DNA folding features prediction using epigenetic data with Recurrent Neural Networks

Michal Rozenwald [1] **Ekaterina Khrameeva** [2] **Aleksandra Galitsyna** [2] **Grigory Sapunov** [1 3] **Mikhail S. Gelfand** [2 4]

Abstract

Technological advances enabled the collection of large epigenetic datasets, including information about various DNA binding proteins and DNA spatial structure. Hi-C experiments have revealed that chromosomes are subdivided into sets of self-interacting domains called Topologically Associating Domains (TADs). TADs are involved in the regulation of gene expression activity, but the mechanisms of their formation are not yet fully understood. In mammals, the genome is folded through the concerted function of architectural proteins cohesin and CTCF. Drosophila, by contrast, lacks CTCF-mediated domain formation. Instead, *Drosophila* TADs appear to be correlated with epigenetic features, such as histone modifications. Here, we focus on Machine Learning methods to characterize DNA folding patterns in *Drosophila* across three cell lines. We present Linear Regression models with four types of regularization, Gradient Boosting, and Recurrent Neural Networks (RNN) to learn chromatin folding characteristics associated with TADs using epigenetic ChIP-Seq data. Our Bidirectional Long Short-Term Memory RNN architecture gained the best prediction scores and has highlighted biologically relevant features. Chriz and H3K4me3 were selected as the most informative features for the prediction of TADs characteristics. The implemented pipeline is called Hi-ChIP-ML, and the code is publicly available. This approach may be adapted to any similar biological dataset of chromatin features across various cell lines and species. *Code:* https://github.com/MichalRozenwald/Hi-ChIP-ML

[1]National Research University Higher School of Economics, 101000 Moscow, Russia [2]Skolkovo Institute of Science and Technology, 143026 Skolkovo, Russia [3]Intento, Inc, Berkeley, CA 94704 [4]Institute for Information Transmission Problems (Kharkevich Institute), RAS, 127051 Moscow, Russia. Correspondence to: Michal Rozenwald <michal.rozenwald@gmail.com>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).

1. Introduction

Investigating the DNA-protein complex of eukaryotic cells called *chromatin* is a challenging task today. Multiple interconnections of chromatin structure with gene regulation, inheritance, and disease have been observed (Lupiáñez et al., 2016). Various cell regulation mechanisms act through the three-dimensional (3D) structure of chromatin. High-throughput experiments capturing contacting fragments of the genome, such as Hi-C, have unraveled many principles of chromosomal folding (Lieberman-Aiden et al., 2009). Hi-C maps have demonstrated that chromosomes are subdivided into sets of self-interacting domains called *Topologically Associating Domains (TADs)* (Ulianov et al., 2016). TADs influence regulatory landscapes within chromosomes at multiple scales and organisms (Szabo et al., 2019). For example, TADs correlate with units of replication timing regulation in mammals (Pope et al., 2014) and with epigenetic domains in *Drosophila* (Sexton et al., 2012). Moreover, Hi-C maps have the potential for various practical and medical applications. For instance, disruption of the chromosomal topology has been reported to affect gliomagenesis and limb malformations in humans (Krijger & De Laat, 2016).

Several studies focused on predicting 3D chromatin architecture using Machine Learning. For example, Cristescu et al. have presented the REcurrent Autoencoders for CHromatin 3D structure prediction (REACH-3D) which reconstructs the chromatin structure and creates an embedding representation (Cristescu et al., 2018). Shashank et al. have constructed a Deep Learning model called SPEID that predicts enhancer-promoter interactions using only sequence-based features (Singh et al., 2019). As partitioning of the genome into TADs is still not fully understood, it provides a unique opportunity to build accurate and interpretable Machine Learning models.

The relationship between TADs and epigenetic marks has been previously investigated using high-throughput ChIP-seq data that localize protein binding sites and histone modifications to DNA *in vivo* (Bushey et al., 2009). Ulianov et al. demonstrated that in *Drosophila*, active transcription plays a key role in the chromosome partitioning into TADs (Ulianov et al., 2016). The authors suggested that nucleosomes tending to interact less often influence the

formation of inter-TADs and TAD boundaries. Active chromatin marks are preferably present at TAD borders, while repressive histone modifications are depleted in inter-TADs, which reveals the correlation between TADs and chromatin marks.

