Published online 28 November 2016

Nucleic Acids Research, 2017, Vol. 45, No. 3 ***1049–1058***
doi: 10.1093/nar/gkw1155

3D genome structure modeling by Lorentzian objective function

Tuan Trieu[1] and Jianlin Cheng[1,2,*]

[1]Computer Science Department, University of Missouri-Columbia, MO 65211, USA and [2]Informatics Institute, University of Missouri-Columbia, MO 65211, USA

Received May 12, 2016; Revised November 01, 2016; Editorial Decision November 03, 2016; Accepted November 04, 2016

ABSTRACT

The 3D structure of the genome plays a vital role in biological processes such as gene interaction, gene regulation, DNA replication and genome methylation. Advanced chromosomal conformation capture techniques, such as Hi-C and tethered conformation capture, can generate chromosomal contact data that can be used to computationally reconstruct 3D structures of the genome. We developed a novel restraint-based method that is capable of reconstructing 3D genome structures utilizing both intra-and inter-chromosomal contact data. Our method was robust to noise and performed well in comparison with a panel of existing methods on a controlled simulated data set. On a real Hi-C data set of the human genome, our method produced chromosome and genome structures that are consistent with 3D FISH data and known knowledge about the human chromosome and genome, such as, chromosome territories and the cluster of small chromosomes in the nucleus center with the exception of the chromosome 18. The tool and experimental data are available at https://missouri.box.com/v/LorDG.

INTRODUCTION

Three-dimensional (3D) chromosome and genome structures have been shown to play important roles in many biological processes (1,2). However, due to the large size of a genome, there is no experimental technology that can directly determine the 3D structure of a genome. Fluorescence *in situ* hybridization (FISH) has been used to measure the distance between genomic regions, but it is limited by its low throughput and low resolution (3–7). Chromosomal conformation capture based techniques empowered by next generation sequencing, such as Hi-C (1), TCC (8), can capture chromosomal fragments of a genome that are in spatial proximity that can be mapped to the genome to generate genome-wide chromosomal interaction/contact data. The contact data, despite not 3D conformation itself, can be used to computationally reconstruct 3D structures of chromosomes and genomes.

Two kinds of methods have been developed to build 3D structures of chromosomes and genomes using chromosomal contact data. The first kind uses the polymer physics of the chromatin to build models that are consistent with observed chromosomal contact data (9,10). The second one treats each chromosomal contact as a restraint, and then solves a spatial optimization problem to find chromosome or genome conformations that satisfy the contact restraints as well as possible (11–19). A common approach used by these methods is to convert chromosomal contacts into spatial distance restraints, and then search for the optimal conformations (models) that satisfy the restraints best according to an objective function. Depending on how these restraints are satisfied, the optimization process can produce one consensus model or an ensemble of models.

One intrinsic characteristic of Hi-C data hindering the conformation search is that chromosomal contacts are often inconsistent because the data are captured from millions of cells of the same cell line and the genome structures of the cells may vary despite the similarity between them. Because of this inconsistency, e.g. some contacts may exist in one genome structure but not in another one, chromosomal contacts and their derived distance restraints cannot be satisfied all together in one structural model. Therefore, during the reconstruction of the genome or chromosome conformation, it may be better not to penalize the violation of inconsistent restraints that are not supposed to be satisfied. In this spirit, we derived an objective function using the Lorentzian function that can reward the satisfaction of consistent restraints whose value is not affected by the violation of inconsistent restraints. Specifically, a form of the bell-shape Lorentzian function was used to quantify the satisfaction of each contact restraint. This function is differentiable and continuous so that the optimization can be solved efficiently using gradient-based optimization techniques. Our recently published method, MOGEN, (17) also implements the idea of not severely penalizing the violation of inconsistent restraints. The method doesn't require a function to translate interaction frequencies of contacts to spatial distances like most restraint-based methods. It is also ro-

*To whom correspondence should be addressed. Tel: +1 573 882 7306; Fax: +1 573 882 8318; Email: chengji@missouri.edu

bust to noise and capable of reconstructing genome models utilizing both inter-and intra-chromosomal contacts. However, the method has several parameters that need to be tuned. Our method introduced here (LorDG - Lorentzian 3D Genome) requires a function to translate interaction frequencies into spatial distances, but has only one parameter that can be found automatically.

We benchmarked LorDG together with existing methods (12–14). These methods use a similar approach in deriving distance restraints from contacts, but differ in how restraints are satisfied. The result shows that our method is significantly more robust to noise or inconsistency than the method that utilizes the squared loss function, and performs better than probabilistic methods on noisy data sets. We also tested our method on real Hi-C data sets of the cell lines GM06990 (1) and GM12878 (20) and validated models with FISH data. The reconstructed models of chromosomes and genomes possess the known features of the human chromosome and genome, such as chromosome territories, the clustering of small chromosomes except chromosome 18, in the nucleus center.

MATERIALS AND METHODS

We model chromosome/genome structures as a series of beads with each bead representing a chromosomal fragment (subsequence) of a specific length (e.g. 1 Mb or 500 Kb). The position of each bead is represented by three coordinates (x, y, z) in 3D space. The number of contacts between beads (i.e. interaction frequency (IF) can be computed from raw reads of Hi-C data (1)). And the Hi-C data of a genome or a chromosome can be summarized as an n-by-n matrix M (contact matrix), where n is the number of beads, and each element in the matrix ($M[i,j]$) contains the IF between bead i and bead j. n determines the resolution of a contact matrix, i.e. a larger n (i.e. a smaller bead length) leads to a higher resolution.

Conversion of interaction frequency to spatial distance

An important component of restraint-based genome/chromosome structure modeling methods is a function to convert chromosomal contacts into spatial distances between beads. These converted distances are called 'wish' distances because they are just theoretical approximations of exact spatial distances in 3D and may be in conflict with each other. The conversion function plays a crucial role in determining the quality of reconstructed models. However, currently, there is no exact function that can accurately capture the relationship between chromosomal contacts derived from experimental Hi-C data and true spatial distances. Moreover, this distance function can vary for different cell types and contact matrices at different resolutions. Despite these challenges, a commonly used approximation function that converts interaction frequency to distance is $d_{ij} = \frac{1}{IF_{ij}^{\alpha}}$, where IF_{ij} and d_{ij} are the interaction frequency and the spatial distance between the two beads, respectively (1,11–14). Some methods try to optimize α (11,13,14) instead of using a fixed value. Our method utilizes this conversion function and searches for the best α within [0.1, 3.0] with a step-size of 0.1. This range covers most previously used values for α (11–13).

