Predicting High-order Chromatin Interactions from Human Genomic Sequence using Deep Neural Networks

Rui Peng

Sunday 3rd December, 2017

Background. Understanding and modeling high-order chromatin organization is a fundamental problem in computational genomics. Chromosome fold into complex shapes by itself mostly by following rules written in the genomic sequence. People have been studying the process closely trying to discover the rules from measured interaction frequencies in Hi-C matrix. Following recent success in using deep learning to model point interactions such as enhancer-promoter interactions directly from genomic sequence, there seems to be similar pathway from genomic sequence to general Hi-C interactions, which leads to modeling general rules behind high-order chromatin interactions.

Aim. We consider the problem of identifying high-order chromatin interactions from human genomic sequence solely. Concretely, we seek to explore how we can use genomic sequence alone to predict pairwise interaction frequencies of loci that forms Hi-C matrix and whether we can identify high-order interaction patterns such as topologically associating domains (TADs) from it.

Data. The dataset we use consists of two main parts, raw genomic sequences and the their pairwise Hi-C interaction measurement. The raw genomic sequence dataset is from GRCh37/hg19 produced by Genome Reference Consortium. The Hi-C measurement data is from GM12878 repository.

Methods. We take a deep learning approach taking genomic sequences as input and perform regression on Hi-C interaction measurements. The deep neural network use a combination of convolution and recurrent neural network LSTM to predict interaction frequency. Once trained, the model can be used to predict Hi-C matrix for any given region and the predicted interaction matrix can be analyzed for higher-order interaction structures.

Results. Our model is capable of extracting sequence interaction patterns and predict interaction matrix visually similar to Hi-C measurement. According so some metrics we define, the level of structural agreement is high and TAD structures are recovered with good accuracy.

Conclusions. We show for the first time that we can use a deep neural network to model interaction between genomic sequences that can predict interaction matrix where high-order chromatin structures, such as TADs, are highly identifiable.

Keywords: Hi-C interaction, computational genomics, chromatin interaction, deep learning.

DAP Committee members:

Jian Ma ⟨*jianma*@*cs.cmu.edu*⟩ (CBD, MLD);
Barnabas Poczos ⟨*bapoczos*@*cs.cmu.edu*⟩ (MLD);
Shashank Singh ⟨*sss*1@*andrew.cmu.edu*⟩ (MLD);

1 Introduction

The human genome consists of over 3 billion nucleotides and is contained in 23 pairs of chromosomes. If the chromosomes are aligned end-to-end, the genome would measure roughly 2 meters long. However, the genome functions within the cell nucleus in sphere shape smaller than 10 μm. This suggests that the genome folds into a complex compact 3D structure. It has been increasingly appreciated that understanding how chromosomes function (e.g. gene expression and replication) heavily rely on detailed knowledge of the 3D structure, which boils down to studying the interaction patterns of genomic sequence. Recent studies Lieberman-Aiden et al. (2009), Fudenberg and Mirny (2012), Gibcus and Dekker (2013), Lajoie et al. (2015) suggest 5 main types of interaction patterns in decreasing level of scale: Cis/Trans interaction ratio, distance-dependent interaction frequency, genomic compartments, topologically associating domains and point interactions. The first 4 are considered higher order and in this work we mainly focus on topologically associating domains (TADs). TADs are sub-Mb (million base pair) structure where interaction frequencies within are elevated and they are heavily associated with gene regulation, according to Shen et al. (2012), Symmons et al. (2014).

To study the shape of human genome, we need to have some ground truth information on how loci on chromosomes interact. Hi-C measurements from Lieberman-Aiden et al. (2009) is a widely used measurement of chromosome conformation capture method that records interaction frequencies between all pairs of loci in the genome. It can be thought as a proximity measure of loci pairs in space with a higher interaction frequency usually indicating a closer distance. If we plot out all interaction frequencies between all loci pairs in chromosomes, we get a interaction map backed by a matrix of interaction frequency values. If we analyze the Hi-C interaction matrix map of a chromosome, we can visually identify some high-order chromatin structures, from genomic compartments to TADs. And more programmatically there are algorithms Lieberman-Aiden et al. (2009), Dixon et al. (2012), Crane et al. (2015) that can identify them from such matrices.

At the same time, we do know those high-order chromatin interaction structures are eventually governed by the underlying genomic sequence. In other words, there should be a direct path from raw genomic sequence to high-order chromatin interactions. Or equivalently, taking the Hi-C interaction matrix as an intermediate step, we should be able to use raw genomic sequence to predict interaction frequencies in the Hi-C matrix and use the predicted Hi-C interaction matrix to discover high-order chromatin interactions.

2 Problem Statement

In this work, we consider the problem of predicting values in Hi-C matrix from raw genomic sequence alone using machine learning with the ultimate goal of predicting high-order chromatin interactions.

