Sorbonne Université
Campus Pierre et Marie Curie
UPMC

MEET-U: Compartment detection Team SB2

Cedric Cornede
Hamid Hachemi
Damien Legros
Rouquaya Mouss
Arnaud Quelin

Partner Team : WB1

28-01-2022

Contents

1 Introduction

The aim of the project is to realize a project from A to Z to address a challenging open question in biology, namely the prediction of the organization of chromatin from Hi-C maps. This project is based on an improvement challenge. The goals are to develop a compartment detection algorithm. As we know as a key observation, the compartments patterns reflect the separation of euchromatin and heterochromatin in the nucleus. For that, it's important to understand how the DNA is organized in the nucleus: it's definitively the home of genetic information. Defining a genomic compartment is in fact defining its spatial distribution or the 3D organization of chromatin in the nucleus.

Chromatin is a complex of DNA and proteins that forms chromosomes within the nucleus of eukaryotic cells. Nuclear DNA doesn't appear in free linear strands. It's highly condensed and wrapped around nuclear proteins in order to fit inside the nucleus. At a dense compaction level, chromatin is called heterochromatin. In these regions of the genome, gene expression is then repressed due to lack of access to the DNA sequence. Some regions must contain a less dense level of compaction in order to allow genes to express themselves, this is called euchromatin. It shows through research that the genome can be split into two compartments, corresponding to an active and inactive one. Thanks to Hi-C it was demonstrate that the genome is partitioned into two distinct compartments A and B [3]. This is why, we can use information about interaction between genomic bins to detect genomic compartments. The Hi-C method detects all pairwise interactions, within and between chromosomes. Because of physical constraints, the euchromatin and heterochromatin can co-locate with itself and end up in the same 3D space in the nucleus.

The spatial organization of the human genome is known to play an important role in the transcriptional control of genes. Various methods have been developed to evaluate the 3D architecture of the nucleus. To interrogate all locus at the same time a method was developed which combines proximity DNA ligation with high-throughput genome-wide sequencing. They used Hi-C to demonstrate that the genome is divided into many domains (genomic compartments) which fall into two distinct compartments (heterochromatin and euchromatin). Two bins located in same compartment are amenable to interact together [6]). Through time, more subcompartments was detected, as in 2011 three-subcompartments was identified [8], in 2014 five-subcompartments noticed thanks to an HMM analysis [6], and recently in 2020 an approach based in Sub-Compartement Identifier algorithms [1]. The method is based on Large-Scale Infos Network Embedding and k-means clustering of Hi-C data. This method is considered to perform better than existing methods for predicting chromosomal compartments.

The goal of our project is to answer the following question: Can we define more than two compartments? What kind of epigenetic markers can help us determine a meaningful number and meaningful boundaries? To complete our analysis, we will also try to answer the following question: The classical resolution for detecting compartments is 100Kb. What if we increase that resolution? Can we extract more meaningful information? We used that kind of approach:

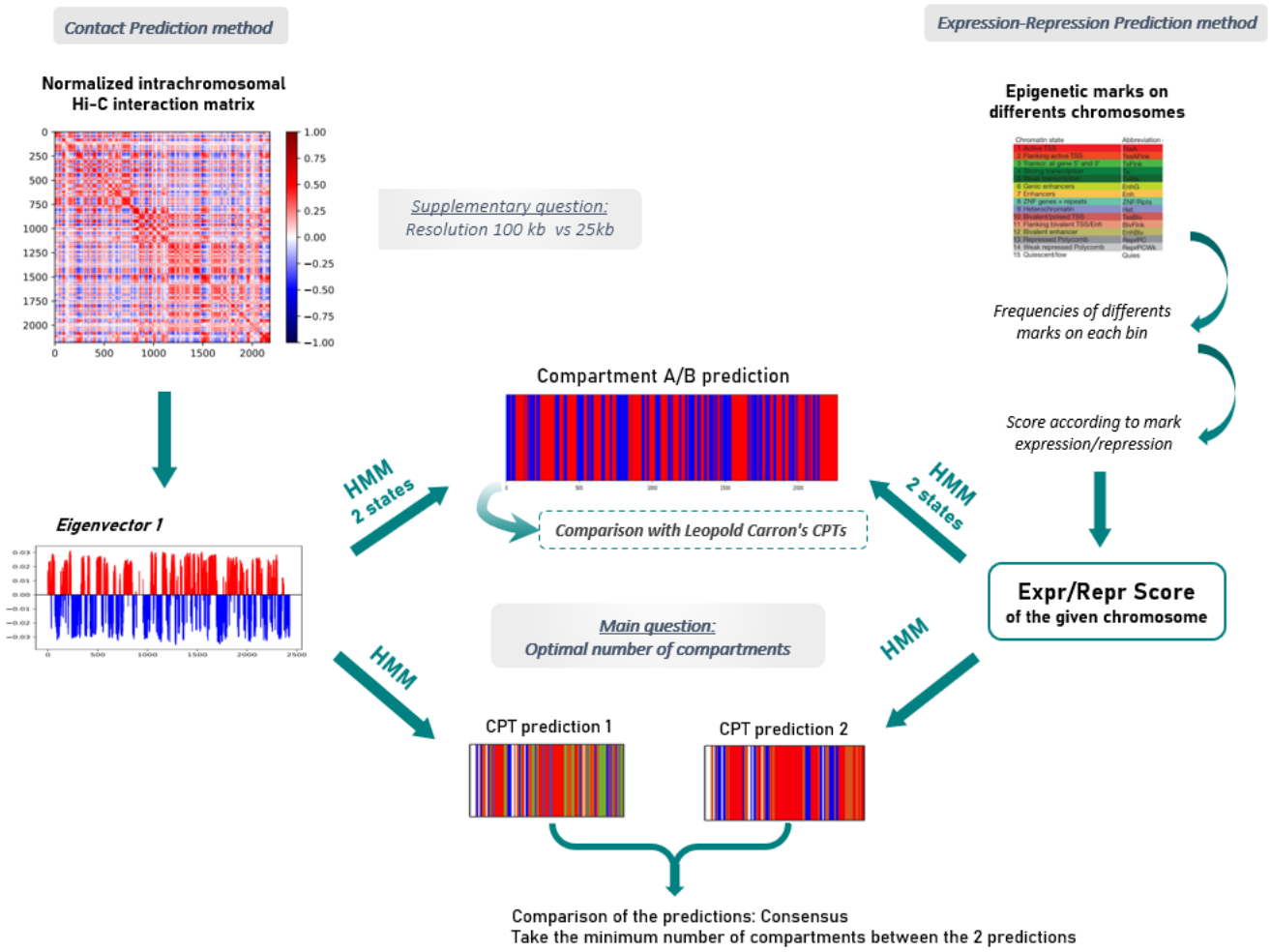

Fig. 1 : Workflow of the analysis

2 Material Methods

2.1 Input Data

Here we present our approach to predict genomic compartments using intra-chromosomal Hi-C interaction data. We have data for two resolutions: 25kb and 100 kb. The structure of each of our data is organized in the same way for each chromosome. We have a certain number of lines made up of three columns that represent the starting and ending genomic coordinates and finally the number of contacts observed between the two bins defined by the associated coordinates :

```
==> chr21_100kb.RAWobserved <==
9400000 9400000 278.0
9400000 9500000 5.0
9500000 9500000 123.0
9400000 9600000 2.0
9500000 9600000 5.0
9600000 9600000 585.0
9500000 9700000 1.0
9600000 9700000 1.0
9700000 9700000 3.0
9400000 9800000 2.0
```

In the data available to us we also have data of epigenetic marks and gene density.

2.2 Compartment prediction

To predict our compartments two different approaches, based on HMM, were pursued during the analysis.

2.2.1 HMM Contact Prediction

For Contact Prediction method we use Hidden Markov Model. Following the process that brought us to the intrachromosomal correlation matrices for each chromosome, we extract the first eigenvector from the matrices. From *hmmlearn library*, we train on these vectors, HMMs with a Gaussian distribution with 100 iterations for numbers of compartments from 2 to 16. To determine the optimal number of compartments, we look at the HMM score. The optimal number of compartments is determined through the HMM score. We consider additional compartments as long as it improves significantly (>1%) the HMM score.

2.2.2 HMM Expression-Repression Prediction

Expr-Rep Prediction method is also based on Hidden Markov Model but this time using epigenetic marks. From the given file with epigenetic marks for the different chromosomes (E116_15_coreMarks_dense) we determined the frequency of the different marks along bins. From these frequencies of the different epigenetic marks we then determine an expression/repression score for all the bins. We established our score from the different states of epigenetic marks as defined in the article Chromatin-state discovery and genome annotation with *ChromHMM* [4]. A score was assigned to each marks according to the intensity of the color assigned in the article. We then summed the scores of marks along a given bin. From this vector we train HMMs with a Gaussian distribution with 100 iterations for numbers of compartments from 2 to 16. As with the previous method, an additional compartment is added when the HMM score increases significantly (>0.5%).

