PLOS ONE

RESEARCH ARTICLE

Sequence-based multiscale modeling for high-throughput chromosome conformation capture (Hi-C) data analysis

Kelin Xia[1,2]*

1 Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore, **2** School of Biological Sciences, Nanyang Technological University, Singapore 637371, Singapore

* xiakelin@ntu.edu.sg

OPEN ACCESS

Citation: Xia K (2018) Sequence-based multiscale modeling for high-throughput chromosome conformation capture (Hi-C) data analysis. PLoS ONE 13(2): e0191899. https://doi.org/10.1371/journal.pone.0191899

Editor: Zhaohui Qin, Emory University Rollins School of Public Health, UNITED STATES

Received: July 18, 2017

Accepted: January 12, 2018

Published: February 6, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Nanyang Technological University (http://www.ntu.edu.sg/) Startup Grant M4081842.110 and Singapore Ministry of Education (https://www.moe.gov.sg/) Academic Research Fund Tier 1 M401110000. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

In this paper, we introduce sequence-based multiscale modeling for biomolecular data analysis. We employ spectral clustering method in our modeling and reveal the difference between sequence-based global scale clustering and local scale clustering. Essentially, two types of distances, i.e., Euclidean (or spatial) distance and genomic (or sequential) distance, can be used in data clustering. Clusters from sequence-based global scale models optimize spatial distances, meaning spatially adjacent loci are more likely to be assigned into the same cluster. Sequence-based local scale models, on the other hand, result in clusters that optimize genomic distances. That is to say, in these models, sequentially adjoining loci tend to be cluster together. We propose two sequence-based multiscale models (SeqMMs) for the study of chromosome hierarchical structures, including genomic compartments and topological associated domains (TADs). We find that genomic compartments are determined only by global scale information in the Hi-C data. The removal of all the local interactions within a band region as large as 10 Mb in genomic distance has almost no significant influence on the final compartment results. Further, in TAD analysis, we find that when the sequential scale is small, a tiny variation of diagonal band region in a contact map will result in a great change in the predicted TAD boundaries. When the scale value is larger than a threshold value, the TAD boundaries become very consistent. This threshold value is highly related to TAD sizes. By the comparison of our results with those previously obtained using a spectral clustering model, we find that our method is more robust and reliable. Finally, we demonstrate that almost all TAD boundaries from both clustering methods are local minimum of a TAD summation function.

Introduction

The chromosome, the physical realization of genetic information, is one of the most complex and important cellular entities [1–7]. Over the past few decades, the significance of its three-dimensional architecture for supporting essential biological functions, such as DNA

Competing interests: The author has declared that no competing interests exist.

replication, transcription, repair of DNA damage and chromosome translocation, has gradually been realized [8–11]. Chromosome conformations are found to be deeply involved in the development of epigenetic organizations, the regulation of genome functions and the epigenetic inheritance of various cell states [8]. A thorough understanding of the chromosome three-dimensional structure is of fundamental importance to the decryption and interpretation of genetic information, and has become one of the most important topics in genomic and epigenetic research. Chromosome conformation capture (3C) technique [12, 13] and its derived methods, including chromosome conformation capture-on-chip (4C) [14, 15], chromosome conformation capture carbon copy (5C) [16] and high-throughput chromosome conformation capture (Hi-C) [17], have been developed and begun to uncover general features of genome organization [17–25].

Recent studies on Hi-C data have demonstrated the existence of two types of structures known as topologically associating domains (TADs) [18, 19] and genomic compartments [17]. TADs are chromosome components that are about 200 kilobases(Kb) to 2 megabases(Mb). They are originally found as the contiguous square regions along the diagonal Hi-C maps with large contact values. More importantly, TADs are very consistent between different cell types and species and their spatial distributions are highly correlated with many genomic features such as histone modifications, coordinated gene expression, lamina and DNA replication timing. Through principle component analysis, two types of genomic compartments, i.e., A and B, have been identified [17]. More specifically, the compartment B is more densely packed with higher contact frequencies. On the contrary, the compartment A is chromosome regions that are more open and accessible. It strongly correlates with the gene loci and higher gene expression. More recently, analysis on the 1Kb resolution Hi-C data indicates the existence of six different subcompartments [26].

Based on Hi-C data, various algorithms and models are proposed to study the hierarchical structure of chromosomes [17, 18, 27–35]. Since TADs are essential to the understanding of relationship between chromosome structure and gene transcription, developing efficient algorithms to detect TADs is an important topic in Hi-C data analysis. Computationally, hidden Markov models (HMMs) are the first method to identify TADs [18]. Based on the contacts located 2Mb upstream and downstream, a directionality index of a locus is calculated in this model and used to capture the sharp transitions at TADs boundaries. After that, the arrowhead algorithm with a "corner score" is proposed [26]. This special score indicates the likelihood of each locus to be at a TAD boundary and can be efficiently evaluated by using dynamic programming. Meanwhile, a resolution parameter is considered to identify TADs at various scales. This algorithm has been incorporated into the software Armatus [27]. Further, a block-wise segmentation model called HiCseg [28] is proposed. This method reduces the problem of maximizing the likelihood with respect to the block boundaries into a 1D segmentation problem, and then employ the standard dynamic programming. More recently, a spectral graph theory based model is developed for the identification of TADs [36]. In this model, Laplacian based graph segmentation is applied iteratively to obtain TADs at the given compactness level. All the above mentioned methods can be roughly divided into two categories, optimization based local models and graph based global models. In local models, TAD indicators, including directionality index, corner score, likelihood of TAD boundaries, block-segmentation, are all evaluated locally within a certain region. In global models, TAD indicators, including eigenvectors, within-cluster variance, cluster distances, among others, are all evaluated globally in the whole domain of Hi-C data.

In this paper, two sequence-based multiscale models (SeqMMs) are introduced. Unlike previous clustering models, we measure the "similarity" of loci by not only their spatial distances but also their sequential distances. With the combination of spectral graph method, we find

that clusters from sequence-based global scale models optimize Euclidean distance relations, and these models can be used in genomic compartment analysis. In contrast, clusters from local scale models optimize genomic distance relations, and these models can be used in TAD analysis. Essentially, our SeqMMs provide a way to explore the hierarchical structures of chromosomes.

Mathematically, genomic compartments are defined from principal component analysis [17], they are global structural features. The loci in the same genomic compartment are spatially close to each other. But their sequential distances can be very large. Based on global scale clustering, we design Type-1 SeqMM and use it for genomic compartment analysis. In contrast, TADs are local structural features. The loci in the same TAD are not only spatially close to each other, but also sequentially adjacent to each other. Their sequential distances are usually within a certain genomic distance. Based on local scale clustering, we introduce Type-2 SeqMM and use it in TAD analysis.

Methods

As a discrete representation of geometries, manifolds, high-dimensional structures, abstract relations and complicated subjects, point cloud data (PCD) are widely used in computer science, engineering, scientific computing and data science. Particularly, PCD and PCD based classification or clustering methods [37], including K-means, hierarchical clustering, spectral clustering, modularity, graph centrality, network approaches, etc, have been constantly used in biomolecular data analysis. However, as demonstrated in Fig 1, biomolecular structure data are essentially different from the general PCD, as they incorporate a unique sequential information. The simulated structure corresponds to chromosome 22 from Human ES Cell line and is generated by using software shRec3D [33].

