Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin

Hao Wang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
hw1@cs.cmu.edu

Geet Duggal
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
geet@cs.cmu.edu

Rob Patro
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
robp@cs.cmu.edu

Michelle Girvan
Department of Physics
University of Maryland
College Park, MD 20742
girvan@umd.edu

Sridhar Hannenhalli
Department of Cell Biology & Molecular Genetics
University of Maryland
College Park, MD 20742
sridhar@umiacs.umd.edu

Carl Kingsford*
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
carlk@cs.cmu.edu

ABSTRACT

Recent chromosome conformation capture (3C) experiments produce genome-wide networks of chromatin interactions to help to study how chromosome structures relate to genomic functions. We investigate whether properties of chromatin interaction graphs based on shortest paths, maximum flows, and dense cores correlate with the spatial proximity in a three-dimensional model of the yeast genome. We demonstrate that within automatically-detected dense subgraphs, which correspond to spatially compact cores of interacting chromatin, these properties are well-correlated with spatial volume. We show that all tested methods are able to identify spatially compact sets when the test sets contain fragments from several chromosomes. We use a framework for systematically evaluating whether a method can accurately assess the spatial enrichment of a set of genomic loci for a hypothesized biological function. In such regions, we observe that the sets of fragments contained in the maximum density subgraph overlap highly with the sets of fragments in the spatially compact cores. Further, we observe that all methods agree on the spatial closeness of the yeast genomic annotations. Together, we show that compared to the more computationally complex and expensive three-dimensional embedding approach, the topological features of 3C graphs can be used to directly detect spatial closeness.

Categories and Subject Descriptors

J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics

*To whom correspondence should be addressed

BCB '13, September 22 - 25, 2013, Washington, DC, USA

Keywords

Chromosome Conformation Capture; Colocalization Test; Genome Structure

1. INTRODUCTION

Chromosome Conformation Capture (3C) [6] is a recently developed experimental method used to study the spatial structure of chromosomes by observing pairwise spatial contacts between regions of chromatin. Such experiments provide counts of observed instances of cross-linking between pairs of genomic segments flanked by restriction enzyme sites, with the interpretation that pairs of segments with high counts were often in close proximity among the population of cells assayed. The 3C technique and its subsequent refinements (4C [22, 9], 5C [2, 8, 25], Hi–C [7, 18, 23, 21], TCC [16]) have been used as tools to explore the genomic structures and features of bacteria [25], yeast [9, 23], fruit fly[21], mouse [7], and human [2, 7, 18]. These interactions have been used to verify the large-scale organization of chromatin territories [18], to investigate cancer and disease related genome alternations [11, 19], and to confirm and postulate instances of long-range regulation [2].

3C data has also been used to identify classes of genomic features that are preferentially spatially colocated in order to find biological functions for which the arrangement of DNA in the nucleus is important. A three-dimensional model of the chromatin is often computed to study the genome structure [9, 23, 2, 25, 3]. The model is usually built to satisfy as many of the 3C interactions as possible, while respecting a variety of other established properties of chromatin, including the volume constraint of the nucleus, the physical constraints of the DNA molecules, and some known biological preferences of the chromosome structure. A three-dimensional model of the chromatin structure is useful, and it can incorporate biological constraints on highly repetitive sequence regions such as centromeres and ribosomal DNAs [9] that are not available in the raw 3C data. However, requiring the computation of an embedded chromatin structure is computationally complex and expensive. In the yeast 4C experiment, for example, there are 4,053 genomic

fragments and 914,746 spatial constraints between pairs of fragments, and it takes more than one day to optimize this complex objective function [9]. Further, a large amount of uncertainty exists in the embedded structure. In order to build the distance constraints among genomic loci, the observed 3C interaction frequencies need to be mapped to distances. Different frequency-to-distance mappings will result in very different constraints, which will then lead to different structures. Moreover, the constraints from the 3C interactions are usually metrically inconsistent. Such inconsistencies are a result of both the noisy 3C data and the nature of the 3C experiment. The interactions come from an aggregate population of cells which may have different and incompatible geometries. Only a small subset of the constraints can be satisfied to some degree, and different subsets of constraints will result in different 3D embeddings [10].

Several approaches have been proposed to infer spatial colocalization of the chromatin structure without computing a chromosome embedding. Earlier methods assume that the number of observed 3C interactions given the number of possible interactions follows a hypergeometric distribution [9, 5]. Such a parametric approach, as pointed out by Witten and Noble [26], makes the inaccurate assumption that every observation of an interaction is an independent event. This assumption does not hold because 3C interactions involving the same or nearby fragments of genes are strongly correlated. To address this, Witten and Noble [26] proposed a non-parametric procedure to evaluate the spatial closeness for sets of genes by randomly resampling sets of restriction fragments as the background distribution. Kruse et al. [17] subsequently proposed a rewiring procedure, randomly shuffling the interactions while preserving the degrees of restriction fragments and the transitivity of the entire graph, to sample from background distribution. All of the methods above are based on the (sometimes weighted [23]) fraction of observed interactions (edge-fraction) between the loci of a given set of genomic features.

Here, we propose and compare a variety of topological metrics as measurements of spatial proximity without the need to compute a three-dimensional embedding of the genome (or any subset of it). Given a set of genomic loci, we explore the properties of all pairwise shortest paths, network flow, and the maximum density subgraph to evaluate spatial proximity of the chromosome structure directly from the 3C graph. Our methods are able to make better use of the observed frequencies of each interaction (edge weights) than the previously proposed edge-fraction approach. For example, our approach based on shortest paths accounts for transitivity of distance constraints and ought to reveal some information about non-observed 3C interactions due to the absence of cross-linking. The maximum density subgraph approach can extract the densest core in a graph and can reveal surprisingly compact regions masked by outliers.

We investigate the topological properties on 4C measurements in budding yeast *Saccharomyces cerevisiae* [9]. These data have been widely studied for colocalization tests on genomic features [9, 5, 26, 17, 3]. We show that all tested topological properties correlate well with the spatial proximity measured by the volume of the convex hull of the dense subgraph of a set of positions. The volume of the convex hull is computed based on an existing three-dimensional model of the yeast genome [9].

