Predicting Hi-C contact matrices through coarse-grained simulations of the chromatin

Maurizio Gilioli

Contents

1 ABSTRACT

The chromatin is one of the most important parts of a cell. In fact, it contains in its volume the largest part of the cell DNA, and a great number of proteins, such as the histones, which help in the functional compaction of the nuclear DNA. However, the direct study of this substance encounters significant difficulties, and the analysis of related data do not give straightforward results. All-atomistic simulation approaches to predict the conformations of the chromatin in time are completely unfeasible, due to the large amount of atoms to simulate. Because of that, it is necessary to adopt a justified coarse-grained approach, which allows for simpler and less complex simulations. To this scope, I had the great pleasure of working in collaboration and on the code written by professor Marco Di Stefano. The aim of the project (which is still on-going) is to predict Hi-C matrices of contact by using coarse-grained simulations for stretches of DNA 2,000,000 bp long. Those maps are generally particularly hard to obtain and of high cost, however, the information that they contain can unveil very interesting mechanisms, such as the promoter-enhancer interactions. At first, a decompaction and a relaxation trajectories were made for 100 replicates. Then, the matrices were produced by tuning the parameters used as weights for the potentials with an iterative-stepwise approach. The results were confronted during the procedure and after with experimentally obtained maps by computing the SCC metric. Despite the fact that it has still not been reached a final and complete result, I believe that this report and its presentation could help significantly in improving the procedure and the produced analysis. To expand this study, we are already thinking about using the method for other *loci* and cell-types, and refining some of the used processes.

2 INTRODUCTION

2.1 Chromatin as the information center of a cell

All the living organisms have, inside the nucleus, the largest portion of their DNA, which is the main molecule through which information is passed from the old generations to the daughter cells. Due to the extreme length of the chromosomes, an assembly of DNA, proteins and RNA, called chromatin, is built in a necessary ordered and functional manner[1]. The most important proteins used to reach this scope are the histones, towards which DNA is wrapped around, forming the nucleosomes. To govern the functioning of the DNA, the histones and the nucleic acid itself are subjected to a variety of modifications, such as methylation. The latter alteration, in mammals, occurs in specific sites of the genome, called CpGs, where a cytosine is connected directly to a guanine. Methylations of regulatory elements have been implicated in determining cell identity and chromatin structure[2,3]. Among the proteins that interact with the DNA, CTCF is one of the most important to be mentioned. It is a protein conserved in eukaryotes and is ubiquitous in mammalians[4], and contains a Zinc-finger that binds to the DNA. The act of binding is performed in cooperation with cohesins, and causes the folding of the chromatin.[4,5].

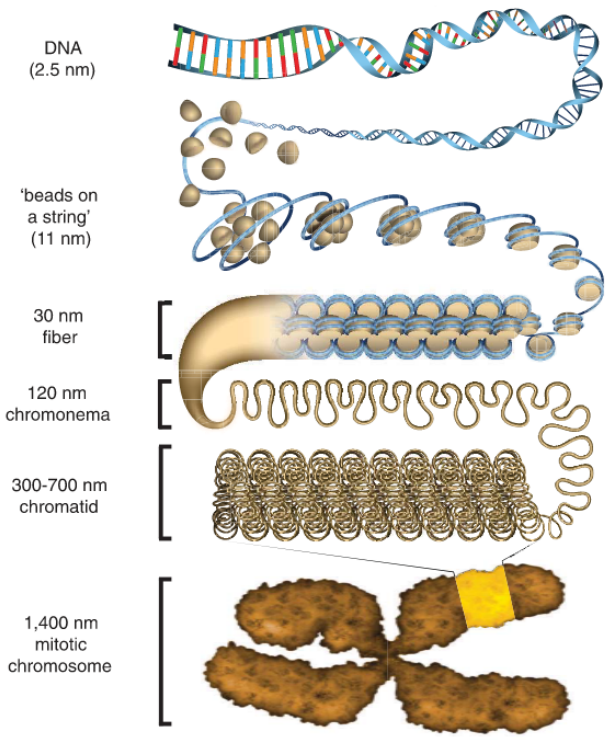


Figure 1. Image representing the hierarchical compaction of DNA in subsequently more compact and dense fibers.

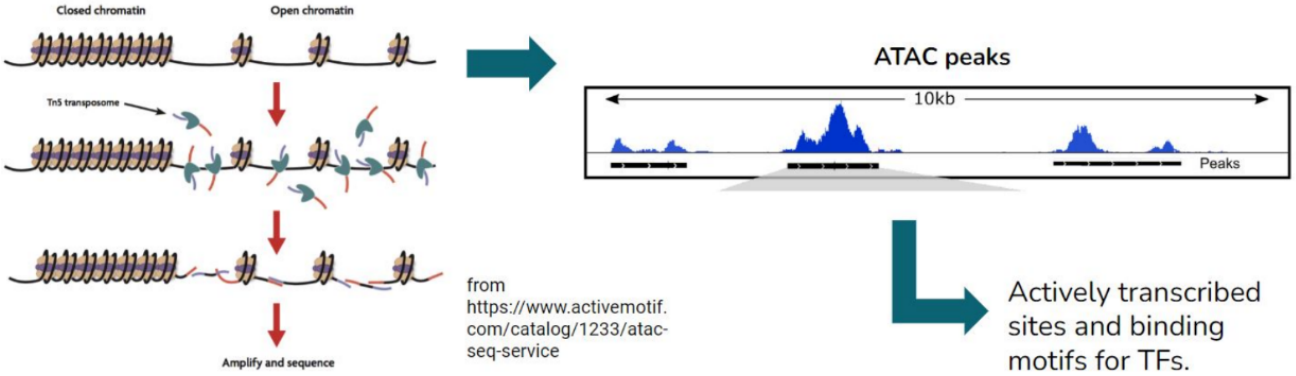
When it comes to its structure, the DNA is thought to fold in a hierarchical manner, as depicted in figure 1. However, this hypothesis makes a simplification: the fibers could have a range of diameters which depend on their activity and their location.[6,7]. Importantly, the diameter of the nucleosomes was determined to be, on average, equal to 10 nm. This type of measure was taken into account when building the Fine Scale (FS) model (chapter 4.2).

2.2 ATAC-sequencing and CTCF sequencing

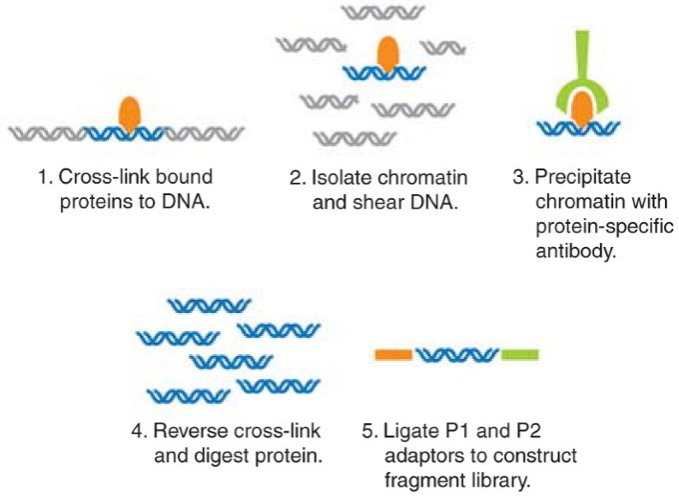
When studying chromatin, it is often very interesting to know where it is more open and accessible, and where the CTCF bins to the structure. To know these information, it is possible to perform ATAC-sequencing and CTCF ChIP-sequencing, respectively. Below it is reported a brief description of the two methods:

- **ATAC-sequencing** is a technology that allows for the identification of open chromatin regions[8,9]. In order to work, it requires the addition of TN-5, a hyper-active transposase. The latter is preloaded with sequencing adapters[9] to induce a contempourary reaction of fragmentation and ligation of the pieces released, in a process called segmentation. The obtained adapted fragments are then amplified and sequenced. Once the reads are generated, a peak-calling algorithm (generally MACS-2[10]) is used to determine which portions of the genome present ATAC peaks, and areas where there are significant enrichments of aligned reads with respect to the background. A significant enrichment of reads is possible only in accessible regions, which are generally also the most active ones and with available sites for transcription factors binding.
- The **CTCF Chip-sequencing** data used for this experiment, named in table 1, were obtained through a classical procedure: it consisted on combining a process of chromatin immunoprecipitation, made wiith antibodies specific for CTCF, and one of DNA sequencing[11]. The scope of the technique is to infer the possible binding sites of the transcription factor on the DNA.

More details about the used data can be found by consulting the ENCODE[12] entries written in table 1.



(a) Basic concept behind the ATAC-sequencing technique. The images on the left and on top were taken from the following link[13].



(b) Pipeline to perform a generic ChIP-sequencing analysis. The image was taken from the work of Anjali Shah[11].

Figure 2. The ATAC and CTCF procedures explanations.

Very interestingly, chromatin portions can be associated to a finite number of states, called chromatin states, on the base of the epigenetic data they produce[14]. Some machine learning approaches, like *ChromHMM*[15] were built with the intention of predicting these configurations. The next chapter (2.3) will talk about that.

