Graph embedding for inferring genome domains from genome 3D structure

Neda Shokraneh
Department of Computer Science
Simon Fraser University

Maxwell W Libbrecht
Department of Computer Science
Simon Fraser University

1 Introduction

Chromatin in the nucleus is segregated into nuclear compartments (5). These compartments have different functional properties and these properties drive many cellular processes, including gene regulation and DNA replication. A popular way to understand genome regulation and compartmentalization is to annotate the genome into several categories of *domains*, such that each domain type has similar activity. For example, Rao et al (5) categorize the genome into subcompartments based on Hi-C data, such that domains having the same pattern of interaction with the rest of the genome are annotated with same subcompartment.

Because the structure of a genome is correlated with epigenetic properties of chromosomes, both structural assays such as Chromatin Conformation Capture (3C) or Hi-C (21; 5; 20) and functional genomic assays (19) such as ChIP-seq and ATAC-seq provide complementary information to find such compartments of a genome. Therefore, integrative analysis methods are needed that integrate both types of information to produce a comprehensive understanding of chromatin regulation.

Segmentation and genome annotation (SAGA) methods are widely-used for integrative analysis of multiple data types. SAGA methods take multiple signal tracks as an input. They segment the genome and output a label for each segment of a genome such that segments with the same label have a similar pattern of genomics signal tracks (3; 4). While most SAGA methods can only handle 1D data sets such as ChIP-seq, two SAGA methods, Segway-GBR and SPIN (2; 6) can incorporate both genomics assays and Hi-C data to infer more accurate domain annotations. Segway-GBR does that through encouraging bins having high interaction in Hi-C data, to get the same label using graph-based regularization. SPIN (6) defines a Markov random field that includes edges defined on Hi-C contacts. Both of these methods have an assumption that bins that have more contact with each other, should get the same functional label.

In this paper, we aim to use recent advances in Graph Neural Networks to incorporate Hi-C data without the specific assumptions made by previous methods. Graph Neural Networks (GNNs) are able to embed graph structures into a low-dimensional feature space. We propose an integrative approach for identifying chromatin domains that leverages GNN methods. We learn latent features for each genome locus through their interaction graph. We concatenate these structural features with data from 1D ChIP-seq and DNase-seq data sets as input to a SAGA algorithm. The resulting domain annotations combine both biochemical activities driven from functional genomics assays and the structure of the genome defined by Hi-C data. These domains form a comprehensive picture of domains in the genome that can be used to discover new categories of domain activity and elucidate their influence on cellular activity such as gene regulation. In particular, we show that our inferred domain states show a more accurate explanation for different functional and structural properties of a genome, including replication timing and distance to nuclear compartments.

2 Method

2.1 Data processing

2.1.1 Hi-C data

We downloaded GM12878 cell line processed Hi-C data with GEO entry 63525 [1]. Then, observed over expected (O/E) contact frequency files for every pair of chromosomes were extracted using a Juicer tool (13). SCI (1) uses interchromosomal interactions to construct a Hi-C graph, but we found out that ignoring intrachromosomal edges results in a larger similarity of embeddings within a chromosome, which makes it unfeasible to cluster same domain

[1] `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525`

Preprint. Under review.

types from different chromosomes together. So, we use both intra and inter chromosomal interactions, however, O/E values are being computed for each pair of chromosomes separately, and they have different visibility across different pairs, which also result in dependency of embeddings on chromosome number. Therefore, we do the normalization on this genome-wide matrix based on mean of its rows, so each entry O_{ij} in matrix is being normalized by $O_i * O_j$ ($O_i = mean_j(O_{ij})$). Finally, we do hyperbolic arcsine transformation on matrix O. We observe that this preprocessing allows us to use both intra and inter chromosomal contact frequency information simultaneously to learn embedding for all bins in the same space.

2.1.2 ChIP-seq data

We downloaded GM12878 cell line (E116) ChIP-seq data sets targeting DNase, H2A.Z and 10 histone modifications (H3K4me1, H3K4me2, H3K4me3, H3k9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H4K20me1) from the Roadmap Epigenomics data portal[2]. To calculate a signal value for each bin, we get a weighted average of signal value over that bin.

2.2 Models

In general, our framework has two steps. In the first step, we infer structural features for genome loci from their 3D structure. In the second step, we input such structural features in addition to functional features to the genome annotation methods. In this section, we first explain a LINE method that we use for graph embedding, and then we shortly describe genome annotation methods.