We apply our methods to both synthetic feature sets selected from the yeast three-dimensional embedding and to real genomic features [9]. To test the intuition that richer graph properties will more accurately identify spatially close sets, we introduce a new framework for systematically evaluating a method's false positive rate, true positive rate, and ability to handle outliers when estimating spatial enrichment.

We show that under a reasonable resampling scheme that controls for chromosome-specific interaction patterns, our colocalization statistical analyses are both robust and unbiased. Moreover, all methods perform well on test sets that contain a variety of chromosomes. For such sets of genomic loci, the method of finding the maximum density subgraph has the added benefit that it finds dense regions in the graph that overlap well with the true spatially compact regions. Lastly, by incorporating interaction frequencies into the tests, we find the telomere sets, which were previously thought to be significantly colocalized [9], are likely to be not. Overall, we illustrate that these proposed graph-theoretic measures can identify spatial closeness well without the need to compute an embedding and can be an alternative indicator of spatial functional enrichment.

2. FORMAL PROBLEM SPECIFICATION

Given a set F of genomic loci (representing genes or other features) and a collection of observed 3C interactions G, we would like to test whether the points in F are significantly spatially close as implied by the 3C interactions. We compute a statistic f on G and F, and we argue that statistical significance of f likely indicates statistically significant spatial proximity. This leads to the following problem:

PROBLEM 1 (SPATIAL PROXIMITY TEST). *Given a set of genomic loci F and a weighted graph $G = (V, E, d)$ of 3C interactions where V is the set of genomic segments produced by the 3C experiment, E is the set of 3C interactions, and $d(e)$ is the weight of interaction e, return **YES** if F is statistically significantly spatially close in three dimensions. Otherwise, return **NO**.*

The input loci F here is a subset of V. Input sets consisting of genes or genomic ranges should be mapped to genomic fragments first, and the test statistic is computed using the fragments. The given edge weight $d(e)$ is typically an estimate of pairwise spatial proximity derived from observed 3C interaction frequency. Problem 1 does not address the issue of outliers within the provided set F. To handle these, we introduce the following problem:

PROBLEM 2 (COMPACT CORE FINDING). *Given input as in Problem 1, return **YES** if some* subset *of F is spatially close in three-dimensions. If so, return the subset.*

3. MATERIALS AND METHODS

3.1 Yeast 3C interaction data

We use the *S. cerevisiae* 4C measurements based on the HindIII restriction enzyme library from Duan et al [9]. The 3C experimental procedure may introduce systematic biases that distort the true frequency of the data. We therefore applied the same false discovery rate (FDR) cutoff of 0.01 to pre-filter the noisy 4C data, and we applied the same frequency-to-distance mapping to convert the interaction

frequencies to distances. Other normalization methods have been proposed for Hi-C experiments [27, 14, 15]. Some are not directly applicable to 4C data because the assumptions for normalization are specific to Hi-C data [15]. Kruse et al. [17] applied the normalization method proposed by Yaffe and Tanay [27] to get an interaction probability for every fragment pair. They then chose a different FDR cutoff to filter the 3C interactions. However, Yaffe and Tanay's normalization method [27] does not take into account the circulation bias [4] that is specific to 4C experiments. Here, in order to compare directly with Duan et al. [9] and Witten and Noble [26], we use their data processing framework.

We test our methods on 3C graphs considering only the inter-chromosomal (fragments from different chromosomes) interactions. Including intra-chromosomal (fragments from the same chromosome) data ought to be beneficial since more information is incorporated for evaluating spatial enrichment. It can reveal unique spatial structure patterns of a specific chromosomes like zippering [9] and long range looping [2]. However, intra-chromosomal interactions are strongly influenced by linear genomic proximity. A high frequency intra-chromosomal interaction can be caused by genomic closeness or spatial closeness, and it is difficult to distinguish the two. Furthermore, intra-chromosomal interactions have higher frequencies than inter-chromosomal interactions, and thus can carry more weight in spatial enrichment estimation. We therefore consider only inter-chromosomal interactions following Witten and Noble [26] and Kruse et al. [17]. We validate this choice further in section 4.5.

3.2 Graph-based proxies for spatial closeness

We evaluate the following topological properties for their statistical correlation with spatial proximity. Given a set of restriction fragments $F \subset V$, we compute the following topological properties $f(F)$:

(a) $f_{\text{edge\_fraction}}(F) = \dfrac{|E(F)|}{|E_a(F)|}$.
Here, $E(F)$ is the set of observed edges with both endpoints in F. $E_a(F)$ is the set of all possible edges among the given set of nodes. If only inter-chromosomal edges are included,

$$|E_a(F)| = \sum_{\substack{i,j,i\neq j \\ i,j\in\text{chromosomes in F}}} c_i c_j,$$

where c_i is the number of fragments in F on chromosome i. $f_{\text{edge\_fraction}}$ is widely used as a proxy for spatial proximity [9, 26], however it is likely sensitive to the effects of outliers or missing measurements in the 3C experiments. This method does not use edge weights directly, but rather considers an edge present only if the false discovery rate (FDR) derived from its observed frequency is less than 0.01.

(b) $f_{\text{sp\_mean}}(F)$ = the mean of all pairwise shortest path lengths between nodes in F. The weight on each edge e here is the distance $d(e)$ computed using the frequency-to-distance mapping as employed by Duan et al. [9]. Computing shortest paths instead of using just observed edge weights addresses the issue of missing 3C interactions in connected triples and longer paths in the 3C graph. $f_{\text{sp\_mean}}$ is therefore robust to this kind of incomplete experimental data.

(c) $f_{\text{flow\_mean}}(F)$ = the mean of the maximum flow value between pairs of nodes in F. The maximum flow on an edge u, v can be thought of as the largest amount of water that can be sent from u to v by treating the edge as a pipe with the capacity of the observed 3C interaction frequency. Unlike when computing shortest paths, the edge weight here is the interaction frequency. Maximum flow avoids the problem that shortest paths can be significantly lengthened or shorted by a single edge deletion or addition. All pairwise maximum flows are efficiently computed via a Gomory-Hu tree [13].