To have an intuitive understanding of the sequential information in PCD analysis, we consider a DNA structure with PDB ID 1ZEW. Using atomic coordinates, a weight matrix is

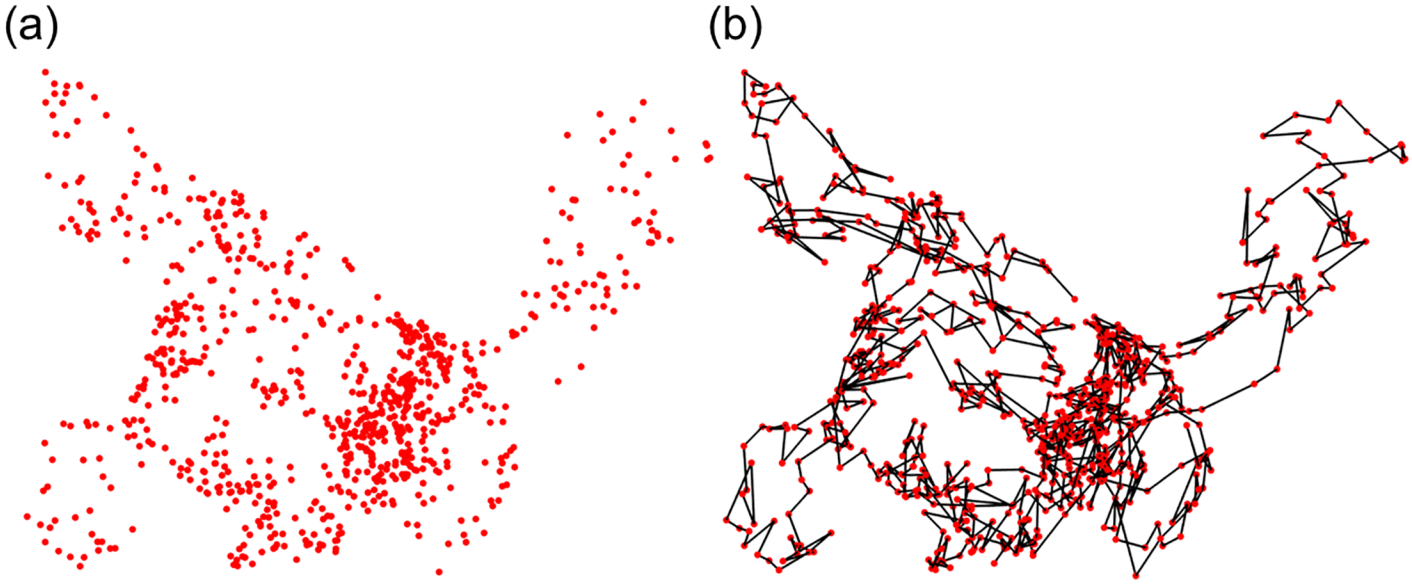

Fig 1. The comparison between point cloud data and biomolecular data. The simulated configuration of Human ES Cell chromosome 22 are generated by using the software shRec3D [33].

https://doi.org/10.1371/journal.pone.0191899.g001

constructed. The weight values are defined by using the rigidity function [38],

$$M = \{M_{ij} = e^{-(r_{ij}/\eta)^2} \mid i = 1, 2, \ldots, N;\ j = 1, 2, \ldots N\}, \quad (1)$$

where r_{ij} is the Euclidean distance between i-th and j-th atoms, N is the total number of atoms and η is the scale parameter that controls the influence range of each atom. In this case, we choose $\eta = 8$ Å. The weight matrix is illustrated in Fig 2(c_2). Two more matrices are

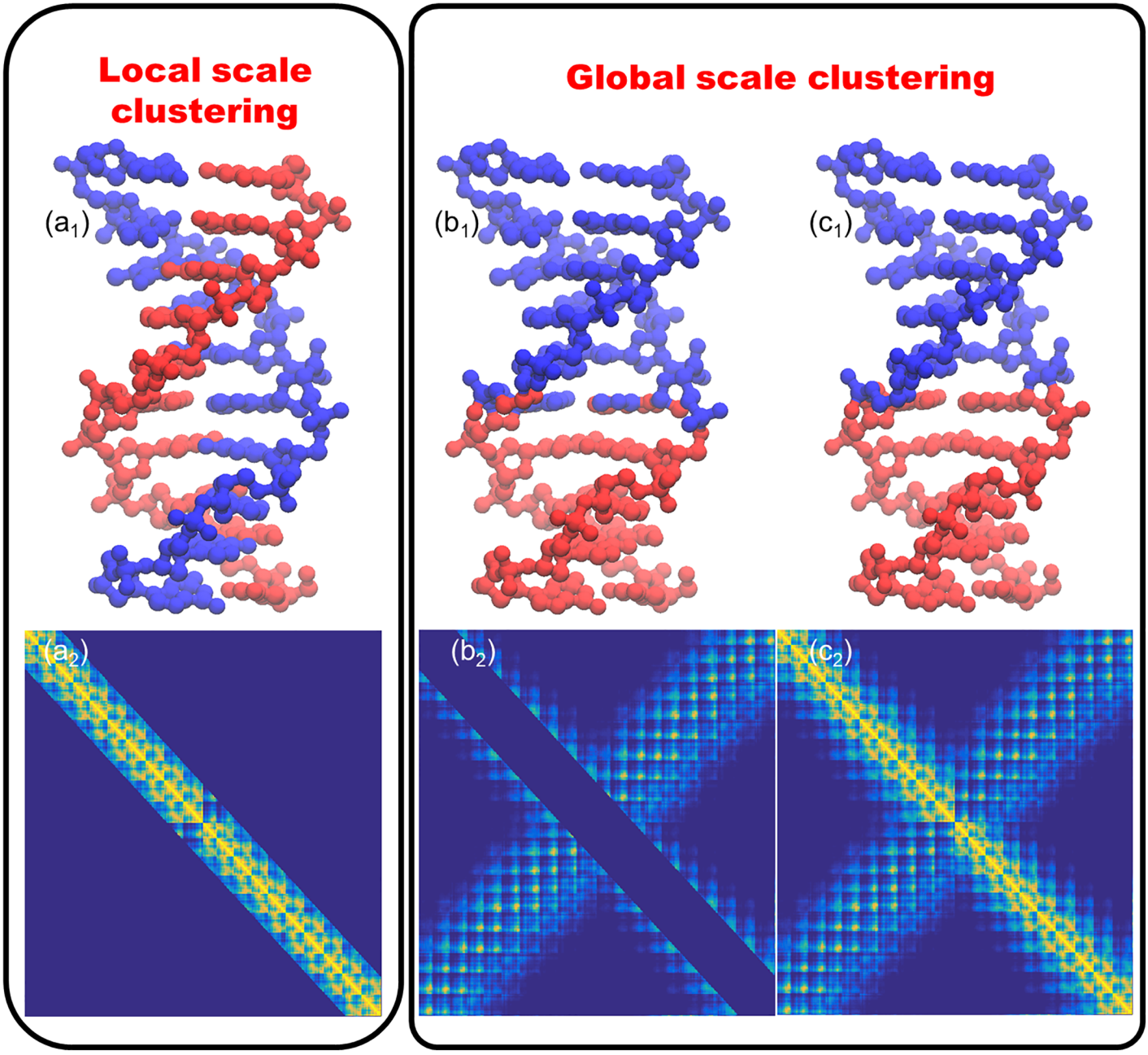

Fig 2. The illustration of two essential different clustering approaches, local scale clustering and global scale clustering. The local scale clustering optimizes sequential distances and is suitable for TAD analysis. The global scale clustering only considers spatial information and can be used in genomic compartment analysis.

https://doi.org/10.1371/journal.pone.0191899.g002

constructed by dividing the weight matrix into a diagonal band region as in Fig 2($\mathbf{a}_2$) and the remaining off-diagonal regions as in Fig 2($\mathbf{b}_2$). Based on these three matrices, we can decompose the DNA structure into two parts using the spectral clustering method [37, 39]. Results are illustrated in Fig 2($\mathbf{a}_1$), 2($\mathbf{b}_1$) and 2($\mathbf{c}_1$). It can be seen that, if we only consider relations between sequentially adjacent atoms, which are represented in the diagonal region, the DNA structure will be clustered into two complementary helix chains. However, if we use the whole matrix or only off-diagonal regions, the DNA structure will be divided in the middle region with two chains in each cluster.

Generally speaking, Fig 2 demonstrates two types of sequence-based models, i.e., sequence-based local models and sequence-based global models. It can be seen that their properties in structure decomposition differ greatly. In the first type, atoms with shorter sequential distances are more likely to be grouped into the same cluster. In the second one, spatially close atoms, i.e., atoms with large weight values, are more likely to be assigned to the same cluster. Mathematically, the sequence-based local model optimizes sequential distances, while the global model optimizes spatial distances or Euclidian distances. All PCD based classification and clustering methods belong to the second type. Therefore, the direct application of these methods in biomolecular data analysis may have some limitations.

