Bioinformatics, 32(9), 2016, 1286–1292
doi: 10.1093/bioinformatics/btv754
Advance Access Publication Date: 31 December 2015
Original Paper

OXFORD

Genome analysis

MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data

Tuan Trieu and Jianlin Cheng*

Computer Science Department, University of Missouri, Columbia, MO 65201, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on 30 August 2015; revised on 16 November 2015; accepted on 19 December 2015

Abstract

Motivation: The three-dimensional (3D) conformation of chromosomes and genomes play an important role in cellular processes such as gene regulation, DNA replication and genome methylation. Several methods have been developed to reconstruct 3D structures of individual chromosomes from chromosomal conformation capturing data such as Hi-C data. However, few methods can effectively reconstruct the 3D structures of an entire genome due to the difficulty of handling noisy and inconsistent inter-chromosomal contact data.

Results: We generalized a 3D chromosome reconstruction method to make it capable of reconstructing 3D models of genomes from both intra- and inter-chromosomal Hi-C contact data and implemented it as a software tool called MOGEN. We validated MOGEN on synthetic datasets of a polymer worm-like chain model and a yeast genome at first, and then applied it to generate an ensemble of 3D structural models of the genome of human B-cells from a Hi-C dataset. These genome models not only were validated by some known structural patterns of the human genome, such as chromosome compartmentalization, chromosome territories, co-localization of small chromosomes in the nucleus center with the exception of chromosome 18, enriched center-toward inter-chromosomal interactions between elongated or telomere regions of chromosomes, but also demonstrated the intrinsically dynamic orientations between chromosomes. Therefore, MOGEN is a useful tool for converting chromosomal contact data into 3D genome models to provide a better view into the spatial organization of genomes.

Availability and implementation: The software of MOGEN is available at: http://calla.rnet.missouri.edu/mogen/.

Contact: chengji@missouri.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The three-dimensional (3D) structures of chromosomes and genomes play an important role in biological processes such as gene regulation, DNA replication, gene–gene interaction, genome methylation, maintenance of genome stability and recurrent chromosomal translocation (Dekker, 2008; Fraser and Bickmore, 2007; Miele and Dekker, 2008; Misteli, 2007). Analyzing 3D structures of chromosomes and genomes is, therefore, an important task in studying functions of genomes. Fluorescence in situ hybridization (FISH) has been used to study the 3D organization of chromosomes and genomes, however, due to its low throughput and low resolution, it cannot be applied to study genomes at a fine or large-scale. Recently, chromosome conformation capture (3C)-based techniques such as Hi-C (Lieberman-Aiden *et al.*, 2009) and TCC (Kalhor *et al.*, 2012) have emerged as powerful tools for capturing the proximity of chromosomal fragments within the same chromosome (intra-

chromosomal contacts) or between different chromosomes (inter-chromosomal contacts). These chromosomal contacts can provide new insights into the 3D organization of genomes (Kalhor *et al.*, 2012; Lieberman-Aiden *et al.*, 2009). Moreover, they can also be used to infer 3D structures of chromosomes and genomes.

Several methods have been proposed to reconstruct 3D structures of chromosomes from chromosomal contacts, which adopt one of the two main strategies (Serra *et al.*, 2015). The first strategy relies on the principles of polymer physics of the chromatin to build models that are consistent with observed chromosomal contacts (Barbieri *et al.*, 2013; Mirny, 2011). The second strategy converts chromosomal contacts into restraints and then constructs models to satisfy these restraints. Methods utilize the second strategy are called restraint-based methods. A common approach used by restraint-based methods is to convert chromosomal contacts into spatial distances first and then solves the problem of satisfying these spatial restraints as an optimization problem. Depending how restraints are satisfied, the optimization procedure could yield one single model (Duan *et al.*, 2010; Lesne *et al.*, 2014; Varoquaux *et al.*, 2014; Zhang *et al.*, 2013) or a set of models (Hu *et al.*, 2013; Rousseau *et al.*, 2011). We developed a restraint-based method (Trieu and Cheng, 2014), in which contacts and non-contacts are supposed to satisfy a distance threshold rather than a specific distance value, which is tackled as an optimization problem. Different from most restraint-based methods, our method does not require converting contacts into distances. When contacts are highly consistent and can be satisfied altogether, our method produces similar models, but when contacts are largely inconsistent, different models are produced.