(d) $f_{\text{max\_dense}}(F)$ = the density of the maximally dense subgraph D contained in F, where

$$\text{density} = \frac{\sum_{e\in E(D)} w(e)}{|V(D)|},$$

and $w(e)$ is the interaction frequency of edge e. The unweighted density can be computed by setting $w(e) = 1$. The definition of density used in $f_{\text{max\_dense}}$ has been widely studied and admits maximization via a polynomial-time algorithm [12]. This statistic emphasizes the importance of a compact core and helps to eliminates the effects of outliers. A density definition like (a) above, the portion of observed edges, is not applicable to maximization since there exists a trivial solution that can maximize the density: a graph with one edge.

3.3 Scheme for spatial enrichment tests

We obtain a p-value for the statistic $f(F)$ in a non-parametric manner similar to that proposed by Witten and Noble [26]. The procedure is described below:

(1) Resample a set $\mathcal{B}$ of 1000 sets of from V. How these sets are sampled depends on whether the input features are fragments, genes, or genomic ranges. In the case that they are fragments, then $|F|$ random restriction fragments are chosen. If the original input is a set of genes, then the same number of genes are randomly selected, and the selected genes are converted into fragments by choosing fragments whose midpoint lies within the selected gene. If the inputs are genomic regions, then new random starting coordinates for the regions are chosen, keeping the length of each region unchanged; we then choose those fragments whose midpoint lies within the regions. In all three cases, we keep the number of elements selected from each chromosome the same as in the input F. Such a procedure controls the fact that different chromosomes may interact with each other quite differently due to the tethered nature of the yeast genome and due to the differences in the chromosomal lengths [24].

(2) For each $B \in \mathcal{B}$ compute $f(B)$ to get a background distribution of values.

(3) Compute the empirical p-value as the fraction of examples $B \in \mathcal{B}$ where we count $f(B) \geq f(F)$ (except for shortest path, where we count $f(B) \leq f(F)$). A set is called statistically significant if this p-value ≤ 0.05.

In order to generate the background distribution, our resampling procedure randomly samples nodes in the graph while keeping the graph topology fixed. An alternative approach would be to randomly rewire the interactions of the

network by fixing the nodes of interest [17]. However, generating a set of random graphs as the null model that preserves the topological structure of the original graph without introduing any artificial bias is quite challenging. Kruse et al. [17] propose a Markov-chain procedure for reshuffling 3C edges until the rewired graph reaches or exceeds the transitivity of the original graph. This is a computationally intensive procedure since it requires many reshuffling steps to obtain a graph with transitivity comparable to that of the observed 3C graph. Further, there might exist other properties, or a combination of properties, that can better describe the topological structure of the yeast network. We thus chose the procedure above in consideration of efficiency, simplicity, and generality.

3.4 Sets of yeast genomic loci of interest

We use the genomic feature sets from Duan et al. [9]. These features include: centromeres, telomeres, breakpoints (including the ancestor breakpoints of *S. cerevisiae* and the evolutionary breakpoints between *S. cerevisiae* (Scer) and *Kluyveromyces waltii* (Kwal)), transfer RNAs (tRNAs) (including the entire tRNA set, two sub-clusters of the tRNA set, and tRNAs outside the two clusters), and origins of early and late DNA replication (including two sets with different identification mechanism). These features were chosen by Duan et al. [9] with both theoretical and experimental support of their clustering behavior.

3.5 Generation of synthetic cores

To test each method's ability to recognize truly spatially close cores masked by outliers, we generate synthetic sets of features with different sizes (20, 50 and 100 fragments) by choosing random segments on the embedded yeast chromatin structure computed by Duan et al. [9]. We sample from Duan et al.'s [9] embedding since it is built based on real 3C data, and it incorporated a variety of known biological constraints. We define a set of segments to be spatially close if they are within a diameter of 400nm, and we define them to be not spatially close otherwise. This diameter is chosen by observing that 400nm is 1/5 of the diameter of the yeast nucleus (2000nm) [9], and such a diameter results in a volume $< 1\%$ of the entire nucleus volume. Additionally, in the histogram of the pairwise Euclidean distances between beads in the Duan et al. [9] yeast embedding, only 11.7% of the distances are within 400nm.

To construct a synthetic set with a spatially close core and with some fraction of outliers, we first pick a certain percent (r_c) of the synthetic set as *core segments*. To do this, we choose a center segment $u \in V$ and then search for restriction fragments that fall within the sphere centered at that point with a radius of 200nm. All fragments inside this sphere will be at most 400nm away from each other. We define $C_a(u)$ as the set of all segments within a 200nm radius of u. We discard $C_a(u)$ if $|C_a(u)| < r_c|F|$. We then randomly pick $C(u) \subset C_a(u)$ as the set of core segments such that $|C(u)| = r_c|F|$. Secondly, we randomly choose the rest of the nodes in F outside the 200nm radius from the center point to be outliers. The reason for choosing some spatially-not-close fragments to add to the synthetic core instead of some random fragments is to enlarge the effect of the outliers.

If all sampled fragments are from the same chromosome, any method based solely on inter-chromosomal interactions will fail to detect the spatial proximity by construction, since no intra-chromosomal interactions are included. We therefore discard sets only containing fragments from a single chromosome since our test data is inter-chromosomal. (See discussion in section 4.5.)

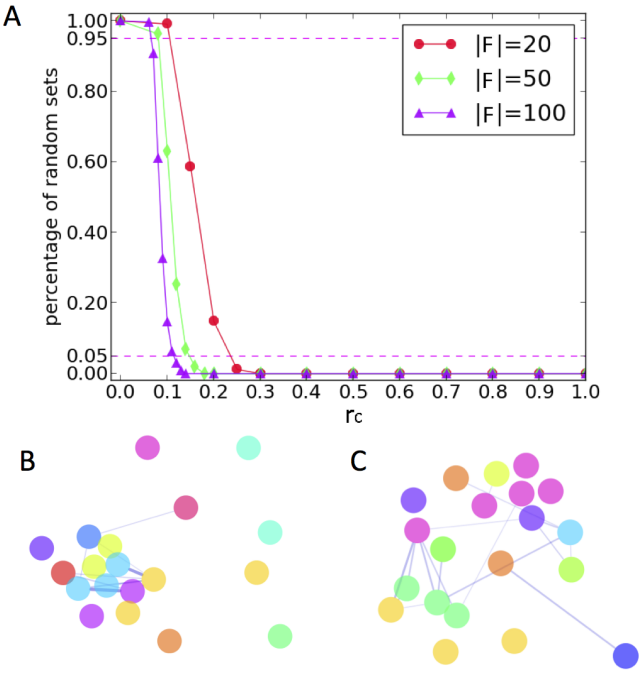

Figure 1: Defining positive examples and negative examples. (A) Cumulative distribution of 1-r_c. (r_c here is the relative size of the largest spatially-close subset of the instance.) None of the random sets contain a spatially close set with $r_c > 0.5$. (B) An example of true positive set ($r_c = 0.6, |F| = 20$). (C) An example of random set ($|F| = 20$). (B) and (C) are drawn using a spring layout. Node colors represent different chromosomes. A less transparent and wider edge represents a higher 3C interaction frequency between fragment pairs.

