NETWORK MODELLING OF TOPOLOGICAL DOMAINS USING HI-C DATA

BY Y.X. RACHEL WANG* AND PURNAMRITA SARKAR† AND OANA URSU‡ AND ANSHUL KUNDAJE§‡ AND PETER J. BICKEL¶

School of Mathematics and Statistics, University of Sydney, Australia*
Department of Statistics and Data Sciences, University of Texas, Austin†
Department of Genetics, Stanford University‡
Department of Computer Science, Stanford University§
Department of Statistics, University of California, Berkeley¶

Chromosome conformation capture experiments such as Hi-C are used to map the three-dimensional spatial organization of genomes. One specific feature of the 3D organization is known as topologically associating domains (TADs), which are densely interacting, contiguous chromatin regions playing important roles in regulating gene expression. A few algorithms have been proposed to detect TADs. In particular, the structure of Hi-C data naturally inspires application of community detection methods. However, one of the drawbacks of community detection is that most methods take exchangeability of the nodes in the network for granted; whereas the nodes in this case, i.e. the positions on the chromosomes, are not exchangeable. We propose a network model for detecting TADs using Hi-C data that takes into account this non-exchangeability. In addition, our model explicitly makes use of cell-type specific CTCF binding sites as biological covariates and can be used to identify conserved TADs across multiple cell types. The model leads to a likelihood objective that can be efficiently optimized via relaxation. We also prove that when suitably initialized, this model finds the underlying TAD structure with high probability. Using simulated data, we show the advantages of our method and the caveats of popular community detection methods, such as spectral clustering, in this application. Applying our method to real Hi-C data, we demonstrate the domains identified have desirable epigenetic features and compare them across different cell types. The code is available at https://github.com/ryxwang/lp_opt.

1. Introduction. In complex organisms, the genomes are very long polymers divided up into chromosomes and tightly packaged to fit in a minuscule cell nucleus. As a result, the packaging and the three-dimensional (3D) conformation of the chromatin have a fundamental impact on essential cellular processes including cell replication and differentiation. In particular, the 3D structure regulates the transcription of genes at multiple levels (Dekker, 2008). At the chromosome level, open (active) and closed (inactive)
compartments alternate along chromosomes (Lieberman-Aiden et al., 2009) to form regions with clusters of active genes and repressed transcriptional activities, the latter typically partitioned to the nuclear periphery (Sexton et al., 2012; Smith et al., 2016). At a smaller scale, chromatin loops make long-range regulations possible by bringing distant enhancers and repressors close to their target promoters.

Recently, one specific feature of chromatin organization known as topologically associating domains (TADs) has attracted much research attention. TADs are contiguous regions of chromatin with high levels of self-interaction and have been found in different cell types and species (Dixon et al., 2012; Sexton et al., 2012; Hou et al., 2012). A number of studies have shown TADs contain clusters of genes that are co-regulated (Nora et al., 2012) and may correlate with domains of histone modifications (Le Dily et al., 2014), suggesting TADs act as functional units to help gene regulation. Disruptions of domain conformation have been associated with various diseases including cancer and limb malformation (Lupiáñez et al., 2015; Meaburn et al., 2009).

While it is not possible to completely observe the 3D conformation, in the past decade several chromosome conformation capture technologies have been developed to measure the number of ligation events between spatially close chromatin regions. Hi-C is one of such technologies and provides genome-wide measurements of chromatin interactions using paired-end sequencing (Lieberman-Aiden et al., 2009). The output can be summarized in a raw contact frequency matrix $M$, where $M_{ij}$ is the total number of read pairs (which are interacting) falling into bins $i$ and $j$ on the genome. These equal-sized bins partition the genome and range from a few kilobases to megabases depending on the data resolution. Since TADs are regions with high levels of self-interactions, they appear as dense squares on the diagonal of the matrix.

A number of algorithms have been proposed to detect TADs, most of which rely on maximizing the intra-domain contact strength. This includes the earlier methods by Dixon et al. (2012) and Sauria et al. (2014), which summarize the 2D matrix as a 1D statistic to capture the changes in interaction strength at domain boundaries; and methods that directly utilize the 2D structure of the matrix to contrast the TAD squares from the background (Filippova et al., 2014; Lévy-Leduc et al., 2014; Weinreb and Raphael, 2016; Malik and Patro, 2015; Rao et al., 2014). All of these methods use an optimization framework and apply standard dynamic programming to obtain the solution. The algorithms typically involve a number of tuning parameters with the number of TADs chosen in heuristic ways. More recently, Cabreros et al. (2016) proposed to view the contact frequency matrix as an
weighted undirected adjacency matrix for a network and applied community
detection algorithms to fit mixed-membership block models.

Statistical networks provide a natural framework for modelling the 3D
structure of chromatin as we can consider it as a spatial interaction network
with positions on the genome as nodes. Network models have gained much
popularity in numerous fields including social science, genomics, and imaging;
the availability of Hi-C data opens new ground for applying network
techniques, such as community detection, in order to answer important ques-
tions in biology. One of the drawbacks of community detection is that most
of the methods take exchangeability of the nodes in the network for granted.
However, modelling Hi-C data is a typical situation where the nodes, i.e. the
positions on the genome, are not exchangeable. In particular, since TADs
are contiguous regions, treating TADs as densely connected communities
imposes a geometric constraint on the community structure.

In this paper, we propose a network model for detecting TADs that in-
corporates the linear order of the nodes and preserves the contiguity of the
communities found. Our main contributions include: i) It has been observed
empirically TADs are conserved across different cell types, but explicit joint
analysis remains incomplete. Our likelihood-based method easily generalizes
to allow for joint inference with multiple cell types. ii) It has been postu-
lated that CTCF (an insulator protein) acts as anchors at TAD boundaries
(Nora et al., 2012; Sanborn et al., 2015). Empirically, TAD boundaries cor-
relate with CTCF sites, and modifications of binding motifs can lead to
TAD disappearance (Sanborn et al., 2015). Our model is flexible enough to
include the positions of CTCF sites as biological covariates. iii) We account
for the existence of nested TADs. iv) The core of our algorithm is based on
linear programming, making it fast and efficient. v) In addition, we provide
theoretical justifications by analyzing the asymptotic performance of the al-
gorithm and using automated model selection for choosing the number of
TADs. The latter saves the need for many tuning parameters. Among these,
i) and ii) are unique features of our method with biological significance.

The rest of the paper is organized as follows. We introduce the model and
the estimation algorithm with asymptotic analysis in Section 2. In addition,
we describe a post-processing step for testing the enrichment of contact
within any TAD found. In Section 3, we first use simulated data to demon-
strate the necessity of taking into account the linear ordering of the nodes
and compare our method with other TAD detection algorithms. We next
present the results of real data analysis for multiple human cell types, in-
dividually and jointly, using a publicly available Hi-C dataset (Rao et al.,
2014). We end the paper with a discussion of the advantages of our method
2. Methods. In this section, we describe a hierarchical network model for detecting nested TADs in a Hi-C contact frequency matrix using cell-line specific CTCF peaks as covariates. At each level of the hierarchy, we show the parameters can be estimated efficiently via coordinate ascent and provide asymptotic analysis of the algorithm. In addition, the model and algorithm can be adapted to identify TADs conserved across multiple cell lines. As further confirmation that the TADs found by the algorithm indeed correspond to regions of the genome with enriched interactions, we post process the candidate regions by performing a nonparametric test.

2.1. Model description. We consider a hierarchical model with a set of maximally non-overlapping TADs at each level. In this section, we focus on describing the model for the base (outermost) level. The model and parameter estimation for the nested levels are identical and will be mentioned at the end of Section 2.2.

Let $M$ denote a $n \times n$ contact frequency matrix. $M$ is first thresholded at the $q$-th quantile to produce a binary adjacency matrix $A$. Thresholding has been a common practice in network modeling to handle weighted matrices, despite the information loss it incurs. At canonical sequencing depth, the signal to noise ratio in Hi-C data is typically high and the resolution is relatively low. Thresholding can improve the signal to noise ratio. We examine the effect and sensitivity of the choice of $q$ in Section 3.