Although reasonable 3D structures of individual chromosomes can be built by some of these methods, it still difficult to reconstruct 3D structures of large genomes such as the human genome consisting of a number of chromosomes from both intra- and inter-chromosomal contacts. To the best of our knowledge, currently, no existing methods had built the structural models of the entire human genome consisting of 23 pairs of chromosomes from Hi-C data, which exhibit the known features of the human genome such as chromosome territories and the clustering of small chromosomes in the center with the exception of chromosome 18.

The limitation could partially be due to the difficulty of handling noisy and often inconsistent inter-chromosomal contacts of low interaction frequency (IF) to assemble individual chromosome structures together while satisfying intra-chromosomal contacts with much higher IF. In order to overcome this problem, we generalized our 3D chromosome reconstruction method (Trieu and Cheng, 2014) to make it capable of reconstructing 3D models of a large genome such as the human genome—utilizing both intra- and inter-chromosomal contact data. We benchmarked the method on two synthetic datasets of the yeast genome (Duan *et al.*, 2010) and of a polymer worm-like chain model (Trussart *et al.*, 2015) and the real Hi-C data of the human genome (Lieberman-Aiden *et al.*, 2009), and implemented it as a software tool called MOGEN for users to reconstruct 3D genome structures.

2 Methods

2.1 Overview of the genome structure reconstruction process

Our modeling goal is to find positions (i.e. x, y, z coordinates) for DNA fragments of chromosomes of a genome in the 3D space that satisfy intra-chromosomal/inter-chromosomal contacts and non-contacts between the fragments observed in Hi-C datasets of the genome as much as possible, which is a probabilistic contact-constrained spatial optimization problem. One challenge of reconstructing the 3D structure of the whole genome from Hi-C datasets is to deal with both erroneous and intrinsically inconsistent contacts caused by the limitations of Hi-C experiments (Yaffe and Tanay, 2011) or variations of genome conformations of a population of cells. Moreover, the highly dynamic nature of genome structures of a population of cells used to generate Hi-C datasets may be a main source of spatially inconsistent, yet non-noisy contacts. In order to deal with a large amount of inconsistent contacts, MOGEN aims to satisfy as many contacts with high probabilities as possible in order to build an ensemble of probable conformations to approximate the dynamic 3D genome structures of a population of cells of the same type.

MOGEN translates Hi-C intra- and inter-chromosomal contact data of a genome into an ensemble of 3D conformations by optimizing a scoring function that measures the realization of intra- and inter-chromosomal contacts and bonding distance between adjacent chromosomal regions in three steps (Fig. 1) : (i) normalize raw Hi-C interaction frequencies into contact likelihood ratios between chromosomal fragments; (ii) initialize coordinates (x, y, z) of fragments such that all contacts are satisfied initially and contacts between adjacent fragments are prioritized to be satisfied during the optimization process; (iii) optimize the scoring function to satisfy as many non-contacts and contacts with high probability as possible. MOGEN also evaluates structural models quantitatively and this evaluation is then used to adjust parameters of MOGEN. A video demonstrating this modeling process is available at: http://calla.rnet.missouri.edu/mogen/.

2.2 Normalization, analysis and preprocessing of contacts

The normalization protocol used in (Lieberman-Aiden *et al.*, 2009) was applied to convert each IF into a normalized IF to reduce the biases and noise in Hi-C data.

The analysis on the Hi-C dataset (Lieberman-Aiden *et al.*, 2009) shows that most of intra-chromosomal contacts have higher IF than most of inter-chromosomal contacts. For instance, 80.2% of intra-chromosomal contacts have IF larger than 0.65, while 84.5% of inter-chromosomal contacts have IF less than 0.65. This might be

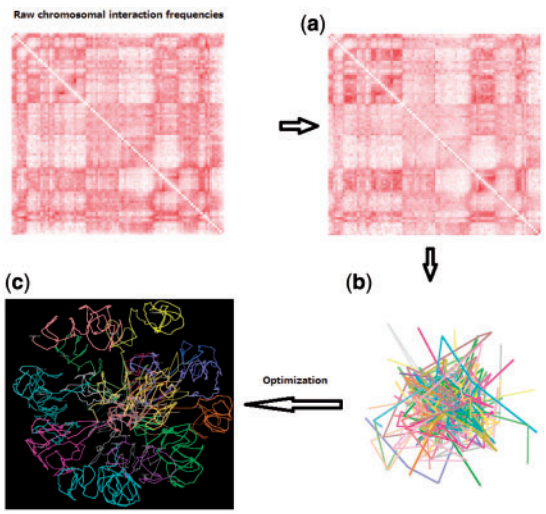

Fig. 1. The three steps of genome structure reconstruction in MOGEN. (**a**) Interaction frequencies between chromosomal fragments are normalized into contact likelihood ratios. (**b**) Initialize coordinates of fragments. (**c**) Maximize a scoring function to satisfy contacts and non-contacts

partially due to the situation that the 3D shapes of individual chromosomes may be more conserved than the relative orientations between chromosomes in individual cells.

