Evaluation and comparison of methods for recapitulation of 3D spatial chromatin structures

Jincheol Park and Shili Lin

Corresponding author. Shili Lin, Department of Statistics, The Ohio State University, Columbus, OH 43210, USA. E-mail: shili@stat.osu.edu

Abstract

How chromosomes fold and how distal genomic elements interact with one another at a genomic scale have been actively pursued in the past decade following the seminal work describing the Chromosome Conformation Capture (3C) assay. Essentially, 3C-based technologies produce two-dimensional (2D) contact maps that capture interactions between genomic fragments. Accordingly, a plethora of analytical methods have been proposed to take a 2D contact map as input to recapitulate the underlying whole genome three-dimensional (3D) structure of the chromatin. However, their performance in terms of several factors, including data resolution and ability to handle contact map features, have not been sufficiently evaluated. This task is taken up in this article, in which we consider several recent and/or well-regarded methods, both optimization-based and model-based, for their aptness of producing 3D structures using contact maps generated based on a population of cells. These methods are evaluated and compared using both simulated and real data. Several criteria have been used. For simulated data sets, the focus is on accurate recapitulation of the entire structure given the existence of the gold standard. For real data sets, comparison with distances measured by Fluorescence *in situ* Hybridization and consistency with several genomic features of known biological functions are examined.

Key words: 3D structure; contact maps; dependency; data resolution

Introduction

Large-scale three-dimensional (3D) organization of chromosomes and the recapitulations of such 3D structures have been hotly pursued in the past few years owing to the rapid advancement of molecular technology. Initiated with the seminal work that launched the Chromosome Conformation Capture (3C) [1], a number of 3C-based high-throughput genome-wide assays have been proposed, including Hi-C [2], TCC [3], *in situ* Hi-C [4], Capture-C [5] and Next-generation (NG) Capture-C [6]. In particular, Hi-C, TCC and *in situ* Hi-C interrogate ‘all-versus-all’ chromatin interactions, in that all interactions between any two fragments in the entire genome are captured in these assays [6]. The interest in the 3D nuclear structure stems from the hypothesis that the nonrandom organization of such structures is linked to the biological functions of the cells [7], and in particular, the organization is cell-type dependent [8]. Table 1 provides a list of commonly used assays for detecting genome-wide chromatin interactions for a population of typically millions of cells.

Many biological questions have mushroomed following the exciting development of 3C-based technologies. One of the fundamental questions is how to unearth physical interactions between distal regulatory elements and their targeted gene promoters [6]. The extent to which answers can be obtained largely depends on the resolutions of the data. With low resolution (e.g. 1 mb) Hi-C data [2], the focus is on large-scale structure in the form of topologically associated domains (TADs) [7]. For higher resolution data (e.g. tens of kb's) [4, 6], it is possible to detect interactions between regulatory elements and those within gene bodies. Balancing between cost and coverage, Hi-C remains the most popular for determining genome-wide contact maps [6], and we note that higher resolution Hi-C data at 40 kb have also been produced [7]. Regardless of the resolution, a genome-wide ‘all versus all’ contact map is usually represented as a matrix of contact counts, also known as a contact matrix, with each entry denoting the interaction frequency between two genomic fragments, called loci, of a fixed length dictated by the resolution. Therefore, the size of the contact matrix is determined by the resolution of the data: the higher

Jincheol Park is a tenure-track Assistant Professor in the Department of Statistics at Keimyung University in Daegu, South Korea.

Shili Lin is a Professor of Statistics in the Department of Statistics at the Ohio State University, USA.

Submitted: 8 August 2017; Received (in revised form): 10 September 2017

© The Author 2017. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Table 1. Commonly used assays for interrogating genome-wide interactions

Assay	All versus all	Selected viewpoints ^a	Reference
Hi-C	✓		[2]
TCC	✓		[3]
In situ Hi-C	✓		[4]
Capture-C		✓	[5]
NG Capture-C		✓	[6]

^aAll interactions connected to hundreds of viewpoints are interrogated.

Table 2. Properties of methods in terms of their ability to handle inherent features in contact maps

Type	Method	Dependency	Over-dispersion	Zero inflation	Cell heterogeneity
Optimization	ChromSDE	✓	✓	✓	X
	ShRec3D	✓	✓	✓	X
Model	BACH	X	X	X	✓
	tPAM	X	X	X	✓
	PASTIS	X	X	X	X
	PRAM	✓	✓	X	✓
	tREX	✓	✓	✓	✓

the resolution, the larger the size of the matrix. To address many of the biological questions of interest, a fundamental task is to recapitulate the underlying 3D spatial structure of chromatin based on a two-dimensional (2D) matrix of interaction map. Such 3D organization of a genome is intimately related to its biological function [7]. It is particularly relevant for the investigation of chromatin packing, which has significant implications in understanding and delineating closed and open chromatin regions [9]. This in turn is important for investigating long-range gene regulations.

