Gauging Association Patterns of Chromosome Territories via Chromatic Median*

Hu Ding and Branislav Stojkovic
Department of Computer Science and Engineering,
State University of New York at Buffalo
{huding, bs65}@buffalo.edu

Ronald Berezney
Department of Biological Sciences,
State University of New York at Buffalo
berezney@buffalo.edu

Jinhui Xu
Department of Computer Science and Engineering,
State University of New York at Buffalo
jinhui@buffalo.edu

Abstract

Computing accurate and robust organizational patterns of chromosome territories inside the cell nucleus is critical for understanding several fundamental genomic processes, such as co-regulation of gene activation, gene silencing, X chromosome inactivation, and abnormal chromosome rearrangement in cancer cells. The usage of advanced fluorescence labeling and image processing techniques has enabled researchers to investigate interactions of chromosome territories at large spatial resolution. The resulting high volume of generated data demands for high-throughput and automated image analysis methods. In this paper, we introduce a novel algorithmic tool for investigating association patterns of chromosome territories in a population of cells. Our method takes as input a set of graphs, one for each cell, containing information about spatial interaction of chromosome territories, and yields a single graph that contains essential information for the whole population and stands as its structural representative. We formulate this combinatorial problem as a semi-definite programming and present novel techniques to efficiently solve it. We validate our approach on both artificial and real biological data; the experimental results suggest that our approach yields a near-optimal solution, and can handle large-size datasets, which are significant improvements over existing techniques.

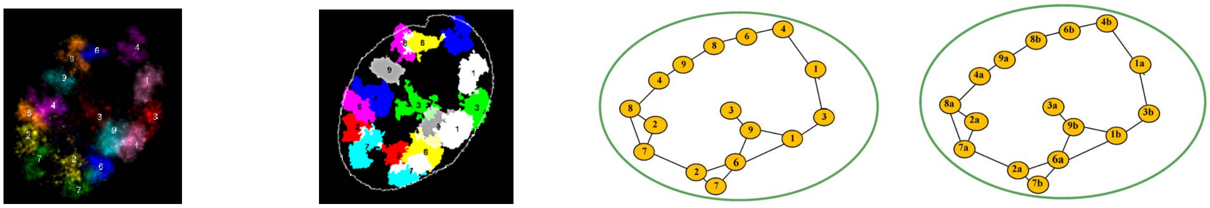
1. Introduction

Cell is the fundamental building block of living organisms. It is a highly complicated system containing the genome and the entire information required for the construction and functioning of the organism. **Chromosome Territories (CTs)** are distinct regions within the cell nucleus where genetic material is confined. Chromosome territories constitute a staple feature of nuclear architecture and are in constant dynamic interaction with other components of cell nucleus. Recent studies show influence of arrangements of chromosome territories on some fundamental cell molecular processes (*e.g.* gene expression, formation of cancer-promoting chromosome translocations *etc.*) [3, 9]. For decades, determining how chromosomes are organized inside the cellular nucleus was a technically challenging process largely relying on manual measurements and observation of experts (see Fig. 1a). Thus, obtaining accurate and robust chromosome interaction patterns from cell nucleus images will significantly facilitate analysis and improve overall understanding of molecular processes in cell nucleus [5, 6, 8, 10].

An effective way of representing high level chromosome territory organization in each cell is to use a graph data structure, where every vertex is an individual chromosome territory and every edge represents the association (spatial proximity) of a pair of neighboring chromosome territories. We call this graph as a *pair graph*. In a pair graph, each chromosome territory is associated with a chromosome number, and each chromosome number has multiplicity of two (one for each chromosome homolog). For example, in Fig. 1a, each chromosome number has two chromosome homologs which share the same color. A graph that can be seen as representative for the set of pair graphs is called a *Chromosome Association Pattern* (CAP).

Despite recent developments in microscopy imaging, and labeling techniques, our abilities to calculate CAP for a given cell population are still limited. The highly heterogeneous nature of association graphs (due to cell dynamics, imperfect synchronization, or various noise introduced in nucleus image processing) and the difficulty of identifying the corresponding chromosome homolog among cells

[*] This research was supported in part by NSF through a grant IIS-1115220.



(a) (b) (c) (d)

Figure 1: 1a shows an image obtained by superimposition along z-direction from a set of 3D microscopic nucleus images. 1b is the segmentation of 1a. 1c is the k-pair association graph for 1a. 1d is a label differentiation for 1c.

in the input complicate the problem. However, the main bottleneck lies in the inherent hardness of the CAP problem itself (since it is closely related to the subgraph matching/isomorphism problem [2] which is notoriously hard for approximation). This suggests that deriving any quality guaranteed solution for CAP is challenging.

1.1. Related Works

Several models in literature are applicable to finding chromosome territories association patterns. The most popular one is *median graph* [16], which has received considerable attentions in recent years [11, 12, 14, 15]. Given a set of input graphs with possible labels associated with vertices, the objective of a median graph problem is to compute a new graph, called median graph, that has the minimum distance to the input graphs. Distance between two graphs is usually defined as their edit distance. A variant, called Generalized Median Graph (GMG) [18], uses a generalized distance function that considers both vertex labels and edge weights, and produces near optimal solutions in polynomial time. In the context of finding association patterns of chromosome territories, GMG, as well as any other median graph method, has two major limitations. Firstly, due to its emphasis on finding the structural similarity between graphs, the GMG tends to match vertices with similar degrees. A possible outcome is that the GMG could match two vertices with different labels. While this is in accordance with the edit distance definition, it gives misleading semantic interpretation in our biological application where vertex labels denote chromosome territories. Secondly, GMG requires graphs with unique predefined labeling, this is in direct contrast with the uncertainty that we are facing in dealing with chromosome territory homologs. Recently, a suitable model based upon integer programming was proposed by Stojkovic *et al*. [20]. Although, it gives better experimental results than GMG in [18], its exponential running time (due to the nature of integer programming formulation) limits its application to only small-size input data sets.