2.3 *ChromHMM* allows the sequential characterization of chromatin states

ChromHMM[15] is a tool which helps in the annotation of genomic DNA by using epigenomic information[14]. It learns chromatin states signatures by using a multivariate hidden Markov model: in each genomic position (segment), it returns the most probable chromatin state and gives other useful information[14,15].

The package works through two functions in particular, which are the following[14]:

1. ***BinarizeBam***: it converts a set of *.bam* files of aligned reads into binarized data files in a specified output directory. The produced data can be used as input for the *LearnModel* function. When using this command, it has to be specified the segment size, which is set by default to be equal to 200 bps.

2. ***LearnModel***: it takes a set of binarized data files, learns chromatin state models, and by default produces data reporting the emission/transition parameters of the states, the abundance of the states at the TSSs (Transcriptional Starting Sites), at the TESs (Transcriptional Ending sites), and other relevant portions of the genome (CPG islands, exons, genes). Additionally, a webpage is generated with links to all the files and images created.

The results obtained from *ChromHMM* are shown in chapter 5.1.

2.4 Hi-C matrices and their patterns

Hi-C maps are useful tools to detect the interactions occuring inside a genome in analysis. Indeed, they allow to gain insights about the stuctural disposition of chromatin domains, loops and regions.[16]. The experimental procedure is described in figure 3. Each position (i, j) in a Hi-C matrix represents the number of contacts occuring between the coordinates i^{th} and j^{th} of the segment. The resolution of an Hi-C map will be dependent on the sequencing process, and have to be decided on the base of the type of information that has to be recovered from the data[16].

The following list of patterns can be found by inspecting an Hi-C matrix[16,17].

1. **Cis/trans interaction ratio**: There are higher interaction frequencies on average between pairs of *loci* in the same chromosome (*cis*), with respect to those among *loci* which reside on different chromosomes (*trans*). The ratio between *cis/trans* interactions could be indicative of the quality of the obtained Hi-C data. The viewed specificity is related to the presence of genomic territories, which govern and establish the disposition of the chromosomes in the nucleus[18]. The cis-interactions can be easily seen in an Hi-C matrix along the diagonal and are depicted in panel *a* of figure 4.

2. **Distance-dependent interaction frequency**: From the visualization of an Hi-C matrix, it is possible to observe that the largest number of interactions are registered at small distances. On the other hand, only a few contacts can be observed with high spacial separations. Because of this recurring pattern, several studies tried to predict this interesting behavior. Importantly, it was found that in a number of situations it is possible to do that: for example, in yeast the probability of interaction could be described with the following equation[16]:

$$p_{\text{interaction}}(x, y) = Z * \text{dist}(x, y)^{-1,5}$$

   Where $\text{dist}(x, y)$ represents the distance between a point x and a position y.

3. **Genomic compartments**: Genomic compartments (which can be seen in panel d of figure 4) have been found to be correlated with chromatin states, involving DNA accessibility, gene density, replication timing, GC content and histone marks[19]. The compartments can pertain to two categories, A and B, and are found by performing a principal components analysis with the matrix generated with Pearson Correlation coefficients (a formula for their calculation can be found in chapter 4.6). In general A-type compartments are defined as the euchromatic gene-dense regions, while B blocks are defined as gene-poor heterochromatic regions. The positions where they are found differ depending on the type and biological conditions of the analyzed cells[16].

4. **Topological domains**: Also called TADs, they can be visually found in Hi-C matrices as larger squared boxes centered on the diagonal of the maps (panel *b* figure 4). They are contiguous portions in which *loci* tend to interact much more with each other than with *loci* outside the region[16]. It is hypothesized that TADs specify elementary regulatory micro-environments in which promoters interact with local enhancers. Finally, some proteins like cohesins and CTCF tend to interact with the genome at the boundaries of the Topological Domains[16].

5. **Point interactions**: Those are connections occuring between small regions, and involve sequences of a few kb length. Biologically speaking, those points could indicate for example the interactions between enhancers and promoters. When considering a specific point connections, the observed value should be compared to the expected number of interactions for the distance in analysis, and the significance should be computed[16].

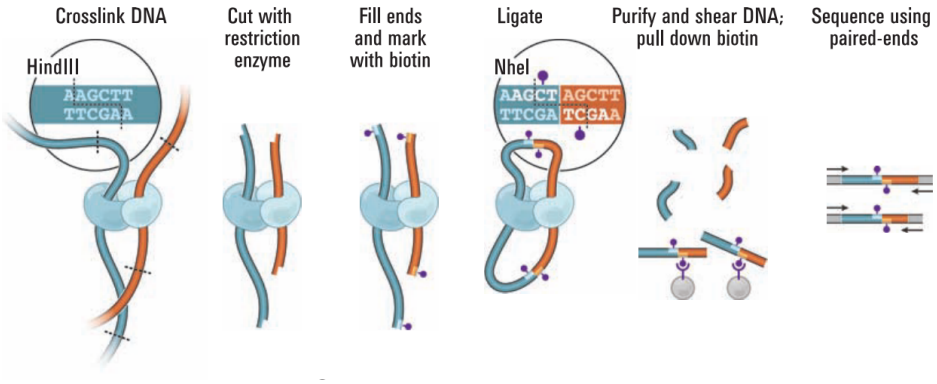


Figure 3. Image taken from the following link[19], representing the Hi-C sequencing technique in a schematized way.

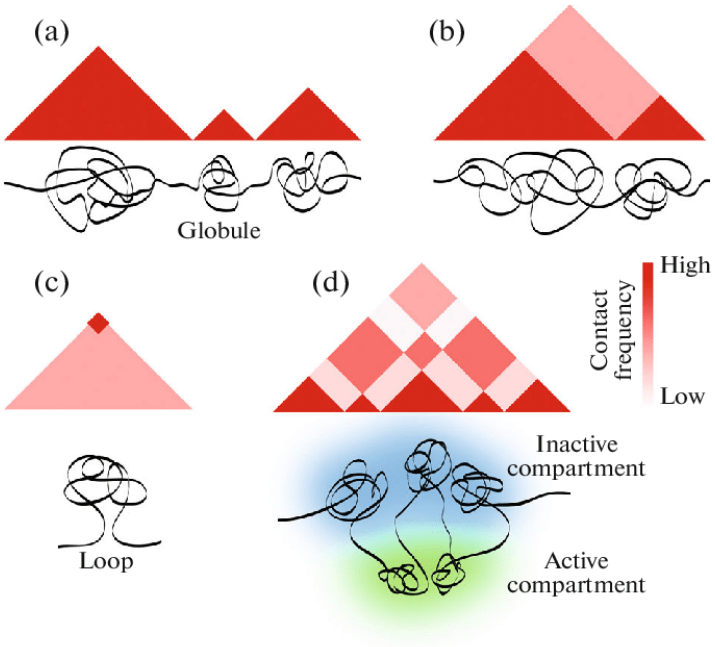


Figure 4. The image was taken from the work of Razin and colleagues[20]. Figure representing the contact typologies described in this chapter. (a) Isolated triangles on a heat map are commonly interpreted as chromatin globules deriving from cis-interactions. (b) a TAD, represented as a combination of small triangles into larger ones. (c) An intense signal at the apex of a triangle suggests the interaction of TAD boundaries and the formation of a chromatin loop. (d) Representation of some chromatin compartments.

2.5 Chromatin Coarse-Graining, chromatin as a polymeric fluid

The coarse-graining procedure performed for this project was inspired by the article written by Kremer and Grest in 1990[21].

In a very interesting chapter, which is summarized in this section, the book "*Giant molecules: here, there, and everywhere*"[22] compares the DNA to a chain of beads, and more in general, to a polymeric fluid inside the cell nucleus. It affirms that the random motion of a DNA fiber could be compared to the stochastic Brownian dynamics of a particle: several small connected segments would move randomly, however maintaining their order and connectivity. It is easy to understand that, given this hypothesis, the most important quantity that should be computed to perform the coarse-graining of a polymer is the length of the rigid segments.

To perform the following steps and derive that value, also called as Kuhn length, it is made the assumption that a DNA polymer moves in a Brownian manner. Although this type of dynamics would largely reduce the volume of the chromatin, still the dimension of the latter would be too high to permit its entire confinement inside the nucleus of any cell. For this reason, it was suggested that some other mechanisms intervene to rule the condensation behavior of the filaments.[22]. For a DNA polymer, the Kuhn length is thought to be approximately equal to 100 nm.

