromosome structures resulting in a completed 3D genome prediction. An example of the output produced by miniMDS can be seen in Figure 1D. miniMDS should be used with caution in organisms where the existence of TADs or TAD-like structures has not been established since the hidden Markov model used for the initial division relies on
the presence of TAD-like structures.

\[
\text{dist}_{i,j} = \begin{cases} 
A_{i,j}^{-0.25} & \text{if } A_{i,j} > 0 \\
0 & \text{if } A_{i,j} = 0 
\end{cases}
\]  

(4)

5.2 Ensemble: Li et al.

The method developed by Li et al. [80] is an ensemble method that incorporates data from lamina-DamID experiments (which are able to detect interactions between the nuclear lamina and genomic regions) with HR-3C data to predict a 3D genomic organization at TAD-level resolution. The data from lamina-DamID experiments allows for the identification of which TAD regions interact with the nuclear envelope (the periphery of the nucleus; abbreviated NE). Briefly, this method uses maximum likelihood estimation to find a set of 3D genome structures that have statistically consistent TAD-TAD and TAD-NE interactions. Specifically, this method uses a variant of expectation maximization described by Tjong et al. [81] to optimize this joint probability. It incorporates additional spatial constraints into the optimization based on known features of the Drosophila melanogaster (fruit fly) genome. These Drosophila-specific constraints are based on microscopy imaging and include the nuclear radius, a maximum distance between chromosome copies, a maximum distance between adjacent TADs, links between heterochromatin regions, links between adjacent TADs, and centromere anchoring to the nucleolus. Due to these additional constraints, this method should only be applied to datasets from Drosophila melanogaster and would not be suitable for solving the 3D-GRP in the general case. An example of the output produced by this tool can be seen in Figure [I]. This method could potentially be applied to other organisms with TADs if the required organism-specific spatial constraints are available.
6 Predicting 3D structures for genomic regions or single chromosomes

There are significantly more tools available that can be used to predict 3D structure of a single genomic region or chromosome from HR-3C data. For the purpose of this manuscript, we will refer to this as 3D regional prediction. The increased number of available tools for 3D regional prediction is likely because it is a much simpler (and often smaller) problem than the 3D-GRP since it does not have to take trans-chromosomal interactions (interactions between genomic regions on different chromosomes) into account. In the majority of cases, it would be computationally infeasible to apply these tools to the 3D-GRP due to their underlying time complexities. It may be possible to overcome this problem by applying a divide-and-conquer approach similar to miniMDS [77].

Table 3: Existing computational tools for predicting 3D organization for a genomic region from HR-3C data. Tools are categorized as either consensus or ensemble and then listed in alphabetical order. Tools marked with an asterisk (*) did not appear to be actively maintained at the time of submission. Column headings are as follows: Name, the tool’s name or abbreviated reference (Panel labels from Figures 2 and 3 are provided in parentheses); Technique, the general algorithmic strategy employed; CHR model, a description of the chromosome model utilized; Additional Data, any additional biological datasets required; a priori Constraints, a descriptor denoting whether a priori information is required and/or assumed; Language, the programming language used to implement the tool; Availability Mode, a description of how the tool was deployed; Website, a link to the tool’s source code. Abbreviations are as follows: IMP, Integrative Modeling Platform (https://integrativemodeling.org/); MDS, Multi-Dimensional Scaling; MLE, Maximum Likelihood Estimation; MCMC, Markov Chain Monte Carlo. A dash indicates "not applicable", "not available", or "none", as appropriate.

See Table3.xlsx

Table 3 provides a list of the computational techniques that utilize HR-3C data to predict a 3D structure for a given genomic region instead of the whole genome. As described above, tools have been categorized in the following ways: DAAM (*), consensus, and/or ensemble. An example of the output produced by each actively maintained consensus and ensemble tool can be found in Figures 2 and 3, respectively. In each case, the images were extracted...
See Figure 2 (attached)

Figure 2: An example of a predicted region organization from each of the existing regional consensus tools. Tool name or abbreviated reference can be found at the top of each panel and the organisms (and specific region, when applicable) are listed at the bottom of the panel. The abbreviation CHR stands for chromosome. Permission was obtained to reprint these images where required.

See Figure 3 (attached)

Figure 3: An example of a predicted region organization from each of the existing regional ensemble tools. Tool name or abbreviated reference can be found at the top of each panel and the organisms (and specific region, when applicable) are listed at the bottom of the panel. The abbreviation CHR stands for chromosome. Permission was obtained to reprint these images where required.

from the corresponding original publication. Permission was obtained to reprint these images where required.

7 Exemplar Regional 3D Prediction Tools

The following section provides a more detailed discussion of an exemplar consensus and an exemplar ensemble method for predicting 3D structure of a single genomic region or chromosome. ShRec3D+ was chosen as the exemplar consensus method since it is the most recent version of one of the popular and highly cited tools, ShRec3D [72, 93]. Chromosome3D was chosen as the exemplar ensemble method since it is the most recent addition to set of ensemble tools presented in Table 3 (Chromosome3D) that has been used by the community to predict 3D regional structures (beyond TAD-level resolution) from real population-based HR-3C data (rather than simulated data).

