Semi-nonparametric Modeling of Topological Domain Formation From Epigenetic Data

Emre Sefer and Carl Kingsford

School of Computer Science, Carnegie Mellon University
{esefer,carlk}@cs.cmu.edu

Abstract. Hi-C experiments capturing the 3D genome architecture have led to the discovery of topologically-associated domains (TADs) that form an important part of the 3D genome organization and appear to play a role in gene regulation and other functions. Several histone modifications have been independently suggested as the possible explanations of TAD formation, but their combinatorial effects on domain formation remain poorly understood at a global scale. Here, we propose a convex semi-nonparametric approach called *nTDP* based on Bernstein polynomials to explore the joint effects of histone markers on TAD formation as well as predict TADs solely from the histone data. We find a small subset of modifications to be predictive of TADs across species. By inferring TADs using our trained model, we are able to predict TADs across different species and cell types, without the use of Hi-C data, suggesting their effect is conserved. This work provides the first comprehensive joint model of the effect histone markers on domain formation.

1. Introduction

The emerging evidence suggests that 3D nuclear architecture is important for the regulation of gene expression and it is tightly linked to the function of the genome. For instance, expression in the beta-globin locus is mediated by folding to bring an enhancer and associated transcription factors within close proximity of a gene [28, 2]. Similarly, loci of mutations that affect expression of genomically far-away genes (eQTLs) are significantly closer in 3D to their regulated genes [7], indicating that 3D genome structure plays a wide-spread role in gene regulation. Lastly, spatial regions that interact with nuclear lamina are generally inactive [12]. Measuring and modeling the 3D shape of a genome is thus essential to obtain a more complete understanding of how cells function.

Chromatin interactions obtained from a variety of recent chromosome conformation capture experimental techniques such as Hi-C [18] have resulted in significant advances in our understanding of the geometry of chromatin structure [11, 25]. These experiments yield matrices of counts that represent the frequency of cross-linking between restriction fragments of DNA at a certain resolution. Analysis of the resulting matrix by Dixon et al. [6] led to the discovery of topologically-associated domains (TADs) which correspond to consecutive, highly-interacting matrix regions typically a few megabases in size that are closely embedded in 3D.

TADs have been identified across different cell cycle phases and in prokaryotes [16]. Several lines of evidence suggest that TADs are a building block of genomic regulatory architecture [15, 27]. Segmental packaging of genome via TADs likely have critical roles in cell dynamics such as long-range transcriptional regulation and cell differentiation [23, 24].

The mechanism by which these TADs form and are demarcated is still largely unknown. A plethora of epigenetic modifications have been identified in metazoan genomes that are associated with 3D genome shape [5], and thus TADs. Several modifications have been found to be specifically correlated with TAD boundaries [6]. For instance, histone modifications with insulator roles such as H3K4me3 and H3K27ac are enriched within TAD boundaries [26], although the causal direction of these associations is still unknown [23]. Despite these analyses, the complete picture of how histone modifications are related to TAD formation is missing. This is partially because previous analyses relating histone marks to domain boundaries have often considered each histone mark independently, without accounting for their combined affects. It is unknown to what extent relationships between the histone markers are important or whether there is a small set of markers that are of primary importance.

Here, we develop and train a joint model, which we call *nTDP*, of how histone modifications are associated with domain boundaries and interiors. We show that we are able to train this model optimally in polynomial time because its likelihood function is convex. The model does not make any assumptions about the effect of each histone mark on domain formation, and instead fits the histone-domain relationship nonparametrically. Using this model, we systematically identify a small set of histone markers that in combination appear to explain TAD boundaries. We find a small number of epigenetic elements account for a large proportion of the accuracy of TAD prediction. All of these identified marks fail to predict domain boundaries when considered independently. We show that these markers are conserved across species and cell types in a very strong way: models trained on mouse continue to work well on human, and models trained on IMR90 cells continue to work on embryonic stem cells.

Our approach, *nTDP*, can form the basis of a unified, explanatory model of the relationship between epigenetic marks and topological domain structures. It can be used to predict domain boundaries for cell types, species, and conditions for which no Hi-C data is available. The model may also be of use for improving Hi-C-based domain finders.

1.1 Additional related work

Previous work mainly focused on analyzing epigenetic data in an unsupervised way. Segway [14] and ChromHMM [9] take as input a collection of genomics datasets and learn chromatin states that exhibit similar epigenetic activity patterns which then have different interpretations such as transcriptionally active, Polycomb-repressed. Libbrecht et al. [17] improve Segway predictions by integrating Hi-C data which is not as abundant as histone data, whereas [13] jointly

infers chromatin state maps in multiple genomes by a hierarchical model. However, none of these methods deal directly with TADs. Even though a subset of their chromatin states overlap with TADs, predicting TADs from them heuristically does not perform well. Additionally, they either ignore the histone densities, or make parametric distribution assumptions such as geometric or normal which are not always reflected in the true data. When modified to run in a supervised setting, they cannot capture the most informative subset of epigenetic elements.