The purpose of predicting Hi-C matrix is to help predict high-order chromatin interaction structures. We know from previous sections that once we have a Hi-C matrix, identifying high-order structures just involves running some algorithm on the Hi-C matrix. So according to this problem definition, we can evaluate the quality of Hi-C matrix prediction from 1) the overall agreement

between predicted Hi-C matrix and measurement matrix and 2) the quality of high-order structure that is identifiable from the predicted Hi-C matrix. And in similar way, a good model for Hi-C matrix prediction will in turn contribute to good high-order chromatin interaction discovery.

3 Related Work

Chromatin interaction prediction has been a hot topic in the computational genomics community. Over the past years, there has been extensive research effort into chromatin interaction modeling, especially using machine learning techniques. On the subject of enhancer-promoter interaction, we see Singh et al. (2016) develop convolutional and recurrent deep neural networks to predict interaction types. For more general point interactions, a locus-specific support vector machine is developed by Nikumbh and Pfeifer (2017) to predict genome-wide interaction partners. Such methods using modern machine learning techniques have all received good results and boosted our confidence in using machine learning to model chromatin interactions.

One key difference in our problem, though, compared to above methods, is that we wish to model beyond point interactions to identifying high-order interaction patterns. We wish to learn broader features that can help determine higher-order chromatin organizations. This difference can pose more challenge to our problem.

Another recent study aiming at chromosome structure prediction that is similar in scope with our problem is Di Pierro et al. (2017). In their work, a neural network is employed to model the relations between epigenetic markers to predict Hi-C interaction frequency. Next, the predicted information is used to model genomic organization along with an energy model to infer spatial conformation of chromosome. Our problem differs with it in that we do not assume epigenetic information to predict Hi-C as we believe those epigenetic signals can be picked up by our machine learning model. In other words, we are addressing a more fundamental problem: what are the patterns and processes in genomic sequence lead to all the spatial organization we are observing.

4 Data

Data for this work is comprised of 2 parts, broadly contributing to source of input and output target to our modeled prediction problem.

Raw genomic sequences.

The dataset for genomic sequence is the February 2009 human genome reference sequence assembly (GRCh37/hg19) produced by Genome Reference Consortium. It contains a sequenced reference human genome for approximately 3 billion base pairs and covers 23 pairs of chromosomes. The raw sequences are marked with the source chromosome and are presented in plain text form from an alphabet of ATCG and N, where ATCG represent the four nucleotide types found in DNA and N denotes unresolvable nucleotide from sequencing step. The dataset if publicly available at `https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg19`.

Hi-C measurement.

The Hi-C measurement of interaction frequency data is from the three-dimensional map of the human genome used by Rao et al. (2014). Specifically, we use the raw observed contact matrices for the combined (primary+replicate) intra-chromosome interaction map for GM12878 cell line. For our use case, we choose the version with 5kb (kilobase) measurement resolution and MAPQ>30 readings. The measurements are divided according to chromosomes and each row in a chromosome-specific file contains three space delimited values (loc_A, loc_B, $freq$). loc_A and loc_B are base pair offset locations of two sequence fragment that are interacting and $freq$ is a positive floating point value of interaction frequency. The Hi-C matrix of a region of interest can be constructed combining the interaction freuqncies $freq$ of all pairs of loc_A and loc_B in the region. The data files can be found at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525`.

5 Method

Methods like SPEID Singh et al. (2016) in prediction enhancer-promoter interaction have shown the power of deep learning models in modeling sequence interaction patterns. As our problem of predicting general interaction anywhere in the Hi-C matrix is effectively a generalization of enhancer-promoter interaction formulation in SPEID, we learn from the architecture of SPEID and adopt a similar deep learning approach.

5.1 SPEID

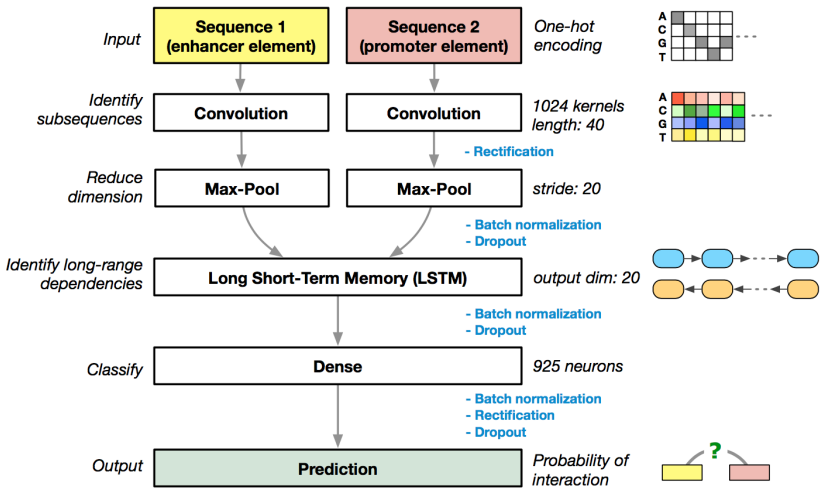


Figure 1: Architecture of deep learning model SPEID introduced for enhancer-promoter interaction prediction. Figure from paper Singh et al. (2016).