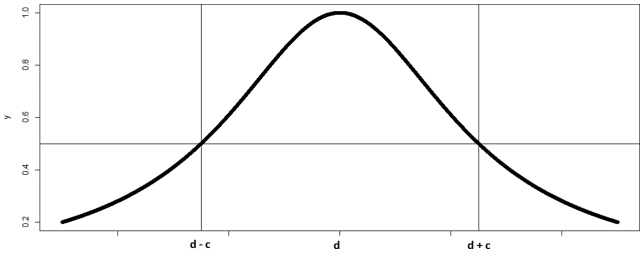

Figure 1. Lorentzian function. When x = d (the contact distance is satisfied), the function is maximized; when x is very far from d (the contact is violated seriously), it becomes flat and its value is very small.

The Lorentzian objective function for spatial optimization

We used a simple form of the Lorentzian function, $\frac{c*c}{c*c+(x-d)^2}$, to quantify the satisfaction of distance restraints, where c and d are constants determining the wideness and the peak point of the bell curve as illustrated in Figure 1. The objective function, fn, for spatial optimization is formulated as below:

$$fn = \sum_{|i-j|=1} \frac{c*c*IF_{\max}}{c*c+\left(x_{ij}-d_{ij}\right)^2} + \sum_{|i-j|\neq 1} \frac{c*c*IF_{ij}}{c*c+\left(x_{ij}-d_{ij}\right)^2},$$

where $IF_{\max}$ is the maximum of IFs, d_{ij} is the wish distance between beads i and j and x_{ij} is the spatial distance between beads i and j in a 3D model. The optimization goal is to find a chromosome or genome conformation that maximizes this objective function, which aims to reduce the difference between x_{ij} and d_{ij}. Although the objective of this function is similar to traditional squared-error function ($\sum IF_{ij}^2(x_{ij} - d_{ij})^2$), they behave quite differently during the optimization process.

A special characteristic of the Lorentzian function is that its derivative is very small almost everywhere except along the two slopes of the curve centered at the wish distance d. In order to maximize the objective function using its gradient, c is set to the average distance of wish distances so that most of restraints will have a non-zero derivative during the optimization process. Although inconsistent restraints that generally have low IFs cannot be satisfied, they barely affect the satisfaction of other restraints because their violation doesn't influence the value of the objective function. Therefore, this objective function can help maximize the number of satisfied restraints that are consistent. In the case of using the squared loss function, the optimization process will try to prevent inconsistent restraints from being seriously violated and thus may not be able to maximize the number of consistently satisfied restraints. Another disadvantage of the squared loss function is that it can be dominated by a few noisy contacts (outliers).

Maximization of the objective function

We used the gradient ascent optimization with adaptive step sizes to adjust the position (x, y, z coordinates) of each bead

in order to maximize the Lorentzian objective function. The search for a new step size was performed only when the objective (scoring) function stops increasing. Starting from a randomly initialized structure, the objective function was maximized iteratively and the final model was outputted when the convergence condition was met. This gradient ascent optimization may produce different models in different runs. Intuitively, we expect that, when there is a major group of consistent restraints, the optimization process will always manage to satisfy this group of restraints. However, when there is no major group of consistent restraints, every group of consistent restraints can be satisfied in different runs. Indeed, as demonstrated in Section 3, when the input was intra-chromosomal contacts of low resolution, the optimization produced similar models because a majority group of consistent restraints existed. But when the input included mostly inconsistent inter-chromosomal contacts, the optimization process produced different models in different runs and these models might represent the structures of sub groups of cells in the whole population of cells used to produce the Hi-C data.

Preparation of synthetic data sets

We simulated 13 artificial chromosomal contact data sets from the theoretical 3D model of the yeast chromosome 4 (19) to test our method and compare it with the four methods implemented in (13). The chromosome was represented by 122 fragments (beads). This model serves as a true model or a gold standard model to test if methods can reconstruct models that are similar to this model. We tried to use the whole model with 3047 beads, but other methods used to compare with our method could not handle the input data at this resolution or took too long to run, even though it only took minutes for our method to run on the same computer. So we reduced the resolution to 122 beads to perform the experiment and comparison.

We simulated the interaction frequencies using Poisson distribution (13) as follow:

$$IF_{ij} = P\left(\beta d_{ij}^{\gamma}\right)$$

Where $\gamma = -1$ and d_{ij} is the Euclidian distance between the fragment i and j in the true model. In (13), the authors selected $\gamma = -3$ by assuming a fractal and/or equilibrium globule polymer model for mammalian DNA (1). However, such a model has been ruled out by the authors later (21). Nevertheless, the choice of γ to produce synthetic data set for testing methods should not matter as long as methods tested on the data are provided with true γ or are able to find the true value of γ (or $\alpha = \frac{1}{\gamma}$). In our comparison of these methods, the methods that cannot search for values for γ were provided with the true value.

To reduce numerical error, d_{ij} is divided by the average of all distances. The scaling factor β doesn't affect the noise level itself because all IFs are scaled by the same factor. However, the Poisson distribution draws only discrete value, therefore, if βd_{ij}^{γ} is a small non-integer number, the ratio $\frac{P(\beta d_{ij}^{\gamma})}{\beta d_{ij}^{\gamma}}$ could be far from 1.0, which means that the obtained $I\ F_{ij} = P(\beta d_{ij}^{\gamma})$ is very noisy. When βd_{ij}^{γ} is large, the ratio $\frac{P(\beta d_{ij}^{\gamma})}{\beta d_{ij}^{\gamma}}$ is often close to 1.0 and the obtained $I\ F_{ij} = P(\beta d_{ij}^{\gamma})$ is less noisy. The role of β is to scale interaction frequencies to control the noise level of simulated interaction frequencies. In our experiment, β in the range [1-20] gives reasonable levels of noise to evaluate methods. Each value of β, we generated one data set and measured the signal to noise ratio (SNR), the higher the SNR, the better quality the data set is. The signal to noise ratio of data sets are reported in the Figure 2. This value of β and the signal to noise ratios are different from (13) because the scale of the model used here and in the model in (13) is different. When d_{ij}^{γ} is large, a small value of β is required to produce sufficient noise.