Semi-definite programming is a well know model for solving optimization problems, especially for some NP-hard problems. Goemans and Williamson [13] introduce the breakthrough for Max-Cut problem via Semi-definite programming. Recently, Arora [1] provides a good survey on approximation algorithms and Semi-definite programming.

1.2. Our Model

Aiming to fix the aforementioned problems in existing approaches, we first enumerate 2^k labeled graphs for each k-pair association graph, with each labeled graph corresponding to a permutation of the original k-pair graph. Note that k is normally a small constant no more than 9 (this is due to the limitation of current labeling techniques in cell biology). For any two labeled graphs, we compute their Jaccard distance. Then we map the $2^k n$ labeled graphs to points in some metric space, such that the pairwise distance of any two points is equal to the Jaccard distance between the corresponding two labeled graphs. Finally, we reduce the *Chromosome Association Pattern* problem to finding a new geometric structure called *Chromatic Median* in the metric space. Although this problem looks similar to some graph optimization problems which minimize (or maximize) the sum of correlations, *e.g.,* correlation clustering [4], the major difference is that those problems do not consider the "chromatic" requirement, which is an essential requirement and a major source of difficulty for our problem.

To solve the chromatic median problem, we first give a Quadratic Integer Programming model, then relax it to a Semi-definite Programming (SDP) problem, and propose a multi-level rounding technique to solve the SDP problem. It should be pointed out that although several rounding techniques exist for SDP [1, 13], the multi-level rounding technique is new, to the best of our knowledge. The technique is naturally applicable to the Chromatic Median model. In this paper, we also show that using adaptive sampling as a speedup tool, it is possible to significantly reduce the running time and still preserving the quality of solutions.

2. Preliminaries

In this section, we introduce some definitions which will be used throughout the paper.

Definition 1 (**k-Pair Association Graph**). *Let $\mathcal{IM}$ be a nucleus image that includes k pairs of chromosomes. An unweighed graph $I = (V, E)$ is a k-pair association graph for $\mathcal{IM}$, if it satisfies*

1. *V has k pairs of vertices, $\{V_1, \cdots, V_k\}$, with each V_i containing exactly two vertices sharing the same index i.*

2. *Any two vertices from V are connect by an edge if the corresponding two chromosome territories are close*

enough to each other in 3D space (based on some biological threshold).

Figure 1c shows the k-pair association graph, where $k = 8$, for the nucleus image given in Figure 1a.

Definition 2 (**Label Differentiation**). *Given a k-pair association graph $I = (V, E)$, $V = \bigcup_{1 \leq i \leq k} V_i$, the label of each vertex pair V_i is differentiated if one vertex in V_i is labeled as $i\textcircled{a}$, and the other is labeled as $i\textcircled{b}$. The graph after Label Differentiation is denoted as $L(I)$.*

From the above definition, it is easy to see that every k-pair association graph has 2^k different label differentiations. Fig. 1d shows an example.

Definition 3 (**Jaccard Distance (JD)**). *For any two given graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ with the same labeled vertex set V, the Jaccard Distance between them is defined as*

$$JD(G_1, G_2) = 1 - \frac{|E_1 \bigcap E_2|}{|E_1 \bigcup E_2|}. \tag{1}$$

Definition 4 (**Chromosome Association Pattern (CAP)**). *Given n k-pair association graphs $\{I_1, \cdots, I_n\}$ with each corresponding to a different cell, the chromosome association pattern problem is to find a new graph $\mathcal{G}_{cap}$ and a proper label differentiation L_i for each I_i, so as to minimize the average Jaccard distance to $\{L_1(I_1), \cdots, L_n(I_n)\}$. The graph $\mathcal{G}_{cap}$ is called the CAP-graph.*

3. Chromatic Median

In this section, we consider a standalone geometric optimization problem, *Chromatic Median*, in metric (*e.g.,* Euclidean) space. In Section 4, we will show how Chromatic Median can be used to solve the CAP problem.

First, we recall that **median point** in geometry is also called *Fermat Weber point*. For any set of n points $\{p_1, p_2, \cdots, p_n\}$ in $\mathbb{R}^d$ space, its median point is defined as

$$Med = \arg\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} ||x - p_i||, \tag{2}$$

where $\arg\min$ means the value of x which minimizes the sum, and $|| \cdot ||$ denotes the Euclidean distance. There is no explicit analytical formula for computing the optimal median point. Consequently, median point is often approximated by using some iterative procedure, such as ***Weiszfeld's algorithm*** [21]. Now, we consider the Chromatic Median.

Definition 5 (**Chromatic Median (CMed)**). *Given n groups of points in $\mathbb{R}^d$ space, $\mathcal{P} = \{P_1, \cdots, P_n\}$ with each P_i containing λ points $\{p_1^i, \cdots, p_\lambda^i\}$, the Chromatic Median of $\mathcal{P}$ is defined as*

$$CMed = \arg\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} dist\{x, P_i\}, \tag{3}$$

where $dist\{x, P_i\} = \min\{||x - p_l^i|| \mid 1 \leq l \leq \lambda\}$.