Since contacts with low IFs could be noisy or less likely to happen, we set cut-off thresholds on IFs to select inter-chromosomal and intra-chromosomal contacts to construct preferred genome structures. Contacts with IF below these thresholds are considered as non-contacts. Theoretically, thresholds for intra- and inter-chromosomes could be the same. But because inter-chromosomal contacts are sparser and harder to be satisfied than intra-chromosomal contacts, we used a lower threshold for inter-chromosomal contacts in order to increase the number of satisfied inter-chromosomal contacts. In our experiment for the Hi-C dataset (Lieberman-Aiden *et al.*, 2009), the cut-off thresholds of 0.65 and 0.58 for intra-chromosomal and inter-chromosomal contacts, respectively, worked best in term of maximizing the number of chromosomal contacts that could be satisfied while preventing small chromosomes in the nucleus center from intermingling.

2.3 A probabilistic chromosomal contact-based scoring function

The core of our probabilistic contact-based method of modeling 3D genome structures (MOGEN) is a data-driven scoring function comprising terms that probabilistically constrain distance ranges between two adjacent (i.e. physically bonded) DNA fragments, between two spatially contacted fragments in the same chromosome, between two non-contacted fragments in the same chromosome, between two contacted fragments from two different chromosomes, and between two non-contacted fragments from two different chromosomes. This function is generalized from the function in (Trieu and Cheng, 2014) to include inter-chromosomal contacts and non-contacts.

$$
\begin{aligned}
Fn = & \sum_{\substack{\text{contacts} \\ \{(i,j):|i-j|\neq 1\}}} \left(W_1[\text{chr1},\text{chr2}] * \tanh(d_c^2 - d_{ij}^2) * \frac{F_{ij}}{\text{totalIF}} \right. \\
& \left. + W_2[\text{chr1},\text{chr2}] * \frac{\tanh(d_{ij}^2 - d_{\min}^2)}{\text{totalIF}} \right) \\
& + \sum_{\{(i,j):|i-j|=1 \,\&\, \text{chr1}=\text{chr2}\}} \left(W_1[\text{chr1},\text{chr2}] * \text{IF}_{\max} * \frac{\tanh\left(\text{da}_{\max}^2 - d_{ij}^2\right)}{\text{totalIF}} \right. \\
& \left. + W_2[\text{chr1},\text{chr2}] * \frac{\tanh\left(d_{ij}^2 - d_{\min}^2\right)}{\text{totalIF}} \right) \\
& + \sum_{\substack{\text{non}-\text{contacts} \\ \{(i,j):|i-j|\neq 1, \text{chr1}=\text{chr2}\}}} \left(W_3[\text{chr1},\text{chr1}] * \frac{\tanh\left(d_{\max\_\text{intra}}^2 - d_{ij}^2\right)}{\text{totalIF}} \right. \\
& \left. + W_4[\text{chr1},\text{chr1}] * \frac{\tanh(d_{ij}^2 - d_c^2)}{\text{totalIF}} \right) \\
& + \sum_{\substack{\text{non}-\text{contacts} \\ \{(i,j):|i-j|\neq 1, \text{chr1}\neq\text{chr2}\}}} \left(W_3[\text{chr1},\text{chr2}] * \frac{\tanh\left(d_{\max\_\text{inter}}^2 - d_{ij}^2\right)}{\text{totalIF}} \right. \\
& \left. + W_4[\text{chr1},\text{chr2}] * \frac{\tanh\left(d_{ij}^2 - d_c^2\right)}{\text{totalIF}} \right)
\end{aligned}
$$

