Sefer *Appl Netw Sci* (2021) 6:54
https://doi.org/10.1007/s41109-021-00396-1

Applied Network Science

RESEARCH Open Access

Hi–C interaction graph analysis reveals the impact of histone modifications in chromatin shape

Emre Sefer*

*Correspondence:
emre.sefer@ozyegin.edu.tr
Computer Science Department, Ozyegin University, Nişantepe, Orman Sk., 34794 Istanbul, Turkey

Abstract

Chromosome conformation capture experiments such as Hi–C map the three-dimensional spatial organization of genomes in a genome-wide scale. Even though Hi–C interactions are not biased towards any of the histone modifications, previous analysis has revealed denser interactions around many histone modifications. Nevertheless, simultaneous effects of these modifications in Hi–C interaction graph have not been fully characterized yet, limiting our understanding of genome shape. Here, we propose ChromatinCoverage and its extension TemporalPrizeCoverage methods to decompose Hi–C interaction graph in terms of known histone modifications. Both methods are based on set multicover with pairs, where each Hi–C interaction is tried to be covered by histone modification pairs. We find 4 histone modifications H3K4me1, H3K4me3, H3K9me3, H3K27ac to be significantly predictive of most Hi–C interactions across species, cell types and cell cycles. The proposed methods are quite effective in predicting Hi–C interactions and topologically-associated domains in one species, given it is trained on another species or cell types. Overall, our findings reveal the impact of subset of histone modifications in chromatin shape via Hi–C interaction graph.

Keywords: Hi–C, Set cover, Bioinformatics, Algorithms, Epigenetics

Introduction

Graph theory has emerged as a robust and competent tool to quantify the connectivity in complex systems (Bullmore and Sporns 2009). Networks can be used to represent the underlying structure of social, physical, and biological systems, which are graphs made up nodes connected by edges. Biological interactions at many different levels of detail can be modeled as networks, ranging from the genomic interactions in a folded genome structure to the relationship of organisms in an ecosystem. A complex system exhibiting network structure is chromatin interaction network, which models higher-order folding of chromatin in the three-dimensional nucleus. Here, we focus on chromatin interaction networks which we define as a set of nodes representing restriction fragments or genome regions, and a set of undirected edges representing the physical interactions between these locations (Babaei et al. 2015).

Chromatin interactions captured via a number of recent chromosome conformation capture experiment methods such as Hi–C (Lieberman-Aiden et al. 2009) have resulted in significant progress in our understanding of the chromatin structure geometry (Rao et al. 2014). Hi–C experiments provide us genome-wide chromatin interactions and contact frequencies, a measure for how often any given pair of loci are adequately close in space to be captured together. As a result, Hi–C yields count matrices representing the cross-linking frequency between DNA restriction fragments at a certain resolution. At a higher level, Hi–C has reported a spatial partition of open and closed chromatin into A and B compartments respectively at the chromosomal level (Dekker et al. 2013). B and A compartments broadly correlate with transcriptionally silent, compacted heterochromatin and transcriptionally active, accessible euchromatin, respectively. Similarly, analysis of the resulting matrix (Dixon et al. 2012) at a higher resolution resulted in the discovery of topologically-associated domains corresponding to highly-interacting, consecutive matrix regions that are close in 3D packed chromatin. Topologically-associated domains (TADs) are ubiquitous unit of genome organization that are highly reproducible features of Hi–C matrices. Higher-order genome organization (including TADs, compartments) is correlated with cell differentiation and long-range regulation of transcription (Nora et al. 2013; Phillips-Cremins et al. 2013).

The emerging evidence has revealed the importance of epigenetics in understanding the basic cellular and molecular mechanisms that take place in chromatin (splicing, transcription, DNA repair and replication) (Gibney and Nolan 2010; Dabin et al. 2016). Even though Hi–C is not biased towards any of the histone modifications, previous analysis has revealed denser interactions around many histone modifications (Dixon et al. 2012; Filippova et al. 2014). Interactions between these one-dimensional histone modifications determine the 3D structure of genome. As an example, insulator proteins, modifications H3K27ac, and H3K4me3 are enriched, H3K27me3 is depleted inside TAD boundaries (Emre et al. 2016), even though these associations' casual direction is not known. Despite such results, the main picture of how histone modifications through their distribution in genome jointly affect 3D genome shape remains poorly understood across species, cell types, and cell cycles. This is partially because the previous analyses relating histone modifications to TADs and A/B compartments have often considered each histone modification independently, without accounting for their combined quantitative effects. It is not fully known to what degree relationships between the histone modifications are important across species and cell types, or whether there is a small set of histone modifications that are of primary importance in explaining observed Hi–C interactions, and thus 3D genome shape.

In this paper, we consider the problem of identifying the relationships between high-order chromatin interactions and histone modifications. Concretely, we aim to understand and predict how Hi–C interactions are formed as a result of these modifications and interactions within them. We propose covering type methods ChromatinCoverage and TemporalPrizeCoverage to decompose Hi–C interaction graph in terms of known histone modifications. Both ChromatinCoverage and its temporal prize collecting variant TemporalPrizeCoverage selects subset of histone modifications based on set multicover with pairs, where each Hi–C interaction is covered by histone modification pairs. We systematically identify 4 histone modifications (H3K4me1, H3K4me3,

H3K9me3, H3K27ac) to be highly predictive of most Hi–C interactions across species and cell types when considered in combination. We complete the missing Hi–C interactions and predict inter and intra-chromosomal Hi–C interactions at a high resolution by using this sparse set of inferred modifications. These histone modifications explain a major proportion of the accuracy of Hi–C prediction, matching with their known roles, which fail to predict Hi–C interaction when considered independently. We show that these modifications are conserved across human and mouse species, as well as embryonic stem cells and GM12878 cells.

Overall, our contributions are as follows: (1) We propose novel covering type formulations to identify subset of histone modifications over Hi–C interaction graph across genome locations and cell cycles, (2) Then, we propose efficient relaxation-based methods with provable optimal guarantees, (3) We show that most of the identified histone modifications exist consistently across different mammals, cell types and cell cycles, (4) We demonstrate the effectiveness of identified histone modifications in predicting Hi–C interactions and TADs. Although Schreiber et al. (2019) discusses the low performance of biological data prediction across cell types, our method's performance across cell types is quite promising.

Related work

Previous research has focused on understanding the subset of genome architectures through epigenetic modifications by ignoring the interactions between modifications. Rao et al. (2014) analyzed the distribution of various genomic elements such as histone modifications, CTCF, enhancers in terms of Hi–C interactions. Another work (Hughes et al. 2014) shows how Hi–C is distributed around regulatory sequences. Mifsud et al. (2015) discusses the degree of overlap between Hi–C interactions and the known promoter and enhancer sites. Among predictive tasks, Al Bkhetan and Plewczynski (2018) predicts 3D chromatin looping interactions within TADs from epigenomics and transcription factor profiles using statistical learning. Ashoor et al. (2020) predicts genomic sub-compartments from Hi–C chromatin interaction data by unsupervised graph embedding. Libbrecht et al. (2015) and Sefer and Kingsford (2019) has considered the impact of epigenetic modifications in predicting TADs, which is different than predicting Hi–C interactions.