From the above Definition 5, we know that the Chromatic Median of a given $\mathcal{P}$ can be obtained in two steps (see Figure 2a): Firstly, select a proper point $p_{l_i}^i$ from each $P_i \in \mathcal{P}$; secondly, compute the geometric median point of $\{p_{l_1}^1, p_{l_2}^2, \cdots, p_{l_n}^n\}$, which is the Chromatic Median of $\mathcal{P}$. Thus, the main difficulty of Chromatic Median is how to find the proper $p_{l_i}^i$ from each P_i. In Section 5, we present a semi-definite programming formulation, which helps us to identify each $p_{l_i}^i$ and yields a 2-approximation solution for Chromatic Median, where the approximation ratio is over the cost function (3).

4. Formulation: from cap to chromatic median

To solve the chromosome association pattern (CAP) problem, we use the following reduction from CAP to the Chromatic Median problem.

1. For each of the n nucleus images $\mathcal{IM}_i$, build a k-pair association graph. Let I_i denote the corresponding graph.
2. Enumerate the 2^k possible Label Differentiations for each I_i, and generate the corresponding 2^k labeled graphs $\{I_1^i, \cdots, I_{2^k}^i\}$. Note that k is usually a small number no larger than 9.
3. For any two labeled graphs $I_s^{i_1}$ and $I_t^{i_2}$, we compute the Jaccard distance between them. Since the Jaccard distance satisfies triangle inequality [7], we map the $2^k n$ labeled graphs to $2^k n$ points in some metric space. For each group of labeled graphs $\{I_1^i, \cdots, I_{2^k}^i\}$, we denote the corresponding points set as $P_i = \{p_1^i, \cdots, p_{2^k}^i\}$. Let $\mathcal{P} = \{P_1, \cdots, P_n\}$.
4. Find the n points $\{p_{l_1}^1, \cdots, p_{l_n}^n\}$ by the semi-definite programming model given in Section 5, where each $p_{l_i}^i$ is the nearest point to Chromatic Median among P_i.
5. Output the Median Graph under Jaccard Distance (see Section 4.1) for $\{I_{l_1}^1, \cdots, I_{l_n}^n\}$.

Mapping graphs into points: In Step 3 of the above reduction, we need to map all labeled graphs into points in some metric space. A natural question is how to construct the space. Our idea is not to explicitly build the metric space; instead, we formulate the Chromatic Median problem as a semi-definite programming (in Section 5) which only requires the pairwise distances among the set of mapped points (*i.e.*, the Jaccard Distances among the labeled graphs) to achieve an approximate solution.

4.1. Median Graph under Jaccard Distance

Next, we introduce *Median Graph under Jaccard Distance*, which is used in the above reduction (Step 5).

Definition 6 (jd-$\mathcal{MG}$)**.** *For a given set of **labeled** graphs $\{G_1, G_2, \cdots, G_n\}$ with all $G_i = (V, E_i)$ sharing the same*

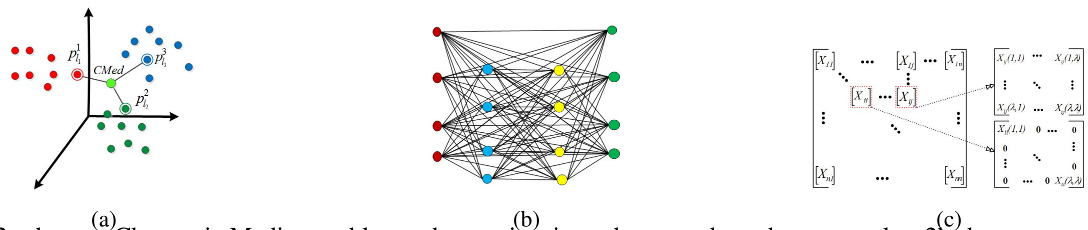


Figure 2: 2a shows a Chromatic Median problem, where points in each group share the same color. 2b shows an example of the column graph corresponding to the 0-1 Quadratic Programming, where the vertices sharing the same color belong to the same column of vertices. 2c is an illustration of the variable matrix.

labeled vertex set V, its Median Graph under Jaccard Distance, jd-$\mathcal{MG}$, is defined as $\tilde{G} = (V, \tilde{E})$, which minimizes the following cost function,

$$\frac{1}{n}\sum_{i=1}^{n} JD(\tilde{G}, G_i), \tag{4}$$

where $JD(\cdot)$ denotes the Jaccard Distance between two graphs (see Definition 3).

If we consider each G_i as one point in some metric space, and the distance in the space is defined as the Jaccard Distance, then jd-$\mathcal{MG}$ is similar to the geometric median point of the point-sets $\{G_1, G_2, \cdots, G_n\}$. Thus, we can use a method similar to Weiszfeld's algorithm to achieve an approximate jd-$\mathcal{MG}$.