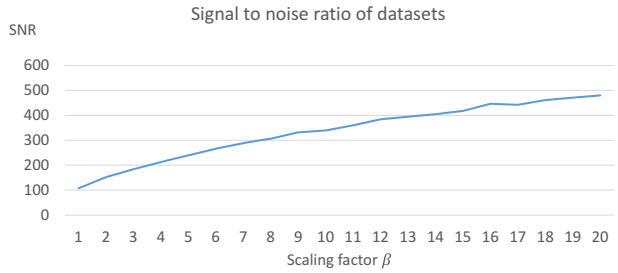

Figure 2. Signal to noise ratio of data sets generated with different value of β.

$$SNR = \frac{\sum IF_{ij}}{\sqrt{\sum \left(\beta d_{ij}^{\gamma} - IF_{ij}\right)^2}}$$

Normalization of the real Hi-C data

We used two real Hi-C data sets in (1,20) to reconstruct chromosome and genome structures with our method. These data sets were generated from millions of human B-cells (GM06990, GM12878). Because the real-world Hi-C data often suffers from several sources of biases and experimental noises (20,21), the data needs to be normalized before 3D model reconstruction. The data set in (20) was already normalized by KR normalization (20). For the data set in (1), we normalized the intra-chromosomal contact data (i.e. contacts between fragments in the same chromosome) using the iterative correction normalization method (21). This iterative correction normalization does not assume any specific source of biases as other normalization methods (20,22) do. It is also highly effective in that it can highlight structural patterns from the background and thus help satisfy restraints derived from contacts better. When dealing with the data set of the whole genome consisting of both intra-chromosomal and inter-chromosomal contacts (i.e. contacts between two different chromosomes), we tried the iterative correction normalization and coverage normalization (1). The coverage normalization led to models with higher scores and more satisfied contacts. Therefore, we applied the coverage normalization protocol (1), which is just one iteration of the iterative correction normalization method.

Measurement of structural similarity

To measure the similarity between reconstructed models and the true synthetic model, we used both Spearman's correlation and root mean square error (RMSE). RMSE has been widely used to measure the similarity between 3D structures. Its formula is given as:

$$RMSE = \sqrt{\frac{1}{n}\sum\left(d_{ij} - d'_{ij}\right)^2},$$

where d_{ij} and d'_{ij} are the distance between beads i and j in in two 3D models and n is the number of pairwise distances. In order to use RMSE, the two structures must be compared at the same scale. We approximately achieved this by scaling all pairwise distances of one model such that the two models have the same average pairwise distance. Another limitation of RMSE is that it can be dominated by a few large errors. Thus, we also used Spearman's correlation between pairwise distances of reconstructed models and the true model. The closer to 1 the correlation, the better is the reconstructed model.

To determine the value of the α parameter of the formula of converting IF to distance, we computed the Spearman's correlation between reconstructed distances and wish distances. We selected the value of α that produced the highest correlation.

RESULTS AND DISCUSSION

Determine the parameter of the function converting IF to wish distance

In 3D model reconstruction by our method, interaction frequencies are first converted to Euclidian distances using the formula $d_{ij} = \frac{1}{IF_{ij}^{\alpha}}$. Because interaction frequencies often contain noise and α is unknown, converted distances (d_{ij}) are not equal to exact true distances and not all of them can be realized in 3D models.

We performed a line search within a range [0.1, 4.0] with a step size of 0.1 to find the best α in term of Spearman's correlation between reconstructed distances and wish distances. We tested this approach on a synthetic dataset to see if our method was capable of determining a reasonable value of α to reconstruct models that were close to the true model. The data set generated with $\beta = 10$ was used as the input. Figure 3 shows how the Spearman's correlation between reconstructed distances and wish distances changed as α varied within [0.1, 4.0]. The correlation was peaked at $\alpha = 1.1$. We measured the similarity of reconstructed models with the true model to see which α really gives the best model. Figure 4 shows these correlations. The model generated with $\alpha = 1.3$ is actually the most similar model to the true model (correlation is 0.995). When $\alpha = 1.1$, our method generated the second most similar model to the true model with correlation of 0.994, which is very close to the most similar model. This suggests that our method can find a reasonable value for α. In general, our method first searches for the best α according to the correlation between reconstructed distances and input interaction frequencies, and then reconstructs models with this α.

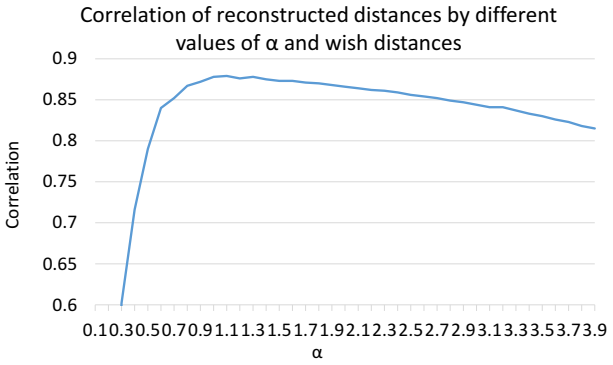

Figure 3. Correlation between reconstructed distances in models with different value of α and wish distances. The correlation was peaked at 1.1.

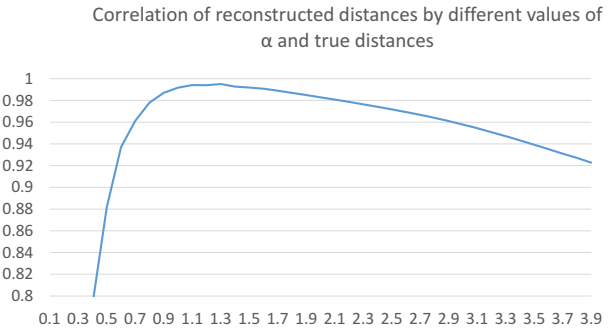

Figure 4. Correlation of models generated with different values of α and the true model. The model generated by $\alpha = 1.1$ is the second most similar model to the true model after the model reconstructed with $\alpha = 1.3$.