A number of data analysis methods (Table 2) have been proposed for 3D chromatin structure recapitulations [10–16]. As the technologies become mature and used routinely in scientific laboratories, tools for data analysis need to be developed to meet the demands. As such, it is imperative that unbiased and comprehensive evaluations be carried out to understand the advantages and disadvantages of the various methods in terms of data resolution and how they handle various inherent features (listed in Table 2) in the contact maps. This is the task that we take up in this article. Specifically, this article considers both optimization-based as well as model-based methods for analyzing Hi-C contact maps that are generated from a population of cells. We will make recommendations on which methods are the most appropriate for analyzing which types of data.

Data, methods and assessment criteria

Data structure and problem

Data produced from the original Hi-C technology and its variations are essentially contact count frequencies, which are typically organized into a square matrix with its dimension determined by the resolution of the data. More specifically, for a given resolution, each of the chromosomes is divided into

segments (referred to as loci) of the desired resolution (e.g. 1 mb in length); interaction frequencies between any two loci in the genome are therefore recorded and entered into the corresponding entries in the matrix. As such, the higher the resolution (more loci), the higher the dimensions of the matrix. One way to visualize the data matrix is through a heatmap. Figure 1 is an example, where the interaction frequencies among loci in chromosomes 14 and 22 in the contact matrix are represented by gray level intensities. Three features underlie the data structure. First, the entries in the matrix are interdependent with complex relationships dictated by the underlying biological mechanism. Second, the entries in the matrix are counts, and therefore, it is natural to model them using a Poisson distribution; however, there may be over-dispersion. Third, the genome-wide contact matrix may be sparse with many 0 entries in the matrix, as there are many loci that do not interact, especially between two different chromosomes. For example, Figure 1 shows low interaction frequencies between loci in chromosomes 14 and 22. Also, as data resolution increases, the matrix becomes sparser. Shown in Figure 1 is a contact matrix based on 1 mb resolution; a higher resolution matrix (say 40 kb) will undoubtedly lead to a sparser matrix. However, for higher resolution data with typical sequencing depth, one has to be mindful of whether a zero in the contact matrix is indeed a structural 0 (i.e. the two loci indeed do not interact) or simply due to insufficient sequencing depth.

Suppose one is interested in the 3D chromatin structure of n loci in the genome such as a particular set of chromosomes or the entire genome. Then, their relative position in the 3D space, up to a scale translation, is of interest. Let us assume that the 3D coordinates of these n loci are denoted by $S \equiv \{s_i = (s_i^x, s_i^y, s_i^z); i = 1, \dots, n\}$. Then, the goal is to make inference about S , that is, to provide estimates of S given the contact matrix $\mathcal{M} = \{y_{ij}, 1 \leq i < j \leq n\}$. Note that the data matrix $\mathcal{M}$ is symmetric; therefore, only the data from the upper triangle are unique and used in the analysis.

There are two major categories of methods for making inference about S from $\mathcal{M}$. One is optimization-based, in which S is sought by minimizing the total difference between pairwise distances in the hypothesized S and the corresponding distances ‘translated’ from the observed $\mathcal{M}$. The other major type of methods is model-based, in which the observed data in $\mathcal{M}$ are assumed to follow some distribution whose characteristics (say mean) are related to the pairwise distances in the hypothesized S . Regardless of which type of methods is used, the following biophysical property plays a central role: the 3D distance between two loci is inversely proportional to their contact frequency. In the following two subsections, we describe some recent and/or well-regarded methods in each of the two categories. We will point out their advantages and disadvantages from a theoretical point of view. Empirical studies of these methods will be presented in later sections.

Optimization-based methods

Metric Multidimensional Scaling (MDS) is a classical method used in estimating coordinates of points given their approximate pairwise Euclidean distances [17]. Nonmetric MDS has also been used for estimating chromatin structures [12]. To use metric MDS in the context of DNA structure inference from Hi-C data, one needs to first ‘translate’ a contact count y_{ij} to a physical pairwise distance d_{ij} representing the Euclidean distance in the classical MDS terminology. With this set of translated

Figure 1. Heatmap of a 1-mb resolution Hi-C contact matrix of chromosomes 14 and 22 jointly; the data were generated using a human lymphoblastoid cell line [2].

distances $\mathcal{D} = \{d_{ij}, 1 \leq i < j \leq n\}$ treated as known, one then aims to find the set of S that minimizes the following unified objective function. That is,

$$\mathcal{S}^* = \operatorname{argmin}_{\mathcal{S}} \sum_{\{(i,j): d_{ij} < \infty\}} w_{ij} f(\|s_i - s_j\|^a - d_{ij}^a) - \lambda \sum_{\{(i,j): d_{ij} = \infty\}} \|s_i - s_j\|^2, \quad (1)$$

where $\|\cdot\|$ is the Euclidean distance; a is a constant typically set to be 1 or 2 and $f(A)$ is a function of A , typically set to be an identity, absolute or squared function (more discussion later); the set of $\{w_{ij}\}$ are weights used to factor in information on reliability of each datum; and λ , when set to be greater than zero, plays the role of regularization to place loci that do not interact to be as far from each other as possible. On the other hand, if $\lambda = 0$ is specified, then no regularization is exercised.

