Compartment prediction

Finding the optimal number of compartments

Natalia Rutecka, Barbara Jurzysta, Aleksandra Możwiłło
University of Warsaw

Meet-U 4EU+ 2021/2022 Report

Partnering with: Team 2 from University of Sorbonne
Quelin Arnaud, Rouquaya Mouss, Damien Legros, Cèdric Cornede, Hamid Hachemi

1 Introduction

A 3D structure of chromatin plays a crucial role in regulation of gene transcription. A pioneering work by Rao et al. [3] introduced a new method, an in situ version of Hi-C, which enables mapping of both intra- and interchromosomal DNA-DNA contacts. The results of a Hi-C experiment consist of a symmetric matrix, where each position represents a number of contacts observed between two loci. Loci are continuous fragments of linear genome of a given size, eg. 100 kb, called resolution.

Analysing Hi-C results from 8 human cell lines (GM12878, IMR90, HMEC, NHEK, K562, HUVEC, HeLa, and KBM7) led to discovery of 2 chromatin compartments, A and B, that were clearly separated in 3D space. It has been shown that the separation into compartments strongly corresponds to the switch between active and inactive histone modifications. Compartment A was found to be highly enriched for open chromatin (active state) and compartment B for closed, inactive chromatin.

Using high resolution data (25 kb) enabled further separation of the two compartments [3]. Compartment A was divided into two subcompartments, A1 and A2 and compartment B to three subcompartments, B1, B2 and B3. An extra subcompartment, called B4, was found on chromosome 19. Despite the fact that the partition to subcompartments is solely based on correlation values, subcompartments tend to have distinct biological properties. The differences between the subcompartments include duration of DNA replication, enrichment for certain histone modifications, GC content and gene length. Interestingly, subcompartment B4 has strong enrichment for both active and inactive chromatin marks, which makes it hard to clearly classify it as composed of heterochromatin or euchromatin.

Subsequent analysis of human reference epigenomes led to identifying average frequencies of 15 distinct chromatin states [1], which made it possible to compare enrichments for certain marks between predicted compartments. Such a comparison is crucial to identify whether the predicted compartments have significant biological differences and should be treated as separate.

In this paper, we describe a new method, called ArmReject, that predicts the optimal partition into chromatin compartments based on Hi-C data. To evaluate our predictions we look at differences in epigenetic marks' distributions between the predicted compartments and report the inferred optimal number of compartments for 5 human cell lines.