As mentioned in the introduction, experimental evidence suggests TAD boundaries tend to coincide with CTCF binding. This motivates us to incorporate the presence of CTCF into our model. Let $Y \in \{0, 1\}^n$ be a binary vector with ones at positions where CTCF binding occurs. We will treat $Y$ as an available covariate, which can be obtained from ChIP-seq data which is cell-type specific.

Let $X$ denote a $n \times n$ binary matrix such that $X_{ab} = 1$ if i) $Y_a = 1, Y_b = 1$ and ii) there is a TAD between position $a$ and $b$. $X_{ab}$ is always 0 when $Y_a Y_b = 0$. This enforces the model to generate TADs which always have CTCF peaks at their boundaries. Thus $X \in \{0, 1\}^{n \times n}$ denotes a binary latent matrix which encodes the positions of all TADs. Also note that, it is possible to have $Y_a Y_b = 1$, but $X_{ab} = 0$, i.e. there was no TAD formed between two CTCF binding sites.

We denote by the parameter vector $\Theta = (\beta, \{\alpha_{ab} : X_{ab} = 1\})$ the probabilities of edges between nodes. If $a \leq i < j \leq b$ for $X_{ab} = 1$, then $P(A_{ij} = 1) = \alpha_{ab}$. (Note that we allow for a different edge probability for each TAD.) Otherwise $P(A_{ij} = 1) = \beta$, which is also referred to as the
background probability. The diagonal of $A$ is set to 0. For simplicity we have assumed the connectivity within each TAD and the background is uniform, although the TADs may contain nested sub-TADs and can be heterogeneous. In general the contact frequency decreases as a function of the distance between two loci. For now one can think of the homogeneity assumption as approximating the actual distribution with a piecewise constant function, and we make use of the original weights in the post-processing step (Section 2.4). Finally, our model does not require the exact number of TADs, but only an upper bound on it. We will make this more concrete in Section 2.2.

**Remark 2.1.** We demonstrate our model using a concrete example. The corresponding edge probability matrix is shown in Figure 1. In this example $Y_i = 1$ for $i \in \{3, 6, 12, 18\}$. We show positions where $Y_a Y_b = 1$ by red dots at the intersection of the grid lines, where the grid lines show the positions of the CTCF sites. Only $X_{ab}$ at these positions are allowed to be one, since according to our model, TADs can only form between two CTCF sites. In this example, there are two TADs between 3 and 6 and between 12 and 18, that is $X_{3,6} = 1$, $X_{12,18} = 1$. The edge probabilities may differ between the two TADs. The model naturally enforces contiguous clusters, and one cannot have a TAD with a hole inside.
2.2. Parameter estimation. Knowing \((X, Y, \Theta)\), the maximization of the log likelihood for \(A\) can be written as

\[
\max_{X \in \{0,1\}^{n \times n}, \Theta} \log p(A; X, Y, \Theta)
\]

\[
= \frac{1}{2} \sum_{i \neq j} \sum_{a < b} Y_a Y_b X_{ab} \mathbb{I}_{i,j \in [a,b]} \left( A_{ij} \log \left( \frac{\alpha_{ab}}{1 - \alpha_{ab}} \right) + \log(1 - \alpha_{ab}) \right)
\]

\[
+ \frac{1}{2} \sum_{i \neq j} \left( 1 - \sum_{a < b} Y_a Y_b X_{ab} \mathbb{I}_{i,j \in [a,b]} \right) \left( A_{ij} \log \left( \frac{\beta}{1 - \beta} \right) + \log(1 - \beta) \right)
\]

s.t. \(\sum_{a < b} Y_a Y_b X_{ab} \leq K\)

and \(\sum_{c \leq a \leq d} Y_c Y_d X_{cd} \leq 1\) for all \(a\) s.t. \(Y_a = 1\).

(2.1)

The first constraint upper bounds the total number of TADs at this level, while the second constraint ensures there is at most one TAD covering each position, thus making the TADs non-overlapping. The likelihood implies it suffices to consider \(X_{ab}\) at positions such that both \(Y_a = 1\) and \(Y_b = 1\), and \(X\) is effectively a \(m \times m\) matrix, where \(m = \sum_a Y_a\). In this way the covariate vector \(Y\) helps reduce the search to a smaller grid.

We maximize the likelihood by considering a relaxed objective function and performing coordinate ascent. First note that taking the derivative of \(\log p(A; X, Y, \Theta)\) with respect to \(\alpha_{ab}\), the estimate of \(\alpha_{ab}\) does not depend on the other parameters and is given by

\[
\hat{\alpha}_{ab} = \frac{\sum_{i,j \in [a,b]} A_{ij}}{(b - a + 1)(b - a)}
\]

(2.2)

Therefore it remains to maximize the likelihood with respect to \(\beta\) and \(X\). Since direct maximization of (2.1) over \(X\) subject to the constraints involve combinatorial optimization, we propose the following relaxed optimization,

\[
\max_{\beta, \pi \in [0,1]^{n \times n}} L(A, Y, \beta, \pi)
\]

\[
:= \max_{\beta, \pi \in [0,1]^{n \times n}} \frac{1}{2} \sum_{i \neq j} \sum_{a < b} Y_a Y_b \pi_{ab} \mathbb{I}_{i,j \in [a,b]} \left[ A_{ij} \log \left( \frac{\hat{\alpha}_{ab}}{1 - \hat{\alpha}_{ab}} \right) + \log(1 - \hat{\alpha}_{ab}) \right]
\]

\[
+ \frac{1}{2} \sum_{i \neq j} \left( 1 - \sum_{a < b} Y_a Y_b \pi_{ab} \mathbb{I}_{i,j \in [a,b]} \right) \left[ A_{ij} \log \left( \frac{\beta}{1 - \beta} \right) + \log(1 - \beta) \right],
\]

s.t. \(\sum_{a < b} Y_a Y_b \pi_{ab} \leq K\)
and \( \sum_{c \leq a \leq d} Y_c Y_d \pi_{cd} \leq 1 \) for all \( a \) s.t. \( Y_a = 1 \).  

(LP-OPT)

The objective and constraints have the same form as (2.1) but with \( \pi \in [0, 1]^{n \times n} \) replacing \( X \in \{0, 1\}^{n \times n} \). Again since \( \pi_{ab} = 0 \) if \( Y_a Y_b = 0 \), the size of \( \pi \) to be estimated is effectively \( m \times m \). This relaxed version can be solved via alternating maximization, also denoted by LP-OPT.

1. For each fixed \( \beta \), (LP-OPT) is linear in \( \pi \) and can be maximized efficiently using linear programing.
2. For each fixed \( \pi \), the objective is maximized at

\[
\hat{\beta} = \frac{\sum_{i,j} A_{ij} - \sum_{a,b} \pi_{ab} Y_a Y_b \sum_{i,j \in[a,b]} A_{ij}}{n(n-1) - \sum_{a,b} \pi_{ab} Y_a Y_b (b-a+1)(b-a)}.
\]

(2.3)

The above two steps are iterated until convergence in \( \beta \).

So far we have described the model and parameter estimation for the outermost level of TADs. Within each of these TADs, we can repeat the same algorithm to detect the secondary (nested) level of TADs and continue iterating.

The likelihood approach allows the method to be easily extended to model conserved TADs across multiple cell lines. Assuming the cell lines are independent, the joint log likelihood can be written as the sum,

\[
\log p(A; X, Y, \Theta) = \sum_{\ell} \log p(A_\ell; X, Y, \Theta_\ell),
\]

(2.4)

where \( X \) represents the latent positions of common TADs, \( Y \) is the set of CTCF peaks common to all cell lines; \( A_\ell \) and \( \Theta_\ell \) are the adjacency matrix and model parameters specific to cell line \( \ell \). Similar to the single cell line case, the parameters can be estimated by using a plug-in estimator for each \( \alpha_\ell \) and alternating between maximizing over \( \pi \) and \( \beta_\ell \), where \( \pi \) is the relaxed form of \( X \).

2.3. Theoretical guarantees. In this section, we analyze the theoretical properties of the algorithm and discuss the asymptotic performance of the estimates. Given that we have relaxed the original likelihood, it is natural to first check whether the solutions of (2.1) and LP-OPT agree. We have the following lemma stating optimizing the relaxed objective is essentially equivalent to optimizing the original one.