The first term in the scoring function enforces distance constraints on contacted, but non-adjacent fragments and also their minimum distances for avoiding clashes according to contact probability proportional to interaction frequency. In other words, the value of the function increases as the distance between two contacted fragments (d_{ij}: distance between midpoints of ith fragment of chromosome chr1 and jth fragment of chromosome Chr2) gets smaller than the contact distance threshold d_c and/or gets larger than the minimum distance threshold between contacted fragments $d_{\min}$. Here, totalIF is the sum of normalized interaction frequencies of the whole genome calculated from the input Hi-C data; F_{ij} is the normalized IF between fragments i and j; and $\text{IF}_{\max}$ is the maximum IF among all contacts. F_{ij} weights the influence of a contact proportionally according to its normalized IF, i.e. a contact with higher IF contributes more to the value of the function, which essentially tries to enforce contacts in a probabilistic way. The second term enforces the minimum and maximum distance on two adjacent fragments on the same chromosome. $\text{da}_{\max}$ is the maximum distance threshold between two adjacent fragments on the same chromosome. The third term enforces constraints on pairs of fragments that are not in contact and in the same chromosome. In particular, the score increases as the distance between non-contacted fragments get larger than d_c and/or smaller than the maximum intra-chromosomal distance threshold $d_{\max_i\text{ntra}}$. The last term exerts forces on distances of pairs of fragments that are not in contact and in different chromosomes to make them less than the maximum inter-chromosomal distance threshold $\text{da}_{\max_i\text{nter}}$ and greater than d_c.

$W_1[\text{chr1},\text{chr2}]$, $_{W2}[\text{chr1},\text{chr2}]$, $W_3[\text{chr1},\text{chr2}]$ and $W_4[\text{chr1},\text{chr2}]$ are adjustable parameters for fragment pairs of chromosomes chr1 and chr2 to weigh and balance the influences of scoring terms associated with inter-chromosomal contacts or non-contacts if chr1 is not equal to chr2 and otherwise with intra-chromosomal contacts or non-contacts. To avoid biases against different chromosomes and reduce the complexity of optimization, the values of W_1, W_2, W_3 or W_4 for any pair of chromosomes are the same. Thus, the number of parameters is much smaller than assigning different weights to different pairs of chromosomes, which makes the parameter adjustment simpler. Theoretically, the weights for different chromosomes can be the same. However, when the same weights were used for all the chromosomes, the qualities of models for different chromosomes were different. This could be due to the different sizes and different ratios of contacts and non-contacts of different chromosomes (e.g. small chromosomes have higher contacts/non-contacts ratio). Therefore, allowing different weights for different chromosomes according to their individual properties can help produce better models for chromosomes. These parameters are adjusted to maximize the number of satisfied contacts and non-contacts and to prevent different chromosomes from intermingling with each other.

2.4 Estimation of parameters in the scoring function

The physical parameters ($d_{\min}$, $\text{da}_{\max}$, $d_{\max_i\text{ntra}}$, $d_{\max_i\text{nter}}$) for the human B-cells in the scoring function were approximately set according to the previous knowledge about the overall sizes of chromosomes and the nucleus gained from the chromosome structural modeling and FISH experiments (Mateos-Langerak *et al.*, 2009; Trieu and Cheng, 2014). The values of $d_{\min}^2$, $\text{da}_{\max}^2$ were set at $0.2\text{and}1.8\,(\mu\text{m}^2)$(micrometer), respectively, the same as in (Trieu and Cheng, 2014). The contact distance threshold d_c^2 was estimated based on the average distance of pairs of regions measured in the FISH data (Mateos-Langerak *et al.*, 2009) as described in Trieu and Cheng (2014). We approximated this value at $6\mu\text{m}^2$ (lower than $7\mu\text{m}^2$ in Trieu and Cheng, 2014) to reduce the size of chromosomes and the intermingling between them. $d_{\max_i\text{ntra}}^2$ was set at $20\mu\text{m}^2$ the same as $d_{\max}^2$ in Trieu and Cheng (2014). The unit of distances is in square for convenience because we use the square of distances in the scoring function to simplify the calculation of gradient and the implementation of the optimization. $d_{\max_i\text{nter}}$ was set to the approximate value of the diameter of the nucleus (i.e.13μm).

The weight parameters W_1, W_2, W_3 and W_4 for all chromosome pairs were estimated according to the two criteria: (i) intra-chromosomal contact and non-contact scores (percentage of contacts and

