Utilizing Machine Learning to Identify Time Asymmetry of DNA Loop Extrusion

Anna Du

January 2024

Abstract

DNA loop extrusion, mediated by cohesin protein complexes, plays a central role in genome organization. However, direct observation of loop extrusion in vivo remains challenging. This study investigates a novel methodology using time reversal asymmetry and machine learning to detect loop extrusion in microscopy data. I aim to do this by analyzing DNA motion in microscopy data, hypothesizing that movies of DNA under loop extrusion appear differently when played forward versus backward. Simulations with and without loop extrusion generate a synthetic dataset to test this hypothesis and determine the feasibility of detection. A Convolutional Neural Network (CNN) is employed to process these DNA motion movies, trained through supervised learning to distinguish between normal and reversed trajectories. The CNN's performance, measured by its accuracy in identifying reversed motion, serves as an indicator of loop extrusion presence in the DNA. The test CNN used here achieved an accuracy consistent with random guessing on simulated data with loop extrusion, suggesting great difficulty in the prediction task. I propose further optimizations such as increasing the frame rate, change in network architecture, and extrusion parameters which may make the task easier. With additional optimization, this approach may enable time reversal and machine learning to analyze the presence of loop extrusion.

Contents

1 Introduction **3**

2 Methods **4**

2.1 Generation of an Equilibrated System with No Loop Extrusion 4

2.2 Polymer Simulations with Loop Extrusion 4

2.3 1D Simulation 5

2.4 3D Simulation 6

2.5 Obtaining CTCF Location Coordinates 6

2.6 Convolutional Neural Networks 7

3 Results and Discussion **8**

4 Acknowledgments **9**

5 References **9**

1 Introduction

DNA loop extrusion is the process by which protein complexes reel in DNA to extrude a loop. By appropriate placement of boundaries, the loops may be directed to obtain locally compared regions of DNA. This is a fundamental process that contributes to the spatial organization of the genome and plays critical roles in gene regulation, somatic recombination, and DNA repair.

The loop extruder cohesin and its interaction with the boundary protein CTCF have been found to lead to the establishment and regulation of compacted regions of DNA known as topologically associated domains (TADs). These domains have been implicated in the control of gene expression by contacting distal enhancer DNA sequences with their target gene promoters, bringing them into close proximity. This spatial proximity could allow for efficient and specific regulation of gene expression (Popay, Dixon, 2022).

Various studies have demonstrated the importance of TADs and loop formation in the development and function of biological organisms. A recent study investigated the impact of altering TAD structure on limb development in mice. They found that deletions, inversions, duplications, and rewiring of TAD boundaries led to changes in gene expression patterns and resulted in limb malformations (Lupianez et al., 2015).

Furthermore, the loop extrusion process has been implicated in somatic recombination and DNA repair. Somatic recombination involves the rearrangement of DNA segments in immune cells to generate diverse antigen receptor genes. Loop extrusion has been implicated in the facilitation of the spatial proximity between antigen receptor gene segments, enabling efficient recombination events that contribute to immune system diversity. Similarly, during DNA repair processes, damaged DNA segments and their repair machinery can be brought into close proximity through loop extrusion, facilitating the repair of DNA lesions (Gabriele et al., 2022).

DNA loop extrusion mediated by cohesin protein complexes is a central organizing principle of our genomes. As such, understanding the mechanisms and functional consequences of loop extrusion provides insights into the fundamental principles governing genome organization and their impact on cellular processes. Further research in this field holds promise for unraveling the intricate relationship between chromatin architecture and genome function. However, the process of loop extrusion has not been observed in vivo and has only been directly observed outside living systems in single-molecule reconstitution assays. Indirect evidence of loop extrusion in vivo has been obtained through the analysis of Hi-C data, which provides information about the genome's spatial organization. Hi-C data scaling analysis has revealed the presence of TADs and suggested the involvement of loop extrusion in their formation.