SPEID is an excellent model providing good sequence pattern extraction and long-range feature combination capabilities as shown in its high prediction accuracy in enhancer-promoter interac-

tion. It extracts general sequence features (such as CTCF motifs) from using one-hot encoding of sequence inputs in conjunction of 1D convolution. Sequence features from two input branches are concatenated after dimensionality reduction introduced by max-pooling layer and fed into a LSTM Hochreiter and Schmidhuber (1997) where they are scanned bidirectionally. The LSTM processes the condensed sequence features by linearly sweeping the combined input features and identify long-range dependencies between them. This LSTM step is responsible for combining interaction signals at various positions in the sequences and are compared to 'regulatory grammar' Quang and Xie (2016). The last dense layer serves as final step combining of interaction signals across the extent of sequences before making probabilistic prediction.
Besides the elegant architecture, it also includes various techniques such as batch normalization Ioffe and Szegedy (2015) and dropout Srivastava et al. (2014) to make the model easier to train and more robust to overfitting.
All the above aspect of SPEID suggests it is a promising starting point for developing models for our problem.

5.2 Our Model

We start by taking the general framework in SPEID and make adaptations to it. Similar to SPEID, the sequences at the interacting loci of interest are passed to the deep learning model as sequence inputs. Instead of performing classifications of whether two sequence interact, we remove the last logistic mapping in the output layer and perform regression on the Hi-C interaction frequencies. We initially start with such simple adaptation, only to find the model struggle to produce similar patterns displayed in Hi-C measurement matrix. We suspect it is due to the much elevated modeling complexity of general interaction in addition to the fact that regression is naturally harder than classification. It seems that the model can benefit from more direct help from other features attainable from sequence. Hence we consider augmenting the model with 3 more branches of auxiliary input:

1. **genomic distance between two input sequences.**
   One key aspect where our problem extends from enhancer-promoter interaction prediction is that we have extra spatial information in the Hi-C matrix we are predicting. This suggests in order to reconstruct the Hi-C matrix of high accuracy, we need to incorporate some spatial information between the two input sequences. From the theory in polymer-physics De Gennes (1979), Fudenberg and Mirny (2012), we know the interaction frequency between two sequences decreases, on average, as their genomic distance increases. This suggests extra information of the genomic distance between the input sequences might be a helpful indicator of the interaction frequency, especially when the large distance provides negative signal that cancels out many positive signals identified by other parts of the model, making prediction much easier for many off-diagonal entries in the Hi-C matrix.

2. **CTCF counts in the sequence between two input sequence locations.**
   One major type of high-order chromatin interaction prevalent in Hi-C matrix is TADs. It has been discovered that CTCF motif bindings sites are usually enriched at TAD boundaries

Dixon et al. (2012), Van Bortle et al. (2014). As TADs are an important type of contributor to sub-Mb structures and interaction frequency fluctuations, CTCF information that tend to co-occur with TADs seems helpful for our model as well. We choose to include the counts of CTCF motifs that lies in the sequence between the two input sequences as another auxiliary input and hope that it shed light on the how many TADs are between the input sequence pair and in turn help determine the interaction frequency between the pair.

3. **helper sequence at genomic middle of two input sequence locations.**
We realize after much trials after including the above 2 auxiliary inputs that our model still predict relatively poorer for input pairs across TAD boundaries. It reflects that CTCF count alone might not be enough to provide adequate hints on TADs. We propose to add another branch of helper sequence that lies in the genomic middle of the two input sequences. The helper sequence (seq_3) goes into feature extraction as well and its extracted features are fed into LSTM along with the features from two input sequences (seq_1 and seq_2). As a result, the LSTM will be learning interaction relations between triplet (seq_1, seq_3, seq_2). Due to the processing order of (seq_1, seq_3, seq_2) in LSTM, the LSTM can leverage long range interaction signals between (seq_1, seq_3) and (seq_3, seq_2) to help with predicting (seq_1, seq_2).

   The introduction of helper middle sequence essentially expands the view of model that includes more sequence features to make use of. In perfect scenario we can include the entire middle sequence of variable length, however, it impose too much computational burden and require heavy make-over of the overall architecture. We instead choose the middle sequence that has the same length as the input sequence to simplify computation and model complexity.