2.2.1 Graph Embedding Methods

Graph embedding methods convert graph into a low dimensional space in which the graph information is preserved (22). We use graph embedding methods to learn latent features for every node of a Hi-C graph to map structural information in the Hi-C graph to structural features for each segment. Ashoor et al (1) predicts subcompartments by applying LINE (10) method on interchromosomal Hi-C graph, followed by k-means clustering on learned embeddings. We also use the LINE method on whole Hi-C graph to learn structural features in this study.

LINE : Large-scale Information Network Embedding (LINE) (10) has been used to map very large networks into low-dimensional vector space. They learn embeddings for nodes in a graph such that pairwise distances in embedding space would be representative of the proximity of pair of nodes in a graph. We chose LINE because it is able to address the embedding problem for large and weighted graphs. LINE can be optimized based on the first-order or second-order proximity loss function.

First-order proximity is the local pairwise proximity between two nodes. The greater weight between nodes indicates more proximity between nodes. By optimizing the first-order proximity objective, we encourage nodes with a large weight between them to get similar embeddings.

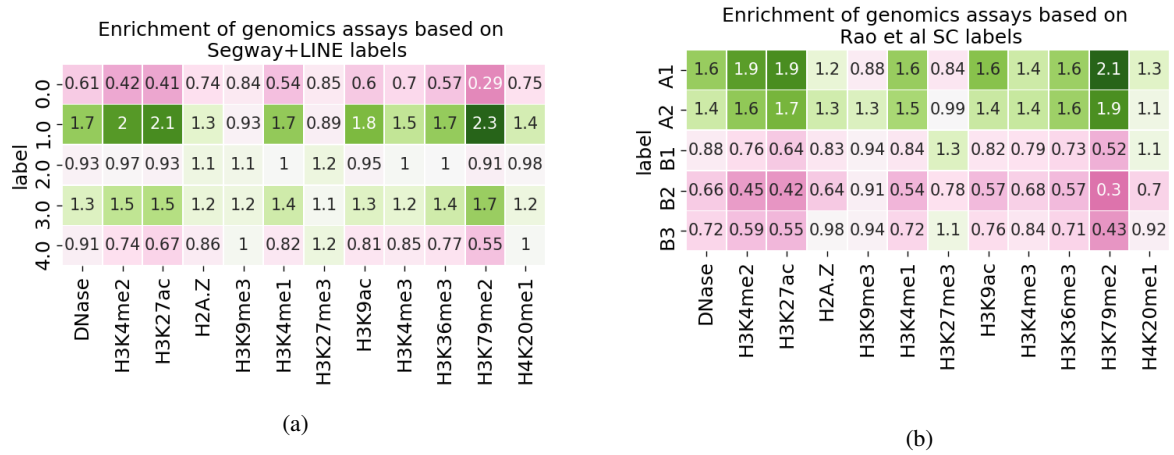
On the other hand, second-order proximity is based on the similarity between how two nodes are connected to all other nodes. For example, if we show first-order proximity of node u with all other nodes with $p_u = (w_{u,1}, ..., w_{u,|V|})$, second-order proximity of nodes u and v is based on similarity between p_u and p_v. To model second-order proximity, they define two vectors corresponding with each node i, one is its embedding (u_i), and the other one is its context (u_i') that comes from the embedding of its neighbors. Then, they define $p_2(.|v_i)$ distribution over all nodes, that determine the probability that the context of each node has been generated using embedding of node 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 second-order proximity, $p_2(.|v_i)$ should be close to empirical distribution coming from graph, which they define as $\hat{p}_2(v_j|v_i) = \frac{w_{ij}}{d_i}, d_i = \sum_j w_{ij}$. They minimize weighted sum of distance between $p_2(.|v_i)$ and $\hat{p}_2(.|v_i)$ over all nodes, that weights show importance of node in a graph, which is equal to degree of a node here. They use KL divergence for measuring closeness of distributions again, and second-order proximity after eliminating constants would be:

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

[2] `https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/`



(a) (b)

Figure 1: Plots of functional genomics assays enrichment based on Segway+LINE annotation (left) and subcompartments (right) (Enrichments have been calculated as ratio of signal mean over bins with a specific label to signal mean over whole genome)

Previous studies on compartmentalization used bins interaction patterns with the rest of the genome to infer their labels. This is equivalent to second-order proximity in the LINE method. We tried both 1st and 2nd orders embeddings, and 2nd order embeddings result in meaningful compartmentalization as expected. So we just use embeddings from second-order proximity in this study.

2.2.2 Segmentation and genome annotation (SAGA)

Segmentation and genome annotation (SAGA) methods aim to assign a label to each locus in a genome. Each of these labels could be divided into a few more detailed labels that have different patterns of input signals. Genome annotation methods use multiple observed signals including different histone modifications, DNase, etc to infer hidden state for loci over a genome. We use Segway in this study that uses the Dynamic Bayesian Network to model this problem. The model is described in (4) completely.