The fundamental goal behind this study is to study a novel method for the detection of loop extrusion in living systems. The idea is to use the fact that movies of DNA under the action of loop extrusion would look different when played forward in time compared to backward. Microscopy data of DNA otion could therefore be used to detect the presence of loop extrusion based on our ability to discern reversed movies from regular movies. To explore the feasibility of this approach, it is necessary to perform simulations both with and without loop extrusion to generate a synthetic dataset. By conducting these simulations, I can assess whether detectability is possible, and if possible, which experimental conditions would be needed to see loop extrusion in living systems.

The methodology rlies on a way to quantify our ability to separate reversed from non-reversed movies of DNA motion. Due to the complexity of the motility, and the need for efficient deployment across large datasets, a convolutional neural network (CNN), which is a type of artificial neural network designed for processing and analyzing multi-dimensional data, including images and videos, is employed. CNNs are trained through a process called supervised learning, where they are presented with labeled training data (here data would be a time-series of DNA motion and the label would be whether the time series was played forward or backward in time) and learn to associate input data with their corresponding output labels. By adjusting the weights and biases through the iterative training process, the CNN gradually learns to recognize and extract relevant features from the input data, enabling it to make accurate predictions or classifications on new, unseen data. The ability of the trained CNN to identify whether it "sees" a difference between flipped and unflipped trajectories based on the time series can be quantified by the prediction accuracy which gives us a measure of the presence of loop extrusion in the trajectories.

2 Methods

2.1 Generation of an Equilibrated System with No Loop Extrusion

Simulations of a polymer were run with 750 base pairs per bead, and locus sizes of 1,515,000 base pairs. Forces such as the bond angle stiffness, radius for contact, bond length, and bond force were applied to the polymer, and run for 1,000 time steps, with the goal of running the polymer simulation to equilibrium without loop extrusion acting upon the system. This data obtained from the equilibrated system with no loop extrusion was used as the basis to start running the simulation with loop extrusion.

2.2 Polymer Simulations with Loop Extrusion

The simulations with loop extrusion maintained the same format of repeating 2,020 beads 50 times, with each repeat containing two CTCF sites, resulting in a total of 100 CTCF locations across the polymer. To conduct these simulations with cohesins loaded to form loop extrusion required a multi-step process in which first (i) generated a 1D simulation to obtain coordinates for cohesin positions, and using that data, (ii) created a 3D simulation, applying all forces necessary to satisfy the cross links implied by the cohesin positions. Finally, we (iii) regularly collected coordinate information regarding CTCF locations in space.

2.3 1D Simulation

The 1D simulation generates information regarding loop growth on the polymer. An array that represents the polymer is generated, with loop extruders initialized in randomized locations on the array. The simulation takes into account the location and direction of the CTCF sites. If the direction of the loop does not align with the CTCF site, the loop will continue to grow (Figure 1a). If two loops intercept at adjacent array values, growth will halt in the direction that is in contact with the other loop, whereas it will continue in the other direction, representing that the loops have joined (Figure 1b). When the loop is in contact with a CTCF site in the direction that stops further growth of the loop, growth of the loop will halt in that direction for future time-steps, while continuing in the other direction (Figure 1c.).