Another set of work has focused on analyzing epigenetic data by deep non-generative models lacking the high-quality interpretation of the relationships. Di Pierro et al. (2017) uses a neural network-based algorithm to predict subcompartment annotations from epigenetic modifications. Li et al. (2019) proposes a bootstrapping deep learning model that predicts interactions only between regulatory elements without utilizing histone modifications. Similarly, Trieu et al. (2020) proposes a deep learning approach to predict the impact of only non-coding sequence variants on 3D chromatin structure. In common, all these methods use deep neural networks which offer little explanation on the relation between model inputs and model output. They also do not model the relationships by a generative framework, limiting the interpretability of the relationships between epigenetic modifications and Hi–C data.

There are a number of differences between our work and the existing work: (1) Some of these methods consider each histone modification independently ignoring the global

dynamics between the modifications, (2) They do not develop explanatory models of Hi–C interactions in terms of histone modifications, so they lack interpretability of these relationships, and (3) They do not quantify the strength of relationships between histone modifications in the Hi–C interaction dataset so existing methods cannot identify the most important subset of histone modifications.

Methods

Problem formulation

Hi–C provides us set of interactions between restriction fragments over the whole genome. More formally, let R be the set of restriction sites over considered genome, and Hi–C provides us an undirected interaction graph $G = (V = R, E)$ where $E = \{E_{uv},\ u < v \in R^2\}$ is set of interactions between restriction sites and E_{uv} is the number of interactions between u and v. These interactions can be analyzed in two ways: (1) We can either work directly at a restriction fragment level where each node is a restriction site and G is an unweighted graph, or (2) We bin the data at a given resolution and analyze the resulting graph $G^{'} = (V^{'} = R^{'}, E^{'})$ where $R^{'}$ represents nonoverlapping genomic regions of fixed length (called a bin), and each edge $E^{'}_{uv}$ is the total number of interactions between restriction sites of bins u and v.

Let M be set of histone modifications that are candidates to explain observed Hi–C interactions and associated biases. Histone modifications are previously shown to be associated with several Hi–C interaction patterns (Dixon et al. 2012). We define c^v_m to be the number of histone modification $m \in M$ around restriction site v which can take binary values if the data is not binned; modification m either exists or not around v. Let $H[v] = \{(m, c^v_m),\ | m \in M, c^v_m > 0\}$ be set of histone modification counts around restriction site v. When analyzed after binning, $H[v^{'}] = \{(m, \sum_{k=1}^{t} c^k_m),\ | m \in M, \sum_{k=1}^{t} c^k_m > 0\}$ where bin $v^{'} = \{v_1, v_2, \ldots, v_t\} \in R^{'}$ includes t restriction sites. Given $H = \{H[v],\ v \in R\}$ (or $H = \{H[v^{'}],\ v^{'} \in R^{'}\}$ if the data is binned), we propose the following problem to identify subset of modifications in terms of which Hi–C data can be explained:

Problem 1 ChromatinCoverage: *Given histone modifications data H and Hi–C interaction graph G over a genome, we infer the minimum weighted set of histone modifications any pair of which can cover all observed Hi–C interactions.*

Problem where data is binned at a given resolution is defined similarly. ChromatinCoverage identifies subset of histone modifications that can cover Hi–C interactions to explain 3D genome shape by taking interacting and non-interacting genomic regions into account. We also propose a variant of ChromatinCoverage: TemporalPrizeCoverage to find consistent spatio-temporal markers by possibly partially covering Hi–C interactions.

ChromatinCoverage: covering chromatin interactions by subset of histone modification pairs

We propose a covering type solution to select a subset of modifications to explain observed Hi–C interactions between restriction sites. We assume each Hi–C interaction to be covered by at least one modification pair. Similar covering problems have been studied in

primer selection and haplotyping (Halldórsson et al. 2004), [24]. Let x_{mn} be a binary variable taking value 1 when modification m interacts with modification n, and let y_m be a binary variable taking value 1 when modification m is in the solution. Without loss of generality, we assume each restriction site to have at least a single modification. Otherwise, we remove restriction sites to which there is no mapped modification. We assume that two histone modifications can interact only when both modifications individually belong to the solution. The resulting Program (1)–(5) is defined as follows:

$$\text{argmin}_{\text{Y}} \sum_{\text{m}\in\text{M}} \text{w}_\text{m}\text{y}_\text{m} \tag{1}$$

$$\text{s.t.} \sum_{(m,c_m^u)\in H[u]} \left(\sum_{(n,c_n^v)\in H[v]} x_{mn} \right) \geq 1, \quad (u,v) \in E \tag{2}$$

$$\begin{aligned} &x_{mn} \leq y_m, \\ &x_{mn} \leq y_n, \quad m \leq n \in M^2 \end{aligned} \tag{3}$$

$$x_{mn} \geq 0, \quad m \leq n \in M^2 \tag{4}$$

$$y_m \geq 0, \quad m \in M \tag{5}$$

where w_m is the cost of adding modification m to the solution. We define w_m as

$$w_m = \frac{\sum_{(u,v)\notin E} \min(c_m^u, c_m^v)}{\binom{R}{2} - |E|} \tag{6}$$

which increases if m exists highly across non-interacting restriction sites. This heuristic weighting scheme penalizes the modifications that are also seen across non-interacting sites which cannot be penalized by the unweighted problem formulation. Constraint (2) ensures that each interaction is covered by at least one modification pair existing in the corresponding sites, and constraint (3) ensures that a modification pair can cover an interaction only when both histone modifications belong to the solution. Since interactions are independent sets of modification pairs, we can replace constraints (3) by the following stronger set of constraints:

$$\sum_{(n,c_n^u)\in H[u]} x_{mn} + \sum_{(n,c_n^v)\in H[v]} x_{mn} \leq y_m, \quad m \in M,\ (u,v) \in E \tag{7}$$

Let $Q = \max_{(u,v)\in E}(|H[u]|\,|H[v]|)$ be maximum size of histone modification pairs that can cover an interaction, ChromatinCoverage is NP-hard, and Program (1)–(5) with the replaced constraint can be approximated by $O(\sqrt{Q\log(|E|)})$ as in Theorem 1 which follows from approximation-preserving reduction to *Minimum Weight Multicolored Subgraph Problem* (MWMCSP) ((Hajiaghayi et al. 2006). This is achieved by solving its LP relaxation and running a randomized rounding, adding each modification m to the solution with probability y_m. If constraints are still not satisfied after rounding, we keep

adding y_m with the maximum number of satisfied constraints increase per unit cost (w_m) until solution is satisfied. Problem is of polynomial size in the order of variables and constraints; the number of variables is $\frac{M(M+1)}{2} + M$, number of constraints is $E + ME$.