**Lemma 2.2.** For every given \( \beta \),

\[
\max_{\pi \in \Pi} L(A, Y, \beta, \pi) = \max_{X \in \mathcal{X}} L(A, Y, \beta, X),
\]

(2.5)
where \( \Pi \) is the feasible set in LP-OPT and \( \mathcal{X} \) is the feasible set in (2.1).

**Proof.** Given \( \beta \), updating \( \pi \) is equivalent to maximizing the function
\[
L(A, Y, \Theta, \pi) = \frac{1}{2} \sum_{i \neq j} \sum_{a < b} Y_a Y_b \pi_{ab} l_{ij} \left[ A_{ij} \log \frac{\hat{\alpha}_{ab}(1 - \beta)}{(1 - \hat{\alpha}_{ab})^\beta} + \log \frac{1 - \hat{\alpha}_{ab}}{1 - \beta} \right] + \text{constant}
\]
where \( l(A; \pi, \beta) = l(A; \pi, \beta) + \text{constant} \). (2.6)

Recalling \( \hat{\alpha}_{ab} \) is independent of all the parameters, \( l(A; \pi, \beta) \) is linear in \( \pi \).

Furthermore, the feasible set for \( \pi \) given in LP-OPT is a convex polyhedron with vertices at \( \mathcal{X} \). Since the optimum for a linear function on a convex polyhedron is always attained at the vertices, it follows then maximizing \( l(A; \pi, \beta) \) with respect to \( \pi \) is equivalent to maximizing \( l(A; \mathcal{X}, \beta) \), which is the original objective.

The above lemma implies it is valid to analyze the solution of (2.1) even though the algorithm solves a relaxed problem. Furthermore, the optimal \( \pi \) for each run of step 1 in the algorithm belongs to the feasible set \( \mathcal{X} \) and defines a set of valid TAD positions (hence no thresholding is needed).

Next we analyze the asymptotics of the alternating optimization algorithm given a reasonable starting value \( \beta_0 \) and the upper bound \( K \) for the following setting. We consider the most general case where each position is allowed a CTCF peak so \( Y_a \) will be omitted for the rest of the section. We focus on a single level of the hierarchical model and assume the \( n \times n \) adjacency matrix \( A \) contains \( K^* \) TADs with \( \{\alpha^*_1, \ldots, \alpha^*_{K^*}\} \) as their connectivity probabilities. Note that to simplify notation, we have changed the subscript for \( \alpha \) to a single index. The background has connectivity probability \( \beta^* \). Let \( \{[s_1, t_1], \ldots, [s_{K^*}, t_{K^*}]\} \) be the TAD locations with the corresponding sizes \( \{n^*_1, \ldots, n^*_{K^*}\}; t_0 = 0, s_{K^*+1} = n + 1 \) for convenience. We consider the case where \( K^* \) is fixed, \( n_k^*/n \rightarrow p_k > 0 \) for all \( k \). In addition the sizes of the inter-TAD regions also follow \( (s_{k+1} - t_k - 1)/n \rightarrow q_k \). Denote the number of inter-TAD regions \( G^* \). Define \( KL(s\|t) = s \log(\frac{s}{t}) + (1 - s) \log(\frac{1 - s}{1 - t}) \).

Assume the given \( \beta \) satisfies the following assumption:

**Assumption 2.3.** \( \beta^* < \beta < \min_k \alpha^*_k \).

**Assumption 2.4.** For large enough \( n \),
\[
\left( (s_j - t_i - 1)^2 - \sum_{i<k<j} (n_k^*)^2 \right) KL(\beta^*\|\beta) < \sum_{i<k<j} (n_k^*)^2 KL(\alpha^*_k\|\beta)
\]
for all \( j > i + 1 \). Note here \((s_j - t_i - 1)\) is the segment between the end of the \(i\)th TAD and the beginning of the \(j\)th TAD.

Note that when \( \beta = \beta^* \), Assumption 2.4 is trivially satisfied.

**Theorem 2.5.** Starting with \( \beta^{(0)} \) satisfying Assumptions 2.3 and 2.4, for any fixed \( K \) and \( K \) large enough such that \( K \geq K^* + G^* \), the optimal \( X \) satisfies

\[
\exp \left\{ \max_{X \in \mathcal{X}} l(A; X, \beta^{(0)}) \right\} = \exp \left\{ l(A; X_0, \beta^{(0)}) \right\} (1 + o_P(1)), \tag{2.7}
\]

where \( X_0 \) is such that \( X_{s_k, t_k} = 1 \) for all \( 1 \leq k \leq K^* \) and \( X_{t_i+1, s_i+1-1} = 1 \) for all \( 0 \leq i \leq K^* \). Furthermore, at the next iteration \( \beta^{(1)} = \beta^* + O_P(n^{-1/2}) \).

We defer the proof to the Appendix. We have the following remarks.

1. Note that each \( X \in \mathcal{X} \) partitions the nodes into \( K + 1 \) classes, given the partition the distribution of the edges follows a block model and the proofs utilize relevant techniques in this literature.
2. The theorem states that given an appropriate initial \( \beta^{(0)} \), the optimal configuration found by the algorithm includes all the TADs as well as the inter-TAD regions. In the next section, we propose a nonparametric test to check enriched interactions within each candidate region called by the algorithm.
3. More importantly, the same optimal \( X_0 \) is found for any choice of fixed \( K \), \( K \) being large enough. This implies the overfitting problem does not pose a serious concern here since increasing \( K \) does not always lead to an increase in the number of candidate TADs. In practice, a reasonable way to choose \( K \) is to increase it incrementally until the number of candidate TADs found starts to saturate.

2.4. Post-processing. After our algorithm detects the (possibly nested) TAD’s, our goal is to see if these indeed have higher contact frequencies than the surrounding region or the parent TAD. Recall that the contact frequency matrix \( M \) has non-negative weights which are truncated to generate the adjacency matrix of the network. These weights \( M_{ij} \), typically decay as \( d = |i - j| \) grows. In order to detect TAD’s with significantly enriched contact frequencies over the surrounding region, we assume the model in Equation 2.8. The main idea is that within a TAD, they decay slowly, whereas in the surrounding regions of a TAD they decay faster. Once we have detected the TADs using our linear program, we use these weights to
prune weakly connected TADs. Consider the base level; let us assume that we have identified a TAD between positions $a$ and $b$ on the genome. Let the upper triangular region of the this TAD be denoted by $R$. Now consider the upper triangular region of the square between $a - \frac{a-b}{2}$ and $b + \frac{a-b}{2}$. Denote this by $S$. We assume the following simple model that dictates how the weights decay within and outside a TAD. Consider two monotonically decaying functions $f, g : \mathbb{N} \to \mathbb{R}^+ \cup \{0\}$, such that $f(d) > g(d) \forall d \in \mathbb{N}$, i.e. $f(d)$ dominates $g(d)$ for any $d$.

Here $\epsilon_{ij}$ are pairwise independent noise random variables.

**Testing:** In order to perform a test, for all $d \in \{1, \ldots, (b-a)\}$, we calculate

\[
\hat{f}(d) = \frac{\sum_{|i-j|=d, i,j \in R} M_{ij}}{b - a + 1 - d}, \quad \hat{g}(d) = \frac{\sum_{|i-j|=d, i,j \in S \setminus R} M_{ij}}{b - a}
\]

Now we take the two sequences $\hat{f}$ and $\hat{g}$ and do a nonparametric rank test (two sample Wilcoxon test) to determine whether $\hat{f}$ dominates $\hat{g}$; if the p-value is smaller than a chosen threshold, we consider the TAD to have significant enrichment over its surrounding neighborhood. Otherwise we discard the TAD. For nested TADs, we are interested in determining whether a TAD found inside a parent TAD (call this $T_0$) is significant. In such cases, the surrounding region $S$ may go across $T_0$. So we simply truncate the outer region so that it does not cross outside $T_0$.

**3. Results.** We first demonstrate key properties of our inference algorithm via simulation experiments, and then provide elaborate real data results.