Our research focuses on understanding the 3D chromatin structure using epigenetic data and Machine Learning techniques. To that end, we analyzed *Drosophila melanogaster* chromatin structure via Linear Regression models, Gradient Boosting Regressors, and Recurrent Neural Networks. The models were trained to predict TAD patterns using ChIP-seq data, and informative epigenetic marks are presented bellow.

2. Methods

Input Data Hi-C datasets for three cultured *Drosophila melanogaster* were collected from Ulyanov et al. (Ulianov et al., 2016). Cell lines Schneider-2 (S2) and Kc167 from late embryos and DmBG3-c2 (BG3) from the central nervous system of third-instar larvae were analysed. *Drosophila* dm3 genome assembly was divided into 5950 sequential genomic regions called *bins*, where each bin corresponded to 20 000 DNA base pairs. Each bin was described by the density of epigenetic features, estimated by ChIP-seq, downloaded from the modENCODE database (Celniker et al., 2009). We selected ChIP-seq features corresponding to epigenetic marks, i.e., transcription factors, insulator protein binding sites, and histone modifications (Chriz, CTCF, Su(Hw), H3K27me3, H3K27ac and additionally RNA-polymerase-II, BEAF-32, GAF, CP190, H3K4me1, H3K4me2, H3K4me3, H3K9me2, H3K9me3, H3K27me1, H3K36me1, H3K36me3, H4K16ac), which had been reported as relevant for TAD formation in previous studies(Chepelev et al., 2012; Wang et al., 2014).

Target Value Topologically Associating Domains annotation software Armatus (Filippova et al., 2014) has a scaling parameter *gamma*, which determines the average size of TADs. When gamma is fixed, each genomic bin is annotated as a part of a TAD, inter-TAD, or TAD boundary. Higher gamma corresponds to smaller TADs sizes on average. We characterized each bin by the scaling parameter called transitional gamma at which this bin switches from being a part of a TAD to being a part of an inter-TAD or a TAD boundary. An illustration of TADs annotation is shown in Figure 1.

Loss Function The target is a continuous variable ranging from 0 to 10 with an unbalanced distribution. Most of the values lie in the interval between 0 and 3. Moreover, the biological nature of bins with maximal transitional gamma is different from the other, since transitional gamma 10 means that the bin never transforms from being a part of a

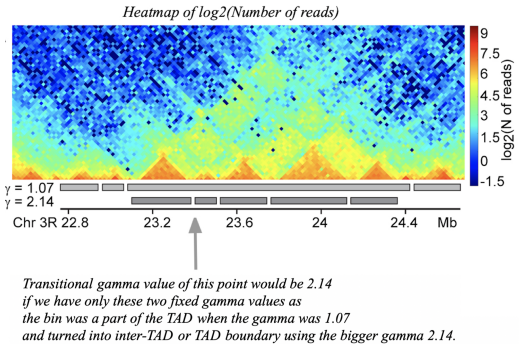

Figure 1. Annotation of TADs at different gamma parameter values is on the left side. The histogram of the target value transitional gamma is presented in the right part of this plot.

TAD to an inter-TAD or TAD boundary.

Hence we have introduced a custom loss function named modified *weighted Mean Square Error (wMSE)*:

$$wMSE = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{true}_i} - y_{\text{pred}_i})^2 \frac{\alpha - y_{\text{true}_i}}{\alpha},$$

where n is the number of data points,
y_{true_i} is the true value for data point number *i*,
y_{pred_i} is the predicted value for data point number *i*,
α is the maximum value of y_{true} increased by 1.

Models To explore the relationships between the 3D chromatin structure and epigenetic data, we built Linear Regression (LR) models, Gradient Boosting (GB) Regressors, and Recurrent Neural Networks (RNN). The Linear models, Gradient Boosting, and Constant Predictions using the mean value of the training dataset were implemented as benchmarks as no other ML pipelines for these datasets have been publicly available yet.