To start, it is defined the difference between a Brownian motion and a straight movement. That discrepancy could be written as follows:

- For a straight motion $R = v(t_2 - t_1)$
- For a Brownian particle $R = l_{\text{eff}}^{1/2}[v(t_2 - t_1)]^{1/2}$

Where $R = |\vec{R}_1 - \vec{R}_2|$ is the difference between the initial position $\vec{R}_1$ and the final one $\vec{R}_2$. The obtained equation can be rewritten as a square-root displacement in the following way:

$$R = l_{\text{eff}}^{1/2}[v(t_2 - t_1)]^{1/2} = \langle(\vec{R}_2 - \vec{R}_1)^2\rangle^{1/2}$$

By easily substituting $v(t_2 - t_1)$ with L, it is possible to derive the equation for a polymer, which becomes

$$R = l_{\text{eff}}^{1/2} * L^{1/2}$$

Where L is the maximal possible length of the polymer, and is called contour length. By computing the squared value of the previous equation, it is possible to obtain

$$R^2 = l_{\text{eff}} * L = \langle \vec{R}^2 \rangle$$
$$\rightarrow l_{\text{eff}} = \frac{R^2}{L} \tag{1}$$

Importantly, the derived equation 1 contains the definition of the Kuhn length l_{eff}. This quantity allows to understand the degree of bendability of the chain. By using a more visual approach, this length could be considered as a memory that is maintained along a path on the polymer. Indeed, keeping in mind the idea of following a "journey" on the chain, the average angle that is obtained at a contour length s, obtained through the intersection of the tangents at the starting and the ending points of the segment (figure 5), is big or low depending on the ratio between the analyzed and the Kuhn lengths. By looking to the equation 2, in general, the lower is the contour segment inspected with respect to the Kuhn length, the higher is the probability of having a low degree angle ($\langle \cos\theta \sim 1 \rangle$). On the contrary, by analyzing larger lengths, it is possible to obtain a wider range of angles, with a calculated cosine that becomes $\langle \cos\theta \sim 0 \rangle$.

$$\langle \cos\theta(s) \rangle = \exp\left(-\frac{s}{l}\right) \tag{2}$$

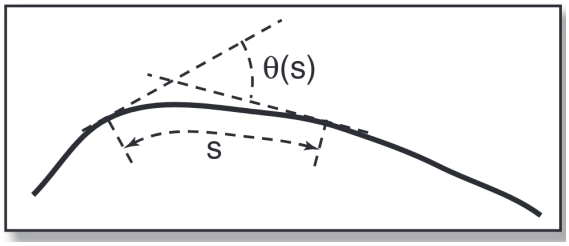


Figure 5. Image taken from Grosberg *et al.*[22]. Angle formed between the tangents at the extremes of a contour length.

To conclude, the Kuhn length is directly related the persistance length, which is another quantity, exploited in the polymer model of this work to compute the angle bending potentials, which are calculated through the equation 5. For this project, the relationship between the persistance and the Kuhn lengths was parametrized as in the formula 6.

Because of the fact that it is not really clear which and how many stages of compaction exist (chapter 2.1), it was decided to generate chains with beads including 5000 bp. The parameters for the most coarse-grained model (CG) were partially obtained by taking into account a Fine Scale (FS) configuration.

.

3 Aim of the project

The current project is a part of the thesis work, whose aim is to predict chromatin matrices of contact through the results obtained with molecular coarse-grained simulations of 2 Mb long chains. Our focus was in particular placed on a region including and surrounding the ANPEP *locus*, starting at position $89,000,000$, and finish at the $91,000,000^{\text{th}}$ base of the 15^{th}. The type of system that was generated had beads incorporating 5000 bp (CG model). As a last step, which has still to be done, a comparison between this modelling approach and others currently available, such as *Hip-Hop* and *cOrigami*[23,24], would give us a better idea about the potential of the proposed method.
The scope of this work, and specifically of this report, is also to gather opinions and useful feedbacks to improve the project, which will continue during the next months.

4 METHODS

All the simulations were performed making use of the simulation software LAMMPS[25,26], and some already made codes of Marco di Stefano, researcher at IGH-CNRS (France).

4.1 Data used during the project

The data of CTCF and ATAC for the IMR90 cell line, included in the paper written by Jimin and colleagues in 2023[24], were used for the project (see table 1). It was decided to exploit the same data of *cOrigami* to allow a better comparison between its predictions and those produced by our modelling, which will be done as a last step. Both the two technical replicates, included in the listed ENCODE[12] entries, were considered.

Cell-Type	CTCF ChiP-seq	ATAC-seq
IMR90	ENCSR000EFI	ENCSR200OML

Table 1. Table referring to the data used for the analysis.

The ATAC-sequencing data were produced following the standard ENCODE procedure[27]. In particular, the processes of read trimming, alignment, and filtering were performed making use of the *Bowtie 2*, *Samtools*, *Sambamba*, *Picard* and *cutadapt* softwares[28]. An explanation of the named processes could be found in the work made by Feng Y. and colleagues in 2020[29].

When it comes to the ChIP-sequencing data, again, the standard procedure of ENCODE was used to produce the online available datasets. An overview about the method can be found at the following link[30]. To sum up, at first the reference genome was indexed with the *BWA*, then, the alignments between the reads and the reference genome (*hg38*[31]) were produced and filtered with the *BWA*, *Samtools*, *Picard*, *BEDTools*, *Phantompeakqualtools* and *SPP* softwares.

In the case of our study, the analyzed region included ANPEP, and it was taken from the $89,000,000^{\text{th}}$ base to the $91,000,000^{\text{th}}$ position of the 15^{th} chromosome.

4.2 The Model

The model creation was done by using and modifying scripts and code written by Marco Di Stefano. The code is included in the *TADphys* unpublished package. The objective of the modelling process was to create a CG polymer model with a resolution of 5000 bp. To start the analysis, some information about the IMR90 cell type had to be collected. For example, the total genome length had to be determined, and was obtained by consulting the UCSC genome browser[32]. Secondly, the dimensions of the nucleus and the nucleolus of the IMR90 cell had to be specified. By consulting some accessible literature[33–35], the volumes of the two structures were respectively equated to 520 µm and 100 µm.
[36]

With the aim of being able to make statistics out of the generated data, and have more chains to analyze, it was decided to produce simulations for 100 replicates. Additionally, three chains were simulated at the same time within each repetition. Everything was included inside a box with a volume written in table 3. It is important to specify that all the simulations were performed making use of periodic boundaries, and were run with the *run_lammps* function included in the *TADphys* package. Three types of potential energies were set and always present in all the simulations that will be presented in this chapter. Those are described in the list below:

- **FENE potential:** The FENE potential is a finite extensible nonlinear elastic potential energy and is usually used for polymer models[25,26]. The first term in the equation is attractive, whilst the second one is repulsive and is a Lennard-Jones (LJ) potential. The first term extends until R_0, the maximum extent of the bond. The second term has a cutoff set at $2^{\frac{1}{6}}\sigma$, were the value of the LJ potential found is the minimum[26]. Indeed, in that position, $V_{\text{LJ}} = -\varepsilon$. As usual, the σ in the LJ formulation is the distance at which the intermolecular repulsive potential between two particles is zeroed.

$$E = -0.5KR_0^2 \ln\left[1 - \left(\frac{r}{R_0}\right)^2\right] + 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \varepsilon \tag{3}$$

The following specifications were made to include the FENE interactions:

1. K (energy/distance^2) $= 30.0$: It weights the contributions deriving from the attractive part of the equation, and represents the stiffness or strength of the bond.

2. R_0 (distance unit) = 1.5: Maximum extension of the bond. This is the maximum distance at which the bond can be stretched.
3. ε (energy unit) = 1.0: This term scales the LJ potential contribution.
4. σ (distance unit) = 1.0: It depicts the equilibrium bond length for the LJ potential. It represents the ideal or equilibrium distance between the bonded particles.

- **Excluded-volume interactions:** To allow for the excluded-volume interactions, a simple Lennard-Jones potential was included, set as depicted in equation 4.

$$E_{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \qquad r < r_c \tag{4}$$

Three parameters are set:

1. ε (energy unit) $= 1.0$
2. σ (distance unit) $= 1.0$
3. LJ cutoff r_0 (distance unit) $= \sigma * 2^{\frac{1}{6}} = 1.12246152962189$

- **Angle bending potentials:** The angle bending associated potentials were set trough the *angle_style* and *angle_value* functions of LAMMPS[26], and were calculated with the formula 5.

$$E = K[1 + \cos(\theta)] \tag{5}$$

In particular, K was equated to the persistence length, which was calculated as in equation 6, and its value is presented in table 3.

$$PL = lk_{CG}/b_{CG}/2 \tag{6}$$

The larger is the value of K, the bigger is the potential generated after the bending of the chain with a specific angle θ.

When it comes to the effective modelling process, the performed steps are described in the following paragraphs:

The computation of the parameters for Coarse Graining: Both parameters for the fine-scaled (FS) model (table 2) and the CG model (table 3) were calculated. The values for some of the variables obtained in the first case were used to compute and derive some of the latter system (see table 3). The number of bp wrapping around a bead in the FS method was considered to be 150, while instead the linker portion was set to be of length 50 bp (table 2). The thickness of a bead, which corresponds to a nucleosome was taken as equal to 10 nm, while instead the default Kuhn length was set to 50 nm. The genome densities ($\rho_{FS} = \rho_{CG} = 0.012\,\mathrm{bp/nm^3}$) were imposed to be the same for both the FS and the CG model; this value has been found experimentally and represents the density of bp inside the human nuclei[37]. Because of that, before computing the parameters for the coarse-grained model, the DNA content of the Kuhn segments in the CG system (Dlk_{CG}) was tuned to match the desired value of DNA content in CG beads (ν_{CG}). The spheres and bonds had all the same length in the FS and the CG models, consequently, the contour lengths (which represent the maximal lengths of the polymer chains) were exactly calculated as the products between the number of beads and their size.

