bioRxiv preprint doi: https://doi.org/10.1101/374132; this version posted July 22, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Flexibility and rigidity index for chromosome packing, flexibility and dynamics analysis

Jiajie Peng [1,*], Jinjin Yang [1] and Kelin Xia [2,3*]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371.
[3] School of Biological Sciences, Nanyang Technological University, Singapore 637371

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The packing of genomic DNA from double string into highly-order hierarchial assemblies has great impact on chromosome flexibility, dynamics and functions. The open and accessible regions of chromosome are the primary binding positions for regulatory elements and are crucial to nuclear processes and biological functions.

Results: Motivated by the success of flexibility-rigidity index (FRI) in biomolecular flexibility analysis and drug design, we propose a FRI based model for quantitatively characterizing the chromosome flexibility. Based on the Hi-C data, a flexibility index for each locus can be evaluated. Physically, the flexibility is tightly related to the packing density. Highly compacted regions are usually more rigid, while loosely packed regions are more flexible. Indeed, a strong correlation is found between our flexibility index and DNase and ATAC values, which are measurements for chromosome accessibility. Recently, Gaussian network model (GNM) is applied to analyze the chromosome accessibility and a mobility profile has been proposed to characterize the chromosome flexibility. Compared with GNM, our FRI is slightly more accurate (1% to 2% increase) and significantly more efficient in both computational time and costs. For a 5kb resolution Hi-C data, the flexibility evaluation process only takes FRI a few minutes on a single-core processor. In contrast, GNM requires 1.5 hours on 10 CPUs. Moreover, interchromosome information can be easily incorporated into the flexibility evaluation, thus further enhance the accuracy of our FRI. In contrast, the consideration of interchromosome information into GNM will significantly increase the size of its Laplacian matrix, thus computationally extremely challenging for the current GNM.

Availability: The software is available at https://github.com/jiajiepeng/FRI_chrFle.

Contact: xiakelin@ntu.edu.sg; jiajiepeng@nwpu.edu.cn

1 Introduction

The packing of chromosome into complicated three-dimensional hierarchial structure has a profound effect on gene expression and other biological functions (Schmitt *et al.*, 2016a). For instance, chromatin loop is formed when cis-regulatory element, such as enhancers, are folded into close spatial proximity with its target promoter. This long-range chromatin contacts are vital to the regulation of gene expression. Recently, the 4D nucleome project is proposed to reveal the packing and dynamics of chromosome and gain insight on the mechanism of gene regulation (Dekker *et al.*, 2017). A major driving force for the project is the advancement of genome-wide C-techniques and C-data for multiple species and tissues (Dekker *et al.*, 2002; Simonis *et al.*, 2006; Zhao *et al.*, 2006; Fullwood *et al.*, 2009; de Wit and de Laat, 2012; Lieberman-Aiden *et al.*, 2009; Dixon *et al.*, 2012; Nora *et al.*, 2012; Jin *et al.*, 2013; Bonev and Cavalli, 2016; Schmitt *et al.*, 2016b; Nagano *et al.*, 2013). With the structure information obtained from C-techniques, researchers begin to understand more about

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

bioRxiv preprint doi: https://doi.org/10.1101/374132; this version posted July 22, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

packing and organization principles of chromosomes. In general, the structure of mammalian chromosomes can be explored from several different scales (Bonev and Cavalli, 2016), including nucleosome, chromatin fiber, chromatin loops (Rao *et al.*, 2014a), topological associated domain (TAD) (Dixon *et al.*, 2012; Nora *et al.*, 2012), genomic compartment (Lieberman-Aiden *et al.*, 2009), chromosome territory (Lieberman-Aiden *et al.*, 2009), etc. A nucleosome is a basic building block for chromatin organization. They interact with each other to form the 30 nm chromatin fibres with solenoid or zigzag shape. These chromatin fibres aggregate and form chromatin loops, in which cis-regulatory element, such as enhancers, are folded into close spatial proximity with its target promoter. From Hi-C data analysis, larger-scale structures, i.e., TAD and genomic compartment, have been defined. TADs, which are about 200 kilobases(Kb) to 2 megabases(Mb), are chromosome components that are highly consistent between different cell types and species. Genomic compartment, which is classified into type A and type B, represents chromosome regions that either densely or sparsely packed. The packing of chromosome into its hierarchial structure is greatly facilitated by various insulator proteins, cohesin complex, mediator, border elements, loop-extruding complexes, other DNA-binding proteins, as well as RNAs. Algorithms and models that based on C-data and ENCODE data are proposed (Lieberman-Aiden *et al.*, 2009; Dixon *et al.*, 2012; Filippova *et al.*, 2014; Lévy-Leduc *et al.*, 2014; Baù *et al.*, 2011; Hu *et al.*, 2013; Zhang *et al.*, 2013; Segal *et al.*, 2014; Lesne *et al.*, 2014; Zhang and Wolynes, 2015; Imakaev *et al.*, 2015). However, even with these progresses, there is a lacking of physical models that quantitatively analyzes the chromosome packing, flexibility and dynamic properties.