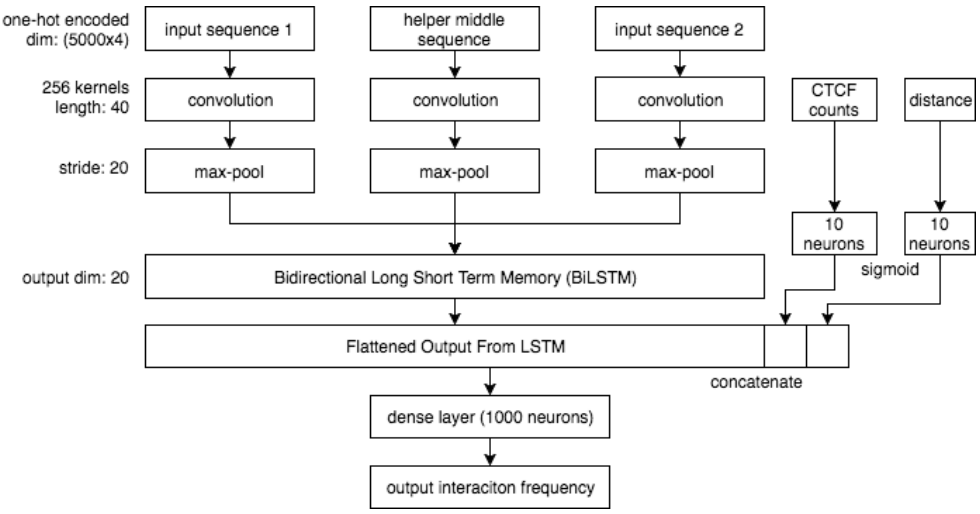


Figure 2: Architecture of our model. Compared to SPEID, it is augmented with components to take auxiliary helper inputs from sequences that serve as more direct hints to the model. Similar rectification, batch normalization and dropout omitted in the graph.

After adding auxiliary inputs, the full model now looks like Figure 2. The extra two layers of dense

neurons right after CTCF and distance input layers are for modeling necessary normalization or preprocessing of the two new input branches. The processed CTCF and distance are merged with the flattened output of the LSTM and form an augmented feature vector. The augmented vector is treated as input to a multi-layer perceptron with one hidden layer and single floating point value output layer representing the final predicted interaction level.

5.3 Training and Testing

Dataset Generation.
From the way we formulate the problem, the model needs to train and test on tuples of data (seq_1, seq_2, seq_3, $dist$, $ctcf$, y). Since such form is not readily available from the original datasets, we perform sampling from the Hi-C matrix for interaction values of y and extract sequences seq_1, seq_2, seq_3 from corresponding locations in the raw sequence dataset and calculate $dist$ and $ctcf$. $dist$ is directly computed as the difference of the genomic locations of seq_1 and seq_2 and is measured in kb. $ctcf$ is computed by feeding the genomic sequences between seq_1 and seq_2 to FIMO Grant et al. (2011) (a software that identifies motif occurrences by matching motif sequence patterns with input sequence) and extract the count of occurrences. The human CTCF motif sequence pattern definition is downloaded from `http://meme-suite.org/doc/download.html`. We use `chrA:Bm-Cm:Dk` to represent sampling from B Mb - C Mb region of chromosome A for D thousand sample tuples of data. We also notice almost all significant values in the Hi-C matrix are no more than certain distance apart from the diagonal. So instead of randomly sample interactions from the entire square matrix of the region, we only sample uniformly from positions within a diagonal band with width w. This directly suggests that we only train on seq_1, seq_2 can not be more than w apart and it can be thought as a way to boost training efficiency – since for pairs that are more than w apart, we are very likely to deduce low interaction frequencies just from the $dist$ input itself. We determine $w = 2.5Mb$ to be a good value since it is just enough to cover most significant interactions.

Training Configurations.
We implement the model in Section 5.2 in Tensorflow/Keras and use GPUs to speed up convolution and dense layer calculation. We use MSE (mean squared error) as loss function, Adam Kingma and Ba (2014) as optimizer with initial learning rate of 1e-5 and train for 50 epochs. We employ a learning rate decay policy of reducing learning rate by a factor of 5 once the training loss plateaus for 3 epochs.
We also leverage a trick to train on $log(y)$ instead of y directly. It has the advantage that the $log(y)$ will be almost normally distributed rather than previously highly skewed and this normality helps lower the difficulty of optimization.

Testing scenarios.
Testing in our problem is defined as using the trained model to predict for pairs of loci that are not seen in the training set. As our ultimate goal beyond prediction is to identify interaction structures from predicted Hi-C matrices, we need to produce predicted Hi-C matrices for some testing regions in the testing step. From the spatial correspondence of training region and testing region, there can be two types of tests we can perform.