Generation of the initial conformation rosettes: Once the coarse graining parameters were calculated, rosettes for all the replicates were built, with *radii* of 12.0 nm inside a cubic box whose edge length was equal to 300 nm. The particle *radius* of the CG model was set to 0.5 nm. In each replicate, three equal chains were built, by setting a different random seed each time. The total number of particles in each chain was of 400 beads. Indeed, if the calculation is made, $2{,}000{,}000/5{,}000 = 400$.

Finding the optimal pressure: When the rosettes were successfully made, a decompaction was performed taking as inputs the compacted configurations. A range of values of pressure was tested from 0.1 to 1 with steps every 0.1. The obtained sizes were compared to the target calculated size (table 3). The precision of the estimation for the pressure was improved by taking more decimals and restricting the length of the range of tested values. For each replicate, a new random seed was generated and stored, again. The simulations done in order to find the optimal pressure parameter were long 1000 steps with a duration of 0.001 ps. Before attempting the decompression, the minimal energy structure was found by taking into consideration a stopping energy tolerance of $1 * 10^{-4}$, a stopping tolerance force of $1 * 10^{-6}$, and a maximal number of iterations and evaluations of 100,000 steps. At the end, the optimal pressure was attested to be at 0.192.

Decompaction and relaxation: Once the optimal value for the pressure was found (0.192) , other two simulations, respectively $5,000,000$ and $25,000,000$ steps long, were performed for each replicate. This time, the step was set to have a temporal length of 0.0012 ps. In both the cases MSD values were collected every 100 steps. A frame was dumped (saved) every 5,000 steps. At the end, the trajectories were collected and analyzed by computing the *RMSD*, the *Rg* and the autocorrelation function as described in chapter 4.4. For the sake of simplicity, to save memory and computational costs, the collection was accomplished by capturing one frame every 50,000 of them.

Computing matrices of contact: Once defined the step at which the simulations were considered to be at the equilibrium (which, in our cases, seemed to be reached at the $50 * 50000 = 2,500,000$ step, as reported in chapter 5.3), some dictionaries produced through the output of *ChromHMM*[14,15] , whose input were CTCF and ATAC-sequencing datasets (chapter 4.1), were used to define the identity of the 5,000 base pairs beads. Then, the best attraction parameters were selected in the iterative manner described by the **??** procedure. To add the attraction potentials between the beads, new Lennard-Jones energies (equation 4) were added to the systems. In particular, the cutoff distance of the potential r_0 was set to be equal to: $r_0 = r_{\text{cutoff}} = \sigma * 2.5$, where σ corresponds to the sum of the *radii* of the interacting particles (set in all the cases to 0.5 nm).

Once the interaction parameters were set, other 1 second long simulations were performed with the intention of generating contact maps. Those simulations were firstly preceded by an energy minimization process, very similar to the one already explained in the "*Finding the optimal pressure*" paragraph. After the small simulation time, the contact matrices were generated, by taking into account just the interactions occurring in the same chain (intra-chain). A contact was established whenever two coarse-grained particles were found at a distance lower then 100 nm.

4.3 Finding enriched states in 5 kbs long beads

To find the enriched states in 5 kb long bins, the procedure described in process 1 was followed. The fold change of a state within a bin was determined by dividing the proportion in the bin by the corresponding proportion in chromosome 15. Once done that, the state with the highest fold-change was assigned to the bin. A visual investigation was performed afterwards to check the quality of the "binarization" procedure by using the IGV visualization software[38].

Algorithm 1: Finding enriched states in 5-kb long bins

```
Result: Enriched states
forall chromosomes do
    Find proportions of the states in the chromosome
foreach bin do
    Calculate bin proportions for each state
    Compute the fold changes
    Assign the state with the highest fold change
Generate .bed files
Visualization in IGV of regions of interests
/* each bin is 5 kb long                                          */
```

4.4 Trajectories analysis

The analysis of the trajectories was done by taking together and considering as independent the results obtained from each single chain of each replicate. For this reason, it was also decided to put together all the outputs and produce the collective plots represented in figures 10b, 11b and 12b.

Three types of analysis were performed:

1. ***Root Mean Square Deviation (RMSD)***: The RMSD is evaluated by using the *MDanalysis* package[39] (*rmsd* in *MDAnalysis.analysis.rms*) and is calculated as follows:

$$RMSD = \rho(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i \left(\vec{x}_i(t) - \vec{x}_i^{REF}\right)^2} \tag{7}$$

   Before performing this type of calculation, the structures were aligned to the first frame (each frame of each replicate was aligned to the first frame of the replicate). This type of alignment was done making use of the *AlignTraj* function[39] . When interpreting the results obtained from RMSD calculation, it is generally considerable true the concept that the smaller is the difference between two structures, the lower is the value of RMSD. The results are written in graph 10a and 10b.

2. R_g: The Radius of Gyration was calculated as written in equation 8 through the *MDanalysis* package (*radius_of_gyration* function). This quantity is a measure of how the mass of an object is spread out relative to a particular axis of rotation. In general, it tells "how spherical" is an object[39,40] ; the higher is the value of the radius of gyration, the lower is the sphericity of the substance. The results are written in graph 11a and 11b.

$$R_g = \sqrt{\frac{\sum_i m_i \vec{r}_i^{\,2}}{\sum_i m_i}} \tag{8}$$

3. ***Autocorrelation function:*** The autocorrelation function represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It can be written as shown in equation 9[41]. The results are depicted in graphs 12a and 12b.

$$r_k = \frac{C_k}{C_0} = \frac{\frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})(A_{t+k} - \bar{A})}{\frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})^2} \tag{9}$$

Where $C_k = \frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})(A_{t+k} - \bar{A})$ is the autocovariance function at lag k and $C_0 = \frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})^2$ is the variance function.

4.5 Algorithms used for comparison

As stated in the methods section 4.2 about the simulations, the attraction potentials are sequentially added to the model in order to improve the predictions. The SCC metric (equation 10) was used to compute the difference between the control contact matrix and the CG derived one (chapter 4.6).

Two methods were then considered to add the variables, which are expressed in algorithm **??** and **??**. Since the states are differentially populated, as stated in chapter 5.1, it is possible to argue that the largest contribution to the result would be given, in a hierarchical manner, by the most populated states. As a consequence of this consideration, it should be possible to fix the values related to the most prevalent states, before considering those that are less present. This type of mechanism is described in the greedy process of algorithm **??**. If that assumption is not accepted, then either all the possible configurations have to be considered, either a better way to test the generated models should be thought.

Because of the fact that beads of different states could act coordinately in defining the quality of the results, the algorithm **??** was devised. It represents a stepwise solution which allows to solve partially the problem at the cost of more simulations to perform. In fact, any choices of parameters are deleted and rethought during the process, in a manner that tries to avoid local *optima*.

Algorithm 2: Step-wise process for matrix comparison

Input: *current_model*: Chain to be simulated associated to *Ps* array of zeroed parameters,
reference: Reference matrix
Output: Final model and set of associated parameters

```
new_models = [];
while any P = 0 in Ps do
    ADDITION;
    foreach P in Ps if P = 0 do
        ind_P ← index of P in Ps;
        foreach val in test_vals do
            addition_Ps ← Ps;
            addition_Ps[ind_P] ← val;
            Build test_model(addition_Ps);
            Get SCC from test_model and reference;
            Add test_model to new_models
    SCCs = [SCC_i for m_i in new_models];
    current_model = new_model_argmax(SCCs);
    Ps ← Ps of current_model;
    REMOVAL;
    removal_models ← [];
    foreach P in Ps do
        if any removal_condition then
            ind_P ← index of P in Ps;
            Ps[ind_P] ← 0;
            Perform ADDITION → Add add_model to removal_models
        removal_model = removal_models_argmax(SCCs);
        Ps_removal ← Ps of removal_model;
    if SCC of removal_model > SCC of add_model then
        current_model ← removal_model;
        Ps ← Ps_removal;
```

4.6 the Stratum Adjusted Correlation Coefficient (SCC) metric

The SCC metric is described in the paper written by Yang and colleagues in 2017[42,43]. It can quantify the similarity between an Hi-C matrix and another. In general, the most common techniques to use in these situations are either to analyze the matrices by eye, or, in a certainly more precise way, to calculate a Pearson/Spearman correlation coefficient. However Hi-C data have certain unique characteristics, including domain structures, such as topological association domain (TAD), A/B compartments and distance dependencies, that require a more precise approach. Indeed, the chromatin interaction frequencies between two genomic loci, on average, decrease substantially as their genomic distance increases. Standard correlation approaches do not take into consideration these structures and may lead to incorrect conclusions[42,43] .

The SCC metric could be seen as a weighted Pearson coefficient, as written in equation 10.