Two of the frequently used methods under this category are ChromSDE [10] and ShRec3D [13]. Their main difference is in the translation of the contact frequencies to 3D Euclidean distances, although they also differ in the parameter setting in the optimization criterion (1). Specifically, for ChromSDE,

$$d_{ij} = \begin{cases} \alpha(1/y_{ij})^\beta & \text{if } y_{ij} \geq 0; \\ \infty & \text{otherwise,} \end{cases}$$

where β is a positive parameter that will be maximized together with the coordinates in S . Indeed, the above translation between the contact frequency and 3D distance observes the biophysical property as described above. In the second step, the weights in (1) are chosen to be inversely proportional to the translated distance, d_{ij} . Further, a is set to equal 2, while two forms of the f function were considered: $f(A) = A$ and $f(A) = |A|$.

These choices make it possible to solve the problem, alleviating the NP-hard label. In particular, for computational efficiency, the optimization criterion in (1) is relaxed as linear or quadratic semidefinite programming [10]. Finally, the penalization parameter λ is pre-set as a constant 0.01 [10], reflecting the regularization feature of ChromSDE.

In ShRec3D [13], a different approach is taken to translate the contact matrix $\mathcal{M}$ to a distance matrix $\mathcal{D}$, which relies on the concept of ‘shortest path’ in graph theory. In a slight abuse of notation, we will use S to denote the points in a 3D space, which will also be described as the nodes of a graph in the 3D space. Nodes s_i and s_j are said to be connected if there exists a path $(i_0, \dots, i_k)$ with $i_0 = i$ and $i_k = j$ for some positive integer k such that $y_{i_q, i_{q+1}} \neq 0$ for all $q = 1, \dots, k-1$. Its path length is then

defined to be $PL(i_0, \dots, i_k) = \sum_{q=0}^{k-1} 1/y_{i_q, i_{q+1}}$. Let $\mathcal{P}_{ij}$ denote the collection of all paths connecting s_i and s_j . Then the ‘translated’ 3D distance is defined as the shortest path length as follows:

$$d_{ij} = \min_{\mathcal{P}_{ij}} \{PL(i_0, \dots, i_k), (i_0, \dots, i_k) \in \mathcal{P}_{ij}\}.$$

For any pair of nodes that are not connected by any path, $d_{ij} = \infty$. The shortest paths and the corresponding path lengths can be found efficiently by using the Floyd–Warshall algorithm [13]. To optimize the objective function (1), in the second step, all weights are assumed to be 1. As ShRec3D does not include a penalty term, λ in (1) is set to be 0; in other words, there is no regularization. Further, the constant a is set to be 1, and $f(A) = A^2$, following the classical MDS paradigm [13].

The major advantage of optimization-based methods is that there is no assumption of independence among the contact frequencies, and there is no concern regarding whether a model is

able to accommodate potential over-dispersion, as such methods are nonparametric. Further, ChromSDE [10] also adopts a regularization procedure, which may address the issue of sparsity. However, finding a single consensus structure overlooks the issue of population heterogeneity, as millions of cells are typically used in a single Hi-C experiment.

Model-based methods

Model-based methods typically use Poisson regression to model the observed contact frequencies, as the data are counts and a Poisson distribution is therefore a natural fit [11–16]. One exception is MCMC5C, which assumes a normal distribution for the counts [9]. Because existing methods predominantly rely on the Poisson distribution, the methods that we have selected in this review all use this distribution. Observing the biophysical property as described above, these methods can be represented using the following unified model for describing the relationship between the mean intensity parameter of the Poisson distribution (ζ_{ij}) for a pair of loci (i and j) and the unobserved 3D distance (d_{ij}) inferred from the 3D structure S :

$$\log \zeta_{ij} = \alpha_0 + \alpha_1 \log d_{ij} + \mathbf{x}_{ij}^T \boldsymbol{\zeta} + fV_{ij}, \quad (2)$$

where α_0 is an offset; $\boldsymbol{\zeta}$ is a vector of regression coefficients for the set of covariates coded in the design matrix $\mathbf{x}_{ij}^T$; and f is a 1/0 flag to signal whether a random effect component, V_{ij} , is included in the model. Most importantly, the restriction of the coefficient $\alpha_1 < 0$ is imposed to reflect the biophysical property, as such a constraint represents an inverse relationship between the 3D distance and the contact frequency [2, 10, 16]. The inclusion of covariates in the model can be used for the purpose of normalization; typical factors considered for this purpose are restriction enzymes, fragment lengths, GC contents and mappability scores, which have been studied for their ability to correct systematic biases [11, 18]. The random effect component, on the other hand, can account for two of the three data structure features, namely, dependency and potential over-dispersion [15, 16]. Specifically, a positive correlation can be induced between two pairs of loci that share a common locus. Moreover, the introduction of the random component can address the issue of over-dispersion in a data-adaptive fashion in that if there is indeed extra variability in the data, then the expectation of the contact counts will be smaller than its variance, a departure from the independent Poisson model. On the other hand, if the overall variability is small, the data are not over-dispersed and will not be forced to have a larger variance [16].