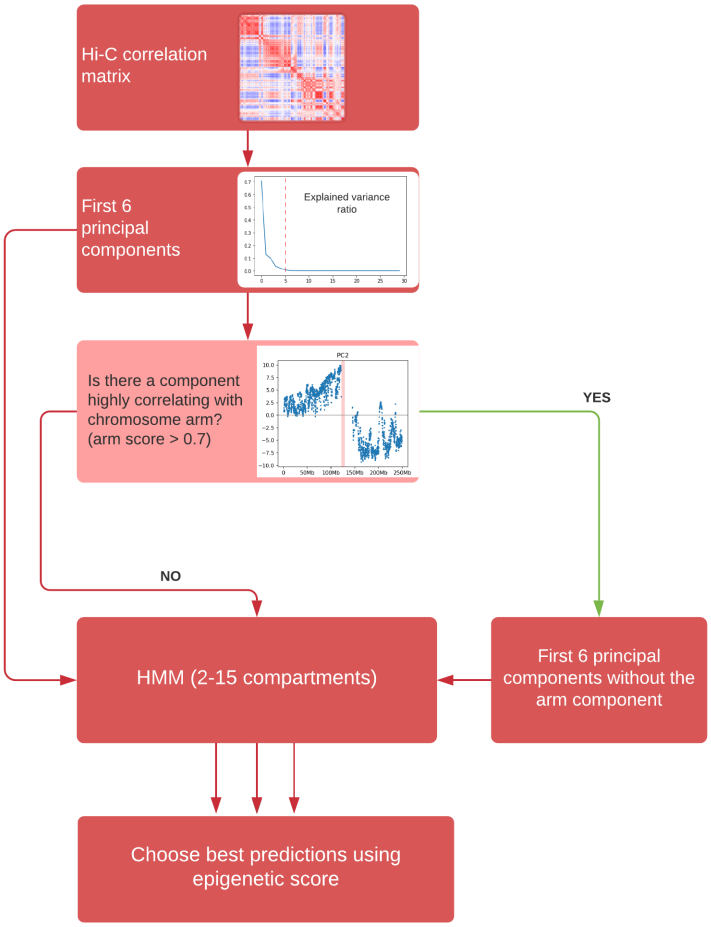


Figure 1: Graph shows workflow of detecting the chromatin compartments. First we calculate a correlation matrix based on HiC data. Then we use PCA to reduce dimensionality of our data to 6 Principal Components that store the biggest part of variability in the data (see plot showing variability explained by each PC). In the next steps we evaluate whether the PC is biased with chromosome arm information and decide whether to out-filter it. Finally, we train a gaussian Hidden Markov Model for a chosen number of states (we will consider 2-15 states/compartments). To evaluate our model we use custom score called epigenetic score (or epi score, in brief).

2 Materials and methods

2.1 Data source

We predict the optimal number of compartments and the partition into compartments for 5 human cell lines: GM12878, IMR90, HMEC, NHEK and HUVEC. Specifically, we use Hi-C matrices of intrachromosomal contacts at 100kB resolution. For evaluation of our models we also use annotations from ChromHMM tool for GM12878. All the

data is available at http://www.lcqb.upmc.fr/meetu/dataforstudent/.

2.2 Baseline model

When developing our baseline model we followed a popular study by Rao et al.[3] First, we filtered the input Hi-C matrix and normalised it using SCN method. Next we calculated observed under expected values. Finally, we computed Pearson correlation between each row and each column of the matrix.

We performed Principal Component Analysis on the correlation matrix and came to a conclusion that first 6 principal components store most of the variability in the data. We trained Gaussian Hidden Markov Models on the components, for numbers of states ranging from 2 to 15. The separation into states was interpreted as a partition into chromatin compartments.

2.3 Epigenetic marks

We validated the compartmentalisation quality based on the composition of 15 epigenetic marks of chromatin states, introduced in [1], which consist of 8 active states and 7 repressed states. To visualise the distribution of the marks in each predicted compartments, we used the results of ChromHMM tool. We normalised the frequency of each state by its average frequency in the genome. The normalisation allowed us to see differences in frequencies of rare epigenetic marks. The average frequencies that we used during the normalisation step can be found in "Cov." column in Figure 2.

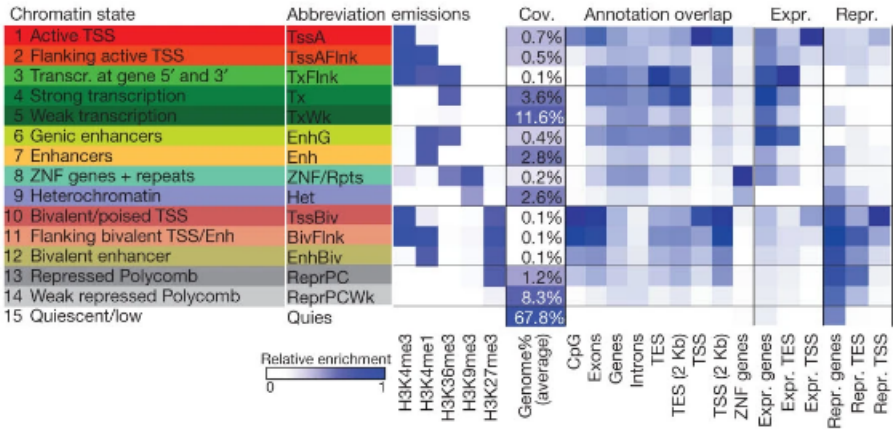


Figure 2: Chromatin states defined by Roadmap Epigenomics Consortium [1]. States are described by histone mark probabilities, their average genome coverage, genomic enrichment and gene activity enrichment.