---

2 Permission to reprint was obtained for the following panels in Figure 2: C (see Appendix A), F (license number: 4751560134989), G (license number 4703250567885), I (license number: 4757200895857), J (license number: 4751551505582), K (license number: 4757201185227). All other panels contain images that are allowed to be reprinted under a Creative Commons License. Additional information pertaining to reprint permissions for each image can be found in Appendix A.

3 Permission to reprint was obtained for the following panels in Figure 3: D (license number: 4757201469502), F (license number: 4751550455329), G (license number: 4751550598254), M (license number: 4751551203420). Additional information pertaining to reprint permissions for each image can be found in Appendix A.
7.1 Consensus: ShRec3D+

ShRec3D+ \cite{72} is a consensus method that is based on ShRec3D \cite{72, 93} and ChromSDE \cite{63}. An overview of the approach taken by ShRec3D+ is as follows. First, interactions are converted into a weighted graph where edge weight (which represents the distance between two vertices) is initially calculated with equation 5 where $\alpha$ is a user-selected value between 0.0 and 2.0. Second, the Floyd-Warshall algorithm is applied to optimize the distances so that the vertices satisfy the triangle inequality. Finally, classical MDS is applied to calculate the $(x, y, z)$ coordinates of each vertex in the graph. An example of the output produced by ShRec3D+ can be seen in Figure 2G.\footnote{Reprinted, with permission from IEEE (license number 4703250567885).} ShRec3D+ does not optimize the value of $\alpha$ like its predecessor ShRec3D. This was done to improve runtime but adds a significant potential for user error in new applications because the user might unintentionally specify an inappropriate value.

$$w_{i,j} = \begin{cases} A_{i,j}^{-\alpha} & \text{if } A_{i,j} > 0 \\ \infty & \text{if } A_{i,j} = 0 \end{cases} \quad (5)$$

7.2 Ensemble: Chromosome3D

Chromosome3D \cite{99} is an ensemble method that models a genomic region as a string of beads. Interaction frequencies are converted to distances based on equation 6 where $K$ is a scaling constant and $\alpha$ is a tuneable parameter with suggested values of 11 and 1/3, respectively. Simulated annealing is then used to find the $(x, y, z)$ coordinates for each bead such that the absolute difference between the predicted distances (based on the $(x, y, z)$ coordinates) and initial calculated distances (based on the interaction frequencies) are minimized. This is repeated twenty times to generate an ensemble of potential 3D genomic structures. This set of structures is ranked using Spearman’s rank correlation coefficient to determine which predicted structures best represent the initially distances calculated based...
on the interaction frequencies. An example of the output produced by Chromosome3D can be seen in Figure 3C. Chromosome3D has been shown to outperform ShRec3D when the input data set is noisy [99], which is a characteristic of HR-3C datasets.

\[
\text{dist}_{i,j} = \frac{K}{A_{i,j}^\alpha/\text{average}(A_{i,j}^\alpha)}
\]  

(6)

8 Future Directions

There is a lack of algorithmically diverse solutions to the 3D-GRP that could be applied to a wide-variety of organisms (we refer to this as generalizable for this manuscript). Five of the six consensus methods use a MDS as a part of their approach for solving the 3D-GRP. MDS presents many potential issues which are described in Section 8.2. The remaining method by Stevens et al. can only be used with single-cell HR-3C data. Additionally, none of the available ensemble methods are usable for solving the 3D-GRP in the general case. Only two of the five ensemble methods provide source code and are actively maintained. These two methods also require additional biological datasets (DamID and/or LAD) for 3D genome prediction. These types of datasets are not commonly gathered with HR-3C assays; therefore, these solutions are not applicable in the general case. Finally, the web application 3D-GNOME only works with the pre-computed HR-3C datasets hosted on the website. As investigations into the 3D genome organization continue, it is possible that the existing tools can not be utilized for applications in organisms with larger, more complicated genomes (when compared to Homo sapiens). The reasons and potential solutions are described in the subsections below.

8.1 Computational Limitations

As mentioned previously, the current formulation of the 3D-GRP is a combinatorial optimization problem. Combinatorial optimization problems are known to be demanding in
terms of computational resources like memory. This is potentially problematic because it adds an upper bound on the number of genomic bins that can be input into existing tools based on available computational resources. This could render certain 3D-GRP solutions impractical for generating high-resolution predictions and/or predictions from organisms with genomes larger than *Homo sapiens*. For instance, these computational limitations cause polymer models to have a genome size and/or resolution limit (i.e. number of "beads"). For these polymer modelling based solutions, the current upper bound on the number of genomic regions that can be predicted has been reported to be 10,000 [111]. The majority of the existing tools have a $O(N^3)$ time complexity since they rely on MDS and/or the Floyd Warshall algorithm. As the resolution of GR-3C data increases so does the value for $N$. This will necessitate investigations into more efficient approaches. Fortunately, combinatorial optimization problems have been extensively studied in computer science and many of the existing solutions for solving these types of problems could be leveraged in 3D-GRP solutions. Existing tools like miniMDS have utilized a divide-and-conquer approach to overcome the computational limitations [77]. It is expected that approaches like this (as well as others that take advantage of parallelism or distributed algorithms) will become more common as advances in HR-3C assays continue to allow researchers to obtain finer genomic resolutions. Future research should focus on establishing solutions that are more computationally efficient and/or take advantage of parallel or distributed algorithms to overcome the current computational limitations.

