Inference of Spatial Organizations of Chromosomes Using Semi-Definite Embedding Approach and Hi-C Data

ZhiZhuo Zhang[1], Guoliang Li[2], Kim-Chuan Toh[3], Wing-Kin Sung[1,2]

[1] School of Computing, National University of Singapore & [3]Department of Mathematics, National University of Singapore & [2]Genome Institute of Singapore
[1]{zhizhuo, ksung}@comp.nus.edu.sg & [3]mattohkc@nus.edu.sg & [2]ligl@gis.a-star.edu.sg

Abstract. For a long period of time, scientists studied genomes assuming they are linear. Recently, chromosome conformation capture (3C) based technologies, such as Hi-C, have been developed that provide the loci contact frequencies among loci pairs in a genome-wide scale. The technology unveiled that two far-apart loci can interact in the tested genome. It indicated that the tested genome forms a 3D chromsomal structure within the nucleus. With the available Hi-C data, our next challenge is to model the 3D chromosomal structure from the 3C-dervied data computationally. This paper presents a deterministic method called ChromSDE, which applies semi-definite programming techniques to find the best structure fitting the observed data and uses golden section search to find the correct parameter for converting the contact frequency to spatial distance. To the best of our knowledge, ChromSDE is the only method which can guarantee recovering the correct structure in the noise-free case. In addition, we prove that the parameter of conversion from contact frequency to spatial distance will change under different resolutions theoretically and empirically. Using simulation data and real Hi-C data, we show that ChromSDE is much more accurate and robust than existing methods. Finally, we demonstrate that interesting biological findings can be uncovered from our predicted 3D structure.

Key words: Chromatin Interaction, 3D genome, Hi-C, Semi-definite Programming

Program and Supplementary online: `http://biogpu.ddns.comp.nus.edu.sg/~zzz/ChromSDE/`

1 Introduction

Genome is usually assumed to be a set of linear chromosomes. This model, however, is oversimplified and it cannot explain the interactions among different genomic elements (e.g., enhancer, promoter, gene). Chromosome actually forms a 3D structure within the nucleus and its spatial organization affects many chromosomal mechanisms such as gene regulation, DNA replication, epigenetic modification and maintenance of genome stability[3, 8, 15, 18–20].

Generally, if two elements in the genome are close in the sequence level, they are also close in the structure level. But the converse statement is not necessarily true. For example, Li et al[15] showed that multiple related genes are located far away in linear model but are organized topologically close through long-range chromatin interactions and transcribed in a single "transcription factory". In the past, the 3D organization of chromosomes was usually studied by florescent in situ hybridization (FISH) which are low throughput and low resolution methods. Recently, several high throughput, high resolution methods [6, 9, 12, 16, 29] derived from the 3C method [4] have been proposed. These methods measure the contact frequencies for loci pairs. Two loci are expected to be spatially nearer if and only if the contact frequency of the loci pair is higher. 4C[29] and 5C[6] can measure the contact frequencies among a subset of loci while Hi-C[16] and its variant (TCC [12]) can capture the contact frequencies in a genome-wide manner.

Given the 3C-derived data, one interesting bioinformatics problem is to infer the 3D structure of the genome. A number of works have been proposed recently. All the current methods have two steps: (1) Converting the contact frequencies between loci to spatial distances and (2) Predicting the 3D chromosomal structure from the spatial distances. Duan et al. [7] converted the contact frequencies extracted from the 4C experiment on yeast to spatial distances and treated the 3D structure modeling problem as a constrained non-convex quadratic optimization problem using an optimization solver called IPOPT[26]. Bau et al. [1] translated the contact frequencies extracted from 5C experiments to spatial distances by inverting the Z-score of contact frequencies and treated the 3D structure modeling problem as finding an equilibrium state of a set of particles using Integrated Modeling Platform (IMP)[23] . With the same platform (IMP), Kalhor et al.[12] claimed that the Hi-C (or TCC) data can be better fitted by learning a set of 3D structures (since the sample has multiple cells where the chromatin structures in different cells are different) instead of one single structure. More recently, two Markov-chain Monte Carlo (MCMC)[21] sampling-based methods, MCMC5C[22] and BACH[10], were proposed to infer the 3D structures by maximizing the likelihood of the observed Hi-C data. Both methods assume that the expected contact frequencies and spatial distances among loci follow the power law distribution. MCMC5C[22] models the observed frequency with Gaussian distribution with respect to the expected frequency. BACH[10] models the observed frequency with Poisson distribution with respect to the expected frequency and takes the enzyme cutting site bias (e.g., CG content, mappability, fragment length) into account.

Although some works have been done, there are unsolved issues in both steps 1 and 2. For step 1, the conversion between the contact frequency and spatial distance has one parameter. Existing methods, except BACH, assume that the parameter is fixed or is known beforehand. We found that the parameter is actually different for different datasets. Thus it is important to have a method to estimate the parameter. For step 2, existing methods infer the 3D chromosomal structure by heuristics. They are not guaranteed to reconstruct the correct structure even in the noise-free case.

To fill in these gaps, we propose a novel chromosome structure modeling algorithm called ChromSDE (Chromosome Semi-Definite Embedding). ChromSDE models the problem as two parts:

1. Assuming that the parameter for the conversion from the contact frequency to the spatial distance is known, ChromSDE formulates the 3D structure modeling problem as a non-convex

non-linear optimization problem similar to the previous works. Instead of directly solving the non-convex optimization which is NP-hard, ChromSDE relaxes it to a semi-definite programming(SDP) problem, whose global optimal solution can be computed in polynomial time. With this formulation, our approach is guaranteed to recover the correct 3D structure in the noise-free case when the structure is uniquely localizable[24].

