3D Chromosome Reconstruction Using Hi-C Data and Graph Autoencoder Networks

Van Hovenga, Daniel Wandeler
University of Colorado-Colorado Springs
1420 Austin Bluffs Pkwy
Colorado Springs, CO 80919
vhovenga@uccs.edu, dwandele@uccs.edu

Abstract

Chromosome confirmation capture (3C) is a method of measuring chromosome topology in terms of loci interaction. The Hi-C method is a derivative of 3C that allows for genome wide quantification of chromosome interaction. From this interaction data, it is possible to infer the 3D structure of the underlying chromosome. In this paper, we use a node embedding algorithm and a graph neural network to predict the 3D coordinates of each genomic loci from the corresponding Hi-C contact data in order to recreate a three dimensional representations of chromosomes in an unsupervised manner. We validate our results using Hi-C and ChIA-PET data from the GM12878 cell line.

Introduction

The nucleus of each eukaryotic cell stores important information about an organism's genetic makeup in the form of chromosomes (Sati and Cavalli 2017). Each chromosome is made of DNA, and the set of all chromosomes is known as the genome. There are well known connections between the makeup of an organism's genome and the traits it possesses, e.g. hair color, skin color, propensity for diseases, etc. (Sati and Cavalli 2017). One particularly effective attribute of a cell's genome is the shape and spatial occupation of its respective chromosomes. The organization of chromosomes facilitates inter-gene communication and regulation, thus contributing to the stability of a genome (Cremer and Cremer 2001). Thus, research into the capturing of topological makeup of chromosomes is valuable.

One such method of capturing chromosome topology is known as Chromosome Confirmation Capture. The 3C method was first developed by Dekker et al. (Dekker 2002). The strategy of 3C relies on the quantification of contacts between two genomic sites. The quantification and analysis of these contacts allows for statistical inference of the three-dimensional (3D) structure of the chromosomes. The measurement of contact sites is accomplished in the following general steps. First, chromatin between several chromosomes are cross linked using a fixative solution. Then, the chromatin is isolated and digested by an enzyme. This results in pairs of crosslinked DNA fragments that may differ linearly but are close in physical space. These separate fragments are then re-ligated, and the crosslinks are reversed, thus resulting in templates. These templates are then amplified and interrogated, usually using polymerase chain reaction (PCR) and DNA sequencing. The resulting data describes the frequency of ligation junctions between genomic loci. These relative contact frequencies describe the proximity of the chromosomes in a 3D environment. See figure 1 for a visual representation of this process.

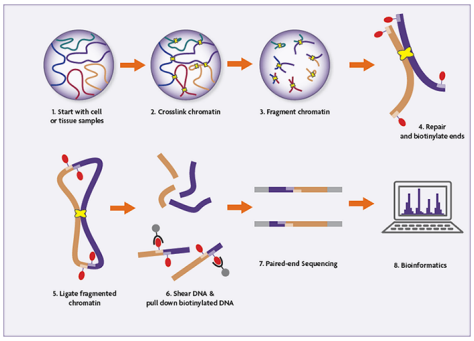

Figure 1: Pipeline for 3C data acquisition.

A defining difference between 3C methods and traditional microscopic methods is the fact that the former relies on large quantities of data to generate inference whereas the ladder relies on isolated observation. For this reason, algorithms and data analysis methods are particularly useful for the analysis of 3C data. The problem of converting interaction frequencies to an inferred 3D chromosome structure is an open problem in bioinformatics, and there are several machine learning based methods for competing this task (Oluwadare, Highsmith, and Cheng 2019). This implementation is a graph autoencoder network as an unsupervised method for reconstructing 3D chromosome models from 3C contact data.

Related Works

There are three methods that are typically used to construct the chromosomal topology from contact data: modeling based, distance-restraint, and geometric based (Park and Lin 2016). Modeling based methods relate contact frequencies to distance via some probabilistic function. From this

relation, the chromosomal structure can be inferred by maximizing the likelihood of the distribution of the loci in 3D space with respect to the chosen probabilistic function (Hu et al. 2013). These methods usually use the polymer physics of the chromatin to guide the assumptions about the function used to model the distribution of loci in space (Barbieri et al. 2013). Distance-restraint methods aim to optimize an objective function with respect to restraints derived from the contact data to find a chromosomal structure that best suits the contact data (Oluwadare, Zhang, and Cheng 2018). Usually, distance-restraint methods use a function to convert the contact matrix to a distance matrix and use these distances as the constraints for the optimization (Liu and Wang 2018). Finally, geometric based methods utilize theorems from distance geometry to recover 3D chromosomal structure from contact data alone. These methods do not rely on assumed distributions or objective functions. Rather, they rely on relationships between the eigenvectors of a distance matrix and the spatial distribution of the nodes to infer the 3D chromosomal structure (Lesne et al. 2014).