### 8.2 Increasing Algorithmic Diversity

Algorithms for solving the 3D-GRP are far from maturity [97]. While there is some algorithmic diversity in the set of existing tools, the full breadth and depth of solutions in each category have yet to be explored. As mentioned above, five of the six consensus methods use a MDS as a part of their approach for solving the 3D-GRP. Many issues have been noted pertaining to the use of MDS as a part of solutions to the 3D-GRP. For instance, because HR-
3C assays represent a heterogeneous population of genome organizations, there is often not a single unique solution. Therefore, the distances calculated by MDS often conflict and cannot be accurately or completely calculated [99]. Furthermore, it is known that standard MDS techniques are inaccurate for sparse high-resolution data [77]. t-Stochastic neighbourhood embedding (tSNE) has been shown to be more accurate than MDS for datasets with these characteristics [112,115] and is a promising technique for new 3D-GRP solutions.

All of the existing methods utilize Euclidean distances in their solutions to the 3D-GRP but the utility of other distance functions (such as relative Sorensen distances, Canberra distances and cosine (similarity) distances) could and should be investigated going forward to increase the accuracy of predicted models. This is especially pertinent in the case of solutions to the 3D-GRP since it is known that Euclidean distances are often not suitable for sparse, high-dimensional datasets [116] which is the case with many whole-genome contact maps. Finally, most of the existing tools model the chromosome as a set of beads or beads-on-a-string/spring. While this seems like a natural representation, the utility of other chromosome models should be investigated.

In general, there is a lack of algorithmic diversity in the existing set of tools for solving the 3D-GRP. Figure 4 provides a visual depiction of the different algorithmic strategies employed by 3D-GRP solutions (purple boxes), 3D regional prediction solutions (orange boxes) and both (green boxes). Additionally, we highlight a few algorithmic strategies that, to the best of our knowledge, have not yet been utilized for predicting 3D structures of the genome or genetic region (grey boxes). In our opinion, these represent promising areas of exploration for new tool development but there are many other algorithms and algorithm types well-suited for combinatorial optimization problems that could also be investigated. While the community has made great strides in developing solutions to the 3D-GRP, a lot of work remains to be done as investigations into the 3D genome organization of non-model organisms begins.
See Figure 4 (attached)

**Figure 4:** An overview of the algorithmic techniques used by existing tools for solving the 3D-GRP (purple boxes), 3D regional prediction (orange boxes) and both (green boxes). A small selection of unexplored algorithmic strategies are indicated with grey boxes. Lines originate at a black dot and represent the hierarchical relationship between each algorithmic approach (more general to more specific).

### 8.3 Applications to other organisms

An increase in algorithmic diversity is necessary to facilitate 3D genome analysis in non-model organisms. As mentioned previously, many methods rely on the presence of previously proposed "hallmarks" of genomic organization like TADs for prediction. This is troubling since the presence of these "hallmarks" has not been verified in a wide variety of organisms. For instance, recently it was found that TADs are not present in certain plant species like *Arabidopsis thaliana* [85] and are therefore not a conserved hallmark of genome organization. Methods like miniMDS that rely on TADs or TAD-like structures for efficient computation would not be applicable to organisms like *Arabidopsis thaliana*.

Many of the existing tools have only been utilized with data generated from standard model organisms such as *Saccharomyces cerevisiae, Mus musculus* or *Homo sapiens*. Table 4 presents an overview of the datasources that have been used by existing tools for solving the 3D-GRP. They are separated with black outlines into the following groups based on their origin: simulated data, parasite, virus, bacteria, yeast, insect, worm, fish, chicken, mice, primate, human, plant. Data sources used in the original manuscript are represented with a grey box. Applications of the tool were determined by reviewing all of the original publications citing articles. The exact number of articles reviewed for each tool is provided in the second column of Table 4. Valid applications of a tool in a different organism and/or dataset than the original paper are indicated with purple (successful) or orange (unsuccessful) boxes. There are many organisms that have Hi-C data available but 3D genomic predictions have not been performed with any of it (Table 4 white boxes). Interestingly, at the time of publication there were over 3200 Hi-C datasets deposited in the Gene Expression Omnibus.
dataset, but complete 3D genome prediction has only been applied to less than 10 unique datasets. This provides an interesting area of future exploration and application in the 3D genomics community.