Structure reconstruction using the synthetic data and comparison with Pastis, ChromeSDE, Shrec3D and MOGEN

We compared our methods with four methods implemented in (13) and with ChromSDE (14), Shrec3D (12) and MOGEN (17) using the synthetic data sets. The first four methods include two classic multidimensional scaling methods (metric multidimensional scaling – MDS2 and non-metric multidimensional scaling – NMDS) and two statistical methods using a Poisson distribution (PM1 and PM2). MDS2 directly infers the coordinates of beads from their pairwise Euclidean distances. NMDS relies on the assumption that, if $IF_{ij} > IF_{kl}$, then d_{ij} should be shorter than d_{kl}, in order to derive an objective function to optimize. PM1 and PM2 model IFs as Poisson random variables, and then try to maximize the likelihood of observing these IFs. While PM1 needs a formula as the prior knowledge to convert the spatial distance to the Poisson intensity, PM2 (also called Pastis) can automatically adjust the formula to infer models that best explain the observed IFs. All four methods generate a consensus model given an input data. The MDS2 method uses the squared loss function (sum of weighted square of errors as the loss function) during optimization. So it serves as a good comparison with our proposed objective function. MDS2, NMDS, PM1 were provided with the true value of α to convert IFs into wish distances before spatial optimization. However, models generated by PM1 are very different from the true models in all cases, we suspected that PM1 has implementation problems and therefore did

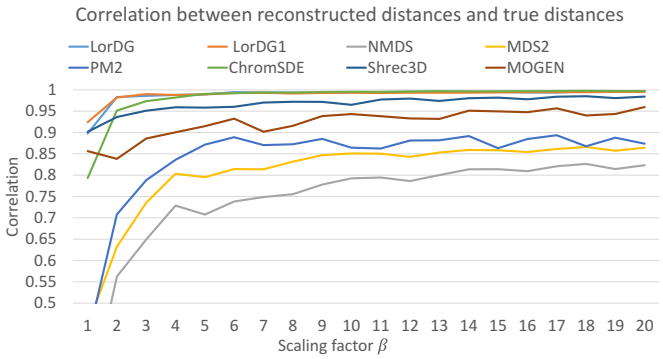

Figure 5. Spearman's correlation of reconstructed model and the true model on the synthetic data with respect to different levels of noise.

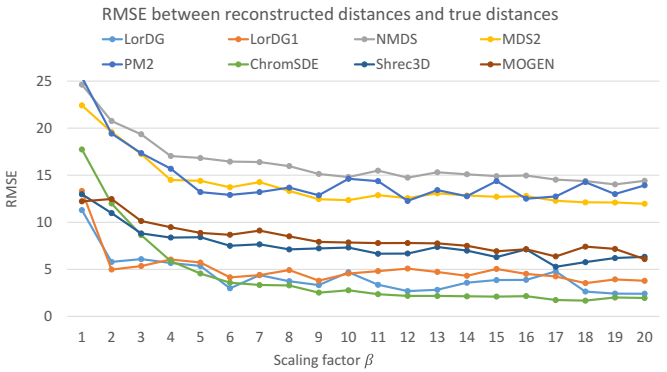

Figure 6. Root mean square error (RMSE) between reconstructed models and the true model of the five methods on the synthetic data with different levels of noise.

not include PM1 in the comparison. MDS2, NMDS and PM2 were run with 500 iterations (default is 100 iterations).

ChromSDE uses semi-definite programming techniques to fit the observed data and jointly optimizes the 3D model and α together. Shrec3D corrects wish distances by a shortest path algorithm and then uses distance geometry to solve for coordinates of 3D models. MOGEN is different from the other methods in that it does not assume a mathematical formula for the relationship between interaction frequencies and spatial distances. It relies on the concept of contact and non-contact to devise soft restraints and build models that satisfies as many soft restraints as possible. The setting of this comparison, where the exact formula for the relationship between interaction frequencies and spatial distances is known, could underestimate the performance of MOGEN and favors other methods. In real Hi-C data, the relationship can be approximated only and therefore, another layer of noise would be produced from the approximation.

To see the effect of using Lorentzian function and to compare it with MDS2 and Shrec3D, we implemented a variant of LorDG, which is called LorDG1, where the objective function is similar to the objective function of MDS2 but the squared-loss function is replaced by the Lorentzian function. The objective function $fn1$ is given as:

$$fn1 = \sum \frac{c * c * IF_{ij}^2}{c * c + \left(x_{ij} - d_{ij}\right)^2}$$

We compared the performance of eight methods LorDG, LorDG1, MDS2, NMDS, PM2, ChromSDE, Shrec3D and MOGEN. LorDG1 also searched for the best α and then built 3D models for each input data. We use the correlation and RMSE of generated models with the true model to compare methods.

The results are shown in Figures 5 and 6. In term of correlation, ChromSDE, LorDG and LorDG1 performed better than the other methods, and LorDG1 seemed to be the most robust method when the input data were very noisy. However, when considering RMSE, ChromSDE was the best method, LorDG was slightly better than LorDG1, and both of them were better than other methods. Overall, ChromSDE, LorDG and LorDG1 performed well and better than the other methods. ChromSDE was the best when the input data were not too noisy, while LorDG and LorDG1 were slightly better than ChromSDE if the input data were very noisy. However, one major limitation of ChromSDE was that it could not scale up well to large structures because of its implementation in Matlab and usage of matrices. ChromSDE could not handle an input data with more than about three hundreds of points in our machine (with 8 GB RAM).

LorDG1 was better than LorDG when the input was very noisy, which was expected because noisy contacts often have low IFs, therefore, assigning IF^2 as weights prioritizes satisfaction of reliable contacts (high IFs). However, assigning weights also forces reconstructed distances to be close to wish distances for contacts with high IFs and makes satisfying contacts with lower IFs more difficult. Because of that, LorDG1 was better than LorDG when the input data are very noisy and LorDG was better when the input was less noisy.

LorDG and LorDG1 outperformed MDS2 using the weighted squared loss function in term of both correlation and RMSE. They also outperformed PM2 and NMDS. Overall, the results show that using the Lorentzian loss function produced better results than using the squared error function or the probabilistic method PM2.

Local maximum in LorDG

The objective function of LorDG is non-convex so that the optimization can converge to a local maximum. We optimized the objective function by gradient ascent and always initialized models randomly. Thus, different runs are very likely to converge to different local maximums. However, our intuition is that if there is a majority of consistent restraints, models at different local maximums should satisfy these restraints and thus be similar to each other. If there is no major group of consistent restraints, models at different local maximums may not be similar to each other. We tested this hypothesis by generating two ensembles of 50 models at two different noise levels $\beta = 1$ and $\beta = 10$ and measured the similarity in term of correlation of models from the same ensemble. At $\beta = 1$ (the noisiest level), the average correlation between models was 0.937 (minimum, maximum and standard deviation were 0.88, 0.99 and 0.0177, respectively). While at $\beta = 10$ (less noisy), the average cor-

relation was 0.996 (minimum, maximum and standard deviation were 0.93, 0.99 and 0.0020, respectively). The results show that, when the input data were less noisy, models generated in different runs were more similar even though they were randomly initialized and might converge to different local maximums.