Variables

N_k	$k \in K$	Number of observations in stratum k;
X_k	$k \in K$	Observations in stratum k in matrix X;
Y_k	$k \in K$	Observations in stratum k in matrix Y;
$r_{1k} = \frac{\sum_{i=1}^{N_k} x_{ik} y_{ik}}{N_k} - \frac{\sum_{i=1}^{N_k} x_{ik} \sum_{j=1}^{N_k} y_{jk}}{N_k^2} = E(X_k Y_k) - E(X_k)E(Y_k)$	$k \in K$	Correlation between X_k and Y_k;
$r_{2k} = \sqrt{\mathrm{var}(X_k) \cdot \mathrm{var}(Y_k)}$	$k \in K$	Square root of the product between the variances of X_k and Y_k;
$\rho_k = r_{1k}/r_{2k}$	$k \in K$	Pearson coefficient related to bin k;

Formula

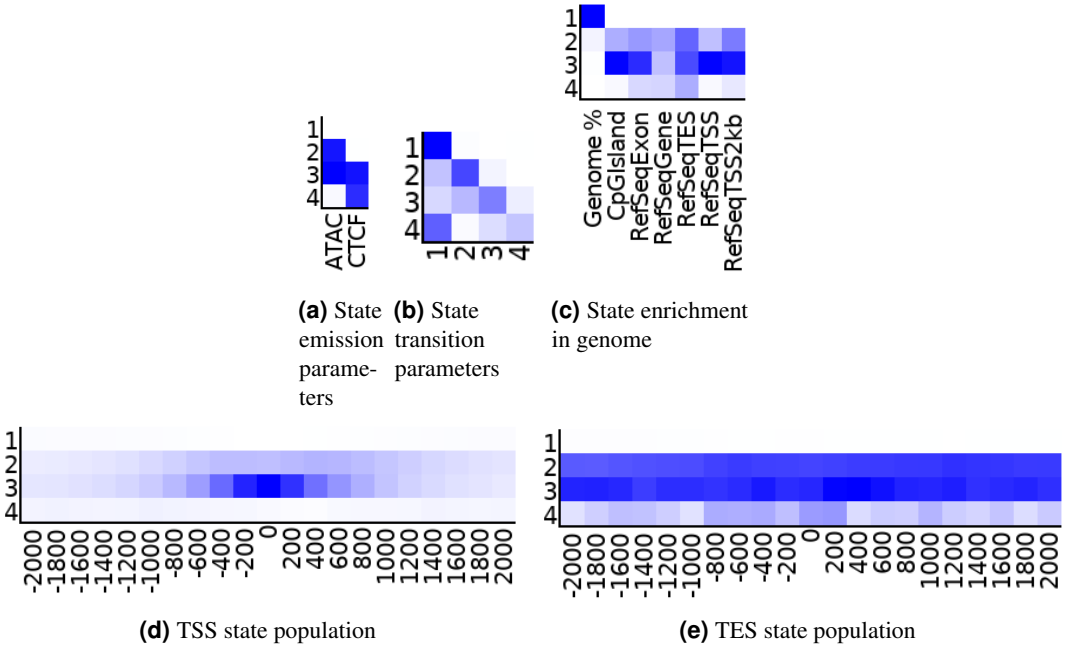
$$\rho_s = \sum_{k=1}^{K} \left(\frac{N_k r_{2k}}{\sum_{k=1}^{K} N_k r_{2k}} \right) \rho_k \tag{10}$$

5 RESULTS AND DISCUSSION

5.1 *ChromHMM* results

In total, 4 states were considered to be present. Two functions in particular were used: *BinarizeBam* and *LearnModel*. The data shown in 4.1 were aligned to *hg38* reference genome[31] . Results are shown in image 6; by taking a look to its subfigures, the following considerations could be done:

1. Clear absence or presence of ATAC and/or CTCF signals could be detected in figure 6a. for this reason, the following states are defined:
   - **State 1**: State without the presence of ATAC and CTCF signal.
   - **State 2**: State with ATAC but not CTCF peaks.
   - **State 3**: State with the presence of both ATAC and CTCF signal.
   - **State 4**: State with CTCF but not ATAC peaks.
2. The states 1 and 2, in particular, tend to perform transitions towards themselves instead of different states (figure 6b)
3. State 2 (with ATAC) and 3 (with ATAC and CTCF) tend to localize in CpG islands, exons and Transcriptional Starting Sites (figures 6c, 6d, 6e) as well as, although to a lesser extent, in Transcriptional Ending Sites (TES).

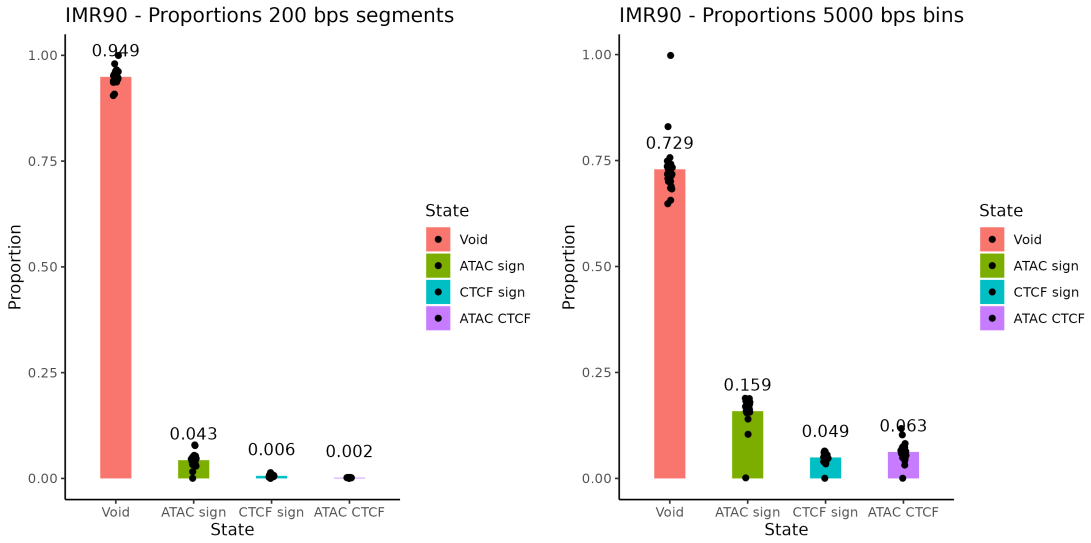


(a) State emission parameters **(b)** State transition parameters **(c)** State enrichment in genome

(d) TSS state population **(e)** TES state population

Figure 6. Results from *ChromHMM* for the IMR90 replicates

The proportions represented in image 7b were obtained for the 5 kbp long bins. In general, the quantities calculated taking into consideration the 200 bp segments were much lower with respect to those calculated with the 5kb bins. The reason is that the first state is in general much more present than the other three, and whenever there are a few occurrences of the rarely present states, the long bins are assigned with a great probability to those states, and as a consequence the proportions tend to smooth out their differences.



(a) Proportions of states in 200 bps segments. The segments were directly found by *ChromHMM*. Each dot represents the proportion relative to a chromosome.

(b) Proportion of states in 5000 bps long bins, obtained through the method described in process 1. Each dot represents the proportion relative to a chromosome.

Figure 7. Proportions found in 200 bp segments and in 5000 bins.

The following image (8) was created by using IGV, a visualization tool[38,44]. After a visual inspection of the results, it was decided to trust the assignment performed. However, some defects become evident while viewing the results: whenever the *ChromHMM* signaled the presence of the 4th state, the relative bin was assigned to it. What happened was that, when the fourth state was found, if the 3th (with both ATAC and CTCF) was not signaled enough, the information about the presence of ATAC peaks was lost. Problems about the precision of the state assignment process couldn't be easily solved and are a direct consequence of the coarse-graining process. Indeed, it's impossible to retain the same amount of information before and after the bin simplification.

Figure 8. IGV snapshot of a portion of the ANPEP region analyzed. The first track reports the alignment results obtained from the ATAC data, the second the CTCF data. The straight traces in the bottom represent the state of the bins in that genomic region. An orange trace represents the state 1 (no ATAC and not CTCF), while instead the purple highlighting signals the third state (with ATAC and CTCF).

5.2 Results obtained while defining the models

The results in this section will be inserted by following the paragraph order written in section 4.2.

- **The computation of the parameters for Coarse Graining**: The values in table 2 and 3 were obtained.

