Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

epigenetics, 3D conformation

Integrative chromatin domain annotation through graph embedding of Hi-C data

Neda Shokraneh [1,*], Mariam Arab [1], and Maxwell Libbrecht [1,*]

[1] Computing Science Department, Simon Fraser University, Burnaby, V5A 1S6, Canada.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The organization of the genome into domains plays a central role in gene expression and other cellular activities. Researchers identify genomic domains mainly through two views: 1D functional assays such as ChIP-seq, and chromatin conformation assays such as Hi-C. Fully understanding domains requires integrative modeling that combines these two views. However, the predominant form of integrative modeling uses segmentation and genome annotation (SAGA) along with the rigid assumption that loci in contact are more likely to share the same domain type, which is not necessarily true for epigenomic domain types and genome-wide chromatin interactions.

Results: Here, we present an integrative approach that annotates domains using both 1D functional genomic signals and Hi-C measurements of genome-wide 3D interactions without the use of a pairwise prior. We do so by using a graph embedding to learn structural features corresponding to each genomic region, then inputting learned structural features along with functional genomic signals to a SAGA algorithm. We show that our domain types recapitulate well-known subcompartments with an additional granularity that distinguishes a combination of the spatial and functional states of the genomic regions. In particular, we identified a division of the previously-identified A2 subcompartment such that the divided domain types have significantly varying expression levels.

Availability: https://github.com/nedashokraneh/genome-structure-function-aggregation

Contact: maxwl@sfu.ca

Supplementary information:

1 Introduction

Chromosomes are segmented into different domain types that are identified through analysis of genomic data types [1]. Genome domain organization plays a central role in many cellular activities including gene expression and DNA replication. The disruption of this regulation leads to developmental defects and diseases like cancer. Therefore, identifying of genomic domain types provides insight into domain-scale mechanisms behind cell differentiation and disease development.

Heitz [2] first classified genomic regions into heterochromatin or euchromatin based on differential compaction at interphase from microscopy observations. Heterochromatin is a condensed region that is not accessible by regulatory elements, while euchromatin is less condensed, more accessible, and highly transcribed [3].

Later, the development of functional assays enabled genome-wide mapping of different proteins and histone modifications [1]. Here, we use the term "functional assay" to mean any experiment that aims to measure biochemical function or activity (such as ChIP-seq, DNase-seq, Repli-seq, and others), whether or not these marks necessarily confer biological function. These functional assays are usually represented as a 1D vector over the genome. Clustering genomic regions according to functional genomic signals at a coarse resolution ($10-100$ kb) characterizes domain types such as active and repressed functional domains [4, 5, 6].

Recently, chromatin conformation capture assays such as Hi-C provided a way to measure the genome-wide contact map between genomic regions [7]. Analysis of Hi-C led to the identification of multiple chromatin domain structures such as compartments, TADs and loops [7, 8, 9]. Compartments are chromatin domain types such that genomic regions within the same compartment have similar interaction patterns with the rest of the genome and are spatially segregated within a nucleus.

© The Author 2015.

Compartments are strongly correlated with local functional genomic data including replication timing, transcriptional activity, epigenetic patterns, and higher-order nuclear structures [10, 11].

The close association between different chromosomal measurements necessitates integrative modeling that incorporates multiple data types in order to understand genome organizational patterns. Segmentation and genome annotation (SAGA) algorithms have emerged as the predominant tool for integrative modeling of multiple 1D genomic signals [12]. A SAGA algorithm such as Segway or ChromHMM assigns a label to each genomic region such that regions with the same label have similar patterns of input signals. Most SAGA algorithms employ a chain-structured probabilistic models such as a hidden Markov model (HMM) structure [13, 14, 15].

Although any SAGA algorithm based on a multivarite hidden Markov model can jointly model different types of 1D genomic signals, this model cannot jointly model 1D functional assays and pairwise Hi-C interaction data together. To remedy this problem, two previous methods were developed to perform such joint modeling. First, Segway-GBR [16] jointly models functional genomic signals and Hi-C intra-chromosomal significant interactions through augmenting a chain-structured probabilistic model present in Segway [14] with graph-based posterior regularization. This regularization uses a graph representing a Hi-C data set to encourage spatial neighbors to receive a similar label. Second, SPIN [17] jointly models spatial genomic signals and Hi-C genome-wide significant interactions using a Hidden Markov random field (HMRF) in order to identify genome-wide spatial compartmentalization patterns. In particular, they construct the Markov random field by connecting both linear and spatial neighbors, as defined by statistically-significant Hi-C interactions. They learn emission and transition probabilities based on observed spatial signals and their spatial neighbors. This approach encourages positions in contact in 3D to have a similar pattern of domain labels to positions that are neighbors in 1D.

Both SPIN and Segway-GBR make the core assumption that positions connected in 3D are more likely to share the same label. This assumption is motivated by the observation that domain labels are correlated in 3D, but it may not hold in all cases. In fact, we show here that both methods fail to jointly capture the functional and higher-order organization of the genome. Thus, we need a domain modeling strategy that avoids this pairwise prior assumption.

The recent development of Hi-C embedding methods provides a way to summarize pairwise chromatin conformation into a small number of 1D genomic tracks [18, 19, 20]. These embedding representations have been shown to be effective for many tasks including identifying subcompartments, improving Hi-C resolution and simulating genomic perturbations.