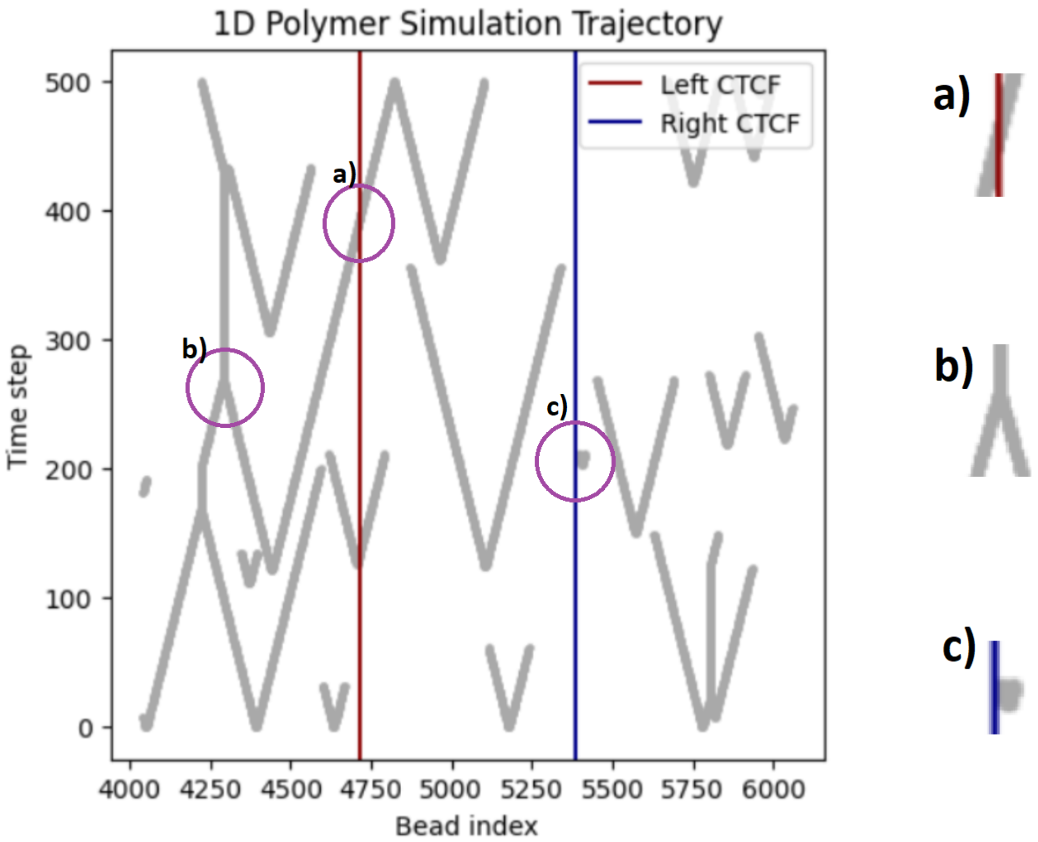

Figure 1: *A visualization of the 1D trajectory of the loop formation and growth across 500 time steps and 2020 beads. The vertical lines represent the locations of the CTCF on the polymer strand. The part at a) shows the unidirectionality of the CTCF protein, thus the growth of a loop from the opposite direction is not stopped by the protein. b) is an example of a collision between two separate loops. c) represents a loop where the position of the CTCF has halted the growth.* The red line represents a right-facing CTCF site, and the blue line, a left-facing CTCF site.

2.4 3D Simulation

The 3D simulation loads the equilibrated system obtained in the simulation with no loop extrusion, and builds upon the results obtained from the 1D simulation. The 3D simulation adds the various forces that act upon the polymer that mimics the physical interactions within DNA, allowing it to undergo conformational changes and dynamics that reflect its behavior in a real biological system. The output of this simulation is a series of files, for each time step, containing coordinate information for every bead.

This graph compares the contact probability to the step separation of the polymer simulation that contains loop extrusion. The bump at around 15 units for the step separation is caused by the bond angle stiffness set in the simulation, which was modulated to ensure the random walk simulation can better emulate the properties of a polymer in a living cell (Figure 2a). The convergence of the curves at approximately 1500 units demonstrates that the system is approaching equilibrium (Figure 2b). The various data points represent different blocks.