2.3 Consensus of our methods

At this step we obtained a prediction of compartments with two different methods. We wanted to see how similar the 2 methods are (predict the same thing). There is still a problem to solve: if method 1 and 2 predicts, for example, 3 compartments. Is the first compartment of method 1 the first compartment of method 2 ? To answer this question we developed a function to find the consensus of our predicted compartments with our methods. We defined Consensus as label association (compartment number) between both methods that gives the best similarity score.

In this function we determine the similarity score for each "scenario" (association of possible compartments) and we keep the one with the highest similarity score. An example of one result that can be obtained for a prediction of 3 compartments with two methods is :

```
Best Scenario :  ([0, 0], [1, 1], [2, 3], [3, 2])
Similarity :  0.5219370860927153
```

The " Best Scenario" can be interpreted as follows:

- The compartment predicted with label 0 of method 1 corresponds to the compartment predicted with label 0 of method 2;
- The compartment predicted with label 1 of method 1 corresponds to the compartment predicted with label 1 of method 2;
- The compartment predicted with label 2 of method 1 corresponds to the compartment predicted with label 3 of method 2;
- The compartment predicted with label 3 of method 1 corresponds to the compartment predicted with label 2 of method 2;

The similarity score for a scenario (label association) is determined as follows : we go through the whole chromosome and calculate the number of times we have a similarity (a bin compartmentalized identically on our two methods, according to our label association ([0,1]->label 0 found in method 1 and label 1 found in method 2 ->similarity)).

3 Results

3.1 Runtime of the analysis

From `main.py` we can run as many chromosomes as we want but in practice we produced the results by batch of chromosomes. Processing time is highly variable depending on chromosome size and resolution. We summarize in the table below the approximate processing times:

Resolution	Chromosome size	Runtime of our analysis (including matrix construction)
100 kb	*short*	~ 15 min
100 kb	*medium*	~ 30 min
100 kb	*long*	~ 1 h
25 kb	*short*	~ 1 h
25 kb	*medium*	~ 3 hrs
25 kb	*long*	~ 6 hrs

Tab. 1 : Runtime of the analysis

3.2 Comparison with Leopold Carron's results

For the two prediction methods applied to all chromosomes for all cell lines, we produce a similarity score with the compartments predicted by L. Carron. Below is an example for a median similarity found with HMM contact (median of 80%):

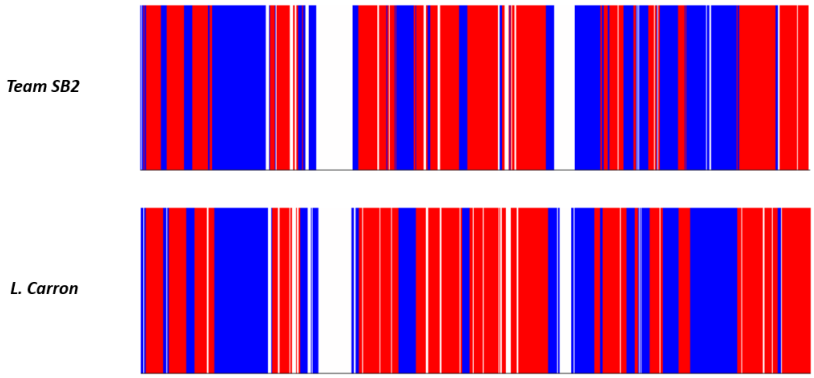

Fig. 2 : Example of comparison for chromosome 17 GM12878 at 100 kb resolution. Similarity score of 80.8% with HMM Contact Prediction

We represent below the distribution of similarity scores for the two prediction methods for all cells at 100kb resolution.

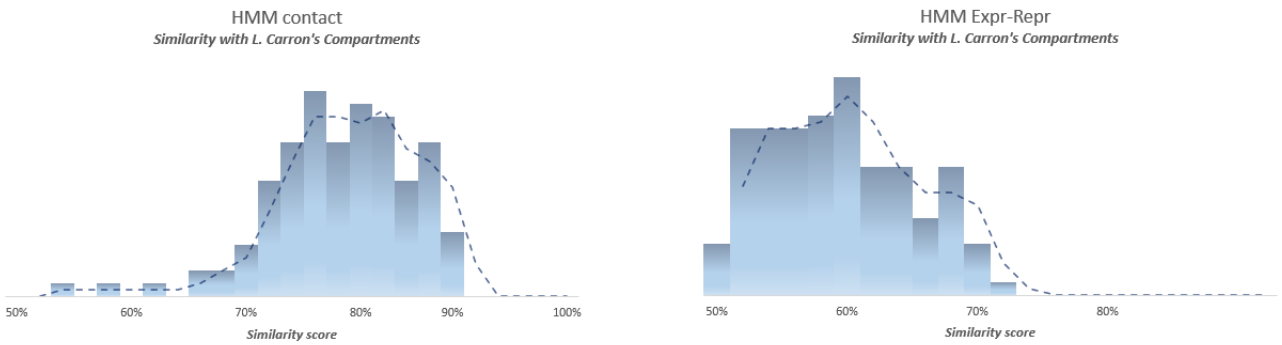

Fig. 3 : Similarity scores for HMM contact and HMM Expr-Repr

We can notice that the similarity rate is higher for the HMM Contact method than for the HMM Expr-Repr method. This is expected because this method is built on the model of the one used for target predictions. Nevertheless, it is interesting to underline that our method built from the Expr-Repr score achieves good scores in some cases, which will be further explored.