The recent approach [3] proposes a supervised learning method based on random forests to predict TAD boundaries from histone modifications and chromatin proteins. In general, this approach is reported to perform quite accurately in predicting boundaries. However, it does not model interior TAD segments and it treats each segment independently ignoring the fact that TADs form as a result of the joint effects of multiple segments. Lastly, it also uses an error function based on gini index ignoring that the marker distributions may not be gaussian.

2. The *nTDP* Model

2.1 The likelihood function

Let V be the ordered set of genome restriction fragments (bins), where each bin v represents the interval $[vr - r + 1\ , vr]$, where r is the Hi-C resolution. Let M be the set of histone modifications (markers) over V. The marker data $H = (h_{vm})$ is a $|V| \times |M|$-matrix where its (v, m)'th entry h_{vm} is the count of the occurrences of marker m inside segment v. Let $d = [s, e]$ be a domain (interval) where s and e are its start and end boundaries respectively, $\{s+1, \ldots, e-1\}$ are the segments inside d, and let $D = \{[s_1, e_1], [s_2, e_2], \ldots, [s_i, e_i]\}$ be a partition of V where none of the domains overlap.

We propose a supervised, semi-nonparametric, high-dimensional model $nTDP$ that uses H to model and predict D. Our model can be seen as a generalization of Conditional Random Field [22, 31] where we have continuous weights instead of binary features and where we model the marker effects semi-nonparametrically.

Specifically, we assume there are 3 types of segments in V that are relevant for modeling: those that are at the domain boundaries ($V_{\mathbf{b}}$), those that are in the interior of domains ($V_{\mathbf{i}}$), and those that are not part of a domain ($V_{\mathbf{e}}$), and we have $V = V_{\mathbf{b}} \cup V_{\mathbf{i}} \cup V_{\mathbf{e}}$. For each marker type m, we have 3 types of *effect functions*, $f^b_m(c, \mathbf{w^b_m}), f^i_m(c, \mathbf{w^i_m}), f^e_m(c, \mathbf{w^e_m})$, that will describe the relationship between marker count c and the fragment type (b, i, e) for marker type m. Here, $\mathbf{w^b_m}, \mathbf{w^i_m}, \mathbf{w^e_m}$ are parameters that we will fit to determine the shape of the effect function. Thus, for example, $f^i_m(c, \mathbf{w^i_m})$ will describe how a count of c for marker m influences whether the fragment is in the interior (i) of a domain.

We assume that these effect functions combine linearly. Therefore, let

$$E^b_{vq} = \sum_{m \in M} f^b_m(c^q_{vm}, \mathbf{w^b_m}) \tag{1}$$

be the total effect of all the markers on fragment v for boundary formation (b). Summations E^i_{vq} and E^e_{vq} are defined analogously for interior (i) and inter-domain fragments (e).

Let W be the union of model parameters $\mathbf{w^b_m}, \mathbf{w^i_m}, \mathbf{w^e_m}$, and let $D^{\text{train}} = \{D^q : q = 1, \ldots, Q\}$ be several domain decompositions (in different sequences or conditions) and let $H^{\text{train}} = \{H^q : q = 1, \ldots, Q\}$ be a set of corresponding histone markers. Under the assumption that the training pairs are independent, the log-likelihood of parameters W given D^{train} is

$$\log\Big(P(D^{\text{train}}|W, H^{\text{train}})\Big) = \sum_q \log\Big(P(D^q|W, H^q)\Big). \tag{2}$$

We define the probability $P(D^q, W, H^q) = \frac{\exp^{F(D^q,W,H^q)}}{\sum_{F'} \exp^{F'}}$ where $F(D^q, W, H^q)$ is the total quality of partition D^q and marker data H^q under model parameters W. Let V^q be the set of segments in pair q. Due to the independence of segments:

$$\log\Big(P(D^q|W, H^q)\Big) = \overbrace{\sum_{d=[s,e]\in D^q} \Big(\sum_{v\in\{s,e\}} \overline{c}_b E^b_{vq} + \sum_{v=s+1}^{e-1} \overline{c}_i E^i_{vq} \Big) + \sum_{v\in V^q_e} \overline{c}_e E^e_{vq}}^{\log\left(P(D^q,W,H^q)\right)=F(D^q,W,H^q)} - \log(Z^q_{|V^q|}) \tag{3}$$

where $Z^q_{|V^q|} = \sum_{D'} P(D', W, H^q)$ is the partition function defined over all possible nonoverlapping partitions D', $\overline{c}_b, \overline{c}_i, \overline{c}_e$ are relative weights of different types of fragments to account for unbalanced training set, and V^q_e is the set of fragments that do not belong to any domain in D^q.