2. For the parameter in our conversion function from the contact frequency to the spatial distance, ChromSDE formulates it as a univariate optimization problem and estimate the correct parameter by a modified version of the golden section search method.

This paper may have significant impact in three aspects. First, the SDP relaxation method in ChromSDE is a powerful relaxation technique, which is theoretically guaranteed to recover the correct structure in the uniquely localizable noise-free case[24]. The SDP approach has been successfully applied in other graph realization problems[2, 14, 27], but to our best knowledge, no one has introduced it in chromosome structure modeling. Second, we prove theoretically and empirically that the conversion parameter changes if we examine the data under different resolutions. Thus, it is inappropriate to assume that the conversion parameter is known. We developed an efficient algorithm to estimate the correct conversion parameter from the input data. Third, we proposed a measure called *Consensus Index* which can quantify if the input frequency data comes from a consensus structure or a mixture of different structures. It is arguable if Hi-C data is appropriate for modeling 3D structures, because the contact frequencies come from a population of cells instead of a single cell. Our simulation shows that if the data is from a consensus structure, the *Consensus Index* is high.

We evaluated our method with simulated data and real Hi-C data. Through simulation study, we showed that ChromSDE can perfectly recover different types of simulated structures in the noise-free setting while other tested programs fail in many cases. Even with noise, ChromSDE still significantly outperforms other tested programs. In addition, we also show that ChromSDE can accurately estimate the conversion parameter and output the *Consensus Index* that can reflect the degree of mixture. Next, real Hi-C data replicates with different cutting enzymes are used to further validate the robustness and accuracy of ChromSDE comparing to other tested programs. The result indicates that ChromSDE can infer a more accurate and robust 3D model than existing methods. Finally, we show that ChromSDE can robustly handle different resolution data and the predicted high resolution 3D structure unveils interesting biological findings.

2 Method

The Hi-C and TCC technologies enable us to obtain paired-end reads from interacting loci in the genome. The interaction data can be summarized by a contact frequency matrix F, in which F_{ij} represents the number of contacts between loci i and j (loci i and j are genomic regions in a fixed bin size such as 1Mbp or 40kb). We expect two loci are spatially close if and only if the contact frequency between them is high. A further note is that the raw Hi-C or TCC interaction frequencies are affected by various biases (GC content, mappability and fragment length), and should be normalized [28].

The chromatin 3D modeling problem is defined as follows: Given a normalized interaction frequency matrix F, infer a 3D structure whose pairwise distances highly correlate with the interaction frequencies in F. This problem can be solved by 1) converting the frequency matrix F into a distance matrix D that describes the expected pairwise distance among the loci; 2) learning a 3D structure from the distance matrix D. Step 1 is based on the observation of Lieberman et al. [16] that the conversion between the frequency matrix F and the distance matrix D follows the power law distribution (Equation 1) where α is a parameter called the conversion factor and D_{ij} and F_{ij} are the distance and frequency between loci i and j.

$$D_{ij} = \begin{cases} (1/F_{ij})^{\alpha} & \text{if } F_{ij} > 0 \\ \infty & \text{otherwise} \end{cases} \tag{1}$$

There are two main challenges in this approach: 1) estimate α; and 2) convert the distance matrix D to the 3D model. In the following two sub-sections, we present ChromSDE that resolves these two challenges. First, assuming that the conversion factor α is known, we describe a method that estimates the 3D structure from the expected distance matrix D. Then, the next section explains how ChromSDE estimates the correct value of the conversion factor α. To note that, the scale between the converted distance and the real physical distance is not considered here, since it does not affect the predicted structure.

2.1 From Distance Matrix To 3D Structure

Assuming the conversion factor $\alpha(> 0)$ is known, the interaction frequency matrix F can be converted to the expected distance matrix D by Equation 1.

The 3D chromatin structure modeling problem aims to compute a set of 3-dimensional coordinates $\{\overrightarrow{x_1}, ..., \overrightarrow{x_n}\}$ for the n loci, such that their distances can fit the distance matrix D well. In other words, we hope to ensure that $\|x_i - x_j\|$ (distance between loci i and j) is approximately the same as D_{ij} for all loci i and j.

Mathematically, this problem can be formulated as three alternative optimization models in Equations (2)-(4), where $\| \cdot \|$ denotes the Euclidean norm. Each equation has two terms. The first term aims to minimize the errors between the embedding distances and the expected distances. The three alternatives apply three different commonly used error functions in the literature: (a) sum of square errors of the distance differences [1, 7], (b) sum of absolute errors of the distance square differences [2, 14] and (c) sum of square errors of the distance square differences [2, 17]. The second term is the same for the three alternatives. It is a regularization term that maximizes the pairwise distances for the loci without any interaction frequency data. It is based on the assumption that the spatial distances of loci pairs not captured by the experiment cannot be too short.

$$\min_{\overrightarrow{x}_1,...,\overrightarrow{x}_n \in \mathrm{R}^3} \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij} \cdot \Big(\|\overrightarrow{x}_i - \overrightarrow{x}_j)\| - D_{ij} \Big)^2 - \lambda \sum_{\{i,j|D_{ij}=\infty\}} \|\overrightarrow{x}_i - \overrightarrow{x}_j\|^2 \tag{2}$$

$$\min_{\overrightarrow{x}_1,...,\overrightarrow{x}_n \in \mathrm{R}^3} \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij} \cdot \Big| \|\overrightarrow{x}_i - \overrightarrow{x}_j)\|^2 - D_{ij}^2 \Big| - \lambda \sum_{\{i,j|D_{ij}=\infty\}} \|\overrightarrow{x}_i - \overrightarrow{x}_j\|^2 \tag{3}$$

$$\min_{\overrightarrow{x}_1,...,\overrightarrow{x}_n \in \mathrm{R}^3} \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij} \cdot \Big(\|\overrightarrow{x}_i - \overrightarrow{x}_j)\|^2 - D_{ij}^2 \Big)^2 - \lambda \sum_{\{i,j|D_{ij}=\infty\}} \|\overrightarrow{x}_i - \overrightarrow{x}_j\|^2 \tag{4}$$

In the formulas, ω_{ij} represents the weight or confidence of the observed data D_{ij}. Since we expect the confidence of D_{ij} is higher when F_{ij} is large, this paper simply set $\omega_{ij} = 1/D_{ij}$. The parameter $\lambda > 0$ in the second term is the regularization coefficient to balance the error term and the regularization term. In practice, we found that the results are stable for $0.001 < \lambda < 0.1$ (Supp Figure 1) and we fix it to 0.01 in this paper.