2.3.1 Epi score

To find the optimal number of compartments for each chromosome, we define *epi score* describing the differences in distributions of epigenetic marks between the predicted compartments. To calculate epi score we use frequencies of epigenetic marks divided by their average coverage. Next we calculate the average of absolute differences between compartments' epigenetic marks (see formula (1)).

$$d_{ij} = \frac{\sum_{m=1}^{15} \left| \frac{e_m^i - e_m^j}{\bar{e}_m} \right|}{15} \qquad \text{epi\_score} = \frac{\sum_{i=1}^{C-1} \sum_{j>i}^{C} d_{ij}}{\binom{C}{2}} \tag{1}$$

$m \in \{1, 2, \ldots, 15\}$ - epigenetic mark
$i \in \{1, 2, \ldots, C\}$ - compartment
e_m^i - frequency of epigenetic mark m in compartment i
$\bar{e}_m$ - average frequency of epigenetic mark m in the whole genome
d_{ij} - distance between compartments i and j

2.4 Arm Component

Analysing first 6 principal components on various chromosomes, we saw that there often exists a principal component that is clearly connected to location on chromosome arm, meaning that it has positive values on one of the arms and positive values on the other. The best example is the second PC of chromosome 1 from GM12878 cell line. The component has positive values in 99.3% of positions located on the first arm and negative values in 92.9% positions on the second arm (see Figure 3).

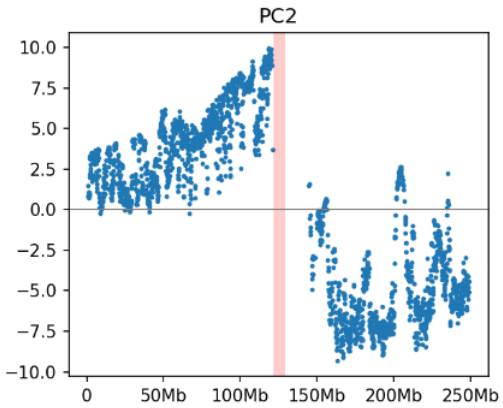


Figure 3: Second principal component of chromosome 1 in GM12878 is storing variability connected to location on chromosome arm.

2.4.1 Arm score

To evaluate if a component contains the information about chromosome arm, we calculate *arm score*, which describes how well we can differentiate the arms using the sign of PC's values. To calculate the score, we first choose the arm, on which the proportion of positive PC values is higher (arm1). Then we calculate arm score using formula (2).

$$\text{arm\_score} = \frac{1}{2}\left(\frac{\sum\limits_{i\in\text{arm1}} \mathbb{1}[PC[i] > 0]}{\text{length(arm1)}} + \frac{\sum\limits_{i\in\text{arm2}} \mathbb{1}[PC[i] < 0]}{\text{length(arm2)}}\right) \tag{2}$$

Depending on the data, arm score can reach values from 0.5 to 1, making it easy to compare its values between the components. For the 2nd principal component of chromosome 1 in GM12878, the calculated arm score is equal to $\frac{1}{2}(0.993 + 0.929) = 0.961$.

2.5 Final model: ArmReject

We used the custom arm score to select a component that stores the most variability connected to position on chromosome arm, called the arm component. Our model enables to define an arm score threshold. If none of the PCs' score is higher than the threshold, we decide that there is no arm component and only run the baseline model. To obtain our results we used arm score threshold equal to 0.7. We also run the baseline model for telocentric chromosomes (when the disproportion of chromosome arms' lengths exceeds 1:10).

In case of finding an arm component, ArmReject removes it from the analysis and trains Gaussian HMM on the remaining 5 components.

3 Results

3.1 Two compartments

We calculated the partitions into 2 compartments using different models and we compared them to the benchmark results. Best score was obtained by HMM model on full correlation matrix - it scored 91.3% accuracy. The model on 6PCs scored 78.4%. ArmReject didn't perform that well, reaching only 72.5%. The results for best performing model compared to benchmark for chromosome 1 from GM12878 are displayed in Figure 4.

Next we further analysed 2 compartments from baseline model and visualised distribution of chromatin states for each compartment (see Figure 5). We saw clear partitioning of compartments according to chromatin states into an active compartment (consisting largely of active states) and an inactive compartment (showing low frequency of the active states).