Based on model (2), we can write down the log-likelihood as follows, noting that the ζ_{ij} contains the potential random effect term:

$$l(\Phi | \mathcal{M} = \{y_{ij}, 1 \leq i < j \leq n\}) \propto \sum_{(i,j) \in \mathcal{I}} \{y_{ij} \log \zeta_{ij} - \log(e^{\zeta_{ij}} - I(|\mathcal{I}| \neq |\mathcal{M}|))\}, \quad (3)$$

where Φ is the collection of all model parameters, including the collection of 3D coordinates S , and $\mathcal{I}$ is an index set with $|\mathcal{I}|$ denoting its cardinality. When the index set encompasses the entire data set, that is, $|\mathcal{I}| = |\mathcal{M}|$, all contact frequencies in $\mathcal{M}$ are being used. For low resolution data, such as data from the originally proposed Hi-C technology [2], there is no excess of zeros for intra-chromosomal interaction data, and therefore, it

is justifiable to use the entire data set. This is essentially the strategy taken by three model-based methods, BACH [11], PASTIS [12] and tPAM [14], all assuming $f=0$ (i.e. no random effect) in addition to assuming no inflation of zero (i.e. Poisson model). While BACH and PASTIS use the entire data set for inference, tPAM, recognizing the potential problem with excess of zeros for inter-chromosomal and/or higher resolution data, considers index set $\mathcal{I} = \{(i,j) : y_{ij} > 0\}$, leading to a truncated Poisson model. This particular choice of an index set is also adopted by tREX for better model fitting. It is further observed that an index set of such nature also leads to computational advantages [15]. Finally, we note that other choices of the index set are also possible, such as only considering counts that exceed a positive threshold to filter out unlikely contacts as adopted in a different investigation [19].

Although the argument of no inflation of zero for intra-chromosomal and low resolution data may be substantiated, dependency among contact counts is inherent in all chromatin interaction data; and therefore, setting $f=0$ may lead to poor fitting of the reduced model in (3) to the observed data. To this end, PRAM [16] and tREX [15] have been proposed to address this problem, essentially setting $f=1$ in (3). While PRAM assumes that the entire set of data can be fitted with a Poisson model, tREX allows the flexibility of using an index set that is a subset of $\mathcal{M}$. In summary, from a modeling point of view, BACH, PASTIS, tPAM and PRAM are all special cases of tREX, which provides a unified approach that can be flexibly adapted to data resolution and genomic features being studied (intra-chromosomal structure only or beyond).

The methods described above diverge into two categories in their statistical inference. PASTIS [12] takes the view of optimization-based approaches by providing a single consensus structure through maximum likelihood estimates of the model parameters. On the other hand, the remaining of the model-based methods all cast the problem into a Bayesian inference framework so that multiple structures may be inferred to address the issue of cell heterogeneity [11–16]. Markov chain Monte Carlo algorithms were devised to sample from the space of chromatin structures; individual or average structures may then be used for understanding biological functions.

Performance assessment

In the presence of gold standard

For the simulation study carried out in the ‘Simulation Study’ section, as the ground truth, that is, the 3D structure S , is known, we can assess how well the estimated structure, $\hat{S}$, conforms to the known underlying structure. Following a common practice [16], we first match $\hat{S}$ to S through appropriate geometric transformations because the methods described in this article can only estimate S up to a scale parameter and the structure may be translated, reflected or rotated. We then use two measures to assess the performances of the methods. The first one is the root mean square of the Euclidean distances (RMSD) between the corresponding points in S and the transformed $\hat{S}$. We also compute the correlation between the pairwise distances in S and those in $\hat{S}$, leading to our second measure for performance assessment.

When experimental data are available

On the other hand, for the real data analyses performed in the ‘Analysis of Human Lymphoblastoid Hi-C Data’ section, there no longer exists any gold standard. Florescence in situ Hybridization (FISH) measurements of pairwise distances for

some pairs exist. Their correlation with the corresponding distances inferred from $\hat{S}$ is then computed as an assessment of the performance of the methods.

Considering structural property

For real data analysis, as the underlying structure is unknown, it is crucial to gauge the accuracy of the reconstructed structure using additional criteria besides FISH, as FISH data are typically quite limited as we will see in the ‘Analysis of Human Lymphoblastoid HI-C Data’ section. One avenue is to assess the structural property of an estimated architecture. To achieve this goal, we define a measure called Chromatin Activity Trending (CAT) at each locus, which is defined as

$$\text{CAT}_{d_0}(i) = \frac{1}{n} \sum_{j=1}^n I(d_{ij} > d_0), \quad (4)$$

where d_{ij} is the 3D distance between loci i and j as defined before; the threshold d_0 controls the degree of localization, and is typically taken to be a small quantile (say < 0.1) of the distribution of estimated distances. We can see that a larger CAT score for a fixed d_0 at locus i ($\text{CAT}_{d_0}(i)$) is indicative of less interaction with other loci locally. This means that the chromatin is more sparsely packed around that locus, which may be viewed as signifying an ‘open’ chromatin and thus greater activity, a property that is of great biological interest [2, 9]. In particular, this measure can be used to study consistency between structural property and genomic function, as we demonstrate in our real data analysis in the ‘Analysis of Human Lymphoblastoid HI-C Data’ section.