Reconstruction of the structures of human chromosomes/genome from real Hi-C data

We applied LorDG to the Hi-C data of the human genome to reconstruct the 3D structures of chromosomes at 1 Mb and 500 Kb resolution and the structure of the entire genome at 1 Mb resolution. Because the interaction of two homologous chromosomes cannot be separated, the structure for each pair of homologous chromosomes was generated from the contacts associated with the pair rather than each individual chromosome, that is each pair of chromosomes with the exception of sex chromosomes X and Y was treated as one single entity during the modeling as most other methods in the literature do.

Reconstruction of chromosome structures

For each contact matrix input of one pair of chromosomes, we generated an ensemble of 50 structures by running LorDG 50 times on the same input data. The models were initialized randomly and might converge to different local maximums in these runs. We measured the quality of structures in term of Spearman's correlation between reconstructed distances and wish distances.

We also calculated the similarity between structures in the same ensemble using Spearman's correlation. The Spearman's correlation between all pairs of structures in the same ensemble were computed and averaged. This averaged correlation is considered as the similarity score of structures in the ensemble. For all chromosomes, the correlation between structures was greater than 0.8, suggesting that structures generated in different runs were similar. For instance, the similarity score of the structures of chromosome 1 at 1 Mb was 0.95, with standard deviation, min and max as 0.02, 0.91 and 0.99, respectively, and the similarity score at 500 Kb resolution was 0.91, with standard deviation, min and max were 0.03, 0.81 and 0.99. Rows (1) and (2) in Table 1 of the Supplementary Material report the similarity scores of structures generated in different runs of each chromosome at 1 Mb and 500 Kb resolution, respectively.

The quality of structures in the ensemble was assessed using the Spearman's correlation between reconstructed distances and wish distances. Most of these correlations were larger than 0.7, indicating that the reconstructed structures are of good quality. Rows (1) and (2) in Table 1 report the quality scores of structures of each chromosome at 1 Mb and 500 Kb resolution, respectively.

Evaluation of the modeling stability against the change of resolution

We compared the structures of the same chromosome reconstructed at 1 Mb and 500 Kb resolution. The structures at 500 Kb resolution were reduced to the size of a 1 Mb structure by averaging every two consecutive points. The Spearman's correlation between all pairs of 1 Mb structures and reduced 500 Kb structures was computed and averaged. For chromosome 1, the Spearman's correlation between 1 Mb structures and 500 Kb structures is 0.9. Figure 7 shows the superimposition of a 1 Mb structure and a 500 Kb structure of Chromosome 1, demonstrating that they are similar. The scores of other chromosomes are shown in Row (3) of Table 1. Except chromosomes 14 and 19, the correlations for other chromosomes are greater than 0.7, suggesting that the models at 1 Mb and corresponding models at 500 Kb are consistent.

Figure 7. The superposition of the structures of chromosome 1 at 1 Mb and 500 Kb resolution.

Comparison with ChromSDE and Verification with FISH data

We used ChromSDE to build a model of chromosome 14 of the cell line GM06990 at 1 Mb and compared it with the 50 models reconstructed by LorDG. The average correlation was 0.91 (minimum, maximum and standard deviation are 0.88, 0.94 and 0.02), which indicates that models reconstructed by ChromSDE and LorDG are highly similar.

We used the chromosome 14 models of LorDG to compare with FISH data obtained in (1). Four 3D-FISH probe positions on chromosome 14 were analyzed, which showed that L3 tended to be closer to L1 than to L2, though L2 lied between L1 and L3, and L2 tended to be closer to L4 than to L3, despite that L3 was between L2 and L4 (1). We measured the distances L1-L3, L2-L3 and L2-L4 in our models. And the distance L2-L3 was consistently larger than L1-L3 and L2-L4 in all models. Figure 8 shows a model reconstructed by LorDG with four probes, L2 is closer to L4 than to L3 and L3 is closer to L1 than to L2. The model reconstructed by ChromSDE is also consistent with the FISH Data. The distances of L1-L3, L2-L3 and L2-L4 in the model of ChromeSDE are 56.2, 81.7 and 34.9, respectively.

Table 1. Evaluation of chromosome structures

Chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
(1)	0.78	0.77	0.78	0.86	0.84	0.81	0.78	0.84	0.84	0.80	0.79	0.80	0.8	0.76	0.81	0.75	0.7	0.75	0.75	0.77	0.88	0.74	0.86
(2)	0.87	0.85	0.77	0.84	0.85	0.82	0.83	0.82	0.81	0.8	0.82	0.82	0.75	0.74	0.74	0.81	0.82	0.7	0.84	0.79	0.86	0.77	0.83
(3)	0.9	0.92	0.79	0.93	0.93	0.94	0.94	0.93	0.84	0.92	0.93	0.95	0.82	0.67	0.77	0.89	0.85	0.84	0.68	0.89	0.72	0.89	0.94

Row (1) Quality of models at 1 Mb resolution. Row (2) Quality of models at 500 Kb. Row (3) Similarity of models at 1 Mb and 500 Kb resolution.

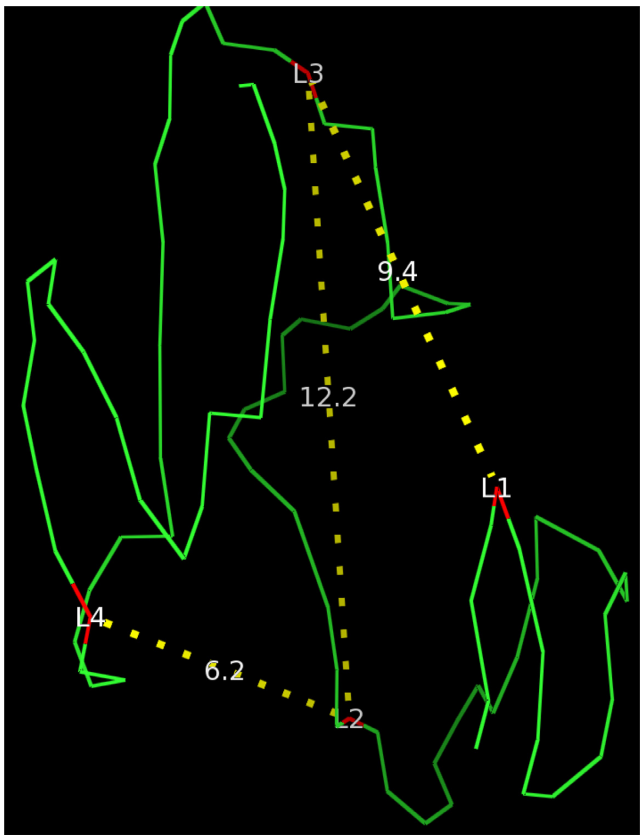

Figure 8. Distances between four fluorescence *in situ* hybridization (FISH) probes in the model of chromosome 14 reconstructed by LorDG. L1, L2, L3 and L4 denote four probes. The distances between probes are labeled along the virtual line segments connected them.