In Hi-C data analysis, sequential information is usually highly relevant. Fig 3 demonstrates a potential problem for global scale clustering in TAD analysis. In this figure, genomic loci are represented by red pentagons. It can be seen that the spatial distance between the two loci in any red circle is much shorter than the one in green circles, while sequential distances are exactly the opposite. If we use the traditional PCD based clustering methods, two loci in the same red circle will always have priority to be clustered into the same TAD. Obviously, this will cause serious interpretation problems, as the sequential distance between the two loci can be much larger than the size of a TAD.

Sequence-based multiscale modeling

It should be noticed that two distance definitions, i.e., Euclidean distance and sequential distance, are greatly different and matter a lot in multiscale modeling. The Euclidean distance is just the three dimensional distance between two elements. In Hi-C data, Euclidean distance between genomic loci is inversely related to their contact frequencies [35]. In contrast, the sequential distance is defined between two elements on chains or polymers. If sequential numbers are assigned to elements, a sequential distance is just the difference between these two integers and it is always an integer. In Hi-C data, sequential distance between two loci is their genomic distance.

Even though graph and network based multiscale models are widely used in biomolecular structure and function analysis [40–49], the measurement defined in these models are in terms of the Euclidean distance. To be more specific, when we discuss atomic scale, residue scale, second structure scale, tertiary structure scale, etc, we are analyzing structural elements based on their sizes measured in Euclidean distances.

In this section, the sequence-based multiscale modeling is proposed for biomolecular data analysis, particularly for Hi-C data analysis. In our multiscale models, a scale parameter N_b is defined not from the Euclidean distance but from the sequential distance. The parameter N_b can be viewed as a cut-off sequential distance. In Hi-C matrices, the parameter N_b specifies the size of the diagonal band region. Further, two sequence-based multiscale models are proposed for analyzing chromosome genomic compartments and TADs. These two models, denoted as

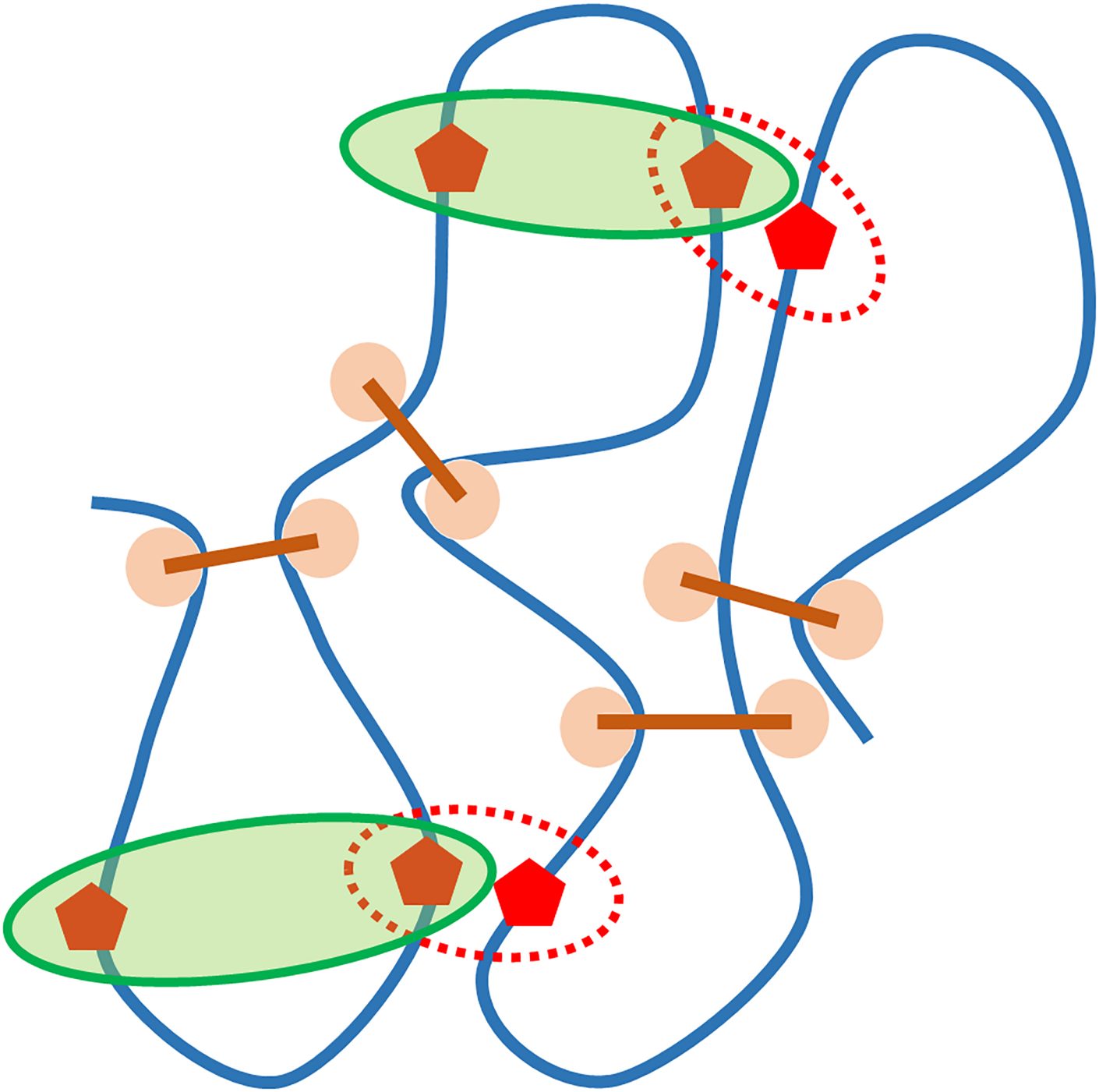

Fig 3. A potential problem for global scale clustering in TAD analysis. Each locus is represented by a red pentagon. Global scale clustering considers only spatial relations, thus groups loci in each red dash circle as a cluster. Biologically, loci in each green circle are more favorable to be clustered into the same TAD, as their sequential distances are much shorter. The missing of sequential information in global scale clustering will cause problems in the TAD analysis.

https://doi.org/10.1371/journal.pone.0191899.g003

Type-1 SeqMM and Type-2 SeqMM, are derived from the perspective of local scale clustering and global scale clustering, respectively.

Type-1 SeqMM. In Type-1 SeqMM, we remove sequentially short-range interactions by changing the value of scale parameter N_b. More specifically, for a Hi-C matrix, a diagonal band

region with size N_b is systematically removed from the model, resulting a new Hi-C matrix as following,

$$M_{ij}^{\text{Seq}} = \begin{cases} M_{ij}, & |i-j| \geq N_b \\ 0, & |i-j| < N_b \end{cases}. \tag{2}$$

Here M_{ij} can be the original or normalized contact frequencies. It can seen that our Type-1 SeqMM is defined by taking away the local interactions from the model and is designed for global scale clustering. An example can be found in Fig 6(**a**) to 6(**c**). We suggest that it can be used in chromosome genomic compartment analysis.

Type-2 SeqMM. In Type-1 SeqMM, when short-range interactions are systematically removed from the biomolecular data, long-range interactions are preserved. Type-2 SeqMM is designed in the exact opposite way,

$$M_{ij}^{\text{Seq}} = \begin{cases} M_{ij}, & |i-j| > 0 \text{ and } |i-j| \leqslant N_b \\ 0, & |i-j| > N_b \\ -\sum_{i \neq j}^{N} M_{ij}^{\text{Seq}}, & i = j \end{cases}. \tag{3}$$

The scale parameter N_b controls the size of the diagonal band region.

Mathematically, our SeqMM matrix in Eq (3) is a weighted Laplacian matrix, which plays an important role in graph representation and spectral clustering [37, 39, 50]. The second smallest eigenvalue and its associated eigenvector from the Laplacian matrix, are known as the Fiedler value (or algebraic connectivity) and the Fiedler vector, respectively. The Fiedler value is an important measurement of the general topological connectivity of a graph. The Fiedler vector gives an optimized classification of a graph into two separated domains [39, 50]. In our Type-2 SeqMM, the local interaction region can be systematically enlarged to model the different scales of interactions.

Type-2 SeqMM is proposed for chromosome TAD analysis. After Hi-C data preprocessing, a weighted Laplacian matrix can be generated by using a suitable scale value N_b. The TAD number in the data is estimated based on size and resolution of the Hi-C matrix. We assume the size of TAD to be around 2Mb, and TAD number N_c can be roughly calculated by dividing the total length over 2Mb. The basic procedure is presented in **Algorithm 1**. It should be noticed that the final number of TADs is usually larger than N_c. The Code is available at S1 File.