Preservation of TADs

TAD is an important feature that has been used to describe and delineate 3D chromatin structures [7]. Loci within the same TAD interact much more than loci located between different TADs. Therefore, evaluating how well TADs are preserved provides another crucial criterion for ascertaining the performance of the methods, which is also adopted in our analysis in the ‘Analysis of Mouse Stem Cell Hi-C Data’ section.

Simulation study

We consider simulated data that can be labeled as low resolution (e.g. 1 mb resolution as in typical Hi-C data [2]) or high resolution (e.g. 10 kb resolution as in *in situ* Hi-C data [4]) by setting the appropriate proportion of zero inflation. For low resolution data, the methods being compared are BACH, ChromSDE, PASTIS, ShRec3D and PRAM (essentially tREX without invoking its truncation feature). For high resolution data, BACH is replaced by its truncated counterpart tPAM and the truncated version of tREX is used; therefore, the methods being compared are tPAM, ChromSDE, PASTIS, ShRec3D and tREX (the truncated version). We choose to compare a different set of methods depending on which resolution the data are simulated under because tPAM and tREX (with truncation) are proposed specifically for higher resolution data only and should be used only in such settings.

Simulation setups

To mimic real data but to reduce computational burden with a large number of settings and replicates, we use the estimated structure on a segment of chromosome 14 (the first 50 loci) as the known architecture, which we denote as CH14⁵⁰. We

generate contact frequencies between loci i and j ($1 \leq i \leq j \leq 50$) from a Poisson distribution with the following model for its log-mean intensity parameter:

$$\log \xi_{ij} \equiv \alpha_0 + \alpha_1 \log d_{ij} + \zeta_i \log(x_{li}x_{lj}) + \zeta_g \log(x_{gi}x_{gj}) + \log(x_{mi}x_{mj}) + \delta(Z_i + Z_j + W_{ij}).$$

In the above model, x_l , x_g and x_m represent three covariates, whose values are simulated as follows: $x_{li} \sim \mathcal{U}(0.3, 0.8)$, $x_{gi} \sim \mathcal{U}(0.3, 0.8)$ and $x_{mi} \sim \mathcal{U}(0.9, 1)$, where $\mathcal{U}$ denotes a uniform distribution. These covariates are chosen to mimic fragment length (x_l), GC content (x_g) and mappability score (x_m) from the real data [2]. Further, δ may take two different values: 0 or 1. When it is zero, the resulting simulation model is referred to as NR. When it is 1, we set $W_{ij} \sim \mathcal{N}(0, 0.05^2)$ to generate over-dispersion; for Z_i and Z_j we consider two distributions: $\mathcal{N}(0, 0.2^2)$ and skewed- t (scale = 0.3, shape = 1, df = 10) and refer to them as R1 and R2, respectively. Guided by real data again, we set $\alpha_0 = 3.5$ and consider the following two α_1 values: -0.5 and -1.0 . Finally, for low resolution data, we assume that the contact frequency random variable $Y_{ij} \sim \text{Poisson}(\xi_{ij})$ guided by real data, as there is no excess of zeros. On the other hand, a zero inflated (20% zero inflation) Poisson distribution is used to generate high resolution data. Replicated data sets of 50 are simulated according to the above description for each combination of the parameter α_1 ($-0.5, -1.0$), resolution (low and high) and random effect component (NR, R1, R2).

Results

Table 3 summarizes the performances of the methods for the data generated with $\alpha_1 = -0.5$. The numbers in the table are the medians of RMSD or Correlation, over 50 replicates between the estimated and those derived from the true structure; the standard deviations are provided in the parentheses. To better visualize the results and to see the variability among the replicates, we also plotted the RMSD against Correlation (Figure 2). From the table and figure, several observations can be made. First, for the two models with random effect components, R1 and R2, tREX and PRAM are seen to have the best performance, with smaller RMSD and larger Correlation, as seen in the last two super-columns of Table 3 and Figure 2C–F. There is a bit of variability from replicate to replicate for all the methods although the superiority of tREX (and PRAM as a special case) can still be clearly seen. These results are not surprising, as tREX is designed to capture dependency and over-dispersion in Hi-C data. What is more interesting is that, Figure 2E–F shows that tREX is robust to the random effect structure, having even greater performance improvement over the other methods when the underlying random effect distribution is a skewed- t rather than a normal. When the interaction frequencies in the contact matrix are assumed to be independent, for low resolution data, PASTIS and ChromSDE are seen to have better performance in terms of Correlation, although they have slightly elevated RMSDs (Figure 2A). Most interestingly, for high resolution data, PASTIS has a large proportion of negative Correlation and larger RMSDs, whereas all the other methods have good performance with large correlations (Figure 2B). Notably, ChromSDE performs well for both low and high resolution data (Figure 2A, B), which demonstrates the advantage of an optimization procedure with regularization. Results for $\alpha_1 = -1.0$ convey the same message qualitatively and are thus omitted, although we would like to point out that PASTIS did not return a