All three formulations (2)–(4) are non-convex non-linear optimization problems, which are NP-hard to solve for their global minimizers. Existing methods solved them by heuristics like MCMC sampling [10, 22] and local search [7, 12, 23]. Here, we show that, by relaxing the solution space of every $\overrightarrow{x}_i$ from R^3 to R^n (n is the number of loci), formulations (3) and (4) become convex semidefinite programming (SDP) problems for which we can compute their global minimizers to any given degree of accuracy in polynomial time. Furthermore, if the expected distance matrix is indeed generated from a 3D object and is noise-free, the above relaxations can reconstruct the optimal R^3 solution by projecting the R^n points to certain R^3 subspace in theory [24]. In practice, even if the distance matrix is not noise-free, we still can find a good approximated solution in the R^3 subspace. The projecting technique to obtain a solution in R^3 will be introduced later.

Formulation of SDP relaxation problems This section describes how to reformulate Equations (3) and (4) as linear and quadratic semidefinite programming (SDP) problems by relaxing the solution space of every $\overrightarrow{x}_i$ from R^3 to R^n. Let K be the kernel matrix for $X = [\overrightarrow{x}_1, \overrightarrow{x}_2, \ldots, \overrightarrow{x}_n]$ (i.e., $K_{ij} = \overrightarrow{x}_i \cdot \overrightarrow{x}_j = K_{ji}$), then every square distance can be expressed in term of K. Precisely, we have: $\|\overrightarrow{x}_i - \overrightarrow{x}_j\|^2 = K_{ii} + K_{jj} - 2K_{ij}$. In addition, we set the center of the points to be the origin, that is:

$$\sum_{i=1}^{n} \overrightarrow{x}_i = 0 \;\Rightarrow\; \|\sum_{i=1}^{n} \overrightarrow{x}_i\|^2 = 0 \;\Rightarrow\; \sum_{i,j} K_{ij} = 0. \tag{5}$$

By our definition of the kernel matrix, K must be symmetric positive semidefinite (i.e., $K \succeq 0$). We first describe the quadratic relaxation (Equation (4)), which is stated as below:

$$\begin{aligned} \min \sum_{\{i,j|D_{ij}<\infty\}} & \omega_{ij}(K_{ii} + K_{jj} - 2K_{ij} - D_{ij}^2)^2 - \lambda \sum_{\{i,j|D_{ij}=\infty\}} (K_{ii} + K_{jj} - 2K_{ij}) \\ \text{s.t. } & \textstyle\sum_{ij} K_{ij} = 0, \quad K \succeq 0. \end{aligned} \tag{6}$$

For Equation (3), the error term contains the absolute value operator $|\cdot|$, which cannot be handled directly by standard SDP solvers. Fortunately, without increasing the problem complexity, we can replace the absolute value operator $|\cdot|$ by adding two sets of slack variables. The linear SDP relaxation of Equation (3) is stated as below:

$$\begin{aligned} \min & \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij}(\varepsilon_{ij}^{+} + \varepsilon_{ij}^{-}) - \lambda \sum_{\{i,j|D_{ij}=\infty\}} (K_{ii} + K_{jj} - 2K_{ij}) \\ \text{s.t. } & K_{ii} + K_{jj} - 2K_{ij} + \varepsilon_{ij}^{+} - \varepsilon_{ij}^{-} = D_{ij}^2 \\ & \textstyle\sum_{ij} K_{ij} = 0, \quad K \succeq 0, \quad \varepsilon_{ij}^{+}, \varepsilon_{ij}^{-} \geq 0. \end{aligned} \tag{7}$$

Note that ε_{ij}^{+} (and ε_{ij}^{-} respectively) represents the penalty when the embedding distance is shorter (and longer respectively) than the expected distance. Moreover, at least one of them must be zero in the final solution since they are non-negative and their summation is minimized.

A general purpose SDP solver, such as SDPT3 [25], can be used to solve the two SDP problems above. However, all the current general purpose SDP solvers (which are all based on interior-point methods) cannot handle large scale SDP problems. They can only comfortably handle distance matrix with around 40,000 expected distances (≈ 200 loci). Fortunately, for convex quadratic SDP such as the problem (6), recently developed advanced algorithm [11] based on partial proximal-point method (with semi-smooth Newton-CG method for solving the subproblems) can handle such a problem very efficiently even when the problem scale is large. In particular, it can handle 10,000,000 expected distances (≈ 3000 loci). In the result section, we present the results for both SDP relaxations in the small scale problems and the results for the quadratic SDP relaxation in the large scale problems (if not specially mentioned, the result is generated by quadratic SDP).

