Identifying TADs using Gaussian HMM Model training

Authors: David Zuravin, Noam Delbari, Amit Halbreich, Yaniv Pasternak, Omer Mushlion, Elisheva Morgenstern

Abstract:

Topologically Associating Domains (TADs) represent a critical aspect of the three-dimensional organization of chromosomes within the nucleus, influencing gene expression and cellular function. Understanding TADs is paramount for insights into the mechanisms of genetic regulation, the maintenance of cellular identity, and the development of disease.In addition recent advance in sequencing technologies have dramatically enhanced our capacity to map the three-dimensional genome architecture, reducing both time and cost while significantly increasing data resolution. Here we present an attempt to identified and characterize TADs and their boundaries by developing a Gaussian Hidden Markov Model.

Intrudcution:

Recent years have seen significant progress in our understanding of how animal genomes are spatial organized, thanks in part to the advancement of chromosome conformation capture[1] (3C) methods and improvements in live cell imaging techniques. These methods have greatly enhanced our understating of the genome's structural layout. The 3C process involves cutting crosslinked chromatin with restriction enzymes, then joining these pieces in a way that reflects their spatial proximity within the nucleus, creating chimeric molecules that signify interactions between genomic regions. [2]Hi-C, a technique that applies the principles of 3C across the entire genome, has increasingly refined our ability to observe the genome's three-dimensional folding at various levels of organization and enabling the quantitative analysis of physical contacts between genomic regions across the entire genome[3]. These advancements led to the discovery of TADs, highlighting the hierarchical organization of the genome and underscoring the spatial context of gene regulation.

TADs are defined as contiguous genomic regions where DNA sequences interact more frequently with each other than with sequences outside the domain, mediated by proteins like CTCF and cohesin[4]. This organization facilitates the proximity of genes and their regulatory elements, such as enhancers, enabling precise control over gene expression crucial for development and differentiation processes.

Identifying the boundaries of TADs is vital for decoding the genome's regulatory architecture. These boundaries often contain insulator[5] elements that block the spread of heterochromatin and restrict enhancer-promoter interactions within TADs, maintaining specific gene expression patterns. Alterations in TAD structures or their boundaries are associated with various diseases, highlighting the importance of understanding TAD [6].

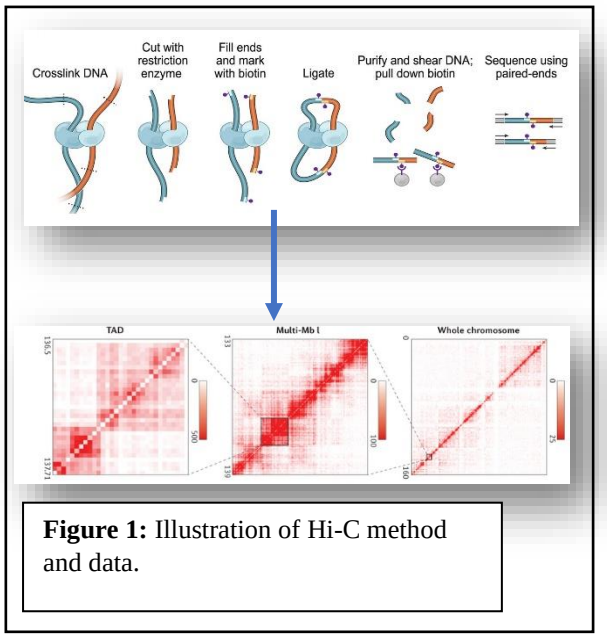


Figure 1: Illustration of Hi-C method and data.

Building upon the foundational understanding of the three-dimensional genome provided by 3C-based methods and Hi-C data,we need to identified Tad from the hi-c data. Hi-C data provides a rich, detailed map of how different parts of the genome interact within the 3D space of the nucleus, revealing the complex network of chromosomal contacts. However, interpreting this massive amount of data to understand the underlying structure of the genome, such as identifying the boundaries of TADs, requires sophisticated analysis methods. We try to do that by using Gaussian Hidden Markov Model (HMM). HMM are particularly adept at recognizing patterns and predicting states within a sequence. By treating the genome as a sequence of interacting regions, HMMs can effectively distinguish between the high-frequency interactions within TADs and the lower-frequency interactions between different TADs, thus identifying the locations of TAD boundaries with a high degree of accuracy.