Recently, graph neural networks have been used for classification of genomic sub-compartments from contact data (Ashoor et al. 2020). Moreover, convolutional networks have been used to predict genomic interactions from sequence data (Schwessinger et al. 2020). To our knowledge, however, graph neural networks have not been used to directly predict the 3D structure of chromosomes from contact data.

Background

Hi-C Chromosome Capture

Hi-C is a 3C derivative that allows for global structural inference of entire genomes. This is accomplished by utilizing biotin-labeled nucleotides in the pre-ligation step of 3C to ensure that only ligation junctions are analyzed (van Berkum et al. 2010). The fragments are then analyzed using massively parallel DNA sequencing to generate reads. These reads are compared to the original genome, and each fragment is given a score based on the quantity of matching nucleotides. This results in a contact matrix that describes the interaction frequencies of the fragments with the original genome. The main advantage to Hi-C analysis is the fact that each locus of the chromosome is compared to each locus on the original genome which allows for insight into the global structure of the chromosome. The Hi-C method allows for all to all comparison of genomic loci, whereas other 3C technologies only allow one to one, one to all, or many to many comparisons (de Wit and de Laat 2012).

Hi-C experiments generate an $N \times N$ symmetric matrix, IF, known as the interaction frequency matrix. Here, N is the number of genomic loci observed in the experiment. The i, j^{th} entry in IF represents the interaction frequency between locus i and locus j. Thus, IF can be interpreted as the adjacency matrix of an edge weighted, undirected graph where each node represents a locus and each vertex represents an interaction frequency. Our task is to label each of these nodes with xyz coordinates in a way which best represents the true structure of the chromosome from which the Hi-C data was derived. This problem is well suited for graph neural networks due to the graphical structure of Hi-C data.

Graph Neural Networks

Graph neural networks (GNNs) can be considered as a generalization of convolutional neural networks. Convolutional networks extract multi-scale, localized spatial features by consolidating grouped features. GNNs generalize the notion of grouped features by consolidating node neighborhoods (ZHOU, JIE, and Publishers 2020). See figure 2 for a visual representation of the GNN structure.

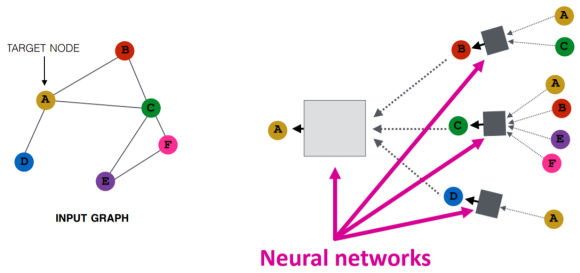

Figure 2: Visual representation of the forward propagation of node A. The network aggregates the information of the second order neighborhoods of A through the dark grey neural networks, and aggregates these outputs to make a prediction for the representation of A using the light grey neural network. These first and second order aggregations can be interpreted as the graph convolutions.

Graph neural networks aim to capture node to node relationships through the aggregation of neighboring feature spaces. The capturing of node-node relationships in graph data has made GNNs particularly useful in the task of node embedding (Cai, Zheng, and Chang 2018), (Goyal and Ferrara 2018).

Methodology

Our method relies on a GNN to find accurate 3D representations of chromosomes from their Hi-C interaction matrices in a self-supervised setting. Our method considers the Hi-C contact map to be the adjacency matrix of an edge weighted graph. From this graph, we utilize a node embedding algorithm to generate trainable node features. These node features are then sent through a message passing GNN to generate predicted 3D coordinates. From these 3D coordinates, we compute the pairwise distances between all coordinates and compare these distances to the ground-truth distances of the input and optimize the mean squared error between the true distances and the output distances.