Here, we propose to leverage Hi-C embeddings to perform integrative identification of chromatin domains. Following SCI [19] we use the graph embedding model LINE [21] to embed Hi-C data into 8 1D structural features that capture genome-wide compartmentalization pattern of genomic bins. We then infer combinatorial domain types by inputting functional and structural genomic signals into hidden Markov model-based SAGA algorithm. We show that, because functional and structural compartmentalization are different genome organizational layers, this approach outperforms existing integrative approaches such as Segway-GBR and SPIN. We use this approach to identify 8 combinatorial domain types and show that they recapitulate chromatin epigenomic and compartmentalization patterns simultaneously and grant additional granularity to previously-identified subcompartments.

2 Materials and Methods

2.1 Dataset

We used 12 functional assays, including ChIP-seq targeting H2A.Z and 10 histone modifications (H3K4me1, H3K4me2, H3K4me3, H3k9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H4K20me1), and DNase-seq, as 1D functional genomic signals. We used Hi-C data to learn genomic structural features, and add spatial constraints to existing integrative models (graph construction from Hi-C data is explained in the supplementary section 2). We also used gene expression, replication timing, and ChIA-PET data to evaluate domain annotations. All data are corresponding to the GM12878 cell line. Details about data sources and processing are provided in supplementary section 1.

2.2 Learning genome-wide global structural features

We use a node embedding model, LINE [21], similar to the subcompartment caller SCI [19] to learn genome-wide global structural features representing the spatial organization of the genome. We describe the general formalization of node embedding models followed by details of the used model, LINE with second-order proximity objective, in this section.

In general, a node embedding model maps each node in the input network graph into a low-dimensional vector space (also known as an embedding space), such that pairwise distances in the embedding space are representative of the proximity of pairs of nodes in a graph. They have four main components, (1) encoder: to map the graph structure into a vector space, (2) proximity definition: to define a proximity measure between pairs of nodes based on the input graph structure, (3) decoder: to map the embeddings of a pair of nodes to a reconstructed proximity measure, and (4) loss function: to penalize the difference between the reconstructed and true proximity measures.

The definition of the proximity measure is the most important component to choose based on the type of graph structure we aim to embed. Given a weighted graph $G = (V, E)$, where V is a set of nodes $(v_1, ..., v_n)$, and E is a set of edges attributed with a w_{ij} representing the weight of the edge between nodes v_i and v_j, the first-order proximity of a pair (i, j) is equal to the weight w_{ij} of the edge connecting that pair. A greater edge weight between nodes indicates more first-order proximity between nodes.

A greater similarity between nodes' interaction patterns with the rest of the nodes in a graph indicates a higher second-order proximity between the two nodes. More formally, a first-order proximity of a node v_i with all other nodes is defined as $p_1(v_i) = (w_{i1}, ..., w_{in})$, and the second-order proximity of a pair (i, j) is defined as the similarity between $p_1(v_i)$ and $p_1(v_j)$.

LINE can be optimized based on either first-order or second-order proximity loss functions. Since we aim to capture the global chromatin structural features, we will use LINE embeddings based on the second-order proximity. We also show that second-order embeddings capture more meaningful global chromatin structural patterns compared to first-order embeddings in the supplementary information. Here, we describe the LINE model based on the second-order proximity in more detail.

To learn the embeddings that capture the second-order proximity between nodes, two vectors corresponding with each node v_i are defined, the embedding of the node, u_i, and the context of the node, u'_i, that comes from the embedding of its neighbors. Then, $p_2(.|v_i)$ is defined for each node v_i as a distribution over all nodes that determine the probability that the context of each node has been generated using embedding of node v_i.

$$p_2(v_j|v_i) = \frac{\exp(\overrightarrow{u}_j'^T \overrightarrow{u}_i)}{\sum_{k=1}^{|V|} \exp(\overrightarrow{u}_k'^T \overrightarrow{u}_i)}.$$

To preserve the second-order proximity, $p_2(.|v_i)$ should be close to the empirical distribution coming from graph structure, defined as $\hat{p}_2(v_j|v_i) = \frac{w_{ij}}{d_i}, d_i = \sum_j w_{ij}$. The objective function aims to minimize the weighted sum of distances between $p_2(.|v_i)$ and $\hat{p}_2(.|v_i)$ over all nodes, defined according to the Kullback-Leibler divergence between $p_2(.|v_i)$ and $\hat{p}_2(.|v_i)$. A weight λ_i is given to each node i equal to its degree. The full objective function is defined as

$$O_2 = \sum_{i \in V} \lambda_i d(\hat{p}_2(.|v_i), p_2(.|v_i)),$$

which is simplified to:

$$O_2 = -\sum_{(i,j) \in E} w_{ij} \log p_2(v_j|v_i).$$

We use LINE's [21] original implementation to learn the global chromatin structural features from O/E chromatin interaction graph (supplementary information, section 3). The optimization strategies used in their implementation makes it applicable to large graphs like genome-wide chromatin interaction graph. Details about choosing the node embedding model and hyperparameters selection are provided in the supplementary information, section 4.