Algorithm 1 Type-2 SeqMM for TAD analysis

Pre-processing: Remove all rows and columns, that summed together equal to zero (or smaller than a pre-defined range); Transform the Hi-C contact frequencies to spatial distances (default function $f(x) = log(1 + x)$);

Step 1: Choose a scale parameter N_b to construct a weighted Laplacian matrix as in Eq 3;

Step 2: Calculate the first N_c eigenvectors. Here N_c is the estimated number of TADs;

Step 3: Employ the K-means algorithm on the N_c eigenvectors to identify N_c clusters;

Step 4: Subdivide each cluster into several TADs until the loci in each TAD are sequentially contiguous.

Results

Genomic compartments

The genomic compartment is defined from the principal component analysis of Hi-C data. Mathematically, the principal component captures the global shape of a structure. In Chen's spectral method [36], it shows that the genomic compartment results from the first principal component (FPC) are identical to the predictions made from the lowest-frequency eigenvector of weighted Laplacian matrices. More interesting, as proved in the elastic network model and normal mode analysis, these lowest-frequency eigenvectors are uniquely determined by the global geometric information of structures [51–54].

Since the FPC describes the global properties of a structure, we use the Type-1 SeqMM for our genomic compartment analysis. We consider the GM06990 chromosome 14 data with resolution 100Kb. This is a classic example used for genomic compartment analysis [17]. Before the principal component analysis (PCA), the chromosome 14 Hi-C matrix is processed. We remove all columns and rows with all zero values and normalize the matrix using the Toeplitz matrix [17]. After that, we construct a new matrix by removing the diagonal band region with $N_b = 60$ from the normalized Hi-C matrix, and calculate its FPC. Further, we compare this new FPC with the original one. Results are shown in Fig 4. The blue line represents the FPC from the original Hi-C matrix and red line represents the FPC from the off-diagonal matrix. It can be seen that they are almost identical to each other. Actually, the Pearson correlation coefficient (PCC) between the two FPCs is as high as 0.991, meaning that the removal of the diagonal band region have almost no influence to the FPC.

To have a more quantitative understanding of the FPC and Hi-C diagonal regions, we continuously change the value of the scale parameter N_b to generate a series of Hi-C matrices at different scales. Then we systematically calculate their FPCs and measure the similarity between these FPCs with the original one by PCCs. Results are shown in Fig 5. It can be seen that PCCs changes very slowly when scale parameter is smaller than 100, which is 10 Mb in genomic distance. State differently, we can get almost the same genomic compartment even when we remove all the Hi-C data within the 10 Mb band region. It should be noticed that almost all the largest Hi-C value, i.e., contact frequencies, are located within this 10 Mb band. These values, however, are irrelevant to the chromosome genomic compartment!

We further test our SeqMM on other GM06990 chromosomes. A very consistent pattern can be observed. Results of chromosomes 1, 5, 9 and 13 are illustrated in Fig 6. It can be seen that the shape decrease of PCCs is usually found at around 100 locus (10 Mb in genomic distance). This indicates a transition between local scales to global scales. Further studies are needed to explain its biological implications.

Topological associated domain

Another very important finding in Hi-C data analysis is the topological associated domain. TADs are megabase-sized local chromatin interaction domains. They have loop structures and are highly stable and conserved across various cell types and species. TAD boundaries are found to be enriched with the protein CTCF, housekeeping genes, transfer RNAs and short interspersed element (SINE) retrotransposons [18, 23, 24, 26]. These components play important roles in establishing and supporting TADs and other architectural structures of the chromosome. Due to the structural and functional importance of TADs, various algorithms have been proposed for the identification of TADs as stated in the introduction part. However, the sequential information is not considered in any of these models.

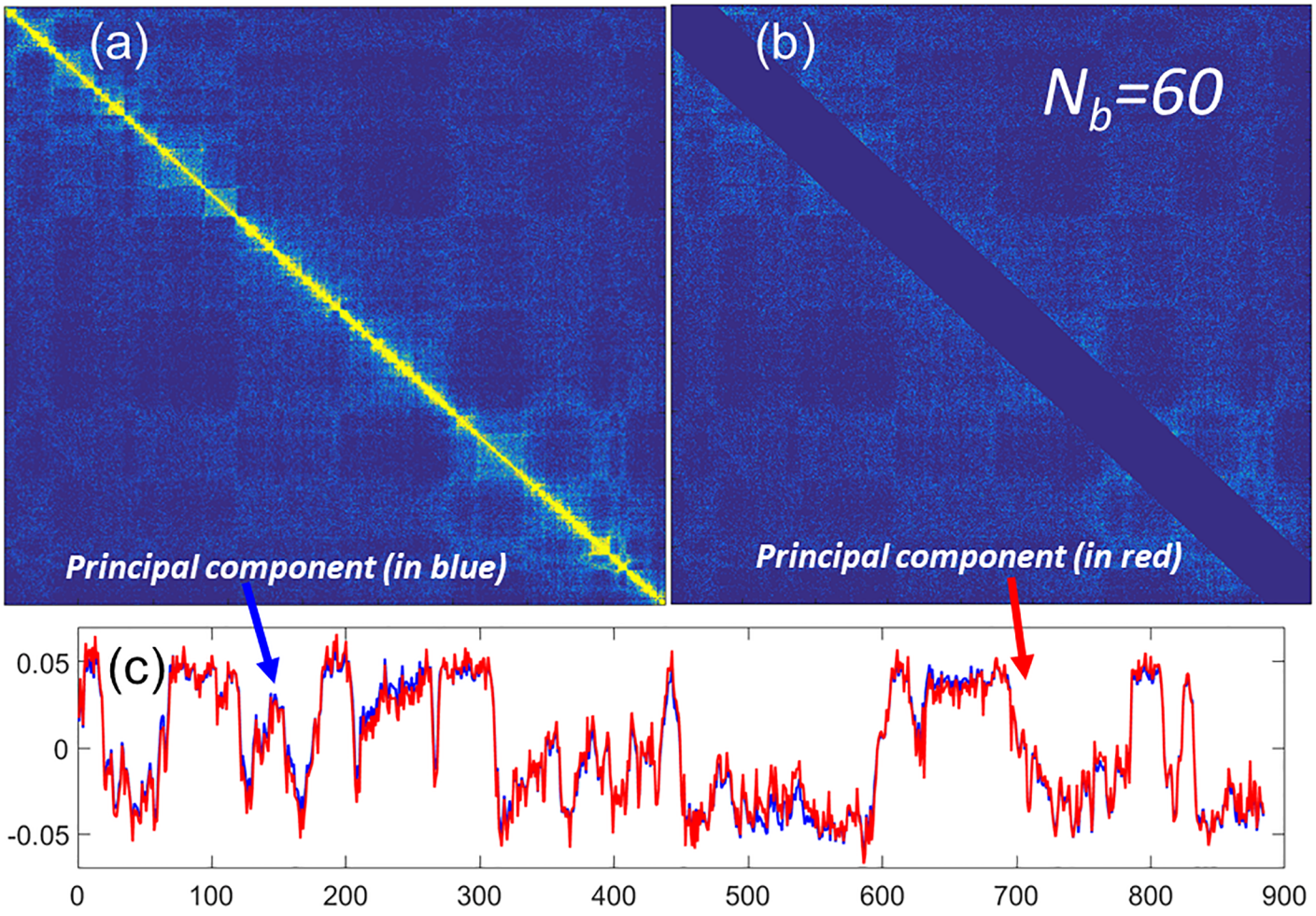

Fig 4. The irrelevance of local interactions in genomic compartment analysis. (**a**) The 100Kb resolution Hi-C data for GM06990 chromosome 14. (**b**) The chromosome 14 Hi-C data with zero contact frequencies in the diagonal band region. The band size, or scale parameter, N_b equal to 60, which amounts to 6 Mb genomic distance. (**c**) The principal components from the two matrices, blue line for (**a**) and red line for (**b**), have almost the same behavior. The Pearson correlation coefficient between these two first principal components is 0.991.

https://doi.org/10.1371/journal.pone.0191899.g004

In this section, Type-2 SeqMM is used to study chromosome TADs. In our Type-2 SeqMM, the clustering is done by using K-means method on eigenvectors from spectral graph model. The basic procedure of the algorithm is illustrated in **Algorithm 1**. To explore the relation between the band size and TAD boundaries from the clustering, we consider a 100Kb resolution Hi-C matrix for chromosome 22 from IMR90 cell line. We systematically change the band size N_b from 20, 80, 140 to 200. The corresponding TAD boundaries are illustrated in Fig 7. It can be seen that the TAD regions evaluated from different scales are not exactly the same and have some variations. Particularly when the band size N_b change from 20 to 80, the calculated TAD regions are quite different. Further, when the band size is larger than 80, although the TAD boundaries are still not the same, they share more and more common values.