Obtaining 3D coordinates from the Kernel Matrix By solving the SDP problems (6) or (7), we obtain the solution as a positive semidefinite kernel matrix K. By computing the eigenvalue decomposition of K, the R^3 coordinates $X = [\overrightarrow{x}_1, \ldots, \overrightarrow{x}_n]$ can be recovered from K (i.e., $K \approx X^T X$). A 3-dimensional representation that approximately satisfies $K_{ij} \approx \overrightarrow{x}_i \cdot \overrightarrow{x}_j$ can be obtained from the top 3 eigenvalues $(\gamma_1, \gamma_2, \gamma_3)$ and eigenvectors $(\overrightarrow{\nu_1}, \overrightarrow{\nu_2}, \overrightarrow{\nu_3})$ of K. That is,

$$\overrightarrow{x}_i = [\sqrt{\gamma_1} \cdot \nu_{1,i} \;\; \sqrt{\gamma_2} \cdot \nu_{2,i} \;\; \sqrt{\gamma_3} \cdot \nu_{3,i}]^T. \tag{8}$$

In the ideal case where the input expected distance matrix is noise-free and dense enough (i.e., it has sufficient constraints to uniquely present a 3D structure), it can be shown that the approximation (8) is the exact solution and all other eigenvalues are equal to zero. This property is called unique localizability [24].

When the input expected distance matrix is noisy, ChromSDE performs further local refinement to the 3D coordinates obtained from the SDP relaxation problems [2]. Specifically, our ChromSDEx algorithm applies a local optimization method such as a quasi-Newton method or a gradient descent method to the original non-convex problem by using the 3D positions obtained from the SDP problems as the starting iteration point. Because the 3D positions produced by the SDP problems are generally close to a local minimizer, a local optimization method can generally converge to a good local minimizer for the original non-convex problems.

To measure if the input distance matrix can be represented as a single 3D structure, we propose a measure called *Consensus Index*, which includes two parts: the first part measures how the input distance matrix satisfying the triangle inequality, and is presented as the ratio between the embedded distance in R^n and the input distance; the second part measures how good the R^3 approximation is, and is presented as the ratio between the sum of top 3 eigenvalues and the sum of all eigenvalues of K. Precisely, Let $D'_{ij} = \sqrt{K_{ii} - 2K_{ij} + K_{jj}}$ be the embedded distance in R^n, then we have:

$$Consensus\ Index = \frac{\sum_{\{i,j|D_{ij}<\infty\}} min(D'_{ij}/D_{ij}, D_{ij}/D'_{ij})}{|\{i,j \mid D_{ij} < \infty\}|} \cdot \frac{\sum_{i=1}^{3} \gamma_i}{\sum_{i=1}^{n} \gamma_i} \tag{9}$$

Note that the *Consensus Index* is between 0 and 1. When the *Consensus Index* trends to 1, this means that the input distance matrix fits a single 3D structure well. The result section showed that the *Consensus Index* is indeed a good indicator on whether the input data corresponds to a single 3D structure or a mixture of 3D structures.

2.2 Searching for the Correct Conversion Factor

In Section 2.1, the conversion factor $\alpha(> 0)$ is assumed to be known. However, the assumption is not valid . Even worse, Lemma 1 shows that the conversion factor changes with different resolutions.

Lemma 1. *Consider the frequency matrix F for loci $x_1, \ldots, x_{2n}$. Let the conversion factor of F be $\alpha > 0$, i.e., distance between loci x_i and x_j is $d_{ij} = (1/F_{ij})^\alpha$. Now, we reduce the resolution by merging adjacent loci, i.e., we generate the frequency matrix F' for the low resolution loci $y_1 \ldots, y_n$ where y_i is formed by merging adjacent loci x_{2i-1} and x_{2i}. Suppose $F'_{ij} = (F_{2i-1,2j-1} + F_{2i-1,2j} + F_{2i,2j-1} + F_{2i,2j})$ and d'_{ij} can be approximated as either arithmetic mean or geometry mean of $\{d_{2i-1,2j-1}, d_{2i-1,2j}, d_{2i,2j-1}, d_{2i,2j}\}$.*

Then the conversion factor α' of F' is less than or equal to α .

Proof. Note that $\log F_{p,q} > 0$ and $\log d_{p,q} < 0$ since $F_{p,q} \geq 1$. Let $d_{\min} = \min_{p \in \{2i,2i-1\}, q \in \{2j,2j-1\}} d_{p,q}$. Since $d'_{ij} \geq d_{\min}$, we have $\log d'_{ij} \geq \log d_{\min}$. We also have

$$F'_{ij} = \sum_{p \in \{2i,2i-1\}, q \in \{2j,2j-1\}} \frac{1}{d_{p,q}^{1/\alpha}} \geq \frac{1}{d_{\min}^{1/\alpha}}.$$

Hence $\log F'_{ij} \geq -\frac{1}{\alpha} \log d_{\min}$. As $d'_{ij} = (1/F'_{ij})^{\alpha'}$, we have $\alpha' = \frac{-\log d'_{ij}}{\log F'_{ij}} \leq \frac{-\log d_{\min}}{-\frac{1}{\alpha}\log d_{\min}} = \alpha$. □

The Lemma 1 implies that the conversion factor of high-resolution Hi-C datasets is usually larger than that of low-resolution Hi-C datasets. Hence, we cannot assume that the conversion factor is a prior or is a fix value for different datasets. In fact, the predicted 3D structure is quite sensitive to the conversion factor. Given the same frequency matrix, different conversion factor leads to different expected distances and finally implies very different 3D structures (Supp Figure 2). Therefore, estimating the correct conversion factor for a frequency matrix F is important.