2.3 Domain annotation

Unsupervised segmentation and genome annotation (SAGA) algorithms cluster the genomic positions based on their observed functional or chemical genomic signals [12] and assign a label to each cluster based on biological interpretation of genomic signals pattern. These algorithms have most commonly been applied at a fine resolution to define $100-1000$ bp segments, resulting in annotations of small-scale phenomena such as protein-coding genes and regulatory elements. However, these algorithms can also be applied on genomic signals defined on larger segments ($10^4 - 10^6$ bp) to find domain-scale patterns and labels such as Polycomb-repressed domains, heterochromatic domains and different types of active domains [6, 1, 16, 22].

SAGA algorithms use structured probabilistic models that model the dependency between latent labels and observed genomic signals, and the dependency between neighbor latent labels on a genome sequence. Two models, Segway-GBR [16] and SPIN [17] extend existing SAGA algorithms to incorporate the dependency between spatial neighbors as well. We characterize such algorithms based on their structures, and provide a guideline for choosing a proper one for different purposes.

2.3.1 GMM

A Gaussian Mixture model is an unsupervised soft clustering method that assumes data are drawn independently from a mixture of Gaussian distributions. More formally, given a data $X = \{X_1 ... X_N\}$ of size N and assuming that data is comprised of K clusters, a latent variable $Z_i \in \{1...K\}$ is defined for each data point $X_i \in X$, where $Z_i = k$ if data point i is from cluster k. Each cluster $k \in \{1...K\}$ is characterized by μ_k and Σ_k, the mean and covariance of its corresponding Gaussian distribution, and π_k, which is a mixing probability of a cluster in data. Then, the probability of that X_i is from a cluster k is

$$p(X_i, Z_i = k) = p(X_i|Z_i = k)p(Z_i = k) = \mathcal{N}(X_i|\mu_k, \Sigma_k) * \pi_k,$$

where $p(X_i|Z_i = k)$ is an emission probability of data point i from a cluster k, and $p(Z_i = k)$ is a constant representing the mixing probability of a cluster k in data. The joint probability of all observed variables X and latent variables Z is defined as

$$p(X, Z) = \prod_{i=1}^{N} P(X_i|Z_i)P(Z_i).$$

Given a model learned parameters μ_k, Σ_k, and π_k for each cluster k, the posterior distribution over K clusters for each data point is calculated during inference to infer the most probable cluster for each data point.

2.3.2 GMM-HMM

Hidden Markov model adds the dependency between neighbor latent variables to the naive probabilistic model such as GMM. So, the probability of that X_i is from a cluster k is defined as $p(X_i, Z_i = k) = p(X_i|Z_i = k)p(Z_i = k|Z_{i-1})$, and the joint probability of all observed variables X and latent variables Z would be:

$$p(X, Z) = \pi_{Z_1} p(X_1|Z_1) \prod_{i=2}^{N} P(X_i|Z_i)P(Z_i|Z_{i-1}),$$

where $P(Z_i = z_i|Z_{i-1} = z_{i-1})$ is captured in a transition probability distribution $A(z_i|z_{i-1})$.

2.3.3 Segway

The Segway model [14] is specifically designed for modeling genomic signals with additional types of latent variables to handle genomic data behaviours. The main feature of a Segway is modeling the distribution of segments length by putting a prior on segments length distribution. This is done by determining a prior information about the range of segments lengths that is modeled by the transition model, and assigning weight scale to the transition model, such that higher weights give more power to the transition model compared to the emission model during the learning and inference.

2.3.4 SPIN (HMRF)

A Markov random field adds higher-order dependencies between latent variables to the naive probabilistic model such as GMM. The dependency between latent variables is captured in a Markov random field represented as a graph $G = (V, E)$, where $V = \{V_1 ... V_N\}$ are corresponding to latent variables $Z = \{Z_1 ... Z_N\}$ and $(i, j) \in E$ if there is a dependency between Z_i and Z_j (SPIN uses the neighboring genomic bins in the linear genome and 3D space to construct the random field). $p(X_i|Z_i)$ is like the emission probability distribution in GMM, and $p(Z_i, Z_j)$ is a transition probability distribution to model the dependency between all neighboring genomic bins. Then, the joint probability of all observed variables X and latent variables Z is defined as

$$p(X, Z) = \prod_{i=1}^{N} P(X_i|Z_i) \prod_{(i,j) \in E} p(Z_i, Z_j).$$

2.3.5 Segway-GBR

HMM and HMRF incorporate the dependency between neighbor latent variables' posterior probability distributions in the definition of the joint probability distribution $P(X, Z)$. An HMM cannot model higher-order dependencies, and HMRF inference is done by approximate algorithms which are not guaranteed to converge, specifically when there is not a high agreement between observed variable and the MRF structure. Therefore, graph-based posterior regularization (GBR) [23] was proposed to incorporate the higher-order dependency information by augmenting the probabilistic model like Segway given a regularization graph such that nearby variables in a regularization graph have similar posterior distributions. This is done by adding a regularizer term to the base