Note that there is no known true distances corresponding to the input chromosome. It has been shown both empirically and theoretically, however, that relationship between distance and interaction frequencies is inversely exponential (Lieberman-Aiden et al. 2009), (Pombo and Nicodemi 2014), (Barbieri et al. 2012), (Chiariello et al. 2016), (Shi and Thirumalai 2019a). Thus, we can estimate the true distances of the chromosome by using formula 1.

$$d(i,j) = \frac{1}{IF_{i,j}^{\gamma}} \tag{1}$$

where $IF_{i,j}$, is the interaction frequency between sites i and j. Here, γ is a hyperparameter called the conversion factor. This conversion is used in several other distance-restraint algorithms and has been shown to be a valid means of generating ground-truth distance data (Trussart et al. 2015). In general, the value for γ is unknown and varies depending on the underlying chromosome. It has been shown, however, that γ lies in the range $[.1, 2]$ for most, common cell types (Shi and Thirumalai 2019a). Thus we will assume that the optimal conversion belongs to the interval $[0.1, 2]$ and perform a grid search in this interval to find the optimal model.

Data

Real Hi-C Data

We tested our method using real Hi-C data. The benefit of evaluating the performance on real Hi-C data is that the real data allows us to test how the algorithm responds to features that are inherent in actual Hi-C data, such as excessive noise, biases, and complicated structures, that may not be present in the synthetic data. We used the GM12878 cell line from (Rao et al. 2014) downloaded from the GSDB repository (Oluwadare et al. 2020) under the GSDB ID OO7429SF. To reduce the effects of noise in the Hi-C data, each cell line was normalized using the Knight-Ruiz normalization technique (Knight and Ruiz 2013). We will tested the performance of our method on all 23 chromosomes at 1 MB resolution.

ChiA-PET Data

Chromatin immunoprecipitation (ChIP) is a technique to investigate protein specific interactions in chromosomes. ChIP relies on antibodies to precipitate specific proteins, histones, or transcription factors from cell populations (Carey, Peterson, and Smale 2009). ChIP can also be combined with sequencing technologies to quantify these interactions. Chromosome Interaction Analysis by Paired-End Sequencing (ChIA-PET) (Li et al. 2010), is an example of such a technology. The main difference between ChiA-PET and Hi-C data is that the ChiA-PET technique measures interactions of a unique protein in the chromosome, whereas the Hi-C technique measures interactions between any proteins in the chromosome.

To further validate our results on the real Hi-C data, we compare the outputs of our method when using Hi-C data to the interaction frequencies of an orthogonal ChiA-PET data set. We performed this validation using ChIA-PET data from the NCBI GEO database (GEO accession: GSE72816) for the RNAPII ChIA-PET data from human GM12878 cells (Barrett et al. 2005). This data measures interactions between the RNA polymerase II multicomplex; a protein complex that is responsible for transcribing genes.

Evaluation

To validate the reconstructive accuracy of our method, we will use the Spearman Correlation Coefficient (dSCC). The formula for dSCC is given by equation

$$dSCC = \frac{\sum_{i\in D'}(X_i - \overline{X})\sum_{i\in\mathcal{D}}(Y_i - \overline{Y})}{\sqrt{\sum_{i\in\mathcal{D}}(X_i - \overline{X})^2 \sum_{i\in\mathcal{D}}(Y_i - \overline{Y})^2}} \tag{2}$$

where $\mathcal{D}'$ is the set of pairwise distances between all loci of the generated model, X_i is the rank of distance i in $\mathcal{D}'$, $\mathcal{D}$ is the set of wish distances corresponding to the input contact frequencies of the chromosome generated by (1), and Y_i is the rank of wish distance i in $\mathcal{D}$. Here, $\overline{X}$, $\overline{Y}$ are the mean of their corresponding ranked vectors in $\mathcal{D}'$ and $\mathcal{D}$ respectively. In general, $dSCC$ values closer to 1 imply higher model similarity.

Note that dSCC is a non-parametric measure of rank correlation. The advantage to evaluating reconstructive performance using a ranked measure of similarity is that the measure is both scale and shift invariant. Intuitively, the model may output a perfect match of the chromosome, but the xyz coordinates may be scaled by a constant and rotated in space. This scaling and rotation would be accounted for in a non-ranked measure of correlation and would likely decrease the correlation value. This decrease in correlation would falsely imply that the generated model is inaccurate when the only dissimilarity between it an the ground truth is the scale and location in space. Since the purpose of modeling the chromosome in 3D space is solely for visualization, the scale and orientation of the output should not matter. Thus, dSCC is an appropriate measure of structural similarity in this context.