A correct conversion factor enables us to convert a frequency matrix to a correct 3D model, and vice versa. Based on this principle, for a frequency matrix F, the goodness of a conversion factor α ($goodness(\alpha, F)$) can be determined by comparing the predicted frequency matrix $\widehat{F}$ and the input frequency matrix F. Figure 1 details the function to compute $goodness(\alpha, F)$.

```
Algorithm ChromSDE
Require: normalized frequency matrix $F$
Ensure: a set of 3D coordinates $X$, conversion factor $\alpha$
$\alpha_{min} = 0.1$, $\alpha_{max} = 3$                          # set search boundary for $\alpha$
$\varphi = \frac{\sqrt{5}-1}{2}$                                  # golden section ratio
repeat
    $\eta = (\frac{\alpha_{max}}{\alpha_{min}})^{\varphi}$        # step size for updating $\alpha$
    $x1 \leftarrow \alpha_{min} \cdot \eta$ , $f1 \leftarrow goodness(x1, F)$
    $x2 \leftarrow \alpha_{max} / \eta$ , $f2 \leftarrow goodness(x2, F)$
    if $f1 > f2$ then
        $\alpha_{min} \leftarrow x2$                              # increase lower bound
    else
        $\alpha_{max} \leftarrow x1$                              # decrease upper bound
    end if
until $(\alpha_{max} - \alpha_{min}) <$ tolerance
$\alpha \leftarrow \alpha_{min}$                                  # final value of $\alpha$
$D \leftarrow (1/F)^{\alpha}$                                     # expected distance matrix
$X \leftarrow$ compute 3D structure using SDP method based on $D$
Function goodness($\alpha$, $F$)
$D \leftarrow (1/F)^{\alpha}$
$X \leftarrow$ compute 3D structure using SDP method based on $D$
$D' \leftarrow$ compute pair-wise distances from $X$
$F' \leftarrow (1/D')^{1/\alpha}$
Return $\sum_{\{(i,j)|F_{i,j}>0\}} -|F'_{i,j} - F_{i,j}|$
```

Fig. 1. Algorithm description for ChromSDE.

Our aim is to compute α that maximizes the goodness function. As there is no obvious well defined gradient for the goodness function, we cannot use methods such as gradient descent or Newton's method to optimize α. Instead, we perform the golden section search method to optimal α, assuming that the goodness function is unimodal in the search interval. Since $d_{ij} = (1/F_{ij})^{\alpha}$, we deduce that α cannot be too small; otherwise the spatial distance will be independent of the frequency (when $\alpha \to 0$). Also, α cannot be too large; otherwise, a small difference in frequencies will lead a very big difference in spatial distances, and small noise will seriously violation of the triangle inequality. In this paper, we assume that $0.1 \leq \alpha \leq 3$. Moreover, we observed that applying

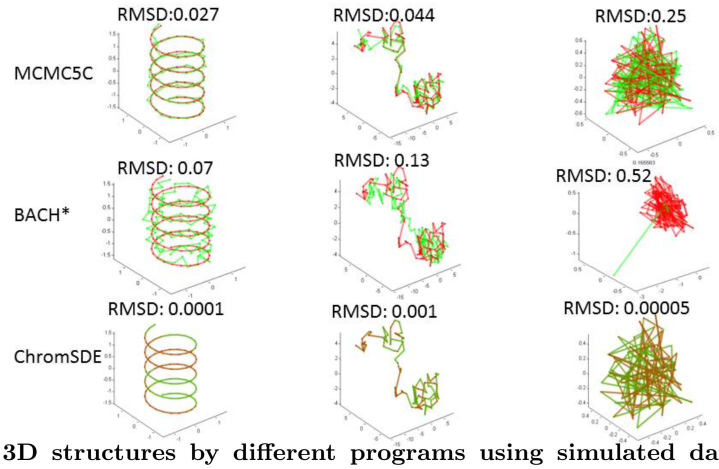

Fig. 2. Predicted 3D structures by different programs using simulated data. Red curve is the true structure and green curve is the predicted structure. ChromSDE uses quadratic SDP here and the linear SDP has the same performance.

the standard golden section search on the logarithm domain of the interval is more efficient(see Supp Figure 3). The algorithm detail is in Figure 1.

3 Result

3.1 Simulation Study

To analyze the performance of ChromSDE, we generated three different types of 3D structures(Supp Figure 4): (1)Helix curve, (2)Brownian motion simulation of a single particle and (3)Uniform random points in a cube. Each structure is represented by 100 points. We assume that the Hi-C technique is sensitive enough to capture interactions with at most 50 nearest neighbours and the conversion factor α is 1, i.e., the contact frequency f of two given points can be computed as $f = (1/d)^{1/\alpha} = 1/d$, where d is the spatial distance between given points. We compared our algorithm with the existing methods MCMC5C[22] and BACH[10], which are the only publicly available standalone programs that are suitable for general Hi-C data. For MCMC5C, it cannot estimate the conversion factor by itself, so we supplied it with the correct value. For BACH, it can estimate the conversion factor with the default starting point equal to 1 (i.e., the correct value in our simulation study). Since there is no enzyme bias in our simulation, we also modify BACH to suppress this feature (called BACH*). For ChromSDE, we assume that the conversion factor is within the range (0.1,3), so we give advantages to the existing programs, but not our ChromSDE.

ChromSDE guarantees optimality in noise-free case Figure 2 shows the true simulated structures and the predicted structures by different programs. For the helix curve, all three programs can recover the structure correctly. For the Brownian motion curve, both ChromSDE and MCMC5C can almost perfectly recover the true structure and BACH* can only reproduce a not-so-accurate but similar structure. For the third case, MCMC5C produced a not-so-accurate structure and BACH* completely failed in this case, while our ChromSDE still can perfectly recover the true structure. The result is not surprising since SDP method is the only one that can guarantee perfect recovery of the true structure when the input data is noise-free and the structure is uniquely localizable. Based on the RMSD(root mean square deviation), ChromSDE also outperforms the other two methods in all the three simulated cases.