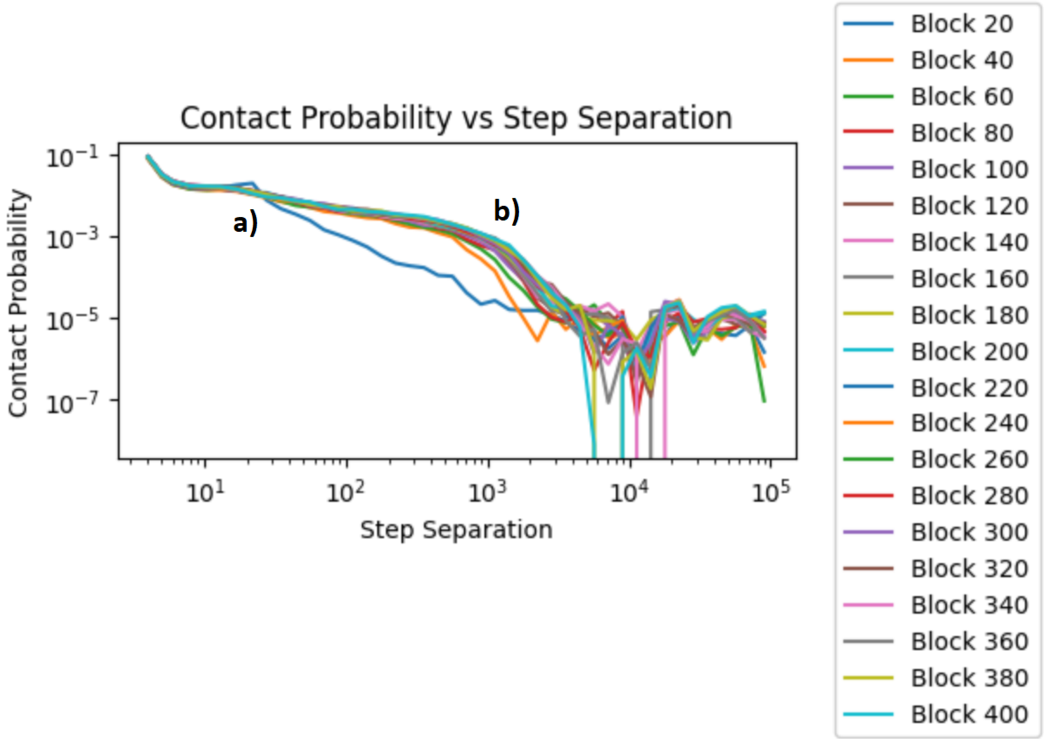

Figure 2: *Contact probability vs step separation*

2.5 Obtaining CTCF Location Coordinates

The 3D simulation program focuses on obtaining CTCF location coordinates and generating trajectory information. The positions of CTCF sites within the polymer are tracked more frequently than the overall coordinates for the entire polymer. To organize and store the collected data, the 3D simulation program writes the trajectory information to a CSV file. This increased frequency of data collection captures the precise movements and interactions of CTCF sites with the polymer but reduces storage requirement as the whole polymer configuration isn't required with such high time-resolution.

Prior to using the data for machine learning, a 'hump' behavior was noted in the probability versus separation curve of chromatin loops (Figure 2b). Analysis of the distance versus time graphs across all three axes revealed converging distances towards zero. This indicates that the cohesin movement stopped at the CTCF boundaries. This convergence confirms that enhancer-promoter distances displayed signatures of binding together.