2.2 Nonparametric form of the effect functions

Because the shape of the marker effect function is unknown, we choose the f functions from the nonparametric family of Bernstein basis polynomials. Bernstein polynomials can approximate any effect function and additionally can handle imposed shape constraints such as monotonicity and concavity.

Let A be the chosen dimension of these polynomials; larger A results in a more expressive family, but more parameters to fit. Let $m_{\max}$ be the maximum possible density of marker m. This is is used to transform the input c^q_{vm} to the range $[0, 1]$; therefore define $p^q_{vm} = c^q_{vm}/m_{\max}$. We model $f^b_m(c^q_{vm}, \mathbf{w^b_m})$ for segment v by a Bernstein polynomial $B_A(p^q_{vm}, \mathbf{w^b_m})$ as in:

$$f^b_m(c^q_{vm}, \mathbf{w^b_m}) = B_A\left(p^q_{vm}, \mathbf{w^b_m}\right) = \sum_{i=0}^{A} w^b_m[i] \overbrace{\binom{A}{i} \left(p^q_{vm}\right)^i \left(1 - p^q_{vm}\right)^{A-i}}^{b_{i,A}(p^q_{vm})} \tag{4}$$

where $b_{i,A}(p^q_{vm})$ are the base Bernstein kernels.

3. Optimal algorithms for training and inference

We must train the parameters W for the above model using data of the form $D^{\text{train}}, H^{\text{train}}$. We will examine these trained parameters (and several good solutions for them) for insights into which markers are most informative for describing D^{train} and thus topological domains.

Problem 1. **Training**: Given a set of marker data H^{train}, likely from several chromosomes and cell conditions, and corresponding set of TAD decompositions D^{train}, we estimate the most likely parameters W according to Eqn. 2.

Problem 2. **Inference**: Given marker data H model parameters W, we estimate the best domain partition D of the track.

3.1 Training

A nice feature of the objective (3) is that it is convex in its arguments, $\{\mathbf{w_m^b}, \mathbf{w_m^i}, \mathbf{w_m^e}\}_{m \in M}$, which follows from linearity, composition rules for convexity, and convexity of the negative logarithm. However, training involves several challenges: (a) computing the partition function $Z^q_{|V^q|}$ in (3), and (b) estimating W so that the weights are sparse. We solve each of these challenges next.

Estimating the partition function. We estimate $Z^q_{|V^q|}$ in (3) recursively in polynomial time since each segment can belong to one of 4 states: domain start (sb), inside a domain (i), domain end (eb), non-domain (e), and state of each segment depends only on the previous segment's state. Let $Y = \{sb, i, eb, e\}$, and $Z^q_{|V^q|} = Z^q_{|V^q|,eb} + Z^q_{|V^q|,e}$ which components can be estimated by:

$$Z^q_{v,x} = \sum_{y \in Y} Z^q_{v-1,y} T_{y,x} \exp^{E^x_{vq}} \tag{5}$$

where $Z^q_{v,sb}$, $Z^q_{v,i}$, $Z^q_{v,eb}$, $Z^q_{v,e}$ represent the partition function up to segment v ending with sb, i, eb and non-domain respectively. T is a 4×4 binary state transition matrix where $T_{y,x} = 1$ if a segment can be assigned to x given previous segment is assigned to state y such as $(y, x) \in \{(sb, i), (sb, eb), (i, i), (i, eb), (eb, sb), (eb, e), (e, sb), (e, e)\}$, otherwise 0. Initial conditions are $Z^q_{0,sb} = Z^q_{0,i} = Z^q_{0,eb} = 0$, $Z^q_{0,e} = 1$. To avoid overflow in estimating $Z^q_{|V^q|}$ and speed it up, we estimate $\log(Z^q_{|V^q|})$ by expressing it in terms of log of the sum of exponentials and forward and backward variables (α, β) similar to Hidden Markov Model [22].

Estimating a sparse set of good histone effect parameters. We would like to augment objective function (2) so that we select a sparse subset of markers from the data and avoid overfitting. If the coefficients $\mathbf{w_m^b} = 0$, then there is no influence of marker m. For this purpose, we will impose grouped lasso type of regularization on the coefficients w_{mk}. Grouped lasso regularization has the

tendency to select a small number of groups of non-zero coefficients but push other groups of coefficients to be zero.