Due to the linear connectivity of our input, bins are sequentially ordered in the genome, and the target variable values are expected to be highly correlated. In addition, the information content of the double-stranded DNA molecule is equivalent if reading in the forward and reverse directions. To utilize the DNA linearity together with independence on the direction, we selected the Bidirectional Long Short-Term Memory (LSTM) Recurrent Neural Networks architecture (Schuster & Paliwal, 1997). Our model takes as input a sequence of bins and outputs the target value of the *middle bin*. The middle bin is an object from the input set with an index *i*, where *i* equals to the floor division of the input set length by 2. Thus the model uses the characteristics of the surrounding bins while predicting the transitional gamma of the middle bin.

To explore the importance of each feature from the input space, we trained the RNNs using only one of the ChIP-seq

features as input. We have also trained models in which columns from the feature matrix were one by one knocked down. Further, we calculated the evaluation metrics and checked if they were significantly different from the results obtained while using the complete set of data. The dataset was randomly split into three groups: train dataset 70%, test dataset 20%, and 10% data for validation.

3. Results

The mean values of the weighted Mean Square Errors for cross-validation of ten experiments for all models and cell lines are presented in Table 1. The best Linear Regression score using L1 and L2 regularization was obtained with alpha equal to 0.2. Gradient Boosting Trees were trained with variable parameters such as the number of estimators, learning rate, maximum depth of the individual regression estimators. The parameters of the best Gradient Boosting Trees were 'n_estimators': 250, 'max_depth': 4 'learning_rate': 0.01.

Experiments on the datasets of five ChIP-seq characteristics have shown stable Linear Regression weight coefficients. We have obtained features prioritization where the most valuable feature was Chriz, while the weights of Su(Hw) and CTCF were the smallest. Chriz, H3K4me1, and H3K27me1 were the most robust influential factors using all the ChIP-seq features together. In addition, Chriz was selected as the only most influencing feature by the Linear Regression model with L1 regularization across datasets.

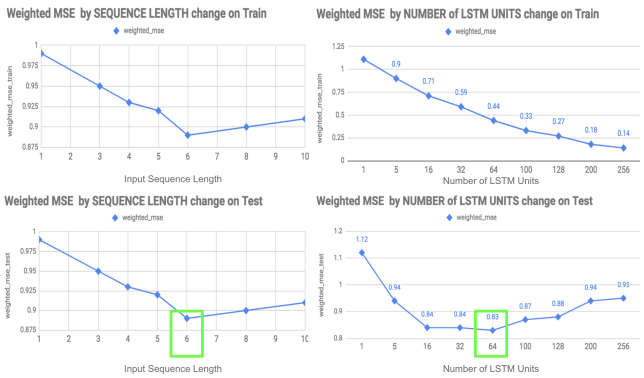

Figure 2. Weighted Mean Square Error of the Bidirectional LSTM. The upper row of graphs outlines the results for the S2 train dataset. The bottom row shows wMSE counted on the S2 test objects. The left half shows the results of training RNN with 64 units for different sizes of sequence length. The right half shows wMSE for training RNNs with an input sequence of 6 bins for a different number of LSTM Units. The green box highlights the best scores.

To explore the weighted Mean Square Error on various input sequence length, we trained Bidirectional RNN models with different input window sizes and numbers of LSTM Units. The result is shown in Figure 2, where the optimal sequence length is equal to the input window size 6 and 64 LSTM Units. This result has a clear biological interpretation as the typical size of TADs in *Drosophila melanogaster* is known to be around 120 Kb, which corresponds to 6 bins of 20,000, providing the best prediction scores.

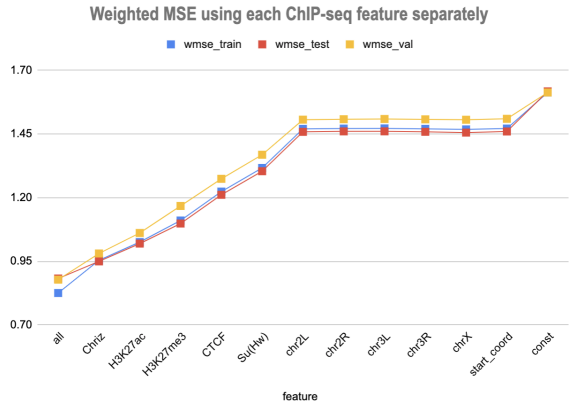

Figure 3. Weighted MSE using one ChIP-Seq feature for each bin of S2 c.l. in the biLSTM RNN. The first mark (*'all'*) corresponds to scores of NNs using the features together, the last mark (*'const'*) represents wMSE using constant prediction.