Implementation

Node Feature Initialization

One challenge that graph neural networks pose in the context of our specific task is the lack of features corresponding to Hi-C data. In general, message passing graph neural networks require node features, whereas Hi-C data only defines a graph structure through weighted edges between featureless nodes. Thus, we must create node features ourselves. These node features have two desirable properties. Firstly, we would like these node features to be correlated in some sense to the underlying graph structure; i.e. nodes within regions of high connectivity should be similar. Secondly, we would like this similarity defined from the graph structure to translate to similarity of node features in Euclidean space so that the 3D Euclidean structure of the chromosome can be inferred from these features. A natural way to accomplish these two goals is to utilize a node embedding algorithm to create vectorized representations of each node and utilize the representations as the input node features.

LINE is a a node embedding algorithm that is specifically adapted to scalable use on large graphs. The general technique of LINE is as follows. Firstly, a conditional node context distribution is defined. This distribution is given by

$$p_2(v_j|v_i) = \frac{exp(u_j \cdot u_i)}{\sum_{k\in\mathcal{N}(v_i)} exp(u'_k, u_i)} \tag{3}$$

where v_i are indexed nodes and u_i are the corresponding feature representations. The empirical distribution $\hat{p_2}$ is then optimized to (7) by minimizing the KL divergence between these two distributions using stochastic gradient descent. Intuitively, LINE maximizes the probability of recreating the underlying graph from the computed node embeddings.

The LINE algorithm accounts for weighted edges in the calculation of node representations, and the representations are specifically designed to preserve similarity in Euclidean space. Thus, LINE effectively creates our desired node features from the edge-weighted graph structure of Hi-C data alone. LINE has been previously showed promising results on Hi-C data for chromosome compartmentalization prediction (Ashoor et al. 2020). Moreover, LINE is a node embedding algorithm that was specifically designed to run effectively on large graphs. Thus LINE offers more potential to scale our algorithm to be used on higher dimensional Hi-C data. For these reasons, we used LINE for our task of 3D chromosomal reconstruction.

Message Passing GNN

After the node features have been initialized, we utilize the representations as inputs into a message passing graph neural network. The advantage to utilizing a message passing GNN to estimate 3D coordinates from the input features as opposed to just using a standard neural network is that GNNs incorporate the graphical structure of the Hi-C data in the output coordinates, whereas standard neural networks have no way of interpreting graphical relationships from the data.

Our method relies on a message passing network inspired by the GraphSAGE algorithm (Hamilton, Ying, and Leskovec 2017). The message passing function of GraphSAGE is defined by averaging the node features of the one hop neighborhood of the target node. The output of the target node is then calculated by multiplying the aggregated messages and the target node features by parameter matrices and summing the two resulting vectors. Unlike GraphSAGE, however, we choose to take the weighted average of the neighborhood of the target node according to edge weights. Specifically, if we are predicting the output of node $\mathbf{x}_i$, then we first compute the messages of the neighborhood of $\mathbf{x}_i$, $M_{\mathbf{x}_i}$ using the equation

$$M_{\mathbf{x}_i} = \frac{1}{\sum_{e_{i,j} \in \mathcal{N}(\mathbf{x}_i)} e_{i,j}} \sum_{\mathbf{x}_j \in \mathcal{N}(i)} e_{i,j} \mathbf{x}_j \tag{4}$$

Where $\mathcal{N}(\mathbf{x}_i)$ is the neighborhood of $\mathbf{x}_i$ and $e_{i.j}$ is the edge weight between node i and node j. We then calculate the output of the layer using the equation

$$\mathbf{x}'_i = W_1 \mathbf{x}_i + W_2 M_{\mathbf{x}_i} \tag{5}$$

Where W_1 and W_2 are parameter matrices that are updated utilizing backpropagation.

After the GraphSAGE layer, the network passes through a three layer multilayer perceptron (MLP) to reduce the node features to three dimensions. These three features are representative of Cartesian coordinates of each node, with each feature representing the x-axis, y-axis, and z-axis respectively. We then compute the pairwise distances between all of these coordinates and compare these output distances to the ground truth wish distances generated by equation (1) using mean squared error (MSE). We then optimize the parameters of the network utilizing backpropagation and the Adam optimizer in order to minimize the MSE between the distances corresponding to the output structure and the wish distances generated by equation (1). The entire algorithm can be visualized in figure 3.