3 Results

3.1 Domain types based on 3D features improve stratification of functional and structural properties of a genome

Earlier studies have shown the correlation between the genome compartmentalization patterns and transcriptional activities (6; 5). We use learned features from applying LINE on the Hi-C graph as structural features and pass it to Segway. Annotation result from Segway stratifies functional genomics assays(figure 1a), although we do not use any functional genomic data as an input to Segway. Segway+LINE labels also represent two active (1: Active1, 3: Active2) and three inactive (0: Inactive1, 2: Inactive2, 4: Inactive3) domain types. We can see that Segway+LINE domains represent a better contrast for enrichment within active or inactive domains. For example, Active1 and Active2 domains show more enrichment difference for most of the assays comparing to A1 and A2. This could be the result of genome-wide annotation of Segway+LINE using both intra and inter chromosomal interaction frequencies data comparing to subcompartments which are based on part of interchromosomal interactions.

We also expect our domain types to reflect genome structural properties such as spatial positions relative to nuclear compartments. Chen et al (18) introduces TSA-seq assay that measures chromosome distances to defined nuclear structures. Using LaminAC TSA-seq data for K562 cell type, we show that our domain types have a different distribution of distance to LaminAC (figure 2 c), and explain 42% of signal variance, comparing to 41% using known subcompartments from Rao et al. We conclude that our learned features are a good and sufficient representation for whole Hi-C matrix.

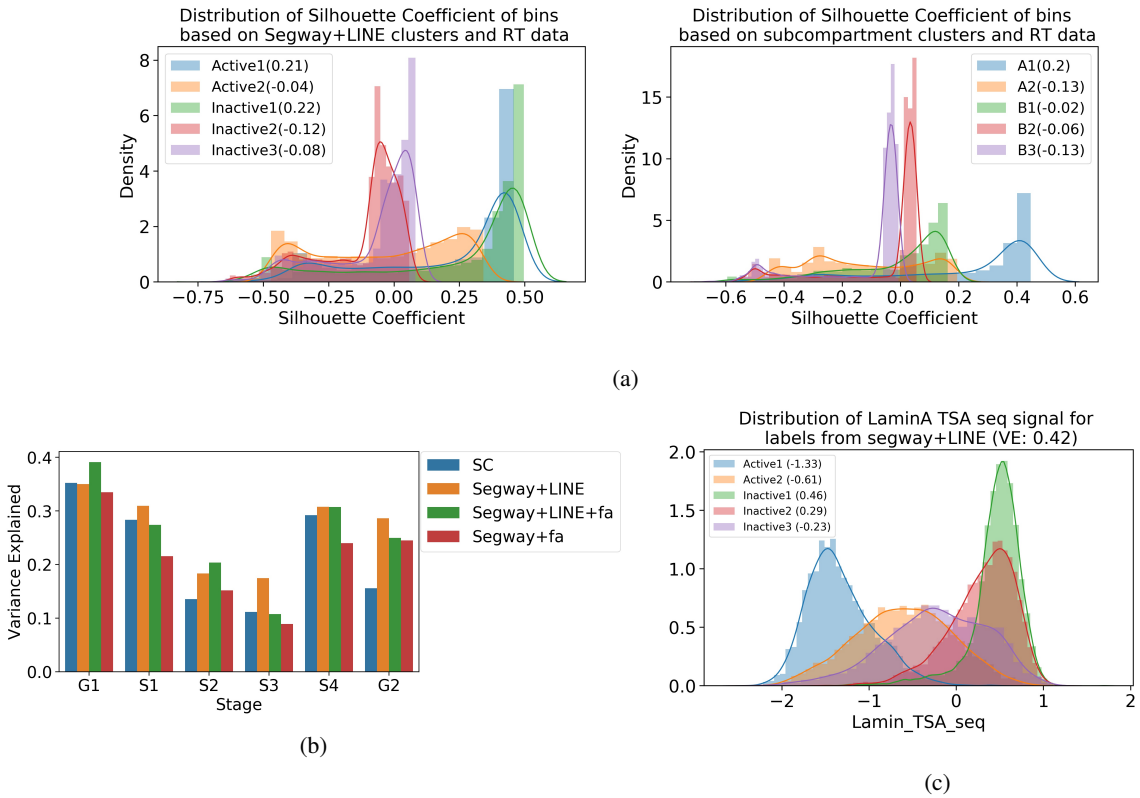


Figure 2: (a) Distribution of Silhouette Index grouped by labels from Segway+LINE annotation (left) and subcompartment labels from Rao et al (right), (b) Variance explained for different phases of RT based on domain labels from different inputs to Segway, (c) Distribution of LaminAC TSA seq signal for labels from Segway+LINE method