**3.1. Simulations.**
Fig 2: (a) Y axis shows the estimated number of clusters $K$, whereas the X axis shows increasing values of $K$. (b) shows the clustering for input $K = 30$.

Fig 3: Clusters identified by (a) LP-OPT and (b) SC. In (a) and (b) different colored squares correspond to different clusters detected by the algorithms. The ideal setting is to see a whole TAD encompassed by one square.

**Data simulated under the simple model.** First following the basic model described in Section 2 and Eq (2.1), we present two sets of experiments to a) show the robustness of our algorithm LP-OPT to the pre-specified number of clusters, and b) compare with the Spectral Clustering (SC) algorithm. For all the simulations in this setting, all TADs have the same linkage probability $\alpha$ and the background has linkage probability $\beta$.

In our first set of experiments (Figure 2 (a) and (b)), we show that with somewhat balanced (but not necessarily equal) block-sizes, LP-OPT returns
the correct TAD’s along with some holes, as shown in Theorem 2.5. Recall
that, in our linear program, we use a constraint to specify an upper bound on
the number of TADs. This constraint is given by \( \sum_{ij} \pi_{ij} \leq K \), where \( \sum_{ij} \pi_{ij} \)
represents the number of TADs. In Figure 2 (a) we plot \( \sum_{ij} \pi_{ij} \) after one
iteration of the linear program, for the adjacency matrix in Figure 2 (b).
To be concrete, we set \( n = 1000, \alpha = .2, \beta = .05 \), and three TADs of sizes
270, 200, and 220. We also created CTCF sites at every 10 nodes for this
experiment. We see that even though \( K \) is increased to thirty, the estimated
number of clusters levels off at 7, which is precisely three TADs plus four
inter-TAD regions, which illustrates our asymptotic result from Theorem 2.5.
These TADs detected by LP-OPT are illustrated in Figure 2 (b). While one
can come up with simple tests to eliminate the “spurious” TADs, we saw
that for real data, our post processing step (see Section 2.4) eliminates them
effectively for both the base level and nested TADs.
In the second set of simulations (Figure 3 (a), (b)), we show that SC often
yields clusters with holes, i.e. clusters that are not contiguous, whereas we
do not. SC is one of the most commonly used algorithms for community
detection in networks. It involves performing spectral decomposition on a
similarity matrix obtained from the data. For networks, one typically uses
the normalized adjacency matrix defined as \( D^{-1/2} A D^{-1/2} \) where \( D \) is the
diagonal matrix of degrees, i.e. \( D = diag(d_i), d_i = \sum_j A_{ij} \). Now for cluster-
ing the nodes into \( K \) blocks, one applies k-means clustering to the top
\( K \) eigenvectors (Rohe et al. (2011)). For Figure 3 (a) and (b), we set \( n = 240, \)
four TADs with sizes (70, 40, 30, 50). The fifth cluster is the background. We
use \( \alpha = .5, \beta = .25 \). In order to have a fair comparison, we do not include
CTCF sites for LP-OPT, since SC is not designed to use them either. For
both methods, we assume the correct number of blocks is given. In order to
be as favorable as possible to SC, we use 4 top eigenvectors, and use k-means
with \( k = 5 \) on these eigenvectors, since the background minus the TADs is
one cluster. The results from conventional SC (choosing 5 top eigenvectors
and using k-means with \( k = 5 \)) are worse and hence omitted. For SC the
plot reflects the clusterings returned: the colors correspond to different clus-
ters. A square corresponds to a maximal contiguous set of nodes assigned
to a cluster. For example, the last TAD (190-240) is assigned to the cyan
cluster by SC. However, SC also assigns some nodes from the penultimate
TAD (160-190) to this cluster, and moreover the small cyan boxes show that
there are many nodes from the last TAD, which are assigned to other clus-
ters, i.e. in this setting, SC is unable to create a contiguous cluster will all
nodes from one TAD.
We want to point out that while we did the above experiments for Figure 3
(a) without the CTCF sites for fairness, including the CTCF sites greatly improves the computational time of LP-OPT. To be concrete, we simulated 10 random networks with the above setting, and obtained the clusterings with and without the CTCF sites. With CTCF sites LP-OPT converges in 0.5 seconds on average, whereas without CTCF sites, the average computation time is 58 seconds.

In Section S1.1 of the supplementary file we include additional comparison with SC for varying signal to noise ratio.

Data simulated using real data distribution. We next used a more realistic framework to simulate Hi-C data for chromosome 21 at a resolution of 40 kb, a typical resolution at which Hi-C data are analyzed. TAD positions were generated artificially and contact frequencies were sampled using empirical distributions from a real Hi-C dataset on chromosome 21 provided in Rao et al. (2014). A detailed description of the framework can be found in Section S1.2 of the supplementary file. CTCF sites were generated as the union of the true TAD boundaries and randomly sampled positions along the chromosome.

Our procedure led to a 1204 × 1204 contact frequency matrix, which was processed using a moving window of length 300 with an overlap of 50. The contact frequencies in each 300 × 300 segment were thresholded at the \( q \)-th quantile to produce a binary adjacency matrix. Between two adjacent windows, any TADs called by the algorithm falling into overlapping regions are resolved as follows. i) If the end point of the TAD is the last CTCF site in the first window, it is extended to the first CTCF site in the second window (similarly if the start point of the TAD is the first CTCF site in the second window; ii) If one TAD is contained in another, the nested one is taken; iii) If two TADs have a significant overlap (Jaccard index > 0.8, defined in Section 3.2), they are merged by taking the intersection. A similar procedure is used on the real data (Section 3.2).

Table 1 compares the TADs found by our algorithm with ground truth using normalized mutual information (NMI) for different choices of the threshold \( q \) and an increasing number of randomly sampled CTCF positions. Note that the last column corresponds to the case where every position is a CTCF site, since the data generated contains 42 true TADs. In other words, we do not provide the algorithm with partial ground truth. As expected, the performance is better when partial ground truth is supplied but remains overall stable for reasonable choices of \( q \). Figure 4 displays a 24mb segment of the simulated data with TADs found by our method. LP-OPT was run without additional CTCF information and still achieved high similarity with ground
In comparison, under similar thresholding levels SC achieves a NMI around 0.86-0.89 when the correct cluster number $K = 43$ (the last cluster being the background) is given, and the TADs found contain holes as described above. In addition, we compare our method with two recently proposed TAD detection algorithms, 3DNetMod (Norton et al., 2018) and MrTADFinder (Yan et al., 2017), which are both based on community detection methods in network analysis. The best NMI achieved by MrTADFinder for a range of tuning parameter values is 0.55. 3DNetMod had difficulty finding TADs on this dataset. These two methods will be included for comparison in the subsequent real data analysis. More details and visual comparisons can be found in Section S1.2 of the supplementary file.

![Normalized mutual information](image)

<table>
<thead>
<tr>
<th># Random CTCF sites</th>
<th>Normalized mutual information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td>$q = 0.88$</td>
<td>0.90</td>
</tr>
<tr>
<td>$q = 0.9$</td>
<td>0.94</td>
</tr>
<tr>
<td>$q = 0.95$</td>
<td>0.97</td>
</tr>
<tr>
<td>$q = 0.98$</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 1: Normalized mutual information measuring the quality of the TADs found vs. ground truth.

(a) Ground truth  (b) LP-OPT

Fig 4: chr21:1-24000000, a 24mb segment. TADs identified by LP-OPT compared with ground truth. Note that LP-OPT was run without CTCF information thresholded at 95-th quantile.
3.2. **Real data.** Using the deep-coverage Hi-C data provided in Rao et al. (2014), we ran LP-OPT to identify cell-type specific TADs in five cell types (GM12878, HMEC, HUVEC, K562, NHEK) and common TADs conserved in all of them. We present here a comprehensive analysis of the results from chromosome 21. Similar analysis was also performed on chromosome 1, the results of which can be found in Section S2.2 of the supplementary file. Following Rao et al. (2014), the raw contact frequency matrix was normalized using the matrix balancing algorithm in Knight and Ruiz (2013). Using data with 10kb resolution, the contact frequency matrix of this chromosome has more than 4800 bins. CTCF peaks for each cell type were obtained from the ENCODE pipeline (Consortium et al., 2012) and converted into a binary vector of the same resolution as the contact frequency matrix, where each entry represents whether or not the corresponding genome bin contains at least one CTCF peak. This led to around 900 non-zero entries in each cell type. In the combined analysis for common TADs, we took the intersection of the cell-type specific CTCF binary vectors, so an entry is one only when the genome bin contains at least one CTCF peak in all cell types.