Theorem 1 *ChromatinCoverage can be approximated by* $O(\sqrt{Q\log(|E|)})$.

Proof

MWMCSP instance: Given an undirected graph $G_M = (V_M, E_M)$ *with a color function that assigns to each edge one or more of n given colors and non-negative vertex weights as input, the aim is to find a set of vertices of* G_M *with minimum weight inducing edges of all n colors. U is color universe where* $\chi = (\chi_1, \ldots, \chi_n)$ *is the family of nonempty "color classes' of edges (without loss of generality we assume that* $\cup_i \chi_i = U$*).*

When mapping ChromatinCoverage into MWMCSP instance, histone modifications to be selected M maps to V_M. Each edge in E_M defines a color class $\chi_{m,n} = \{(u,v)|(u,v) \in E, m \in H[u], n \in H[v]\}$ on Hi–C interactions E for corresponding histone pairs m and n. MWMCSP can be approximated by $O(\sqrt{mlog(n)}))$, which becomes $O(\sqrt{Q\log(|E|)})$ in our case where $m = Q = \max_{(u,v)\in E}(|H[u]|\,|H[v]|)$ is the maximum size of a color class, and $n = |E|$ is number of colors. □

Binning variant

If we bin the Hi–C data to a given resolution, there will be multiple Hi–C interactions to be explained by multiple histone pairs. Similar to the unbinned case, we assume each interaction to be explained by a single histone pair. There can also be self-interactions in G' as each node is a binning over multiple restriction sites. In this case, problem becomes Multiset Multicover variant of the problem in Sect. 2.2 where constraint (2) is replaced by:

$$\sum_{(m,c_m^u)\in H[u']} \sum_{(n,c_n^v)\in H[v']} c_m^u c_n^v x_{mn} \geq E'_{u'v'}, \quad (u', v') \in E' \tag{8}$$

We define weights in objective function in Eq. (1) as $w_m = \frac{\sum_{(u,v)\notin E'} c_m^u c_m^v}{\binom{R}{2} - |E|}$. To our best knowledge, this variant of the problem has not been defined before. This problem can again be solved by LP relaxation and randomized rounding. However, such scheme now does not give $O(\sqrt{Q\log(|E|)})$ approximation guarantee as in the unbinned case.

TemporalPrizeCoverage: spatio-temporal flexible marker selection

Genome shape, and thus Hi–C interactions tend to change over time. For instance, TADs at S and G1 cell cycle phases are not exactly the same (Naumova et al. 2013). Moreoever, it may not always be ideal biologically to cover all Hi–C interactions. For instance, some Hi–C interactions are false positives due to noise in Hi–C experiments (Rao et al. 2014). Chromatin marker data also include false positives as they are obtained mainly via noisy ChIP-seq experiment. Lastly, some of Hi–C interactions

may actually depend on other types of chromatin markers that are not considered in the study. Due to these reasons; (1) Having the flexibility to skip covering certain interactions by paying a penalty, (2) Selecting consistent set of markers to model spatio-temporal dynamics of Hi–C interaction network become quite important and biologically reliable.

In order to model the flexible but consistent spatio-temporal dynamics of genome shape through Hi–C interactions, we consider a spatio-temporal system at ordered set of time points $t = 1, 2, \ldots, T$ (such as cell cycles) over which we want to idenfity consistent set of chromatin markers. Let p_{uvt} be penalty of not covering the interaction between genomic loci/segments u and v at time t, we propose spatio-temporal prize collecting problem TemporalPrizeCoverage to select markers flexibly over multiple time points. TemporalPrizeCoverage is defined as in (9)–(14):

$$\text{argmin}_{\text{Y}} \sum_{\text{t}\in\text{T}} \sum_{\text{m}\in\text{M}} \text{w}_{\text{mt}}\text{y}_{\text{mt}} + \underbrace{\beta \sum_{t\in T} \sum_{(u,v)\in E} p_{uvt} z_{uvt}}_{\text{Uncovered edge penalty}} + \underbrace{C \sum_{t=2}^{T} \sum_{m\in M} (y_{mt} - y_{m(t-1)})^2}_{\text{Spatio-temporal penalty}} \tag{9}$$

$$\text{s.t.} \sum_{(m,c_m^u)\in N[u]} \sum_{(n,c_n^v)\in N[v]} x_{mnt} + z_{uvt} \geq 1, \quad (u,v) \in E, t \in T \tag{10}$$

$$\sum_{(n,c_n^u)\in N[u]} x_{mnt} + \sum_{(n,c_n^v)\in N[v]} x_{mnt} \leq y_{mt}, \quad m \in M, (u,v) \in E, t \in T \tag{11}$$

$$x_{mnt} \geq 0, \quad m \leq n \in M^2, t \in T \tag{12}$$

$$y_{mt} \geq 0, \quad m \in M, t \in T \tag{13}$$

$$z_{uvt} \geq 0, \quad (u,v) \in E, t \in T \tag{14}$$

where the objective (9) includes additional penalties $\beta \sum_{t\in T} \sum_{(u,v)\in E} p_{uvt} z_{uvt}$ to penalize Hi–C interactions that are not covered by selected set of markers, and $C \sum_{t=2}^{T} \sum_{m\in M} (y_{mt} - y_{m(t-1)})^2$ to enforce spatio-temporal consistency. Here β is a sparsity parameter that adjusts the tradeoff between marker costs and uncovered edge penalties. This tradeoff corresponds to collecting the prize on an edge by including the edge and evading its marker cost by excluding it. Similarly, C adjusts the tradeoff between inconsistent spatio-temporal markers and optimal interaction coverage by marker pairs. The rest of constraints are same as ChromatinCoverage, except they are defined for each individual time step t.

To our best knowledge, TemporalPrizeCoverage has not been studied before. It is also NP-hard, and it can be solved by LP relaxation and randomized rounding of y_{mt}'s similar to ChromatinCoverage. However, such scheme now does not give any approximation guarantee as in ChromatinCoverage. One problem variant is PrizeCoverage where there is only single time step, so spatio-temporal consistency penalty in the objective disappears. PrizeCoverage is also NP-hard.

Results

Implementation and datasets

We use Hi–C data from embryonic stem (ES) cells in mouse and human (Schmitt et al. 2016), GM12878 cells only in human (Rao et al. 2014) covering autosomal chromosomes. We download the genome assemblies from the UCSC genome browser. We use Juicer (Durand et al. 2016) to process the Hi–C sequencing reads of species to obtain the Hi–C contact pairs based on the corresponding genome assembly. We obtain histone modifications for human and mouse from NIH Roadmap Epigenomics (Bernstein 2010) and UCSC Encode [30]. In the binned case, we bin Hi–C, ChIP-Seq histone modifications at 1 kb resolution, and log-transform each bin's Reads Per Kilobase per Million (RPKM) values. This transform decreases the distorting effects of higher values. In case of more than one replicates, we average the RPKM-level for each bin to obtain a single histone modification file which minimizes the batch-related differences. Then, such normalized values are turned into binary values simply by thresholding 0.5. In the unbinned case, we map histone modification sites to neighboring Hi–C interaction's restriction sites. A modification is said to belong to a restriction site if the distance between modification and restriction site is less than 100. After such mapping, a histone modification either exists or not at a given restriction site. Results are based on unbinned case unless otherwise noted. When multiple time steps are considered by TemporalPrizeCoverage, we focus on human ES cell dataset across four phases (early G1, mid G1, S, M) of the cell cycle (Schmitt et al. 2016). In this case, TemporalPrizeCoverage consider all these 4 time steps simultaneously, and we utilize the same set of histone modifications as attributes across these different cell phases.

