The Role of Protein Occupancy in DNA Compartmentalization

Kevin Zhao and Vishnu Emani, mentored by Martin Falk

Abstract

The organization of DNA throughout the genome is a complex process to study. Analysis reveals a checker-board pattern of separation at a megabase-pair scale, called compartments, which are captured well by the largest eigenvector of the Hi-C contact matrix. The sign of the eigenvector correlates with active and repressed areas of the genome. These compartments have been characterized into two categories, called A and B compartments, which are hypothesized to be spatially separated based upon the protein occupancy in the region. This project explores the factors that govern DNA compartmentalization, including the relationship between compartments and protein occupancy. In order to analyze contacts within the genome, Hi-C data was loaded and the eigenvectors of the contact matrix were computed. Protein occupancy in murine cortical neurons and neural progenitor cells was measured via ChIP-Seq. Using this data, we calculated the influence of several proteins on the sign of the Hi-C eigenvector via regression and Support Vector Machines (SVMs). Based on our findings, we tried to develop a simple model for compartments and explored this via simulations. We developed simple simulations of compartments based on ChIP-Seq data, and compared the results to compartments identified in experimental Hi-C maps. The results demonstrate a high correlation between the eigenvectors of the simulated and experimental Hi-C maps. In conclusion, the computational methods are effective at determining the proteins which most significantly contribute to compartmentalization.

Introduction

Every cell has a nucleus that contains DNA. DNA is a long strand with three billion base pairs and fits into the nucleus of every cell, which is only six microns across. Different forces act on DNA within the nucleus, which allows it to take different shapes and structures. Understanding of how three billion base pairs fold in space and undergo cell division has long been believed to be extremely relevant to reveal biological functions at the gene level as well as the global nuclear level, including gene regulation, the control of chromatin interactions, the maintenance of genetic information and the safe transfer of chromosomes from generation to generation (Belton et al., 2002).

There are many layers of organization of chromatin, which occur on different scales. On a smaller scale, DNA forms Topologically Associated Domains, caused by the formation of small loops. On the larger megabase scale, analysis reveals an organizational layer of chromatin called compartments. These compartments have been characterized into two categories, called A and B compartments, which are correlated to active and repressed regions of the genome. Chromatin loci in the A compartment clusters with other A compartment chromatin loci, and the opposite is true for chromatin in the B compartment. An open question in the field is the identification of molecular agents that lead to the physical separation of A and B compartments.

Prior research has found certain proteins to be associated with active and repressed DNA(Vermeulen et al., 2010; Soldi et al., 2013; Ji et al., 2015). Our project's initial goal was to build predictive models of genome compartmentalization based on protein occupancy data. Before discussing our results, we outline relevant methods.

Methods Overview

Many experimental techniques have been developed to explore the spatial organization of chromatin. Some of these methods use fluorescent (light) microscopy to measure the shape and distribution of chromosomes with a fine resolution (Dehghani et al., 2005), while others such as Fluorescent In Situ Hybridization (FISH) use probes that bind to a particular DNA sequence to investigate the relationship between the structure and sequence of the genome (Solovei et al, 2004). Another method, Chromosome Conformation Capture (3C), relies on molecular assays to relate nuclear architecture with the DNA sequence (Belton et al., 2002; Fullwood et al., 2009; Horike et al., 2005). Most 3C based methods only focus on interactions between certain loci, but Hi-C is a more comprehensive technique that allows “all-versus-all profiling”. This is done by crosslinking chromatin with formaldehyde, fragmenting it, and then religating only DNA fragments that are covalently linked together, which then creates a library with all possible pairwise interactions between fragments. For visualization, Hi-C displays a heat map to indicate the probability of different interactions. This visualization allows the identification of A and B compartments, whose spatial separation manifests as a checkerboard pattern in the Hi-C heat map.