To have a more quantitative understanding of this, we systematically change the scale parameter N_b from 2 to 351 (the size of the normalized Hi-C matrix) and calculate the TAD boundaries. Results are shown in Fig 8. We can find that when the value of scale parameter N_b is small, a tiny change of its value can result in huge variations of the predicted TAD boundaries. However, when the scale parameter is larger than a certain value, the fluctuations in the

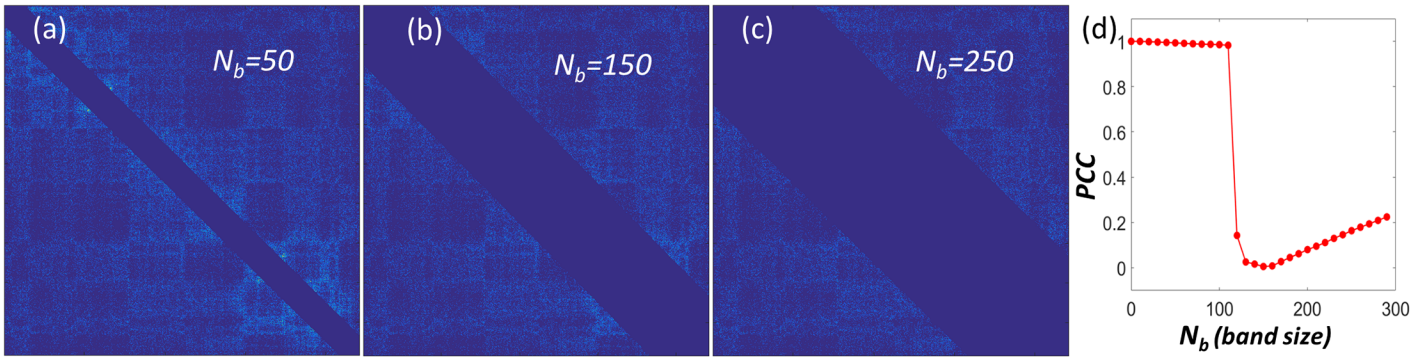

Fig 5. The Type-1 SeqMM for genomic compartment analysis of GM06990 chromosome 14. It is found that the local interactions within a band size around 10 Mb genomic distance contribute very little to genomic compartments. (**a**)-(**c**) The illustration of the Hi-C matrices in Type-1 SeqMM. The sizes of the diagonal band region removed from the Hi-C data go from N_b = 50, N_b = 150 to N_b = 250. (**d**) The PCCs between the first principal components from the original Hi-C matrix and the Hi-C matrices from our Type-1 SeqMM. A high PCC value is observed when the band size is smaller than 10 Mb, meaning the removal of data in this band region has almost no significant influence in the genomic compartments.

https://doi.org/10.1371/journal.pone.0191899.g005

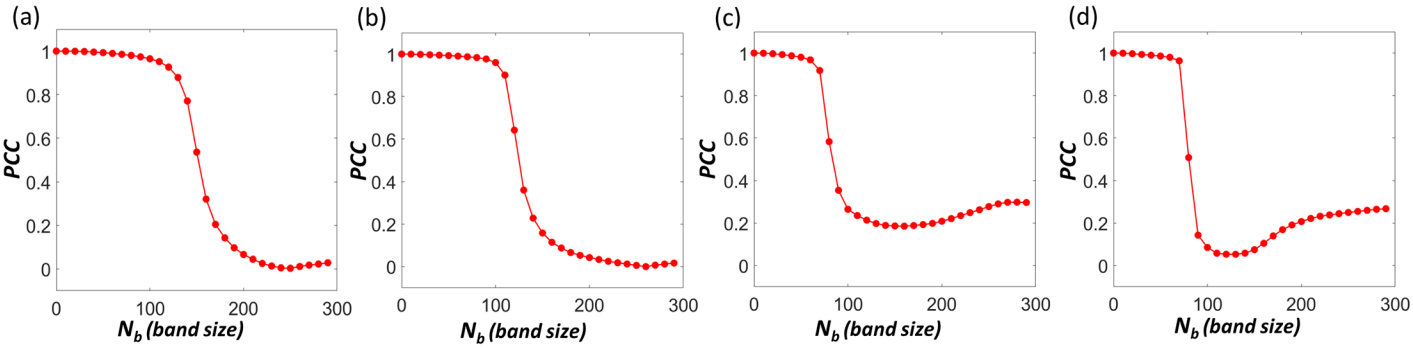

Fig 6. The Type-1 SeqMM for genomic compartment analysis of GM06990 chromosomes 1, 5, 9 and 13. (**a**)-(**d**) The PCCs between the first principal components from the original Hi-C matrix and the Hi-C matrices from our Type-1 SeqMM. Form (**a**) to (**d**), the PCC results are for GM06990 chromosomes 1, 5, 9 and 13, respectively. A consistent behavior has been observed. High PCCs are obtained when band sizes are smaller than around 10 Mb genomic distance. The results confirm our finding that local interactions within a special band region have very little contribution to genomic compartments.

https://doi.org/10.1371/journal.pone.0191899.g006

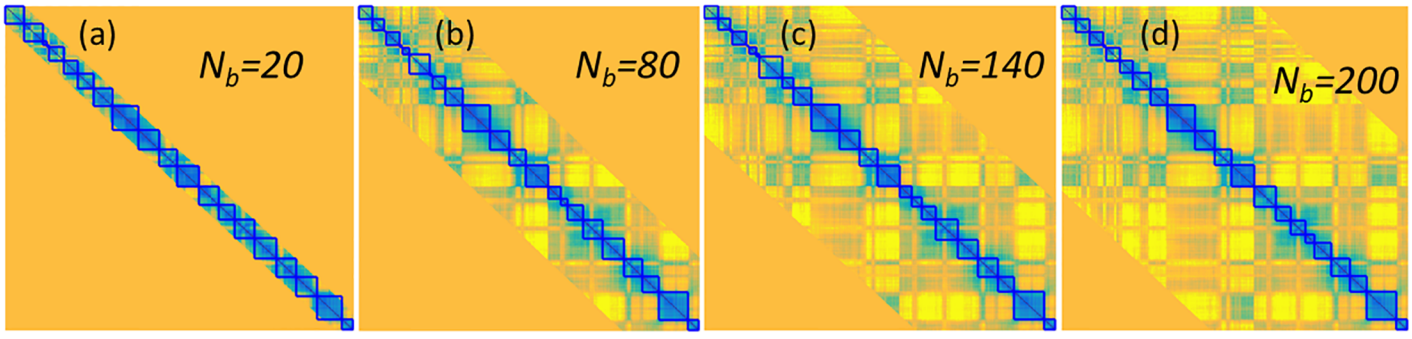

Fig 7. The illustration of TADs calculated for our Type-2 SeqMM. The 100Kb resolution Hi-C matrix for chromosome 22 from IMR90 cell line is considered. All predicted TAD boundaries are marked by blue lines. From (**a**) to (**d**), the band sizes are N_b = 20, 80, 140 and 200, respectively. The predicted TAD boundaries are relatively consistent when N_b is larger than 80. To facilitate a better visualization, the matrices values are correlation coefficients of the normalized Hi-C matrix [17].