1. Testing on part of training region. The test data essentially comes from some region of the training set but does not include exact loci pairs in the training set. We call it **in-region testing**. This type of test validates the capability of the model to reconstruct the Hi-C matrix from a small fraction of values in matrix. It is effectively a summarization task where the model learns key features to summarize the interaction patterns.

2. Testing on a different region outside the training region. The test data comes from either a different region but same chromosome or some region in another chromosome. We call it **out-region testing** and it can be out-region-in-chromosome or out-region-out-chromosome. This type of test challenges the capability of the model to predict on sequences that are potentially drastically different from training sequences. It is a generalization task where the model needs to learn transferable features that can work well on unseen inputs.

Out-region tests are considered harder than in-region tests since it requires the learned features to be highly transferable. Such tests are thus thought to be stretch tests.

6 Analysis

In this section we describe how we measure the performance of our model. We hold the general purpose to predict Hi-C matrix from sequence so that we can use the predicted Hi-C matrix for interaction structure prediction. Despite our model effectively predict the Hi-C matrix from predicting point interactions between sequence pairs, the usual error metric for point predictions such as MSE no longer fit in our context. Instead, we propose two metrics that are more indicative of overall agreement of predicted Hi-C matrix and truth Hi-C matrix in terms of the high-order interaction patterns it identifies. We believe they are quite indicative of high-order interaction patterns are reasonably interpretable.

1. **Trend score: agreement in insulation score profile.**
   We want to quantify the general spatial shapes that can be identified in the Hi-C matrix. For this purpose, we make use of insulation score profile. Insulation score profile Mizuguchi et al. (2014) is a useful 1-dimensional summarization of the interaction frequencies close to the diagonal of the 2-dimensional Hi-C matrix. It is computed as average of all interaction frequencies in square sliding along the matrix diagonal and generate a set of local interaction aggregations at various locations. In mathematical terms, for contact matrix C, insulation score $R_j(s)$ quantifying the relative frequency of contacts occurring over a bin j at a distance s is:

$$R_j(s) = \sum_{k=j-s2}^{k=j+s2} \frac{C_{k,k+s}}{mean_s(\sum_{k=j-s2}^{k=j+s2} C_{k,k+s})} \tag{1}$$

   This carries more intuition if we consider in terms of sequence – the score profile assigns a value for each locus, representing the aggregated interaction frequencies between it and all its neighboring locus within the a certain window size. The profile is usually presented as

a curve (example in Figure 3 with high and low values where the higher values indicating the locus is in a strongly interacting neighborhood. This matches our purpose well in that the changes in insulation score profile can help indicate strongly interacting regions and their boundaries.

We can compute insulation score profiles for both predicted Hi-C matrix and truth matrix and get two profile curves. Now we only need to measure the agreement of the profile curves, in particular their trends as trends directly affect matches in the strongly interacting regions. We report a commonly used statistical measurement of trend similarity between curves called Spearman's rank correlation.

For two curves X_i and Y_i with their rank rgX_i, rgY_i, the Spearman's rank correlation r_s is defined as follows: (ρ is the Pearson correlation coefficient, $cov(rg_X, rg_Y)$ is covariance of ranked variables and σ denotes standard deviation):

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}. \tag{2}$$

Spearman's rank correlation r_s measures the degree of monotonicity between X_i and Y_i. In simpler terms, it assesses how the values in two profile curves go up or down together. It ranges from -1 to +1 with +1 indicating a perfect match in trends. In later sections, we refer to this metric as **trend score**.

Since we can get different insulation score profiles with different window sizes of aggregation, we report r_s agreement score in insulation profiles in different scales for the following window sizes {200kb, 400kb, 800kb}.

2. **TAD score: agreement in TADs discovered.**

Another way to quantify matches in interaction structures is to directly measure the TADs that can be discovered in both Hi-C matrices. We pick TADs as our primary interaction structure type of interest because TAD contributes to the majority of interaction shapes in Hi-C matrices. Since TADs can be directly obtained from insulation profile curve by extracting local minima as TAD boundaries, we can run TAD calling from insulation profile curves of predicted and truth Hi-C and compare the quality of match between two sets of discovered TADs. The boundary determination process is illustrated in Figure 3.