Property	**Formula**	**Value**
c	*const.*	19
v_{FS} (DNA content of a monomer in b.)	*const.*	150+50 bp = 200 bp
b_{FS} (Diameter of a bead in nm)	*const.*	10 nm
lk_{FS} (Kuhn length of the chain in FS)	*const.*	50 nm
ρ_{FS} (Genome density)	*const.*	0.012 bp/nm^3[37]
N_{FS} (Number of monomers to represent the chromosome)	$\frac{\text{DNAcontent}}{v_{FS}} * \text{ncopies}$	30000 mon.
N^k_{FS} (Number of Kuhn lengths of the chain)	$\frac{N_{FS} * b_{FS}}{lk_{FS}}$	6000 k. l.
ρ^k_{FS} (Genome density in Kuhn lengths)	$\frac{\rho_{FS} \cdot b_{FS}}{v_{FS} \cdot \text{lk}_{FS}}$	$1.2e-05$ 1/nm³
L_{FS} (Polymer contour length)	$N_{FS} * b_{FS}$	300000 nm
Le_{FS} (Entanglement length of the chain in nm)	$lk_{FS} * \left(\frac{c}{\rho^k_{FS} * lk^3_{FS}}\right)^2$	8022.22 nm
Number of monomers in a Kuhn length FS	lk_{FS}/b_{FS}	5 mon.
Blk_{FS} (Bead content of a Kuhn length FS)	$(lk_{FS} \cdot b_{FS})/v_{FS}$	2.5 nm^2/bp
Dlk_{FS} (DNA content of a Kuhn length FS)	$(lk_{FS} \cdot v_{FS})/b_{FS}$	1000 bp

Table 2. Parameters calculated for the Fine Scale (FS) model

Property	Formula	Value
c	*const.*	19
v_{CG} (DNA content of a monomer in b.)	*const.*	5000 bp
Dlk_{CG} (DNA content of a Kuhn length CG)	*tuned const.*	33791 bp
ϕ_{CG} (Volumetric density of the chain in the CG model for IMR90 cell-type)	*const.*	0.1
ρ_{CG} (Genome density in bp/nm³)	*const.*	0.012 bp/nm^3
b_{CG} (Diameter of a bead in nm)	$\sqrt{\left(\sqrt{\frac{Dlk_{CG}}{Blk_{FS}}}\right)/\rho_{CG}\cdot\frac{6}{\pi}\cdot\phi_{CG}}$	43.0155 nm
lk_{CG} (Kuhn length of the chain in CG)	$\sqrt{Dlk_{CG}*Blk_{FS}}$	290.65 nm
Number of monomers in a Kuhn length CG	lk_{CG}/b_{CG}	6.75687 mon.
N_{CG} (Number of monomers to represent the chromosome)	$\frac{\text{DNAcontent}}{v_{CG}}*\text{ncopies}$	1200 mon.
side$_{CG}$ (size of the cubic simulation box)	$\frac{(N_{CG}\cdot v_{CG}/\rho_{CG})^{1/3}}{b_{CG}}$	18.4515 nm
N^k_{CG} (Number of Kuhn lengths of the chain)	$(N_{CG}*b_{CG})/lk_{CG}$	177.597 k. l.
ρ^k_{CG} (Genome density in Kuhn lengths bp/nm)	$\frac{\rho_{CG}\cdot b_{CG}}{v_{CG}\cdot \text{lk}_{CG}}$	3.55194e-07 bp/nm
L_{CG} (Polymer contour length)	$N_{CG}*b_{CG}$	51618 nm
Le_{CG} (Entanglement length of the chain in nm)	$lk_{CG}*\left(\frac{c}{\rho^k_{CG}*lk^3_{CG}}\right)^2$	1379.51 nm

Table 3. Parameters calculated for the coarse-grained (CG) model

- **Finding the optimal pressure**: The values of pressure and the respective sizes are plotted in figure 9. To perform this step, just 5 replicates of the 100 total replicates were used for simplicity.

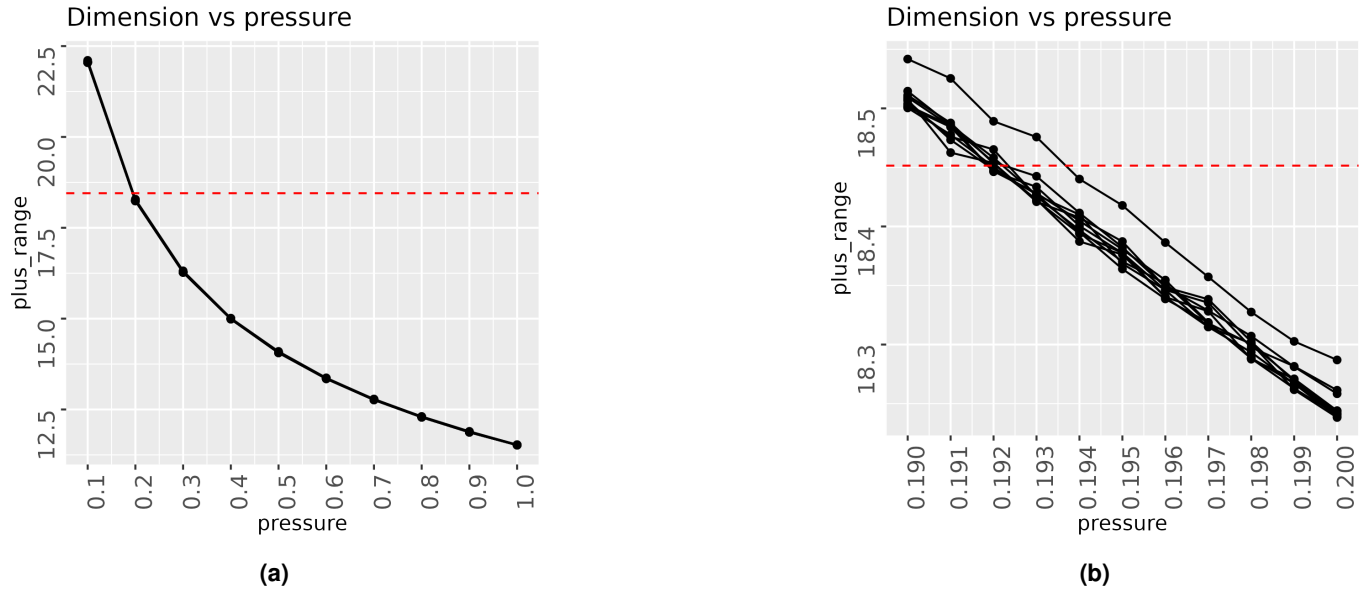
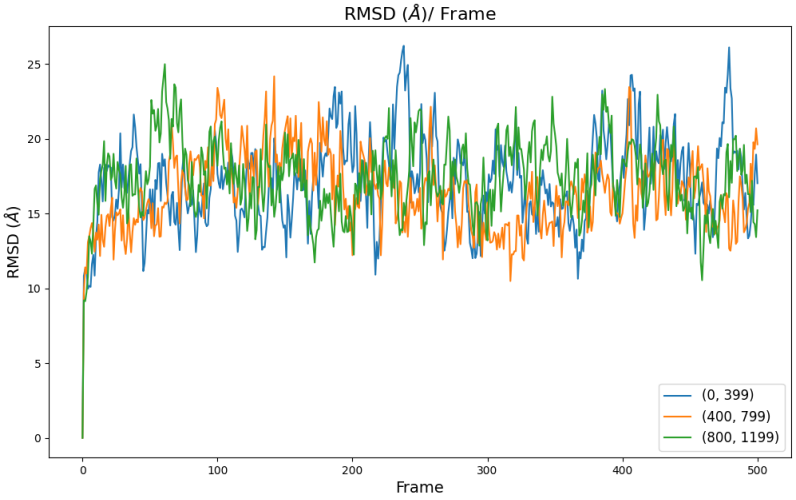


Figure 9. The side estimates obtained for different pressure values are represented in the two graphs. As said in the Methods chapter, the range of trial values for the pressure was shortened by adding more decimals to the quantity. For example, at first the values between 0.1 and 0.9, with a difference of 0.1, were tested (a); after only the region between 0.19 and 0.2 was investigated (b).

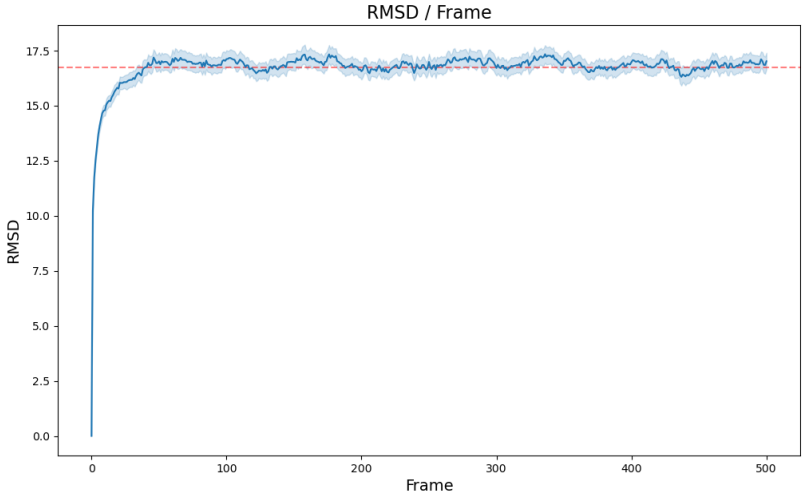
5.3 Trajectory analysis results

Results from the RMSD analysis are reported in figure 10a and 10b. The R_g results are instead plotted in images 11a and 11b. Finally the autocorrelation function graphs are reported in 12a and 12b. Although the RMSD seems to be less variable then the R_g profile, in reality the interval is approximately large the same. By looking to the RMSD profile, it can be argued that the final distension is obtained at the 50$^{\text{th}}$ frame, which corresponds to the $50*50000 = 2,500,000$ step. As stated in 4.2, these evaluations

were done in absence of specific interactions between the beads of the polymer. As a consequence, the equilibrium that was thought to be obtained was due to the reaching of the maximal and most stable extension of the chain.