In structure biology, it is well-known that biomolecular flexibility, dynamics and functions are tightly related to their structures. For instance, in an intrinsically disordered protein, its well-organized regions are usually very rigid and highly stable. In contrast, its disorder parts, such as hanging chains and extruding loops, are normally very flexible and easy to interact with others. In fact, flexible regions in biomolecules are always more dynamic and tend to have interactions with ligands or other biomolecules. Experimentally, flexility can be quantitatively measured in terms of Debye-Waller factor (or B-factor). Various models have been proposed to reveal the deep connection between biomolecular structure and flexibility (Flores *et al.*, 2007; Emekli *et al.*, 2008; Keating *et al.*, 2009; Shatsky *et al.*, 2004; Flores and Gerstein, 2007; Tama *et al.*, 2000; Halle, 2002; Kundu *et al.*, 2002; Kondrashov *et al.*, 2007; Song and Jernigan, 2007; Hinsen, 2008; Park *et al.*, 2013; Demerdash and Mitchell, 2012; Zhang and Brüschweiler, 2002; Lin *et al.*, 2008; Huang *et al.*, 2008; Li and Brüschweiler, 2009). Among these methods, flexibility-rigidity index (FRI) (Xia *et al.*, 2013; Opron *et al.*, 2014) is one of the most accurate and efficient model in B-factor prediction. In FRI model, a biomolecular structure is viewed as an equilibrium state in which all the interactions from the surrounding environment and within the molecule are well balanced. The unique position of each atom in the structure is the outcome from the "fight" with all the other atoms. Therefore, instead of resorting to the complicated protein interaction Hamiltonian as in other methods, FRI measures the biomolecular flexibility by its topological connectivity or packing density (Halle, 2002).

Compared with the classic models, such as Gaussian network model (GNM) and anisotropic network model (ANM), FRI has significantly increased the accuracy and dramatically reduced the computational time. In both GNM and ANM, a large matrix, either Laplacian matrix or Hessian matrix, is constructed. Their B-factor prediction formula requires the calculation of all the eigenvalues and eigenvectors of the matrix, which can be time-consuming not to mention about the memory cost. Free from eigenvalue decomposition, FRI only has the computational complexity of $\mathcal{O}(N^2)$ with N the total number of atoms. Fast FRI (fFRI) (Opron *et al.*, 2014) can further reduce the computational complexity to $\mathcal{O}(N)$ with almost no scarifying the accuracy of the model. It only takes fFRI 30 seconds on a single-core processor to calculate the flexibility of an HIV virus capsid with 313 236 residues (Opron *et al.*, 2014). Multiscale FRI (Opron *et al.*, 2015a), Generalized FRI (Nguyen *et al.*, 2016) and multiscale weighted colored graph (MWCG) based FRI (Bramer and Wei, 2018) can further increase the accuracy of FRI models. Especially, MWCG based FRI has set a new accuracy benchmark for protein flexibility analysis. More interestingly, FRI model has been applied to protein-ligand binding affinity prediction in drug design (Nguyen *et al.*, 2017). The results from the FRI based machine learning model has significantly outperformed all traditional models. This reveals that the rigidity strengthening can be a potential mechanism for protein-ligand binding (Nguyen *et al.*, 2017). More recently, a virtual particle based FRI model is proposed for analyzing the dynamics of extremely large-sized biomolecular complexes and organelles, especially the ones from Electron Microscopy Data Bank (EMDB) (Xia and Wei, 2016; Xia *et al.*, 2018). The success of the FRI model in these subcellular structures have motivated us to propose the FRI based model for chromosome packing, flexibility and dynamics analysis.

Recently, Gaussian network model has been used to quantitatively characterize the chromatin accessibility. It is known that packing of DNA into a compressed form causes accessibility problems for transcription and DNA replication. In general, the chromatin packing density is negatively correlated with transcriptional activity. For tightly packed chromatin regions, their DNA is less accessible to transcriptional machinery. For loosely packed chromatin regions, the DNA is more accessible and transcription is much easier. In this way, the chromatin accessibility can be characterized by nuclease hypersensitivity, which is directly measured by MNase, DNase, FAIRE, as well as ATAC. Based on the Hi-C data, a mobility profile is defined from GNM and has been found to be highly correlated with DNase-seq and ATAC-seq values (Sauerwald *et al.*, 2017). Essentially, the mobility profile can be viewed as a characterization of flexibility properties of the chromosome structures.

In the current paper, we propose a FRI based model for chromosome flexibility analysis. Similar to biomolecular flexibility, we assume the rigidity and flexibility of each loci is solely determined by its local packing density. In this way, the rigidity index for each locus can be directly evaluated by summarizing its connectivity with its local neighbours. The connectivity strength between any two loci can be evaluated from its contact frequency from the Hi-C matrix. The flexibility index is inversely related to the rigidity index and can be used to characterize the chromatin accessibility. Our flexibility model is validated by comparing with the chromosome packing information from the chromatin accessibility. A high correlation coefficient is found between our flexibility index, DNase-seq and ATAC-seq values. Compared with GNM (Sauerwald *et al.*, 2017), our model is slightly more accurate with average 1%-2% increase in accuracy and significantly more efficient in both computational time and resources. Moreover, interchromosome interaction can be incorporated into our FRI model. As expected, the inclusion of the interchromosome interactions further enhances the performance of our model. Essentially, our FRI model provides a direct link between the chromosome packing density with chromosome flexibility.