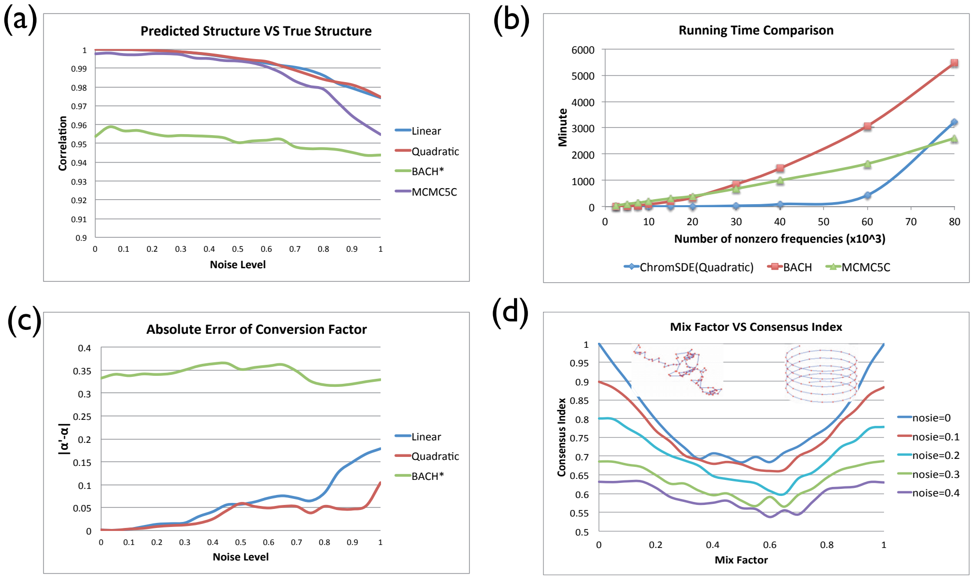

Fig. 3. Performance of different methods on simulated data. (a) Spearman correlation between the pair-wise distance matrices of the predicted structure and the true structure under different noise level. (b) Running times of tested programs given different number of pairs of observed frequency (test stop at 80000 pair-wise frequencies, 1600 points). (c) The absolute error of the estimated value of conversion factor under different noise levels. (d) The *Consensus Index* predicted by ChromSDE(quadratic model) under different degree of mixture of helix curve(right) and Brownian motion curve(left).

ChromSDE outperforms the existing methods in noisy-data The previous section showed that ChromSDE can recover the optimal chromatin structure in the noise-free case. Now, we test whether ChromSDE is robust in a noisy data setting. To study this, we simulated noisy contact frequency data in different noise level based on the Brownian curve structure. For any two loci i and j, the noisy frequency $\tilde{F_{ij}}$ is deviated from the true frequency $F_{ij} = 1/D_{ij}$ (D_{ij} is the spatial distance between loci i and j) by adding a uniform random noise δ within a given noise level. Precisely, $\tilde{F_{ij}} = F_{ij}(1 + \delta)$ where $|\delta|$ is smaller than the noise level.

Figure 3 shows the performance of the programs with different noise levels under different measurements. Figure 3(a) shows that, when the noise level increases, the Spearman correlation between the pairwise distances from the predicted structure and those from the true structure generally decreases. ChromSDE and MCMC5C perform similarly when the noise level < 0.6 and ChromSDE (both linear SDP and quadratic SDP) outperforms others when the noise level is higher than 0.6. Similar result is observed when we measure the RMSD between the predicted structure and the true structure(Supp Figure 5(a)). In Figure 3(c), we observed that ChromSDE can estimate the conversion factor quite accurately (deviation <0.1) when noise level <0.7. In contrast, the estimated conversion factor from BACH* tends to be incorrect (deviation around 0.35). This may be the reason why BACH* has worse performance comparing to others across different noise levels. Moreover, ChromSDE is faster than BACH and comparable to MCMC5C even though ChromSDE needs to search for the correct conversion factor but MCMC5C does not (Figure 3(b)). In summary, the result shows that the linear SDP and quadratic SDP models perform quite consistently and ChromSDE is more robust and accurate than existing methods.

***Consensus Index* indicates the degree of mixture of 3D structures** In Hi-C and TCC experiments, the data is from a population of cells, and each potentially has different 3D chromosomal structure. The method section proposed to use the *Consensus Index* to determine if the data is from a consensus 3D structure. To show that the *Consensus Index* is a good indicator of the degree of mixture, we generated a frequency matrix F_{merge} by merging the frequency matrix from the helix curve F_1 and the Brownian motion curve F_2 under different mix factor γ (i.e., $F_{merge} = \gamma \cdot F_1 + (1 - \gamma) \cdot F_2$). Figure 3(d) shows that the $Consensus\ Index$ is affected by both the noise level and mix factor. For the same noise level, the $Consensus\ Index$ approaches the minimum when the mix factor is close to 0.5. This indicates that the $Consensus\ Index$ is the lowest when the two structures are highly mixed. For different noise levels, the $Consensus\ Index$ decreases as the noise level increases. Also we note that the estimated conversion factors by ChromSDE are quite consistent with its true value even under different mix factors and noise levels(Supp Figure 5(c)).

3.2 Real Hi-C Data Study

Validate ChromSDE using two enzyme replicates From the literature, two different enzymes(Hind3, NcoI) were used to generate Hi-C replicate data from the mouse ES cell(mESC)[5] and the human GM06990 cell(GM)[16]. Each enzyme replicate is an independent observation of the chromosome structure in the same cell type. Hence, we expect the result produced by a robust algorithm using one enzyme data can be validated using the other enzyme data.