https://doi.org/10.1371/journal.pone.0191899.g007

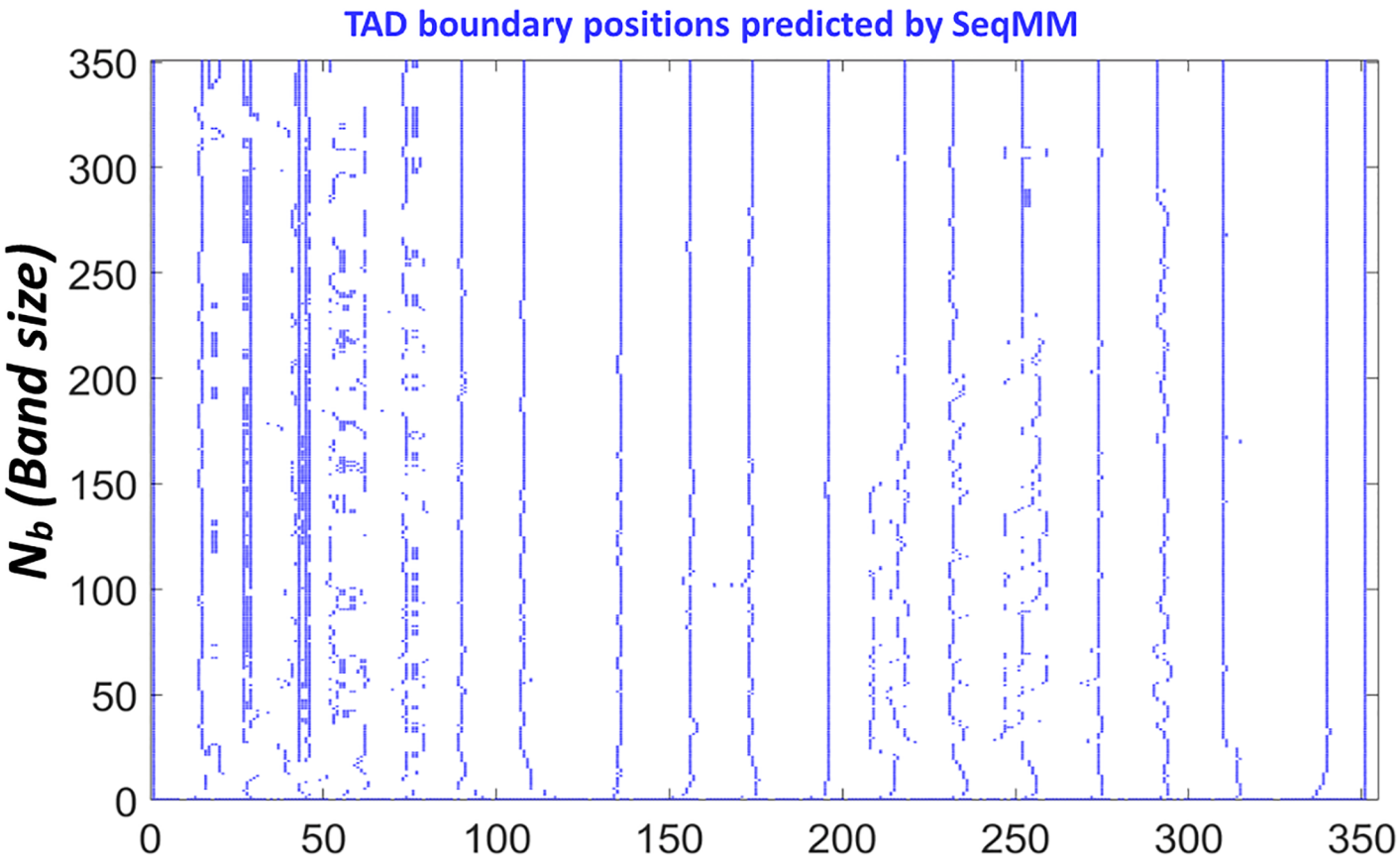

Fig 8. The illustration of TAD boundaries calculated by our Type-2 SeqMM. The 100Kb resolution Hi-C matrix for chromosome 22 from IMR90 cell line is considered. All predicted TAD boundaries are marked by the blue color. It can be seen that, the predicted TAD boundaries are relatively consistent.

https://doi.org/10.1371/journal.pone.0191899.g008

predicted TAD boundaries are greatly reduced. The threshold value is roughly about 20, which is 2 Mb in genomic distance.

We further apply the spectral approach used in Chen's method [36] on the multiscale Laplacian matrices in Eq (3). Results are shown in Fig 9. It can be seen that the variation of the predicted TAD boundaries by his method is much larger than that of our Type-2 SeqMM. More interestingly, the amplitude of variation below the threshold (2 Mb) is much larger than the one after the threshold, which is the same as in our model. Biologically, the threshold value should be highly related to the TAD properties. This is because when the band sizes N_b of our multiscale matrices are smaller than the size of TADs, local interactions within TADs are removed from our models, resulting in a much larger variation in predicted TAD boundaries. However, when the band size is larger than 2Mb, all TAD-related local interactions will be considered, thus a much consistent TAD boundaries should be expected. Stated differently, since TADs are mainly determined by local interactions within the 2Mb band region, the calculated TAD boundaries should always be the same for multiscale matrices with N_b larger than 2Mb. In this sense, our Type-2 SeqMM is much more robust and reliable than Chen's method [36] as a much smaller variation is observed in our model when N_b is larger than 2Mb. Mathematically, in Chen's spectral method, the global scale clustering is iteratively used to subdivide contact matrix or matrix region into two subregions until the algebraic connectivities within the submatrices are all smaller than certain threshold. Therefore, this method optimizes only spatial distances between different loci.