2 Methods

2.1 Flexibility-rigidity index

Basic setting of flexibility-rigidity index The flexibility and rigidity property of a systems is highly related to their inner network structure. We consider a system of N-elements with coordinates $\{\mathbf{r}_j|\mathbf{r}_j \in \mathbb{R}^3, j = 1, 2, \cdots, N\}$. The connectivity between the i-th and j-th element can be characterized by a general correlation kernel $\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij})$ satisfying,

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = 1 \quad \text{as} \quad \|\mathbf{r}_i - \mathbf{r}_j\| \to 0, \tag{1}$$

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = 0 \quad \text{as} \quad \|\mathbf{r}_i - \mathbf{r}_j\| \to \infty. \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/374132; this version posted July 22, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The parameter η_{ij} is a resolution parameter (or scale parameter). And the correlation kernel can be chosen as any real-valued monotonically decreasing radial basis function. The commonly used kernel functions (Xia *et al.*, 2013; Opron *et al.*, 2014, 2015b) include generalized exponential functions

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = e^{-\left(\|\mathbf{r}_i - \mathbf{r}_j\|/\eta_{ij}\right)^{\kappa}}, \quad \kappa > 0, \tag{3}$$

and generalized Lorentz functions,

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = \frac{1}{1 + (\|\mathbf{r}_i - \mathbf{r}_{ij}\|/\eta_j)^{\upsilon}}, \quad \upsilon > 0. \tag{4}$$

For i-th element, the rigidity index μ_i is defined as the summation of its connectivity with all other atoms,

$$\mu_i = \sum_{i \neq j} w_j \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}). \tag{5}$$

Here the weight parameter w_j is used to characterize the atom properties, for example, it can be chosen as the atomic number. The flexility index is inversely related to the rigidity index. For i-th element, the flexility index f_i is defined as,

$$f_i = \frac{1}{\mu_i} = \frac{1}{\sum_{i \neq j} w_j \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij})}. \tag{6}$$

The flexibility index f_i is linearly related to i-th B-factor,

$$B_i^t = af_i + b, \quad \forall i = 1, 2, \cdots, N \tag{7}$$

where $\{B_i^t\}$ is the predicted B-factor. Linear regression model can be used to determine the fitting parameter a and b. In biomolecular flexibility analysis, a coarse-grained C_α representation is usually considered. In this case, we can set the weight parameter $w_i = 1, i = 1, 2, ..., N$ and scale parameter $\eta_{ij} = \eta, i, j = 1, 2, ..., N$ with η a fixed scale value. The fast FRI (fFRI) model is proposed for modeling extremely large biomolecules (Opron *et al.*, 2014). Essentially, a cell list algorithm (Allen and Tildesley, 1987) is employed and the rigidity index in Eq. (5) is calculated between atoms within a certain cut-off distance.

Rigidity function and flexibility function Mathematically, flexibility and rigidity index are discrete values defined on atoms. They can be generalized into continuous representations, i.e., flexibility function and rigidity function (Xia *et al.*, 2013; Opron *et al.*, 2014). More specifically, the rigidity function can be defined as follows,

$$\mu(\mathbf{r}) = \sum_{j=1}^{N} w_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_{ij}). \tag{8}$$

Essentially, a continuous rigidity function can be viewed as a density distribution and equals to Gaussian surface (Liu *et al.*, 2015) when $\kappa = 2$ and w_i is chosen as the atomic number. A continuous flexibility function can be defined in a similar way as the flexibility index,

$$F(\mathbf{r}) = \frac{1}{\sum_{j=1}^{N} w_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_{ij})}. \tag{9}$$

It should be noticed that the flexibility function is well defined only in the region when rigidity index is nonzero.

Multiscale flexibility-rigidity index For a system with a hierarchial structure and multiscale properties, a unique scale parameter value is usually not suitable. Therefore, multiple correlation kernels with different scale values are considered. The multiscale flexibility (Opron *et al.*, 2015a) can be generalized as follows,

$$f_i^n = \frac{1}{\sum_{j=1}^{N} w_j^n \Phi^n(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}^n)}, \tag{10}$$

where w_j^n, $\Phi^n(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}^n)$ and η_{ij}^n are quantities associated with n-th kernel. The linear regression is used to minimize the objective function,

$$\text{Min}_{a^n, b} \left\{ \sum_i \left| \sum_n a^n f_i^n + b - B_i^e \right|^2 \right\}, \tag{11}$$

where $\{B_i^e\}$ are the experimental B-factors. Usually, for each kernel function, a unique scale parameter η is used. The multiscale properties can be well characterized by several different η values.

Multiscale weighted colored graph based FRI As a special case of graph labeling, graph coloring is to assign labels (or "colors") to nodes or edges of a graph under a certain rule or some constraints. In the weighted colored graph model, a protein graph is labelled by its edge types and subgraphs are defined according to these labels (Bramer and Wei, 2018). The atoms in a protein structure are predominately from several atom types, including carbon (C), nitrogen (N), oxygen (O), and sulfur (S). Hydrogen and ion atoms (such as Mn^{2+}, Mg^{2+}, Fe^{2+}, Zn^{2+}, etc) are not considered due to their absence from most PDB files. With this setting, all nodes in a colored protein graph can be labeled by element in an atom set (C, N, O, S) and all edges can be colored by element-specific pairs in a set (CC, CN, CO, CS, NC, NN, NO, NS, OC, ON, OO, OS, SC, SN, SO, SS). It should be noticed that edges in a colored protein graph are directed (Bramer and Wei, 2018). A CN pair is different from a NC pair in the colored graph model. A subgraph contains only the same type of directed pairs. For example, all NC pairs together form a subgraph. Further, these colored graphs are combined with FRI models, particularly the mFRI, to evaluate the protein B-factors. More specifically, if we are interested about flexility for all C atoms, we can consider subgraphs made from CC, CN, CO and CS pairs. For each subgraph, more than one scale values can be used. In this way, a multiscale weighted colored graph based FRI model is constructed.