probabilistic model loss. The regularizer can be added either during training to learn both model parameters and posterior distributions with respect to the regularizer graph, or during the inference to infer the posterior distributions given the original model parameters with respect to the regularizer graph. We train a regular Segway, and use GBR during the inference similar to the Segway-GBR paper [16]. While the learning and inference of probabilistic models is based on maximizing the objective which is equal to the likelihood of observed data $\mathcal{L}(\theta) = log\, p_\theta(X)$, GBR adds a regularizer term to the objective to encourage the similar posterior distributions for neighboring latent variables:

$$\text{maximize}_{\theta,q}\ \mathcal{J}(\theta, q) = \mathcal{L}(\theta) - D(q(Z) \parallel p_\theta(Z|X)) - \lambda_G \sum_{(u,v)\in E_R} w(u,v) D(q_u^M(Z_u) \parallel q_v^M(Z_v)),$$

where $q(Z)$ is an auxiliary joint distribution over latent variables Z, and $q_u^M(Z_u)$ is a marginal distribution over a latent variable Z_u, and D is the Kullback-Leibler divergence. The second term encourages the auxiliary joint distribution and inferred joint distribution over latent variables given observed variables to be similar, and the third term encourages the marginal distribution of neighbor latent variables in a regularization graph to be similar. E_R is a set of edges of the regularization graph, and $w(u, v)$ is a weight of edge between latent variables Z_u and Z_v, and λ_G controls the strength of regularization.

2.3.6 Combinatorial HMM

By applying a node embedding model to genome-wide Hi-C data, we map each genomic bin into a vector space. A vector space with dimension 8 is enough to capture the structural information in whole Hi-C data (supplementary section 3). We input the learned 8 structural features with 12 functional genomic signals to existing SAGA algorithms, GMM, HMM, or Segway. Our final results are outputs of an HMM which is a proper choice for the annotation on domain-scale.

2.4 Evaluation methods

2.4.1 Gene expression

The most important function of domain structures is a transcription regulation, and we expect that domain states have a different transcriptional patterns. Therefore, we annotate each gene with a most frequent domain label among the gene region, and calculate a *percentage of variance explained* (VE) based on domain labels.

The percentage of variance explained is a fraction of signal variance that is explainable using an annotation prediction. Given a genome annotation $a_{1:n} \in \{1...K\}^n$ (K is the number of labels) and a signal vector $s_{1:n} \in \mathbb{R}^n$, the mean value of the signal over positions with label k is calculated as

$$\mu_k = \frac{\sum_{i=1}^{n} \mathbb{1}(a_i = k) s_i}{\sum_{i=1}^{n} \mathbb{1}(a_i = k)}.$$

So, the predicted signal vector is $s_i^p = \mu_{a_i}$. VE is computed as the difference of total variance $\text{var}(s_i)$ and the variance of residuals of the prediction $\text{var}(s_i - s_i^p)$:

$$\text{VE} = \frac{\text{var}(s_i) - \text{var}(s_i - s_i^p)}{\text{var}(s_i)} = 1 - \frac{\sum_{i=1}^{n} (s_i - s_i^p)^2}{\sum_{i=1}^{n} (s_i - \overline{s})^2}.$$

VE is bounded by the range $[0, 1]$, and higher values indicate more agreement between annotation and signal values.

To reduce the influence of large outlier gene expression values, we normalize gene expression values using the inverse hyperbolic sine function.

2.4.2 Replication timing

We used 6-phase replication timing (RT) data to evaluate the domain annotations. The other important functional implication of domains is the RT pattern. The RT data is in 1000 bp resolution, such that each 1000 bp-length region is replicated with a proportion p_i in a phase i. We calculated a consistency between 6-phases RT and labels in two ways: first, we calculated VE for each phase separately (the value of RT at each phase is a 1D signal) and calculated a mean of VE among 6 phases (RT score1). Second, we assigned each 1000 bp-length region with a RT label, which is a phase that a region was mostly replicated in. We calculated the similarity between domain labels and RT labels using Adjusted Rand Index (ARI) score [24] (RT score2).

2.4.3 Agreement with ChIA-PET loops

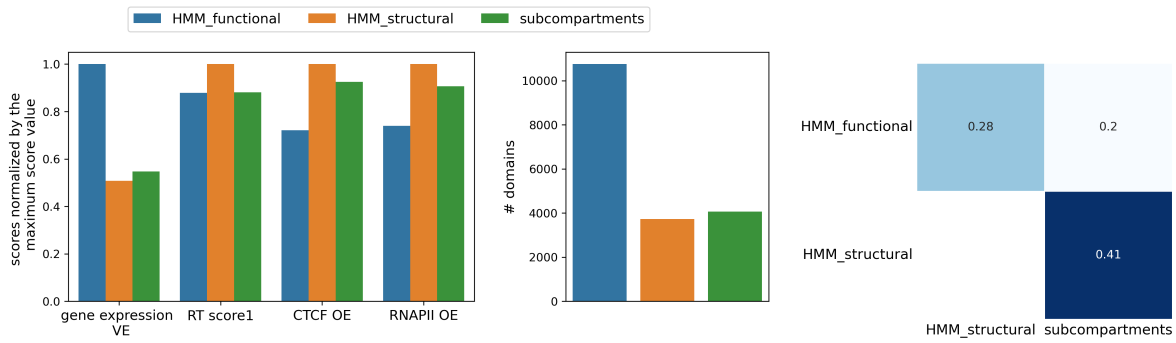
We also assessed the domains with respect to higher-order chromatin structures, including loops from CTCF and RNAPII ChIA-PET data. If domain structures capture the higher-order structure of the chromatin, we expect ChIA-PET loops to usually have the same label on their ends. Therefore, we grouped loops based on the labels assigned to their ends, resulting in a matrix, $O_{L\times L}$, where L is a number of labels. Given a normalized coverage of labels $C = (C_1, ..., C_L)$, the expected number of loops between labels i and j is $E_{ij} = \text{total}_O * C_i * C_j$, where $\text{total}_O = \sum_{i,j} O_{ij}$. The expected and observed number of loops with a same label on their ends are $\sum_i E_{ii}$ and $\sum_i O_{ii}$, respectively. The ratio of observed over expected number of loops with a same label shows the enrichment of loops occurring between similar domain states. The higher ratio indicates the higher consistency between the domain annotation and the higher-order structure input. We call observed over expected enrichment of CTCF and RNAPII ChIA-PET loops as CTCF OE and RNAPII OE scores respectively.