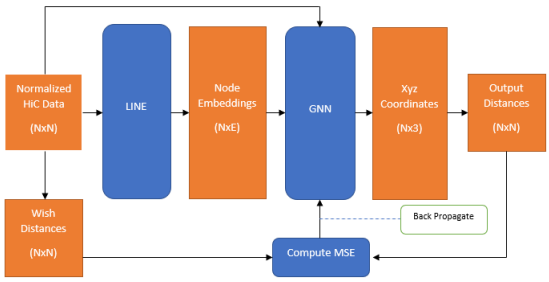

Figure 3: Flow chart of our entire algorithm.

This chart assumes we have N nodes and E features per node. In our implementation, we utilized $E = 128$. Our GNN architecture consists of one SAGE layer as defined above that outputs a hidden representation of size 64 for each node. The SAGE layer is followed by a three layer MLP that reduces the representation to size 3 for each node; the first layer reduces from 64 to 32, the second layer reduces from 32 to 10, and the third layer reduces from 10 to 3. Every layer besides the final 10 to 3 layer is followed by a ReLU activation. The GNN block of figure 3 can be visualized in figure 4.

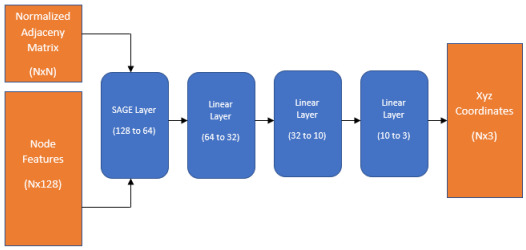

Figure 4: Flow chart of the GNN block of figure 3.

Validation

Validation On Real Hi-C Data

To validate our method, we utilized all 23 chromosomes from the GM12878 cell line at 1 MB resolution. To find the optimal model of a given chromosome, performed a grid search for the optimal value of γ in equation (1). I.e, we generated a model for each value of γ in the set $\{.1, .2, \cdots, 1.9, 2.0\}$ and selected the model that maximized

the dSCC between the output distances and the wish distances generated by equation (1). The results of this experiment are shown in figure 5.

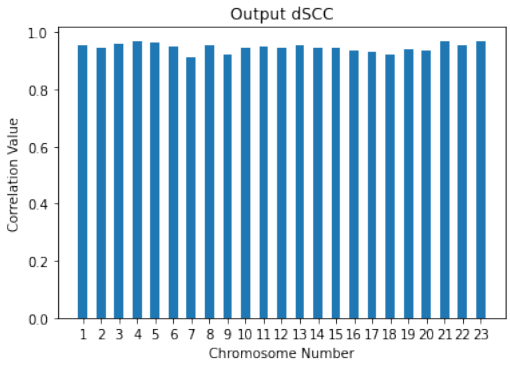

Figure 5: The dSCC between the output distances and wish distances for each chromosome at 1 MB resolution. A value closer to 1 implies higher reconstructive accuracy.

From this figure, it is clear that all dSCC values are above .9, thereby implying accurate reconstruction for each chromosome as generated by our method. Figures show the actual models generated by our method for chromosomes 4, 14, and 23.

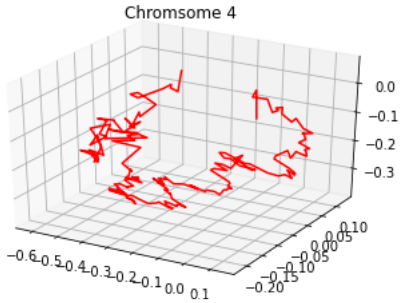

Figure 6: Generated model for chromosome 4.

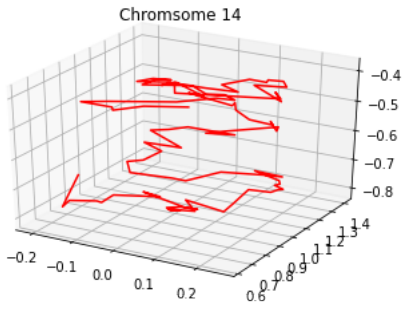

Figure 7: Generated model for chromosome 14.