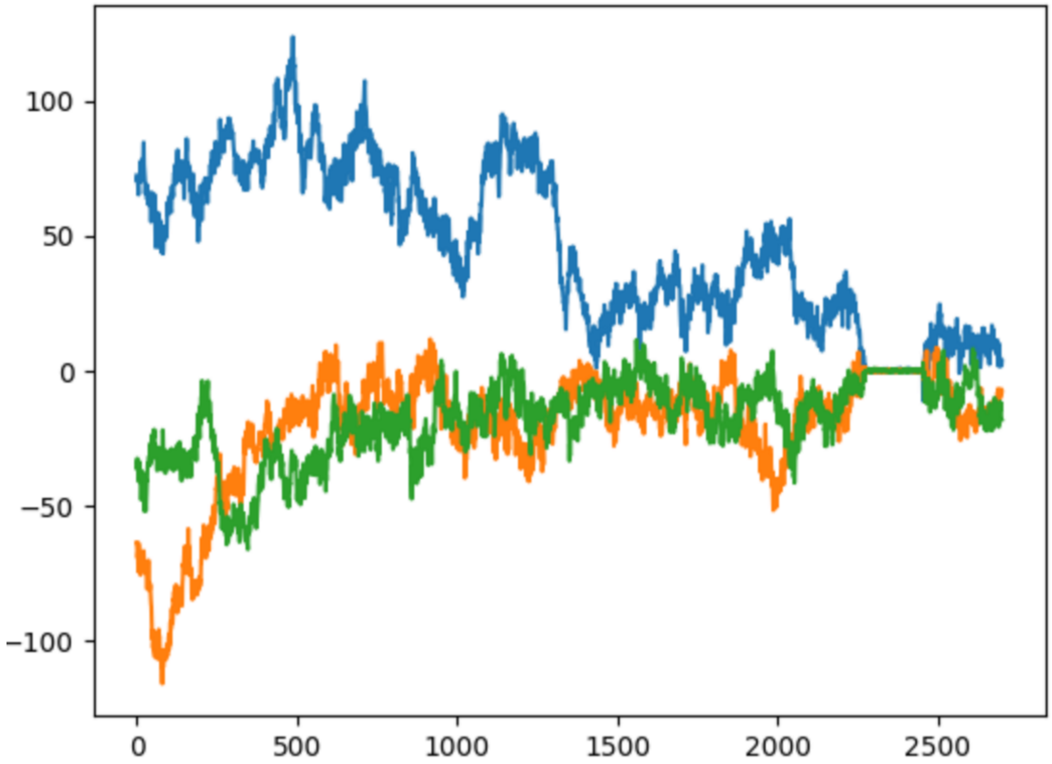

Figure 3: *X, Y, and Z distances between the two trajectories vs time step*

2.6 Convolutional Neural Networks

The subsequent phase in the project involves the use of time reversal as a method to ascertain whether the trajectories of DNA are impacted by loop extrusion. The movement of a DNA polymer without loop extrusion is passive and isotropic, and thus, the relative motion between any two positions on the DNA would be independent of whether it is observed forward or backward in time. This is not the case for simulations that include loop extrusion, as the process of zipping two loci together generally causes the distance to decrease over time (although this effect can be subtle). The goal would be to observe how well the trained CNN model is able to predict and distinguish the flipped and non-flipped trajectories, which is only possible if there is loop extrusion in the simulation.

Both forward and reverse temporal trajectories of DNA motion are utilized to train the CNN. CNNs utilize supervised learning, where the network is exposed to labeled training data – in this case, a time-series dataset of DNA trajectories is labeled based on whether it is run forward or backward in time, enabling the CNN to associate input data with corresponding output labels. CNN's ability to discern differences between flipped (reversed) and unflipped (forward) trajectories based on the time series data provides a quantitative measure of loop extrusion presence in the observed trajectories.