3 Results

3.1 Functional and structural features confer complementary information

To ascertain the need for combinatorial domain annotations, we evaluated whether features of chromatin types and compartments confer complementary or equivalent information. To do so, we created domain annotations using each type respectively, and we found that they confer complementary information. In particular, we found that structural domains are broader than functional domains and regulate chromatin functions acting in domain-scale.

To identify domains based on functional features, we applied three baseline annotation models (GMM, HMM, and Segway) on functional genomic signals, and identify 6 domain types. The average length of Segway domains is longer than HMM domains and the number of HMM domains is also less than GMM domains due to the HMM and Segway duration models (Fig S. 3a). We compared these three models in terms of explaining different cellular activities and chromatin features described in the evaluation methods (Fig S. 3a) and found that HMM domains are the best in $4/5$ scores, therefore, we chose HMM domains (HMM_functional) as a baseline for functional domain types. We also chose 6 domain types based on the proportion of variance explained (VE) for input signals (Fig S. 3b), as VE does not improve significantly for any of the input signals for the number of domain types greater than six. This is also consistent with known chromatin types in a domain-scale [1, 6, 16].



(a) (b)

Fig.1: (a) Comparison of HMM_functional, HMM_structural, and subcompartments [8] domain types based on the proportion of variance explained for gene expression (gene expression VE), the average proportion of variance explained for replication timing 6 phases signals (RT score1), agreement with CTCF and RNAPII ChIA-PET loops (CTCF OE, RNAPII OE), and the number of domains along the whole genome. (b) Similarity (adjusted rand index (ARI)) between each pair of HMM_functional, HMM_structural, and subcompartments [8] domain types.

To identify domains based on structural features, we learned genomic structural signals using the LINE embedding model for all genomic bins in a shared space (Section 2.2). We applied an HMM on the learned structural signals (section 2.3.6) and identified 6 domain types (HMM_structural).

We compared HMM_functional, HMM_structural and well-known subcompartments from Rao et al [8] in terms of their sizes and other chromatin features including gene expression and replication timing profiles and loops from ChIA-PET data. We observed that structural domains are broader than functional domains, and there is less number of structural domains (Fig. 1a). The chromatin features that are regulated in domain-scale such as replication timing and ChIA-PET loops are better explained by structural domain types compared to functional domain types, while a more local feature, gene expression, is better explained by functional domain types (Fig. 1a).

To evaluate which types of cellular activities are captured by functional and structural domains respectively, we compared each domain type to multiple data sources: the average transcription of genes in each domain type (Fig S. 4), and the overlap of functional and structural domain types with Rao et al subcompartments (Fig S. 5). The structural domain types are highly similar to subcompartments, which confirms that the learned structural signals capture a genome-wide compartmentalization pattern. B2 and B3 subcompartments are only mixed and divided in another way into two structural domain types. This is also observable from structural embeddings UMAPs that B2 and B3 subcompartments are not distinguishable by learned structural embeddings (Fig S. 2a). On the other hand, functional domain types are less similar to subcompartments and structural domain types. Two active functional domain types overlap with both active subcompartments (Fig S. 5), and the other four functional domain types overlap with A2 and inactive subcompartments in an ordered continuous manner; for example, the labels 3 and 2 are mainly a mix of A2 and B1, and the label 0 is a mix of B1, B2, and B3 subcompartments. Furthermore, the two active subcompartments and two active structural domain types are not distinguished by their gene expression pattern, while two active functional domain types have different gene expression patterns. This is also validated by lower proportion of variance explained for gene expression by structural domain types compared to functional domain types (Fig. 1a).

We also calculated the similarity between subcompartments, functional and structural domain types using Adjusted Rand Index (ARI) score [24]. We found that functional domain types are less similar to subcompartments than structural domain types (Fig. 1b).

All these observations show that chromatin types and the structural compartmentalization of chromatin are different chromatin organization layers and complementary to each other, and thus an integrative approach is needed to model both patterns together.