To further evaluate the reconstructive accuracy of our method, the dSCC of the generated structures from each chromosome were compared to the dSCC of structures generated by three other methods other methods—3DMax (Oluwadare, Zhang, and Cheng 2018), Pastis (Varoquaux et al. 2014), and LorDG (Trieu and Cheng 2017). The performance of these methods were acquired from the GSDB data base (Oluwadare et al. 2020). The result of this comparison is shown in figure 9. Although our method is outperformed by 3DMax, our results still surpass that of the two other methods.

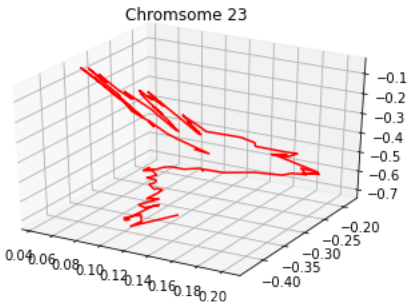

Figure 8: Generated model for chromosome 23.

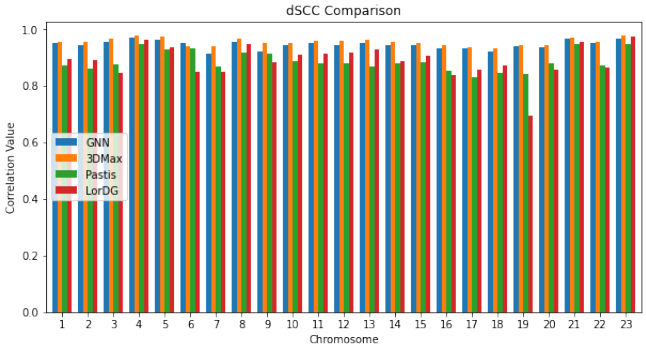

Figure 9: Comparison of the reconstructive accuracy of our method to three other distance-restraint algorithms.

Validation On ChIA-PET Data

Note that the results from figure 5 imply a high correlation between the output model and the input wish distances. Thus, we know our method can accurately estimate 3D coordinates from a set of distances. Since this correlation is calculated between the output and the input as opposed to between the output and a known 3D structure corresponding to the chromosome, this high correlation value does not necessarily imply that these models are representative of the actual structure of the chromosome. In general, however, the 3D structure of the chromosome is unknown, thereby making it difficult to validate the ability for chromosome reconstruction methods to produce representative models. We circumvent this issue by testing and comparing our generated models to orthogonal ChIA-PET data.

The ChIA-PET data provided by Barret et al. consists of contact maps measuring interactions between the RNAPII complex in all 23 chromosomes of the GM12878 cell line

(Barrett et al. 2005). From these contact maps, RNAPII loops were identified by considering contact regions that have an interaction frequency greater than or equal to 5. To validate our method, we split this ChIA-PET data into two sets: one containing looped regions and one containing non-looped regions. We then calculated the distances of our output models between the identified looped and non-looped regions separately for each chromosome. If our models are representative of the true structure of the chromosome, then distances corresponding to looped regions of our output models should typically be smaller than distances corresponding to non-looped regions. Figure 10 shows the box plots for the looped and non-looped regions for all chromosomes combined.

From this figure, it is clear that the distribution of distances corresponding to the looped regions is centered around smaller values, thereby implying that the outputs of our method are consistent with the true structure of the chromosomes.

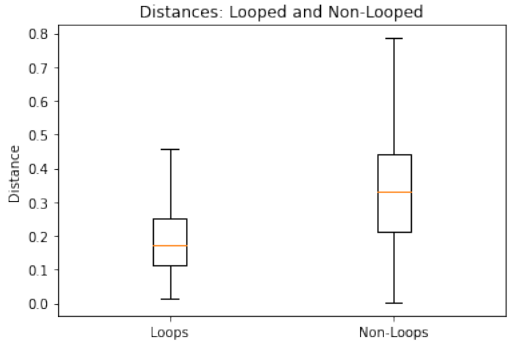

Figure 10: Box plots for the looped and non-looped regions for all chromosomes combined in the GM12879 cell line.

Conclusion

In this paper, we presented a method for inferring the spatial organisation of chromosomes from Hi-C contact data that relies on node embeddings and a graph neural network inspired by GraphSAGE. Our method relies heavily on a graphical interpretation of Hi-C data. Previous chromosome reconstruction methods ignore the graphical structure of Hi-C data and consider only contacts themselves as features, whereas our method considers the contacts themselves along with the graphical connectivity between contacts as well. This consideration of the graphical structure allows for our algorithm to consider neighborhoods of target nodes when predicting their respective 3D coordinates. We evaluated the performance of our method on the GM12878 cell line and found that our method accurately maps distances to 3D coordinates. Moreover, we found that the distributions of these distances are consistent with what we should expect from orthogonal ChIA-PET data.