We introduce two types of regularization. First, we require that many of the weights be 0 using an L_2-norm regularization term. Second, we want the effect functions $\{f\}$ to be smooth. Let $P = \{b, i, e\}$. We modify our objective to trade off between these goals:

$$\underset{W}{\operatorname{argmin}} \; -\sum_{q} \log\Big(P(D^q|W,H^q)\Big) + \overbrace{\lambda_1 \sum_{p\in P}\Big(\sum_{m\in M}\|\mathbf{w_m^p}\|\Big)^2 + \lambda_2 \sum_{p\in P}\sum_{m\in M} R(f_m^p)}^{\text{Regularization}} \tag{6}$$

where λ_1, λ_2 are the regularization parameters, and $R(f_m^p)$ is the smoothing function for effect of marker m at $p \in P$:

$$R(f_m^p) = \int_x \left(\frac{\partial^2 f_m^p(x, \mathbf{w_m^p})}{\partial x^2}\right)^2 dx \tag{7}$$

Group lasso in (6) uses the square of block l_1-norm instead of l_2-norm group lasso which does not change the regularization properties [1]. Second-order derivative in (7) can be expressed more explicitly as a convex quadratic function of $\mathbf{w_m^p}$. Its derivation can be found in Appendix.

We note that (6) is convex, but it is a nonsmooth optimization problem because of the regularizer. We solve it efficiently by using an iterative algorithm from multiple kernel learning [1]. By Cauchy-Schwarz inequality:

$$\sum_{p\in P}\Big(\sum_{m\in M}\|\mathbf{w_m^p}\|\Big)^2 \le \sum_{p\in P}\sum_{m\in M}\frac{\|\mathbf{w_m^p}\|^2}{\gamma_{mp}} \tag{8}$$

where $\gamma_{mp} \ge 0$, $\sum_{m\in M}\gamma_{mp} = 1,\ p \in P$, and the equality in (8) holds when

$$\gamma_{mp} = \frac{\|\mathbf{w_m^p}\|}{\sum_{m\in M}\|\mathbf{w_m^p}\|}, \quad p \in P \tag{9}$$

This modification turns the objective into the following which is jointly convex in both $\mathbf{w_m^p}$ and γ_{mp}:

$$\underset{W}{\operatorname{argmin}} \; -\sum_{q} \log\Big(P(D^q|W,H^q)\Big) + \sum_{p\in P}\sum_{m\in M}\Big(\lambda_1\frac{\|\mathbf{w_m^p}\|^2}{\gamma_{mp}} + \lambda_2 R(f_m^p)\Big) \tag{10}$$

$$\text{s.t.} \sum_{m\in M}\gamma_{mp} = 1.0, \quad p \in P \tag{11}$$

$$\gamma_{mp} \ge 0,\ m \in M, p \in P \tag{12}$$

We solve this by alternating between the optimization of $\mathbf{w_m^p}$ and γ_{mp}. When we fix γ_{mp}, we can find the optimal $\mathbf{w_m^p}$ by any quasi-newton solver such as L-BFGS [19] which runs faster than the other solvers such as iterative scaling or conjugate gradient. When we fixed $\mathbf{w_m^p}$, we can obtain the best γ_{mp} by the closed form equation (9). Both steps iterate until convergence.

3.2 Training extensions

We can model a variety of shape-restricted effect functions by Bernstein polynomials that cannot be easily achieved by other nonparametric approaches such as smoothing splines [20]. We add the following constraints to ensure monotonicity:

$$w_m^b[i] \leq w_m^b[i+1], \quad i = 0, \ldots, A-1 \tag{13}$$

which is a realistic assumption since increasing the marker density should not decrease its effect. We can also ensure concavity of the effect function by:

$$w_m^b[i-1] - 2w_m^b[i] + w_m^b[i+1] \leq 0, \quad i = 1, \ldots, A-1 \tag{14}$$

which has a natural diminishing returns property: the increase in the value of the effect function generated by an increase in the marker density is smaller when output is large than when it is small. Our problem is different than smoothing splines since our loss function is more complicated than traditional spline loss functions due to partition function estimation in (5) which makes it hard to directly apply the smoothing spline methods [29]. In addition, these nonnegativity and other shape constraints can be naturally enforced in our method.