2.2 FRI based chromosome packing density analysis

2.2.1 Data processing

Hi-C data normalization One of the major challenges for Hi-C data analysis is the systematic bias from the experimental setting that complicates the interpretation of observed contact frequencies (Yaffe and Tanay, 2011; Imakaev *et al.*, 2012; Ay *et al.*, 2014; Witten and Noble, 2012). Bias can be introduced from procedures including crosslinking, chromatin fragmentation, biotin-labelling and religation. Particularly, systematic biases that can substantially affect the Hi-C experimental results come from three major sources (Yaffe and Tanay, 2011), including distance-between restriction sites, the GC content and sequence mappability. Accounting for these biases is the first and most important step in C-data analysis. Various methods have been proposed to remove these biases, including HiCNorm (Hu *et al.*, 2012), Vanilla-Coverage normalization (Rao *et al.*, 2014a; Lieberman-Aiden *et al.*, 2009), iterative correction and eigenvector decomposition (ICE) (Imakaev *et al.*, 2012), Matrix-balancing method (Knight and Ruiz, 2013), etc. In this paper, we use the Vanilla-Coverage normalization.

Distance matrix construction To construct the chromosome three-dimensional structure from the Hi-C data, computational models usually employ a reciprocal function to describe the relation between interaction frequency and spatial distance of two loci (Zhang *et al.*, 2013; Boulos *et al.*,

bioRxiv preprint doi: https://doi.org/10.1101/374132; this version posted July 22, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2013; Wang *et al.*, 2013; Segal *et al.*, 2014; Siahpirani *et al.*, 2016; Filippova *et al.*, 2014; Imakaev *et al.*, 2015; Lesne *et al.*, 2014; Chen *et al.*, 2016; Tjong *et al.*, 2016; Zhu *et al.*, 2018). More specifically, it assumes that the conversion between the contact frequency matrix $U = \{u_{ij}; i, j = 1, 2, ..., N\}$ and the distance matrix $D = \{d_{ij}; i, j = 1, 2, ..., N\}$ follows a power law distribution $d_{ij} = 1/(u_{ij})^{\alpha}$. The coefficient α is a parameter called conversion factor, and d_{ij} and u_{ij} are the distance and contact frequency between loci i and j, respectively. The relation can be expressed as follows,

$$d_{ij} = \begin{cases} \frac{1}{(u_{ij})^{\alpha}}, & u_{ij} > 0; \\ \infty, & u_{ij} = 0. \end{cases} \tag{12}$$

In this paper, we assume the conversion factor $\alpha = 1$. Even though it is unphysical to assume the spatial distance equals to infinity when a contact frequency is zero, this definition works well for the kernel functions in our FRI models.

2.2.2 Algorithm

We assume that the interaction between two loci in a chromosome structure can be characterized by the general correlation kernel as in Eqs. (1) and (2). We assume all loci have similar properties, thus $w_{ij} = 1$ and $\eta_{ij} = \eta$. The value of scale parameter η are linearly related to the locus resolution. Generally, a small η value is used in modeling high resolution data, whereas a large scale value in modeling low resolution ones. As shown in Algorithm 1, the process of chromosome flexibility analysis includes four components: normalizing the Hi-C data; transfering Hi-C contact frequency to relative distance; calculating the chromosome rigidity; calculating the chromosome flexibility.

```
Algorithm 1 Chromosome flexibility analysis with FRI
Input: Hi-C data
Output: Chromosome flexibility of each locus
Pre-processing: Normalize the Hi-C data with the Vanilla-Coverage model;

Step 1: Transfer Hi-C contact frequency to relative distance. We use conversion factor $\alpha = 1$. That is if the contact frequency between $i$-th and $j$-th locus is $u_{ij}$, the relative distance between them is $d_{ij} = \frac{1}{u_{ij}}$;

Step 2: Calculate the chromosome rigidity. The generalized Gaussian kernel is used, $u_i = \sum_j e^{-\left(d_{ij}/\eta(s)\right)^{\kappa}}$. In our simulation, we always set $\kappa = 1$. The scale parameter $\eta(s)$ is linearly related to Hi-C data resolution. More specifically, we set $\eta$ as $5.0 * 10^{-4}$, $2.0 * 10^{-4}$, and $1.0 * 10^{-4}$ for resolution 100kb, 50kb, and 25kb, respectively;

Step 3: Calculate the chromosome flexibility. That is $f_i = \frac{1}{u_i}$.
```

3 Results and discussions

3.1 Data Description

To evaluate the performance of our FRI models, we consider the Hi-C dataset for GM12878 and IMR90 cell line with GEO accession number GSE63525 from Rao's paper (Rao *et al.*, 2014b). GM12878 is a lymphoblastoid cell line produced from the blood. IMR90 is a cell line derived from human foetal lung. We also evaluate the Spearman correlation coefficient (SCC) between chromosome flexibility and chromosome accessibility measurements, including Dnase and ATAC. The DNase-seq data are obtained from ENCODE project. The ATAC-seq data is obtained from GEO database (GEO accessions GSM1155959 for GM12878 and GSM1418975 for IMR90). For both experimental datasets, bed-formatted peak files are used. We bin the data into the same resolution as used in the Hi-C data by adding all peak values within each locus. The binned data were then smoothed using moving average with a window size of 200 kb in the same way as GNM (Sauerwald *et al.*, 2017).

3.2 FRI based chromosome flexibility analysis

As stated in the introduction, there is a strong correlation between chromosome packing density, accessibility and flexibility. To illustrate their relations, we consider two 5kb resolution Hi-C chromosome data from GM12878. Chromosome packing density is measured by chromosome flexibility index from our FRI. Chromatin accessibility is characterized by DNase-seq and ATAC-seq values (Sauerwald *et al.*, 2017). After the normalization of DNase-seq and ATAC-seq data, flexibility index is linearly fitted with their values. The results are demonstrated in Figure 1. For DNase-seq, it can be seen that there is a very good agreement between the predicted values and experimental ones. For ATAC-seq, the results are not as good as DNase-seq models. But a comparably good agreement is still observed. To further evaluate our model, we consider all the 23 chromosomes from GM12878 and IMR90, and calculate the Spearman correlation coefficient (SCC) between experimental results and theoretical predictions. The resolution of the Hi-C data is 25kb. Figures 2 and 3 demonstrate the SCCs for cell lines GM12878 and IMR90, respectively. The results for DNase-seq and ATAC-seq are listed in subfigures (a) and (b). In DNase-seq models, the average SCCs of all 23 chromosomes for GM12878 and IMR90 are 0.860 and 0.847 respectively. In ATAC-seq models, the average SCCs for GM12878 and IMR90 are 0.633 and 0.674 . To avoid confusion, the parameter values of κ and η in our FRI model are set as 1.0 and $5.0 * 10^{-4}$ for GM12878, 1.0 and $3.0 * 10^{-3}$ for IMR90, respectively.