The Bidirectional Long Short-Term Memory Recurrent Neural Networks with 64 LSTM Units and sequences of 6 bins taken as input data scored better than all other models. The BiLSTM Recurrent Neural Networks that we explored were able to capture and utilize the sequential relationship of the input objects in terms of the physical distance in the DNA.

4. Discussion

The ejection of the chromodomain protein called Chriz (Eggert et al., 2004) strongly influenced the prediction scores. Similarly, the RNNs that used Chriz as input produced better wMSE scores (Figures 3 and 4), and all Linear models assigned the highest regression weight to the Chriz input signal. This protein is known to be specific for the interbands of *Drosophila melanogaster* chromosomes (Chepelev et al., 2012). Ulianov et al. demonstrated strong enrichment of Chriz at TAD boundaries and inter-TAD regions (Ulianov et al., 2016). Additionally, in (Hug et al., 2017; Ramírez et al., 2018; Sexton et al., 2012), the insulator proteins Chriz and BEAF-32 were enriched at boundaries. This explains why the third rank of the predictability score was achieved by BEAF-32.

The application of the Recurrent Neural Networks using each of the described ChIP-seq features separately has revealed a strong predictive power of active histone modifications such as dimethylation of histone H3 lysine 4 (H3K4me2). This result aligns with the fact that H3K4me2 was previously shown to strongly define the transcription

Table 1. Weighted MSE on cross-validation of all methods for each cell line and while using them together.

METHOD	SCHNEIDER-2	KC167	DMBG3-C2	ALL
CONSTANT PREDICTION	1.62 ± 0.09	1.53 ± 0.06	1.36 ± 0.05	1.51 ± 0.04
LINEAR REGRESSION	1.14 ± 0.08	1.01 ± 0.06	0.91 ± 0.08	1.04 ± 0.04
LINEAR REGRESSION + L1	1.12 ± 0.07	1.04 ± 0.06	0.95 ± 0.07	1.05 ± 0.04
LINEAR REGRESSION + L2	1.12 ± 0.07	1.01 ± 0.06	0.9 ± 0.08	1.03 ± 0.04
LINEAR REGRESSION + L1 + L2	1.11 ± 0.07	1.02 ± 0.06	0.91 ± 0.07	1.03 ± 0.04
GRADIENT BOOSTING	1.07 ± 0.06	0.98 ± 0.07	0.86 ± 0.08	0.96 ± 0.04
BILSTM 64 UNITS & 6 BINS	0.86 ± 0.04	0.83 ± 0.04	0.73 ± 0.01	0.78 ± 0.01

factor binding regions in different cells (Wang et al., 2014). Histone modifications H3K4me3, H3K27ac, H3K4me1, H3K4me3, H4K16ac, and other active chromatin marks are also enriched in inter-TADs and at TAD boundaries. In addition, HMs have been used to distinguish various genomic regions. For instance, H3K27ac, and H3K4me1 were selected to identify if an enhancer is poised or active (Barski et al., 2007; Creyghton et al., 2010; Rada-Iglesias et al., 2011).

As for the models' performance using Su(Hw) and CTCF, the outcome is in line with published research (Ulianov et al., 2016). Ulianov et al. found that for prediction of TAD boundaries, the binding of insulator proteins Su(Hw) and CTCF performed much worse than the active chromatin marks. In *Drosophila*, the absence of strong enrichment of CTCF at TAD boundaries and preferential location of Su(Hw) in TADs implies that CTCF- and Su(Hw)-dependent insulators are not the major determinants of TAD boundaries. In agreement with the previous knowledge, our results demonstrate that the impact of Su(Hw) and CTCF is low for both insulator proteins.

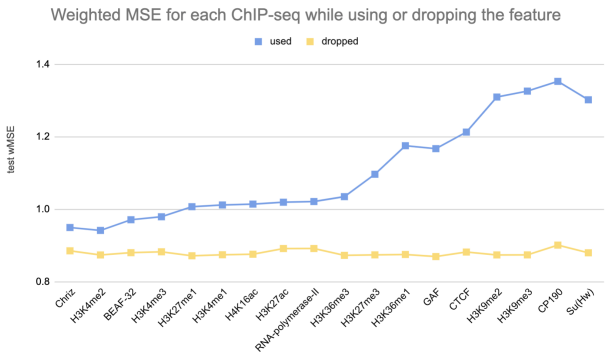

Figure 4. Weighted MSE on the S2 test dataset using each ChIP-Seq either as a single feature (blue line) or ejecting it from the biLSTM RNN input (yellow line).