non-contacts that are satisfied) should be high (over 60%); and (ii) chromosome territories (separable space of each chromosome) must be largely enforced, i.e. the 3D space of chromosomes should not be seriously intermingled, which can be achieved when inter-chromosomal non-contact scores are high (over 60%). The rationale behind these two criteria is that each chromosome largely occupies its own territories without significant intermingling with others, but neighboring chromosomes could border each other closely (Cremer and Cremer, 2010; Parada and Misteli, 2002); and that inter-chromosomal contacts are much weaker than intra-chromosomal contacts and thus should have a lower priority of being satisfied. In searching for values of these parameters, they were initialized to 0.05, and then, depending on which contact or non-contact scores were low, their corresponding weight parameters would get increased. The intra-chromosomal contact and non-contact scores of chromosomes were expected to be higher than 60% in order to ensure good chromosome structures. The parameters corresponding to inter-chromosomal contact scores were increased to get inter-chromosomal contacts scores as high as possible, while at the same time inter-chromosomal non-contact scores must be maintained higher than 60%. This would maximize interactions between chromosomes while maintaining chromosome territories. For the dataset of the human genome used here, the range of values was finally settled within [0.05, 14.5], but it may vary for different datasets. The detailed values and configuration of the parameters are described in the Supplementary Material.

2.5 Structure initialization

The (x, y, z) coordinates of fragments of each chromosome were randomly initialized with values in a small, symmetric range $[-0.5, 0.5]$ to facilitate the gradient-ascent optimization, subject to the condition that the difference between the coordinates of two adjacent fragments were not larger than a small number (i.e. 0.025) to make sure the proximity of adjacent fragments was satisfied during the optimization. This simple initialization is important for the successful structure optimization when the input data contains noisy inter-chromosomal contacts.

2.6 Structure optimization

An important component of the method is an optimizer that maximizes the scoring function Fn above to generate genome structures. We used the gradient ascent optimization with adaptive step sizes to generate genome structures starting from randomly initialized structures. The search for a new step size was performed only when the scoring function stopped increasing. For convenience, MOGEN allows users to specify the initial step size. Calculating the value of the scoring function and its gradient with respect to (x, y, z) coordinates was the bottleneck of the optimization process so parallel threads were used in MOGEN to speed up the calculation.

2.7 Implementation

MOGEN was implemented in Java and can run on different platforms. The core of the tool is an optimizer to optimize the scoring function as described in Section 2.6. Structures were initialized as in Section 2.5. The initial structure was further optimized by gradient ascent according to the scoring function. The final structures are stored in the Protein Data Bank (PDB) format and can be visualized by popular visualizers such as PyMOL or Chimera. For each structure, MOGEN also computes contact and non-contact scores and outputs these scores in a separate text file. The values of the thresholds and weights used by MOGEN are configured in text files. The implementation details are described in Supplementary Material.

3 Results

We validated our method on two independent synthetic datasets, one from a polymer worm-like chain model (Trussart *et al.*, 2015) and one from a yeast genome model consisting of several chromosomes (Duan *et al.*, 2010). The validation on the dataset of the worm-like chain model is described in Section 3.1 and the validation on the dataset of the yeast genome is included in the Supplementary Material.

3.1 Validation on synthetic polymer worm-like chain models and comparison with TADbit

We used a synthetic dataset in Trussart *et al.* (2015), which simulated both noise and structural variability in Hi-C data using a polymer worm-like chain model of length ~1 Mb represented by 202 bins of size 5 kb, to validate our method. The model also simulated Topologically Associated Domain (TAD)-like architecture (Dixon *et al.*, 2012). There are 7 levels of structural variability represented by 7 sets of 100 models (called true structures), denoting as Set 0, 1, 2, 3, 4, 5 and 6 varied from low structural variability to high variability. And at each level of structural variability, there are 4 levels of noise (50, 100, 150 and 200, respectively). In total, there are 28 contact matrices (i.e., four matrices for each set of 100 true structures). Matrices were normalized as in Trussart *et al.* (2015). For each input matrix, we generated an ensemble of 100 structures. We checked if there was a structure in the ensemble that was similar to the true structures. Following Trussart *et al.* (2015), the Spearman correlation was used to assess if two structure were similar. The structure in the ensemble that is most similar to each true structure was selected, and the correlations between the true structures with their selected counterparts were averaged and used as the accuracy of the ensemble of generated structures. Figure 2 shows these accuracies for the ensembles of structures generated at different levels of noise and structural variability. It can be seen that the accuracy of generated structures is high when the structural variability is low, even though the noise level is high. However, the average quality of generated structures decreases as the level of structural variability increases. Overall, MOGEN is robust against noise and performs well at reasonable levels of structural variability.