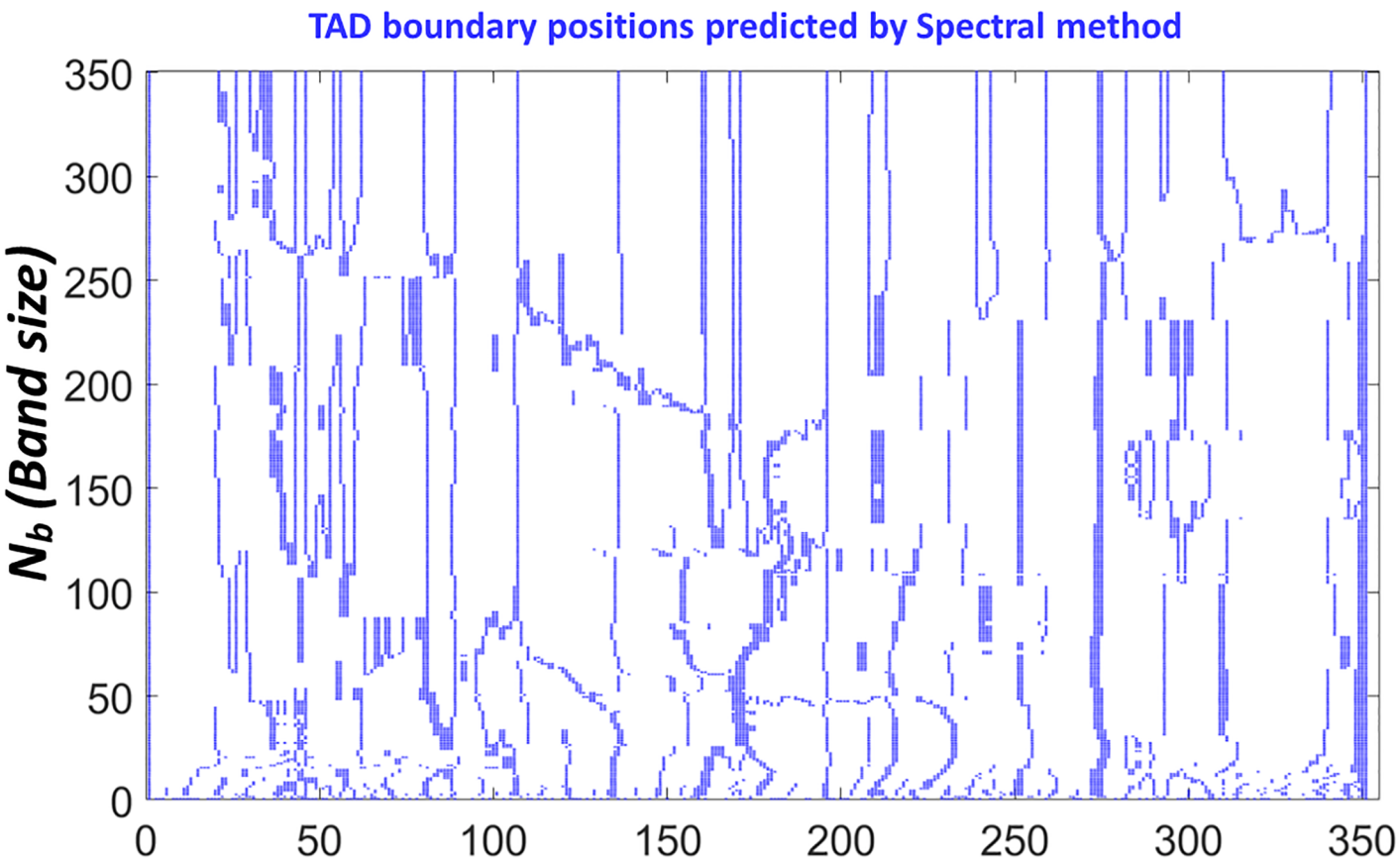

Fig 9. The illustration of TAD boundaries calculated by Chen's spectral method [36] for our Type-2 sequence-based Hi-C matrix models. The 100Kb resolution Hi-C matrix for chromosome 22 from IMR90 cell line is considered. All predicted TAD boundaries are marked by the blue color. It can be seen that, the predicted TAD boundaries have a much larger variation compared with our results in Fig 8.

https://doi.org/10.1371/journal.pone.0191899.g009

Further, even with the difference between the two models, both methods capture the local minimum of a TAD summation function. We consider the 100Kb resolution Hi-C matrix for chromosome 22 from IMR90 cell line. The band size N_b is chosen as 60, which is amount to 6 Mb in genomic sequence. We summarize the contact matric values along the direction that is perpendicular to the matrix diagonal. Results are shown as the black lines in Fig 10(**a**) and 10(**b**). The TAD boundaries from Chen's method and our Type-2 SeqMM are illustrated by blue and red lines. It can be seen that nearly all of these lines are located at the local minima of the summation function. More interestingly, the two methods share many common TAD boundaries. This indicates that the situation illustrated in Fig 3 does not widely exist. This can also be confirmed from the behavior of off-diagonal values. Usually, the off-diagonal values decrease very quickly outside the TAD regions, meaning the distance between a locus from a TAD and a locus outside this TAD is usually very large.

Conclusion

In this paper, we discuss a sequence-based multiscale clustering model for biomolecular data analysis. Biomolecules and their complexes are hierarchical structures made from one or several polymer chains. With the sequential information embedded in these polymer chains, biomolecular data are fundamentally different from the general point cloud data. Traditional clustering methods derived from point cloud data, fall short when sequential information

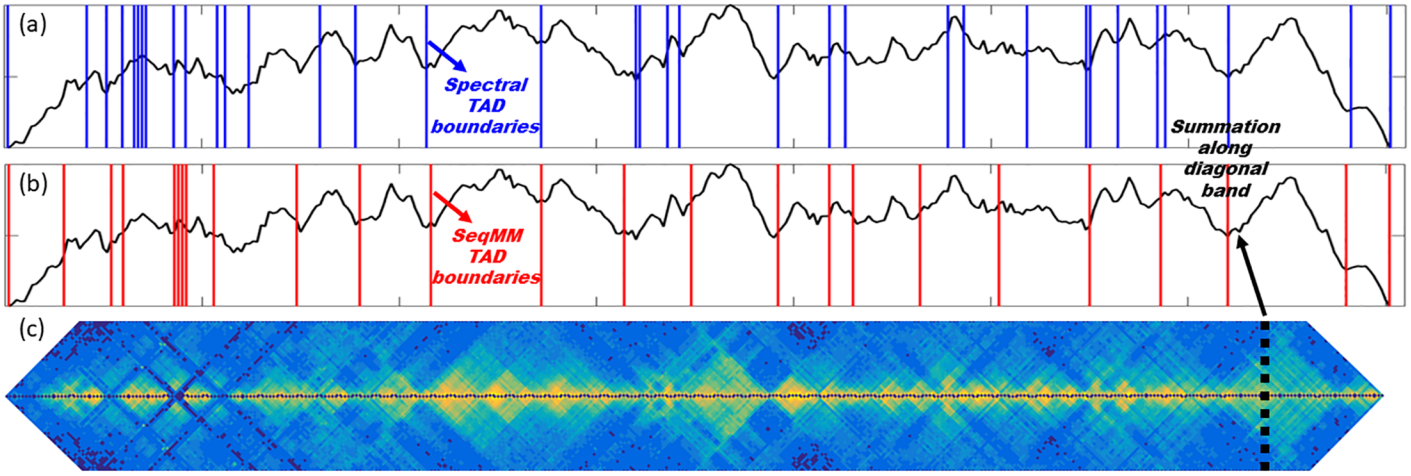

Fig 10. The comparison of the predicted TAD boundaries from Chen's spectral method [36] (blue lines) and our Type-2 SeqMM (red lines). Even though the predicted TAD boundaries from two methods vary a lot, almost all of them are local minimal values of a TAD summation function. **(a)** The blue lines are TAD boundaries calculated from Chen's method [36]. **(b)** The red lines are TAD boundaries calculated from our Type-2 SeqMM with band size 6Mb. **(c)** The diagonal band region from the normalized Hi-C matrix. Again the band size is 6 Mb. We summarize the Hi-C values along the direction that is perpendicular to the matrix diagonal, as indicated by the black dash line. The summation results are represented by the black lines in both **(a)** and **(b)**.

https://doi.org/10.1371/journal.pone.0191899.g010

matters. To overcome this problem, we propose a sequence-based multiscale model for biomolecular structure analysis. We generate a series of structural matrices by gradually and systematically removing the short-range or long-range interactions. These new matrices focus on different sequential scales and their clustering has different biological interpretations. Two SeqMMs have been applied to Hi-C data analysis. We find that the genomic compartments only relate to the global scale information. The removal of a diagonal band region as large as 10 Mb has very little influence to the finally compartment results. Further, we study TADs with our local scale models. We find that when sequence scale is small, a tiny variation of its value will result in great changes in TAD boundaries. However, when the scale value is larger than a threshold value, the TAD boundaries become very consistent. This threshold value is highly related to the sizes of TADs. Interestingly, our method is much more robust than a previous spectral clustering method in the TAD analysis.

Supporting information

S1 File.
(ZIP)

Acknowledgments

The author would like to thank Amartya Sanyal for valuable discussions. This work was supported in part by Nanyang Technological University Startup Grant M4081842.110 and Singapore Ministry of Education Academic Research fund Tier 1 M401110000.

Author Contributions

Conceptualization: Kelin Xia.

Data curation: Kelin Xia.

Formal analysis: Kelin Xia.

Funding acquisition: Kelin Xia.

Investigation: Kelin Xia.

Methodology: Kelin Xia.

Project administration: Kelin Xia.

Resources: Kelin Xia.

Software: Kelin Xia.

Supervision: Kelin Xia.

Validation: Kelin Xia.

Visualization: Kelin Xia.

Writing – original draft: Kelin Xia.

Writing – review & editing: Kelin Xia.