Loop realization at peak loci

We used the data set in (20) to test if LorDG was able to reconstruct loops at peak loci predicted in the data set. There are four loops on four different chromosomes (chr.11, chr.13, chr.14, chr.17) of the cell line GM12878 that were confirmed by 3D-FISH (20), and thus, we chose to build 3D model for fragments that contains these four loops to test LorDG. Three FISH probe positons L1, L2 and L3 were analyzed to confirm these loops. Peak loci L1, L2 lie close to each other and make up a loop while L2 and L3 are far away from each other although they have the same genomic distance as L1-L2. 3D models of fragments of 10 Mb long at 10 kb resolution that contain the loops were built by LorDG. We observed these four loops in all 3D models of the fragments reconstructed by LordDG. Locations of the fragments and FISH probes and their distances in 3D models are included in Table 2. The distance L1-L2 is always significantly shorter than the distance L2-L3. Figure 9 shows loops in 3D models. L1-L2 form loops while L3 is further away from L2. This result validates 3D models reconstructed by LorDG and demonstrates the potential of using LorDG to predict loops or verify loops predicted by other computational methods.

Reconstruction of the 3D structures of the entire human genome

From the normalized IF matrix of the whole genome at 1 Mb resolution, we generated an ensemble of 50 genome structures. The similarity between structures in the ensemble was computed. Unlike chromosome structures, genome structures in the same ensemble were different from each other. The average Spearman's correlation between all pairs of structures was 0.32, with standard deviation, min and max of 0.04, 0.23 and 0.50, respectively. However, the chromosome structures extracted from different genome structures were still similar to each other. For instance, the average Spearman's correlation between the structures of Chromosome 1 extracted from the 50 genome structures was 0.86, with standard deviation, min and max of 0.05, 0.62 and 0.98, respectively. This result suggests that the dissimilarity between genome structures comes mostly from the variability of the orientations of chromosomes rather than the structures of individual chromosomes. This is consistent with the fact that inter-chromosomal contacts often have very small IF in comparison with intra-chromosomal contacts, e.g. mean of IFs of inter-chromosomal contacts is 0.77 while mean of IFs of intra-chromosomal contacts is 4.99 and 77.7% of inter-chromosomal contacts with IF less than 1.0 while 84.2% of intra-chromosomal contacts with IF larger than 1.0. Because of this, it is likely that inter-chromosomal contacts are less conserved than intra-chromosomal contacts in different cells.

Despite the dissimilarity between genome structures, they still share some similar structural patterns that are consistent with some known knowledge about the human genome. Figure 10 shows a reconstructed genome structure that has the chromosome territory feature (23,24). That is, although chromosomes touch with each other at their borders, they generally maintain their own space and do not mix up with each other.

We calculated the center of the mass of each chromosome and of the whole genome and the distances between these centers. These distances were averaged over all genome structures. The heat map of averaged distances is shown in Figure 11. The center of the mass of the genome is labeled as 0 and chromosome X and Y are labeled as 23 and 24, while the other chromosomes are labeled by their index from 1 to 22. The intensity of red color corresponds to spatial proximity. Small chromosomes (17, 19, 20, 21 and 22) except chromosome 18 are closer to each other and cluster together near the center of the mass of the genome. Chro-