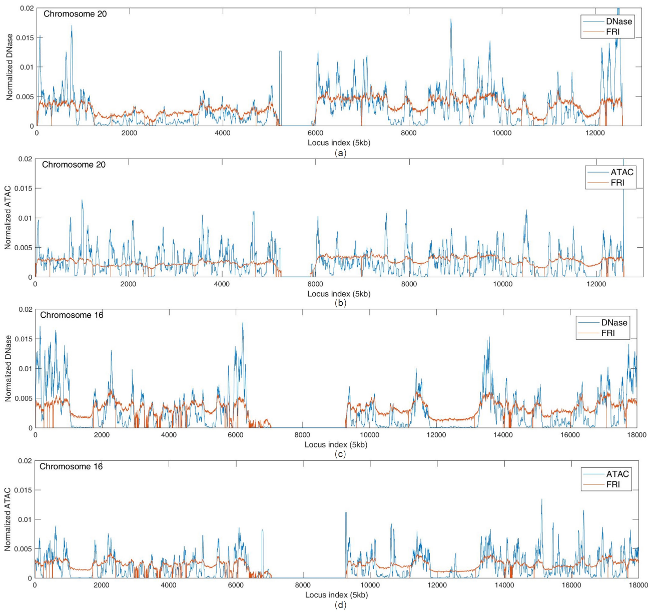

Fig. 1. The illustration of the correlation between chromosome flexibility and chromatin accessibility. Two Hi-C data with resolution 5kb for chromosomes 20 (a, b) and 16 (c, d) of GM12878 are considered. The chromosome flexibility (red line) is calculated from our FRI model. The chromatin accessibility (blue color) is measured by DNase-seq (a, c) and ATAC-seq (b, d).

Currently, GNM is the only method that has been used to predict the chromatin accessibility from Hi-C data (Sauerwald *et al.*, 2017), to the best of our knowledge. In GNM, a mobility profile is generated from each chromosome. It is found that the mobility profile is linearly related to

bioRxiv preprint doi: https://doi.org/10.1101/374132; this version posted July 22, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

chromatin accessibility characterized by DNase-seq and ATAC-seq data. Larger mobility profile values indicate higher accessibility, whereas smaller values are associated with lower accessibility (Sauerwald *et al.*, 2017). Essentially, the mobility profile from GNM is similar to flexibility index in our FRI. To compare the performance of FRI and GNM in chromatin accessibility analysis, we calculate the SCCs for all 23 chromosomes in GM12878 and IMR90 using the GNM codes (Sauerwald *et al.*, 2017). The corresponding results from GNM are listed in Figures 2 and 3. It can be observed that, in nearly all the chromosome cases, SCCs from our FRI are higher than those from GNM in both DNase-seq and ATAC-seq. More specifically, in DNase-seq models, the average GNM (FRI) SCCs of all 23 chromosomes for GM12878 and IMR90 are 0.838 (0.860) and 0.833 (0.847), respectively. In ATAC-seq models, the average GNM (FRI) SCCs for GM12878 and IMR90 are 0.618 (0.633) and 0.667 (0.674). We can see that the average SCCs of FRI method are around 1% to 2% higher than GNM method in both DNase-seq and ATAC-seq.

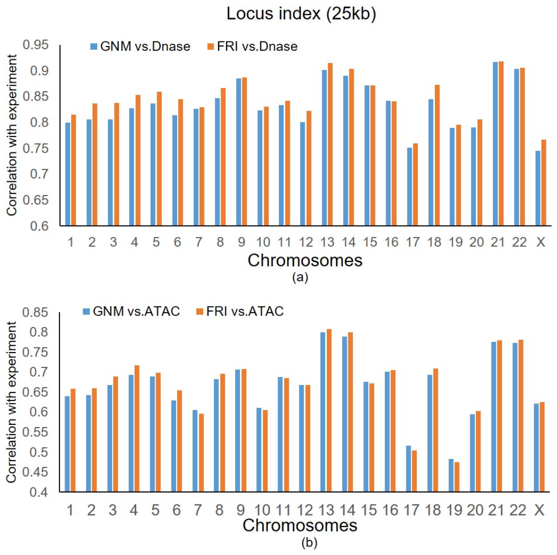

Fig. 2. SCCs between chromosome flexibility and chromatin accessibility for GM12878 cell line. (a) SCCs of FRI (orange bar) and GNM (blue bar) for DNase-seq data. (b) SCCs of FRI (orange bar) and GNM (blue bar) for ATAC-seq data.