We implement ChromatinCoverage and its variant TemporalPrizeCoverage in Python, and use Gurobi to solve LP relaxations (Optimization 2020). Datasets and code can be found on http://www.github.com/seferlab/chrocoverage. ChromatinCoverage and TemporalPrizeCoverage are reasonably fast: Both methods can solve coverage formulations even without binning in less than 10 minutes on a laptop with 2.6 GHz Dual-Core Intel Core i5 processor and 16 Gb Ram. We prevent overfitting and optimize regularization parameters by following a 5-fold nested cross-validation where Hi–C and histone modifications datasets are split into 5 groups. In this case, the outer cross-validation step trains ChromatinCoverage on all chromosomes except the chromosome to be predicted. Then, we perform inner cross-validation step to calculate the regularization parameters within each outer cross-validation loop.

Four histone modifications are predictive of most Hi–C interactions

We find only 4 histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27ac) out of 16 modifications to be enough to explain the most of genome-wide Hi–C interactions of human ES cells by ChromatinCoverage. This is accurate for both 5 kb binned and unbinned cases. Figure 1 shows the percentage of covered interactions by increasing number of histone modifications, where more than 93% of interactions are covered by these 4 histone modifications for most chromosomes in human ES cells. Table 1 shows the histone modifications used in our experiments across species and cell types. As we range the number of included modifications from 1 to 16, coverage increase

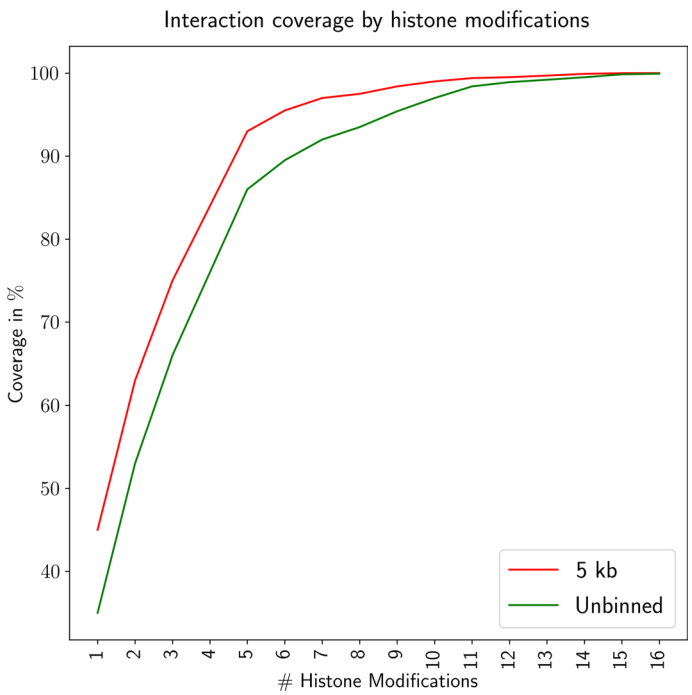


Fig. 1 Hi–C coverage by percentage by increasing number of histone modifications on 5 kb and unbinned cases

nearly stabilizes after 4 modifications, with some additional small increase up to 8 histone modifications by H3K79me2, H4K20me1, H3K36me3, H3K27me3 for 5 kb binned case. These set of histone modifications are highly conserved when we repeat this procedure across human GM12878 and mouse ES cells. Similarly, Fig. 2 shows the fraction of chromosomes histone modifications appearing in ChromatinCoverage solution on human ES cells, where ChromatinCoverage is run independently for each chromosome. In line with the previous figure, H3K4me1, H3K4me3, H3K9me3, H3K27ac still appear to be most important histone modifications across different chromosomes. The fraction of chromosomes in which these modifications appear important are significantly greater than the fraction of chromosomes in which the remaining modifications appear as important. The results show the similarity of histone modifications explaining the interactions across different chromosomes.

Among these important set of 8 markers, H3K4me1, H3K27ac are enhancer-specific markers. H3K9me3 is part of heterochromatin and is known to have repressive roles, whereas H3K4me3 is an activating marker. Among the rest of markers, H3K36me3 and H3K79me2 are known for their activator roles, wheres H3K27me3 is associated with polycomb repression similar to H3K9me3. Lastly, H4K20me1 is associated with transcriptional activation. The same subset of histone modifications are important for human IMR90 cells as well, where activating marker H3K9ac and H3K36me3 that is associated with active gene bodies and elongation are also part of the important set of 6 markers.

We also found the similar set of 4 histone modifications to be predictive of most Hi–C interactions when considered jointly across cell cycles via TemporalPrizeCoverage.

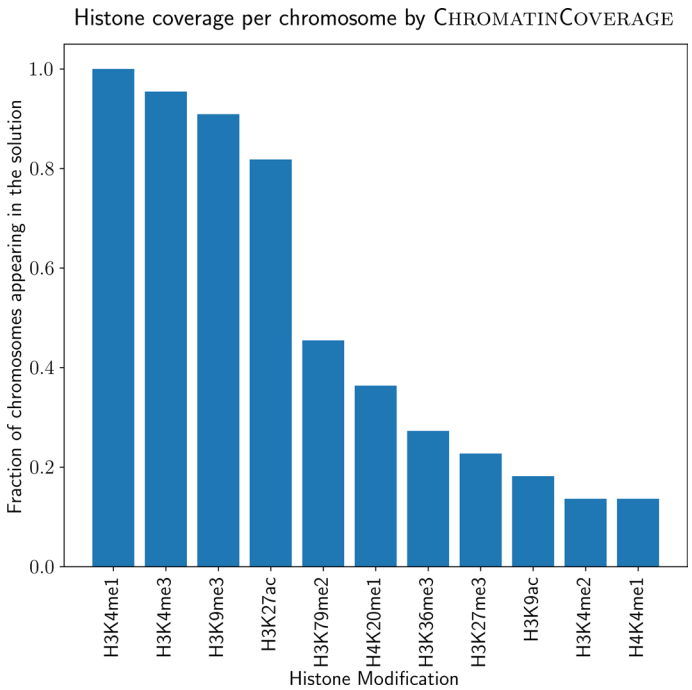


Fig. 2 The fraction of chromosomes histone modifications appearing in ChromatinCoverage solution on human ES cells, where ChromatinCoverage is run independently for each chromosome

In human ES cells, TemporalPrizeCoverage identifies H3K4me1, H3K4me3, H3K9me3, H3K27ac and activator H3K36me3 modification as important in covering Hi–C interactions when prize collecting penalty parameter $\beta = 0$, and spatio-temporal consistency penalty $C = 1$. Four of these markers are same as the 4 histone modifications identified by ChromatinCoverage. Even though prize collecting penalty parameter makes the problem formulation flexible, subset of identified markers do not change significantly when β is ranged from 0 to larger values. These results show that histone modifications considered in this study are enough to cover majority of Hi–C interactions across cell cycles.