In particular, if we have predicted TADs $\mathcal{X} = \{(X_1^s, X_1^e), \ldots, (X_m^s, X_m^e)\}$ and ground truth TADs $\mathcal{Y} = \{(Y_1^s, Y_1^e), \ldots, (Y_n^s, Y_n^e)\}$ where (X_1^s, X_1^e) denotes a mathematical interval of base pair offsets from start to end, we define a quality of match t, representing the fraction of ground truth TADs discovered by prediction:

$$t = \frac{k}{n} \tag{3}$$

$$k = \sum_{i=1}^{n} \mathbb{1}\left\{\exists (X_j^s, X_j^e) \in \mathcal{X} \ s.t. \ \frac{(X_j^s, X_j^e) \cap (Y_i^s, Y_i^e)}{(X_j^s, X_j^e) \cup (Y_i^s, Y_i^e)} > 0.95\right\} \tag{4}$$

In this definition, k is the number of matches in the predicted TADs. The intersection-over-union criterion accounts for the positions and sizes of TADs being compared and is analogous to how bounding boxes are matched in object detection tasks in computer vision. The 0.95 threshold is chosen empirically so that it gives an the appropriate amount of slack for slight mismatch of interval boundaries. This metric t is referred as TAD score in later sections.

Technically, the number of TADs extracted from insulation score profile depends on a cutoff insulation score level of choosing. In our experiments, we always make sure we choose proper thresholds so that same number of TADs are extracted from two insulation profile curves.

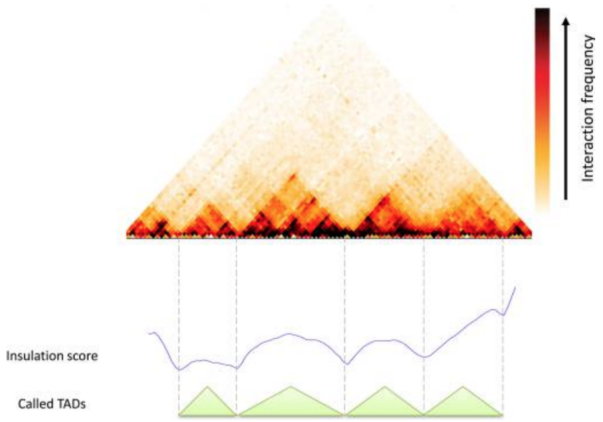


Figure 3: TAD boundaries can be extracted as local minima in the insulation score curve.

7 Results

We present results from 3 key models we have trained. They all use the same neural network architecture described in Section 5.2 but differ in the training data, respectively:

- model 1: `chr22:40m-50m:120k`
- model 2: `chr22:25m-50m:300k`
- model 3: `chr22:0m-50m:300k + chr20:0m-50m:300k`

We use three main testing regions:

- `chr22:40m-45m`, to show in-region performance for all 3 models.
- `chr22:35m-40m`, to show out-region-in-chromosome performance for model 1 and in-region performance for model 2 and 3.
- `chr21:35m-40m`, to show out-region-out-chromosome performance for all 3 models.

Table 1: Performance metrics of trained model on different test regions.
In-region tests are marked green and out-region tests are marked red.
(1: ws means window size for the trend score. 2: when #TADs in ground truth is discovered.)

model	test region	trend score [1] ws=200kb	ws=400kb	ws=800kb	TAD score (#TAD) [2]
chr22:40m-50m:120k	chr22:40m-45m	0.96	0.98	0.99	0.86(7), 0.83(6)
	chr22:35m-40m	-0.14	-0.07	0.15	0(6), 0(5), 0(4)
	chr21:35m-40m	0.13	0.54	0.61	0.16(6), 0(5)
chr22:25m-50m:150k	chr22:40m-45m	0.69	0.81	0.95	0.57(7), 0.66(6), 0.6(5)
	chr22:35m-40m	0.79	0.94	0.97	0.4(5), 0.5(4)
	chr21:35m-40m	0.31	0.58	0.64	0(5), 0(4)
chr22:0m-50m:300k + chr20:0m-50m:300k	chr22:40m-45m	0.51	0.72	0.93	0.33(6), 0.4(5), 0.5(4)
	chr22:35m-40m	0.64	0.92	0.96	0.33(6), 0.4(5)
	chr21:35m-40m	0.15	0.61	0.63	0(6), 0(5)

Following the setup, we present results as following:

1. Numeric results in terms of trend score and TAD score are summarized in Table 1:

2. Besides metrics, we also plot the predicted matrix along with ground truth Hi-C matrix for the test regions for more visual feedback on the matching of interaction shapes in the matrix.

   Figure 4 shows an in-region test `chr22:40m-45m` prediction from all 3 models and the ground truth Hi-C matrix, showing good match in visual shapes. Figure 5 shows an out-region test `chr21:35m-40m` for all 3 models, showing little structure discovered.

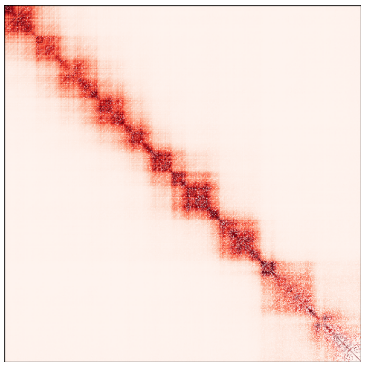
3. To show that the visual feedback and metrics we defined agree, in Figure 6 we plot a comparison of prediction vs truth matrix generated from model 1 with TADs identified in the matrices along with the insulation score profiles. We see that visually similar matrices have similar insulation profiles and can extract similar TADs.