There are several other aspects of our method that would be advantageous for 3D chromosome structure prediction as well. Firstly, the GraphSAGE network is learned inductively, meaning that the weights of the parameter matrices are universal across all nodes. This means that with higher dimensional data, one could learn the parameters of our network utilizing sub-graphs as mini batches for optimization. This would decrease the computational load of our algorithm when implemented on higher resolution data when compared to other distance-restraint methods. Secondly, most other distance restraint methods suffer when the input matrices are sparse. This is due to the fact that the zeros in the input matrices are effectively featureless, which in turn creates lower-coverage training data. Since features are generated externally in our method, however, the sparsity of inputs does not affect the coverage of features available for training. In fact, our algorithm would likely benefit from higher sparsity since it is from sparsity in the adjacency matrix that the graphical structure of the data is defined.

There are also several aspects of our method that could potentially be improved. Firstly, our current implementation only considers the one-hop neighborhood of the target node for the aggregation of messages in our GNN. By expanding the GNN to consider both one-hop and two-hop neighborhoods, it is likely that the performance of our method would be improved as higher order graphical structure would be considered when predicting coordinates. Secondly, our method relies on a relatively simple way of converting contacts to distances. There are several other ways with which this can be achieved that are somewhat more sophisticated than our method. Also, our message passing layer is followed by a simple MLP. By trying more complicated network layers, such as convolutional or gated layers, it may be possible to further improve the performance of our method. Finally, it has been shown that larger interaction frequency values do not necessarily imply shorter distances due to variability in the spatial organization of chromosomes in the cell population used to generate Hi-C data (Shi and Thirumalai 2019b). For this reason, Hi-C methods have been developed to consider only the interactions of single cells. This single-cell Hi-C data is much more sparse than multi-cell Hi-C data which presents further issues 3D reconstruction. Our method could potentially be used for single-cell Hi-C data to resolve the issue of heterogeneous cell populations.

References

Ashoor, H.; Chen, X.; Rosikiewicz, W.; Wang, J.; Cheng, A.; Wang, P.; Ruan, Y.; and Li, S. 2020. Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. *Nature communications* 11(1): 1173.

Barbieri, M.; Chotalia, M.; Fraser, J.; Lavitas, L.-M.; Dostie, J.; Pombo, A.; and Nicodemi, M. 2012. Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences* 109(40): 16173–16178.

Barbieri, M.; Chotalia, M.; Fraser, J.; Lavitas, L.-M.; Dostie, J.; Pombo, A.; and Nicodemi, M. 2013. A model of the large-scale organization of chromatin. *Biochemical Society transactions* 41(2): 508–512.

Barrett, T.; Suzek, T. O.; Troup, D. B.; Wilhite, S. E.; Ngau,

W.-C.; Ledoux, P.; Rudnev, D.; Lash, A. E.; Fujibuchi, W.; and Edgar, R. 2005. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic acids research* 33(suppl_1): D562–D566.

Cai, H.; Zheng, V. W.; and Chang, K. C. 2018. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering* 30(9): 1616–1637.

Carey, M. F.; Peterson, C. L.; and Smale, S. T. 2009. Chromatin immunoprecipitation (chip). *Cold Spring Harbor Protocols* 2009(9): pdb–prot5279.

Chiariello, A. M.; Annunziatella, C.; Bianco, S.; Esposito, A.; and Nicodemi, M. 2016. Polymer physics of chromosome large-scale 3D organisation. *Scientific reports* 6(1): 1–8.

Cremer, T.; and Cremer, C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews.Genetics* 2(4): 292–301.

de Wit, E.; and de Laat, W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes development* 26(1): 11–24.

Dekker, J. 2002. Capturing Chromosome Conformation. *Science (American Association for the Advancement of Science)* 295(5558): 1306–1311.

Goyal, P.; and Ferrara, E. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* 151: 78–94.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs .

Hu, M.; Deng, K.; Qin, Z.; Dixon, J.; Selvaraj, S.; Fang, J.; Ren, B.; and Liu, J. S. 2013. Bayesian Inference of Spatial Organizations of Chromosomes. *PLoS computational biology* 9(1): e1002893.