3.6 Constructing spatially close sets (positive examples)

In order to evaluate a statistic's power for detecting spatial enrichment, a positive example—a set of fragments that can be called significantly more spatially close than expected by chance, and a true negative—a set that cannot be called significantly compact, should be well defined.

Intuitively, a set containing a large compact core (r_c is big) is likely to be called spatially enriched as a whole. We therefore find a r_c cutoff such that a set with a core at least that large can be called significantly spatially compact. To find such a cutoff, we estimate the size of the largest set with a diameter of 400nm in 1000 randomly selected fragment sets with different set sizes ($|F| = 20, 50, 100$). This distribution is an estimation of the probability of observing a compact core of a particular size in randomly chosen samples (Figure 1A). None of the samples in the random sets

Table 1: Pearson correlation coefficient between the tested topological properties and the cubic root of the volume of the covex hull on the entire randomly sampled fragment sets and on the maximum density subgraphs within in the sets.

	$\|F\| = 20$		$\|F\| = 50$		$\|F\| = 100$	
	whole set	dense core	whole set	dense core	whole set	dense core
edge-fraction	-0.08	-0.79	-0.11	-0.88	-0.10	-0.94
average shortest path	0.05	0.79	0.03	0.91	0.07	0.96
average max flow	-0.05	-0.16	-0.09	-0.58	-0.11	-0.74
weighted density	-0.06	0.05	-0.05	-0.35	-0.02	-0.51
unweighted density	-0.07	0.41	-0.08	0.20	-0.07	0.16

contains a compact core with $r_c > 50\%$. We thus define a set of embedded fragments generated as in the previous section to be a positive example if the largest close core within the set has a size $\geq 0.5|F|$. An example of a true positive set is shown in Figure 1B. For different set sizes $|F| = 20, 50, 100$, we generate 1000 positive sets with r_c varying from 0.5 to 1.0.

3.7 Constructing negative examples

Analogous to the definition of positives, we could define the negative set with another cutoff of r_c such that a set containing a core with r_c less than some value is not significantly spatially close. However, such a filtering scheme makes the example less random. This introduces the new problem that a method that can reject this 'far' set is not necessarily capable of rejecting the true null hypothesis of random loci. Therefore we define the negative set as a set of randomly chosen fragments (or genes). An example of a random set is shown in Figure 1C.

4. RESULTS

4.1 Graph-based statistics correlate well with embedded distances on dense cores

We randomly sample 1000 sets ($|F| = 20, 50, 100$) of restriction fragments from G. As an indication of the true spatial proximity, we compute the volume of the convex hull [1]

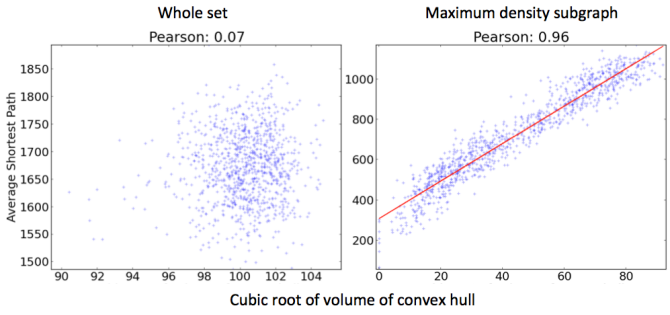

Figure 2: Scatter plot relating average shortest paths and the cubic root of volume of convex hull (nm) on the entire randomly sampled fragment sets and on the maximum density subgraph within the sets ($|F| = 100$). Average shortest path on the maximum density subgraph strongly correlates with the cubic root of the volume, while no correlation is observed on the entire set.

using the embeddings from Duan et al. [9] for every random set. We then compute the correlation between the topological properties and the true spatial proximity. We take the cubic root of the volume so that the unit of the topological properties and the unit of the true spatial proximity are on the same scale.

Although no strong correlations are observed when the properties are computed on the entire set (Table 1), if the maximum density subgraphs within the sets are found first, and the embedded distances and the topological properties are computed on these subgraphs, strong correlations appear between the embedded distance and the edge-fraction, average shortest path, and average maximum flow (Figure 2, Table 1). (Here property (d), the density is computed on the entire maximum density subgraph, not on the maximum density subgraph of the maximum density subgraph). Edge-fraction, max flow and density all inversely correlated with the embedded distance, while shortest path positively correlates with the embedded distance. Intuitively, we expect a spatially compact set to be denser, to have a larger average max flow, and to have shorter average shortest path, and we find the sets returned by maximum density subgraph to have these properties. These correlations increase as $|F|$ increases. This is because not enough 3C interactions are included in smaller sets to accurately evaluate the density and edge fraction, and noisy data has a larger effect.

The density of the maximum density subgraph has a weaker correlation compared to other tested properties. As we observe, the density grows as the graph size increases. For instance, a complete non-weighted graph of size n has a density of $(n-1)/2$. Edge-fraction, on the other hand, assigns all complete graphs of all different sizes the same score of 1. Therefore, the positive correlation between the density and the graph size weakens the correlation between the density and the embedded distance.

The weighted density correlates better with real spatial proximity than the unweighted density (Table 1). The unweighted density does not correlate as expected with spatial proximity. For set size 50 and 100, weighted density is inversely correlated with the cubic root of the volume, while the unweighted density are positively correlated with the cubic root of the volume. The inverse correlation is expected since denser regions should have a smaller volume. The correlation for both weighted density and non-weighted density are positive when set size is 20, which is probably due to the sparsity of the small graph as discussed above. These results illustrate that a more precise evaluation of the spatial proximity within set can be achieved by considering the interaction frequencies rather than just edge presence or absence as done by Witten and Noble [26].