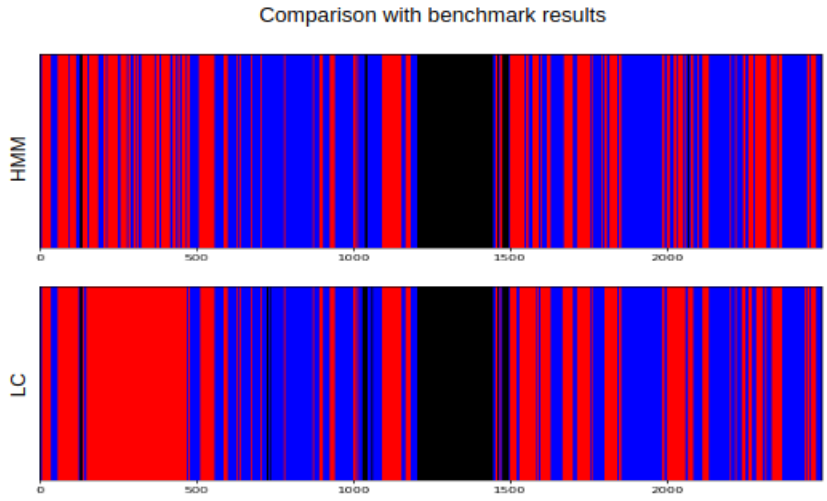


Figure 4: Results from baseline approach using Hidden Markov Model (marked as HMM) on correlation matrix compared to the benchmark results (marked as LC). Red and blue colours correspond to predicted compartment A or B, black colour corresponds to out-filtered regions.

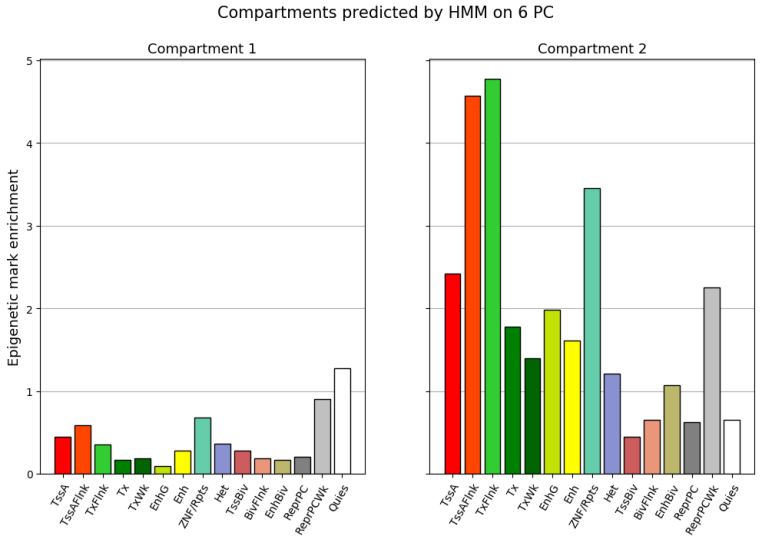


Figure 5: Distribution of chromatin states calculated by the baseline model for 2 compartments on chromosome 1 from GM12878 cell line. Here we see a clear separation into an inactive compartment 1 and an active compartment 2.

3.2 Increasing number of compartments

First, we calculated our predictions for 3+ compartments using the baseline model. We noticed that there always existed two or more compartments that had little to

no difference in epigenetic marks' distribution. An example for 3 compartments on chromosome 1 from GM12878 is shown in Figure 6.