References

1. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS biology. 2005; 3(5): e157.
2. Hou CH, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Molecular cell. 2012; 48(3):471–484. https://doi.org/10.1016/j.molcel.2012.08.031 PMID: 23041285
3. Duan ZJ, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, et al. A three-dimensional model of the yeast genome. Nature. 2010; 465(7296):363–367. https://doi.org/10.1038/nature08973 PMID: 20436457
4. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012; 148(3):458–472. https://doi.org/10.1016/j.cell.2012.01.010 PMID: 22265598
5. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic acids research. 2010; 38(22):8164–8177. https://doi.org/10.1093/nar/gkq955 PMID: 21030438
6. Zhang YB, Wong CH, Birnbaum RY, Li GL, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013; 504(7479):306–310. https://doi.org/10.1038/nature12716 PMID: 24213634
7. Sanyal A, Baù D, Martí-Renom MA, Dekker J. Chromatin globules: a common motif of higher order chromosome structure? Current opinion in cell biology. 2011; 23(3):325–331. https://doi.org/10.1016/j.ceb.2011.03.009 PMID: 21489772
8. Cavalli G, Misteli T. Functional implications of genome topology. Nature structural & molecular biology. 2013; 20(3):290–299. https://doi.org/10.1038/nsmb.2474
9. Chen HM, Chen J, Muir LA, Ronquist S, Meixner W, Ljungman M, et al. Functional organization of the human 4D Nucleome. Proceedings of the National Academy of Sciences. 2015; 112(26):8002–8007. https://doi.org/10.1073/pnas.1505822112
10. Le Dily F, Baù D, Pohl A, Vicent GP, Serra F, Soronellas D, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes & development. 2014; 28(19):2151–2162. https://doi.org/10.1101/gad.241422.114
11. Pope BD, Ryba T, Dileep V, Yue F, Wu WS, Denas O, et al. Topologically associating domains are stable units of replication-timing regulation. Nature. 2014; 515(7527):402–405. https://doi.org/10.1038/nature13986 PMID: 25409831
12. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. science. 2002; 295 (5558):1306–1311. https://doi.org/10.1126/science.1067799 PMID: 11847345
13. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. Genes & development. 2012; 26(1):11–24. https://doi.org/10.1101/gad.179804.111

14. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, De Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature genetics. 2006; 38(11):1348–1354. https://doi.org/10.1038/ng1896 PMID: 17033623

15. Zhao ZH, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. Nature genetics. 2006; 38(11):1341–1347. https://doi.org/10.1038/ng1891 PMID: 17033624

16. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research. 2006; 16(10):1299–1309. https://doi.org/10.1101/gr.5571506 PMID: 16954542

17. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science. 2009; 326(5950):289–293. https://doi.org/10.1126/science.1181369 PMID: 19815776

18. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485(7398):376–380. https://doi.org/10.1038/nature11082 PMID: 22495300

19. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012; 485(7398):381–385. https://doi.org/10.1038/nature11049 PMID: 22495304

20. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome research. 2014; 24 (11):1854–1868. https://doi.org/10.1101/gr.175034.114 PMID: 25122612

21. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015; 518(7539):331–336. https://doi.org/10.1038/nature14222 PMID: 25693564

22. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome research. 2015; 25(4):582–597. https://doi.org/10.1101/gr.185272.114 PMID: 25752748

23. Bonev B, Cavalli G. Organization and function of the 3D genome. Nature Reviews Genetics. 2016; 17 (11):661–678. https://doi.org/10.1038/nrg.2016.112 PMID: 27739532

24. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nature Reviews Molecular Cell Biology. 2016; 17:743–755. https://doi.org/10.1038/nrm.2016.104 PMID: 27580841

25. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013; 502(7469):59–64. https://doi.org/10.1038/nature12593 PMID: 24067610

26. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159(7):1665–1680. https://doi.org/10.1016/j.cell.2014.11.021 PMID: 25497547

27. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms for Molecular Biology. 2014; 9(1):14. https://doi.org/10.1186/1748-7188-9-14 PMID: 24868242

28. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014; 30(17):i386–i392. https://doi.org/10.1093/bioinformatics/btu443 PMID: 25161224

29. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, et al. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nature structural & molecular biology. 2011; 18(1):107–114. https://doi.org/10.1038/nsmb.1936

30. Hu M, Deng K, Qin ZH, Dixon J, Selvaraj S, Fang J, et al. Bayesian inference of spatial organizations of chromosomes. PLoS Comput Biol. 2013; 9(1):e1002893. https://doi.org/10.1371/journal.pcbi.1002893 PMID: 23382666

31. Zhang ZZ, Li GL, Toh KC, Sung WK. 3D chromosome modeling with semi-definite programming and Hi-C data. Journal of computational biology. 2013; 20(11):831–846. https://doi.org/10.1089/cmb.2013.0076 PMID: 24195706

32. Segal MR, Xiong H, Capurso D, Vazquez M, Arsuaga J. Reproducibility of 3D chromatin configuration reconstructions. Biostatistics. 2014; 15(3):442–456. https://doi.org/10.1093/biostatistics/kxu003 PMID: 24519450

33. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J. 3D genome reconstruction from chromosomal contacts. Nature methods. 2014; 11(11):1141–1143. https://doi.org/10.1038/nmeth.3104 PMID: 25240436

34. Zhang B, Wolynes PG. Topology, structures, and energy landscapes of human chromosomes. Proceedings of the National Academy of Sciences. 2015; 112(19):6062–6067. https://doi.org/10.1073/pnas.1506257112

35. Imakaev MV, Fudenberg G, Mirny LA. Modeling chromosomes: Beyond pretty pictures. FEBS letters. 2015; 589(20PartA):3031–3036. https://doi.org/10.1016/j.febslet.2015.09.004 PMID: 26364723

36. Chen J, Hero AO, Rajapakse I. Spectral identification of topological domains. Bioinformatics. 2016; p. 1–7.

37. Xia KL, Wei GW. A review of geometric, topological and graph theory apparatuses for the modeling and analysis of biomolecular data. arXiv preprint arXiv:161201735. 2016;.

38. Opron K, Xia KL, Wei GW. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. Journal of Chemical Physics. 2014; 140:234105. https://doi.org/10.1063/1.4882258 PMID: 24952521

39. Von Luxburg U. A tutorial on spectral clustering. Statistics and computing. 2007; 17(4):395–416. https://doi.org/10.1007/s11222-007-9033-z

40. Xia KL, Feng X, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules I: Cartesian representation. Journal of Computational Physics. 2014; 275:912–936. https://doi.org/10.1016/j.jcp.2013.09.034

41. Feng X, Xia K, Tong Y, Wei GW. Geometric modeling of subcellular structures, organelles and large multiprotein complexes. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:1198–1223. https://doi.org/10.1002/cnm.2532 PMID: 23212797

42. Xia KL, Wei GW. Persistent homology analysis of protein structure, flexibility and folding. International Journal for Numerical Methods in Biomedical Engineerings. 2014; 30:814–844. https://doi.org/10.1002/cnm.2655

43. Xia KL, Wei GW. Multidimensional persistence in biomolecular data. Journal Computational Chemistry. 2015; 36:1502–1520. https://doi.org/10.1002/jcc.23953

44. Xia KL, Opron K, Wei GW. Multiscale multiphysics and multidomain models— Flexibility and Rigidity. Journal of Chemical Physics. 2013; 139:194109. https://doi.org/10.1063/1.4830404 PMID: 24320318

45. Xia KL, Opron K, Wei GW. Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM). The Journal of chemical physics. 2015; 143(20):204106. https://doi.org/10.1063/1.4936132 PMID: 26627949

46. Xia KL, Zhao ZX, Wei GW. Multiresolution topological simplification. Journal Computational Biology. 2015; 22:1–5. https://doi.org/10.1089/cmb.2015.0104

47. Cang ZX, Wei GW. Element specific persistent homology for the analysis and prediction of protein folding stability upon mutation. Bioinformatics. 2017; 33:3549–3557.

48. Cang ZX, Wei GW. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLOS Computational Biology. 2017; 13(7):e1005690. https://doi.org/10.1371/journal.pcbi.1005690 PMID: 28749969

49. Cang ZX, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. International Journal for Numerical Methods in Biomedical Engineering. 2017; https://doi.org/10.1002/cnm.2914 PMID: 28677268

50. Chung F. Spectral graph theory. American Mathematical Society; 1997.

51. Lu MY, Ma JP. The role of shape in determining molecular motions. Biophysical journal. 2005; 89(4):2395–2401. https://doi.org/10.1529/biophysj.105.065904 PMID: 16055547

52. Ming DM, Kong YF, Lambert MA, Huang Z, Ma JP. How to describe protein motion without amino acid sequence and atomic coordinates. Proceedings of the National Academy of Sciences. 2002; 99(13):8620–8625. https://doi.org/10.1073/pnas.082148899

53. Tama F, Brooks CL III. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. Annu Rev Biophys Biomol Struct. 2006; 35:115–133. https://doi.org/10.1146/annurev.biophys.35.040405.102010 PMID: 16689630

54. Xia KL. Multiscale virtual particle based elastic network model (MVP-ENM) for biomolecular normal mode analysis. Physical Chemistry Chemical Physics, 2017; 20(1):658–669 https://doi.org/10.1039/C7CP07177A PMID: 29227479