3.3 Optimal number of sub-compartments

The optimal number of compartments predicted by our HMM contact method is higher than the one predicted by the HMM Exp/Repr method. The number of compartments most often predicted by the Expr/Repr method is 2 compartments (in 57 % of cases). The optimal number of compartments predicted by HMM contact varies from 4 to 9, with 6 or 7 compartments representing two-thirds of the predictions.

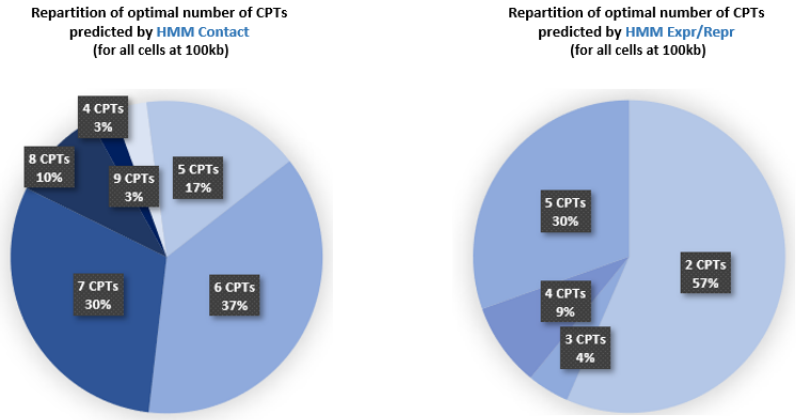

Fig. 4 : Optimal number of compartments predicted for HMM contact and HMM Expr-Repr

We observed that HMM Contact method predicts on average many more compartments than HMM Exp-Repr method. This can be explained by the fact that these two methods approach a different point of view.

HMM Contact Prediction method tells us about the spatial compartmentalization of the chromosome in the nucleus.Expr-Repr Prediction method tells us about the expression-repression compartmentalization of the chromosome, and therefore indirectly about the spatial compartmentalization. In the case of 2 compartments, there is a clear separation of the chromosome which is a very expressed part (euchromatin) and another very repressed part (heterochromatin). In the case of more than two compartments, we are in a situation of less clear-cut segregation, with sub-compartments which correspond to more or less condensed or decondensed portions of chromosomes (constitutive heterochromatin, optional heterochromatin, active euchromatin, inactive euchromatin). Thus HMM Contact underlines the spatial compartmentalization of chromosomes in the nucleus, where HMM Expr/Repr allows us to see the type of chromatin. With the HMM Contact method, we notice that the chromosomes segregate in most cases into 6-7 compartments in the nucleus. With the HMM Expr/Repr method, we note that in most cases, chromatins have 2 distinct states: heterochromatin and euchromatin. But when this is not the case, we notice that the compartmentalization takes place in 5 states which correspond to the different known states of chromatin (constitutive heterochromatin, facultative heterochromatin, active euchromatin and inactive euchromatin). We hypothesize that the 5th compartment is an intermediate state that cannot be clearly identified (lack of knowledge?, intermediate state?, quiescent state?). The predictions in 3 and 4 compartments are relative to different distributions of different states of chromatin previously described. Indeed, chromatin can be found in different states in the nucleus:

Euchromatin can be found in two forms:

- active euchromatin, which is the least condensed form of chromatin and accounts for 10% of euchromatin. It can be transcribed;
- inactive euchromatin, which has an intermediate level of condensation and cannot be transcribed, representing 90% of euchromatin

For the heterochromatin we distinguish:

- constitutive heterochromatin, corresponding to portions of chromosomes apparently inactive in all cells and carrying repetitive sequences, the largest parts of which are found near centromeres and telomeres. These regions are usually replicated late;

- facultative heterochromatin, corresponding to variable portions of heterochromatin. Depending on the type of cell, this or that portion is condensed, and therefore inactivated. In the same way, for the same cell, the same portion of heterochromatin can be condensed or not, depending on the state of differentiation of the cell.