We performed TAD calling for three levels, each level with its own quantile threshold parameter. At the base level, we processed the chromosome using a moving window of length 300 (3mb) with an overlap of 50. The contact frequencies in each 300 x 300 segment was thresholded at 90% quantile ($q_1 = 0.9$) to produce a binary adjacency matrix. Note that by using a moving window, we avoided using one universal threshold for the entire chromosome, which contains active and inactive regions with different chromatin interaction patterns. Any overlaps between two adjacent windows are resolved using the rules described in Section 3.1. The TADs called at the base level were then post-processed using the nonparametric test described in Section 2.4, and only those passing a p-value cutoff (in this case 0.05) were retained for further TAD calling. For the second level, we thresholded the contact frequencies inside the base-level TADs at 50% quantile ($q_2 = 0.5$), followed by running the algorithm and post-processing. The same steps were followed for the third level with $q_3 = 0.5$. For all three levels, the p-value cutoff was chosen to be 0.05. As a side note, correcting for multiple testing at a false discovery rate (FDR) of 0.05 made almost no difference at the base level. However, the same FDR cutoff led to fewer TADs being called at the nested levels. This is unsurprising as the power of the nonparametric test decreases as the number of data points available decreases at the nested levels.

The combined analysis for conserved TADs was performed in the same way, using the algorithm described in Section 2.2. The nonparametric test
was run on the called regions for all cell types, and we required all the p-values to be smaller than the cutoff 0.05.

**Choice of thresholds.** We first checked the robustness of the results using different thresholding levels and biological replicates. Table 2 shows the number of TADs identified under different scenarios and with significant overlap. To compare two TADs $S, T$ from two different sets, we measure the Jaccard index $J(S, T) = \frac{|S \cap T|}{|S \cup T|}$. When the Jaccard index is high enough, there is a one-to-one correspondence between TADs in the two sets. The first two rows in the table show different thresholds at the base level still lead to quite consistent results. Varying $q_2, q_3$ between 0.4-0.6 does not lead to noticeable changes and the results are hence omitted. Since two biological replicates (primary and replicate) are available for GM12878, we examined the consistency between them and the combined data, and the results are shown in row 3 and 4 of the table. Finally, as the current results were obtained using normalized data, we compared them with the case using the raw contact frequency matrix (row 5). This case still shows a reasonable degree of consistency despite having the lowest amount of overlap among all.

<table>
<thead>
<tr>
<th># TADs</th>
<th>$q_1 = 0.85$ (GM12878)</th>
<th>$q_1 = 0.9$ (GM12878)</th>
<th>Jaccard index $&gt; 0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85</td>
<td>81</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>$q_1 = 0.85$ (HMEC)</td>
<td>$q_1 = 0.9$ (HMEC)</td>
<td>Jaccard index $&gt; 0.7$</td>
</tr>
<tr>
<td></td>
<td>123</td>
<td>114</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>Primary (GM12878)</td>
<td>Replicate (GM12878)</td>
<td>Jaccard index $&gt; 0.7$</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>83</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Primary (GM12878)</td>
<td>Combined (GM12878)</td>
<td>Jaccard index $&gt; 0.7$</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>81</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Normalized (GM12878)</td>
<td>Raw (GM12878)</td>
<td>Jaccard index $&gt; 0.7$</td>
</tr>
<tr>
<td></td>
<td>81</td>
<td>94</td>
<td>61</td>
</tr>
</tbody>
</table>

**Table 2**

Number of TADs detected under different scenarios and with significant overlap

**Enrichment of histone marks at boundaries.** One of the most commonly used criteria for checking the accuracy of TAD boundaries is to count the number of histone modification peaks nearby (Filippova et al., 2014; Weinreb and Raphael, 2016) and taking higher levels of histone activity as indicators for the start and end points of TADs. The histone data are available in Kellis et al. (2014) and the processed data were downloaded from https://sites.google.com/site/anshulkundaje/projects/encodehistonemods. From bin indices, we obtain the coordinates of TAD boundaries by taking the mid-
point of every genome bin. Table 3 shows the average number of peaks within 15kb upstream or downstream from each detected boundary point for various types of histone modification. We compared LP-OPT with 3DNetMod Norton et al. (2018), MrTADFinder Yan et al. (2017), and the Arrowhead domains originally reported in Rao et al. (2014). We found that MrTADFinder produced domains quite different from the other three methods (Fig S3 in the supplementary file) and the domain boundaries show less enrichment of histone marks. We have thus omitted the method from further comparison. The tuning parameters for 3DNetMod were chosen so that the number of TADs found is roughly comparable to the other two methods. In Table 3, counting the number of times each method achieves the highest enrichment, LP-OPT and 3DNetMod outperform Arrowhead with LP-OPT being slightly better than 3DNetMod. In addition, we note that LP-OPT is significantly faster than 3DNetMod, taking about 10 minutes on chromosome 21 (and 40 minutes on chromosome 1) using one core on a 3.1 GHz Intel Core i5 processor. In comparison, 3DNetMod takes more than 40 minutes on chromosome 21 (and 4 hours on chromosome 1) requiring four cores on the same processor. The results of 3DNetMod are also quite sensitive to the choice of tuning parameters.

Conserved and cell-type specific TADs. Although commonly used, the metric in Table 3 does not consider epigenetic features inside each TAD, which are particularly important for confirming shared regulatory structures and mechanisms across different cell types. We first examine the histone modification peaks within highly conserved TADs, which are defined as i) TADs identified in the combined analysis of all cell types (denote this set $I_c$), and ii) if for $S \in I_c$, $\max_{T \in I_i} J(S, T) > 0.7$ for all $i$, where $I_i$ is the set of TADs identified in cell type $i$. Out of the 50 TADs found in the combined analysis, 29 of them satisfy ii).

Figure 5 shows the signal tracks for all five cell types inside one of the 29 conserved TADs (chr21:35275000 - 35725000) for two types of histone modifications (UCSC genome browser). The signal peaks are visibly correlated between cell types. Using ChIP-seq signals from the ENCODE pipeline, the average pairwise correlations between cell types for this TAD were calculated for different histone modifications. For H3k9ac, H3k27ac, H3k4me1 and H3k4me3, the average correlations are 0.69, 0.80, 0.58, 0.89 respectively. Figure 6 compares the average pairwise correlations inside all 29 conserved TADs with 50 randomly chosen regions of length 290kb (median length of the conserved TADs) on chromosome 21 for two instances of histone modification. The two-sample Wilcoxon test has p-values 0.05 and 0.006 for
H3k27ac and H3k4me1; the results for H3k9ac and H3k4me3 are similar.