We compared the performance of MOGEN and TADbit (Baù and Marti-Renom, 2012) on the synthetic dataset. The structures generated by TADbit were provided together with the synthetic dataset. There is one structure generated by TADbit for each input matrix. For each structure generated by TADbit, we selected its most similar counterpart in the ensemble of true structures, and used the corresponding correlation between the two to compare with the correlation between the most similar pair of structures generated by MOGEN and true structures. We performed the comparison on the lowest noise level and the highest noise level, respectively. Figures 3 and 4 show the results on the lowest and highest noise levels, respectively. Overall, their performances are comparable across different levels of noise and structural variability, while MOGEN seems to be somewhat more robust against noise and TADbit seems to be less sensitive to structural variability. But the difference is not significant.

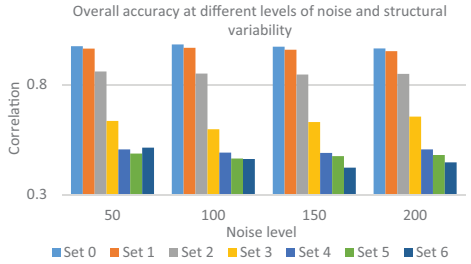

Fig. 2. Accuracy of structures generated by MOGEN at different levels of noise and structural variability

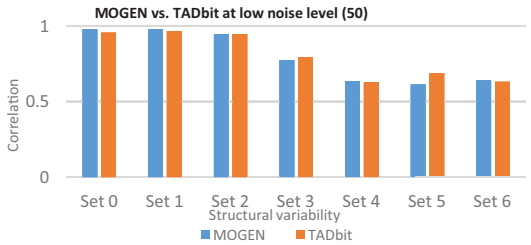

Fig. 3. Comparison between MOGEN and TADbit when noise level is low

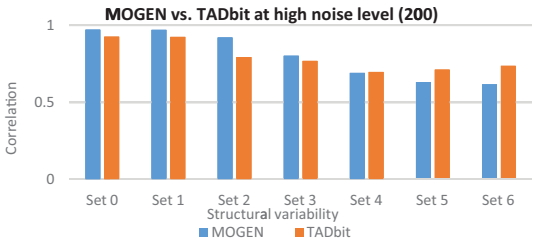

Fig. 4. Comparison between MOGEN and TADbit when noise level is high

3.2 The structures of the human genome reconstructed from real Hi-C data

We used MOGEN to generate an ensemble of 500 genome structures from the Hi-C data of the normal human lymphoblastoid cell line (GM06990) (Lieberman-Aiden *et al.*, 2009). The dataset was normalized as described in Section 2.2. The distance parameters were set as in Section 2.4. We then adjusted weights as described in Section 2.4 to generate structures. As for this dataset, like all existing methods, MOGEN could not distinguish homologous chromosomes and identify interactions between all pairs of 46 chromosomes. So we built the genome structures consisting of one complete set of 23 chromosomal models, each corresponding to the average structure of a pair of homologous chromosomes. These structures have the known features of the human genome. The generated structures are available at http://calla.rnet.missouri.edu/mogen/.

Figure 5 visualizes a 3D genome structure (chromosomes 1–23 were labeled as letter A-X, respectively, with the letter L omitted). It is shown that the model exhibits the known genome structural feature of chromosome territories as expected, i.e. each chromosome largely occupies its own space instead of mixing together, even though there are some strong interactions between some regions of a few pairs of small chromosomes in the center of the nucleus. This demonstrates that the parameters of the modeling process had been adjusted effectively to realize the separate territories of individual chromosomes during the process of satisfying intra- and inter-chromosomal contacts.

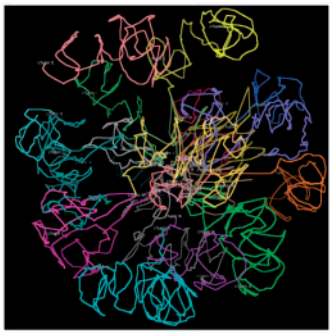

Fig. 5. A genome structure with chromosomes colored in different colors

A striking feature of the global topology of the human genome is that smaller chromosomes (chromosome 16, 17, 19, 20, 21 and 22) except chromosome 18 form a core at the center surrounded by larger chromosomes (chromosome 1–15, X) lying near the periphery (Fig. 6a).