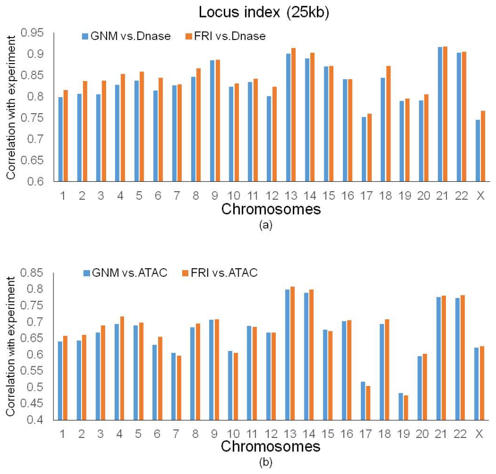

Fig. 3. SCCs between chromosome flexibility and chromatin accessibility for IMR90 cell line. (a) SCCs of FRI (orange bar) and GNM (blue bar) for DNase-seq data. (b) SCCs of FRI (orange bar) and GNM (blue bar) for ATAC-seq data.

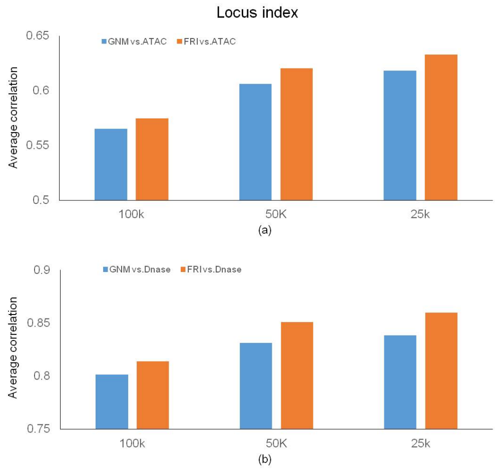

Fig. 4. Average SCCs between chromosome flexibility and chromatin accessibility for all chromosomes from GM12878 cell line. (a) Average SCCs of FRI (orange bar) and GNM (blue bar) for ATAC-seq data. (b) Average SCCs of FRI (orange bar) and GNM (blue bar) for DNase-seq data.

3.3 Robustness of FRI for chromosome flexibility analysis

In the above section, we only consider the Hi-C data with resolution of 25kb. To further test robustness of FRI method, we consider the Hi-C data from the GM12878 cell line in different resolutions, i.e., 50kb and 100kb. Correspondingly, scale parameter η is set as $2.0 * 10^{-4}$ and $1.0 * 10^{-4}$, respectively. The SCCs are calculated for both ATAC-seq and DNase-seq data. The results are illustrated in Figure 4(a) and Figure 4(b). The GNM results are also listed for comparison. Consistent with the above results, our predictions are highly accurate and is constantly better than GNM. Interestingly, the accuracy of both models increases with the resolution.

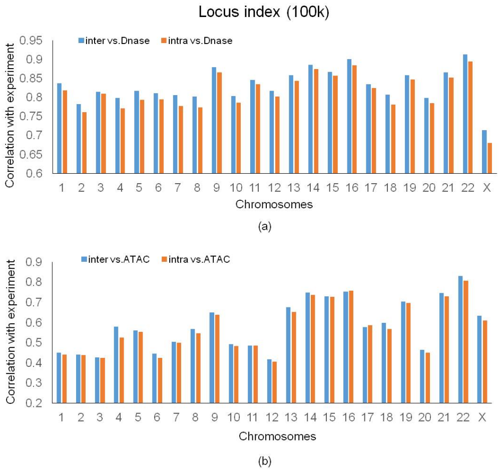

Fig. 5. SCCs of FRI with inter-chromosome interactions for GM12878 cell line. (a) SCCs of FRI (orange bar) and GNM (blue bar) for DNase-seq data. (b) SCCs of FRI (orange bar) and GNM (blue bar) for ATAC-seq data.

3.4 Effect of inter-chromosome interactions on chromosome flexibility

Both intra-chromosome and inter-chromosome interactions have a great impact on the chromosome packing density, thus directly influence chromosome flexibility. However, the consideration of the inter-chromosome information will dramatically increase the computational costs, especially when matrix operations are involved. Therefore, it is prohibitively

bioRxiv preprint doi: https://doi.org/10.1101/374132; this version posted July 22, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expensive for GNM to incorporate the effect of inter-chromosome interactions (Sauerwald *et al.*, 2017). Since FRI method involves only simple algebraic operations, it brings great promise to tackle the challenge from inter-chromosome interactions.

The Hi-C data from GM12878 with resolution 100kb is used. These are the highest resolution data with inter-chromosome interactions that we can obtain. Figure 5 demonstrates the SCCs between chromosome flexibility and chromatin accessibility for DNase-seq (a) and ATAC-seq (b). Two situations are considered for comparison. One is with only intra-chromosome interactions and the other is with both intra- and inter-chromosome interactions. It can be seen clearly that the incorporate of the inter-chromosome information can further increase the accuracy of our chromosome flexibility model. For DNase-seq, the average SCC is 0.832 for both intra and inter case and it is 0.02 higher than the intra model (0.831). For ATAC-seq, the average SCC is 0.587 for both intra and inter case and it is 0.01 higher than the intra model (0.586).

3.5 Algorithm efficiency comparison between FRI and GNM on chromosome flexibility analysis

A significant advantage of FRI over all previous models in flexility analysis is its great efficiency and low computational cost. To compare the algorithm efficiency for FRI and GNM in chromosome flexibility analysis, we measure running times for all 23 chromosomes from GM12878 cell line. Both FRI and GNM are implemented with MATLAB. A linux server with Xeon(R) E5-2690 CPU (2.60GHz) and 512 GB memory is used.