**Table 4**: An overview of the data sources that have been used by existing tools for solving the 3D-GRP. Tool name is provided in the first column and follows the same ordering presented in Table 2. The number of citations that were examined is given in column 2. Data sources are listed in the first row and have been separated (black outlines) into the following groups based on their origin: simulated data, parasite, virus, bacteria, yeast, insect, worm, fish, chicken, mice, primate, human, plant. Grey boxes represent the datasource that was used in the original manuscript. Applications of the tool in other organisms are indicated with purple (successful) or orange (unsuccessful) boxes. Datasets that have not been applied to a tool are indicated with a white box.

See Table 4 (attached)

None of the existing tools have been applied to organisms with a ploidy greater than 2. As such, it is not clear whether these tools can be effectively utilized for predicting 3D genome structure in organisms with higher ploidies such as *Triticum aestivum* (bread wheat; hexaploid) [117]. Additionally, many of the 3D regional tools do not effectively deal with datasets from polyploid organisms and therefore, could not be applied to polyploid datasets (or extended to solve the 3D-GRP irregardless of computational complexity). This can be seen when looking at the applications presented in the original manuscripts of the regional tools where most chose to use either a haploid organism, pre-phased data, or a genomic region from the X chromosome of male cells. How to effectively deconvolute interaction signals from distinct chromosome copies (the ploidy problem) still remains a large, unanswered question in the field. While it may be possible to address this problem during read mapping and/or pre-processing steps, solutions built-in to 3D-GRP tools should also be investigated.

9 Conclusion

There has been a great deal of success predicting 3D genome organizations from HR-3C data originating from model organisms like *Saccharomyces cerevisiae* and *Homo sapiens*. 
Addressing the challenges outlined in Section 8 above will be crucial as the field continues to evolve and be extended to non-model organisms (especially ones with larger, non-standard genomes). The set of existing tools for solving the 3D-GRP is far from mature and cannot be applied to analyze 3D genome organization across various species. A tool that can be used to predict 3D genome structure across organisms is urgently needed. Many of the existing solution approaches in computer science for overcoming the difficulties associated with optimization problems like the 3D-GRP have not yet been explored. These types of solutions are likely to be an area of major development in the coming years within the 3D genome community. While a great deal of foundational work has been done, there is a clear lack of generalizable, algorithmically diverse computational tools for predicting the complete 3D genome organization from HR-3C data.

10 Key Points

- Many computational solutions exist for predicting 3D genome organizations in a select few model organisms.

- These existing tools cannot necessarily be applied to non-model organisms due to inherent constraints imposed by the underlying techniques.

- New tools are required to facilitate 3D genome organization studies in non-model organisms.

- There are many promising algorithmic areas that have not yet been applied to the 3D-GRP.

- There are many existing Hi-C datasets that have not been used to predict 3D genomic organization with existing tools.
11 Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada [RGPIN 37207 to AK, Vanier Canada Graduate Scholarship to KM].

12 Author Information

Kimberly MacKay is a Ph.D. candidate in the Department of Computer Science, University of Saskatchewan.

Anthony Kusalik is a Professor in the Department of Computer Science at the University of Saskatchewan. He also is the director of the Bioinformatics Program at the University of Saskatchewan.

13 Acknowledgements

We would like to thank Daniel Hogan, Morgan W.B. Kirzinger, Dr. Matthew Links, Dr. Ian McQuillan and Dr. Steve Robinson for reading the manuscript and providing invaluable feedback.
Appendix A

- Permission to reprint the images in Figure 1:
  - Diament and Tuller [74]: Creative Commons Attribution (CC BY) license.
  - Duan et al. [65]:
    * License Number: 4751560452148; Issued on January 17, 2020
  - Kapilevich et al. [76]:
    * License Number: 4703250178326; Issued on Nov. 6, 2019
  - miniMDS [77]: Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license.
  - Segal and Bengtsson [50]: Creative Commons Attribution (CC BY) license.
  - Stevens et al. [78]:
    * License Number: 4751560316055; Issued on January 17, 2020
  - Chrom3D [79]: Creative Commons Attribution (CC BY) license.
  - Kalhor et al. [31]:
    * License Number: 4703251132971; Issued on Nov. 6, 2019
  - Li et al. [80]: Creative Commons Attribution (CC BY) license.
  - LorDG [73]: Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license.
  - Tjong et al. [81]: Use of original figures published in PNAS is permitted for noncommercial use (such as review articles).
  - 3D-GNOME [82, 83]: Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license.

- Permission to reprint the images in Figure 2.
– BACH [70]: Creative Commons Attribution (CC BY) license.

– Chiariello et al. [89]: Creative Commons Attribution 4.0 International (CC BY 4.0) license.

– ChromSDE [63]:
  * Publisher contacted for clarification, license will be added prior to publication.

– PASTIS [71]: Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license.

– RPR [64]: Creative Commons Attribution (CC BY) license.