Methods

Scoring Methods

We employed several scoring methods to quantify the interaction between region such one of them will be the input for our model. At the end we choose to use the DI score to feed our HMM model as the input.

Directionality Index (DI) Score: This metric quantifies the directionality of chromatin interactions to delineate boundaries. Positive DI score for bin indicates strong interactions with previous bins (vice versa for Negative DI score). The values of DI score were calculated for each genomic bin by comparing the number of upstream and downstream interactions (Appendix 1).

Insulation Score: The score represents the total interactions within the interval. Areas with low values in the insulation profile, which indicate high insulation, are identified as TAD boundaries [7]. We computed insulation scores using a sliding window approach on Hi-C matrix values above the main diagonal, evaluating the change in interaction frequencies across consecutive genomic bins (Appendix 1).

Incorporating the DI Score method into a Gaussian HMM leverages the model's strength in dealing with sequential data that exhibit hidden states. In the context of genomic data, these hidden states correspond to different chromatin interaction patterns. The Gaussian HMM utilizes Gaussian probability distributions to model the observed interaction frequencies (as informed by the DI Score) and to capture the variability within these interactions.

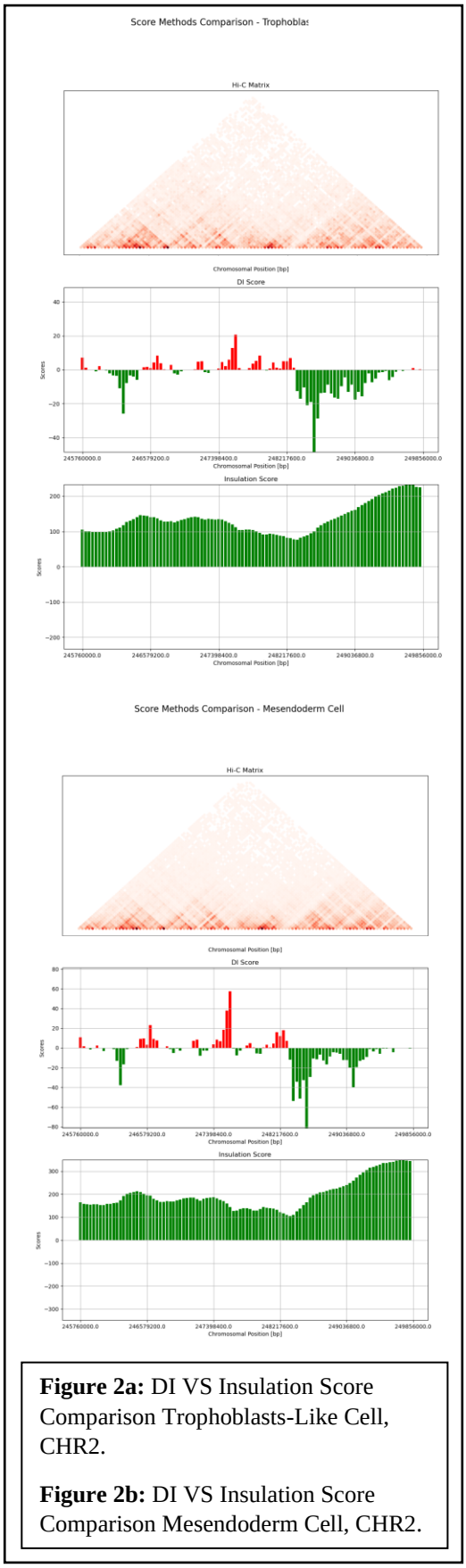


Figure 2a: DI VS Insulation Score Comparison Trophoblasts-Like Cell, CHR2.

Figure 2b: DI VS Insulation Score Comparison Mesendoderm Cell, CHR2.