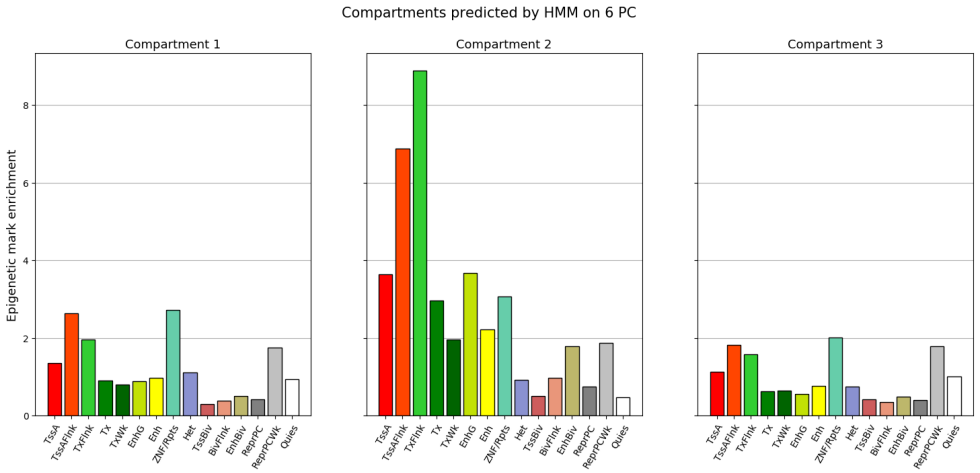


Figure 6: Distribution of chromatin states calculated by the baseline model for 3 compartments on chromosome 1 from GM12878 cell line. Compartment 2 is the active compartment (same as compartment 2 in Figure 5) and compartments 1 and 3 are almost identical and corresponding to the inactive compartment (compartment 1 in Figure 5). Interestingly, the similar compartments are located on the opposite arms of chromosome.

When applying ArmReject to the same data, we noticed that after removing the arm component, the inferred compartments were much different in terms of epigenetic marks (see Figure 7), which resulted in an increase of epi score.

However, using ArmReject did not always improve the predictions. Therefore, for the final results we calculated predictions using both methods - baseline model and ArmReject model and chose better results based on epi score. The runtime of both methods using data from one cell line took about 14 minutes.

3.3 The optimal number of compartments

Finally, we decided to calculate an optimal number of compartments, that would result from a consensus across all chromosomes in a given cell line. For each number of compartments we added the scores received for each chromosome. Then, we chose the number of compartments with highest sum of scores. The resulting scores for 2-15 compartments in GM12878 are visualised in Figures 8a and 8b.

The consensus number of compartments was equal to 3 in GM12878 cell line and 2 in the rest of the analysed cell lines (IMR90, HMEC, NHEK and HUVEC). However, one needs to remember that the model predicts compartments for every chromosome independently and therefore the compartment labels aren't universal across chromosomes.

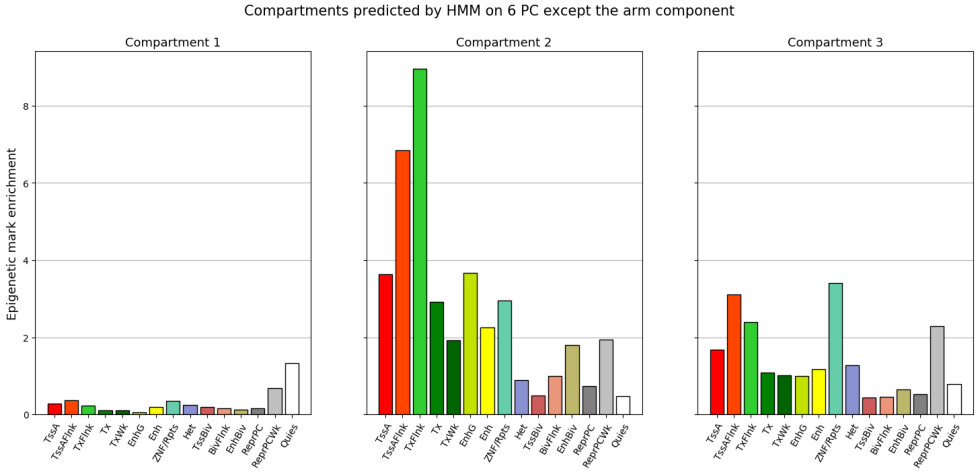
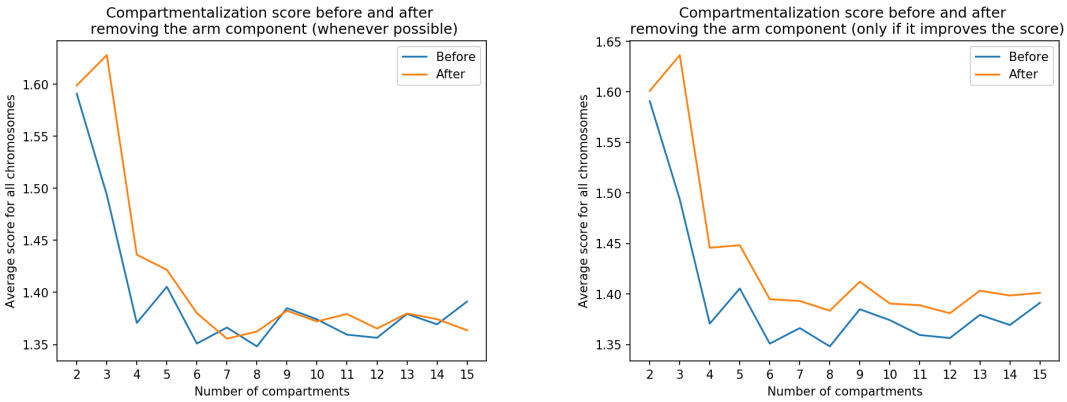