We compare the difference between the number of compartments predicted for each chromosome in each cell with our two methods and obtain the following table:

	GM12878	HMEC	HUVEC	IMR90	NHEK
Chr. 1	5	4	4	4	7
Chr. 2	4	4	5	5	6
Chr. 3	4	3	4	3	5
Chr. 4	3	2	3	4	3
Chr. 5	3	4	5	4	5
Chr. 6	0	2	1	2	2
Chr. 7	3	4	5	5	7
Chr. 8	3	5	4	4	4
Chr. 9	0	1	2	1	1
Chr. 10	4	6	5	5	5
Chr. 11	3	4	6	4	6
Chr. 12	4	6	6	5	6
Chr. 13	0	1	1	2	1
Chr. 14	0	1	2	0	1
Chr. 15	3	5	3	3	2
Chr. 16	2	2	3	1	2
Chr. 17	4	5	4	5	6
Chr. 18	0	1	1	1	1
Chr. 19	4	4	4	3	4
Chr. 20	4	6	4	3	5
Chr. 21	0	1	1	1	1
Chr. 22	2	0	1	2	2
Chr. X	3	5	5	5	5

Tab. 2 : Difference between the number of compartments predicted by the two prediction methods

In the case where the two predictions return the same number of compartments, we have a gap of 0 in the table. In this case we have a prediction with high confidence because there is a spatial localization of the different types of chromatin, which "co-localizes" with the spatial localization of the chromosome.

3.4 Visualization 3D

The visualization was carried out on *Pymol*. We can see below the visualization for three chromosomes.