Model Architecture

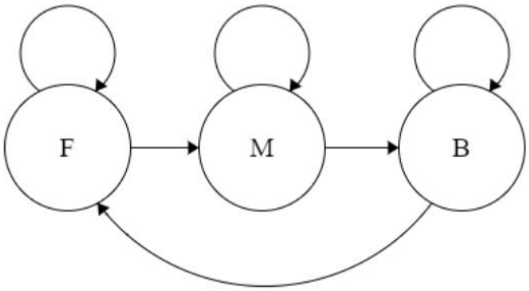


Figure 3: HMM Model Diagram

Legend

- F = Forward State
- M = Middle State
- B = Backward State

The Gaussian HMM was configured to recognize variations in interaction frequencies, with state transitions representing changes in chromatin organization indicative of TAD boundaries. The model parameters were estimated using a Baum-Welch algorithm, an expectation-maximization method that iteratively refines estimates to maximize the likelihood of the observed sequence of interactions. The model's states were characterized by Gaussian distributions, chosen for their flexibility in capturing the variability of interaction frequencies across genomic regions.

HMM Transitions

The model incorporates three primary states representing different types of genomic interactions:

1. **Backward (B):** This state models the regions where there is a predominant interaction towards the 5' end of the chromosome. The transitions into this state are typically triggered by a significant increase in the DI score, indicating a shift

from interactions across a boundary to more localized interactions within a TAD. This state is crucial for identifying the start of a TAD, where interactions begin to intensify in a backward direction.

2. **Forward (F):** The Forward state captures regions with interactions predominantly extending towards the 3' end of the chromosome. Transitions into the Forward state occur when the DI score decreases, reflecting a transition from within a TAD to its boundary. This state is essential for detecting the end of a TAD, where interactions start focusing forward beyond the current domain.

3. **Middle (M):** This state is indicative of regions within the core of a TAD, where interactions are not biased significantly in either the forward or backward directions. The Middle state typically corresponds to areas with close to zero DI score, suggesting balanced interaction across the region without the influence of directional biases typical of TAD boundaries.

To improve our model's accuracy in finding TAD boundaries, we fixed the state transition rules to better match the actual behavior of genomic interactions.

The adjusted transition matrix was constrained to reflect the transitions depicted in the accompanying HMM diagram (Fig. 3), thereby allowing only transitions from the Forward (F) state to the Middle (M) state, and from the Middle (M) state to the Backward (B) state and Backward state (B) to Forward (F) state, as well as allowing every state transition back to itself. By setting specific transition matrix probabilities to zero, we prohibited biologically implausible transitions, thus improving the accuracy of our TAD boundary predictions.

HMM Model Transition Matrix:

$$T = \overbrace{\begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{bmatrix}}^{F \quad M \quad B}$$

Utilizing Viterbi algorithm to predict the states from the Gaussian Hidden Markov Model. we accurately identified TAD boundaries by a function which captured the critical transition from the Backward (B) to Forward (F) state.

Data

The chromatin interaction data were derived from Hi-C experiments conducted on 14 primary human tissue samples and 4 cell lines. This data included high-resolution interaction matrices that record the frequency of chromosomal contacts within and between chromosomes. The data we used underwent steps to normalize interaction[8] frequencies and remove technical biases such as GC content, mappability, and DNA fragment length effects. Our data were taken from Schmitt Et Al[9]

The processed data provided a standardized basis for comparing interaction patterns and modeling TAD structures across different human tissue types:

- Prefrontal Cortex
- Hippocampus
- Lung
- Small Bowel
- Liver
- Bladder
- Ovary
- Psoas Muscle
- Aorta
- Left Ventricle
- Right Ventricle
- Spleen
- Adrenal Gland
- Pancreas

In addition, we compared interaction patterns and modeling TAD structures across different cell types:

- Mesendoderm-Cell (MES)
- Trophoblasts-Like Cell (TRO)
- Embryonic Stem Cells (H1)
- Neural Progenitor Cell (NPC)

Results

Using Gaussian Hidden Markov Models (HMMs), we successfully identified distinct regions within the chromatin that are consistent with known topologically associating domains (TADs). The HMMs effectively identified these domains by detecting changes in interaction patterns, as evidenced by shifts in the Gaussian state probabilities across chromosomal regions.