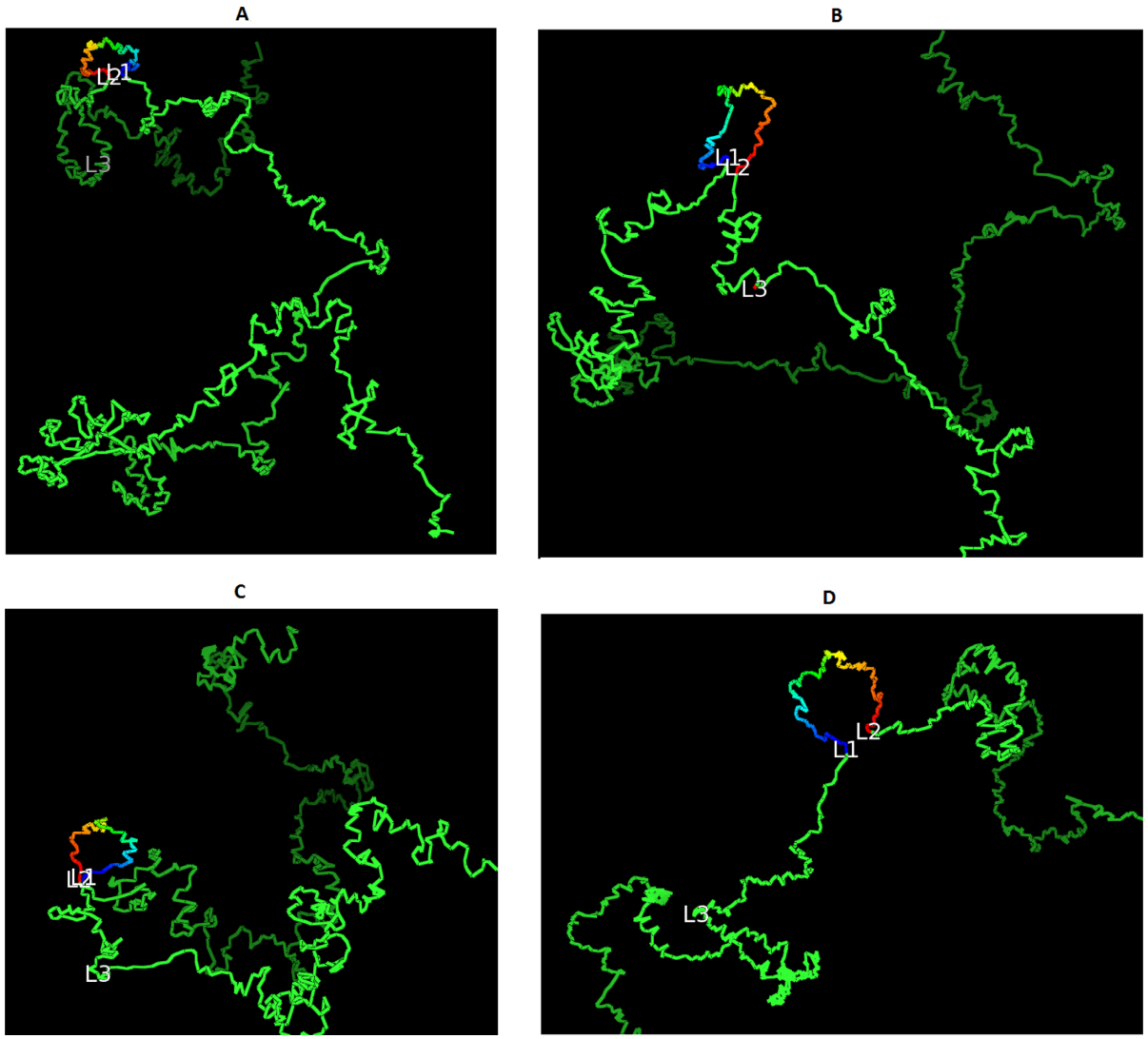

Figure 9. Loop and peak loci (L1, L2) on fragments of (**A**) Chr. 17, (**B**) Chr. 14 (**C**) Chr. 11 and (**D**) Chr.13.

Table 2. Fragments that contain loops and distances of peaks in 3D models

Fragment	Chr.	Start (Mb)	End (Mb)	L1 (position, Mb)	L2 (position, Mb)	L3 (position, Mb)	L1-L2	L2-L3
1	17	60.00	70.00	66.76–66.79	67.22–67.25	67.68–67.71	0.3	4.7
2	14	65.00	75.00	71.60–71.63	72.20–72.23	72.80–72.83	0.3	2.6
3	11	125.00	135.00	130.72–130.75	130.29–130.32	129.86–129.89	0.3	3.3
4	13	80.00	90.00	86.37–86.40	85.46–85.49	84.55–84.58	0.8	6.6

Columns 3 and 4 list the start and end position of each fragment, Columns 5, 6, and 7 the start/end position of probes L1, L2 and L3, and Columns 8 and 9 the distance of L1-L2 and L2-L3.

mosome 18 lies near the periphery. This is consistent with previous studies (25–27).

In contrast, the large chromosomes lie near the periphery. We also observed a striking feature that large chromosomes have telomeres and/or elongated regions intruding into the genome center to interact with small chromosomes as shown in Figure 12. To quantify the satisfaction of inter-chromosomal contacts, we computed the inter-chromosomal contact matrix between chromosomes and then calculated the correlation between this matrix and the distance heat map between chromosomes (lower distances are expected for higher IFs). The correlation is −0.386 indicating that a portion of inter-chromosomal contacts had been satisfied.

Running time

LorDG is relatively fast. We ran all experiments in a PC with Intel Core i5-2400 3.1 Ghz and 8.00 GB RAM. It took about a minute to reconstruct a chromosome structure at 1 Mb or 500 Kb resolution. For genome structures at 1 Mb resolution with 2911 beads, it took LorDG about 2 minutes to generate a model.

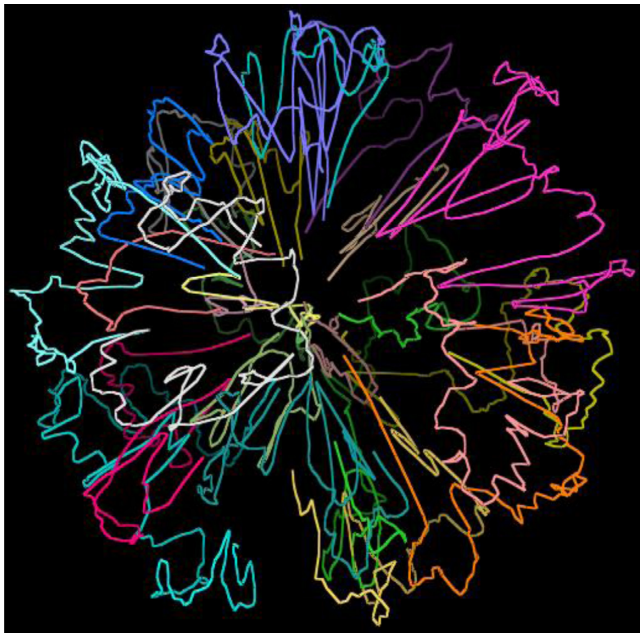

Figure 10. A genome structure with chromosome in different colors demonstrating the existence of chromosome territories.

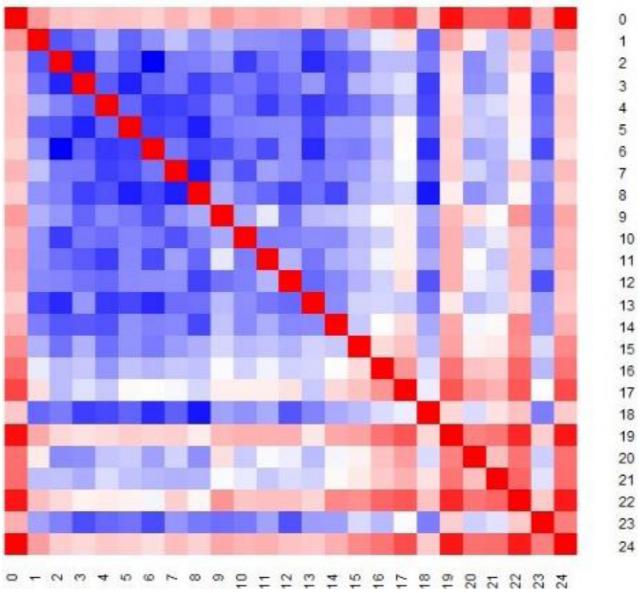

Figure 11. Distances between centers of the mass of chromosomes and of the genome. The intensity of red is proportional to proximity. Small chromosomes (17, 19, 20, 21 and 22), except chromosome 18, cluster near the center of the genome, as shown by their close proximity to the center of the genome.