Having analyzed TADs with consistent overlaps across all cell types, we now consider TADs which are specific to individual cell types. A TAD is considered specific to that cell type \( i \) if i) \( S \in I_i \); ii) \( \max_{T \in I_j} J(S, T) < 0.4 \) for all \( j \neq i \). This criterion leads to 28 TADs, each specific to one of the cell types. The median length of these TADs is 210kb, smaller than that of the conserved TADs. As an illustration, Figure 7 shows the histone modification tracks inside two TADs specific to K562 and GM12878 respectively. In these two regions, the histone modifications show a higher level of activity for the two specific cell types. To evaluate whether this is a systematic trend, we next calculated the total signal level for each of the 28 TADs under different types of histone modifications. For each type, we counted the number of TADs which have the highest total signal level in the cell type they are associated with. Comparing to a null distribution under which the cell type with the highest total signal levels is selected randomly, we computed the

<table>
<thead>
<tr>
<th></th>
<th>GM12878</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># domains</td>
<td>H3k9ac</td>
<td>H3k27ac</td>
<td>H3k4me1</td>
<td>H3k4me3</td>
</tr>
<tr>
<td>LP-OPT</td>
<td>81</td>
<td>1.35</td>
<td>1.68</td>
<td>1.16</td>
<td>1.29</td>
</tr>
<tr>
<td>Arrowhead</td>
<td>96</td>
<td>1.40</td>
<td>1.60</td>
<td>1.29</td>
<td>1.22</td>
</tr>
<tr>
<td>3DNetMod</td>
<td>129</td>
<td>1.36</td>
<td><strong>1.82</strong></td>
<td>1.25</td>
<td>1.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>HUVEC</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># domains</td>
<td>H3k9ac</td>
<td>H3k27ac</td>
<td>H3k4me1</td>
<td>H3k4me3</td>
<td>Pol II</td>
</tr>
<tr>
<td>LP-OPT</td>
<td>106</td>
<td><strong>1.07</strong></td>
<td><strong>1.16</strong></td>
<td><strong>2.17</strong></td>
<td><strong>0.84</strong></td>
<td></td>
</tr>
<tr>
<td>Arrowhead</td>
<td>59</td>
<td>1.02</td>
<td>1.08</td>
<td>2.06</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>3DNetMod</td>
<td>125</td>
<td>0.96</td>
<td>1.08</td>
<td>1.82</td>
<td>0.68</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>HMEC</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># domains</td>
<td>H3k9ac</td>
<td>H3k27ac</td>
<td>H3k4me1</td>
<td>H3k4me3</td>
<td>Pol II</td>
</tr>
<tr>
<td>LP-OPT</td>
<td>114</td>
<td>1.06</td>
<td><strong>1.49</strong></td>
<td>3.05</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>Arrowhead</td>
<td>44</td>
<td>1.02</td>
<td>1.46</td>
<td><strong>3.09</strong></td>
<td><strong>0.81</strong></td>
<td></td>
</tr>
<tr>
<td>3DNetMod</td>
<td>122</td>
<td>0.92</td>
<td>1.24</td>
<td>2.67</td>
<td>0.77</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NHEK</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># domains</td>
<td>H3k9ac</td>
<td>H3k27ac</td>
<td>H3k4me1</td>
<td>H3k4me3</td>
<td>Pol II</td>
</tr>
<tr>
<td>LP-OPT</td>
<td>112</td>
<td>1.21</td>
<td>1.32</td>
<td>2.70</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Arrowhead</td>
<td>78</td>
<td>0.99</td>
<td>1.12</td>
<td>2.19</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>3DNetMod</td>
<td>136</td>
<td><strong>1.42</strong></td>
<td><strong>1.55</strong></td>
<td><strong>2.91</strong></td>
<td><strong>1.03</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>K562</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># domains</td>
<td>H3k9ac</td>
<td>H3k4me1</td>
<td>H3k4me3</td>
<td>Pol II</td>
<td></td>
</tr>
<tr>
<td>LP-OPT</td>
<td>91</td>
<td>0.76</td>
<td><strong>2.31</strong></td>
<td><strong>0.98</strong></td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Arrowhead</td>
<td>82</td>
<td>0.57</td>
<td>1.82</td>
<td>0.82</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>3DNetMod</td>
<td>101</td>
<td><strong>0.79</strong></td>
<td>2.25</td>
<td>0.95</td>
<td><strong>0.71</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3**

Average number of histone modification peaks ±15kb upstream or downstream from the boundary points.
Fig 5: histone signal tracks within chr21:35275000 - 35725000

(a) Histone modification H3k27ac, average pairwise correlation 0.80

(b) Histone modification H3k4me3, average pairwise correlation 0.89

Fig 6: Comparing conserved TADs with random regions on chr21; pairwise correlations between all cell types for a) H3k27ac and b) H3k4me1.

p-values using a binomial distribution in Table 4. This suggests the cell-type specific TADs tend to be regions with more active histone modifications.

Without CTCF information. As a final remark, we tested whether LP-OPT could reproduce consistent results without CTCF information. Applying the algorithm to a 5mb segment of chromosome 21 (chr21:26000000-31000000) using GM12878 data, 15 TADs were identified (vs. 13 TADs with
Table 4

<table>
<thead>
<tr>
<th></th>
<th>H3k9ac</th>
<th>H3k27ac</th>
<th>H3k4me1</th>
<th>H3k4me3</th>
</tr>
</thead>
<tbody>
<tr>
<td># TADs with the highest total signal level</td>
<td>13</td>
<td>16</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>p-value</td>
<td>$1.5 \times 10^{-3}$</td>
<td>$1.7 \times 10^{-5}$</td>
<td>$4.2 \times 10^{-7}$</td>
<td>$3.9 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

For each type of histone modification, the number of TADs (out of 28) such that they have the highest total signal level in the cell type they are associated with.

CTCF covariate) at the same p-value cutoffs. 11 pairs of these have a Jaccard index greater than 0.7, suggesting the results are reasonably stable. However, we also note the computational time in this case is significantly longer, as the search space for optimization is considerably larger without incorporating the CTCF sites.

4. Discussion. The 3D structure of chromatin provides key information for understanding the regulatory mechanisms. Recently, technologies such as Hi-C have revealed the existence of an important type of chromatin structure known as TADs, which are regions with enriched contact frequency and have been shown to act as functional units with coordinated regulatory actions inside. In this paper, we propose a statistical network model to identify TADs treating genome segments as nodes and their interactions in 3D
as edges. Unlike many traditional networks with exchangeable distributions, our model incorporates the linear ordering of the nodes and guarantees the communities found represent contiguous regions on the genome. Our method also achieves two important biological goals: i) Considering the empirical observation that TADs boundaries tend to correlate with CTCF binding sites, our method offers the flexibility to include CTCF binding data (or other ChIP-seq data) as biological covariates. ii) The likelihood-based approach allows for joint inference across multiple cell types. On the theoretical side, we have shown asymptotic convergence of the estimation procedure with appropriate initializations. In practice, we observe the algorithm always converges in a few iterations. Due to the linear nature of the algorithm, our method is computationally efficient; it takes less than 10 minutes to complete the computation on chr21 with CTCF information on a laptop, whereas methods like TADtree (Weinreb and Raphael, 2016) can take up to hours.

Some areas for future work include extending the theoretical analysis to increasing $K$, and considering modelling higher order interactions between TADs. Our current way of finding conserved and cell-type specific TADs involves computing overlaps between domains and choosing heuristic cutoffs. While we have shown using epigenetic features that the conserved and cell-type specific TADs found have desirable features, it would be more ideal to statistically model the extent of overlaps between different types of TADs.

Acknowledgements. The authors would like to thank Nathan Boley for helpful initial discussions. This research was funded in part by the ARC DECRA Fellowship to Y.X.R. Wang, the HHMI International Student Research Fellowship and Stanford Gabilan Fellowship to O. Ursu, NSF Grant DMS 17130833 and NIH Grant 1U01HG007031-01 to P. Bickel.

SUPPLEMENTARY MATERIAL

Supplement: Supplementary Information for “Network modelling of topological domains using Hi-C data” (http://). Additional simulations and real data results can be found in the supplementary file.

APPENDIX A: PROOFS

Each $X \in \mathcal{X}$ partitions the nodes into $K + 1$ classes, thus we define the corresponding node labels as $Z = (Z_1, \ldots, Z_n)$, with $Z_i = k$ if $Z_i \in [s_k, t_k]$, $Z_i = K^* + 1$ if $Z_i$ does not fall inside any TAD. The set of feasible $Z$ is a subset of $\{1, \ldots, K^* + 1\}^n$ and can be seen as the latent node labels in a block model. Let $\mathcal{X}$ and $\mathcal{Z}$ be the feasible sets for $X$ and $Z$ respectively.
$X^*$ and $Z^*$ are the true latent positions and the corresponding node labels. Following block model notations, define a $(K^* + 1) \times (K^* + 1)$ matrix $H^*$, where $H_{k,k} = \alpha_{s_k,t_k}^*$ for $1 \leq k \leq K^*$, and $H_{k,l}^* = \beta^*$ otherwise. For any label $Z$, let $R(Z, Z^*)$ be the confusion matrix with

$$R_{k,l}(Z, Z^*) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i = k, Z^*_i = l).$$

Finally set $E = n \times n$ matrix of 1.