– ShRec3D [72, 93]:
  * License Number: 4751560134989; Issued on January 17, 2020

– ShRec3D+ [94]:
  * License Number: 4703250567885; Issued on Nov. 6, 2019

– SuperRec [37]: Creative Commons Attribution 4.0 International (CC BY 4.0) license.

– 3DChrom [87]:
  * License Number: 4757200895857; Issued on January 27, 2020

– tPAM [95]:
  * License Number: 4751551505582; Issued on January 17, 2020

– Zhang et al. [97]:
  * License Number: 4757201185227; Issued on January 27, 2020

• Permission to reprint the images in Figure 3.

– BACH-MIX [70]: Creative Commons Attribution (CC BY) license.

– Caudai et al. [98]: Creative Commons Attribution (CC BY) license.
- Chromosome3D [99]: Creative Commons Attribution 4.0 International (CC BY 4.0) license.

- ChromStruct 4 [100]:
  * License Number: 4757201469502; Issued on January 27, 2020

- GEM [101]: Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license.

- Giorgetti et al. [102]:
  * License Number: 4751550455329 ; January 17, 2020

- IMP [67, 69, 103]:
  * License Number: 4751550598254 ; January 17, 2020

- InfMod3DGen [104]: Creative Commons Attribution (CC BY) license.

- ISD [105]: Creative Commons Attribution (CC BY) license.

- MBO [106]: Creative Commons Attribution (CC BY) license.

- MCMC5C [68]: Creative Commons Attribution (CC BY) license.

- Meluzzi and Arya [107]: Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license.

- Nagano et al. [108]:
  * License Number: 4751551203420 ; January 17, 2020

- TADbit [103, 109]: Creative Commons Attribution (CC BY) license.

- Trieu and Cheng [110]: Creative Commons Attribution (CC BY) license.
Supplementary Information

13.1 Extended Methodology

- Manuscripts describing existing tools were identified through a google scholar (https://scholar.google.com/) literature search using the key phrases: "3D genome reconstruction problem", "3D genome prediction", "3D genome structure" and "3D genome organization". The results were restricted to papers published from 2006 (the year 5C was first described [22]) to September 1, 2019 and filtered to only include peer-reviewed manuscripts describing software for predicting 3D genome organizations or 3D regional organizations.

- Applications of existing tools were identified by examining all manuscripts that cited the original publications (as of January 8, 2020).

- Organisms with available Hi-C datasets (for Table 4) were determined by using "Hi-C" as a search term in the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/). Results were extracted on January 8, 2020.
References

[1] A. Rasim Barutcu, Andrew J. Fritz, Sayyed K. Zaidi, André J. van Wijnen, et al. C-ing the genome: A compendium of chromosome conformation capture methods to study higher-order chromatin organization. *Journal of Cellular Physiology*, 231(1):31–35, 2016.

[2] Masahiko Kuroda, Hideyuki Tanabe, Keiichi Yoshida, Kosuke Oikawa, et al. Alteration of chromosome positioning during adipocyte differentiation. *Journal of Cell Science*, 117:5897–5903, November 2004.

[3] Ishita S Mehta, Manelle Amira, Amanda J Harvey, and Joanna M Bridger. Rapid chromosome territory relocation by nuclear motor activity in response to serum removal in primary human fibroblasts. *Genome Biology*, 11(1):R5, January 2010.

[4] Ishita S Mehta, Christopher H Eskiw, Halime D Arican, Ian R Kill, et al. Farnesyltransferase inhibitor treatment restores chromosome territory positions and active chromosome dynamics in Hutchinson-Gilford progeria syndrome cells. *Genome Biology*, 12(8):R74, August 2011.

[5] Ishita S Mehta, Mugdha Kulashreshtha, Sandeep Chakraborty, Ullas Kolthur-Seetharam, et al. Chromosome territories reposition during DNA damage-repair response. *Genome Biology*, 14(12):R135, December 2013.

[6] Ferhat Ay, Evelien M. Bunnik, Nelle Varoquaux, Sebastiaan M. Bol, 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 Research*, 24:974–988, March 2014.

[7] Job Dekker. Regulation of gene expression through chromatin interaction networks. *Blood Cells, Molecules, and Diseases*, 38(2):135, 2007.
[8] Lyubomira Chakalova, Emmanuel Debrand, Jennifer A. Mitchell, Cameron S. Osborne, et al. Replication and transcription: shaping the landscape of the genome. *Nature Reviews Genetics*, 6(9):669–677, 2005.

[9] Yun Li, Ming Hu, and Yin Shen. Gene regulation in the 3d genome. *Human Molecular Genetics*, R2:R228–R233, 2018.

[10] Peter R. Cook and Davide Marenduzzo. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic Acids Research*, 46(19):9896–9906, 2018.

[11] Hyejung Won, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, 538(7626):523–527, 2016.

[12] Phillippa C. Taberlay, Joanna Achinger-Kawecka, Aaron T.L. Lun, Fabian A. Buske, et al. Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations. *Genome Research*, 26(6):719–731, 2016.