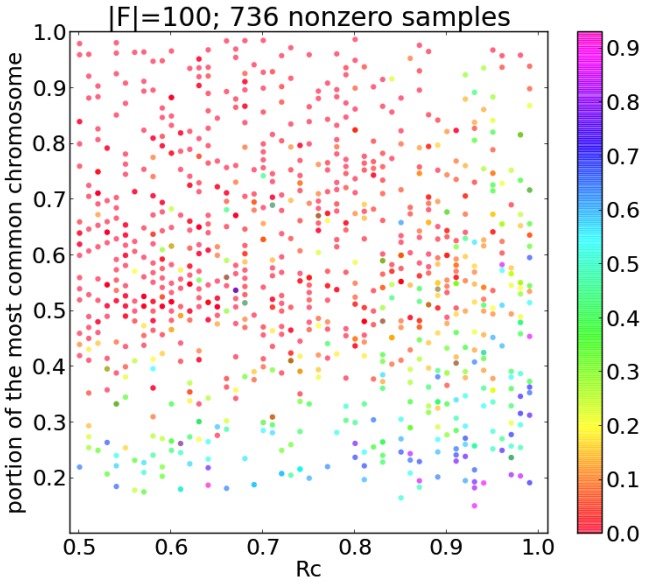

Figure 3: Scatter plot of the spatial Jaccard score between the true compact core and the dense core on the positive sets with size $|F| = 100$ (similar results observed for $|F| = 20, 50$). The x axis is r_c, y axis is the portion of the most common chromosome, and the color represents the value of the spatial Jaccard score. The spatial Jaccard score is high when r_c approaches 1 and when the test sets contains fragments from a variety of chromosomes.

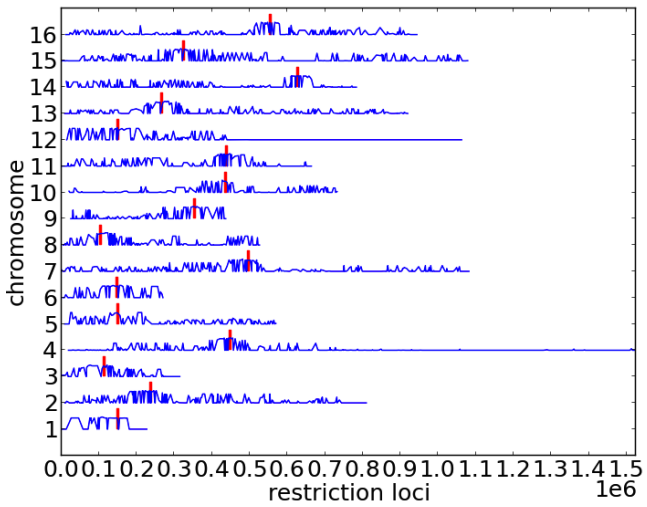

Figure 4: Maximum spatial Jaccard scores of fragments (J_f) observed in the maximum density subgraphs of all test set with $|F| = 100$. Centromere regions for different chromosomes are marked with red rectangles. J_f is high near centromere regions.

High correlations can occur if the densest subgraphs of all sets strongly overlap with each other. Under such circumstances, the conclusion that the topological properties drive the high correlation is not valid. To make sure the correlation is not caused by dense, highly overlapping regions, we compute the Jaccard similarity coefficient on nodes in maximum density subgraphs for all pair of sets. More than 99% of the pairwise Jaccard scores are less than 0.5 for all set sizes. Moreover, 99.1% pairs share zero nodes in their maximum density subgraphs when $|F| = 20$. The proportion of zero overlap is 95.3% for $|F| = 50$ and 82.4% for $|F| = 100$. This test illustrates that we have covered distinct regions on the chromosome and that the correlation between the approximate embedded distance and the tested topological properties holds in general.

The results above not only demonstrate that maximum density subgraphs correspond to spatially compact cores, but they also indicate that the tested topological properties are a good approximation for spatial proximity of these cores.

4.2 The maximum density approach identifies true compact cores

Further, to evaluate the ability of the maximum density subgraph approach to find spatially compact cores, we want to show both that the nodes inside the maximum density subgraph (dense core) overlap well with the nodes in a known true spatial compact core (true core), and that the volume of the dense core agrees well with the volume of the true core. We use a spatial Jaccard score to measure a combination of both properties. The spatial Jaccard score J_{vol} between nodes in the dense core D and nodes in the true core C is defined as:

$$J_{\text{vol}}(D, C) = \frac{\text{volume}(D \cap C)}{\text{volume}(D \cup C)},$$

where volume(X) is the volume of the convex hull of a set X of points.

We compute the spatial Jaccard score on all positive sets and observe that the spatial Jaccard score increases if the portion of the fragments from the most common chromosome decreases and if r_c increases. The score is generally between 0.5 and 0.9 when the portion of the most common chromosome is around 20% (Figure 3). This indicates that the dense core overlaps well with the true core in terms of volume when the set is not denominated by a single chromosome.

The spatial Jaccard score drops down to near zero when the portion of the most common chromosome is over 50%. The reason that maximum density subgraph cannot extract the most dense regions for such test sets is due to the absence of intra-chromosomal interactions. A detailed discussion about our reason to exclude intra-chromosomal interactions is in section 4.5.