Figure 1: Experimental Hi-C Contact Map

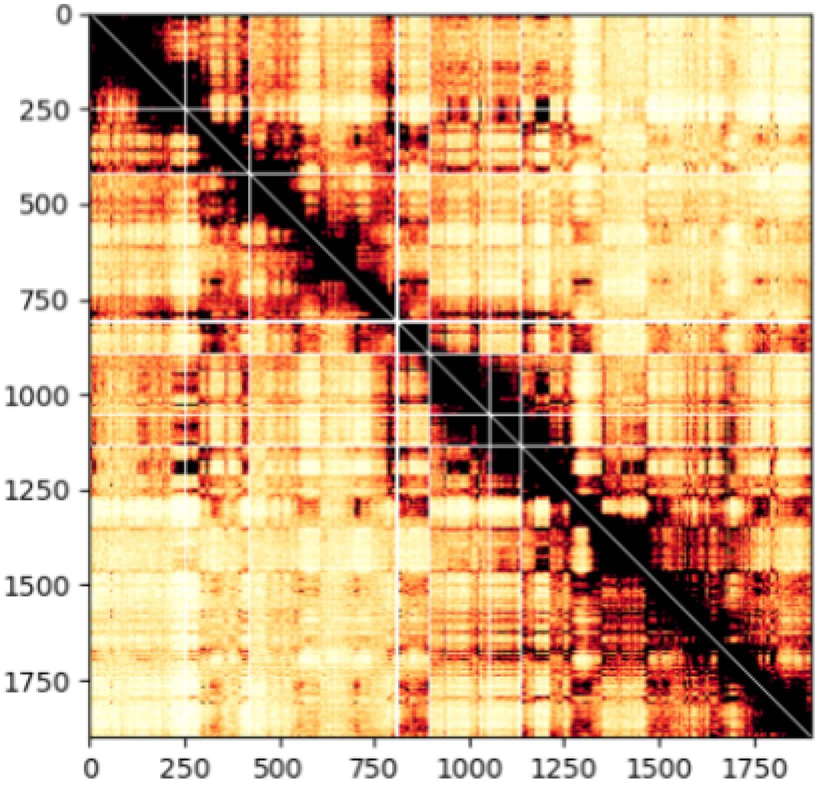


Figure 1 Caption: Hi-C contact map for CN (cortical neuron) cells, where axes are in units are in units of 100 kilobases. The distinct rectangles are examples of compartments. Darker regions correspond to more interactions and lighter regions correspond to less interactions.

One disadvantage with the Hi-C matrix is that it contains too much data which makes it difficult to be analyzed efficiently. In this project, we utilize a Principal Component Analysis(PCA) method that uses eigenvectors that summarize the behavior of the Hi-C matrix to dramatically reduce the computational burden. The sign of the largest resulting eigenvector has been shown to indicate compartments (Imakaev et al, 2012).

Proteins bind to DNA at various locations. ChIP-Seq (Barski et al., 2007) measures how likely it is for a protein to bind to DNA at a particular location, which is known as the protein occupancy. We downloaded Hi-C and ChIP-Seq data from Bonev et al. (Bonev et al., 2017) for cortical neurons (CN) and neural progenitor cells (NPC). The eigenvector with the largest eigenvalue was calculated from the Hi-C matrix, and ChIP-Seq tracks, also from the Bonev dataset, were loaded from bigwig files.

With this data, our goal was to identify proteins whose occupancy (as measured by ChIP-Seq) could predict compartmentalization (as measured by the largest eigenvector of the Hi-C matrix) in both CNs and NPCs.

Results

After downloading ChIP-Seq and Hi-C data from Bonev et al., we retained only the data from Chromosome 1, to minimize computation time. Outliers in the ChIP-Seq that had an (absolute value) z-score of more than 3 were removed. Next, the ChIP-Seq tracks were centered around the mean, and the variance inflation factor (VIF) of each protein was calculated. The protein with the highest VIF was removed since a high VIF indicates a strong correlation with other proteins. After that, the VIFs were recalculated for the remaining proteins until all VIFs are less than five. At that point, there was minimal multicollinearity between the remaining proteins. Then, all ChIP Seq tracks were normalized between 0 and 1 so that they had the same range.