3.2 Combinatorial domain types better capture the chromatin higher-order organization compared to existing integrative models

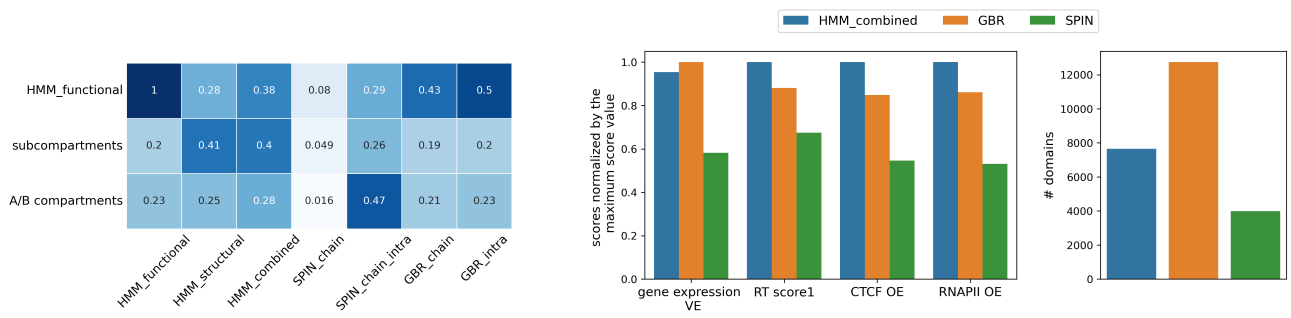
To infer domain types that jointly capture chromatin types and conformation, we input functional and learned structural genomic signals to the HMM (section 2.3.6); we term the result combinatorial domains (HMM_combined).

We compared combinatorial domains to existing methods that utilize both measurements of chromatin type and conformation, Segway-GBR and SPIN. The graph-based posterior regularization (GBR) can be added to any type of probabilistic model, and we use the GMM-GBR version to be consistent with two alternatives, SPIN, and HMM_combined, that utilize GMM as their basic component as well.

First, we evaluated the agreement of each of the integrative annotations, HMM_combined, SPIN, and GBR with chain and Hi-C edges, with well-known annotations, HMM_functional, subcompartments [8], and A/B compartments [1] (Fig. 2a). We found that the SPIN model cannot effectively handle functional data as input; doing so results in the model assigning most of the genome to one or two labels (Supplementary Section 4, Figs S. 6b, Fig. 2a). GBR domains have a high agreement with original functional domain types, are just a little broader as a result of regularization, and do not have a higher agreement with subcompartments and compartments compared to the original HMM_functional domain types (Fig. 2a). HMM_combined domains have a high agreement with both original functional domain types, and subcompartments, showing that it captures both chromatin states and conformation simultaneously.

Second, we compared three aggregative models in terms of explaining different cellular activities and chromatin features (Fig. 2). We can see that HMM_combined outperforms both SPIN and GBR in all metrics except for the proportion of variance explained for the gene expression. This is due to the high similarity of the GBR annotation to the original functional

[1] A-type and B-type subcompartments are merged into the A and B compartments, respectively.



(a) (b)

Fig.2: (a) Agreement of annotation from different aggregative models with well known annotations, HMM_functional, subcompartments, and A/B compartments, calculated using adjusted rand index (ARI). (b) Comparison of HMM_combined, GBR, and SPIN domain types based on the proportion of variance explained for gene expression (gene expression VE), the average proportion of variance explained for replication timing 6 phases signals (RT score1), agreement with CTCF and RNAPII ChIA-PET loops (CTCF OE, RNAPII OE), and the number of domains along the whole genome.

domain types. Thus, GBR mostly captures local chromatin types that are sufficient to explain the local cellular activities, such as gene expression.

Although SPIN and GBR can take both 1D signals and interactions data as an input, they cannot model chromatin local types and higher-order conformation together. This limitation is likely due to the strong pairwise prior assumption made by both models, which assumes that positions close in 3D are more likely to be of the same domain type. Therefore, it is necessary to embed the higher-order organization of the chromatin into a vector space at first, to preserve the information of both chromatin local types and higher-order conformation in inferred domain types.

3.3 Combinatorial domains capture different chromatin organization layers

We found that the domain types learned by the combinatorial HMM recapitulate subcompartments and identify with additional granularity among these compartment types. As a result of the aggregation of functional and structural genomic signals, it is expected to find more than 6 domain types. Therefore, we evaluated combinatorial domain annotations with 6-10 types accordign to gene expression variance explained and RT score1 (Fig. S 7a). We choose 8 domain types because gene expression VE score does not change by increasing the number of domain types, RT score1's change is small, and we were not able to interpret further domain types. To further our understanding of the inferred combinatorial domain types, we compared them to functional genomic signals, overlap with subcompartments, average gene expression, number of genes per 1 Mb region, and their coverage along the genome (Fig. 3).

These eight combinatorial domain types can be classified into 4 active and 4 inactive types. These types largely recapitulate known subcompartments. Label 6 corresponds to the A1 subcompartment (associated with broad transcription): it is highly overlapped with A1 and is enriched with transcribed genes.

Combinatorial domains reveal a distinction between two types of A2 subcompartments (associated with cell type-specific transcription), represented by labels 0 and 5. Label 0 is mostly overlapped with A2 and has a lower gene density compared to the label 6, however its genes are highly transcribed. Label 5 is also mostly overlapped with A2, however, it is less enriched with active marks, and its genes are moderately expressed compared to the label 0. It shows that the A2 subcompartment can be further divided into two domain types with different epigenomic and gene expression patterns.