Table 3. Performance comparison of five methods under the model setup $\beta_1 = -0.5$ using median^a RMSD and correlation between the true structure and the estimated structures

		NR		R1		R2	
		RMSD	Correlation	RMSD	Correlation	RMSD	Correlation
0%	PRAM	4.44 (1.106)	0.90 (0.036)	4.94 (0.876)	0.77 (0.080)	4.87 (0.888)	0.72 (0.085)
	BACH	4.25 (0.981)	0.84 (0.030)	6.48 (1.120)	0.51 (0.100)	7.18 (1.031)	0.31 (0.107)
	ShRec3D	5.18 (1.497)	0.87 (0.018)	6.30 (1.059)	0.67 (0.094)	NA ^b	NA ^b
	PASTIS	4.71 (1.478)	0.95 (0.008)	6.47 (1.188)	0.60 (0.080)	6.87 (1.087)	0.38 (0.091)
	ChromSDE	5.08 (1.645)	0.95 (0.008)	6.10 (1.224)	0.60 (0.076)	6.59 (1.032)	0.39 (0.089)
20%	tRex	4.54 (1.199)	0.89 (0.033)	5.37 (0.926)	0.73 (0.070)	5.37 (1.077)	0.68 (0.099)
	tPAM	5.01 (1.275)	0.90 (0.039)	6.90 (1.168)	0.51 (0.111)	7.21 (0.941)	0.33 (0.089)
	ShRec3D	5.75 (1.269)	0.83 (0.028)	7.51 (1.233)	0.58 (0.088)	NA ^b	NA ^b
	PASTIS	7.69 (1.845)	-0.18 (0.604)	7.39 (1.082)	0.54 (0.116)	7.45 (0.932)	0.31 (0.090)
	ChromSDE	5.13 (1.560)	0.94 (0.012)	6.95 (1.123)	0.57 (0.081)	7.27 (1.039)	0.37 (0.079)

^aThe number in the parentheses following the median denotes standard deviation over the replications.

^bThe NAs are owing to the fact that ShRec3D failed to run for some data generated under R2.

negative correlation for higher resolution data with this α_1 value.

Analysis of human lymphoblastoid Hi-C data

Comparison with FISH measurements

We evaluate the performances of the methods using the Hi-C data library of human lymphoblastoid cell line [2]. In addition to the Hi-C data, experimental validation data based on FISH are also available from the same cell line [2], which are treated as the gold standard for our assessment. Specifically, there are FISH measurements of 3D distances between 9 pairs of loci, involving 4 loci on chromosome 14 and 4 loci on chromosome 22. Correspondingly, we extracted the subset of the Hi-C library that corresponds to chromosomes 14 and 22, and applied each of the methods to construct the genomic 3D structure involving these two chromosomes. For methods that only produce a single consensus structure (ChromSDE, PASTIS and ShRec3D), we first normalized the contact matrix by using HiCNorm [20] and used the normalized data as input to the methods. For the rest of the methods, we used fragment lengths, GC contents and mappability scores in the model to account for potential biases in lieu of a prior normalization step.

For each method, based on the estimated 3D structures, we obtained the estimated distances for the pairs for which FISH measurements of distances are available. Then Pearson correlation as well as Spearman correlation between the inferred distances and the measured distances by FISH measurements are calculated and plotted in Figure 3. The results clearly show that tREX has a much higher Spearman correlation than the rest of the methods, although it has similar Pearson correlation with BACH and PASTIS. The two optimization-based methods have both lower Pearson and Spearman correlations.

Whole genome analysis and correlation with genomic features

To further evaluate the methods, we applied them to the Hi-C data in the whole genome chromosome by chromosome. As FISH data are only available for several loci on chromosomes 14 and 22 while chromatin packing density, measured by the CAT score, has important implications for transcription regulation, we analyzed the CAT scores to explore whether locus-specific

spatial characteristics predicted by our estimated structure are consistent with known genomic and epigenetic features. For each of the methods, we calculated the CAT_{d₀} scores for each locus over a range of d_0 values, which were selected as various quantiles (0.01 to 0.1 with an increment of 0.01) of the distribution of the distances derived from S . We then performed a clustering analysis based on the k-means algorithm. Based on the criterion of average silhouette width [21], the best clustering result was when there were two clusters, which we interpret as representing densely packed (smaller CAT scores) and sparsely packed (larger CAT scores) groups. As the latter may be correlated with open chromatin, we further investigated whether this classification is associated with important genomic features. Specifically, we considered H3K36me3, H3K4me3 and DNaseI sensitivity (measuring chromatin accessibility). As can be seen in Figure 4, for both BACH and tREX (Figure 4A, E), the sparsely packed group is associated with greater chromatin activities (signal intensity), with the Wilcoxon rank sum tests all achieving significance at the 0.001 level, although the tREX results are a bit more significant. In contrast, for ChromSDE, PASTIS and ShRec3D (Figure 4B–D), the associations, if any, are much weaker. Specifically, for H3K36me3, ChromSDE is significant at the 0.01 level while PASTIS and ShRec3D are significant at the 0.05 level. While ChromSDE is also significant at the 0.05 level for DNaseI, the other two methods are not significant at this level. Further, none of them are significant at the 0.05 level for H3K4me3.