3.3 Inferring domains using the trained model

Given marker data H over a single track and W, the inference log-likelihood is:

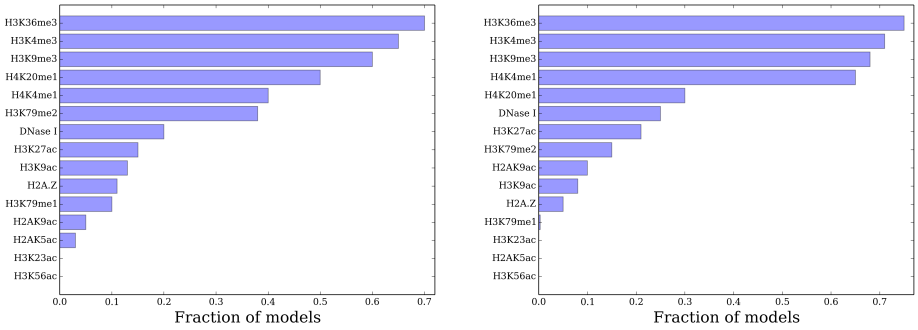
$$\underset{D}{\text{argmax}} \ \log\Big(P(D|W,H)\Big) = \sum_{d=[s,e]\in\overline{D}} r_{se}x_{se} \ + \ \sum_{v\in V} E_v^e y_v \tag{15}$$

where $\overline{D} = \{[s,e] \,|\, s,e \in V,\, e-s \geq 1\}$ is the set of all potential domains of length at least 2 and $r_{se} = E_s^b + E_e^b + \sum_{v=s+1}^{e-1} E_v^i$. The intuition is that variable $x_{se} = 1$ when the solution contains interval $[s,e]$, and variable $y_v = 1$ if v is not assigned to any domain. The $\log(Z_{|V|})$ term is removed during inference since it is same for all D. We solve (16)–(17) to find the best partition D:

$$\underset{D}{\text{argmax}} \sum_{d=[s,e]\in\overline{D}} r_{se}x_{se} \ + \ \sum_{v\in V} E_v^e\Big(1 - \sum_{[s,e]\in M[v]} x_{se}\Big) \tag{16}$$

$$\text{s.t.} \ \ x_{se} + x_{s'e'} \leq 1 \qquad \forall \text{ domains } [s,e],\, [s',e'] \text{ that overlap} \tag{17}$$

where $M[v]$ is the set of intervals that span fragment v. We replace y_v in (15) with $1 - \sum_{[s,e]\in M[v]} x_{se}$ since each segment can be assigned to at most a single domain. (17) ensures that inferred domains do not overlap. Problem (16)–(17) is *Maximum Weight Independent Set* in interval graph defined over domains which can be solved optimally by dynamic programming in $O(|V|^2)$ time.



(a) human IMR90 (b) human ES

Fig. 1: Fraction of histone modifications appearing in a best scoring four-modification model in (a) human IMR90, (b) human ES. Best scoring is defined as reaching at least 95% of NVI score of the model with all modifications.

4. Results

4.1 Experimental setup

We binned ChIP-Seq histone modification and DNase-seq data at 40 kb resolution, estimate RPKM (Reads Per Kilobase per Million) measure for each bin, and transform values x in each bin by $\log(x+1)$, which reduces the distorting effects of high values. In the case of 2 or more replicates, the RPKM-level for each bin is averaged to get a single histone modification file, in order to minimize batch-related differences. We convert any data mapped to hg19 (mm8) to hg18 (mm9) using UCSC liftOver tool. We define TADs over human IMR90, human embryonic stem (ES), and mouse ES cells Hi-C data [6] at 40 kb resolution after normalization by [30]. We use consensus domains from Armatus [10] as the true TAD partition by selecting threshold γ where maximum Armatus domain size is closest to the maximum Dixon et al. [6] domain size ($\gamma = 0.5$ for IMR90, $\gamma = 0.6$ for human ES, and $\gamma = 0.2$ for mouse ES cells).

We solved the training optimization problem by L-BFGS [19]. We use the public implementation of *Armatus* [10], and obtain histone modifications from NIH Roadmap Epigenomics [4] and UCSC Encode [8]. Code and datasets can be found at `http://www.cs.cmu.edu/~ckingsf/research/ntdp`. *nTDP* is reasonably fast: we train on all human IMR90 chromosomes in less than 3 hours on a MacBook Pro with 2.5Ghz processor and 8Gb Ram. The iterative procedure in general converges in fewer than 10 iterations.

We prevent overfitting by following a two-step nested cross-validation which has inner and outer steps. The outer K-fold cross-validation, for example, trains on all autosomal human chromosomes except the one to be predicted. Within each loop of outer cross-validation, we perform $(K-1)$-fold inner cross-validation to estimate the regularization parameters.

4.2 *nTDP* finds a small subset of modifications predictive of TADs

We identified a minimal set of histone marks that can model TADs as follows: we run *nTDP* independently on each chromosome of human IMR90 to obtain 21 sets of marks. These sets overlap significantly across all chromosome pairs (hypergeometric $p < 0.05$ for all pair-wise comparisons), and a total of 16 modifications cover all chromosomes. Despite the regularization, the weights of several of these marks are still very close to 0, so we identify a non-redundant subset of the modifications by Bayesian information criterion (BIC) [22] which penalizes model complexity more strongly.