ChromatinCoverage can predict Hi–C interactions and interaction network characteristics

We are able to detect false positive interactions on human ES cells by applying fivefold nested cross validation independently on each chromosome where both Hi–C interaction and histone modifications datasets are split into 5 groups. We predict Hi–C interactions by ChromatinCoverage over previously identified 4 histone modifications. Here, true interaction network is Hi–C interaction network of a human ES cells chromosome, whereas the predicted interaction network is another chromosome's Hi–C interaction network. We evaluate the performance by F1 score which is the harmonic mean of precision and recall scores, showing the tradeoff between both scores. According to matrix in Fig. 3, chromosome 4 has the best performance with 0.75 F1. Our experiments across chromosomes show that the performance decreases but it is reasonably well when we train on one chromosome and test on another one; training with interactions on chromosome 6 and predicting interactions on chromosome

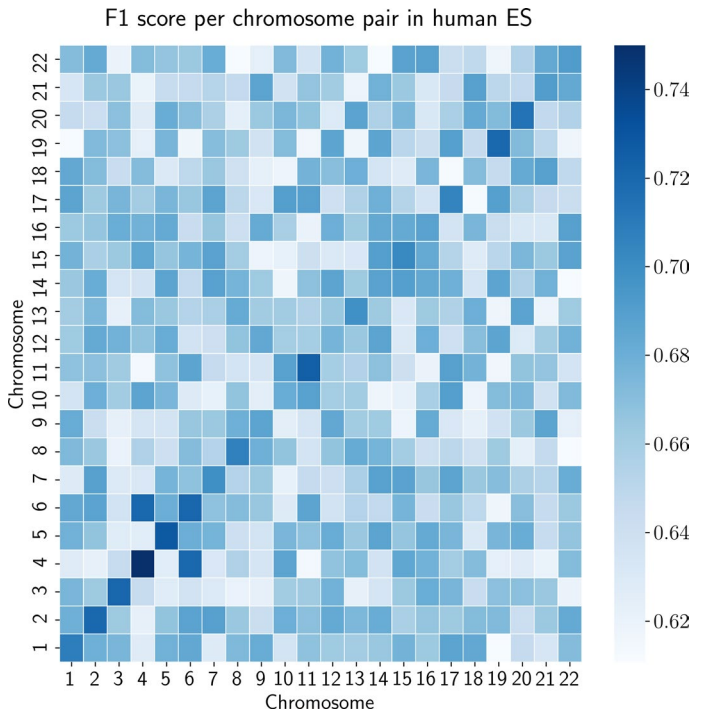


Fig. 3 F1 score for Hi–C interaction prediction per chromosome pair on human ES cells. Fivefold nested cross-validation is independently applied to each chromosome

Table 1 Histone modifications used in our experiments

Species & Cell type	Histone modifications
Human ES	H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K79me2, H3K36me3, H4K20me1, H3K27me3, H3K56ac, H3K23ac, H2AK5ac, H2A.Z, H3K9ac, H3K4me2, H4K8ac, H3K18ac
Human IMR90	H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K79me2, H3K36me3, H4K20me1, H3K27me3, H3K56ac, H3K23ac, H2AK5ac, H2A.Z, H3K9ac, H3K4me2, H4K8ac, H3K18ac
Mouse ES	H3K4me3, H3K27ac, H3K36me3, H3K4me1, H3K9me3, H3K27me3, H3K9ac

4 gives F1 score of 0.72. The results show the similarity of the identified modifications across different chromosomes. This suggests the similarity of properties governing chromosomal contacts across chromosomes, but, there can be close-grained differences that are not being captured by ChromatinCoverage.

Apart from F1 score comparison, we also measure the similarity between the predicted and true interaction network by comparing network characteristics as in Table 2. Here, we report the metrics for chromosome 4 which has the best performance according to Fig. 3, but the results on other chromosomes are not significantly different. ChromatinCoverage can predict modularity (Newman 2006) precisely which increases for highly-clustered networks. True interaction network shows certain degree of clustering as shown by the modularity scores. We found the scale-free

Table 2 Similarity metrics of true and inferred Hi–C interaction networks on chromosome 4

Species & Cell type	Similarity metric	True network	Inferred network
Human ES	Modularity Newman (2006)	0.73	0.67
	Avg. clustering coefficient	0.23	0.261
	Scale-free exponent	2.072	2.254
	Assortativity	0.141	0.121
	Diameter	4	4
Mouse ES	Modularity Newman (2006)	0.65	0.60
	Avg. clustering coefficient	0.24	0.232
	Scale-free exponent	2.12	2.087
	Assortativity	0.115	0.122
	Diameter	4	4

exponent of the interaction network's power law degree distribution as discussed in Clauset et al. (2009). Both true and predicted interaction networks degree distribution exhibit power law behaviour. Additionally, the predicted interaction network has the same diameter as the original interaction network which shows similar characteristics to larger social networks (Leskovec et al. 2007). ChromatinCoverage can also predict assortativity and average clustering coefficient closely to the original network.

Similarly, we measure the true and estimated interaction network modularities for each chromosome in human ES cells. Even though the modularity does not change significantly across chromosomes, interaction networks of certain chromosomes have more similar modularity scores than the rest of the interaction networks. Shorter chromosomes 19, 20, 21, and 22 have smaller modularity scores, showing more distributed 3D shapes of these chromosomes as a result of less clustered interaction network structure. Overall, correlation between modularity scores and length of chromosomes in terms of bases is 0.82 showing the possible importance of chromosome length in denser 3D genome shape formation via more interactions.

TemporalPrizeCoverage can predict Hi–C interactions over cell cycles

We can predict Hi–C interactions across four cell cycles in human ES cells by TemporalPrizeCoverage over the previously identified histone modifications. In Fig. 4, we test the performance in terms of precision by ranging spatio-temporal consistency parameter C from 0 to 50 where $\beta = \{1, 2.5, 5, 10\}$ are used. We report results in chromosome 4, but results for other chromosomes are also similar. The best performance is obtained for $C = 2$ and $\beta = 2.5$ which shows interactions across near cell cycles exhibit similar interaction structure which has been previously observed (Naumova et al. 2013). The precision decreases as we increase C values beyond 2 showing the importance of cell-cycle specific interactions which cannot be fully captured by higher C values. The best values are obtained when prize-collecting penalty $\beta = 2.5$, so skipping to cover certain fraction of interactions by the previously identified histone modifications have higher precision than the case where all interactions must be covered.