For each $G_i = (V, E_i)$ (note that G_i is a labeled graph), we denote each E_i as a binary vector ν_i of length $\binom{|V|}{2}$, where each coordinate indicates the existence of the corresponding edge. That is, if the corresponding edge exists in E_i, the value of the coordinate is 1, or 0 otherwise. Compute the median point of $\{\nu_1, \nu_2, \cdots, \nu_n\}$ using Weiszfeld's algorithm (see Section 3), and denote it by $\tilde{\nu}$. Note that $\tilde{\nu}$ is not necessarily a binary vector, instead, the value of each coordinate could be fractional, *i.e.*, between 0 and 1. For any threshold $\rho \in [0, 1]$, denote the rounded binary vector by $\tilde{\nu}(\rho)$, where the value of each coordinate of $\tilde{\nu}(\rho)$ is assigned to be 1 if the corresponding coordinate value of $\tilde{\nu}$ is larger than or equal to ρ, or 0 otherwise. Correspondingly, denote the graph as $\tilde{G}(\rho) = (V, \tilde{E}(\rho))$, where the edges set $\tilde{E}(\rho)$ is indicated by $\tilde{\nu}(\rho)$. We can search the proper threshold ρ_0 from $[0, 1]$, such that $\rho_0 = \arg\min_{0\le\rho\le1} \frac{1}{n}\sum_{i=1}^{n} JD(\tilde{G}(\rho), G_i)$. Finally, output $\tilde{G}(\rho_0)$ as an approximate jd-$\mathcal{MG}$.

5. Semi-definite Model for Chromatic Median

In this section, we show that it is not necessary to find the exact chromatic median $CMed$. Instead, it is sufficient to first find the n points $\{p^1_{l_1}, \cdots, p^n_{l_n}\}$, where each $p^i_{l_i}$ is the nearest to $CMed$ among all the points in P_i, and then output the Median Graph under Jaccard Distance for $\{I^1_{l_1}, \cdots, I^n_{l_n}\}$ using the idea described in Section 4.1. Since it is hard to find the embedding space for the points [17], we do not explicitly compute $CMed$. Instead, we minimize the total pairwise distance, $\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - p^i_{l_i}||$, so as to avoid finding $CMed$. The following lemma shows the relation between the total distance to $CMed$ and the total pairwise distance.

Lemma 1.

$$1 \le \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - p^i_{l_i}||}{\sum_{i=1}^{n} ||CMed - p^i_{l_i}||} \le 2 \tag{5}$$

Proof. We prove the two sides of (5) separately.

Since $CMed$ is the geometric median point of $\{p^1_{l_1}, \cdots, p^n_{l_n}\}$, by (2), we know that for any $1 \le j \le n$,

$$\sum_{i=1}^{n} ||CMed - p^i_{l_i}|| \le \sum_{i=1}^{n} ||p^j_{l_j} - p^i_{l_i}||. \tag{6}$$

Averaging the right hand side of (6) over j, we have the left side of (5), *i.e.*, $\sum_{i=1}^{n} ||CMed - p^i_{l_i}|| \le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - p^i_{l_i}||$.

Meanwhile, by triangle inequality, for any $1 \le i, j \le n$, we have

$$||p^j_{l_j} - p^i_{l_i}|| \le ||p^j_{l_j} - CMed|| + ||p^i_{l_i} - CMed||. \tag{7}$$

Summing both sides of (7) over i and j, we have

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - p^i_{l_i}|| \\ \le &\sum_{i=1}^{n}\sum_{j=1}^{n} (||p^j_{l_j} - CMed|| + ||p^i_{l_i} - CMed||).\end{aligned} \tag{8}$$

Meanwhile, for the right side of (8), we have

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n} (||p^j_{l_j} - CMed|| + ||p^i_{l_i} - CMed||) \\ = &\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - CMed|| + \sum_{i=1}^{n}\sum_{j=1}^{n} ||p^i_{l_i} - CMed|| \\ = &\, n\sum_{j=1}^{n} ||p^j_{l_j} - CMed|| + n\sum_{i=1}^{n} ||p^i_{l_i} - CMed|| \\ = &\, 2n\sum_{i=1}^{n} ||CMed - p^i_{l_i}||.\end{aligned} \tag{9}$$

From (8) and (9), we immediately have the right side of inequality (5), *i.e.*, $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - p^i_{l_i}|| \le 2\sum_{i=1}^{n} ||CMed - p^i_{l_i}||$. This completes the proof. □

From Lemma 1, we get the following lemma.

Lemma 2. *Let $\{p^1_{l'_1}, p^2_{l'_2}, \cdots, p^n_{l'_n}\}$ be the n points minimizing the total pairwise distances with each $p^i_{l'_i} \in P_i$, and Med' be their geometric median point. Then,*

$$\sum_{i=1}^{n} ||Med' - p^i_{l'_i}|| \le 2\sum_{i=1}^{n} ||CMed - p^i_{l_i}||. \tag{10}$$

Proof. Firstly, we have the following two inequalities:

$$\sum_{i=1}^{n} ||Med' - p^i_{l'_i}|| \le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l'_j} - p^i_{l'_i}||; \tag{11}$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l'_j} - p^i_{l'_i}|| \le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - p^i_{l_i}||, \tag{12}$$

where (11) is obtained in a similar manner as the left side of (5), and (12) follows from the fact that $\{p^1_{l'_1}, p^2_{l'_2}, \cdots, p^n_{l'_n}\}$ minimize the total pairwise distances. Furthermore, by combining (11) and (12), we have $\sum_{i=1}^{n} ||Med' - p^i_{l'_i}|| \le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ||p^j_{l_j} - p^i_{l_i}||$. By combining this inequality and the right side of (5), we have (10), *i.e.*, $\sum_{i=1}^{n} ||Med' - p^i_{l'_i}|| \le 2\sum_{i=1}^{n} ||CMed - p^i_{l_i}||$. □

The above lemma suggests that minimizing the total pairwise distances enables us to obtain a 2-approximation for chromatic median. Next, we introduce a quadratic programming model for finding the desired n points which minimize the total pairwise distances.