Comparative Analysis Across Cell Types

We've further explored the conservation of TAD boundaries across different cell types.
Despite the variability in chromatin structure across cell types, our analysis demonstrated a high degree of conservation of TAD boundaries.

This finding suggests that the structural organization into TADs is a fundamental aspect of chromosomal architecture that is maintained across different cellular contexts as we can see in Fig. 5,7,8.

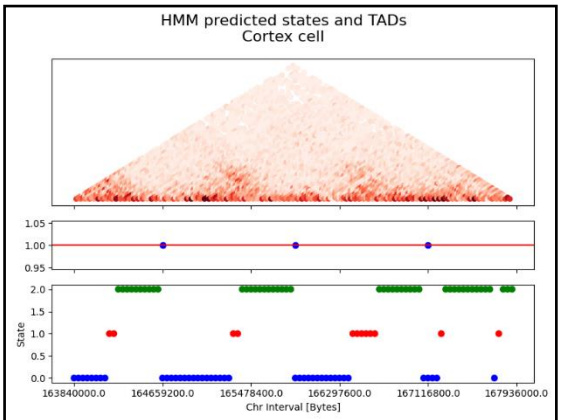


Figure 4: HMM TADs boundaries estimation Chr2, Cortex Cell.

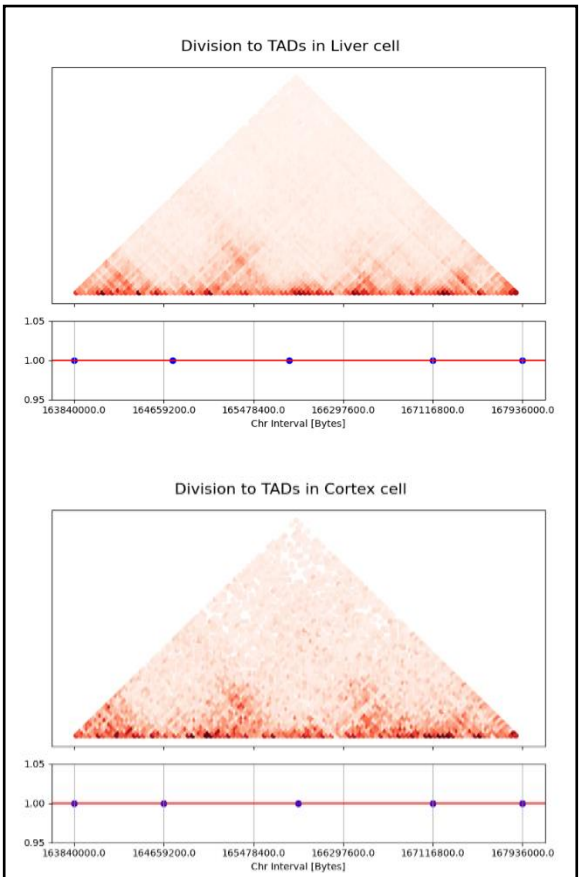
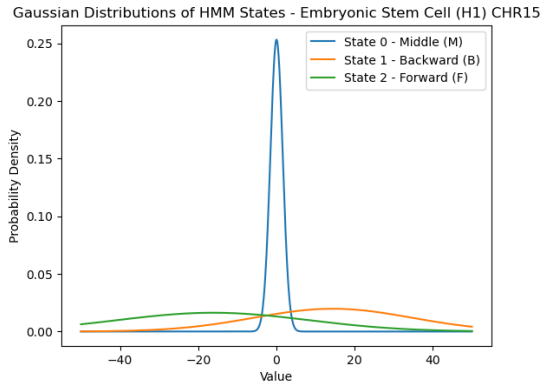


Figure 5: HMM TADs boundaries Chr2 Comparison: Cortex VS Liver Cells.