With appropriate concentration, it suffices to consider $l(A; \pi, \beta)$ at expectation $E(A)$. Define

$$G(R, \beta) = \sum_{k=1}^{K^*} (R^2)_{k,k} KL \left( \frac{(RH^* R^T)_{k,k}}{(R^2)_{k,k}}, \beta \right)$$

for some $Z \in Z$ and its corresponding $R$. For simplicity of notation, we assume $A$ has diagonal entries generated in the same way as non-diagonal entries, which does not affect the asymptotic results. We have the following lemma for the maximum of $G(\cdot, \beta)$.

**Lemma A.1.** Suppose $\beta$ satisfies Assumptions 2.3 and 2.4. Then for all $K \geq K^* + G^*$ and $n$ large enough,

$$\max_{Z \in Z} G(R(Z, Z^*), \beta) = \frac{1}{n^2} \sum_{k=1}^{K^*} (n_k^*)^2 KL(\alpha_k^* \| \beta) + \frac{1}{n^2} \sum_{i=0}^{K^*} (s_{i+1} - t_i - 1)^2 KL(\beta^* \| \beta).$$

The maximum is unique at $R_0$ such that $X_{s_k, t_k} = 1$ for all $1 \leq k \leq K^*$ and $X_{t_i, s_{i+1}} = 1$ for all $0 \leq i \leq K^*$. Furthermore, for any $R_1 \neq R_0$ and $n$ large enough,

$$\left. \left. \frac{\partial G((1 - \epsilon)R_0 + \epsilon R_1, \beta)}{\partial \epsilon} \right|_{\epsilon=0^+} \leq -C < 0$$

for some $C > 0$.

**Proof.** For each feasible $Z$, let $\{[l_1, m_1], \ldots, [l_K, m_K]\}$ be the corresponding TAD positions defined by $Z$. For each row of $R(Z)$,

$$\frac{(R^2)_{k,k}}{(R^2)_{k,k}} KL \left( \frac{(RH^* R^T)_{k,k}}{(R^2)_{k,k}}, \beta \right)$$

$$\leq \max \left\{ \left( \frac{(m_k - l_k)^2}{n^2} - \sum_{i=1}^{K^*} R_{ki}^2 \right) KL(\beta^* \| \beta), \sum_{i=1}^{K^*} R_{ki}^2 KL(\alpha_k^* \| \beta) \right\}$$

(A.3)
by Assumption 2.3 and the convexity of $K(\cdot\|\beta)$. Also for the $k$th row of $R$, define

$$i_k = \min\{i : [s_i, t_i] \cap [l_k, m_k] \neq \emptyset\}, \quad j_k = \max\{i : [s_i, t_i] \cap [l_k, m_k] \neq \emptyset\}.$$  

(A.4)

We first consider the case where the set above is nonempty. For two adjacent rows $k$ and $k + 1$, it suffices to consider the case $j_k = i_{k+1}$. Denote \(S_{k,k+1} = \sum_{l=k}^{k+1} (RER^T)_{i,l} KL\left(\frac{(RER^T)_{i,j}}{(RER^T)_{i,j}} \| \beta\right).\) By (A.3), $S_{k,k+1}$ is upper bounded by one of the following:

1. \((m_{k+1} - l_k)^2/n^2 - \sum_{q=k}^{k+1} \sum_{l=i_q}^{j_q} R_{q,l}^2\) $KL(\beta^*\|\beta)$.

2. \[\sum_{l=i_k}^{j_k-1} R_{k,l}^2 KL(\alpha^*_l\|\beta) + R_{k,j_k}^2 KL(\alpha^*_{j_k}\|\beta) + \left((m_{k+1} - l_k)^2/n^2 - \sum_{l=i_{k+1}}^{j_{k+1}} R_{k+1,l}^2\right) KL(\beta^*\|\beta),\]

which is itself upper bounded by

\[\sum_{l=i_k}^{j_k-1} R_{k,l}^2 KL(\alpha^*_l\|\beta) + \max\left\{\left(\frac{(m_{k+1} - s_{j_k})^2 - (n_{j_k})^2}{n^2} - \sum_{l=i_{k+1}}^{j_{k+1}} R_{k+1,l}^2\right) KL(\beta^*\|\beta),\right\}\]

3. \[\left((m_k - l_k)^2/n^2 - \sum_{l=i_k}^{j_k} R_{k,l}^2\right) KL(\beta^*\|\beta) + R_{k+1,i_{k+1}}^2 KL(\alpha^*_{i_{k+1}}\|\beta) + \sum_{l=i_{k+1}+1}^{j_{k+1}} R_{k+1,l}^2 KL(\alpha^*_l\|\beta).\]

Similar to the case above, this is bounded by \[\sum_{l=i_{k+1}+1}^{j_{k+1}} R_{k+1,l}^2 KL(\alpha^*_l\|\beta) + \max\left\{\left(\frac{(t_{j_k} - l_k)^2 - (n_{j_k})^2}{n^2} - \sum_{l=i_k}^{j_k-1} R_{k,l}^2\right) KL(\beta^*\|\beta),\right\}\]
\[
\left( \frac{n^*}{n} \right)^2 KL(\alpha_j^* \| \beta) + \left( \frac{(s_j - l_k)^2}{n^2} - \sum_{l=i_k}^{j_k-1} R_{k,l}^2 \right) KL(\beta^* \| \beta) \right). 
\]

4. \( \sum_{q=k}^{j_k} \sum_{l=i_q}^{j_q} R_{q,l}^2 KL(\alpha_q^* \| \beta) \).

If the set in (A.4) is empty,

\[
(RER^T)_{k,k} KL \left( \frac{(RH^* R^T)_{k,k}}{(RER^T)_{k,k}} \right) = \left( \frac{m_k - l_k}{n} \right)^2 KL(\beta^* \| \beta) 
\leq (s_{l+1} - t_l)^2 KL(\beta^* \| \beta)
\]

for some \( 1 \leq l \leq K^* \).

The above cases show for any \( Z \in \mathcal{Z} \), an upper bound for \( G(R(Z, Z^*), \beta) \) is of the form

\[
\sum_{k=1}^{L} \left( \frac{(s_{j_k} - t_{i_k})^2}{n^2} - \sum_{i_k < l < j_k} (n_l^*)^2 \right) KL(\beta^* \| \beta) + \sum_{l \in \mathcal{I}} (n_l^*)^2 KL(\alpha_l^* \| \beta),
\]

where \( \mathcal{I} \) is an index set such that \( \mathcal{I} \cap \bigcup_{k=1}^{L} [i_k, j_k] = \emptyset \). By Assumption 2.3, this is bounded by

\[
\frac{1}{n^2} \sum_{k=1}^{K^*} (n_k^*)^2 KL(\alpha_k^* \| \beta) + \frac{1}{n^2} \sum_{l=0}^{K^*} (s_{l+1} - t_l - 1)^2 KL(\beta^* \| \beta)
\]

with equality achieved only at \( R_0 \) for any \( K \geq K^* + G^* \). The second part of the lemma can be checked with differentiation.

Let \( [l_k, m_k] \) be the \( k \)-th domain in a configuration \( Z \) corresponding to the \( k \)-th row in \( R \). Next we state a concentration lemma for the averages \( \hat{\alpha}_{l_k, m_k}(Z) \). Denote \( O_{l_k, m_k}(Z) = (m_k - l_k)^2 \hat{\alpha}_{l_k, m_k}(Z) \) and \( \Delta_k(Z) = O_{l_k, m_k}(Z)/n^2 - (RH^* R^T(Z))_{k,k} \).