After that, two different methods were utilized to determine the most influential proteins on the eigenvector of the Hi-C matrix. First, linear regression was conducted, with the 7 proteins as the independent variables and the eigenvector as the dependent variable. The importance of a protein was evaluated based on the coefficients of the linear model. The protein with the smallest (absolute value) coefficient was removed, and the coefficients were recalculated using an updated regression model with the remaining proteins. This process was repeated until there were only three proteins left. A linear SVM (Boser et al., 1992) was then used to classify the data based on the positivity of the eigenvector. This method attempted to determine the influence of each protein on the sign of eigenvector rather than the value itself. The importance of a protein was based on the coefficients of the decision boundary. Similar to the previous method, the protein with the smallest (absolute value) coefficient was removed after each iteration until there were only three proteins left.

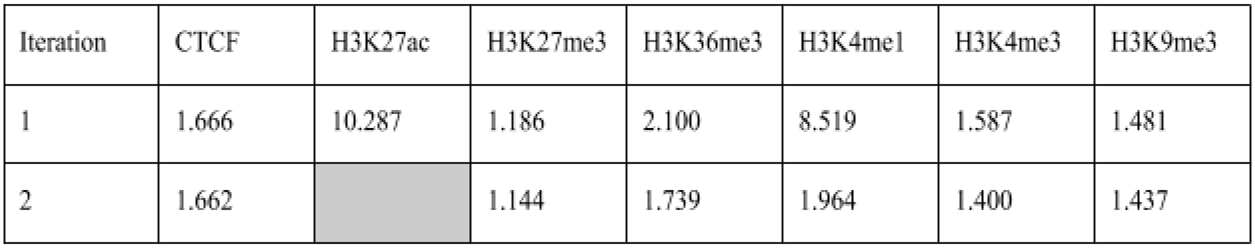
Seven proteins identified in previous studies to be associated with active and inactive regions of the genome (Vermeulen et al., 2010; Soldi et al., 2013; Ji et al., 2015) were fitted into the linear regression model[1]. For CN cells, the only protein that was highly correlated with the others was H3K27ac as indicated by it's high VIF value of 10.287, so it was eliminated[2]. The rest of the proteins all had VIF values below five (Table 1). Using the regression method, CTCF, H3K4me3, and H3K27me3 were eliminated sequentially (Table 2). Out of the three remaining proteins, H3K9me3 had the greatest (absolute value) coefficient of -1.36 and was considered the most influential protein. Using the classification approach, the same set of proteins were eliminated, although H3K27me3 was eliminated before H3K4me3. H3K9me3 was still the most influential protein with the largest coefficient of -11220.823 (Table 3).

For NPC cells, H3K27ac was also the only protein with a VIF above 5 which suggested that it was highly correlated with others (Table 4). CTCF, H3K27me3, and H3K36me3 were eliminated in the same order using both regression (Table 5) and classification (Table 6). H3K4me3 was the most influential protein in both methods. The tables below show the VIFs and coefficients for each protein.

[1] The seven proteins that were used are CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3.

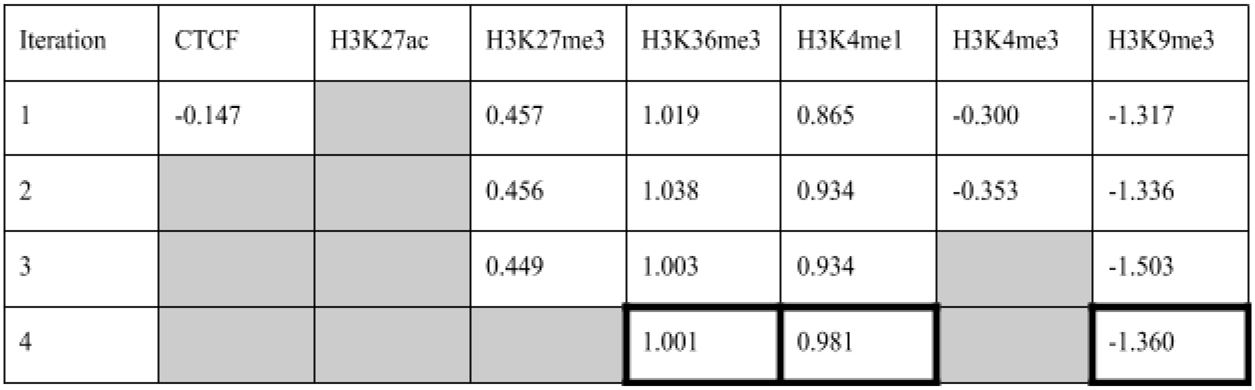
[2] Cells are shaded gray after the protein was eliminated.