[13] Qianli Dong, Ning Li, Xiaochong Li, Zan Yuan, et al. Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. *The Plant Journal*, 94(6):1141–1156, 2018.

[14] Annette Denker and Wouter de Laat. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes & Development*, 30(12):1357–1382, 2016.

[15] Elzo de Wit and Wouter de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, 26:11–24, 2017.

[16] Satish Sati and Giacomo Cavalli. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, 126(1):33–44, 2017.
[17] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, February 2002.

[18] Melissa J. Fullwood and Yijun Ruan. ChIP-based methods for the identification of long-range chromatin interactions. *Journal of Cellular Biochemistry*, 107(1):30–39, 2009.

[19] Hugo Würtele and Pierre Chartrand. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Research*, 14(5):477–495, 2006.

[20] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38:1341–1347, October 2006.

[21] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, 38(11):1348–1354, 2006.

[22] Josée Dostie, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, et al. Chromosome conformation capture carbon copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, 2006.

[23] Jim R Hughes, Nigel Roberts, Simon McGowan, Deborah Hay, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics*, 46(2):205–212, 2014.

[24] Nicola H. Dryden, Laura R. Broome, Frank Dudbridge, Nichola Johnson, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Research*, 24(11):1854–1868, 2014.
[25] Roland Jäger, Gabriele Migliorini, Marc Henrion, Radhika Kandaswamy, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nature Communications, 6:6178, February 2015.

[26] Pelin Sahlén, Ilgar Abdullayev, Daniel Ramsköld, Liudmila Matskova, et al. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biology, 16:156, 2015.

[27] Petros Kolovos, Harmen JG van de Werken, Nick Kepper, Jessica Zuin, et al. Targeted chromatin capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. Epigenetics & Chromatin, 7:10, 2014.

[28] Maxwell R. Mumbach, Adam J. Rubin, Ryan A. Flynn, Chao Dai, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods, 13:919–922, 2016.

[29] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, et al. Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science, 326(5950):289–293, 2009.

[30] Jon-Matthew Belton, Rachel P. McCord, Johan Harmen Gibcus, Natalia Naumova, et al. Hi–C: A comprehensive technique to capture the conformation of genomes. Methods, 58:268–276, 2012.

[31] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, et al. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nature Biotechnology, 30(1):90–98, 2012.

[32] Imielinski lab (New York Genome Center) collaboration. Pore-C: using nanopore reads to delineate long-range interactions between genomic
loci in the human genome. Technical report, Oxford Nanopore Technologies, (Poster) 2018. URL https://nanoporetech.com/resource-centre/pore-c-using-nanopore-reads-delineate-long-range-interactions-between-genomic-loci?keys=MinION

[33] Jinlei Han, Zhiliang Zhang, and Kai Wang. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics*, 11:21, 2018.

[34] Steven Wingett, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*, 4:1310, 2015.

[35] Bryan R. Lajoie, Job Dekker, and Noam Kaplan. The hitchhiker’s guide to Hi-C analysis: Practical guidelines. *Methods*, 72:65–75, 2015.

[36] Kimberly MacKay, Anthony Kusalik, and Christopher H. Eskiw. GrapHi-C: graph-based visualization of Hi-C datasets. *BMC Research Notes*, 11(1):418, 2018. doi: https://doi.org/10.1186/s13104-018-3507-2.

[37] Yanlin Zhang, Weiwei Liu, Yu Lin, Yen Kaow Ng, et al. Large-scale 3D chromatin reconstruction from chromosomal contacts. *BMC Genomics*, 20(Suppl 2):186, 2019.

[38] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, 2012.

[39] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43:1059–1065, 2011.

[40] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, et al. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.
[41] Nicolas Servant, Nelle Varoquaux, Edith Heard, Emmanuel Barillot, et al. Effective normalization for copy number variation in Hi-C data. *BMC Bioinformatics*, 19:313, 2018.

[42] Philip A. Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *Journal of Numerical Analysis*, 33(3):1029–1047, 2012.

[43] Wenyuan Li, Ke Gong, Qingjiao Li, Frank Alber, et al. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics*, 31(6):960–962, 2015.

[44] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, et al. Normalization of a chromosomal contact map. *BMC Genomics*, 13:436, 2012.

[45] John C. Stansfield, Kellen G. Cresswell, Vladimir I. Vladimirov, and Mikhail G. Dوزmorov. HiCcompare: an R-package for joint normalization and comparison of Hi-C datasets. *BMC Bioinformatics*, 19(279), 2018.

[46] Hongqiang Lyu, Erhu Liu, and Zhifang Wu. Comparison of normalization methods for Hi-C data. *BioTechniques*, ahead of print, 2019. URL https://doi.org/10.2144/btn-2019-0105.

[47] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485 (7398):376–380, 2012.

[48] Yinxiu Zhan, Luca Mariani, Iros Barozzi, Edda G. Schulz, et al. Reciprocal insulation analysis of hi-c data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Research*, 27(3): 479–490, 2017.
[49] Elzo de Wit. TADs as the caller calls them. *Journal of Molecular Biology*, In press, 2019.