Label 3 is the last active type, which has a lower coverage and is less transcribed compared to other active types. These domains are mixed between the A1 and B1 subcompartments and feature moderate activating signals and gene expression. This domain type confirms the existence of genomic regions interacting with the active compartment but without high transcription.

The other four combinatorial labels correspond to inactive domain types. Label 7 corresponds to the B1 subcompartment (associated with facultative repression): it is highly enriched with H3K27me3, representative of facultative Heterochromatin [1, 16], and mostly overlaps the B1 subcompartment. It includes a moderate number of genes; however, they are highly inactivated.

Label 4 corresponds to the B2 subcompartment (associated with nucleolar domains): it is enriched with the H3K9me3 mark, representative of constitutive heterochromatin, is highly overlapped with the B2 subcompartment, and depleted from genes.

Combinatorial domains reveal a distinction between two types B3 subcompartments (associated with constitutive repression). Label 2 is enriched with both inactive marks, H3K27me3 and H3K9me3, and is overlapped with the B3 subcompartment. Label 1 is depleted of all epigenomic marks and is overlapped with the B3 subcompartment as well.

We also visualized the same plot for HMM_functional 8 domain types (Fig S. 7b), and we can see that such discrimination of overlap with subcompartments and gene expression patterns is not observable from functional domain types. These observations show that, although there is a link between 1D functional domain types and 3D compartment domains [1, 25], they can supplement each other through integrative analysis.

4 Discussion and Conclusion

The link between 1D and 3D organization of the chromatin have been studied in different scales through integrative modeling [17, 16, 26], and predictive models [27, 25, 28]. In this study, we focus on integrative modeling of functional genomic signals and chromatin conformation from Hi-C data to identify chromatin domain types that capture both chromatin epigenomic and conformation patterns.

Previous integrative models [17, 16] relied on the assumption that spatially close genomic regions are more probable to be assigned the same label. These methods are sensitive to the procedure we use to determine spatially close genomic regions, as they will push non-colocalized regions to get the same label if we incorporate random contacts. On the other hand, methods that infer genome subcompartments are based on clustering genomic regions' interaction patterns, which is equivalent to the second-order proximity of the nodes (genomic regions) in a Hi-C

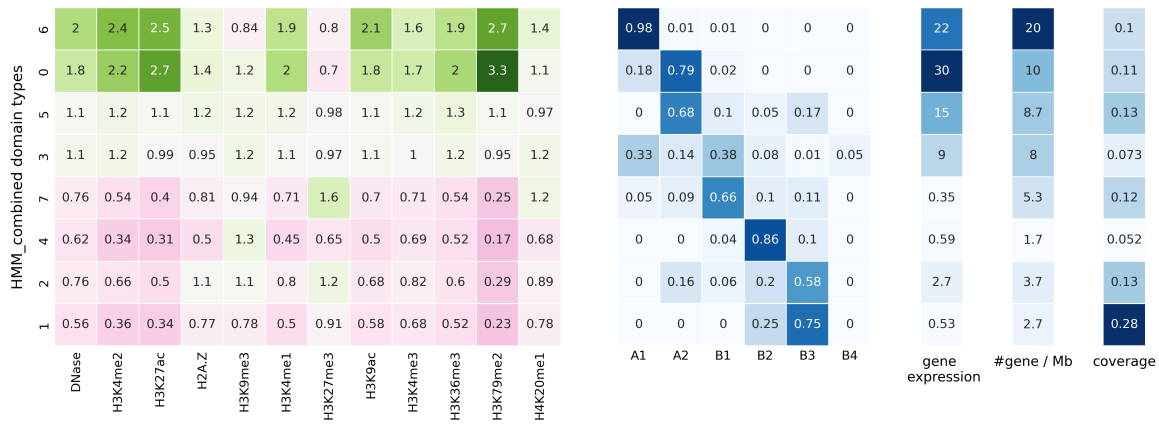


Fig.3: Summary of different genomic features for HMM_combined domain types, including enrichment of functional genomic signals (ratio of the average of the signal over each domain type to the average of the signal over the whole genome), overlap with subcompartments (percentage of overlap with different subcompartments for each domain type), an average of gene expression (average of expression (RPKM) of genes in each domain type), the density of genes (average number of genes within each 1 Mb region of each domain type), and their coverage (genome coverage of each domain type).

graph. Therefore, we propose a framework that has two steps, learning structural features from Hi-C data that preserve second-order proximity of genomic regions in a Hi-C graph, and using functional and learned structural genomic signals to infer chromatin combinatorial domain types. We identify 8 combinatorial domain types that recapitulate well-known subcompartments [8] with additional granularity and have distinct patterns of different cellular activities such as gene expression and replication timing profile.