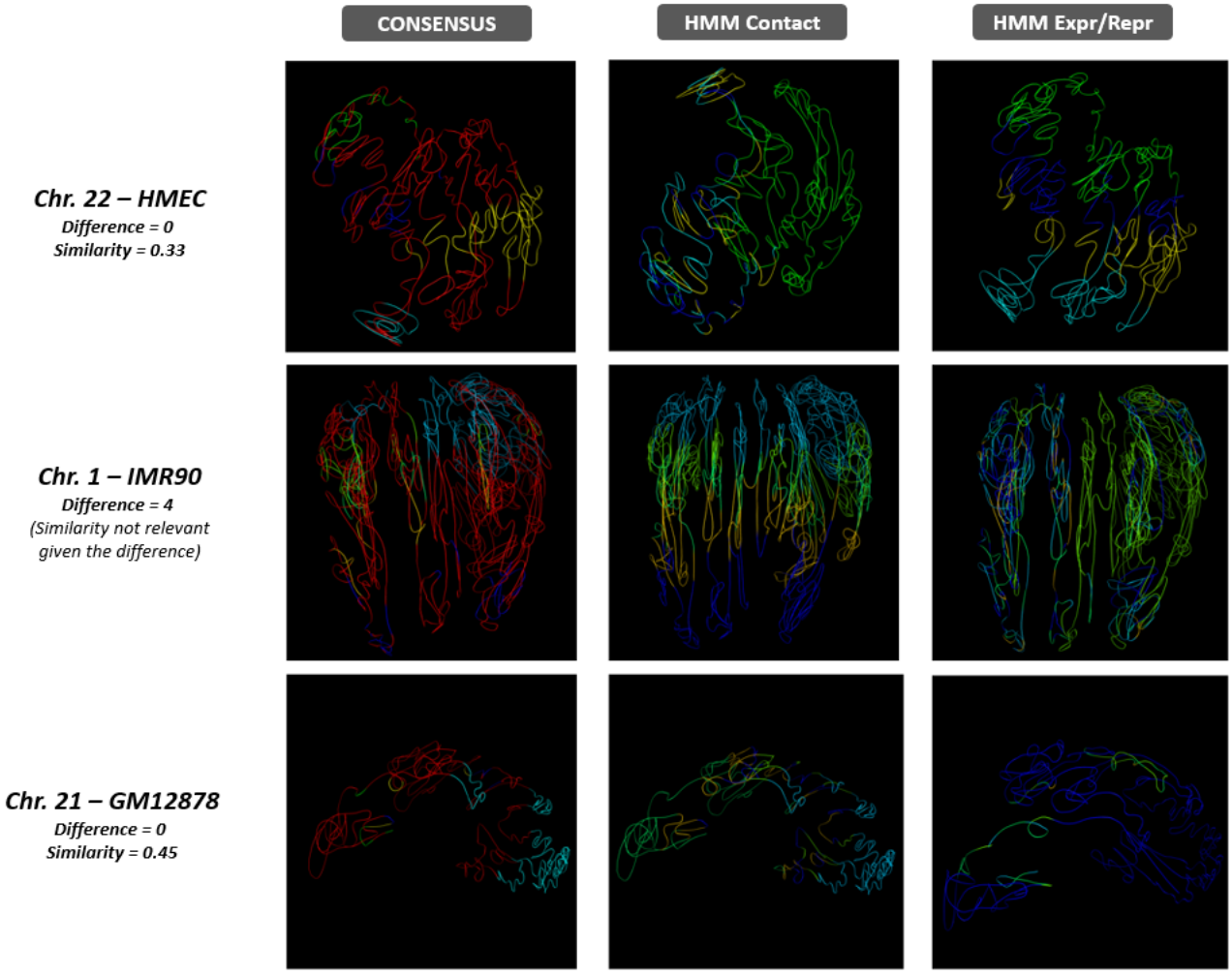

Fig. 5 : Visualization of the consensus and the two methods with *Pymol* for three chromosomes

We can see in this case that the best consensus prediction is obtained for Chr. 21 - GM12878, and the worst for Chr.1 - IMR90.

3.5 Warsaw CPT Team 1 : Discussion and comparison

Our partner team from Warsaw used a HMM to predict compartments too on the distance matrix for 100 Kb resolution of four cell lines (GM12878, HMEC, HUVEC, IMR90). Warsaw Team 1 determined and plotted the frequency of epigenetic marks of their predicted compartments obtained with their methods. We therefore used our partner team code to obtain the frequency diagrams on our own predictions to see if epigenetic marks help us determine a meaningful number and great boundaries of compartments.

If the compartments have different mark profiles, this means that there is a heterogeneous spatial distribution of epigenetic marks, therefore epigenetic marks can help us determine a meaningful number and great boundaries of compartments. Otherwise, the spatial distribution of epigenetic marks is homogeneous, and epigenetic

marks cannot help us determine a meaningful number and great boundaries of compartments.

We will perform checks on different results:

If we take the example of the chromosome 21 of the HUVEC cell, with the HMM Expr prediction we get an optimal number of compartments of 5 and if we plot the frequency of epigenetic marks for our 5 compartments with the Warsaw method we get :

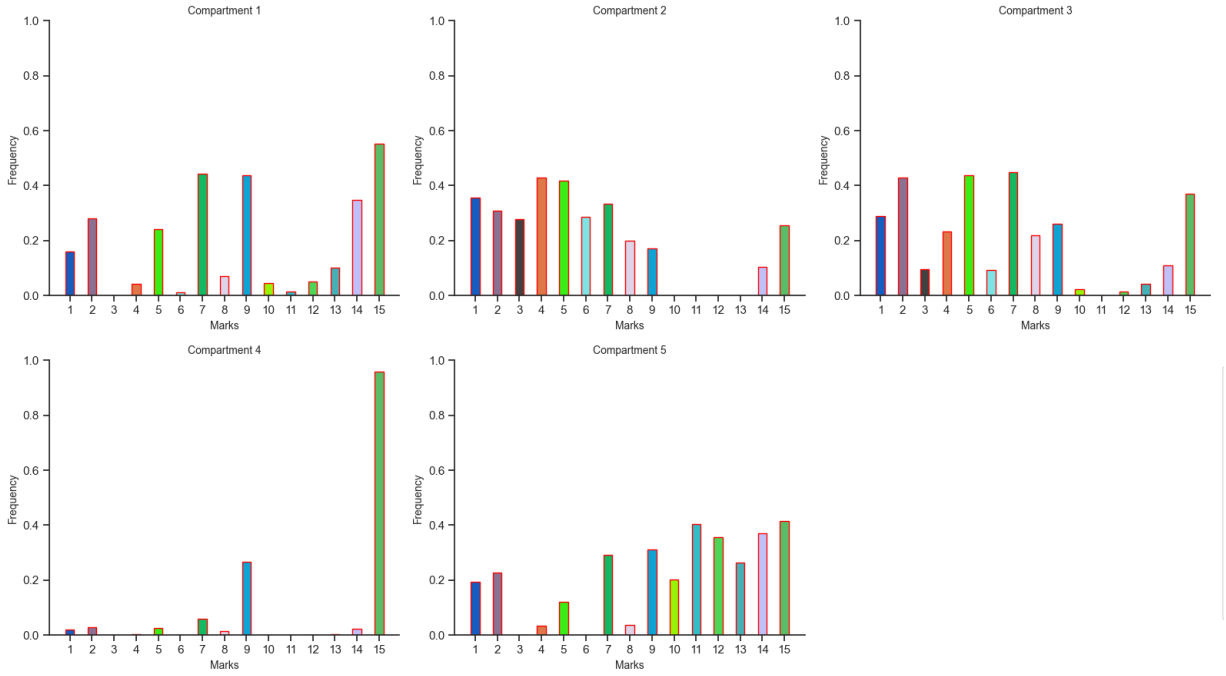

Fig. 6 : Frequency of epigenetic marks with the Team Warsaw method - Case 1

On this figure we can see that the compartments have different mark profiles, so in this case epigenetic marks can help us determine a meaningful number and meaningful boundaries of compartments.