8 Discussion

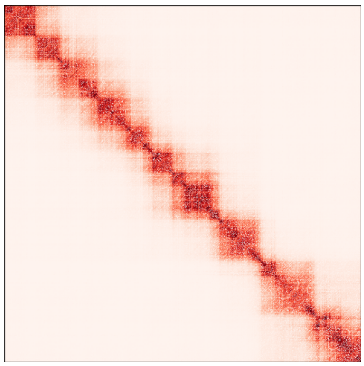
We can see a natural division of performance between in-region tests and out-region tests.

8.1 In-region tests.

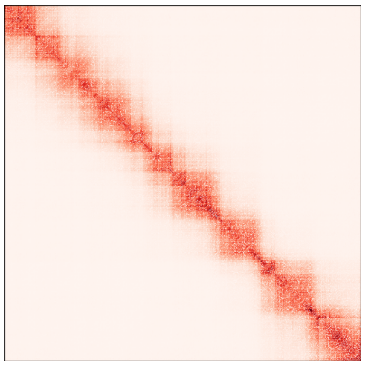
For in-region tests, the trend score is usually significantly greater than 0.5, showing a strong positive match in trends. The TAD score also high for many cases, at least identifying half of TADs with high precision. Both metrics indicate a good agreement in high-level interaction patterns between prediction matrix and truth matrix. Such agreement is also visually confirmed in Figure 6 and 4. We tribute these good performance to the ability of the deep learning model to summarize well from relatively small number of samples. For example in Figure 4, model 1 and 2 reconstruct the



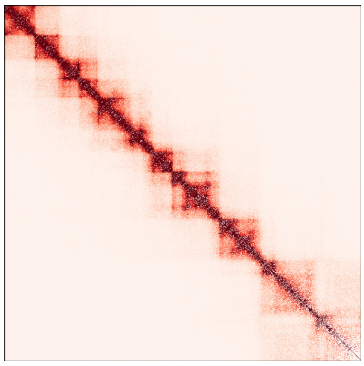
(a) model 1: trained on chr22:40m-50m.



(b) model 2: trained on chr22:25m-50m.

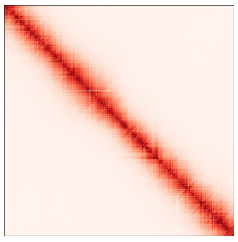


(c) model 3: trained on chr22:0m-50m+ chr20:0m-50m.

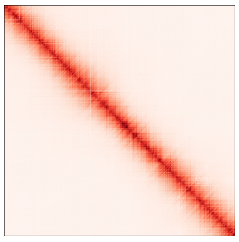


(d) ground truth Hi-C.

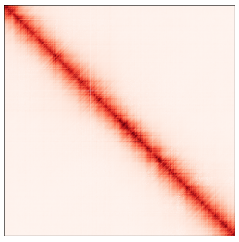
Figure 4: It-region test `chr22:40m-45m`. Predicted Hi-C matrices from 3 models, compared with ground truth.



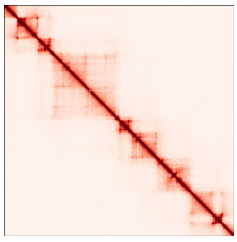
(a) model 1: trained on chr22:40m-50m.



(b) model 2: trained on chr22:25m-50m.

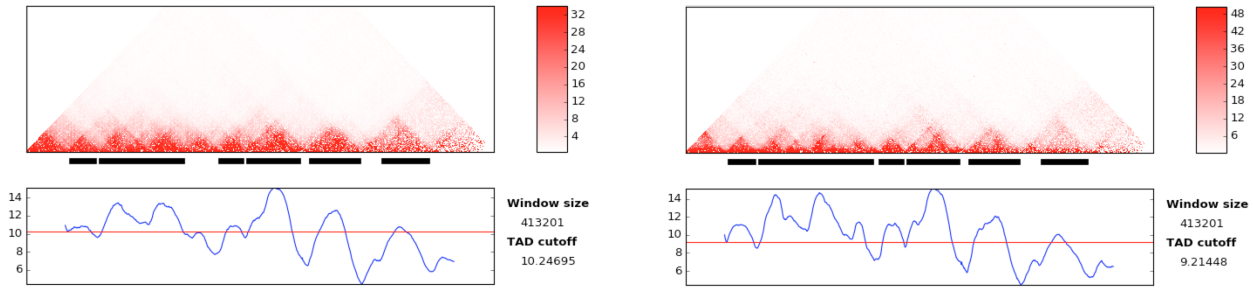


(c) model 3: trained on chr22:0m-50m+ chr20:0m-50m.