[50] Mark R. Segal and Henrik L. Bengtsson. Reconstruction of 3D genome architecture via a two-stage algorithm. *BMC Bioinformatics*, 16:373, 2015.

[51] Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[52] Joseph B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964.

[53] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[54] Robert W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345, 1962.

[55] Stephen Warshall. A theorem on boolean matrices. *Journal of the ACM*, 9(1):11–12, 1962.

[56] Donald B. Johnson. Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, 24(1):1–13, 1977.

[57] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[58] M. Srinivas and Lalit M. Patnaik. Genetic algorithms: A survey. *Computer*, 27(6):17–26, 1994.

[59] Jonathan Barzilai and Jonathan M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.
[60] Scott Kirkpatrick, C. Daniel Gelatt, and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[61] Harold Szu and Ralph Hartley. Fast simulated annealing. *Physics Letters A*, 122(3-4):157–162, 1987.

[62] Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[63] Zhizhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of Computational Biology*, 20(11):831–846, 2013.

[64] Yoshito Hirata, Arisa Oda, Kunihiro Ohta, and Kazuyuki Aihara. Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. *Scientific Reports*, 6:34982, 2016.

[65] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, et al. A three-dimensional model of the yeast genome. *Nature*, 465:363–367, 2010.

[66] James Fraser, Mathieu Rousseau, Mathieu Blanchette, and Josée Dostie. *Computing Chromosome Conformation*, pages 251–268. Humana Press, Totowa, NJ, 2010. ISBN 978-1-60761-854-6. doi: 10.1007/978-1-60761-854-6_16.

[67] Davide Baù and Marc A. Marti-Renom. Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods*, 58:300–306, 2012.

[68] Mathieu Rousseau, James Fraser, Maria A Ferraiuolo, Josée Dostie, et al. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, 12:414, 2011.

[69] Davide Baù and Marc A. Marti-Renom. Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosome Research*, 19:25–35, 2011.
[70] Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, et al. Bayesian inference of spatial organizations of chromosomes. *PLOS Computational Biology*, 9(1):e1002893, 2013.

[71] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30(12): i26–i33, 2014.

[72] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, et al. 3D genome reconstruction from chromosomal contacts. *Nature Methods*, 11:1141–1143, 2014.

[73] Tuan Trieu and Jianlin Cheng. 3D genome structure modeling by Lorentzian objective function. *Nucleic Acids Research*, 45(3):1049–1058, 2017.

[74] Alon Diament and Tamir Tuller. Improving 3D genome reconstructions using orthologous and functional constraints. *PLOS Computational Biology*, 11(5):e1004298, 2015.

[75] Hideki Tanizawa, Osamu Iwasaki, Atsunari Tanaka, Joseph R. Capizzi, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research*, 38(22):8164–8177, 2010.

[76] Viacheslav Kapilevich, Shigeto Seno, Hideo Matsuda, and Yoichi Takenaka. Chromatin 3D reconstruction from chromosomal contacts using a genetic algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[77] Lila Rieber and Shaun Mahony. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics*, 33(14):i261–i266, 2017.

[78] Tim J. Stevens, David Lando, Srinjan Basu, Liam P. Atkinson, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544:59–64, 2017.

[79] Jonas Paulsen, Monika Sekelja, Anja R. Oldenburg, Alice Barateau, et al. Chrom3D:
three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. 

Genome Biology, 18:21, 2017.

[80] Qingjiao Li, Harianto Tjong, Xiao Li, Ke Gong, et al. The three-dimensional genome organization of Drosophila melanogaster through data integration. Genome Biology, 18:145, 2017.

[81] Harianto Tjong, Wenyuan Li, Reza Kalhor, Chao Dai, et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. PNAS, 113(12):E1663–E1672, 2016.

[82] Przemyslaw Szalaj, Paul J. Michalski, Przemystaw Wroblewski, Zhonghui Tang, et al. 3D-GNOME: an integrated web service for structural modeling of the 3D genome. Nucleic Acids Research, 44:W288–W293, 2016.

[83] Przemysław Szałaj, Zhonghui Tang, Paul Michalski, Michal J. Pietal, et al. An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. Genome Research, 26:1697–1709, 2016.

[84] Mariana Sotelo-Silveira, Ricardo A. Chávez Montes, Jose R. Sotelo-Silveira, Nayelli Marsch-Martínez, et al. Entering the next dimension: Plant genomes in 3D. Trends in Plant Science, 23(7):598–612, 2018.

[85] Pengfei Dong, Xiaoyu Tu, Po-Yu Chu, Peitao Lü, et al. 3D chromatin architecture of large plant genomes determined by local A/B compartments. Molecular Plant, 10(12): 1497–1509, 2017.

[86] W. Kabsch. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Foundations and Advances, 32(5):922–923, 1976.

[87] Tong Liu and Zheng Wang. Measuring the three-dimensional structural properties of
topologically associating domains. In IEEE, editor, *Proceedings, IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 21–28, 2018.