Figure 7: Distribution of chromatin states calculated by the ArmReject model for 3 compartments on chromosome 1 from GM12878 cell line. Unlike in Figure 6 we can see that the inactive component separated into two components with distinct epigenetic distributions, and they are equally distributed on the both arms.



(a) (b)

Figure 8: (a) Average epi scores for baseline model and ArmReject model. (b) Average epi scores for baseline model and a model, which uses ArmReject only when it increases the score. Average score was calculated for all chromosomes in line GM12878. It is notable that removing the arm component shifts the optimal number of compartments from two to three.

For that reason, our final goal was to find whether it is possible to define three types of compartments that are present in each GM12878 chromosome or do the biological qualities of the compartments differ among chromosomes. We decided to perform UMAP embedding [2] and clustering of compartments for all GM12878 chromosomes. We found more than 3 distinct clusters (see Figure 9), meaning that there is no clear correspondence between the partitions in all chromosomes. However, most of the com-

partments from each chromosome are distributed between different clusters, meaning that many chromosomes share similar compartmentalisation.

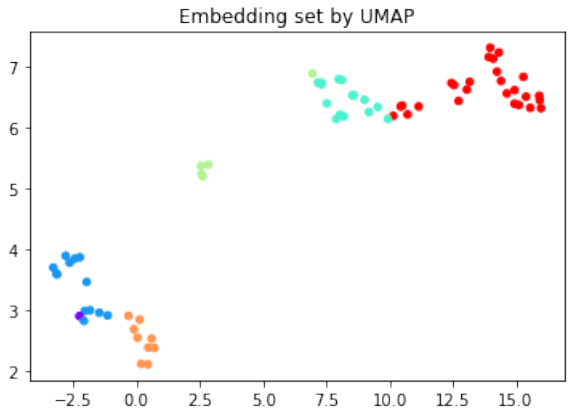


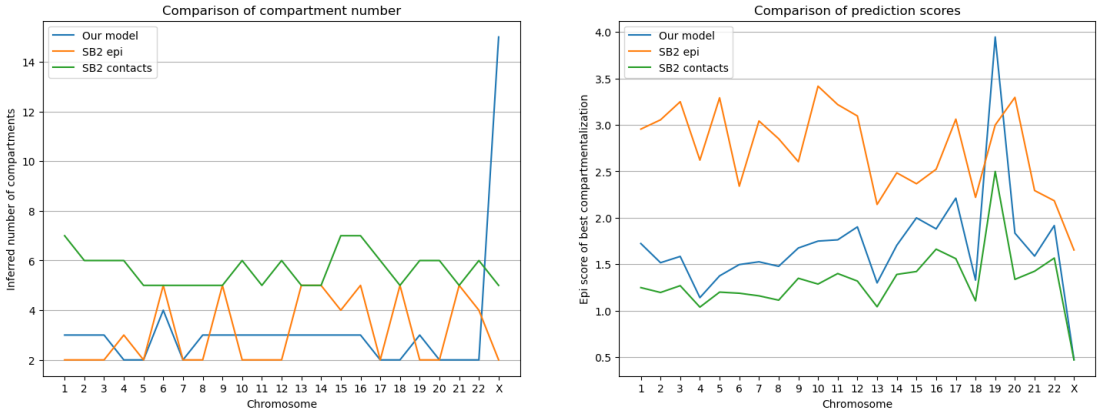
Figure 9: UMAP embedding and clustering of predicted compartments (using their normalised epigenetic marks distribution) for all chromosomes of GM12878. Each colour corresponds to different cluster found by k-Means method.