Table 1: Variance Inflation Factors for CN Proteins



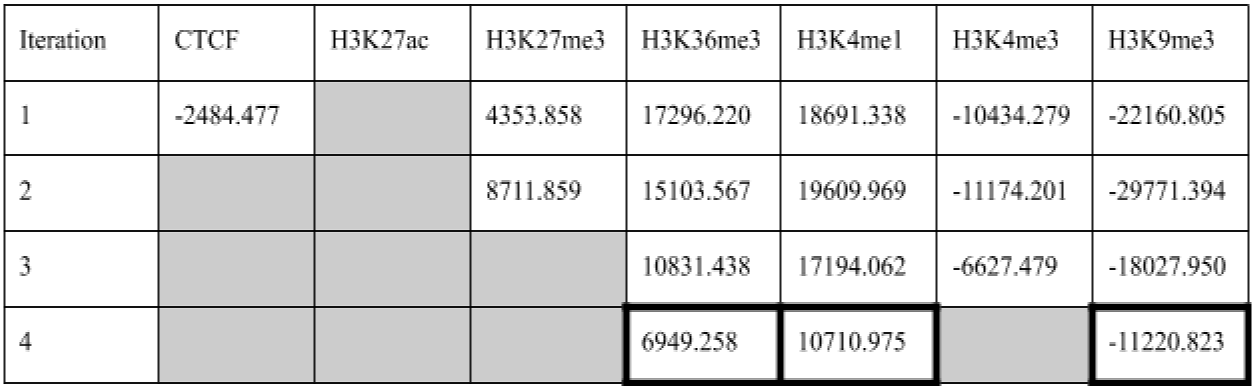
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	1.666	10.287	1.186	2.100	8.519	1.587	1.481
2	1.662		1.144	1.739	1.964	1.400	1.437

Table 2: Regression Coefficients for CN Proteins



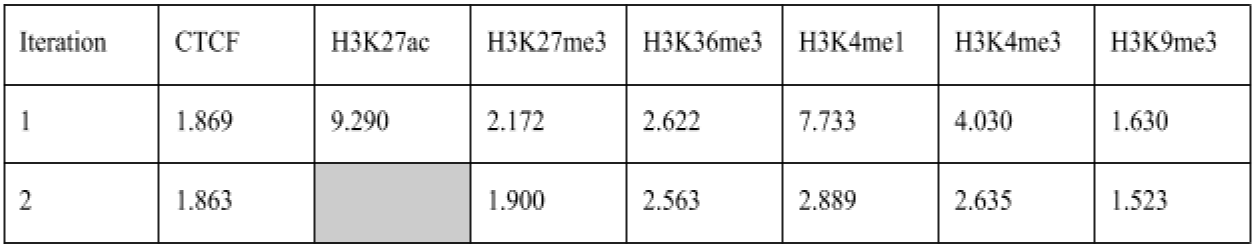
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	-0.147		0.457	1.019	0.865	-0.300	-1.317
2			0.456	1.038	0.934	-0.353	-1.336
3			0.449	1.003	0.934		-1.503
4				1.001	0.981		-1.360