As we increase the number of included modifications from 1 to 16, the BIC decrease nearly stops after 4 modifications, with some additional small reduction up to 6 modifications. The sets of 4 and 6 modifications that were most informative are: {H3K36me3, H3K4me1, H3K4me3, H3K9me3} and {H3K4me3, H3K79me2, H3K27ac, H3K9me3, H3K36me3, H4K20me1}. These non-redundant set of elements are preserved when we repeat this procedure between species. We find that only these $4-6$ modifications are needed to accurately predict TADs.

These marks are common in good models. The 4 modifications {H3K36me3, H3K4me1, H3K4me3, H3K9me3} are also enriched among a collection of high quality training solutions. We measure the agreement between estimated and true partitions by normalized variation of information $NVI = \frac{VI}{\log |V|}$ [21] where VI measures the similarity between two partitions and lower score means better performance. We analyze the fraction of models with 4 histone modifications for which NVI score is at least 95% of optimum NVI score obtained by running *nTDP* over all modifications as in Figure 1a–1b. We find 161, 139 solutions satisfying this criteria among 1820 candidates for human IMR90 and human ES histone modifications respectively. We find the 4 histone modifications above to be significantly overrepresented in the set of models for both human IM90 and ES cells (hypergeometric $p < 0.0001$). These significance values combined with the results above suggest the importance of the identified modifications in TADs.

These marks have nearly optimal coherence score. We assess the performance of various subset of modifications by the coherence score which is the exponential of the negative mean log-likelihood of each chromosome on the test set, and it is normalized by the best model coherence score as in Table 1. As such it is a relative measure of the quality of various models. The coherence score using only the set {H3K36me3, H3K4me1, H3K4me3, H3K9me3} is almost as high as the score for all 28 histone modifications in human IMR90. Restricting the effect function shape to be nonnegative and concave slightly improves the score. Our analysis indicates that the remaining modifications carry either redundant information or are less important for TADs.

4.3 Predicting TADs from histone marks in human

nTDP is able to predict domain boundaries accurately using 4 histone marks alone in both human IMR90 and human ES cells. We compare TAD prediction performance of *nTDP* with the chromatin state partition predicted by Seg-

Table 1: Normalized coherence scores of various marker subsets

Allowed modifications (human IMR90 to IMR90)	Coherence score (Normalized)
28 histone modifications + Concave + Nonnegative *	1.00
28 histone modifications + Concave	0.99
28 histone modifications	0.97
H3K4me3, H3K79me2, H3K27ac, H3K9me3, H3K36me3, H4K20me1	0.94
H3K36me3, H3K4me1, H3K4me3, H3K9me3 + Concave + Nonnegative	0.94
H3K36me3, H3K4me1, H3K4me3, H3K9me3 + Concave	0.93
H3K36me3, H3K4me1, H3K4me3, H3K9me3	0.92

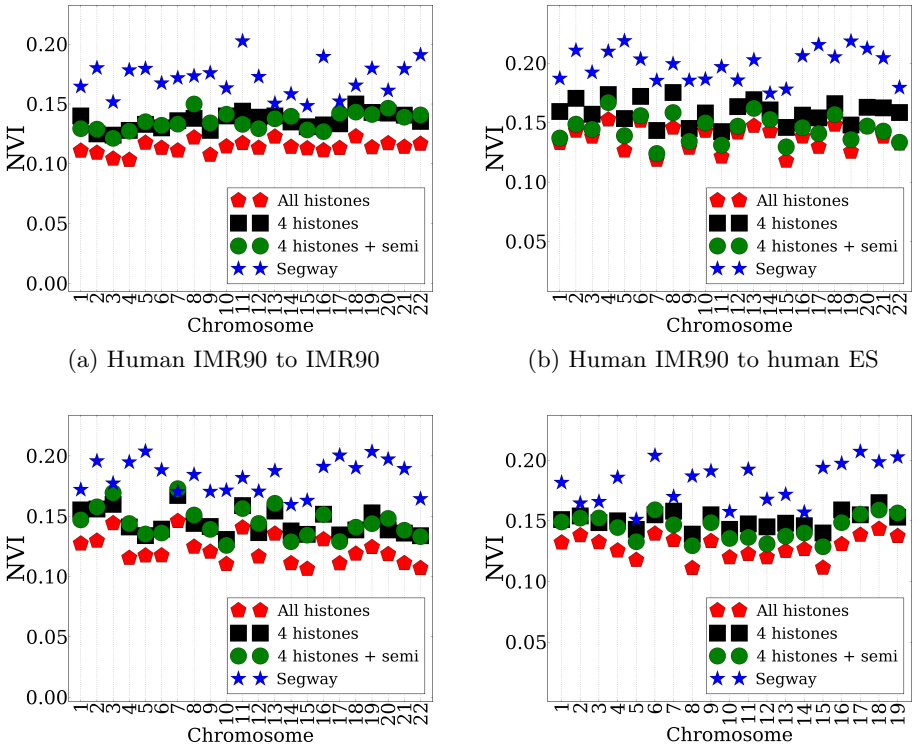
way [14] in terms of NVI. Even though Segway does not predict TADs directly, its chromatin state partition can still be used as a baseline. Training with all 28 histone modifications instead of with the identified 4 modifications does not lead to a major performance increase as shown in Figure 2a even though it increases the training time approximately 4 times for human IMR90 cells. Restricting the effect function to be monotonic and concave only slightly increases the performance. Chromatin states inferred by Segway do not directly correspond to TADs which leads to a lower TAD prediction performance even though they have other meaningful interpretations.