Let $\{p^1_{l'_1}, p^2_{l'_2}, \cdots, p^n_{l'_n}\}$ be the n points minimizing the total pairwise distances with each $p^i_{l'_i} \in P_i$. We have an indicator variable x^i_j for each point p^i_j. x^i_j is equal to 1 if $j = l'_i$ (*i.e.*, $p^i_j = p^i_{l'_i}$), or 0 otherwise. For any $i_1 \ne i_2$ and any $1 \le s, t \le \lambda$, we compute the coefficient $w^{i_1,i_2}_{s,t} = ||p^{i_1}_s - p^{i_2}_t||$. We have the following 0-1 Quadratic Programming model. For ease of understanding, we can imagine that there is a **column graph** built for the model (see Figure 2b).

1. For each $i, 1 \le i \le n$, there is a corresponding column of vertices, where each indicator variable x^i_j denotes one vertex in the column.
2. For any two vertices from different columns, connect them with an edge, where the weight is $w^{i_1,i_2}_{s,t}$.
3. Finally, the 0-1 Quadratic Programming is to find a subgraph of the column graph with minimum total weight, where the subgraph contains exactly one vertex from each column.

0-1 **Quadratic Programming**

$$\min\ f = \sum_{i_1=1}^{n}\sum_{i_2=i_1+1}^{n}\sum_{s=1}^{\lambda}\sum_{t=1}^{\lambda} w^{i_1,i_2}_{s,t} x^{i_1}_s x^{i_2}_t \tag{13}$$

$$x^i_j \in \{0,1\}, \forall 1 \le i \le n, 1 \le j \le \lambda \tag{14}$$

$$\sum_{j=1}^{\lambda} x^i_j = 1, \forall 1 \le i \le n \tag{15}$$

The above 0-1 quadratic programming indicates that if we compute its optimal solution, we immediately obtain exactly one indicator variable, say $x^i_{l'_i}$, from each $\{x^i_1, x^i_2, \cdots, x^i_\lambda\}$ with value 1. Let $p^i_{l'_i}$ be the corresponding point. Then by Lemma 2, we know that the geometric median point of $\{p^1_{l'_1}, p^2_{l'_2}, \cdots, p^n_{l'_n}\}$ is a 2-approximation of Chromatic Median of $\mathcal{P}$ over the cost function (3). Thus, we immediately have the following theorem.

Theorem 1. *The optimal solution of the above 0-1 quadratic programming yields a 2-approximation for the chromatic median problem.*

5.1. Semi-definite Programming Model

Since the 0-1 quadratic programming is challenging to solve optimally [19], we convert it to a semi-definite programming model. We first build an equivalent 0-1 semi-definite programming (SDP), and then use rounding technique to solve it. There are two steps for constructing the 0-1 SDP model.

- Firstly, we build the variable matrix $X \in \mathbb{R}^{n\lambda \times n\lambda}$. For any $1 \le i_1, i_2 \le n$, we denote the sub-matrix formed by $((i_1-1)\lambda+1)$-th to $(i_1\lambda)$-th row and $((i_2-1)\lambda+1)$-th to $(i_2\lambda)$-th column as $X_{i_1i_2}$. We also denote the entry in s-th row and t-th column of $X_{i_1i_2}$, where $1 \le s, t \le \lambda$, as $X_{i_1i_2}(s,t)$. We let $X_{i_1i_2}(s,t) = x^{i_1}_s x^{i_2}_t$. Since for each $1 \le i \le n$, we just select one of $\{x^i_1, x^i_2, \cdots, x^i_\lambda\}$ to be 1, and all others to be 0, we can assign each diagonal sub-matrix $X_{i,i}$ to be diagonal matrix, and the trace to be 1. See Figure 2c.

- Secondly, we build the coefficient matrix $W \in \mathbb{R}^{n\lambda \times n\lambda}$. For any $1 \le i_1, i_2 \le n$, we denote the sub-matrix formed by $((i_1-1)\lambda+1)$-th to $(i_1\lambda)$-th row and $((i_2-1)\lambda+1)$-th to $(i_2\lambda)$-th column as $W_{i_1i_2}$. We also denote the entry in the s-th row and the t-th column of $W_{i_1i_2}$ (where $1 \le s, t \le \lambda$) as $W_{i_1i_2}(s,t)$. We let $W_{i_1i_2}(s,t) = w^{i_1,i_2}_{s,t}$.

Through the above construction for variable matrix X and coefficient matrix W, we have the following model.

0-1 **Semidefinite Programming**

$$\min\ f = tr(WX) \tag{16}$$

$$X \succeq 0, X \in \mathbb{R}^{n\lambda \times n\lambda} \tag{17}$$

(14) is rewritten as $X_{i,j}(s,t) = 0, 1$ for $\forall i, j, s, t$, and (15) is rewritten as $X_{i,i}(s,t) = 0$ for $s \ne t$, and $tr(X_{i,i}) = 1$. From Theorem 1, we immediately have the following theorem.