(d) ground truth Hi-C.

Figure 5: Out-region test `chr21:35m-40m`. Predicted Hi-C matrices from 3 models, compared with ground truth.



(a) Insulation profile curve and extracted TADs from predicted Hi-C matrix from model 1.

(b) Insulation profile curve and extracted TADs from ground truth Hi-C matrix.

Figure 6: Agreement between high trend score, high TAD score and visual similarity of Hi-C matrices. Example shown on test region `chr22:40m-45m`. Trend score=0.98, TAD score=5/6.

target matrix very well despite only seeing 6% of data points from the test region; model 3 has produced decent results with only seeing 3%.
It appears the model is very effective in learning the long range interaction patterns and produce a Hi-C matrix that is highly indicative of high-order chromatin interaction patterns. We have established empirical proof that we can build a model to predict high-order interaction patterns from only sequence inputs with high accuracy. This should be a new learning for the research field.

8.2 Out-region tests.

For all the out-region tests, we see much lower trend score than in-region and TAD scores are almost always 0. The visual similarity between predicted Hi-C matrix and true Hi-C matrix is not strong either. We do see the positive trend score in out-region tests and it suggests the model has learned some features at least to produce positively correlated trends, yet it is insufficient. The performance degradation is very likely due to the fact that the models failed to learn a transferable sequence interaction pattern to apply to unseen regions. And by the comparison of model 2 and model 3, it seems increasing the amount of training data (adding more samples from more regions) does not help with either metrics.
From the perspective of machine learning and optimization, the fact that the features learned from training set can not effectively transferred to test set indicates our model has been trapped in local minimum. We have tried a few ways to counter it including 1) tuning model architectures (reducing number of neurons, changing activation functions, etc) 2) increasing the amount of training data to the maximum the server can hold in memory 3) trying to increase the variance of training samples, and they have all failed to suggest significant improvement. This make us think, as a suggestion for future work, that maybe we need to work on a better formulation of the problem. So far the work is all based on a pairwise point interaction prediction framework, and thus in optimization steps the model gets feedback on how to produce closer interaction frequencies for points in the Hi-C matrix. However, the ultimate goal of our problem is to perform structural prediction and feedback from structural disagreement has not been properly incorporated into training the model.

As a future direction, we suggest:

- building a better error measure for structural difference into the objective of the model so that the model can use back propagation improve towards structural similarity.
- devising a new framework beyond pairwise point prediction that can include more direct structural elements in the model setup.

9 Conclusion

In conclusion, we have explored the task of predicting Hi-C matrix directly from sequence and extract high-order chromatin interaction structures from the predicted Hi-C matrix. Our proposed deep learning method has shown good ability to extract complex sequence features and work well in reconstructing Hi-C regions it has been trained on and can show strong structural agreement in predicted Hi-C matrix both from visual confirmation and from 2 metrics we propose. It has answered the questions we proposed and generated empirical evidence that high-order chromatin interaction patterns can be extracted from raw genomic sequence. At the same time, we do admit that this deep learning approach lacks adequate generalization ability to work well on all Hi-C matrix prediction cases. We have suggested directions for future work and certainly welcome other researchers in this field to join force on this fundamental challenge.

References

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J., and Meyer, B. J. (2015). Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature*, 523(7559):240–244.

De Gennes, P.-G. (1979). *Scaling concepts in polymer physics*. Cornell university press.

Di Pierro, M., Cheng, R. R., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, page 201714980.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

Fudenberg, G. and Mirny, L. A. (2012). Higher-order chromatin structure: bridging physics and biology. *Current opinion in genetics & development*, 22(2):115–124.

Gibcus, J. H. and Dekker, J. (2013). The hierarchy of the 3d genome. *Molecular cell*, 49(5):773–782.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The hitchhikers guide to hi-c analysis: practical guidelines. *Methods*, 72:65–75.

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., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293.

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H. D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J., et al. (2014). Cohesin-dependent globules and heterochromatin shape 3d genome architecture in s. pombe. *Nature*, 516(7531):432–435.

Nikumbh, S. and Pfeifer, N. (2017). Genetic sequence-based prediction of long-range chromatin interactions suggests a potential role of short tandem repeat sequences in genome organization. *BMC bioinformatics*, 18(1):218.

Quang, D. and Xie, X. (2016). Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11):e107–e107.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120.

Singh, S., Yang, Y., Poczos, B., and Ma, J. (2016). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *bioRxiv*, page 085241.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome research*, 24(3):390–400.

Van Bortle, K., Nichols, M. H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z. S., and Corces, V. G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome biology*, 15(5):R82.