Different resolutions (5kb, 25kb, 50kb, 100kb and 250kb) are considered and results are listed in Figure 6. Here the values represent the total computational time for all 23 chromosomes together. It can be seen clearly that FRI method is much more efficient than GNM method. For all resolutions, the running time of FRI is significantly less than that of GNM. Previously in GNM, it takes about 1.5 hour per chromosome at 5kb resolution using 10 CPUs (Sauerwald *et al.*, 2017). In contrast, it only costs FRI several minutes on a single CPU. These results are consistent with the model complexity (Xia *et al.*, 2013; Opron *et al.*, 2014). Essentially, FRI uses only simply algebraic operations, whereas GNM requires not only matrix operations but also eigenvalue decomposition.

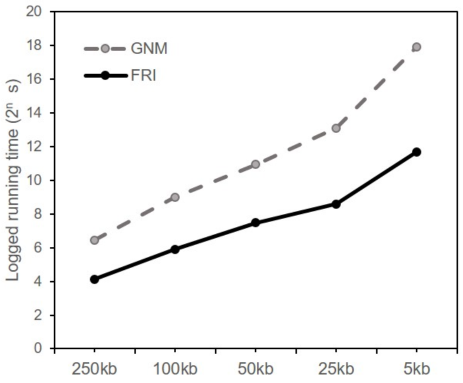

Fig. 6. Running time of GNM and FRI on 23 chromosomes. The Hi-C data in different resolutions are considered. A significant reduce of the computational time in FRI can be clearly observed.

4 Conclusion

In this paper, the flexibility and rigidity index (FRI) model is introduced for the first time to analyze the chromosome packing, flexibility and dynamics. We evaluate the flexibility index for each locus. It is found that the flexibility index can be used to characterize the chromosome flexibility. A high correlation is found between our chromosome flexibility and accessibility measurements, including DNase and ATAC. Compared with the Gaussian network model (GNM), FRI is not only more accurate, but also significantly more efficient in both computational times and costs. Moreover, FRI can incorporate the inter-chromosome information into the flexibility evaluation, thus further enhance the model accuracy.

Acknowledgements

The author Kelin Xia would like to thank Amartya Sanyal for his introduction and discussion of Hi-C experiments and data analysis.

Funding

This work was supported in part by Nanyang Technological University Startup Grant M4081842.110, Singapore Ministry of Education Academic Research fund Tier 1 M401110000, National Natural Science Foundation of China (Grant No. 61702421, 61332014, 61772426).

References

Allen, M. P. and Tildesley, D. J. (1987). *Computer Simulation of Liquids*. Oxford: Clarendon Press.

Ay, F., Bailey, T. L., and Noble, W. S. (2014). Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, **24**(6), 999–1011.

Baù, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., Dekker, J., and Marti-Renom, M. A. (2011). The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, **18**(1), 107–114.

Bonev, B. and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, **17**(11), 661–678.

Boulos, R. E., Arneodo, A., Jensen, P., and Audit, B. (2013). Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Physical review letters*, **111**(11), 118102.

Bramer, D. and Wei, G.-W. (2018). Multiscale weighted colored graphs for protein flexibility and rigidity analysis. *The Journal of chemical physics*, **148**(5), 054103.

Chen, J., Hero, A. O., and Rajapakse, I. (2016). Spectral identification of topological domains. *Bioinformatics*, pages 1–7.

de Wit, E. and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes & development*, **26**(1), 11–24.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, **295**(5558), 1306–1311.

Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'shea, C. C., Park, P. J., Ren, B., *et al.* (2017). The 4D nucleome project. *Nature*, **549**(7671), 219.

Demerdash, O. N. A. and Mitchell, J. C. (2012). Density-cluster NMA: A new protein decomposition technique for coarse-grained normal mode analysis. *Proteins:Structure Function and Bioinformatics*, **80**(7), 1766–1779.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Emekli, U., Dina, S., Wolfson, H., Nussinov, R., and Haliloglu, T. (2008). HingeProt: automated prediction of hinges in protein structures. *Proteins*, **70**(4), 1219–1227.

Filippova, D., Patro, R., Duggal, G., and Kingsford, C. (2014). Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, **9**(1), 14.

Flores, S. and Gerstein, M. (2007). FlexOracle: predicting flexible hinges by identification of stable domains. *BMC bioinformatics*, **8**(1).

Flores, S., Lu, L., Yang, J., Carriero, N., and Gerstein, M. (2007). Hinge atlas: relating protein sequence to sites of structural flexibility. *BMC bioinformatics*, **8**.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., *et al.* (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**(7269), 58–64.

Halle, B. (2002). Flexibility and packing in proteins. *PNAS*, **99**, 1274–1279.

Hinsen, K. (2008). Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, **24**, 521 – 528.

Hu, M., Deng, K., Selvaraj, S., Qin, Z. H., Ren, B., and Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, **28**(23), 3131–3133.

Hu, M., Deng, K., Qin, Z. H., Dixon, J., Selvaraj, S., Fang, J., Ren, B., and Liu, J. S. (2013). Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol*, **9**(1), e1002893.

bioRxiv preprint doi: https://doi.org/10.1101/374132; this version posted July 22, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Huang, S. W., Shih, C. H., Lin, C. P., and Hwang, J. K. (2008). Prediction of nmr order parameters in proteins using weighted protein contact-number model. *Theoretical Chemistry Accounts*, **121**(3-4), 197–200.

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. *Nature methods*, **9**(10), 999–1003.

Imakaev, M. V., Fudenberg, G., and Mirny, L. A. (2015). Modeling chromosomes: Beyond pretty pictures. *FEBS letters*, **589**(20PartA), 3031–3036.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C. A., Schmitt, A. D., Espinoza, C. A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**(7475), 290–294.

Keating, K. S., Flores, S. C., Gerstein, M. B., and Kuhn, L. A. (2009). StoneHinge: hinge prediction by network analysis of individual protein structures. *Protein Science*, **18**(2), 359–371.