3.3 Consistency of individual chromosome structures across different genome structures

By visualizing structures, it is observed that individual chromosome structures in different genome structures in the ensemble were similar. For instance, the average structural similarity score (the GDT-HA score) of the structural models of chromosome 1 (i.e. the largest chromosome of the human genome) calculated by a modified version of the TM-score program (Trieu and Cheng, 2014; Zhang and Skolnick, 2004) is 0.71 (with minimum, maximum and standard deviation are 0.5, 0.98 and 0.06, respectively, which is much higher than the similarity score (generally < 0.2) of two random structures of the same length. The structures of individual chromosomes also have the two-compartment feature (Lieberman-Aiden *et al.*, 2009; Trieu and Cheng, 2014) (Fig. 7), e.g. the two chromosomal compartments (euchromatin and heterochromatin) of chromosome 11 identified by the principal component analysis as in Lieberman-Aiden *et al.* (2009) and Trieu and Cheng (2014) were spatially clustered together as expected. These results may suggest that the global shape of individual chromosomes may be largely preserved in the varied genome structures of the population of cells used to generate the Hi-C data.

3.4 Variability of the relative orientation and position of chromosomes in genome structures

Despite the relatively high similarity between chromosome structures in different genome models, the similarity between the overall genome structures in the ensemble is very low. The average GDT-HA score between genome structures is 0.16, suggesting that the conformations of the whole genomes, particularly, the orientations of chromosomes in the ensemble, vary a lot. This is not surprising because the genome conformations of single cells in the cell population used to generate the Hi-C data likely vary. Indeed, different from the majority of intra-chromosomal contacts being satisfied, satisfying inter-chromosomal contacts is more difficult possibly due to the intrinsic inconsistency within inter-chromosomal contacts caused by the dynamics of the genome structures in the population

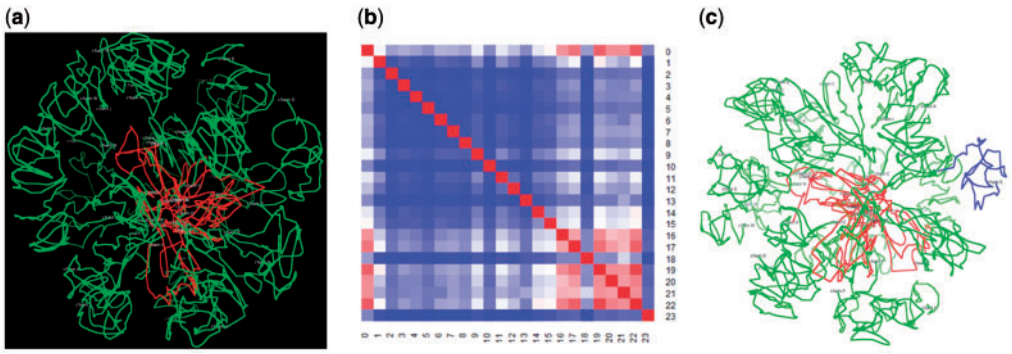

Fig. 6. (**a**) Small chromosomes except chromosome 18 (in red) cluster in the center. (**b**) Heatmap of average distances between centers of the mass of chromosomes (1–23) and the center of the mass of the genome (0) (computed from all genome structures); the intensity of red corresponds to proximity. (**c**) The small chromosome 18 (blue) lies near the periphery

Fig. 7. Two-compartment feature in a structure of chromosome 11 extracted from a genome structure. Fragments of compartments were obtained from principal component analysis and then colored in red and green

of cells and possibly higher noise and variation in inter-chromosomal contacts.

However, despite the variability in the genome structures, the Pearson's correlation between the inter-chromosomal interaction matrix calculated from the Hi-C data and that derived from all reconstructed 3D models in the ensemble is 0.48 with an significant e-value $P < 2.2e-16$. This suggests that each genome structure still captures a significant portion of inter-chromosomal contacts that can be consistently realized in one genome structural model, which may correspond to the realistic structure of one cell or a subset of cells and/or to the structural patterns conserved in all genome structures. Indeed, there is a consistent global topology of genome models that smaller chromosomes are preferably clustered together to form a more rigid core of the genome, and that larger chromosomes lie in the periphery of the genome surrounding small chromosomes (Fig. 6a).

Figure 6b illustrates the average distance between the centers of the mass of individual chromosomes (indexed from 1 to 23) and the center of the mass of the genome structure (indexed as 0) calculated from all genome models. Except for chromosome 18, the centers of small chromosomes were near the center of the genome while the centers of larger chromosomes were further away from it, which is consistent with the previous study that small chromosomes frequently co-localize in the center of the nucleus (Boyle *et al.*, 2001; Tanabe *et al.*, 2002). However, as an exception, the location of the small but gene-poor chromosome 18 is near the periphery also agrees with the previous FISH study (Croft *et al.*, 1999) (Fig. 6c).