We applied four different programs ChromSDE, BACH*, BACH and MCMC5C to predict the 3D structures of different chromosomes in the two cell lines using the Hi-C data from two replicates. For ChromSDE, BACH* and MCMC5C, the input is a normalized frequency matrix using the normalization pipeline by Yaffe and Tanay[28]. For BACH, we provide the raw Hi-C frequency and enzyme cutting point feature data. More detail can be found in the supplementary material.

We compute Spearman correlation between the normalized frequency of one enzyme data and the estimated frequency ($frequency \sim 1/distance$) of the predicted structure from the other enzyme data. (We use Spearman correlation instead of Pearson correlation since the Spearman correlation is independent to the conversion between frequency and distance, hence it is fair to every tested program.) Figure 4(a) shows that ChromSDE (both Linear SDP and Quadratic SDP) outperforms the other programs by at least 5% across all four tested Hi-C datasets. Especially, in the mESC dataset, ChromSDE obtains the average correlation of 0.9 across all chromosomes but other tested programs only obtain correlation at most 0.82. What's more, Figure 4(b),(c) and Supp Figure 6 showed the 3D structures of different chromosomes predicted by ChromSDE are highly reproducible and the conversion factors estimated by ChromSDE are more consistent than the ones estimated by BACH and BACH* across different chromosomes and different enzymes (Supp Table 1).

Besides, we observed that all the tested programs perform worse in GM than in mESC and the $Consensus\ Index$ is around 0.9 in mESC and is only 0.7 in GM(Supp Figure 7). It indicates that mESC has a consensus 3D structure for its genome and GM is relatively diverse or has higher noise level due to the low sequencing depth.

ChromSDE can generate consistent 3D structures from different genomic resolutions We further tested ChromSDE on different genomic resolution data. Figure 5(a) showed that ChromSDE can predict similar structures of chromosome 13 under different resolutions using mESC Hind3 data (average Spearman correlation is 0.97, average RMSD is 0.08). In contrast, other existing programs cannot reproduce similar structures with different resolution data, especially for MCMC5C which cannot estimate the correct conversion factor(Supp Figure 8). We also showed that the conversion factor for each predicted structure in Figure 5(a). It demonstrated that the conversion factor increases as the resolution increases (also supported by BACH in Supp Figure 8). This further confirms the correctness of Lemma 1 even though the frequency has been normalized under different genomic resolutions.

To demonstrate the application of our predicted 3D structure, we generated a high resolution chromosome 3D structure for the region chr13:21Mb-25Mb (Figure 5(b)) using ChromSDE and

(a)

Validate Correlation	Quadratic SDP	Linear SDP	BACH	BACH*	MCMC5C
mESC_Ncol	**0.9096**	0.9095	0.7918	0.7871	0.8187
mESC_Hind3	**0.9080**	0.9070	0.8061	0.7999	0.7676
GM_Ncol	0.7674	**0.7899**	0.6740	0.7244	0.6942
GM_Hind3	**0.7832**	0.7763	0.7066	0.7358	0.7293

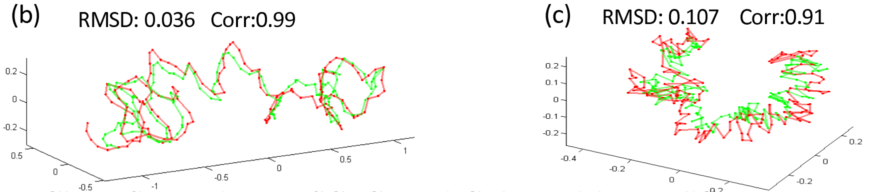

Fig. 4. Validate ChromSDE using mESC, GM Hi-C data with two different enzymes (Hind3, NcoI). (a) Average Spearman correlation across all chromosomes between inverse 3D distance and contact frequency from testing dataset. For each dataset, the best performer is highlighted. (b) Alignment between predicted structures of chromosome 1 of mESC Hind3(red) and mESC NcoI(green) by ChromSDE. (c) Alignment between predicted structures of chromosome 1 of GM Hind3(red) and GM NcoI(green) by ChromSDE.

mouse ES cell Hind3 data (40kbp resolution, estimated α is 0.83). Hist1h genes are highlighted with yellow color in the 3D structure, and we find that two groups of Hist1h genes are separated quite far away(∼1.5Mbp) in the linear genomic locations. In contrast, the promoters of two groups of Hist1h genes are spatially close to each other. To test if these two groups of genes interact each other for transcription, we checked the Pol2 ChIA-PET data available in our lab. We found that there are strong interactions(red dash line) between these two promoter regions mediated by Pol2, which indicates that the histone genes are co-regulated in the mouse ES cell.

Moreover, we found that the dense region and the loose region in the predicted 3D structure can be used to indicate the level of activity of those regions (from the snapshot of UCSC genome browser[13]). Dense regions (purple and blue color) correspond to repressive chromatin state in the cell, and there are few active histone modification and transcriptionx factor-binding events occurring in those regions. In contrast, loose regions(green and yellow color) correspond to active chromatin state in the cell, and there are a lot of histone modification and transcription factor-binding events occurring in those regions. Also, we found that loose regions usually containing more genes and are associated with early replication timing than the dense regions. It is also noted that the purple region is associated with LaminB1 binding and late replication timing, which suggests that Lamin may plays a part in the histone genes regulation and DNA replication.