Theorem 2. *The optimal solution of the above 0-1 semi-definite programming is equivalent to the optimal solution of the 0-1 quadratic programming. Consequently, it yields a 2-approximation for the chromatic median problem.*

5.2. Multilevel Rounding Algorithm

If we solve the semi-definite programming on the model builded in Section 5.1, the output will be a positive semi-definite matrix $X \in \mathbb{R}^{n\lambda \times n\lambda}$. Since X is not necessarily a 0-1 matrix, to obtain a feasible solution to CAP, we need to perform a rounding procedure on X to convert it into a 0-1 matrix. An intuitive idea for this is that, for each sub-matrix $X_{i,i}$, $1 \leq i \leq n$, round the largest diagonal elements to 1, and others to 0. A main issue with this approach is that it could cause a significant loss on the quality of solution. Below, we introduce a multiple-level rounding algorithm to achieve a better solution.

Algorithm overview: Our algorithm performs $\log \lambda$ rounding steps iteratively. At each iteration, for each $X_{i,i}$, the algorithm first forms an index set, Ind_i, which corresponds to the smaller half of the nonzero diagonal elements. Then the algorithm rounds this half of diagonal elements to 0, and puts these as constraints in the next iteration. Finally, after $\log \lambda$ iterations, there is only one positive number among the diagonal elements for each $X_{i,i}$, which is the final rounding result.

Multilevel Rounding Algorithm

1. Initialize n index sets $\{Ind_1, \cdots, Ind_n\}$ with each $Ind_i = \emptyset$, and $\bar{\lambda} = \lambda$.
2. Perform the following steps $\log \lambda$ times, and then output the final X.
   (a) Solve the 0-1 SDP via some SDP solver, and output the semi-definite matrix $X \in \mathbb{R}^{n\lambda \times n\lambda}$.
   (b) For each sub-matrix $X_{i,i}$, $1 \leq i \leq n$, excluding the indices from Ind_i, find the $\frac{\bar{\lambda}}{2}$ indices that have the smallest diagonal elements. Add the indices into Ind_i.
   (c) Add the following constraints to the 0-1 SDP: For each $1 \leq i \leq n$, the elements in the j-th row and j-th column of $X_{i,i}$ are equal to zero, if $j \in Ind_i$.
   (d) Set $\bar{\lambda}$ to be $\frac{\bar{\lambda}}{2}$.

5.3. Speedup via Adaptive Sampling

Obviously, directly solving a large-size SDP is extremely computationally expensive. Thus, we introduce a speedup technique using *Adaptive Sampling*.

Main idea: We first randomly select a small number (*i.e.*, $m \ll n$) of point-sets $\mathcal{P}'$ from $\mathcal{P}$, which is denoted as $\{P_{i_1}, \cdots, P_{i_m}\}$, and then find the approximate chromatic median $CMed'$ for $\mathcal{P}'$ using the SDP introduced in Section 5.1. Using $CMed'$, for each P_i we can remove a large subset which is far away from $CMed'$. Thus, the size of $\mathcal{P}$ can be reduced significantly. Finally we get the solution for the whole $\mathcal{P}$. The key step is how to select the small sample $\mathcal{P}'$. Intuitively, the m point-sets need to be well separated in the space so as to preserve the distribution of $\mathcal{P}$. Our strategy is to select the sample adaptively: select the m point-sets iteratively; in each iteration, choose the one that has the largest distance to those already selected point-sets.

6. Evaluations

We present our experimental results in two subsections. In the first part we evaluate the performance of our method on synthetic datasets, with each dataset consisting of a number of randomly generated k-pair association graphs. In the second part, we apply our method for gauging the associations of chromosome territories in a population of cells belonging to *WI 38 human lung fibroblast cell line*. All of the experimental results are obtained on a 2.4GHz Linux workstation using SDPT3 as the SDP solver.

6.1. Synthetic Datasets

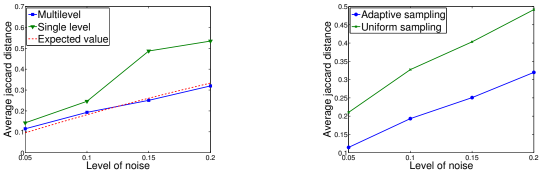
Data generation method. For each synthetic dataset, we randomly generate a k-pair association graph, $I = (V, E)$, as the ground truth. Then we generate a set of k-pair association graphs based on the ground truth with some percentage of input noise. For example, if the input noise percentage is $0 \leq p \leq 1$, we randomly delete $p|E|$ edges from E, and add $p|E|$ new edges for each graph. According to Definition 3, the expected Jaccard distance to the ground truth graph for each input graph would be $1 - \frac{1-p}{1+p} = \frac{2p}{1+p}$. Moreover, we regard it as the expected objective value for the CAP-graph obtained by our method. Throughout this section we denote the number of (chromosome territory) pairs as k, the number of input k-pair association graphs in each dataset as n, and the input noise percentage as p.

Experimental Results. In order to show that our method scales well, we measure its performance on the datasets, which have significantly larger size than the previously available techniques, *i.e.*, varying n from 100 to 1000, and k from 6 to 12. In addition, to show the robustness of our method, we test it on different levels of noise, *i.e.*, $p \in \{5\%, 10\%, 15\%, 20\%\}$. We compute the Jaccard distance between the CAP-graph generated by our method and the corresponding ground truth. For each data size, we run 10 times and take the average result. The average output results respect to different noise levels are given in Table 1 . The presented values of Jaccard distance is very close to the expected value of the objective function (*i.e.*, $\frac{2p}{1+p}$). Moreover, we note that our method performs well on larger datasets compared to [20], which has input size $n = 45$.