3.5 Contact patterns between elongated regions of chromosomes

In light of this core–periphery architecture of the 3D human genome, we investigated how chromosomes may interact to form this organization. We found that most interactions between two chromosomes occurred between telomeres (within 10 MB from the two ends of a chromosome), between telomeres and centromeres (within 5 MB from the omitted centromere loci of a chromosome), and between centromeres as shown in Imakaev *et al.* (2012). Figure 8a visualizes the heatmap of average inter-chromosomal distances between chromosomal regions computed from all 3D genome models (the intensity of red indicates proximity). And if two chromosomes are in contact, it is more likely that their telomeres are in contact. Centromere–centromere or centromere–telomere interactions are also more likely than regions outside of centromeres and telomeres, although the difference between centromere–centromere distances and centromere–telomere distances is not significant. Figure 8b illustrates interactions between chromosome 4, 6, 11 and 14 occurring in telomeres and centromeres. In addition to the enriched interaction among telomeres and centromeres, large chromosomes lying near the periphery generally have elongated regions, centromeres and/or telomeres that extend towards the nucleus center, where small and gene-rich chromosomes are located (Fig. 8c). And this interaction pattern is a consistent structural feature among most 3D genome models even though the overall orientations and relative locations between chromosomes are highly varied. As a result, if a large chromosome interacts with other chromosomes in the nucleus center, some specific regions (telomeres, centromeres and/or an elongated region) rather than other regions are often involved in interactions.

4 Conclusions

We developed a method called MOGEN to reconstruct 3D genome structures from Hi-C data. We tested MOGEN on the synthetic datasets of a yeast genome and of a polymer worm-like chain model and then applied it to reconstruct 3D genome structures of human B-cells. The rigorous testing shows that MOGEN is a useful method for translating chromosomal Hi-C data into 3D structures to facilitate the study of the spatial organization of genomes and its role in biological processes such as gene regulation, epigenetic modification and DNA replication.

Fig. 8. (**a**) Interaction between telomeres and centromeres. The intensity of red indicates proximity. The row (left to right)/column represents regions going from telomere to centromere. (**b**) Telomeres and centromeres of chromosome 4, 6, 11, 14 in a genome structure interact with each other (the white circle). (**c**) Large chromosomes extend toward the genome center (the white circle)

Funding

This work is partially supported by a National Science Foundation CAREER award (grant no: DBI1149224) to J.C.

Conflict of Interest: none declared.

References

Barbieri,M. *et al.* (2013) A model of the large-scale organization of chromatin. *Biochem. Soc. Trans.*, **41**, 508–512.

Baù,D. and Marti-Renom,M.A. (2012) Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods San Diego Calif.*, **58**, 300–306.

Boyle,S. *et al.* (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.*, **10**, 211–219.

Cremer,T. and Cremer,M. (2010) Chromosome territories. *Cold Spring Harb. Perspect. Biol.*, **2**,

Croft,J.A. *et al.* (1999) Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.*, **145**, 1119–1131.

Dekker,J. (2008) Gene regulation in the third dimension. *Science*, **319**, 1793–1794.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Duan,Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.

Fraser,P. and Bickmore,W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413–417.

Hu,M. *et al.* (2013) Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.*, **9**, e1002893.

Imakaev,M. *et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Kalhor,R. *et al.* (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.

Lesne,A. *et al.* (2014) 3D genome reconstruction from chromosomal contacts. *Nat. Methods*, **11**, 1141–1143.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Mateos-Langerak,J. *et al.* (2009) Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci.*, **106**, 3812–3817.

Miele,A. and Dekker,J. (2008) Long-range chromosomal interactions and gene regulation. *Mol. Biosyst.*, **4**, 1046–1057.

Mirny,L.A. (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.*, **19**, 37–51.

Misteli,T. (2007) Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787–800.

Parada,L.A. and Misteli,T. (2002) Chromosome positioning in the interphase nucleus. *Trends Cell Biol.*, **12**, 425–432.

Rousseau,M. *et al.* (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, **12**, 414.

Serra,F. *et al.* (2015) Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.*, **589**, 2987–2995.

Tanabe,H. *et al.* (2002) Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutat. Res.*, **504**, 37–45.

Trieu,T. and Cheng,J. (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res.*, **42**.

Trussart,M. *et al.* (2015) Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. *Nucleic Acids Res.*, gkv221.

Varoquaux,N. *et al.* (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, **30**, i26–i33.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhang,Z. *et al.* (2013) 3D chromosome modeling with semi-definite programming and Hi-C data. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **20**, 831–846.