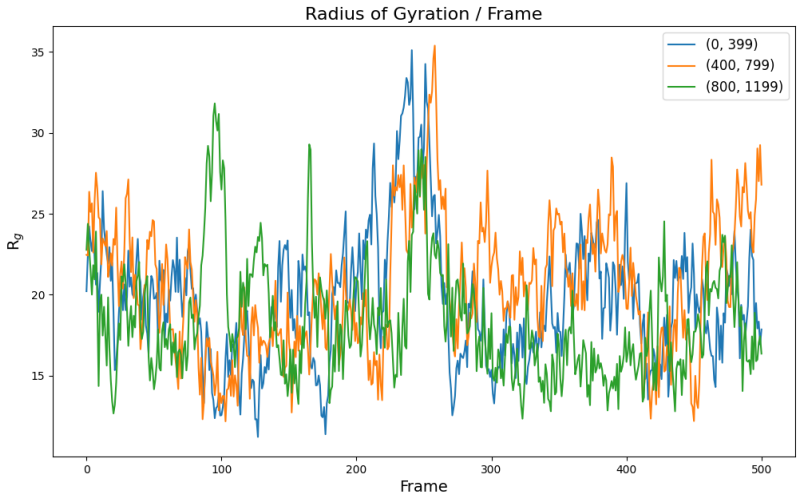


(a) Graph representing the **RMSD** of the chains pertaining to the first replicate. As it is possible to see from the legend, the first, the second and the third chain are represented respectively in blue orange and green.

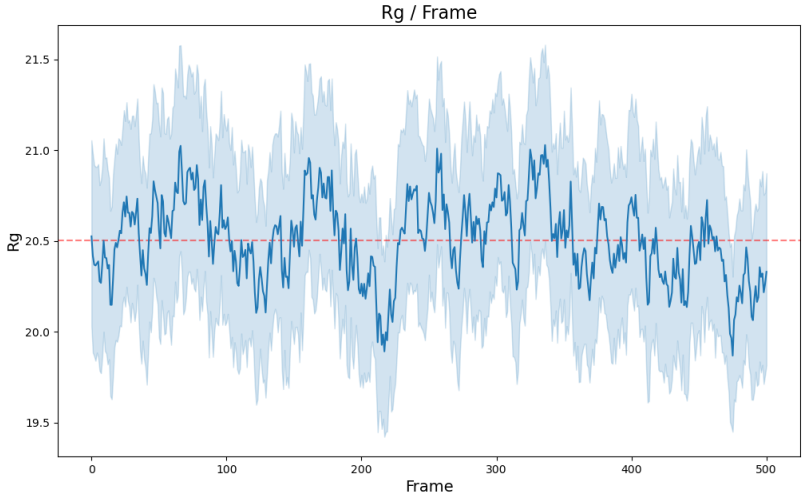


(b) Figure representing the collective behaviour of all the chains of all the 100 replicates. It is possible to observe a plateau at approximately $50 * 50000$ steps. The red dashed line represents the average value.

Figure 10. RMSD profiles

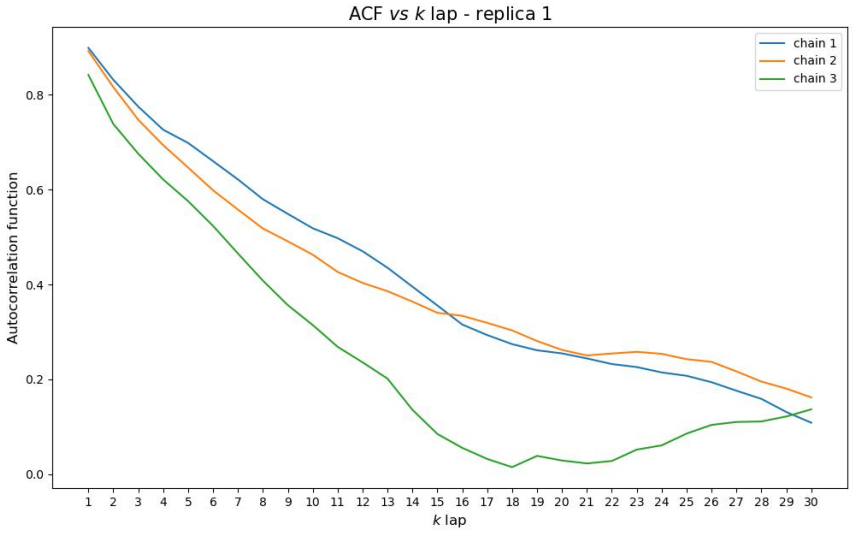


(a) Graph representing the R_g of the chains pertaining to the first replicate. As it is possible to see from the legend, the first, the second and the third chain are represented respectively in blue orange and green.

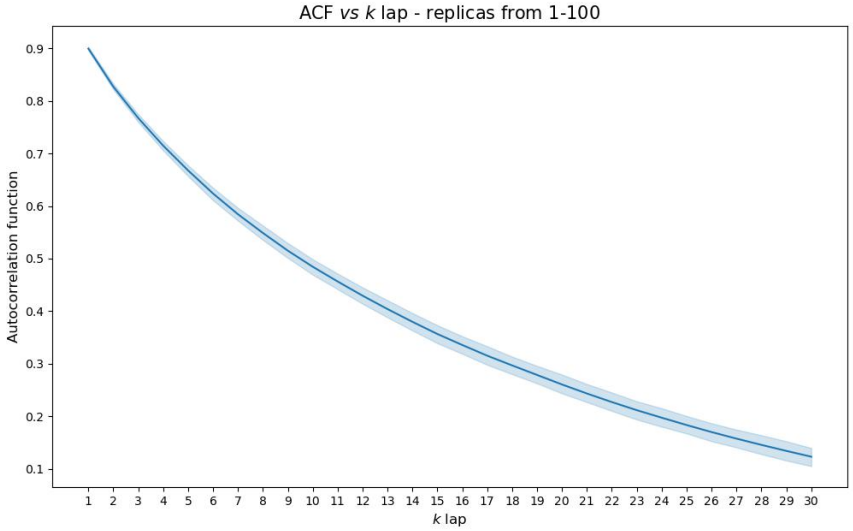


(b) Figure representing the collective behavior of all the chains of all the 100 replicates. The red dashed line represents the average value.

Figure 11. R_g profiles



(a) Graph representing the **autocorrelation function** of the chains pertaining to the first replicate. As it is possible to see from the legend, the first, the second and the third chain are represented respectively in blue orange and green.



(b) Figure representing the collective behavior of all the chains of all the 100 replicates. All the chains of all the replicates were considered independent from each other and taken as singular examples.

Figure 12. Autocorrelation function results

5.4 Results from model selection

The steps described in algorithm **??** were partially followed. A number of configurations for the model were tested, one after the other (loops). To assess each time the best version of a loop, the variations, obtained by tuning the value associated to the same variable, were confronted in terms of the produced SCC quantities. In table 4, the performed loops are listed, and only the characterizations giving the best results are shown.

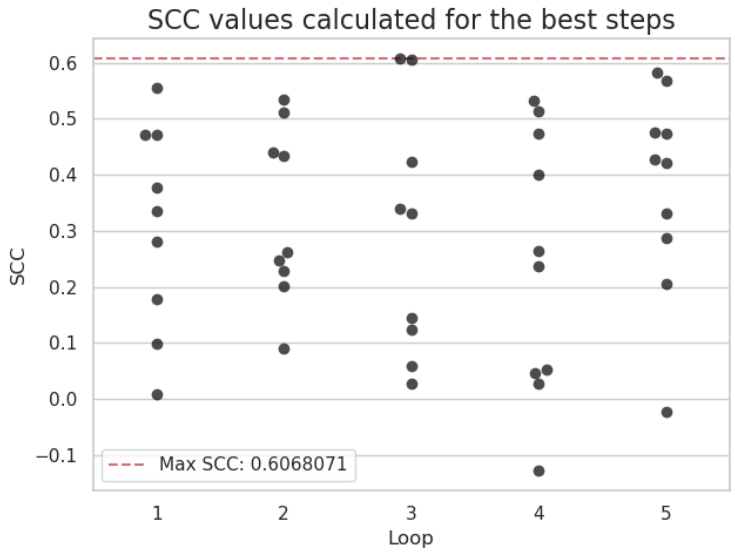


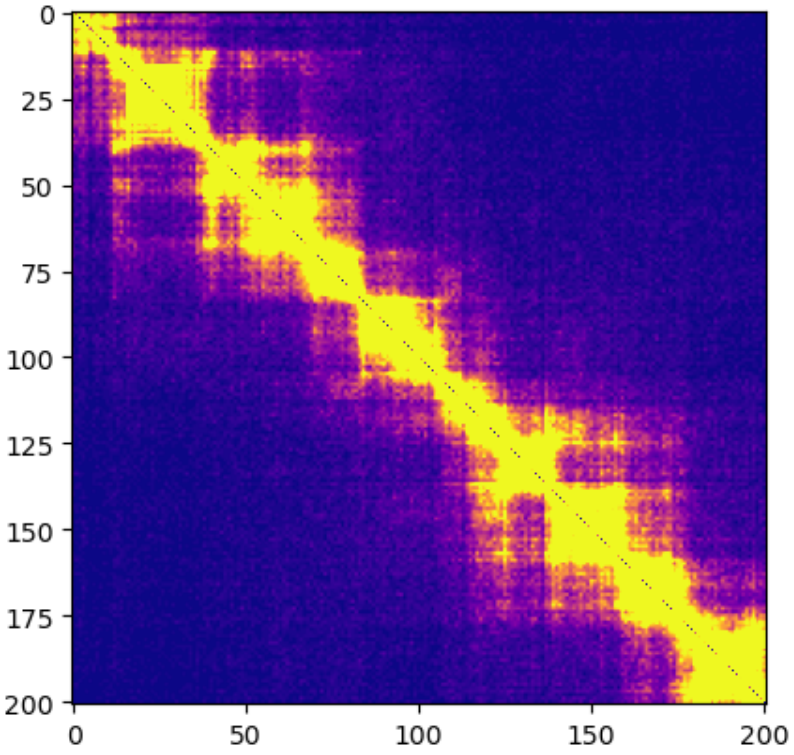
Figure 13. The results obtained from the best steps of the loops. The information reported are also in table 4.