We find combinatorial effects of histone modifications to be important for accurate domain prediction since none of the modifications can achieve NVI score better than 0.2 when considered independently. To verify that there are not inherent structures in the data that can lead to an easy prediction, we randomly shuffle domains in the training set by preserving their lengths without shuffling modifications, which NVI score is never better than 0.3 in all chromosomes showing the importance of histone modification distributions in TADs.

nTDP also predicts TADs accurately across different species as well as across different cell types as in Figures (2b)–(2d). For example, if we train on human IMR90 data, the model we obtain is still able to recover domains in human ES cells (Figure 2b). This holds true across species as well: training on human ES data, for example, produces a model that can work well on mouse ES data.

4.4 Multiscale analysis of the predicted TADs

We find that our predicted TADs match true TADs more accurately at different scales defined by different *Armatus* γ's as in Figure 3a. We observe a slight performance improvement if we define true TAD partition at lower *Armatus* γ values in human IMR90 which correspond to longer TADs. This figure suggests that some of our wrong TAD predictions may actually correspond to longer TAD blocks which we erroneously interpret as incorrect due to a scale mismatch.



(a) Human IMR90 to IMR90

(b) Human IMR90 to human ES

(c) Human ES to human ES

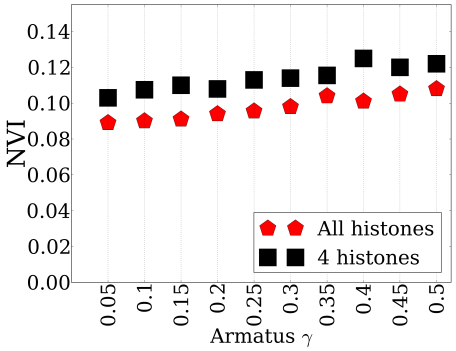
(d) Human ES to mouse ES

Fig. 2: TAD prediction performance on different chromomes a) human IMR90 to human IMR90: infer each human IMR90 chromosome by training with all IMR90 chromosomes except the one to be inferred, b) human IMR90 to human ES, c) human ES to human ES, d) human ES to mouse ES are defined similarly.

5. Conclusion

We formulate semi-nonparametric modeling of TADs in terms of histone modifications, and propose an efficient provably optimal solution *nTDP* for training and inference. Experimental results on human and mouse cells show that a common subset of histone modifications can accurately predict TADs across cell types and species. Via our trained model, we also accurately predict TADs without using any Hi-C data which is especially useful for understanding the 3D genome conformation on species with limited Hi-C data.

Funding. This research is funded in part by the Gordon and Betty Moore Foundations Data-Driven Discovery Initiative through Grant GBMF4554 to Carl Kingsford, by the US NSF (1256087, 1319998), and by the US NIH (HG006913, HG007104). C.K. received support as an Alfred P. Sloan Research Fellow.



(a) Performance at different scales

Fig. 3: Multiscale analysis of the predicted TADs a) Performance over true TAD partitions at different scales obtained via different *Armatus* γ in human IMR90.