**Lemma A.2.** For \( \epsilon \leq 3 \),

\[
P \left( \max_{Z \in \mathcal{Z}} \max_{1 \leq k \leq K} |\Delta_k(Z)| \geq \epsilon \right) \leq 2(K)^n \exp \left( -C_1(H^*)^2 \epsilon^2 n^2 \right).
\]

(A.6)

Let \( Z_0 \in \mathcal{Z} \) be a fixed set of labels, then for \( \epsilon \leq 3m/n \),

\[
P \left( \max_{Z: |Z-Z_0| \leq m} \max_{1 \leq k \leq K} |\Delta_k(Z) - \Delta_k(Z_0)| > \epsilon \right)
\]
\[ \leq \frac{n}{m}(K)^{m+1}\exp(-C_2(H^*)\frac{n^3\epsilon^2}{m}). \quad (A.7) \]

\(C_1(H^*)\) and \(C_2(H^*)\) are constants depending only on \(H^*\).

**Proof.** The proof follows from Bickel and Chen (2009) with minor modifications. \(\square\)

**Proof of Theorem 2.5.** Suppose \(\beta^{(0)}\) satisfies Assumptions 2.3 and 2.4. We consider the most general setup where every position is a CTCF binding site. The likelihood objective is given by

\[ l(A; Z, \beta^{(0)}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \neq j} A_{ij} \log \frac{\hat{\alpha}_{l_k,m_k}(Z)(1 - \beta^{(0)})}{(1 - \hat{\alpha}_{l_k,m_k}(Z))\beta^{(0)}} + \log \frac{1 - \hat{\alpha}_{l_k,m_k}(Z)}{1 - \beta^{(0)}} \]

\[ = \frac{1}{2} \sum_{k=1}^{K} (m_k - l_k)^2 KL(\hat{\alpha}_{l_k,m_k}(Z)\|\beta^{(0)}). \quad (A.8) \]

Let \(R_0\) (and the corresponding \(X_0, Z_0\)) be the optimal configuration in Lemma A.1.

We first consider \(X\) far away from \(X_0\). Define

\[ I_{\delta_n} = \{X \in \mathcal{X} : G(R(X), \beta^{(0)}) - G(R_0, \beta^{(0)}) < -\delta_n\}, \]

where \(\delta_n\) is a sequence converging to 0 slowly. First by (A.6) in Lemma A.2,

\[ \left| l(A; X, \beta) - n^2G(R(X), \beta) \right| \leq C n^2 \sum_{k=1}^{K} \left| \frac{O_{l_k,m_k}(X)}{n^2} - (RH^*R^T(X))_{k,k} \right| \]

\[ = o_p(n^{2-\gamma}) \quad (A.9) \]

for some \(\gamma < 1/2\). It follows then

\[ l(A; X, \beta^{(0)}) - l(A; X_0, \beta^{(0)}) \leq o_p(n^{2-\gamma}) - n^2\delta_n, \]

and

\[ \exp \left\{ \max_{X \in I_{\delta_n}} l(A; X, \beta^{(0)}) - l(A; X_0, \beta^{(0)}) \right\} \quad (A.10) \]
\[
\begin{align*}
&\leq \sum_{X \in I_{\delta_n}} \exp \left\{ l(A; X, \beta^{(0)}) - l(A; X_0, \beta^{(0)}) \right\} \\
&\leq \exp(o_P(n^{-2\gamma}) - n^2 \delta_n + n \log K) = o_P(1) \quad (A.11)
\end{align*}
\]

choosing \( \delta_n \to 0 \) slowly enough.

Next consider the case \( X \in I_{\delta_n}^c \) and \( X \neq X_0 \). By (A.7) in Lemma A.2,

\[
\begin{align*}
&\mathbb{P} \left( \max_{X \neq X_0} \left| \Delta(Z) - \Delta(Z_0) \right| > \epsilon \frac{|Z - Z_0|}{n} \right) \\
&\leq \sum_{m=1}^{n} \mathbb{P} \left( \max_{Z: |Z-Z_0|=m} \left| \Delta_k(Z) - \Delta_k(Z_0) \right| > \frac{\epsilon m}{n} \right) \\
&\leq \sum_{m=1}^{n} 2n^m K^{m+1} \exp \left( -Cmn \right) \to 0. \quad (A.12)
\end{align*}
\]

It follows then if \( |Z - Z_0| = m \), \( \frac{\| \Delta(Z) - \Delta(Z_0) \|_\infty}{m/n} = o_P(1) \), and \( \frac{1}{n^2} \| O(Z) - O(Z_0) \|_\infty \geq \frac{m}{n} (C + o_P(1)) \) since \( \| RH^*R^T(Z) - RH^*R^T(Z_0) \|_\infty \geq C \frac{m}{n} \). Note that in the set \( I_{\delta_n}^c \), \( |Z - Z_0| \to 0 \). Then (A.2) implies

\[
G(R(Z), \beta^{(0)}) - G(R_0(Z_0), \beta^{(0)}) < -C \frac{m}{n} \quad (A.13)
\]

if \( |Z - Z_0| = m \). Since \( G(R, \beta^{(0)}) \) is the population version of \( \frac{1}{n^2} l(A; R, \beta^{(0)}) \) and \( O(Z)/n^2 \) approaches \( RH^*R^T(Z) \) uniformly in probability, by the continuity of the derivative,

\[
\frac{1}{n^2} \left( l(A; R(Z)_0), \beta^{(0)} \right) - l(A; R(Z), \beta^{(0)}) = \Omega_P(m/n) \quad (A.14)
\]

for \( |Z - Z_0| = m \). It follows then

\[
\begin{align*}
&\exp \left\{ \max_{X \in I_{\delta_n}^c, X \neq X_0} l(A; X, \beta^{(0)}) - l(A; X_0, \beta^{(0)}) \right\} \\
&\leq \sum_{X \in I_{\delta_n}^c, X \neq X_0} \exp \left\{ l(A; X, \beta^{(0)}) - l(A; X_0, \beta^{(0)}) \right\} \\
&\leq \sum_{m=1}^{n} \left( \frac{n}{m} \right) (K + 1)^m e^{-\Omega_P(mn)} = o_P(1) \quad (A.15)
\end{align*}
\]
REFERENCES

Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–
girvan and other modularities. Proceedings of the National Academy of Sciences, 106
Irineo Cabreros, Emmanuel Abbe, and Aristotelis Tsirigos. Detecting community struc-
tures in hi-c genomic data. In Information Science and Systems (CISS), 2016 Annual
ENCODEx Project Consortium et al. An integrated encyclopedia of dna elements in the
Jesse R Dixon et al. Topological domains in mammalian genomes identified by analysis
Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative
Chunhui Hou, Li Li, Zhaohui S Qin, and Victor G Corces. Gene density, transcription, and
insulators contribute to the partition of the drosophila genome into physical domains.
Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. IMA Journal of
François Le Dily et al. Distinct structural transitions of chromatin topological domains
 correlate with coordinated hormone-induced gene regulation. Genes & development, 28
Celine Lévy-Leduc, Maud Delattre, Tristan Mary-Huard, and Stephane Robin. Two-
dimensional segmentation for analyzing hi-c data. Bioinformatics, 30(17):i386–i392,
2014.
Erez Lieberman-Aiden et al. Comprehensive mapping of long-range interactions reveals
Darío G Lupiáñez et al. Disruptions of topological chromatin domains cause pathogenic
Laraib Iqbal Malik and Rob Patro. Rich chromatin structure prediction from hi-c data.
Karen J Meaburn, Prabhukar R Gudla, Sameena Khan, Stephen J Lockett, and Tom
Mistel. Disease-specific gene repositioning in breast cancer. The Journal of cell biology,
Elphege P Nora et al. Spatial partitioning of the regulatory landscape of the x-inactivation
Heidi K Norton, Daniel J Emerson, Harvey Huang, Jesi Kim, Katelyn R Titus, Shi Gu,
Danielle S Bassett, and Jennifer E Phillips-Cremins. Detecting hierarchical genome
Suhas SP Rao et al. A 3d map of the human genome at kilobase resolution reveals
Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional
Adrian L Sanborn, Suhas SP Rao, Su-Chen Huang, Neva C Durand, Miriam H Huntley,
Andrew I Jewett, Ivan D Bochkov, Dharmaraj Chinnappan, Ashok Outosky, Jian Li,
et al. Chromatin extrusion explains key features of loop and domain formation in wild-
type and engineered genomes. Proceedings of the National Academy of Sciences, 112