To determine whether certain chromosomal regions correspond to dense regions that overlap well with true compact regions, we look at the maximum spatial Jaccard score of a fragment within the maximum density subgraph for every test set. Formally, every test set F contains a true compact core C_F and a maximum dense subgraph D_F; these result in a spatial Jaccard score $J_{\text{vol}}(D_F, C_F)$. Let $\mathcal{F}$ represent all test sets. For a restriction fragment $r \in \bigcup_{F \in \mathcal{F}} D_F$, the maximum spatial Jaccard score is defined as:

$$J_f(r) := \max_{F \in \{S | S \in \mathcal{F}, r \in D_S\}} J_{\text{vol}}(D_F, C_F).$$

A high J_f indicates that there exists some highly overlapping

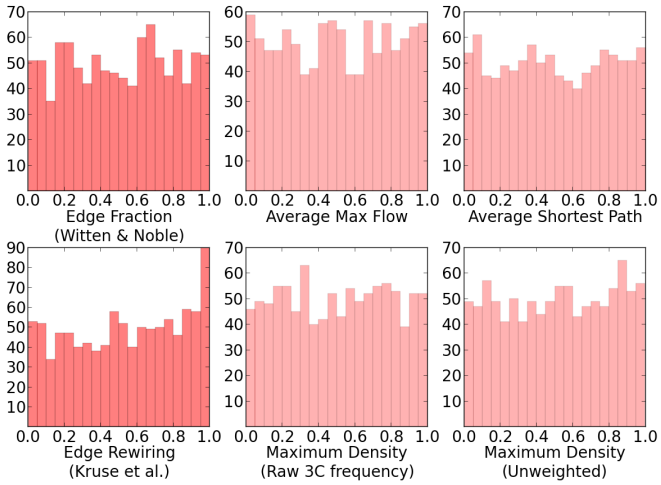

Figure 5: Histogram of p-values for different methods on random gene sets [26]. All sets of genes are chosen from a list of target genes of all known transcription factors, the size of the gene set is determined by the number of target genes of a randomly selected transcription factor. All statistics in section 3.2 achieve a near uniform distribution on this null set, indicating the methods are not biased.

cores involving this fragment. We observe fragments with a high spatial Jaccard scores often locate near centromere regions (Figure 4). More than half of the chromosomes have a average non-zero J_f scores greater than 0.4 within a 20,000 bp window of the centromere. Moreover, 100% of the non-zero J_f scores of short chromosomes (such as chromosome 1, 3 and 9) near the centromere are > 0.5, and the average non-zero J_f scores are greater than 0.6. Yet there are 432 fragments outside of the centromere regions (100,000 bp window) with $J_f > 0.5$. In summary, fragments with high Jaccard scores mainly locate near centromere regions, but can be also found in other areas on the chromosome.

4.3 An unbiased null hypothesis

When tested on randomly generated gene sets containing randomly selected genes without any functional relationships or colocalized properties [26], all tested topological statistics produced a uniform p-value distribution (Figure 5). Evaluating the p-value distribution on the null sets is a standard approach to check whether a statistic has good control for type I error [26, 17, 20]. A uniform distribution of p-values is expected if the statistic is valid. Similar results are observed on sets of randomly chosen fragments.

4.4 Topological properties as spatial proxies for spatial enrichment test

We test the power of each topological property as spatial proxies for evaluating spatial enrichment of a given set on the positive sets with different set sizes. We observe all methods correctly call compact cores significant when the portion of fragments from the most common chromosome is less than 30% (Figure 6). All methods except for edge-rewiring achieve a true positive rate of 100% when the portion of fragments from the most common chromosome is less than 20%, and a positive rate above 80% when the portion of the most common chromosome is less than 30%. Similar to the result of the dense core overlap, the spatial enrichment evaluation is more accurate when sets of fragments are from several chromosomes.

The true positive rates of all methods decrease when the portion of the most common chromosome increases, and it reaches a low level when the majority of the test set consists of fragments from the same chromosome. If most of the fragments are from one or two chromosomes, there will be very few inter-interactions among the set to accurately estimate whether the set is more spatially close than expected by chance.

The edge rewiring method proposed by Kruse et al. [17] is the most conservative among all tested statistics: The true positive rate on $|F| = 20$ is below 20%. This is probably because edge rewiring controls for the global transitivity of the entire graph in the random rewiring procedure, while the transitivity in local subgraphs in yeast might vary radically.

4.5 Rationale for including only inter-chromosomal edges

In line with previous studies [9, 17], our tests are on the set of inter-chromosomal edges. Ideally, both intra-chromosomal and inter-chromosomal edges would be used. However, testing for spatial enrichment including intra-chromosomal interactions remains a challenge.

First, a cutoff of 400nm is potentially unsuitable for a primarily intra-chromosomal set of fragments. Based on the embedding, more than 50% of the intra-chromosomal distances are less than 400nm, while less than 10% of the inter-chromosomal distances are less than 400nm. Thus 400nm is not a 'surprisingly close' cutoff for intra-chromosomal distances. However, it is difficult to find a distance cutoff that captures sets that contains both significantly spatially close intra-chromosomal and inter-chromosomal structures.

Second, the close spatial proximity between intra-chromosomal pairs is due in large part to the genomic proximity between these pairs of loci. Distinguishing spatially close sets caused by polymer packing from otherwise more interesting close sets such as fragments involving long range loops is not straightforward.

Finally, inter-chromosomal and intra-chromosomal interaction frequencies are not on the same scale: intra-chromosomal interactions have a much higher expected frequency than inter-chromosomal interactions. It is thus necessary to place inter-chromosomal interactions and intra-chromosomal interactions on the same scale, and to place intra-chromosomal interactions with different genomic distances on the same scale. One such approach is to set q-values as the edge weights, where a null model has already taken genomic proximity into consideration. Another approach is to set the edge weight as z-scores conditioned on different genomic distances [20]. Experiments run on the current test sets do not provide evidence that these approaches perform well in estimating spatial enrichment. Methods of averaged shortest path and average maximum flow only achieve a true positive rate of $< 10\%$ when the edge weights are either q-values or z-scores. It is possible that these approaches are too conservative and are not sensitive enough to detect spatially enrichment. It is also possible that since the embedding is computed on the constraints based on frequencies (mapped to