One problem with joint modeling of 1D and 3D chromatin domain organization is a gap between their spatial resolutions. For example, functional genomic signals such as ChIP-seq are available in finer resolution ans their domain organization can be studied at a fine resolutions (5-100 kb), while chromatin 3D domain types, or subcompartments, have been studied in lower resolutions ($\geq$ 100 kb). Such gap can be reduced by higher resolution Hi-C data or more powerful embedding models to learn embeddings in higher resolution. For example, we use both intra and inter-chromosomal Hi-C data to learn genome-wide embeddings in a shared space. The inter-chromosomal Hi-C data are very noisy in resolutions higher than 100 kb, so the learned structural features are limited to this resolution. One future work is to learn genome-wide structural features in higher resolutions through more powerful embedding approaches.

Funding

This work was supported by NSERC RGPIN/06150-2018 and a Michael Smith Foundation Scholar Award.

References

[1]Wendy A Bickmore and Bas van Steensel. Genome architecture: domain organization of interphase chromosomes. *Cell*, 152(6):1270–1284, 2013.

[2]Emil Heitz. *Das heterochromatin der moose*. Bornträger, 1928.

[3]Shiv IS Grewal and Songtao Jia. Heterochromatin revisited. *Nature Reviews Genetics*, 8(1):35–46, 2007.

[4]Robert E Thurman, Nathan Day, William S Noble, and John A Stamatoyannopoulos. Identification of higher-order functional domains in the human encode regions. *Genome research*, 17(6):917–927, 2007.

[5]Jessica L Larson and Guo-Cheng Yuan. Epigenetic domains found in mouse embryonic stem cells via a hidden markov model. *BMC bioinformatics*, 11(1):1–12, 2010.

[6]Guillaume J Filion, Joke G van Bemmel, Ulrich Braunschweig, Wendy Talhout, Jop Kind, Lucas D Ward, Wim Brugman, Inês J de Castro, Ron M Kerkhoven, Harmen J Bussemaker, et al. Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, 143(2):212–224, 2010.

[7]Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[8]Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[9]Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[10]Feng Liu, Chao Ren, Hao Li, Pingkun Zhou, Xiaochen Bo, and Wenjie Shu. De novo identification of replication-timing domains in the human genome by deep learning. *Bioinformatics*, 32(5):641–649, 2016.

[11]Johannes Nuebler, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A Mirny. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115(29):E6697–E6706, 2018.

[12]Maxwell W Libbrecht, Rachel CW Chan, and Michael M Hoffman. Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. *PLOS Computational Biology*, 17(10):e1009423, 2021.

[13]Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012.

[14]Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery

in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473, 2012.

[15]Nathan Day, Andrew Hemmaplardh, Robert E Thurman, John A Stamatoyannopoulos, and William S Noble. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426, 2007.

[16]Maxwell W Libbrecht, Ferhat Ay, Michael M Hoffman, David M Gilbert, Jeffrey A Bilmes, and William Stafford Noble. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome research*, 25(4):544–557, 2015.

[17]Yuchuan Wang, Yang Zhang, Ruochi Zhang, Tom van Schaik, Liguo Zhang, Takayo Sasaki, Daniel Peric-Hupkes, Yu Chen, David M Gilbert, Bas van Steensel, et al. Spin reveals genome-wide landscape of nuclear compartmentalization. *bioRxiv*, 2020.

[18]Kyle Xiong and Jian Ma. Revealing hi-c subcompartments by imputing inter-chromosomal chromatin interactions. *Nature communications*, 10, 2019.

[19]Haitham Ashoor, Xiaowen Chen, Wojciech Rosikiewicz, Jiahui Wang, Albert Cheng, Ping Wang, Yijun Ruan, and Sheng Li. Graph embedding and unsupervised learning predict genomic sub-compartments from hic chromatin interaction data. *Nature communications*, 11(1):1–11, 2020.

[20]Kevin Dsouza, Alexandra Maslova, Ediem Al-Jibury, Matthias Merkenschlager, Vijay Bhargava, and Maxwell Libbrecht. Hi-c-lstm: Learning representations of chromatin contacts using a recurrent neural network identifies genomic drivers of 3d genome conformation. 2021.

[21]Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015.

[22]Eugenio Marco, Wouter Meuleman, Jialiang Huang, Kimberly Glass, Luca Pinello, Jianrong Wang, Manolis Kellis, and Guo-Cheng Yuan. Multi-scale chromatin state annotation using a hierarchical hidden markov model. *Nature communications*, 8(1):1–9, 2017.

[23]Maxwell Libbrecht, Michael Hoffman, Jeff Bilmes, and William Noble. Entropic graph-based posterior regularization. In *International Conference on Machine Learning*, pages 1992–2001. PMLR, 2015.

[24]William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[25]Michele Di Pierro, Ryan R Cheng, Erez Lieberman Aiden, Peter G Wolynes, and José N Onuchic. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, 114(46):12126–12131, 2017.

[26]Anastasiya Belyaeva, Saradha Venkatachalapathy, Mallika Nagarajan, GV Shivashankar, and Caroline Uhler. Network analysis identifies chromosome intermingling regions as regulatory hotspots for transcription. *Proceedings of the National Academy of Sciences*, 114(52):13714–13719, 2017.

[27]Jean-Philippe Fortin and Kasper D Hansen. Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16(1):180, 2015.

[28]Jialiang Huang, Eugenio Marco, Luca Pinello, and Guo-Cheng Yuan. Predicting chromatin organization using histone marks. *Genome biology*, 16(1):1–11, 2015.