If we take the chromosome 8 of the GM12878 cell, with the HMM-Contact prediction we get an optimal number of compartments of 5 and if we plot the frequency of epigenetic marks for our 5 compartments with the Warsaw method we get :

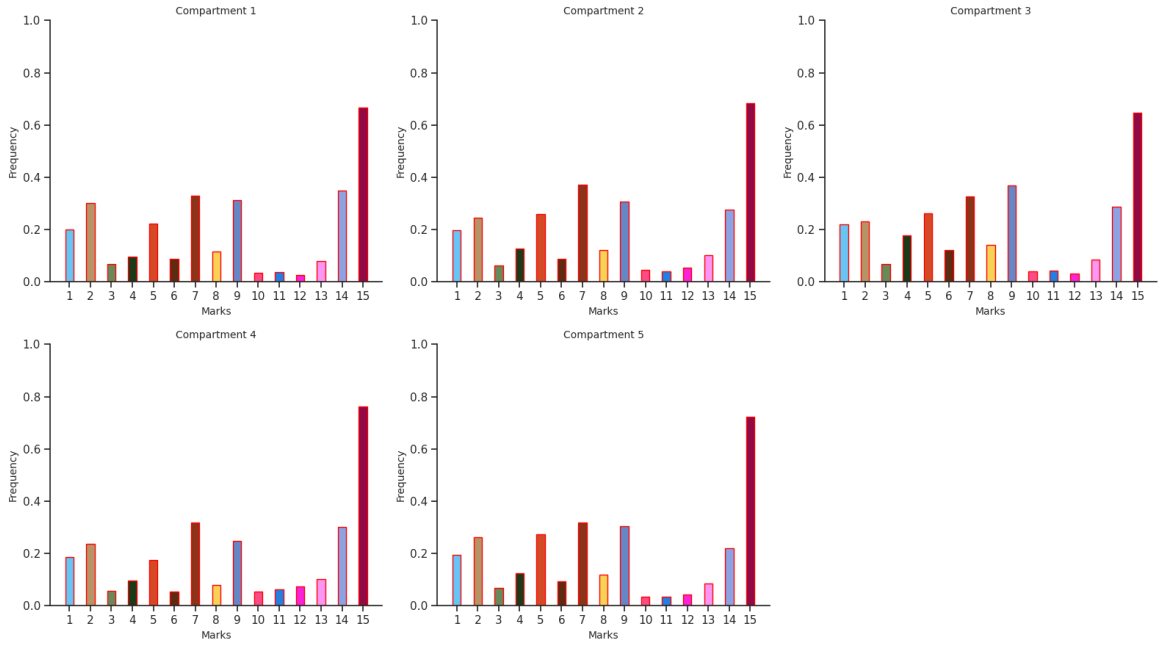

Fig. 7 : Frequency of epigenetic marks with the Team Warsaw method - Case 2

In this case we can see that each predicted compartment has a very similar profile (almost identical) in frequency of epigenetic marks, so in this case epigenetic marks can not help us determine a meaningful number and meaningful boundaries of compartments.

3.6 Supplementary question : resolution 25kb

We also performed our predictions with a resolution of 25kb. When we considered this resolution, we realized that our Contact prediction method allows us to identify TADs, as you can see below :

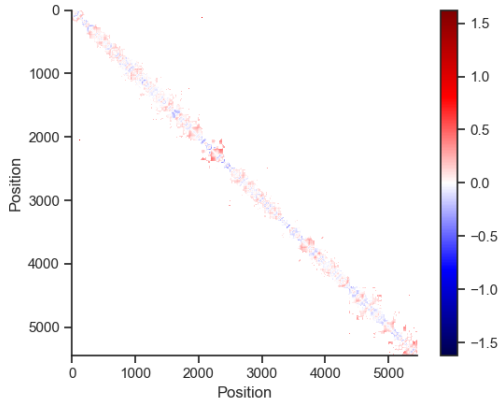

Fig. 8 : O/E matrix of chromosome 7 from GM12878

In order to express a gene, certain regions of a compartment need to be close to each others. These regions will then clump together and form clusters of chromatin. When these two regions are far apart in the sequence, a chromatin loop will form. If more than two regions need a certain proximity, a more complex cluster will form: a TAD (topological associated domain). We can identify the TAD with this resolution. We decide to compare the 25kb results with the results obtained at the 100kb resolution. The following table summarizes the analyzes performed on certain chromosomes of the GM12878 cell line:

	100 Kb			25 Kb		
Chromosome number	**4**	**11**	**21**	**4**	**11**	**21**
Nb cpts Contact Prediction	6	5	5	7	6	6
% similarity with L. Carron	58.7	72.8	80.9	73.8	81.9	84.4
Nb cpts Expr-Repr Prediction	3	2	5	2	3	3
% similarity with L. Carron	59.9	57.9	64.9	66.3	61.7	64
Nb cpts Consensus Prediction	3	2	5	2	3	3
% Consensus similiraty	42,6	62,2	45,2	66,2	42,6	45,7

Tab. 3 : Comparison of results obtained between resolution 100kb and resolution 25kb

Looking at the results above, it seems that the 25kb resolution seems on average to improve 2-compartments predictions/detection (% similarity with Leopold). Nevertheless, it seems that increasing the resolution decreases the detection of sub-compartments. (Consensus similarity for number of compartments predicted Consensus>2).

4 Discussion

Regarding the prediction of compartments, we raise as a path for improvement to make an association of labels between the different chromosomes (intra-chromosomal data) and to retain the consensus to refine our results. We could then compare this to inter-chromosomal data to refine our results and have the optimal compartmentalization.

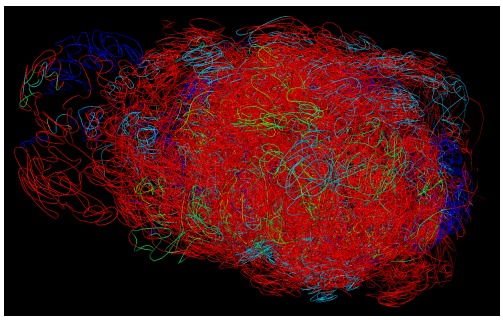

Fig. 9 : 3D visualization of consensus predictions for all chromosomes

Inter-chromosomal data, could improve our results by reducing the areas of doubt, with the analysis of these data. What is important to notice is the red zone, which corresponds to the non-consensus. The other colors are senseless. Our labels are specific to each chromosome, so we can't have a visualization of the compartments at the scale of all the chromosomes. This is also one of the interests we could have in processing inter data.

We could have a 3D visualization of the compartmentalization of all the chromosomes, which would allow us to refine our analysis, or even to be able to carry out additional analyzes from this 3D compartmentalization.

In conclusion, we can say that our HMM Contact Prediction method underlines the spatial compartmentalization of chromosomes in the nucleus, where HMM Expr/Repr Prediction method allows us to see the type of chromatin. With the HMM Contact method, we notice that the chromosomes segregate in most cases into 6-7 compartments in the nucleus. With the HMM Expr/Repr method, we note that in most cases, chromatins have 2 distinct states: heterochromatin and euchromatin. But when this is not the case, we notice that the compartmentalization takes place in 5 states which correspond to the different known states of chromatin (constitutive heterochromatin, facultative heterochromatin, active euchromatin and inactive euchromatin). We hypothesize that the 5th compartment is an intermediate state that cannot be clearly identified (lack of knowledge?, intermediate state?, quiescent state?). The predictions in 3 and 4 compartments are relative to different distributions of different states of chromatin previously described.

References

[1] Haitham Ashoor, Xiaowen Chen, Wojciech Rosikiewicz, Jiahui Wang, Albert Cheng, Ping Wang, Yijun Ruan, and Sheng Li. Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. *Nat Commun*, 11(1):1173, March 2020. Cc_license_type: cc_by Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Data processing;Machine learning;Software Subject_term_id: data-processing;machine-learning;software.

[2] Léopold Carron. Qu'est-ce qu'un compartiment génomique ?, June 2020.

[3] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, May 2012. Number: 7398 Publisher: Nature Publishing Group.

[4] Jason Ernst and Manolis Kellis. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*, 12(12):2478–2492, December 2017. Number: 12 Primary_atype: Protocols Publisher: Nature Publishing Group Subject_term: Chromatin;Chromatin analysis;Epigenomics;Gene regulation;Software Subject_term_id: chromatin;chromatin-analysis;epigenomics;gene-regulation;software.

[5] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, May 2012. Number: 7398 Publisher: Nature Publishing Group.

[6] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, December 2014.

[7] M. Jordan Rowley, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G. Corces. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Molecular Cell*, 67(5):837–852.e7, September 2017.

[8] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43(11):1059–1065, October 2011.