Analysis of mouse stem cell Hi-C data

As presented above, TAD is an important concept in chromatin 3D structure and thus a crucial criterion for evaluating the performance of methods for their ability to preserve such distinct features. To this end, we consider a mouse embryonic stem cell Hi-C data set [7] based on which TADs have been established. This data set is presented at 40 kb resolution; the data that are publicly available have already been biased-corrected. Our focus is on the segment of chromosome 2 from base pair 73720001 to 75440000, as this segment is known to contain two TADs [7]. As loci within the same domain interact with each other much more than across domains, the two domains should be well separated in the 3D space and can be clearly visualized. Applications of the methods yielded the estimated 3D structure depicted in Figure 5. Note that because this data set is of a

Figure 2. The plots of pairs of RMSD and correlation between true underlying structure and estimates of PRAM (or tREX), BACH (or tPAM), ShRec3D, PASTIS and ChromSDE for the in silico data of $\alpha_1 \equiv -0.5$ with various random components. No zero-inflation: (A) NR, (C) R1, (E) R2; 20% zero-inflated: (B) NR, (D) R1, (F) R2.

higher resolution than the typical 1 mb, BACH's counterpart for higher resolution data, tPAM, was used in the analysis. The first domain, denoted by dark gray balls, consists of the 19 loci within the segment from base pair 73720001 to 74480000; the second domain, denoted by light gray balls, consists of the remaining 24 loci within the segment from base pair 74480001 to 75440000. It can be seen from the figure that, other than ShRec3D (Figure 5D), the rest of the methods preserve the two domains. Given the similarities between ShRec3D and ChromSDE in their optimization procedure, it is likely that the shortest-path algorithm may inadvertently shorten the distances between loci residing in these two different domains.

Discussion

In this article, we compare several methods for reconstructing spatial 3D chromatin structures from 2D contact maps obtained from current state-of-the-art genome-wide 3C-based technologies. These methods differ in their fundamental approach (optimization-based versus model-based), in their output (a single consensus structure versus many realizations) and in their adaptability to data resolution. Our evaluation was in part based on simulated data that mimic some aspects of contact map data features, dependency, over-dispersion and zero inflation when higher resolution data are considered. We also evaluate

Figure 3. Plots of pairs of Pearson and Spearman correlations between FISH measurements and estimates obtained from various methods.

Figure 4. Boxplots of genome-wide signal intensity of genomic features. For each of the three features, the left box is for the densely packed group (smaller CAT scores), while the right one is for the sparsely packed group (larger CAT scores), predicted from the chromatin structure reconstructed by (A) BACH, (B) ChromSDE, (C) PASTIS, (D) ShRec3D and (E) tREX.

Figure 5. The estimated 3D structures from several methods: (A) tPAM, (B) ChromSDE, (C) PASTIS, (D) ShRec3D and (E) tREX.

the methods using two real data sets with features that permit comparisons: availability of FISH measurements, genomic feature data or well-established TADs.

For low resolution data, all methods considered appear to perform reasonably well, except ShRec3D, which lags behind the other methods considerably (Table 3 top panel and Figure 3). In fact, ShRec3D failed to produce a 3D structure for data generated under a setting in which the contact counts are not independent. For high resolution data, ShRec3D again stands out for its inability to analyze data simulated under a nonindependent setting (Table 3 bottom panel). It also failed to separate loci known to belong to two TADs, while the other methods are all seen to be able to preserve the TAD structures (Figure 5). Although PASTIS performed well in most of the analyses, its peculiar behavior with a negative correlation between estimated and true distances for a high resolution setting needs further evaluation. Finally, compared with methods that only produce a single consensus structure (ChromSDE, PASTIS and ShRec3D), features extracted from methods that produce the 3D structure based on multiple realizations (BACH and tREX) appear to be better correlated with several genomic functions (Figure 4). These results also point to the potential of integrating such genomic features into model-based methods to produce even more accurate 3D structures.

Not surprisingly, methods that produce a single consensus structure are much more computationally efficient compared with those that sample multiple structures. Among the latter, methods that only analyze nonzero count data are more efficient. For data produced from a population of cells, ignoring the zeros may be appropriate [15]; however, one should keep in mind that this issue should be revisited for single-cell Hi-C data

[8], especially because there is a great deal of developments recently [22–24], as one needs to distinguish structural zeros from zeros due to insufficient sequencing depth. A further disadvantage of methods that rely on multiple realizations is that their computational constraints render truly genome-wide analyses not yet realizable. Current analyses are done chromosome by chromosome, or at best, with a couple of chromosomes analyzed jointly [16].