As an example, we will present the HMM States distribution for Embryonic Stem Cell Chr15:



Covariances & Means of HMM States

$$\mathbb{E}_{Means} = \overbrace{[0.0264 \quad 14.446 \quad -16.23]}^{M \qquad B \qquad F}$$

$$\mathbb{C}_{Covs} = [2.479 \quad 406.518 \quad 598.147]$$

$$\mathbb{P}_{States} = [0.5495, 0.2344, 0.1782]$$

Figure 6: HMM States distributions Chr15, Embryonic Stem Cell.

The Gaussian distributions in the plot represent the state probability densities for the three states (Middle, Backward, Forward, respectively) characterized by the HMM analysis.

State 0 - Middle (M): With a mean close to zero (0.0264) and a relatively small variance (covariance of 2.479), this state has a narrow distribution centered around the mean.

State 1 - Backward (B): The mean of this state is positive (14.446), and it has a larger variance (covariance of 406.518) compared to the Middle state.

State 2 - Forward (F): This state has a negative mean (-16.23) with the largest variance (covariance of 598.147) among the three states.

$\mathbb{P}_{States}$ – refers to the distribution of the states (M, B, F), respectively over Embryonic Stem Cell, CHR15.

Cells originating from different tissues

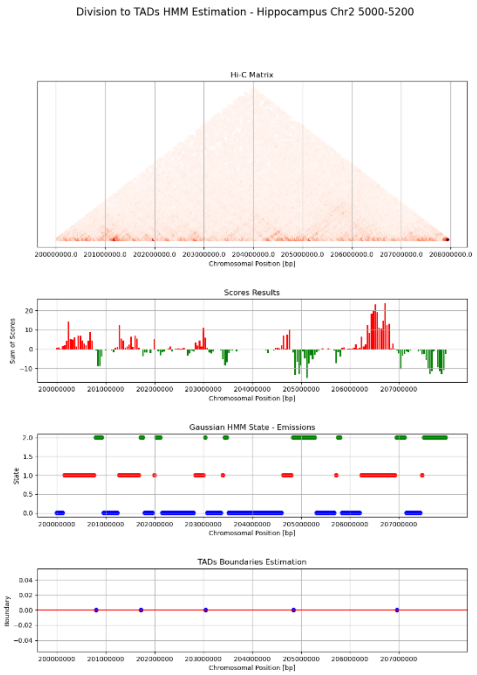


Figure 7a: TADs estimation Hippocampus Tissue Cell, CHR2.

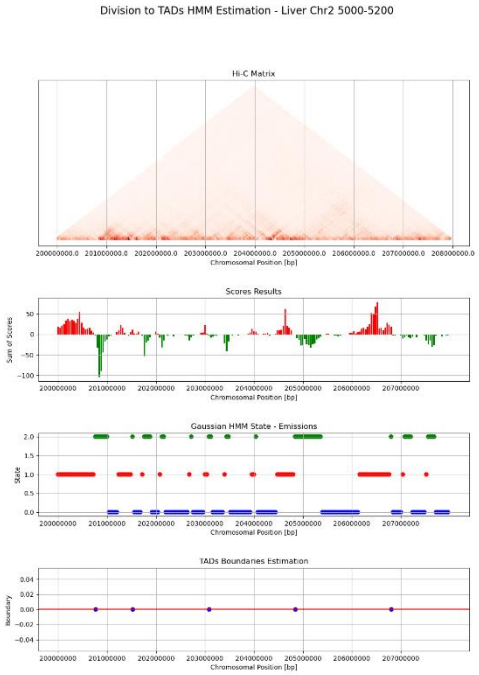


Figure 7b: TADs estimation Liver Tissue Cell, CHR2.