[88] Cheng Peng, Liang-Yu Fu, Peng-Fei Dong, Zhi-Luo Deng, et al. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Research*, 41(19):e183, 2013.

[89] Andrea M. Chiariello, Carlo Annunziatella, Simona Bianco, Andrea Esposito, et al. Polymer physics of chromosome large-scale 3D organisation. *Scientific Reports*, 6: 29775, July 2016.

[90] James Fraser, Mathieu Rousseau, Solomon Shenker, Maria A. Ferraiuolo, et al. Chromatin conformation signatures of cellular differentiation. *Genome Biology*, 10:R37, 2009.

[91] Maria A. Ferraiuolo, Mathieu Rousseau, Carol Miyamoto, Solomon Shenker, et al. The three-dimensional architecture of hox cluster silencing. *Nucleic Acids Research*, 38(21): 7472–7484, 2010.

[92] Chenchen Zou, Yuping Zhang, and Zhengqing Ouyang. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome Biology*, 14:40, 2016.

[93] Jean-Baptiste Morlot, Julien Mozziconacci, and Annick Lesne. Network concepts for analyzing 3D genome structure from chromosomal contact maps. *EPJ Nonlinear Biomedical Physics*, 4:2, 2016.

[94] Jiangeng Li, Wei Zhang, and Xiaodan Li. 3D genome reconstruction with ShRec3D+ and Hi-C data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2):460–467, 2018.
[95] Jincheol Park and Shili Lin. Statistical inference on three-dimensional structure of genome by truncated Poisson architecture model. *Ordered Data Analysis, Modeling and Health Research Methods*, 149:245–261, 2015.

[96] Jincheol Park and Shili Lin. Impact of data resolution on three-dimensional structure inference methods. *BMC Bioinformatics*, 17:70, 2016.

[97] Rongrong Zhang, Ming Hu, Yu Zhu, Zhaohui Qin, et al. Inferring spatial organization of individual topologically associated domains via piecewise helical model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[98] Claudia Caudai, Emanuele Salerno, Monica Zoppè, and Anna Tonazzini. Inferring 3D chromatin structure using a multiscale approach based on quaternions. *BMC Bioinformatics*, 16:234, 2015.

[99] Badri Adhikari, Tuan Trieu, and Jianlin Cheng. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC Genomics*, 17:3210–3214, 2016.

[100] Claudia Caudai, Emanuele Salerno, Monica Zoppe, Ivan Merelli, et al. ChromStruct 4: A python code to estimate the chromatin structure from Hi-C data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[101] Guangxiang Zhu, Wenxuan Deng, Hailin Hu, Rui Ma, et al. Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic Acids Research*, 46 (8):e50, 2018.

[102] Luca Giorgetti, Rafael Galupa, Elphège P. Nora, Tristan Piolot, et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157:950–963, 2014.
[103] Marie Trussart, François Serra, Davide Baù, Ivan Junier, et al. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. *Nucleic Acids Research*, 43(7):3465–3477, 2015.

[104] Siyu Wang, Jinbo Xu, and Jianyang Zeng. Inferential modeling of 3D chromatin structure. *Nucleic Acids Research*, 43(8):e54, 2015.

[105] Simeon Carstens, Michael Nilges, and Michael Habeck. Inferential structure determination of chromosomes from single-cell Hi-C data. *PLOS Computational Biology*, 12(12):e1005292, 2016.

[106] Jonas Paulsen, Odin Gramstad, and Philippe Collas. Manifold based optimization for single-cell 3D genome reconstruction. *PLOS Computational Biology*, 11(8):e1004396, 2015.

[107] Dario Meluzzi and Gaurav Arya. Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Research*, 41(1):63–75, 2012.

[108] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502:59–64, October 2013.

[109] Francois Serra, Davide Baù, Mike Goodstadt, David Castillo, et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors reveals structural features of the fly chromatin colors. *PLOS Computational Biology*, 13(7):e1005665, 2017.

[110] Tuan Trieu and Jianlin Cheng. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research*, 42(7):e52, 2014.

[111] Mariliis Tark-Dame, Roel van Driel, and Dieter W. Heermann. Chromatin folding – from biology to polymer models and back. *Journal of Cell Science*, 124:839–845, 2011.
[112] Laurens van der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(Oct):3321–3245, 2014.

[113] Laurens van der Maaten and Geoffrey Hinton. Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55, 2012.

[114] Laurens van der Maaten. Learning a parametric embedding by preserving local structure. In *Proceedings, Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS)*, pages 384–391, Clearwater, Florida USA, 2009. PMLR.

[115] Laurens van der Maaten and Geoffrey Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[116] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory — International Conference on Database Theory 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[117] Aleksey V. Zimin, Daniela Puiu, Richard Hall, Sarah Kingan, et al. The first near-complete assembly of the hexaploid bread wheat genome, triticum aestivum. *GigaScience*, 6(11):1–7, 2017.