4 Comparison of results with Team-SB2

Team SB2 developed two models for calculating the optimal number of chromatin compartments. First model calculates HMM based on the first eigenvector of the correlation matrix. Second model calculates HMM model using the epigenetic marks data. Further we will refer to those models as 'contact model' and 'epi model'. Both of their models returned higher number of compartments than our model for all chromosomes except for chrX (see Figure 10a). We also compared the models using our epi score. We expected the epi model to be overfitting in terms of epi score since it is based on the same information (see Figure 10b). We see that our model gets higher scores than the contact model, but the epi model has the highest epi scores for almost all chromosomes.

Furthermore, we decided to visualise the models results using molecular visualisation system to see how the compartments are arranged in 3D space (see Figure 11). We can see that epi model does not return the expected results, because the compartments seem to be randomly shattered in space. The models based on the HiC data have more intuitive arrangement in the space.

Summing up it is hard to come up with a fair scoring system to compare the prediction of the models and find the optimal compartment number. Despite the epi model having the highest value of epi score, its results in 3D space do not seem to reflect any meaningful structural division.



(a) (b)

Figure 10: Comparison of ArmReject model and two models developed by SB2 team. (a) Comparison of chosen number of compartments for each chromosome. (b) Epi score comparison between the models for each chromosome of GM12878 cell line

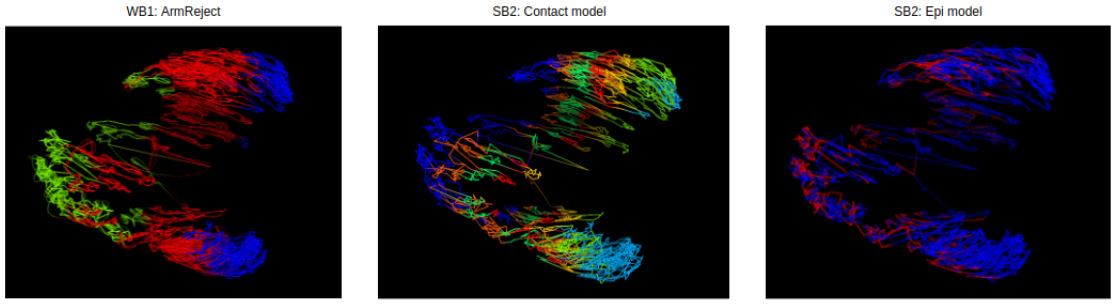


Figure 11: Visualisation of ArmReject model and two models developed by SB2 team using PyMOL tool. From the left we see: model ArmReject developed by our team, next SB2 Team models following model based on correlation matrix (label Contact model) and model calculated on epigenetic marks (label Epi model).

5 Discussion

Although further investigations are needed, our analyses on arm information brought interesting results on increasing number of potential chromatin compartments. It is still not clear whether there are any biological issues concerning rejecting information about chromosome arm in detection of compartments. Furthermore, decision on how much information we should introduce to our model like how many PCs we should choose and what data we should out-filter is still a topic we can investigate more. Biological data is often very noisy or models we consider are not capable of catching the biological complexity of the issue and detect non-obvious dependencies. Proper interpretation of the chromatin compartments roles and correctness requires expert

knowledge in field of epigenetics. Nonetheless, we believe that it is well justified to conclude that our model improves the detection of increased number of compartments.

References

[1] A. Kundaje et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330.

[2] L. McInnes et al. "UMAP: uniform manifold approximation and projection for dimension reduction". In: (2020).

[3] S. S. Rao et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In: *Cell* 159.7 (2014), pp. 1665–1680.