Table 1: Results for the multi-level rounding

Noise Level	5%	10%	15%	20%
Output	0.1146	0.1933	0.2507	0.3197
Expect	0.0952	0.1818	0.2609	0.3333

Multilevel *vs* Single Rounding. To determine the performance of the multilevel rounding technique, we show its comparison with the results obtained by single level round-



(a) (b)

Figure 3: In 3a, the three curves are the average objective value generated by multilevel rounding, the average objective value by single level rounding, and the expected value $\frac{2p}{1+p}$ respectively. In 3b, the two curves are for adaptive sampling, and uniform sampling respectively.

ing on the same input set. Figure 3a shows the result. The Jaccard distance values for multilevel rounding are significantly lower than those obtained by single level rounding.

Adaptive *vs* Uniform sampling. To speedup our algorithm, we have introduced the adaptive sampling procedure in Section 5.3. This step can be considered as a part of preprocessing phase. To evaluate the advantage of the adaptive sampling strategy, we give a comparison with the results obtained by uniform sampling. The results are shown in Figure 3b, where the results by adaptive sampling are significantly better than those by uniform sampling.

Comparison with previous results. We compare our method with the reported evaluation results for MAG [20] on synthetic datasets. In our experimental settings, we have taken datasets with ranging data size n from 100 to 1000. Also, for generating noise, we perform both insertion and deletion operations on the edges from the ground truth. On the contrary, MAG's performance is only evaluated on comparatively smaller input ($n = 45$) for the case when only deletion operation on edges is considered. In order to compare with MAG, we additionally generate 3 synthetic datasets mimicking the procedure described in [20], *i.e.*, $k = 6$, and $p \in \{15\%, 30\%, 40\%\}$, where only deletion of edges is allowed. In order to maintain consistency with [20], we use Sorenson index as the evaluation parameter. Given two labeled graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$, Sorensen index is defined as $\frac{2 \cdot |E_1 \cap E_2|}{|E_1| + |E_2|}$. Results are shown in Table 2, where our method achieves more than 10% improvements over all the three datasets, and the improvement becomes larger when the noise level increases.

Table 2: Performance comparison of MAG [20] and our method on synthetic datasets

Noise level	15%	30%	40%
Our model	0.94	0.88	0.83
MAG [20]	0.84	0.72	0.62
Improvement	11.90%	22.22%	33.87%

6.2. Biological Data

Cell nucleus image acquisition process. The usage of advanced fluorescence labeling and image processing techniques has enabled researchers to investigate interactions of chromosome territories at large spatial resolution. Current limits of microscopic image acquisition process allow at most 9 pairs of chromosome territories to be labeled per cell nucleus. In this experiment, we focus on the study of associations of 8 chromosome territory pairs (with chromosome numbers from 1 to 9, except for 5) in WI 38 human lung fibroblast cells. We run our algorithm on dataset consisting of 90 microscope nucleus images. Each raw microscope image is a set of 2-D slices, where each slice corresponds to an image of the 3-D cell nucleus acquired at a certain focal plane. Next, denoising, enhancement, segmentation (see Figure 1b), and 3-D volume reconstruction of the set of 2-D slices are done. Individual k-pair association graph for each cell is derived using nearest border to border distances among chromosome territories.

Comparison with previous results. Like many biological problems, there is no ground truth for the CAP-graph on the cell nucleus images. To alleviate this problem, we take the average similarity (or dissimilarity) between the generated CAP-graph and the input dataset as the quality evaluation criteria. In order to produce an equal-footing comparison of our method with [18] and [20], we provide three values for our output. The objective of our model is to minimize the average Jaccard distance, while [18] shows Jaccard similarity, and [20] shows Sorensen index. Table 3 shows all of the three evaluations. The last row shows the improvement of our result (if we denote the Jaccard similarity (Sorensen index) we obtain as s_1, and the Jaccard similarity (Sorensen index) obtained from [20] or [18] as s_2, then the improvement would be $(s_1 - s_2)/s_2$). From Table 3, it is clear that our method outperforms all existing approaches. Furthermore, Figure 4a shows the CAP-graph obtained by our algorithm.

Table 3: Performance comparison on biological dataset

Models	*JD*[1]	*JS* [2]	*SI* [3]
Our Model	0.65	0.35	0.52
MAG in [20]	-	-	0.46
GMG in [18]	-	0.27	0.44
Improvement	-	29.6%	13%

[1] Jaccard distance.
[2] Jaccard similarity.
[3] Sorensen index.