3.2 Domain types from aggregation of functional assays and 3D features improve representation of replication timing profile

The replication timing (RT) program of a genome is highly dependent on its 3D structure. Previous studies showed that the active compartment is replicated earlier than the inactive one; however, the difference between the RT program of subcompartments has not been observed. We show that our domain types using the Segway+LINE method stratify the genome according to its replication time better than existing methods. We use Silhouette Index (SI) metric to assess the goodness of compartmentalization with respect to the RT profile. We use 6 phase RT data and calculate SI indices based on Rao et al subcompartment and Segway+LINE clusters (figure 2 a). SI values range from -1 to 1, and a value greater than 0 indicates that the sample is fitting good to its assigned cluster. Subcompartment clusters show good SI for the A1 subcompartment, as it is replicated significantly earlier than other subcompartments; however other subcompartments do not preserve specific RT patterns. We observe that Segway+LINE labels show better RT profile discrimination not only for the Active1 domain but also for Inactive1.

We wanted to assess the effect of each of the structural and functional features on defining domain types that respect the RT program. Therefore, we run Segway using different sets of inputs including just structural features learned from LINE on Hi-C graph (Segway+LINE), just functional features described in section 2.1.2 (Segway+fa) and their combination (Segway+LINE+fa). We calculated variance explained for six RT profile phases signals based on domain types inferred from each of these input types. As expected, structural features can explain RT profile variance better than functional properties like histone modifications. However, we can see that aggregation of learned structural features with functional properties results in annotations with improved variance explanation for the RT signals. We also observe that compartmentalization based on Segway+LINE which just uses Hi-C data as an input like subcompartments from Rao et al, result in better explanation for RT signals variance (figure 2b).

References

[1] Ashoor, Haitham, et al. "Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data." Nature communications 11.1 (2020): 1-11.

[2] Libbrecht, Maxwell W., et al. "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 (2015): 544-557.

[3] Ernst, Jason, and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization." Nature methods 9.3 (2012): 215-216.

[4] Hoffman, Michael M., et al. "Unsupervised pattern discovery in human chromatin structure through genomic segmentation." Nature methods 9.5 (2012): 473.

[5] Rao, Suhas SP, et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell 159.7 (2014): 1665-1680.

[6] Wang, Yuchuan, et al. "SPIN reveals genome-wide landscape of nuclear compartmentalization." bioRxiv (2020).

[7] Wu, Zonghan, et al. "A comprehensive survey on graph neural networks." IEEE Transactions on Neural Networks and Learning Systems (2020).

[8] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.

[9] https://data.4dnucleome.org/browse/

[10] Tang, Jian, et al. "Line: Large-scale information network embedding." Proceedings of the 24th international conference on world wide web. 2015.

[11] Carty, Mark, et al. "An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data." Nature communications 8.1 (2017): 1-10.

[12] https://www.nature.com/articles/nature23884

[13] Durand, Neva C., et al. "Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments." Cell systems 3.1 (2016): 95-98.

[14] Xiong, Kyle, and Jian Ma. "Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions." Nature communications 10 (2019).

[15] Al Bkhetan, Ziad, and Dariusz Plewczynski. "Three-dimensional epigenome statistical model: genome-wide chromatin looping prediction." Scientific reports 8.1 (2018): 1-11.

[16] Zhu, Yun, et al. "Constructing 3D interaction maps from 1D epigenomes." Nature communications 7.1 (2016): 1-11.

[17] Qi, Yifeng, and Bin Zhang. "Predicting three-dimensional genome organization with chromatin states." PLoS computational biology 15.6 (2019): e1007024.

[18] Chen, Yu, et al. "Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler." Journal of Cell Biology 217.11 (2018): 4025-4048.

[19] Fortin, Jean-Philippe, and Kasper D. Hansen. "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data." Genome biology 16.1 (2015): 180.

[20] Yaffe, Eitan, and Amos Tanay. "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture." Nature genetics 43.11 (2011): 1059.

[21] Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." science 326.5950 (2009): 289-293.

[22] Cai, Hongyun, Vincent W. Zheng, and Kevin Chen-Chuan Chang. "A comprehensive survey of graph embedding: Problems, techniques, and applications." IEEE Transactions on Knowledge and Data Engineering 30.9 (2018): 1616-1637.