Additionally, we analyzed the impact of weighting heuristic defined in Eq. 6 in interaction prediction as in Fig. 5 where $\beta = 2.5$ and $C = 2$ as identified by cross-validation

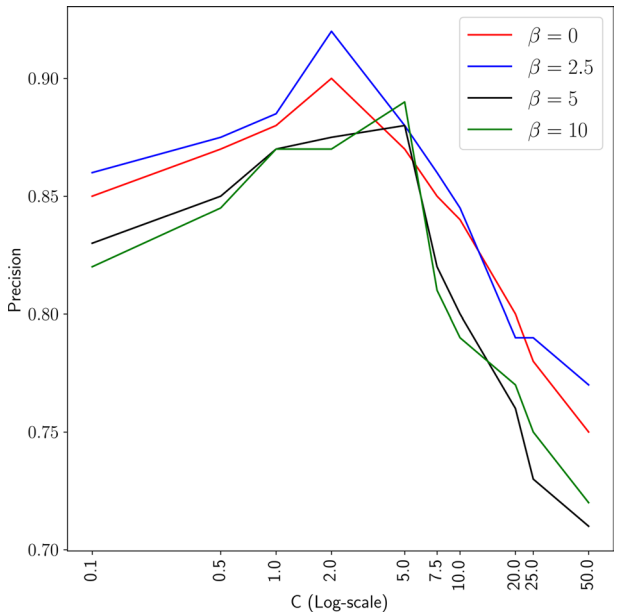


Fig. 4 The impact of weighting heuristic (Eq. 6) in precision on human ES cells across chromosomes

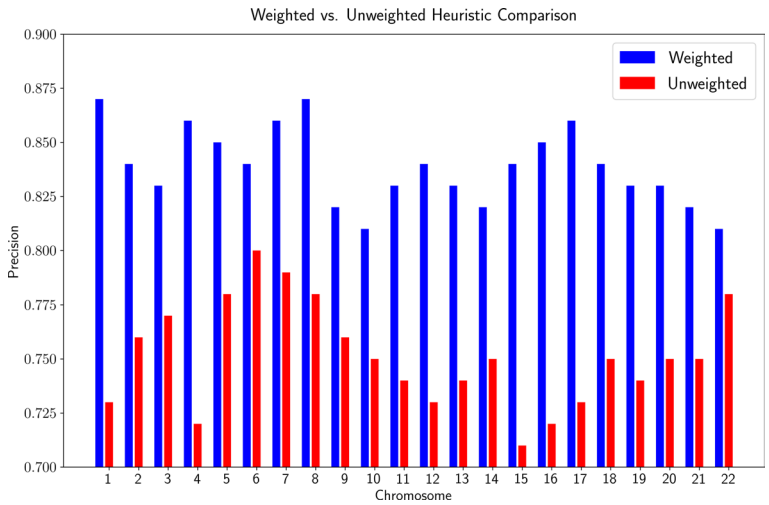


Fig. 5 The impact of weighting heuristic (Eq. 6) in precision on human ES cells across chromosomes

on human ES cells. Compared to the unweighted (equal-weighted) case, our heuristic weighting scheme penalizes the modifications that are also seen across non-interacting sites which cannot be penalized by the unweighted problem formulation. According to Fig. 5, interaction predictions made by our weighting heuristic have higher precision than the unweighted case across all chromosomes where we train and test on the same chromosome. As a result, penalizing modifications that are also seen across non-interacting genome sites brings us different subset of modifications that perform better in interaction prediction. The difference between both approaches is more apparent on chromosome 1 and 4, which are among the longest chromosomes and their structure

can be the main reason for such difference. Even though not shown here, weighting heuristic also outperforms equal-weighted case in terms of recall.

ChromatinCoverage can accurately predict chromosomal structures from identified histone modifications: TADs and hubs

We also evaluate the performance of ChromatinCoverage in predicting chromosomal structures. Figure 6 shows TAD prediction performance from histone modifications on human ES cells. TADs are known to be important in 3D genome shape formation (Dixon et al. 2012), so accurate prediction of TADs from histone modifications show the importance of these modifications in TAD and resultantly genome shape formation. The performance is evaluated via Normalized Variation of Information (NVI) (Meilă 2007), where Variation of Information defines the two partitions similarity, and a better performance is associated with a lower score. NVI takes values between 0 and 1, and lower NVI score means predicted TADs match greatly with the provided true TADs in the genome. TAD prediction from histone modifications is a two-step process: Once Hi–C graph interactions are predicted by ChromatinCoverage, we use Armatus (Filippova et al. 2014) to detect TADs over the predicted Hi–C interactions. Then, predicted TADs are compared with true TADs in terms of NVI where true TADs are detected over true Hi–C interaction matrix via Armatus. TAD prediction performance of training with all histone modifications is almost same as training only with 4 modifications which again suggests the overall quality of the previously identified 4 modifications in TAD prediction. In general, NVI scores between 0.1–0.15 as implied by ChromatinCoverage TAD predictions exhibit a good performance. NVI score is also comparable across chromosomes, the best TAD prediction performance is observed on chromosome 20 which can be due to relatively better predictive distribution of TADs in chromosome 20.

Relatedly, we predict hubs over ChromatinCoverage inferred interaction network in each chromosome of human ES cells. In graphs, hubs are the nodes with a

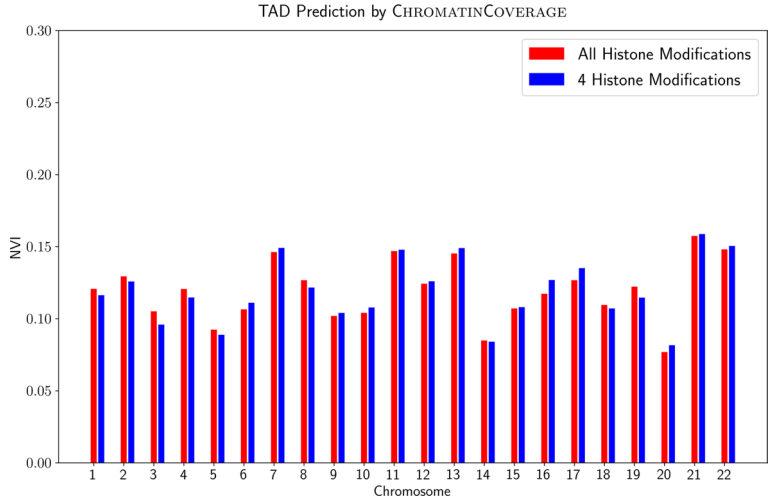
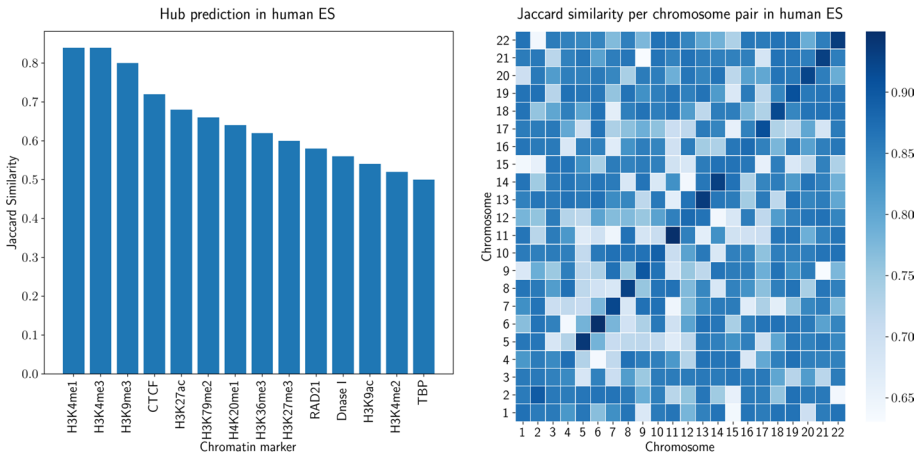