3 Results and Discussion

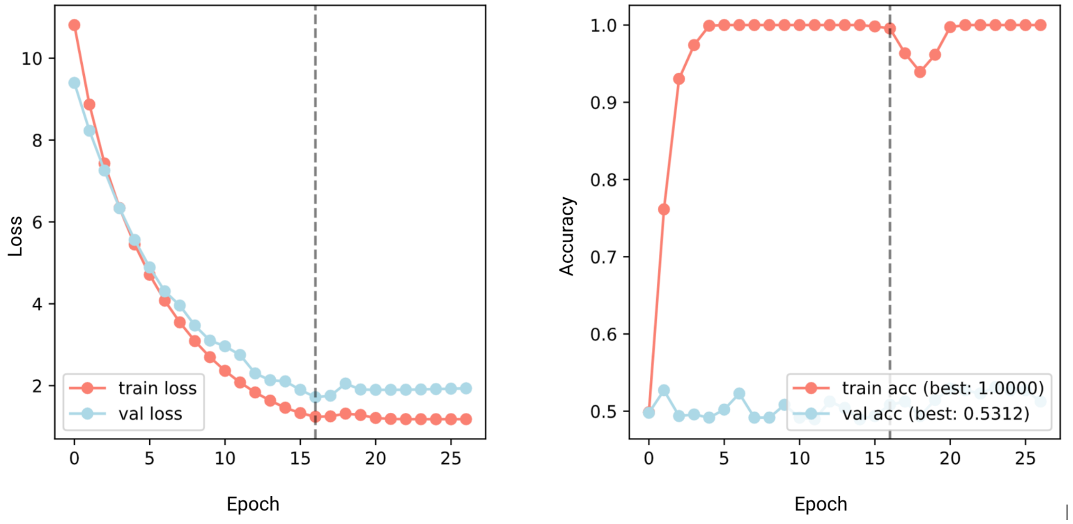

Figure 3: *Loss (left) and accuracy (right) curves as a function of epoch from the convolutional neural network. The dashed line represents the epoch that resulted in the lowest loss, to prevent over-fitting.*

The primary objective of using the Convolutional Neural Network (CNN) model in this study was to determine whether it is possible to accurately distinguish between forward and reverse chromatin loops. This low loss rate reflects the model's ability to closely align its predictions with the actual data, demonstrating a high level of accuracy and reliability in its output. Such an outcome signifies the effectiveness of the model's learning algorithm and demonstrates potential applicability in real-world scenarios. However, the accuracy results indicate a limitation in this capability. With an accuracy of only 53%, the model does not significantly exceed the performance of random guessing (which would be 50% in a binary classification task like this). This outcome suggests that the current CNN model is not effectively differentiating between the two types of loops.

This study's findings, particularly the Convolutional Neural Network (CNN) model's 53% accuracy rate in differentiating forward and reverse trajectories, suggest two possibilities:

either an absence of observable asymmetry in the process or a limitation in the model's capability to detect such asymmetry. Given the inherent asymmetry in the loop extrusion process, it's plausible that this can be attributed to the slow extrusion speed of the simulation, which could potentially be hard to distinguish from the noise.

The primary goal moving forward is to enhance the model's accuracy, ensuring more reliable and definitive distinctions between loop types. The current model may require refinement. This could involve utilizing more layers or different types of layers, or exploring alternative CNN architectures. The field of deep learning is rapidly evolving, with new algorithms and approaches being developed continually. Investigating these new algorithms could provide a fresh perspective on the problem. A faster frame rate in the data collection process might also reduce the noise-to-signal ratio, offering clearer insights into the dynamic behavior of chromatin loops. By refining the model, future studies can aim to significantly increase the accuracy of chromatin loop classification. This could potentially advance our understanding of chromatin dynamics and contribute to the broader field of genomic research.

4 Acknowledgments

I would like to thank the MIT PRIMES program for this amazing research opportunity, as well as my mentors, Mr. Henrik Pinholt and Dr. Aleksandra Galitsyna, for their invaluable expertise and encouragement.

5 References

1. Gabriele, M., Brandão, H. B., Grosse-Holz, S., Jha, A., Dailey, G. M., Cat toglio, C., Hsieh, T.-H. S., Mirny, L., Zechner, C., & Hansen, A. S. (2022). *Dynamics of CTCF and cohesin mediated chromatin looping revealed by live-cell imaging*. Science. http://dx.doi.org/10.1101/2021.12.12.472242

2. Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., ... Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, *161*(5), 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

3. Popay, T. M., & Dixon, J. R. (2022). Coming full circle: On the origin and evolution of the looping model for enhancer–promoter communication. *Journal of Biological Chemistry*, *298*(8), 102117. https://doi.org/10.1016/j.jbc.2022.102117

4. Sabaté, T., Lelandais, B., Bertrand, E., & Zimmer, C. (2023). Polymer simulations guide the detection and quantification of chromatin loop extrusion by imaging. *Nucleic Acids Research*, *51*(6), 2614–2632. https://doi.org/10.1093/nar/gkad034