Table 3: Classification Coefficients for CN Proteins



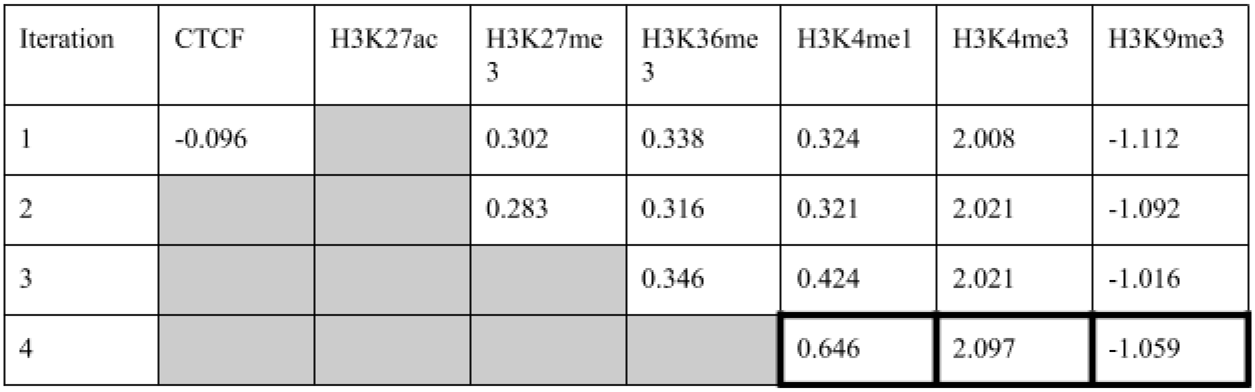
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	-2484.477		4353.858	17296.220	18691.338	-10434.279	-22160.805
2			8711.859	15103.567	19609.969	-11174.201	-29771.394
3				10831.438	17194.062	-6627.479	-18027.950
4				6949.258	10710.975		-11220.823

Table 4: Variance Inflation Factors for NPC Proteins



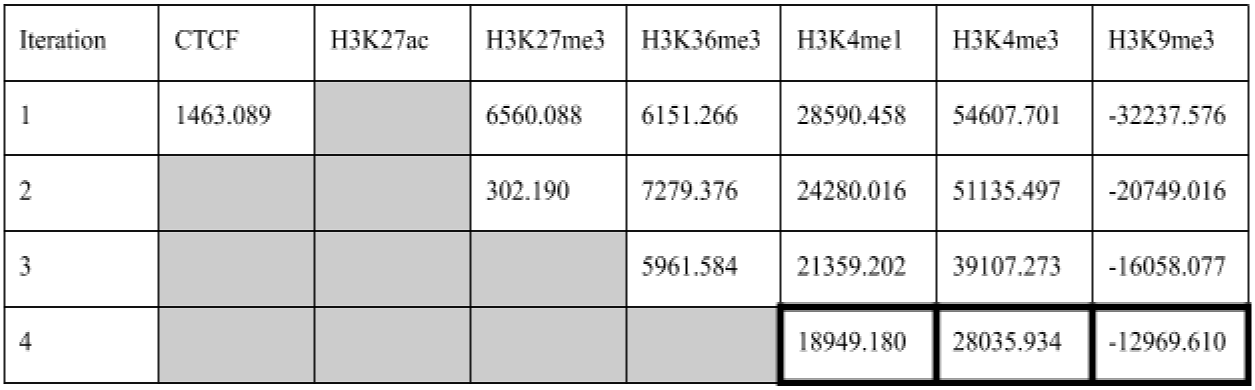
Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	1.869	9.290	2.172	2.622	7.733	4.030	1.630
2	1.863		1.900	2.563	2.889	2.635	1.523

Table 5: Regression Coefficients for NPC Proteins



Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	-0.096		0.302	0.338	0.324	2.008	-1.112
2			0.283	0.316	0.321	2.021	-1.092
3				0.346	0.424	2.021	-1.016
4					0.646	2.097	-1.059

Table 6: Classification Coefficients for NPC Proteins



Iteration	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
1	1463.089		6560.088	6151.266	28590.458	54607.701	-32237.576
2			302.190	7279.376	24280.016	51135.497	-20749.016
3				5961.584	21359.202	39107.273	-16058.077
4					18949.180	28035.934	-12969.610