Fig. 6 Performance of TAD prediction by NVI on human ES cells for histone modifications

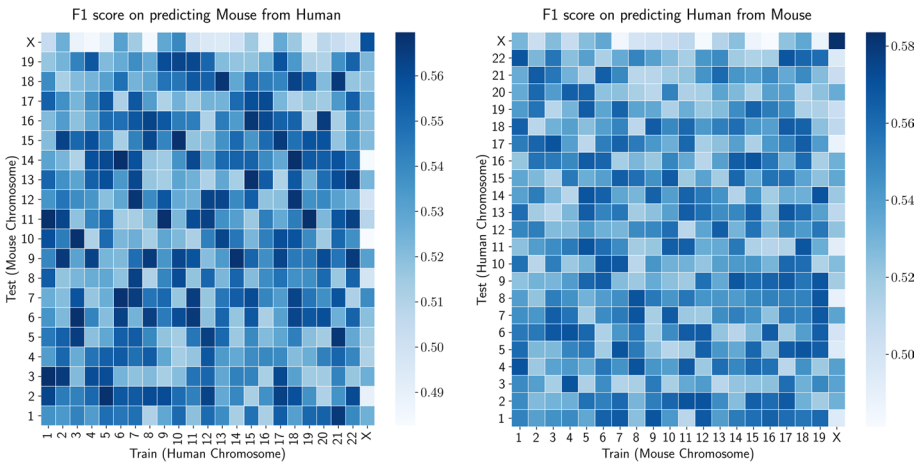
number of links that greatly exceeds a given threshold. Here, we define hubs as the topmost 10% nodes in the Hi–C interaction graph in terms of number of connections, and present prediction results by Jaccard similarity which is defined as the size of the intersection of true and inferred hubs divided by the size of the union of the true and inferred hubs. In this case, true hubs are the topmost 10% nodes in the true Hi–C interaction graph in terms of number of connections, whereas inferred hubs are the topmost 10% nodes in the inferred Hi–C interaction graph. Jaccard similarity takes values between 0 and 1, and higher score means better hub prediction performance. When considered independently, H3K4me1 is the most informative predictor for hubs in human ES as in Fig. 7a where average Jaccard similarity among all chromosomes is presented. As in Hi–C interaction prediction, histone modifications are more effective than transcription factor binding sites in hub prediction. Similarly, Fig. 7b shows Jaccard similarities for hubs between chromosome pairs, where we train ChromatinCoverage with a single chromosome, and predict hubs on a different chromosome. Even though set of chromatin markers change depending on the cell type, we find similar AUC results on human GM12878 and IMR90 cells. As a result, cell-type specific chromatin marker information is required for the prediction of chromatin interaction hubs. The ability of ChromatinCoverage to properly extract hubs suggest its high-quality performance in Hi–C interaction graph topology reconstruction. The emergence of hubs in Hi–C interaction graph can be explained by its scale-free properties. Hubs in interaction graph correspond to genome segments that interact greatly with other genome segments. As a result, hubs are more important and more central in 3D genome shape than non-hub nodes.



(a) The impact of the topmost effective 10 independent markers in human ES, where the average of all chromosomes is taken.

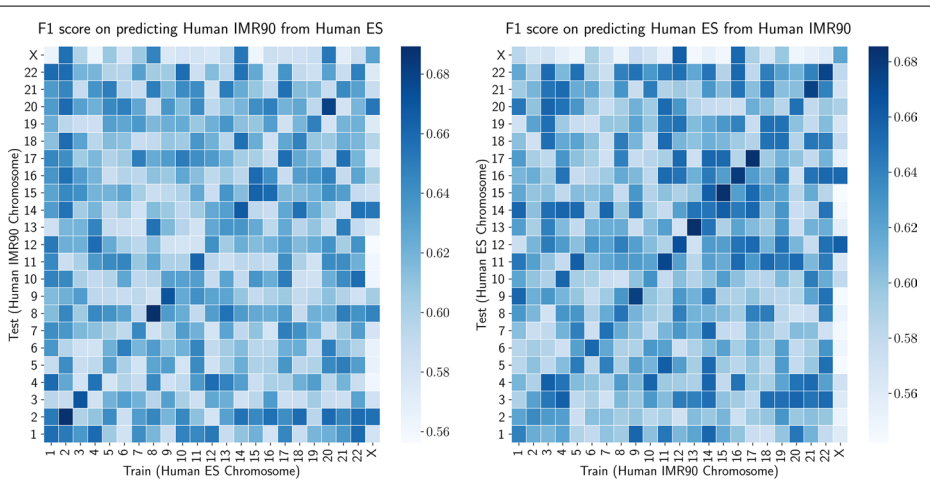
(b) Jaccard similarity matrix for chromosome pairs on Hi-C hub prediction in human ES cells. 5-fold nested cross-validation is independently applied to each chromosome.

Fig. 7 **a** The impact of the topmost effective 10 independent markers in human ES, where the average of all chromosomes is taken. **b** Jaccard similarity matrix for chromosome pairs on Hi–C hub prediction in human ES cells. Fivefold nested cross-validation is independently applied to each chromosome



(a) Mouse ES from Human ES (b) Human ES from Mouse ES

Fig. 8 F1 score for predicting **a** mouse ES from human ES, **b** human ES from mouse ES cells. Fivefold nested cross-validation is independently applied to each chromosome



(a) Human GM12878 from ES (b) Human ES from GM12878

Fig. 9 F1 score for predicting **a** human GM12878 from human ES, **b** human ES from human GM12878 cells. 5-fold nested cross-validation is independently applied to each chromosome

Histone modifications are important in Hi–C interaction prediction across species and cell types

We predict Hi–C interactions on mouse ES chromosomes over ChromatinCoverage trained with chromosome-wide human ES cells. Figure 8a shows *F*1 score in mouse chromosomes. Our cross-species prediction shows that the performance can vary from one training species to another. Prediction performance is the worst for X chromosome, and intrachromosomal interactions can be predicted more accurately than the interchromosomal ones. Similarly, Fig. 8b repeats the same analysis on human ES cells by training ChromatinCoverage over mouse ES chromosomes. Prediction performance between species is lower than the performance between cell types on the same species, showing

effective histone modifications may differ across species. Results prove the importance of histone modification on genome shape, since the performance decreases even when modifications are transferred from a closer species.