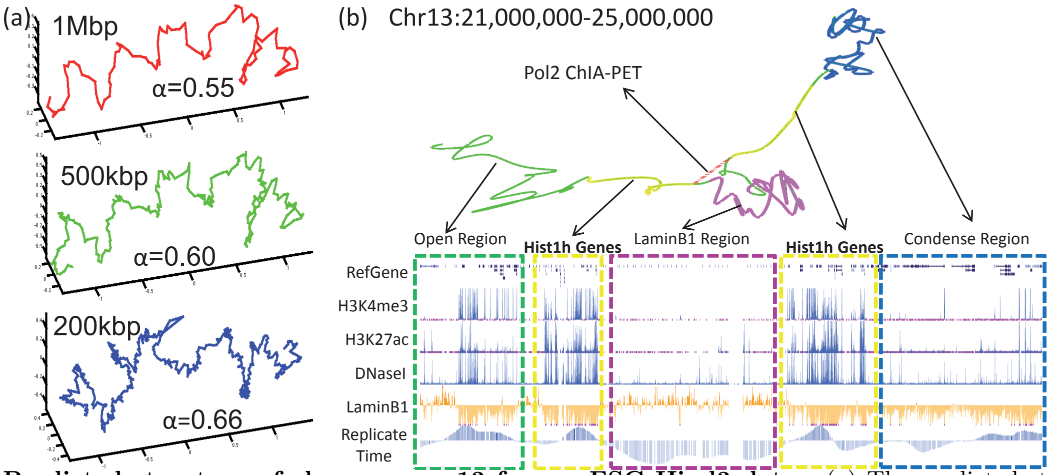

Fig. 5. Predicted structure of chromosome 13 from mESC Hind3 data. (a) The predicted structure of chromosome 13 under 1Mbp,500kbp,200kbp resolutions. (b) The predicted structure of the region chr13:21Mb-25Mb under 40kbp resolution and the different signal tracks of mESC from UCSC genome browser[13].

4 Discussion

In this study, we presented a method ChromSDE to reconstruct the consensus/dominate chromatin 3D structure of the given HiC data. To our best knowledge, ChromSDE is the only method which can guarantee recovering the correct structure in the noise-free case. In the noisy case, ChromSDE is much more accurate and robust than existing methods in both simulation and real data study. In addition, ChromSDE can automatically estimate the conversion factor, which is proved to change under different resolutions theoretically and empirically. Furthermore, we demonstrate that interesting biological findings can be uncovered from our predicted 3D structure.

We also developed the *Consensus Index* to determine how good the data can be explained by a single 3D structure. However, *Consensus Index* may not be informative when the noise level of the data is high or the mixing structures are similar. When the mixing structures are similar to each other then ChromSDE will learn the average structure. One future research is to recover all the mixing structures using Hi-C data.

Acknowledgements

The authors would like to thank Dr.Hugo Willy for the insightful discussions. This work was supported in part by the MOEs AcRF Tier 2 funding R-252-000-444-112.

References

1. D. Bau and M. A. Marti-Renom. Genome structure determination via 3c-based data integration by the integrative modeling platform, 2012.
2. P. Biswas, T.C. Liang, K.C. Toh, Y. Ye, and T.C. Wang. Semidefinite programming approaches for sensor network localization with noisy distance measurements. *Automation Science and Engineering, IEEE Transactions on*, 3(4):360–371, 2006.
3. J. Dekker. Gene regulation in the third dimension. *Science Signalling*, 319(5871):1793, 2008.
4. J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.
5. J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.
6. J. Dostie and J. Dekker. Mapping networks of physical interactions between genomic elements using 5c technology. *Nat Protoc*, 2(4):988–1002, 2007.
7. Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, 2010.
8. P. Fraser and W. Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143):413–417, 2007.
9. M.J. Fullwood and Y. Ruan. Chip-based methods for the identification of long-range chromatin interactions. *Journal of cellular biochemistry*, 107(1):30–39, 2009.
10. M Hu, K Deng, ZS Qin, J Dixon, S Selvaraj, J Fang, B Ren, and JS. Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS Computational Biology*, In press, 2012.
11. K.F. Jiang, D.F. Sun, and K.C. Toh. A partial proximal point algorithm for nuclear norm regularized matrix least squares problems. *National University of Singapore*, preprint, 2012.
12. R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, 30(1):90–98, 2012.
13. D. Karolchik, A.S. Hinrichs, and W.J. Kent. The ucsc genome browser. *Current protocols in bioinformatics*, pages 1–4, 2009.
14. N.H.Z. Leung and K.C. Toh. An sdp-based divide-and-conquer algorithm for large-scale noisy anchor-free graph realization. *SIAM Journal on Scientific Computing*, 31(6):4351–4372, 2009.
15. G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, 2012.

16. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.
17. Y.J. Liu, D. Sun, and K.C. Toh. An implementable proximal point algorithmic framework for nuclear norm minimization. *Mathematical programming*, pages 1–38, 2009.
18. A. Miele and J. Dekker. Long-range chromosomal interactions and gene regulation. *Mol. BioSyst.*, 4(11):1046–1057, 2008.
19. T. Misteli. Spatial positioning: A new dimension in genome function. *Cell*, 119(2):153–156, 2004.
20. T. Misteli et al. Beyond the sequence: cellular organization of genome function. *Cell*, 128(4):787, 2007.
21. R.M. Neal. Probabilistic inference using markov chain monte carlo methods. 1993.
22. M. Rousseau, J. Fraser, M.A. Ferraiuolo, J. Dostie, and M. Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC bioinformatics*, 12(1), 2011.
23. D. Russel, K. Lasker, B. Webb, J. Velazquez-Muriel, E. Tjioe, D. Schneidman-Duhovny, B. Peterson, and A. Sali. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol*, 10(1), 2012.
24. A.M.C. So and Y. Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2):367–384, 2007.
25. K.C. Toh, M.J. Todd, and R.H. Tütüncü. Sdpt3–matlab software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, 11(1-4):545–581, 1999.
26. A. Wächter and L.T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.
27. K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. volume 2, pages II–988–II–995 Vol. 2–. IEEE, 2004.
28. E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 2011.
29. Z. Zhao, G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 38(11):1341–1347, 2006.