Different Cell Types Comparison

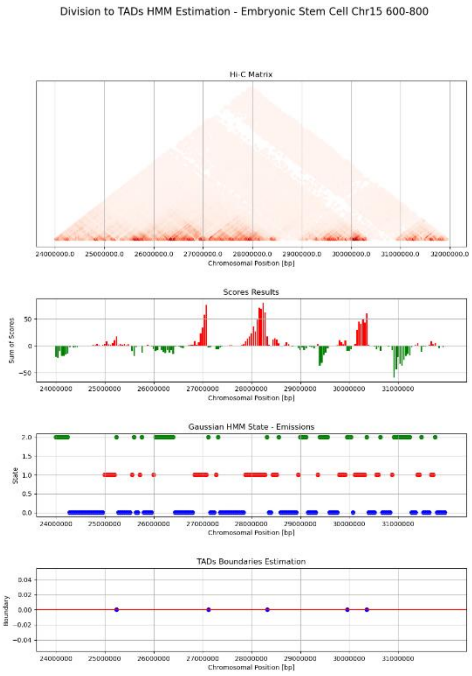


Figure 8a: TADs estimation Embryonic Stem Cell, CHR15.

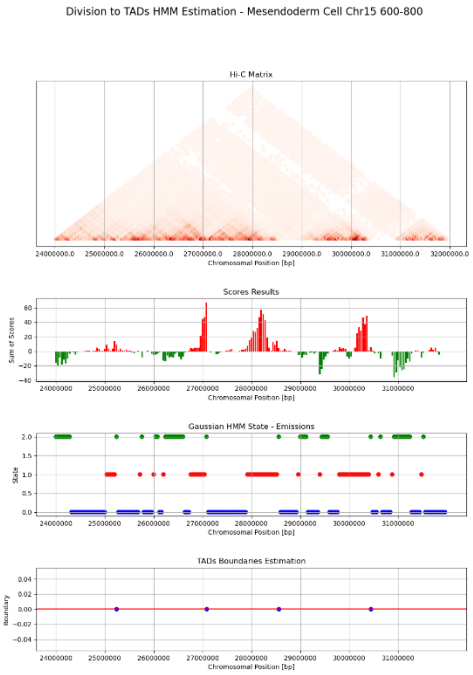


Figure 8b: TADs estimation Mesendoderm Cell, CHR15.

Discussion

Robustness and Stochasticity of the model

Although the results in each run of the Model were close, the inherent stochastic nature of HMMs introduces an element of variability: the model does not guarantee identical results across different runs. This characteristic is particularly pronounced when the model's initialization parameters, such as state transition probabilities and emission distributions, are randomly assigned.

Variability in Chromatin Interactions

The differences in the magnitude of DI scores and the sharpness of TAD boundaries observed in the Hippocampus-Liver comparison (Fig. 8) could be indicative of tissue-specific regulatory landscapes.

For instance, the liver's metabolic diversity may necessitate a more complex or varied TAD organization than the hippocampus. Such findings have profound implications for our understanding of the etiology of diseases where misregulation of boundary elements disrupts gene expression patterns. On the other hand, differences in signals might root from the measuring methods used for different cells and normalization of the data, and not related to specific biological properties of cells (Fig. 9).

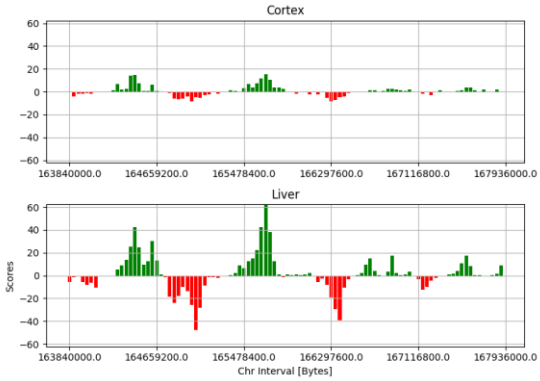


Figure 9: Cortex-Liver Chr2 DI Score Magnitude Comparison

Data Normalization

In a hackathon setting, we rely on pre-processed and normalized datasets. However, this normalization can introduce biases, especially when the methods or scales of normalization are not tailored to the specificities of the dataset and can affect on our prediction. The normalized data might have obscured unique genomic signatures crucial for accurate TAD boundary detection.