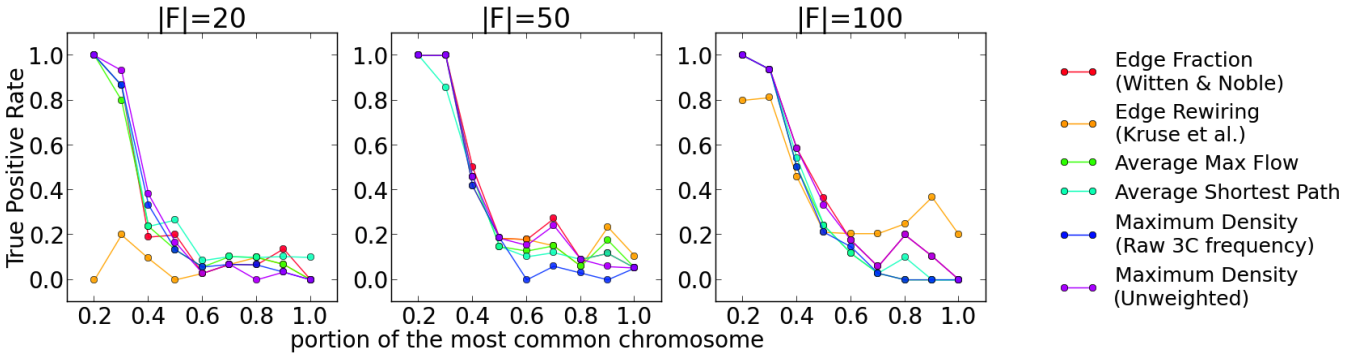

Figure 6: True positive rate of different methods on positive examples with different set sizes ($|F| = 20, 50, 100$). The x axis is the portion of the most common chromosome. The true positive rate is high when the fragments of the test sets are from different chromosomes, and is low when the fragments are mainly from the same chromosome. Edge rewiring is the most conservative method.

Table 2: P-values (before Bonferroni correction) of different methods on the yeast feature sets from Duan et al. [9] Numbers marked in red bold are p-values considered significant. Asterisks after the numbers indicate significance after Bonferroni correction.

features	edge-fraction	mean shortest path	mean flow	unweighted maximum density subgraph	weighted maximum density subgraph
centromeres	**0.00E+00***	**0.00E+00***	**2.67E-02**	**0.00E+00***	**0.00E+00***
telomeres all	**1.23E-02**	8.56E-01	1.00E+00	9.97E-01	9.72E-01
early firing CIB5-independent origins	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***	**2.00E-03***
late firing CIB5-dependent origins	1.66E-01	7.72E-01	8.42E-01	3.38E-01	6.30E-01
early firing Rad53-regulated origins	**3.80E-03***	**1.20E-02**	5.40E-02	**2.00E-03***	**4.00E-03***
late firing Rad53-regulated orgins	4.04E-01	5.48E-01	6.96E-01	6.90E-01	8.40E-01
breakpoints (Scer)	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***	**2.00E-03***
breakpoints (Scer and Kwal)	**3.80E-02**	1.2E-01	5.40E-02	**4.00E-03***	**3.20E-02**
trnas	**2.00E-03***	**1.60E-02**	**2.20E-02**	**1.00E-02**	**4.00E-02**
trna cluster bright	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***
trna cluster dim	**0.00E+00***	**6.00E-03**	**1.00E-02**	**3.40E-02**	2.34E-01
trna cluster other	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E00

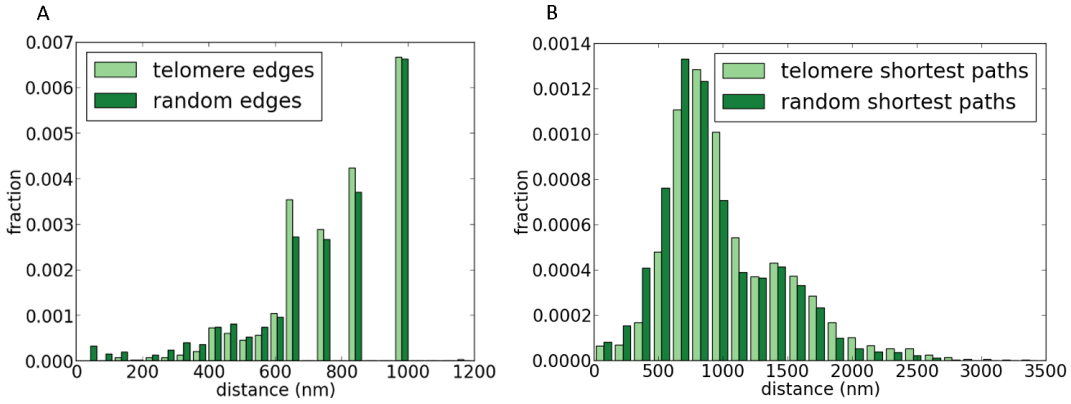

Figure 7: (A) Edge lengths and (B) pairwise shortest-path distribution between points in the telomeres set. Interactions among telomeres are overall low-frequency (long-distance) interactions and the distribution of the neither $d(e)$ nor shortest paths are significantly different from a randomly generated set.

distances), not on z-scores or q-values, setting edge weights as raw frequencies will thus perform better by construction.

We thus construct the 3C graph by only including inter-chromosomal interactions. Good estimations are achieved in sets containing fragments from a variety of chromosomes. As mentioned, such regions often locate near centromeres (Figure 4). These regions are known to be spatially compact and clustered around the spindle pole body (SPB) with multiple chromosomes [28]. The fact that our tests perform well in these cases helps to validate that the sets we identify as spatially close correspond to truly spatially close sets.

4.6 Evaluation of the spatial closeness of various yeast feature sets

For the yeast feature sets [9], most statistics tested here agree with the results presented by Witten and Noble [26]. The edge-fraction method using only the inter-chromosomal interaction data is equivalent to the resampling method proposed by Witten and Noble [26]. Before Bonferroni correction, this test finds that all tested features except the two sets of late-firing origins (late firing CIB5-dependent origins, late firing Rad53-regulated origins) and the tRNA outside two clusters (trna cluster other) are statistically co-located. Among all the methods tested, the non-weighted maximum density approach agreed with the edge-fraction approach in the most instances (Table 2). This is not surprising as these are the methods that ignore edge weights. On the other hand, the mean flow statistic is more conservative.