In summary, H3K9me3 was indicated as the protein most predictive of compartmentalization in CNs, while H3K4me3 was the most predictive in NPCs. Furthermore, for both methods and both cell types, prediction accuracy did not suffer much when using just the top three most influential proteins versus the seven initial proteins.With all seven proteins, the regression score for both cell types was around 0.6, while the score only dropped by 0.01 when the top three were used. When trained with seven proteins, our SVM's test accuracy was 0.839, with 293 points classified incorrectly out of 1817 for the CN datasets. Using only the three most influential proteins, the score slightly decreased, dropping to 0.824, with 326 wrong out of 1852. There was a similar trend in the NPC proteins. With all seven, the SVM test accuracy was 0.878, with only 221 incorrect out of 1812. The accuracy score barely changed when only the three most influential proteins were used, dropping to 0.874, with 237 wrong out of 1882.

The results of the SVM were cross-validated in two different ways. We first cross-validated by varying the percentage of data withheld for testing our SVM between 20% and 40%. With this method, the accuracy score only differed by at most 0.02 in both cell types. The other method used the KFold approach, which evenly divided the data set into K subsets. In each split, K-1 subsets were selected to be the training set and the remaining subset served as the testing set. The number of splits for KFold were between 2 and 10. There was occasionally an unusual score at a particular iteration, but the average SVM accuracy never changed by more than 0.01, for both cell types.

Using the data from the regression and classification models, a simulation was conducted to confirm the influence of the top proteins. The complexity of the DNA molecule was simplified by modeling it as a polymer. The simulation included some generic forces for a polymer: a random thermal force, a harmonic bond force, and repulsive forces between monomers. Data from the simulation was used to compute the contact matrix. The eigenvector of the simulated contact map was compared against the experiment eigenvector to determine how realistic the simulation was in modeling the nucleus.

There were two main simulation models: variable stickiness and stochastic stickiness. In each simulation, the main force that was modified was the attraction between monomers, referred to as the "stickiness", and was based on the ChIP-Seq track. This was based on the assumption that regions which interact more with proteins should interact more with other monomers. In the first model, the "stickiness" was a continuous variable based on the value of the CN H3K9me3 ChIP-Seq track, transformed by various functions. In the second model, the "stickiness" was binary, where monomers were assigned to be either "sticky" or "non sticky" with a random probability based on the ChIP-Seq track.

Figure 2: Variable Stickiness Model Simulated Contact Map

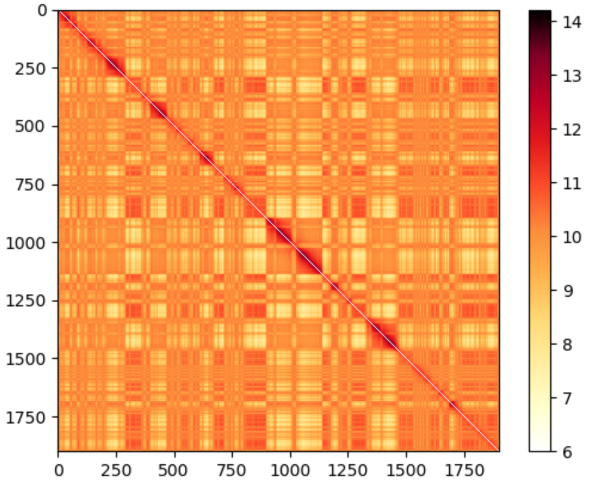
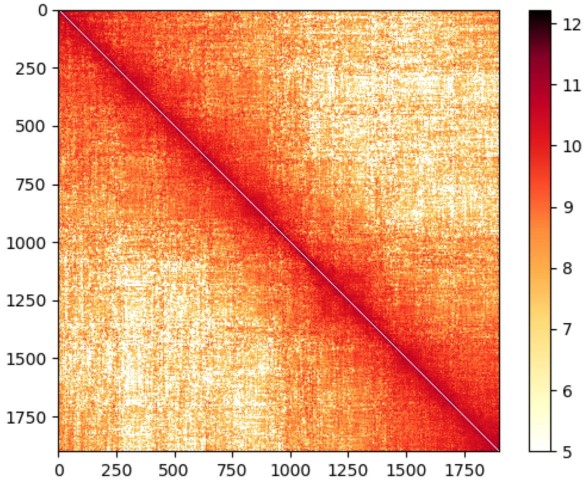