5. Conclusion

In this work, Recurrent Neural Networks, Gradient Boosting Trees, and Linear Regression models were applied for the prediction of chromatin folding patterns using epigenetic data in *Drosophila melanogaster* across three cell lines. Explicit accounting for the linearly ordered bins in the DNA molecule improved the prediction significantly, as the best results were obtained by the Bidirectional LSTM RNN. Furthermore, the optimal length of the input sequence was equal to six, which is biologically meaningful as it corresponds to the average TAD size in *Drosophila*.

Feature importance analysis of the input ChIP-seq data was conducted and provided prioritization that aligned with the existing chromatin research. All models selected Chriz signal as one of the most influencing and histone modification H3K4me2 was shown to increase the wMSE score of the prediction strongly.

Exploration of models transferability between a broad set of cell types might be an interesting development direction for this research, as well as the integration of various biological features, such as raw DNA sequence, to the presented models.

The implemented pipeline called Hi-ChIP-ML is open-source. The methods can be used to explore the 3D chromatin structure of various species and may be easily adapted to any similar biological dataset. The code is freely available at:
https://github.com/MichalRozenwald/Hi-ChIP-ML

Acknowledgments

This study was supported by the Russian Science Foundation, grant number 19-74-00112, and Skoltech Fellowship in Systems Biology for A.Galitsyna.

The HiC data for the *Drosophila melanogaster* cell lines were generously provided by the Razin Lab (Institute of Gene Biology RAS, Moscow, Russia).

*A paper describing this work was accepted to *The International Symposium on Bioinformatics Research and Applications (ISBRA)* rescheduled to December 1-4, 2020 due to COVID-19.

References

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

Bushey, A. M., Ramos, E., and Corces, V. G. Three subclasses of a drosophila insulator show distinct and cell type-specific genomic distributions. *Genes & development*, 23(11):1338–1350, 2009.

Celniker, S. E., Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Research*, 22(3):490–503, 2012.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010.

Cristescu, B.-C., Borsos, Z., Lygeros, J., Martínez, M. R., and Rapsomaniki, M. A. Inference of the three-dimensional chromatin structure and its temporal behavior. *arXiv:1811.09619*, 2018.

Eggert, H., Gortchakov, A., and Saumweber, H. Identification of the drosophila interband-specific protein z4 as a dna-binding zinc-finger protein determining chromosomal structure. *Journal of Cell Science*, 117(18): 4253–4264, 2004.

Filippova, D., Patro, R., Duggal, G., and Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, 9(1):14, 2014.

Hug, C. B., Grimaldi, A. G., Kruse, K., and Vaquerizas, J. M. Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell*, 169 (2):216–228, 2017.

Krijger, P. H. L. and De Laat, W. Regulation of disease-associated gene expression in the 3d genome. *Nature Reviews Molecular Cell Biology*, 17(12):771–782, 2016.

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

Lupiáñez, D. G., Spielmann, M., and Mundlos, S. Breaking tads: how alterations of chromatin domains result in disease. *Trends in Genetics*, 32(4):225–237, 2016.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, 2014.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283, 2011.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. High-resolution tads reveal dna sequences underlying genome organization in flies. *Nature Communications*, 9(1):1–15, 2018.

Schuster, M. and Paliwal, K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45 (11):2673–2681, 1997. doi: 10.1109/78.650093.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472, 2012.

Singh, S., Yang, Y., Poczos, B., and Ma, J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology*, 7(2): 122–137, 2019.

Szabo, Q., Bantignies, F., and Cavalli, G. Principles of genome folding into topologically associating domains. *Science Advances*, 5(4):eaaw1668, 2019.

Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., Penin, A. A., Logacheva, M. D., Imakaev, M. V., Chertovich, A., et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research*, 26(1):70–84, 2016.

Wang, Y., Li, X., and Hu, H. H3k4me2 reliably defines transcription factor binding regions in different cells. *Genomics*, 103(2-3):222–228, 2014.