References

1. Bach, F.R.: Exploring large feature spaces with hierarchical multiple kernel learning. In: Advances in Neural Information Processing Systems. pp. 105–112 (2009)
2. Baù, D., Marti-Renom, M.A.: Structure determination of genomic domains by satisfaction of spatial restraints. Chromosome Research 19(1), 25–35 (2011)
3. Bednarz, P., Wilczyński, B.: Supervised learning method for predicting chromatin boundary associated insulator elements. Journal of Bioinformatics and Computational Biology 12(06), 1442006 (2014)
4. Bernstein, B.E., et al.: The NIH roadmap epigenomics mapping consortium. Nature Biotechnology 28(10), 1045–1048 (2010)
5. Bickmore, W.A., van Steensel, B.: Genome architecture: Domain organization of interphase chromosomes. Cell 152(6), 1270 – 1284 (2013)
6. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398), 376–380 (2012)
7. Duggal, G., Wang, H., Kingsford, C.: Higher-order chromatin domains link eQTLs with the expression of far-away genes. Nucleic Acids Research 42(1), 87–96 (2014)
8. ENCODE Project Consortium, et al.: An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414), 57–74 (2012)
9. Ernst, J., Kellis, M.: ChromHMM: automating chromatin-state discovery and characterization. Nature Methods 9(3), 215–216 (2012)
10. Filippova, D., Patro, R., Duggal, G., Kingsford, C.: Identification of alternative topological domains in chromatin. Alg Mol Biol 9(1), 14 (2014)
11. Gibcus, J.H., Dekker, J.: The hierarchy of the 3D genome. Molecular Cell 49(5), 773–782 (2013)
12. Guelen, L., et al.: Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453(7197), 948–951 (2008)
13. Ho, J.W., et al.: Comparative analysis of metazoan chromatin organization. Nature 512(7515), 449–452 (2014)

14. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., Noble, W.S.: Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods 9, 473–476 (2012)
15. Hou, C., Li, L., Qin, Z.S., Corces, V.G.: Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains. Molecular Cell 48(3), 471 – 484 (2012)
16. Le, T.B.K., et al.: High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342(6159), 731–734 (2013)
17. Libbrecht, M.W., Ay, F., Hoffman, M.M., Gilbert, D.M., Bilmes, J.A., Noble, W.S.: Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell type-specific expression. Genome Research 25, 544–557 (2015)
18. Lieberman-Aiden, E., et al.: Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326(5950), 289–293 (2009)
19. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming 45(1-3), 503–528 (1989)
20. McKay Curtis, S., Ghosh, S.K.: A variable selection approach to monotonic regression with Bernstein polynomials. J Applied Statistics 38(5), 961–976 (2011)
21. Meilă, M.: Comparing clusterings—an information based distance. Journal of Multivariate Analysis 98(5), 873–895 (2007)
22. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT Press (2012)
23. Nora, E.P., et al.: Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? BioEssays 35(9), 818–828 (2013)
24. Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., Bland, M.J., Wagstaff, W., Dalton, S., McDevitt, T.C., Sen, R., Dekker, J., Taylor, J., Corces, V.G.: Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 153(6), 1281 – 1295 (2013)
25. Rao, S.S., et al.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7), 1665–1680 (2014)
26. Sefer, E., Duggal, G., Kingsford, C.: Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations. In: RECOMB, vol. 9029, pp. 293–308. Springer (2015)
27. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., Cavalli, G.: Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148(3), 458–472 (2012)
28. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., de Laat, W.: Looping and interaction between hypersensitive sites in the active β-globin locus. Molecular Cell 10(6), 1453–1465 (2002)
29. Wahba, G.: Spline models for observational data, vol. 59. SIAM (1990)
30. Yaffe, E., Tanay, A.: Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics 43(11), 1059–1065 (2011)
31. Zhou, J., Troyanskaya, O.G.: Global quantitative modeling of chromatin factor interactions. PLoS Comput Biol 10(3), e1003525 (2014)

Appendix

$R(f_m^p)$ can be written more explicitly as in (18) according to [20]:

$$\frac{\partial^2 f_m^p(x, \mathbf{w_m^p})}{\partial x^2} = A(A-1)\sum_{i=0}^{A-2}(w_m^p[i+2]-2w_m^p[i+1]+w_m^p[i])\binom{A-2}{i}x^i(1-x)^{A-2-i} \tag{18}$$

which turns $R(f_m^p)$ into (19):

$$\int_0^1 \left(\frac{\partial^2 f_m^p(x)}{\partial x^2}\right)^2 dx = A^2(A-1)^2\sum_{i=0}^{A}\sum_{j=i}^{A}(w_m^p[i]w_m^p[j])\left(\sum_{q=\overline{e}_i}^{\min(i,2)}\sum_{r=\overline{e}_j}^{\min(j,2)}(-1)^{q+r}\binom{2}{q}\binom{2}{r}T_{j-r}^{i-q}(x)\right) \tag{19}$$

where $\overline{e}_p = \max(0, 2-A+p)$, $T_{j-r}^{i-q}(x)$ is defined below and $\beta(i+j-q-r+1, 2A-3-i-j+q+r)$ is the beta function:

$$T_{j-r}^{i-q}(x) = \binom{A-2}{i-q}\binom{A-2}{j-r}\underbrace{\int_0^1 x^{i-q}(1-x)^{A-2-i+q}x^{j-r}(1-x)^{A-2-j+r}dx}_{\beta(i+j-q-r+1, 2A-3-i-j+q+r)} \tag{20}$$

$R(f_m^p)$ is convex which follows from semidefiniteness of the resulting polynomial.