We also examine the impact of cell types in Hi–C interaction prediction as in Fig. 9a–b respectively. Prediction performance between Human GM12878 and human ES is better than the prediction between species suggesting that common subset of histone modifications explain genome shapes of different cell types. There is no significant performance difference between training on human ES vs. human GM12878.

Conclusions

We investigate how histone modifications and interactions between them explain Hi–C interactions, thus the three-dimensional genome organization. Experiment results on mouse and human imply that a common set of histone modifications accurately predict Hi–C interactions across cell types, species and cycles. By using our methods, we can also accurately infer Hi–C interactions only from histone modifications, that is mainly useful to understand the 3D genome shape on species with limited Hi–C data. ChromatinCoverage is also effective in identifying chromosomal structures such as topologically-associated domains. Our cross-chromosome experiments show that the performance decreases when training on one chromosome and testing on another. The features identified as important across different chromosomes are quite similar, suggesting that the overall properties governing chromosomal interactions are similar across chromosomes. Predictions made by our methods can be verified both experimentally and biologically. Overall, the analysis performed in this work provides good insights on the impact of histone modifications and interaction between them in the 3D genome shape.

In the future, ChromatinCoverage can be extended to more recent multilocus chromatin interaction experiments through a hypergraph formalism instead of a graph. Besides, the problem can be casted as a constrained supermodular minimization problem where covered interactions will be a supermodular function of the added modifications. Additionally, our method can be extended to handle datasets other than histone modifications, such as RNA-Seq, transcription factor binding sites, etc. We expect these future methods to possibly enhance the Hi–C interaction prediction performance.

Abbreviations

ES: Embryonic stem; MWMCSP: Minimum weight multicolored subgraph problem; NVI: Normalized variation of information; RPKM: Reads Per kilobase per million; TAD: Topologically-associated domain.

Acknowledgements

A preliminary version of this paper has appeared in Complex Networks 2020 (The 9th International Conference on Complex Networks and their Applications) conference proceedings. Author would like to thank the reviewers for their helpful comments.

Author Contributions

ES formulated the problem, implemented the tool, and analyzed the results. ES wrote the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by Ozyegin University research grant.

Availability of data materials

Datasets and code can be found on Github repository http://www.github.com/seferlab/chrocoverage.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 9 January 2021 Accepted: 2 July 2021
Published online: 17 July 2021

References

Al Bkhetan Z, Plewczynski D (2018) Three-dimensional epigenome statistical model: genome-wide chromatin looping prediction. Sci Rep 8(1):5217
Ashoor H, Chen X, Rosikiewicz W, Wang J, Cheng A, Wang P, Ruan Y, Li S (2020) Graph embedding and unsupervised learning predict genomic sub-compartments from hic chromatin interaction data. Nature Commun 11(1):1173
Babaei S, Mahfouz A, Hulsman M, Lelieveldt BPF, de Ridder J, Reinders M (2015) Hi–C chromatin interaction networks predict co-expression in the mouse cortex. PLoS Comput Biol 11(5):1–21
Bernstein BE et al (2010) The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 28(10):1045–1048
Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10(3):186–198
Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703. https://doi.org/10.1137/070710111
Dabin J, Fortuny A, Polo SE (2016) Epigenome maintenance in response to dna damage. Mol Cell 62(5):712–727. https://doi.org/10.1016/j.molcel.2016.04.006
Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14(6):390–403
Di Pierro M, Cheng RR, Lieberman Aiden E, Wolynes PG, Onuchic JN (2017) De novo prediction of human chromosome structures: epigenetic marking patterns encode genome architecture. Proc Nat Acad Sci 114(46):12126–12131
Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398):376–380
Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL (2016) Juicer provides a one-click system for analyzing loop-resolution Hi–C experiments. Cell Syst 3(1):95–98
Emre S, Geet D, Carl K (2016) Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations. J Comput Biol 23(6):425–438
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74
Filippova D, Patro R, Duggal G, Kingsford C (2014) Identification of alternative topological domains in chromatin. Algorith Mol Biol 9(1):14
Gibney ER, Nolan CM (2010) Epigenetics and gene expression. Heredity 105(1):4–13. https://doi.org/10.1038/hdy.2010.54
Hajiaghayi MT, Jain K, Lau LC, Măndoiu II, Russell A, Vazirani VV (2006) Minimum multicolored subgraph problem in multiplex pcr primer set selection and population haplotyping. In: Alexandrov VN, van Albada GD, Sloot PMA, Dongarra J (eds) Computational Science - ICCS 2006. Springer, Berlin, Heidelberg, pp 758–766
Halldórsson BV, Bafna V, Edwards N, Lippert R, Yooseph S, Istrail S (2004) A survey of computational methods for determining haplotypes. Lect Notes Comput Sci 2983:26–47
Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, De Gobbi M, Taylor S, Gibbons R, Higgs DR (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet 46(2):205–212
Konwar KM, Mandoiu II, Russell A, Shvartsman AA Improved algorithms for multiplex PCR primer set selection with amplification length constraints, pp 41–50
Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. ACM Trans Knowl Discov Data 1(1):2. https://doi.org/10.1145/1217299.1217301
Li W, Wong WH, Jiang R (2019) DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res 47(10):60
Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS (2015) Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell type-specific expression. Genome Res
Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293
Meilă M (2007) Comparing clusterings–an information based distance. J Multivar Anal 98(5):873–895
Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA et al (2015) Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. Nat Genet 47(6):598–606
Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J (2013) Organization of the mitotic chromosome. Science 342(6161):948–953. https://doi.org/10.1126/science.1236083
Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103
Nora EP, Dekker J, Heard E (2013) Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? In: BioEssays: news and reviews in molecular, cellular and developmental biology
Optimization G (2020) Gurobi Optimizer Reference Manual. http://www.gurobi.com

Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 153(6):1281–1295

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al (2014) A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665–1680

Schmitt A, Hu M, Jung I, Xu Z, Qiu Y, Tan C, Li Y, Lin S, Lin Y, Barr C, Ren B (2016) A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep 17(8):2042–2059

Schreiber J, Singh R, Bilmes J, Noble WS (2019) A pitfall for machine learning methods aiming to predict across cell types. bioRxiv

Sefer E, Kingsford C (2019) Semi-nonparametric modeling of topological domain formation from epigenetic data. Algorith Mol Biol 14(1):4

Trieu T, Martinez-Fundichely A, Khurana E (2020) Deepmilo: a deep learning approach to predict the impact of non-coding sequence variants on 3d chromatin structure. Genome Biol 21(1):79

Publisher's Note