Hyperparameters: Bin size and window size

In our approach to identify TAD boundaries, we employed a scoring method that utilized a known idea where a TAD should be of size 2MB window, comprising 1MB on each side of a given point[9]. This relatively large window size was chosen to capture a broad range of interactions, assuming that significant genomic features influencing TAD boundary formation would be evident within this scale. However, this size could potentially oversmooth the data, obscuring finer genomic details crucial for identifying precise TAD boundaries. In addition our raw data was divided into 40kb bins. This constraint limited our ability to explore the benefits of higher

resolution that smaller bin sizes might offer, such as more detailed insights into genomic structures and potentially more precise identification of TAD boundaries. Future studies should focus on systematically evaluating the effects of different bin and window sizes on the detection of TAD boundaries.

Improve the Dataset

The accuracy of TAD boundary predictions depends heavily on the quality and scope of the data used. By including more diverse Hi-C datasets from various cell types and conditions, we can better analyze how TAD boundaries are conserved or vary. Additionally, incorporating other genomic data types, like ChIP-seq for histone modifications or ATAC-seq for assessing chromatin accessibility, would improve the inputs for our models. This enhancement would increase the precision and biological relevance of our TAD predictions.

Verifying Gaussian HMM Predictions with Independent Datasets

To make the Gaussian HMM predictions more reliable, it's crucial to verify them using separate datasets. Comparisons with TAD boundaries identified in other studies can serve as an external validation mechanism, ensuring that the model's predictions are not an overfit to the training data.

Conclusions

This Hackathon project presents the Gaussian HMM as an effective tool for identified TAD boundaries and in our experiments we saw a conserved nature of chromosomal organization across different cell types.

In summary, while our study has made strides in the predictive modeling of TADs using Gaussian HMMs, the challenges of data normalization, the intricacies of bin selection, the reliance on single-score predictions, and the need for extensive datasets are all critical considerations that need addressing. Future research should focus on these areas to refine TAD prediction methods.

Appendix:

Scores:

1. "***DI Left − Right***" **Score**:

$$\frac{L-R}{|L-R|}\left(\frac{(L-E)^2}{E}+\frac{(R-E)^2}{E}\right)$$

L is the sum of interaction from left and R is the sum of interaction from right (in 2[MB] window): $E = \frac{L+R}{2}$

$DI > 0$ bias Upstream , $DI < 0$ bias Downstream

2.Insulation Score:

We slide 400[kb]×400[kb] square along the main diagonal of the matrix and assign the mean of the square values to the 50[kb] bin on the main diagonal.

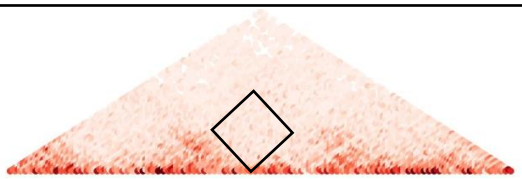


Figure 6: Sliding window square - Insulation Score visualization.

- Note: kb – kilo-base ($\times 1000$).

References

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

[2] Lieberman-Aiden, Erez et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science (New York, N.Y.)* vol. 326,5950 (2009): 289-93. doi:10.1126/science.1181369.

[3] Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558), 1306–1311.

[4] Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485(7398):381-385. Published 2012 Apr 11. doi:10.1038/nature11049

[5] Matharu, Navneet K, and Sajad H Ahanger. "Chromatin Insulators and Topological Domains: Adding New Dimensions to 3D Genome Architecture." Genes vol. 6,3 790-811. 1 Sep. 2015, doi:10.3390/genes6030790

[6] Lupiáñez DG, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161(5):1012-1025. doi:10.1016/j.cell.2015.04.004

[7] Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015 Jul 9;523(7559):240-4. doi: 10.1038/nature14450. Epub 2015 Jun 1. PMID: 26030525; PMCID: PMC4498965.

[8] Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012 Dec 1;28(23):3131-3. doi: 10.1093/bioinformatics/bts570. Epub 2012 Sep 27. PMID: 23023982; PMCID: PMC3509491.

[9] Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, Ren B. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. 2016 Nov 15;17(8):2042-2059. doi: 10.1016/j.celrep.2016.10.061. PMID: 27851967; PMCID: PMC5478386.