New association patterns discovered. For the association pattern, we compare the edges in our association pattern with those in the graph [22] generated by GMG [18] using the same dataset. Comparison shows that our method can discover all edges (or associations) yielded by GMG. Additionally, our method finds the following new association edges, where Table 4 shows them and their frequency among the dataset. We also label out these new edges in

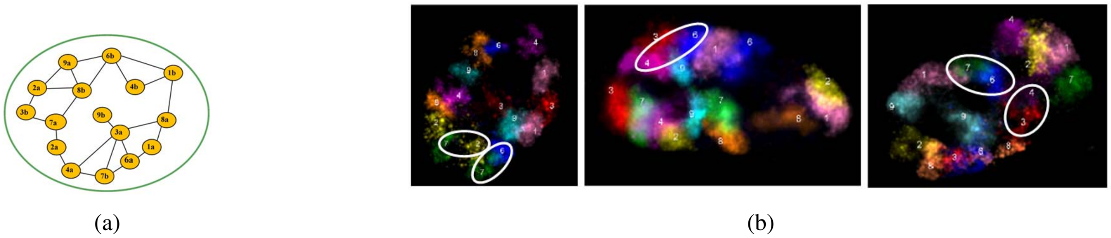


Figure 4: The CAP-graph obtained is shown in 4a. The labels for the new discovered edges are shown in 4b.

the nucleus image (Figure 4b). Note that since the nucleus images are in 3D and the association patterns may appear in different slices, we only show the associations in some slices. From Figure 4b, it is easy to see that the labeled pairs of chromosome territories are indeed close to each other, which also confirms the effectiveness of our approach.

Table 4: Additional association edges discovered by our method

Edges	$4-6$	$3-4$	$6-7$	$7-2$
Frequency	56.1%	53.7%	51.2%	51.2%

7. Summary and Discussion

In this paper we present an efficient method for recognizing the pattern of associations for chromosome territories in the cell nucleus. Our technique is able to find the CAP-graph with better quality than existing ones. We validate our approach by evaluating its performance on both synthetic and real biological datasets. Experiments using synthetic datasets reveal the scalability and high efficiency of our method, while the experiment on a cell nucleus image dataset shows the accuracy of our method for finding the association pattern of human chromosome terrirtories.

References

[1] Sanjeev Arora: Semidefinite Programming and Approximation Algorithms: A Survey. *ISAAC*, 6-9, 2011.

[2] H. A. Almohamad and S. O. Duffuaa. A linear programming approach for the weighted graph matching problem. *IEEE PAMI*, 15(5):522-525, 1993.

[3] R. Berezney. Regulating the mammalian genome: the role of nuclear architecture. *Advances in Enzyme Regulation*, 42:39-52, 2002.

[4] N. Bansal, A. Blum and S. Chawla: Correlation Clustering. *FOCS*, 238, 2002.

[5] S. Boyle, S. Gilchrist, J. Bridger, N. Mahy, J. Ellis, and W. Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.*, 10(3):211-219, 2001.

[6] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Mller, R. Eils, C. Cremer, M. Speicher, and T. Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS. Biol.*, 3(5):826-842, 2005.

[7] K. Clarkson, Nearest-Neighbor Searching and Metric Space Dimensions. *survey*, 2005

[8] J. Croft, J. Bridger, S. Boyle, P. Perry, P. Teague, W. Bickmore. Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell. Biol.*, 145(6):1119-1131, 1999.

[9] T. Cremer and M. Cremer. Chromosome territories, nuclear architecture and gene regulation in mamalian cells. *Nature Reviews Genetics*, 2:292-301, 2001.

[10] T. Cremer, M. Cremer, S. Dietzel, S. Mller, I. Solovei, and S. Fakan. Chromosome territories–a functional nuclear landscape. *Curr. Opin. Cell. Biol.*, 18(3):307-316, 2006.

[11] M. Ferrer, D. Karatzas, E. Valveny, I. Bardaj, and H. Bunke. A generic framework for median graph computation based on a recursive embedding approach. *CVIU*, 115(7):919-928, 2011

[12] M. Ferrer, E. Valveny, F. Serratosa, K. Riesen, and H. Bunke. Generalized median graph computation by means of graph embedding in vector spaces. *Pattern Recognition*, 43(4):1642-1655, 2010.

[13] M. X. Goemans and D. P. Williamson: Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *J. ACM* 42(6):1115-1145, 1995.

[14] A. Hlaoui and S. Wang. Median graph computation for graph clustering. *Soft. Comp.*, 10(1):47-53, 2006.

[15] D. Justice and A. Hero. A binary linear programming formulation of the graph edit distance. *IEEE PAMI*, 28(8):1200-1214, 2006.

[16] X. Jiang, A. Munger, and H. Bunke. On Median Graphs: Properties, Algorithms, and Applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(10):1144-1151, 2001.

[17] S. Khot, and R. Saket. Hardness of Embedding Metric Spaces of Equal Size. *APPROX-RANDOM*, 218-227, 2007.

[18] L. Mukherjee, V. Singh, J. Peng, J. Xu, M. J. Zeitz, and R. Berezney. Generalized Median Graphs: Theory and Applications. *ICCV*, 1-8, 2007.

[19] S. Sahni. Computationally Related Problems. *SIAM J. Comput.*, 3(4):262-279 , 1974.

[20] B. Stojkovic, Y. Zhu, J. Xu, A. Fritz, M. J. Zeitz, J. Vecerova, and R. Berezney. Computing Maximum Association Graph in Microscopic Nucleus Images. *MICCAI*, (2):530-537, 2010

[21] E. Weiszfeld. On the point for which the sum of the distances to n given points is minimum. *Tohoku. Math. Journal.*, 43:355-386, 1937.

[22] M. J. Zeitz, L. Mukherjee, S. Bhattacharya, J. Xu, and R. Berezney. A probabilistic model for the arrangement of a subset of human chromosome territories in wi38 human fibroblasts. *Journal of Cellular Physiology*, 221(1):120-129, 2009