Figure 3: Stochastic Stickiness Model Simulated Contact Map



Figures 2 and 3 Caption: Simulated Hi-C contact map for variable (Figure 2) and stochastic (Figure 3) stickiness models, where axes are in units are in units of 100 kilobases. The colorbar shows ln(number of contacts). Maps are created by averaging contacts over 100 different simulations with 5000 (Figure 2) and 2500 (Figure 3) configurations per simulation. Compartments are visible for the variable stickiness model, but not obviously present in the stochastic stickiness model.

The contact map of the variable stickiness model was similar to the experimental contact map, and there were visible compartments (Figure 2). On the other hand, the contact map for the stochastic stickiness model did not have a clear separation between any DNA regions (Figure 3). One potential cause is that the stochastic simulations need to be run longer in order to see a pattern.

Discussion and Conclusion

In this project, we used protein occupancy data to predict the compartment eigenvector in Hi-C experiments in murine neural progenitor cells (NPCs) and cortical neurons (CNs). We analyzed Hi-C data from Bonev et al. and computed eigenvectors for both cell types. Protein occupancies in both cell types were measured via ChIP-Seq.

Using regression and SVMs, we found that H3K9me3 was the most effective protein at predicting compartment eigenvectors in CNs and H3K4me3 was the most useful factor for NPCs. Based on our findings, we then used the protein occupancy data from these important proteins to run two different types of polymer simulations. We found that one model, the stochastic stickiness model, was not good at reproducing the experimental data. This suggests that proteins are consistently present at varying levels in every cell, and not on or off in an individual cell. However, this may have been the result of insufficient simulation time. Additionally, depending on the cell, we found that different proteins were more important in predicting compartment eigenvectors. This suggests that future work should consider more cell types in a similar analysis.

Reference:

1. Vermeulen M, Eberl HC, Matarese F, Marks H, Denissov S, Butter F, Lee KK, Olsen JV, Hyman AA, Stunnenberg HG, Mann M. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. Cell. 2010 Sep 17;142(6):967-80.

2. Monica Soldi and Tiziana Bonaldi. The Proteomic Investigation of Chromatin Functional Domains Reveals Novel Synergisms among Distinct Heterochromatin Components. Molecular & Cellular Proteomics March 1, 2013.
3. Xiong Ji, Daniel B. Dadon, Brian J. Abraham, Tong Ihn Lee, Rudolf Jaenisch, James E. Bradner, and Richard A. Young. Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. PNAS March 24, 2015 112 (12) 3841-3846
4. Dehghani H, Dellaire G, Bazett-Jones DP. Organization of chromatin in the interphase mammalian cell. Micron. 2005; 36:95–108. PubMed: 15629642
5. Solovei I, Cavallo A, Schermelleh L, Jaunin F, Scasselati C, Cmarko D, Cremer C, Fakan S, Cremer T. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). Exp Cell Res. 2002; 276:10–23. PubMed: 11978004
6. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002; 295:1306–1311. PubMed: 11847345
7. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E, Ruan Y. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462:58–64. PubMed: 19890323
8. Horike S, Cai S, Miyano M, Cheng JF, Kohwi-Shigematsu T. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet. 2005; 37:31–40. PubMed: 15608638
9. Belton JM1, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012 Nov;58(3):268-76. Epub 2012 May 29
10. Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young, Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, Keji Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell Volume 129, Issue 4, P823-837, 18 May 2007.
11. Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 9, 999–1003 (2012)
12. Boser, B. E., Gayon, I. M. and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, Pittsburgh, PA.

13. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, Cavalli G. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell. 2017 Oct 19;171(3):557-572.e24.