Knight, P. A.; and Ruiz, D. 2013. A fast algorithm for matrix balancing. *IMA journal of numerical analysis* 33(3): 1029–1047.

Lesne, A.; Riposo, J.; Roger, P.; Cournac, A.; and Mozziconacci, J. 2014. 3D genome reconstruction from chromosomal contacts. *Nature methods* 11(11): 1141–1143.

Li, G.; Fullwood, M. J.; Xu, H.; Mulawadi, F. H.; Velkov, S.; Vega, V.; Ariyaratne, P. N.; Mohamed, Y. B.; Ooi, H.-S.; Tennakoon, C.; et al. 2010. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology* 11(2): 1–13.

Lieberman-Aiden, E.; van Berkum, N. L.; Williams, L.; Imakaev, M.; Ragoczy, T.; Telling, A.; Amit, I.; Lajoie, B. R.; Sabo, P. J.; Dorschner, M. O.; Sandstrom, R.; Bernstein, B.; Bender, M. A.; Groudine, M.; Gnirke, A.; Stamatoyannopoulos, J.; Mirny, L. A.; Lander, E. S.; and Dekker, J. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (American Association for the Advancement of Science)* 326(5950): 289–293.

Liu, T.; and Wang, Z. 2018. Reconstructing high-resolution chromosome three-dimensional structures by Hi-C complex networks. *BMC bioinformatics* 19(Suppl 17): 496.

Oluwadare, O.; Highsmith, M.; and Cheng, J. 2019. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biological Procedures Online* 21(1): 7.

Oluwadare, O.; Highsmith, M.; Turner, D.; Lieberman Aiden, E.; and Cheng, J. 2020. GSDB: a database of 3D chromosome and genome structures reconstructed from Hi-C data. *BMC molecular and cell biology* 21(1): 60.

Oluwadare, O.; Zhang, Y.; and Cheng, J. 2018. A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. *BMC genomics* 19(1): 161.

Park, J.; and Lin, S. 2016. Impact of data resolution on three-dimensional structure inference methods. *BMC bioinformatics* 17(1): 70.

Pombo, A.; and Nicodemi, M. 2014. Physical mechanisms behind the large scale features of chromatin organization. *Transcription* 5(2): e28447.

Rao, S. S. P.; Huntley, M. H.; Durand, N. C.; Stamenova, E. K.; Bochkov, I. D.; Robinson, J. T.; Sanborn, A.; Machol, I.; Omer, A. D.; Lander, E. S.; and Aiden, E. L. 2014. A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell (Cambridge)* 159(7): 1665–1680.

Sati, S.; and Cavalli, G. 2017. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma* 126(1): 33–44.

Schwessinger, R.; Gosden, M.; Downes, D.; Brown, R. C.; Oudelaar, A. M.; Telenius, J.; Teh, Y. W.; Lunter, G.; and Hughes, J. R. 2020. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature methods* 17(11): 1118–1124.

Shi, G.; and Thirumalai, D. 2019a. Conformational heterogeneity in human interphase chromosome organization reconciles the FISH and Hi-C paradox. *Nature communications* 10(1): 1–10.

Shi, G.; and Thirumalai, D. 2019b. Conformational heterogeneity in human interphase chromosome organization reconciles the FISH and Hi-C paradox. *Nature communications* 10(1): 3894.

Trieu, T.; and Cheng, J. 2017. 3D genome structure modeling by Lorentzian objective function. *Nucleic acids research* 45(3): 1049–1058.

Trussart, M.; Serra, F.; Baù, D.; Junier, I.; Serrano, L.; and Marti-Renom, M. A. 2015. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. *Nucleic acids research* 43(7): 3465–3477.

van Berkum, N. L.; Lieberman-Aiden, E.; Williams, L.; Imakaev, M.; Gnirke, A.; Mirny, L. A.; Dekker, J.; and Lander, E. S. 2010. Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of visualized experiments : JoVE* (39).

Varoquaux, N.; Ay, F.; Noble, W. S.; and Vert, J.-P. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics (Oxford, England)* 30(12): i26–i33.

ZHOU, Z. L.; JIE; and Publishers, M. . C. 2020. *INTRODUCTION TO GRAPH NEURAL NETWORKS*, volume 45. MORGAN CLAYPOOL PUBLISH, 1 edition. ISBN 1939-4608.