Knight, P. A. and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, **33**(3), 1029–1047.

Kondrashov, D. A., Van Wynsberghe, A. W., Bannen, R. M., Cui, Q., and Phillips, J. G. N. (2007). Protein structural variation in computational models and crystallographic data. *Structure*, **15**, 169 – 177.

Kundu, S., Melton, J. S., Sorensen, D. C., and Phillips, J. G. N. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.*, **83**, 723 – 732.

Lesne, A., Riposo, J., Roger, P., Cournac, A., and Mozziconacci, J. (2014). 3D genome reconstruction from chromosomal contacts. *Nature methods*, **11**(11), 1141–1143.

Lévy-Leduc, C., Delattre, M., Mary-Huard, T., and Robin, S. (2014). Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**(17), i386–i392.

Li, D. W. and Brüschweiler, R. (2009). All-atom contact model for understanding protein dynamics from crystallographic b-factors. *Biophysical journal*, **96**(8), 3074–3081.

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**(5950), 289–293.

Lin, C. P., Huang, S. W., Lai, Y. L., Yen, S. C., Shih, C. H., Lu, C. H., Huang, C. C., and Hwang, J. K. (2008). Deriving protein dynamical properties from weighted protein contact number. *Proteins: Structure, Function, and Bioinformatics*, **72**(3), 929–935.

Liu, T. T., Chen, M. X., and Lu, B. Z. (2015). Parameterization for molecular Gaussian surface and a comparison study of surface mesh generation. *Journal of molecular modeling*, **21**(5), 113.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**(7469), 59–64.

Nguyen, D. D., Xia, K. L., and Wei, G. W. (2016). Generalized flexibility-rigidity index. *Journal of Chemical Physics*, **144**, 234106.

Nguyen, D. D., Xiao, T., Wang, M. L., and Wei, G. W. (2017). Rigidity strengthening: A mechanism for protein–ligand binding. *Journal of chemical information and modeling*, **57**(7), 1715–1721.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., *et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398), 381–385.

Opron, K., Xia, K. L., and Wei, G. W. (2014). Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *Journal of Chemical Physics*, **140**, 234105.

Opron, K., Xia, K. L., and Wei, G. (2015a). Communication: Capturing protein multiscale thermal fluctuations. *The Journal of chemical physics*, **142**(21), 211101.

Opron, K., Xia, K. L., and Wei, G. W. (2015b). Communication: Capturing protein multiscale thermal fluctuations. *Journal of Chemical Physics*, **142**(211101).

Park, J. K., Jernigan, R., and Wu, Z. (2013). Coarse grained normal mode analysis vs. refined gaussian network model for protein residue-level structural fluctuations. *Bulletin of Mathematical Biology*, **75**, 124 –160.

Rao, S. S., 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.* (2014a). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**(7), 1665–1680.

Rao, S. S., 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.* (2014b). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**(7), 1665–1680.

Sauerwald, N., Zhang, S., Kingsford, C., and Bahar, I. (2017). Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. *Nucleic acids research*, **45**(7), 3663–3673.

Schmitt, A. D., Hu, M., and Ren, B. (2016a). Genome-wide mapping and analysis of chromosome architecture. *Nature reviews Molecular cell biology*, **17**(12), 743.

Schmitt, A. D., Hu, M., and Ren, B. (2016b). Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, **17**, 743–755.

Segal, M. R., Xiong, H., Capurso, D., Vazquez, M., and Arsuaga, J. (2014). Reproducibility of 3d chromatin configuration reconstructions. *Biostatistics*, **15**(3), 442–456.

Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004). FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. *Journal of Computational Biology*, **11**(1), 83–8106.

Siahpirani, A. F., Ay, F., and Roy, S. (2016). A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. *Genome biology*, **17**(1), 114.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature genetics*, **38**(11), 1348–1354.

Song, G. and Jernigan, R. L. (2007). vgnm: a better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.*, **369**(3), 880 – 893.

Tama, F., Gadea, F. X., Marques, O., and Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Structure, Function, and Bioinformatics*, **41**(1), 1–7.

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., Le Gros, M. A., *et al.* (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences*, **113**(12), E1663–E1672.

Wang, H., Duggal, G., Patro, R., Girvan, M., Hannenhalli, S., and Kingsford, C. (2013). Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 306. ACM.

Witten, D. M. and Noble, W. S. (2012). On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic acids research*, **40**(9), 3849âĽ“–3855.

Xia, K. L. and Wei, G. W. (2016). A review of geometric, topological and graph theory apparatuses for the modeling and analysis of biomolecular data. *arXiv preprint arXiv:1612.01735*.

Xia, K. L., Opron, K., and Wei, G. W. (2013). Multiscale multiphysics and multidomain models - Flexibility and Rigidity. *Journal of Chemical Physics*, **139**, 194109.

Xia, K. L., Li, Z. M., and Mu, L. (2018). Multiscale persistent functions for biomolecular structure characterization. *Bulletin of mathematical biology*, **80**(1), 1–31.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**(11), 1059–1065.

Zhang, B. and Wolynes, P. G. (2015). Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, **112**(19), 6062–6067.

Zhang, F. L. and Brüschweiler, R. (2002). Contact model for the prediction of nmr nh order parameters in globular proteins. *Journal of the American Chemical Society*, **124**(43), 12654–12655.

Zhang, Z. Z., Li, G. L., Toh, K. C., and Sung, W. K. (2013). 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of computational biology*, **20**(11), 831–846.

Zhao, Z. H., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., *et al.* (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics*, **38**(11), 1341–1347.

Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., Peng, J., Kaplan, T., and Zeng, J. (2018). Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic acids research*, **46**(8), e50–e50.