In conclusion, ChromSDE is a computationally efficient method for constructing a single consensus structure, and its performance is not particularly sensitive to data resolution. However, care should be exercised when interpreting a single consensus structure for data generated from a population of cells. Perhaps methods for constructing single consensus structures would be better-suited for single-cell data. On the other hand, tREX is computationally much more expensive, but its performance by and large appears to be superior compared with others in almost all analyses perform.

Looking forward, 3C-based and other similar technologies, without a doubt, will continue to be innovated to produce more and more high resolution and high-quality data. The recent surge in single-cell Hi-C technology is one example that has garnered a great deal of attention. Accordingly, novel statistical and computational methods will need to be advanced to meet the demands of processing the data to extract information to the greatest extent possible. One of the immediate tasks is the development of methods that have the capability of exploring the space of genome-wide 3D structures utilizing the totality of data generated from a population of cells. An equally pressing task is the curation of existing methods for their aptness, and the further development of novel methods if needed, for

analyzing single-cell Hi-C data. Not surprisingly, preliminary exploration has shown that many single-cell contact matrices are extremely sparse. This calls for revisit of the question on how zeros should be handled, a critical issue likely to garner immediate attention. Lastly, as chromatin structures are known to be dynamic, time-course Hi-C data have begun to be produced. Analyzing such data adds yet another layer of difficulty and thus poses unique challenges that require the development of sophisticated methodology.

Key Points

- A plethora of analytical methods have been proposed to reconstruct the underlying 3D spatial chromatin structure using 2D contact maps produced by 3C-based technologies.
- These analytical tools include optimization-based as well as model-based methods. An alternative classification of the methods is whether a single consensus structure or multiple realizations are being sought after.
- We evaluate and compare these methods using both simulated and real data sets based on several criteria, including the use of known structures for simulated data, and FISH measurements and genomic features for real data.
- An optimization-based method, ChromSDE, is seen to perform well and is not particularly sensitive to data resolution owing to the use of a penalty term for contact counts that are zeros.
- A model-based method, tREX, appears to be superior to others in most of the settings studied; the results from the real data analyses are consistent with FISH measurements and several genomic features of known biological functions.

Acknowledgments

We would like to thank the reviewers for constructive comments and suggestions, which have led to improved presentations of the material.

Funding

This work was supported in part by the National Institute of Health grant R01GM114142. This research was also supported in part by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education grant NRF-2016R1D1A1A02937304.

References

- Dekker J, Rippe K, Dekker M, et al. Capturing chromosome conformation. *Science* 2002;**295**:1306–11.
- Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93.
- Kalhor R, Tjong H, Jayathilaka N, et al. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotech* 2012;**30**:90–8.
- Rao SSP, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80.
- Hughes JR, Roberts N, McGowan S, et al. Analysis of hundreds of cis -regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Publ Group* 2014;**46**:205–12.
- Davies JOJ, Telenius JM, McGowan SJ, et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* 2016;**13**:74–80.
- Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.
- Nagano T, Lubling Y, Stevens TJ, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**:59–64.
- Rousseau M, Fraser J, Ferraiuolo M, et al. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics* 2011;**12**:414.
- Zhang Z, Li G, Toh KC, et al. Inference of spatial organizations of chromosomes using semi-definite embedding approach and hi-c data. In: *Proceedings of the 17th International Conference on Research in Computational Molecular Biology, RECOMB'13*. Springer-Verlag, Berlin, Heidelberg, 2013, 317–32.
- Hu M, Deng K, Qin Z, et al. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol* 2013;**9**:e1002893.
- Varoquaux N, Ay F, Noble WS, et al. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics* 2014;**30**:26–33.
- Lesne A, Riposo J, Roger P, et al. 3d genome reconstruction from chromosomal contacts. *Nat Meth* 2014;**11**:1141–3.
- Park J, Lin S. *Statistical Inference on Three-Dimensional Structure of Genome by Truncated Poisson Architecture Model*. Cham: Springer International Publishing, 2015, 245–61.
- Park J, Lin S. Impact of data resolution on three-dimensional structure inference methods. *BMC Bioinformatics* 2016;**17**:70.
- Park J, Lin S. A random effect model for reconstruction of spatial chromatin structure. *Biometrics* 2017;**73**:52–62.
- Kruskal JB, Wish M. *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications, 1978.
- Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 2011;**43**:1059–65.
- Niu L, Li G, Lin S. Statistical models for detecting differential chromatin interactions mediated by a protein. *PLoS One* 2014;**9**:e97560.
- Hu M, Deng K, Selvaraj S, et al. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* 2012;**28**:3131–3.
- Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65.
- Stevens TJ, Lando D, Basu S, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature* 2017;**544**:59–64.
- Ramani V, Deng X, Qiu R, et al. Massively multiplex single-cell hi-c. *Nat Methods* 2017;**14**:263–6.
- de Wit E. Capturing heterogeneity: single-cell structures of the 3d genome. *Nat Struct Mol Biol* 2017;**24**:437–8.