Witten and Noble [26] correctly identified the telomere set as not spatially close only after Bonferroni correction. In contrast, the other methods do not rely on multiple hypothesis correction to get the correct answer, and the p-values from the other methods are all close to 1.0. Telomeres tend to form five to eight foci inside the nucleus during interphase [9]. However, most interactions within the telomere set are low-frequency interactions. Thus edges of the subgraph of the telomere set are long-distance edges, and the distribution of the pairwise shortest paths is not significantly smaller when compared to a random set (Figure 7). The fact that $f_{\text{edge\_fraction}}$ cannot exploit edge weights may have lead to a false indication that this feature is statistically significantly colocalized.

tRNAs are also observed to have clustering behavior in the nucleolus [9]. Duan et al. [9] found two clusters of tRNAs with 3C interaction data: one colocalized with centromeres (trna cluster bright), and the other colocalized with rDNAs (trna cluster dim). Although the trna cluster dim is considered significantly spatially enriched by the method of edge-fraction, the weighted maximum density subgraph does not identify it as spatially close. Again, it is plausible that they are not significantly colocalized. The interactions between points in this set are of lower frequency and thus the cluster appears to be 'dim' in the heat map. Taking into account the edge frequencies, as in the weighted maximum density subgraph approach, might lead to a more accurate estimate.

5. CONCLUSIONS AND FUTURE WORK

We proposed several novel topological properties as proxies for testing for spatially compact regions of chromatin. These methods avoid the costly process of computing a 3D embedding. The shortest path and the maximum flow approach implicitly apply inferred information from the 3C graph, while the maximum density subgraph approach reduces the effect of outliers. The topological properties we chose here are directly related to spatial proximity of the 3C structures and are easy to compute. Alternative properties of the 3C graph could result in an equally good or better estimation of spatial proximity. One particularly interesting approach for future work is to measure proximity via a diffusion process on the 3C graph, providing robust estimates of node proximity in the graph.

We illustrate that the tested topological properties can be used to infer true spatial proximities in the chromosome structure by first showing in section 4.1 that graphical proximities within dense cores are strongly correlated with proximities in their corresponding embeddings. We then show in section 4.2 that dense regions found by the maximum density subgraph overlap well with true spatial compact cores when the test sets contains fragments from several chromosomes.

To evaluate the power of the graphical properties for testing spatial enrichment, we first show in section 4.3 that all methods result a uniform p-value distribution on the true negative sets and are thus unbiased and valid. We then systematically evaluated the performance of all methods by testing them on both synthetic test sets (section 4.4) and yeast feature sets (section 4.6). We have shown that Problem 1 (Spatial Proximity Test) can be solved equally well by many statistics based on different topological properties when test sets involve fragments from several chromosomes. We also demonstrate that, under such circumstances, the weighted maximum density method is a good solution to both Problem 1 and Problem 2 (Compact Core Finding) since the cores it finds overlap significantly with synthetically generated cores. The maximum density subgraph solves a slightly different problem from the other methods and previous approaches. It finds the densest subset of a given set of fragments, while the other methods evaluate the given set as a whole.

While the framework for using topological properties to infer spatial enrichment is generally effective, the proposed methods cannot accurately evaluate the spatial enrichment in regions that fragments mainly come from the same chromosome due to the absence of the intra-chromosomal interactions when constructing the 3C graph. We discussed in section 4.5 the challenge of including intra-chromosomal interactions. A more comprehensive test set that includes single chromosomal examples that cannot be simply explained by genomic proximity is yet to be developed.

Overall, we show that incorporating richer topological features such as flow, shortest path, and maximum density subgraphs provides insight into finding regions that are truly spatially enriched when the 3C graph contains sufficient interactions. These topological features can be efficiently computed using well-known graph algorithms.

6. ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation [CCF-1053918, EF-0849899, and IIS-0812111], National Institutes of Health [1R21AI085376 and 1R21HG006913], and a University of Maryland Institute for Advanced Studies New Frontiers Award. C.K. received support as an Alfred P. Sloan Research Fellow. We thank Daniela Witten and Kai Kruse for providing their codes for comparison.

7. REFERENCES

[1] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[2] D. Baù et al. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, 18(1):107–114, 2010.

[3] S. Ben-Elazar, Z. Yakhini, and I. Yanai. Spatial localization of co-regulated genes exceeds genomic gene clustering in the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.*, 41(4):2191–2201, 2013.

[4] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13(1):436, 2012.

[5] Z. Dai and X. Dai. Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Res.*, 40(1):27–36, 2012.

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

[7] 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.

[8] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome conformation capture carbon copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16(10):1299–1309, 2006.

[9] Z. Duan et al. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, 2010.

[10] G. Duggal, R. Patro, E. Sefer, H. Wang, D. Filippova, S. Khuller, and C. Kingsford. Resolving spatial inconsistencies in chromosome conformation measurements. *Alg. Mol. Biol.*, 8(1):8, 2013.

[11] G. Fudenberg, G. Getz, M. Meyerson, and L. A. Mirny. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.*, 29(12):1109–1113, 2011.

[12] A. V. Goldberg. Finding a maximum density subgraph. Technical report, CSD-84-171, Berkeley, CA, USA, 1984.

[13] R. E. Gomory and T. C. Hu. Multi-terminal network flows. *J. Soc. Ind. Appl. Math.*, 9(4):pp. 551–570, 1961.

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

[15] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, 2012.

[16] R. Kalhor et al. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, 30(1):90–98, 2012.

[17] K. Kruse, S. Sewitz, and M. M. Babu. A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. *Nucleic Acids Res.*, 41(2):701–710, 2013.

[18] 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.

[19] R. P. McCord, A. Nazario-Toole, H. Zhang, P. S. Chines, Y. Zhan, M. R. Erdos, F. S. Collins, J. Dekker, and K. Cao. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford Progeria syndrome. *Genome Res.*, 23(2):260–269, 2013.

[20] J. Paulsen, T. G. Lien, G. K. Sandve, L. Holden, O. Borgan, I. K. Glad, and E. Hovig. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Res.*, 2013.

[21] T. Sexton et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3):458–472, 2012.

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

[23] H. Tanizawa et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nuc. Acids Res.*, 38(22):8164–8177, 2010.

[24] H. Tjong, K. Gong, L. Chen, and F. Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res.*, 22(7):1295–1305, 2012.

[25] M. A. Umbarger et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell*, 44(2):252–264, 2011.

[26] D. M. Witten and W. S. Noble. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.*, 40(9):3849–3855, 2012.

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

[28] C. Zimmer and E. Fabre. Principles of chromosomal organization: lessons from yeast. *J. Cell Biol.*, 192(5):723–733, 2011.