CONCLUSION

We present a novel method to reconstruct 3D chromosome and genome structures from Hi-C data. The core of the method is an objective function that can tolerate inconsistent/noisy restraints and maximize the number of realistically satisfied restraints. The method was tested on both synthetic and real Hi-C data sets. And it performed very well among a panel of eight methods on the synthetic data. The models reconstructed by our method from real Hi-C data sets were validated by 3D-FISH data. And they exhibit known features of the human chromosome and genome such as, chromosome territories and the cluster of small chromosomes, except chromosome 18, in the nucleus center. This result demonstrates that our method is useful for modeling the architecture of the human chromosome and genome from Hi-C data.

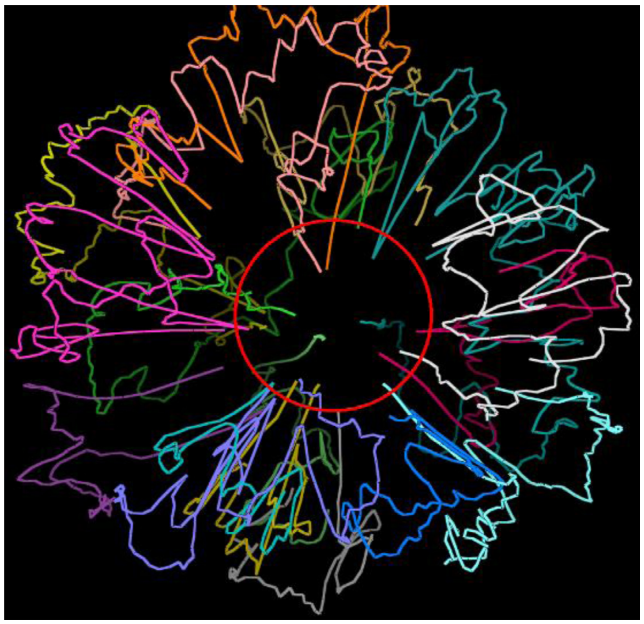

Figure 12. Telomeres and/or elongated regions of large chromosomes intrude into the nucleus center (the red circle), where small chromosomes are located, but not shown for the purpose of clarity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation [grant no. DBI1149224] to JC. Funding for open access charge: the National Science Foundation, USA.

FUNDING

National Science Foundation [DBI1149224 to J.C.]. Funding for open access charge: National Science Foundation [DBI1149224 to J.C.].
Conflict of interest statement. None declared.

REFERENCES

1. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
2. van Steensel,B. and Dekker,J. (2010) Genomics tools for unraveling chromosome architecture. *Nat. Biotech.*, **28**, 1089–1095.

3. Cremer,T. and Cremer,C. (2006) Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories. *Eur. J. Histochem.*, **50**, 161–176.
4. Edelmann,P., Bornfleth,H., Zink,D., Cremer,T. and Cremer,C. (2001) Morphology and dynamics of chromosome territories in living cells. *Biochim. Biophys. Acta*, **1551**, M29–M39.
5. Seitan,V.C., Faure,A.J., Zhan,Y., McCord,R.P., Lajoie,B.R., Ing-Simmons,E., Lenhard,B., Giorgetti,L., Heard,E., Fisher,A.G. *et al.* (2013) Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res.*, **23**, 2066–2077.
6. Gilbert,N., Boyle,S., Fiegler,H., Woodfine,K., Carter,N.P. and Bickmore,W.A. (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell*, **118**, 555–566.
7. Williamson,I., Berlivet,S., Eskeland,R., Boyle,S., Illingworth,R.S., Paquette,D., Dostie,J. and Bickmore,W.A. (2014) Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.*, **28**, 2778–2791.
8. Kalhor,R., Tjong,H., Jayathilaka,N., Alber,F. and Chen,L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.
9. Mirny,L.A. (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.*, **19**, 37–51.
10. Barbieri,M., Chotalia,M., Fraser,J., Lavitas,L.-M., Dostie,J., Pombo,A. and Nicodemi,M. (2013) A model of the large-scale organization of chromatin. *Biochem. Soc. Trans.*, **41**, 508–512.
11. Rousseau,M., Fraser,J., Ferraiuolo,M.A., Dostie,J. and Blanchette,M. (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, **12**, 414.
12. Lesne,A., Riposo,J., Roger,P., Cournac,A. and Mozziconacci,J. (2014) 3D genome reconstruction from chromosomal contacts. *Nat. Methods*, **11**, 1141–1143.
13. Varoquaux,N., Ay,F., Noble,W.S. and Vert,J.-P. (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, **30**, i26–i33.
14. Zhang,Z., Li,G., Toh,K.-C. and Sung,W.-K. (2013) 3D chromosome modeling with semi-definite programming and Hi-C data. *J. Comput. Biol.*, **20**, 831–846.
15. Hu,M., Deng,K., Qin,Z., Dixon,J., Selvaraj,S., Fang,J., Ren,B. and Liu,J.S. (2013) Bayesian Inference of Spatial Organizations of Chromosomes. *PLoS Comput. Biol.*, **9**, e1002893.
16. Trieu,T. and Cheng,J. (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res.*, **42**, e52.
17. Trieu,T. and Cheng,J. (2015) MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. *Bioinformatics*, doi:10.1093/bioinformatics/btv754.
18. Baù,D. and Marti-Renom,M.A. (2012) Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods*, **58**, 300–306.
19. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
20. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
21. Sanborn,A.L., Rao,S.S.P., Huang,S.-C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *PNAS*, **112**, E6456–E6465.
22. Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.
23. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
24. Cournac,A., Marie-Nelly,H., Marbouty,M., Koszul,R. and Mozziconacci,J. (2012) Normalization of a chromosomal contact map. *BMC Genomics*, **13**, 436–436.
25. Parada,L.A. and Misteli,T. (2002) Chromosome positioning in the interphase nucleus. *Trends Cell Biol.*, **12**, 425–432.
26. Cremer,T. and Cremer,M. (2010) Chromosome Territories. *Cold Spring Harb. Perspect. Biol.*, **2**, a003889.
27. Tanabe,H., Habermann,F.A., Solovei,I., Cremer,M. and Cremer,T. (2002) Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutat. Res.*, **504**, 37–45.