Loop	Model	SCC
Loop 1	$E_{33} = 0.7$	0.5549067
Loop 2	$E_{33} = 0.7$, $E_{22} = 0.8$	0.5352417
Loop 3	$E_{33} = 0.7$, $E_{22} = 0.8$, $E_{44} = 0.6$	**0.6068071**
Loop 4 (removal)	$E_{22} = 0.8$, $E_{44} = 0.6$	0.5324294
Loop 5	$E_{33} = 0.7$, $E_{22} = 0.8$, $E_{44} = 0.6$, $E_{11} = 0.6$	0.5822261

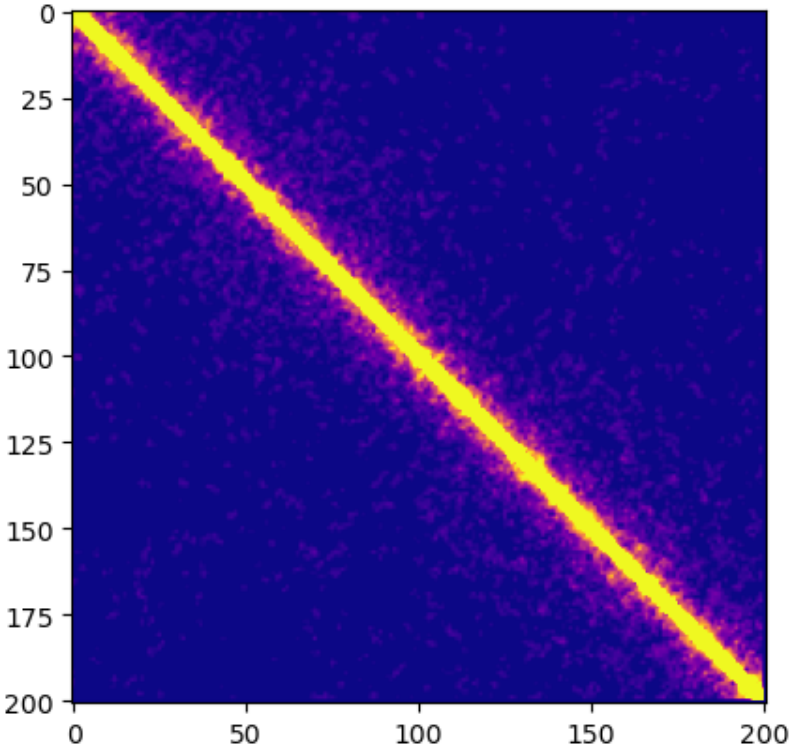
Table 4. Results from the first tested loops. Interactions between beads of the same type are written in the form "$E_{nn} = value$"

In contrast with the initial guesses, state 1 does not appear in the best model found until now. This meant that the "greedy" approach, described in 4.5, should be discarded, and more combinations should be considered in order to find the best combination of parameters. The data in the table could also be visualized in figure 13.

Although this process has not been finished yet, I decided to produce the contact map of the best model tested ($E_{33} = 0.7$, $E_{22} = 0.8$, $E_{44} = 0.6$), to make the first considerations that would aid in the next future (figure 14). Despite the evident difference existing among the two matrices, it is possible to see larger densities of points where the original matrix signals the presence of TADs, and, in general on the boundaries of the squared shapes. The value obtained of SCC, which is a Pearson Correlation coefficient weighted with factors that depend on the population of the distance bends, is quite high, and indicate a moderate correlation. Expectedly, the contacts measured by the simulated model were very lower with respect to those of the original one.



(a) Original matrix



(b) Simulated matrix

Figure 14. Representation of the real and the simulated matrices.

6 CONCLUSIONS

To conclude, several simulations of 2 Mb chromatin regions containing the ANPEP locus were performed. The length was considered as composed by beads with fixed dimension (chapter 4.2), which were assigned to one of the state presented in the *ChromHMM* results (chapters and 2.3 and 4.3) whose input were ATAC-seq and CTCF Chip-Seq data (chapter 4.1). After the tuning of the parameters associated to the interaction potentials generated between beads of the same type, very interesting correlation coefficients between the simulated matrices and the true experimental matrices were found. The maps were compared making use of the SCC coefficient, however, another possible way to see the differences would be to compute the Spearman correlation coefficient. Ideally, it would be interesting to see if there are cases where the two metrics produce different results, and to understand which of them is better in what situations. As a future perspective, it could be considered the extension of the analysis towards new cell-types and/or new *loci*. In particular, we would be interested in investigating the MYC, SOX9, ITG45, MSX2, NT5E genes, and the GM12878 cell-type. Also, the tuning process for the parameters could be improved and automated better. To allow a better comparison with the already present models, the results obtained with the model could be compared to those resulting from other very interesting simulation softwares, such as *Origami* and *Hip-Hop*[23,24].

7 Glossary

TAD	Topological domains
bp	Base Pairs
Bead	The complex formed by the DNA and the histone proteins
TSS	Transcriptional Starting Site
TES	Transcriptional Ending site
FS	Fine Scale
CG	Coarse Grained. It is used to refer to the model with 5,000 bp beads
k. l.	Kuhn length
mon.	Monomer
PL	persistence length
LJ	Lennard-Jones
FENE potential	Finite Extensible Nonlinear Elastic potential
RMSD	Root Mean Square Deviation
Rg	Radius of Gyration
Map	The term is used as a synonym for the term matrix

References

1. Paro, P. D. R., Grossniklaus, P. D. U., Santoro, D. R. & Wutz, P. D. A. Biology of Chromatin. In *Introduction to Epigenetics [Internet]*, DOI: 10.1007/978-3-030-68670-3_1 (Springer, 2021).
2. Liau, B. B. *et al.* Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell* **20**, 233–246.e7, DOI: 10.1016/j.stem.2016.11.003 (2017).
3. Shareef, S. J. *et al.* Extended-representation bisulfite sequencing of gene regulatory elements in multiplexed samples and single cells. *Nat. Biotechnol.* **39**, 1086–1094, DOI: 10.1038/s41587-021-00910-x (2021).
4. Kim, S., Yu, N.-K. & Kaang, B.-K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. & Mol. Medicine* **47**, e166–e166, DOI: 10.1038/emm.2015.33 (2015). CTCF

   conserved in eukariotes, ubiquitous in mammalians. 55000 65000 sites present in mamm.

   zinc finger to bind to DNA

   "~ 50% are intergenic, whereas 35% are intragenic and the rest are promoter proximal" (Kim et al., 2015, p. 1) because of this we suppose that it is able to remodel chromatine

   when to insulator, no enh prom comunication

   –> no spurious signals and inappropriate singlas

   generates topological domains, binds to nuclei to stabilie the chromatine

   binding in regions of housekeeping genes –> maintainance of stable architectures

   cooperates with cohesin

   generates a functional complex

   methylation generates lower binding capability –> loosened efficiency in binding. Not all the zinc fingers behaves in the same way.
5. Hsieh, T.-H. S. *et al.* Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nat. Genet.* **54**, 1919–1932, DOI: 10.1038/s41588-022-01223-8 (2022).
6. Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Sci. (New York, N.Y.)* **357**, eaag0025, DOI: 10.1126/science.aag0025 (2017).
7. Robinson, P. J. J., Fairall, L., Huynh, V. A. T. & Rhodes, D. EM measurements define the dimensions of the "30-nm" chromatin fiber: Evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci.* **103**, 6506–6511, DOI: 10.1073/pnas.0601212103 (2006).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218, DOI: 10.1038/nmeth.2688 (2013).
9. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17**, 1518–1552, DOI: 10.1038/s41596-022-00692-9 (2022). Atac

   –> position histones

   –> TF binding motifs

   –> enhancers.
10. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137, DOI: 10.1186/gb-2008-9-9-r137 (2008).
11. Shah, A. Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system. *Nat. Methods* **6**, ii–iii, DOI: 10.1038/nmeth.f.247 (2009).
12. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, DOI: 10.1038/nature11247 (2012).
13. ATAC-Seq Services - End-to-End Open Chromatin Analysis Service. https://www.activemotif.com/catalog/1233/atac-seq-service.

14. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492, DOI: 10.1038/nprot.2017.124 (2017).
Epigenetic marks –> annotation of genomes –> different cell types
spatial organization –> chromatine states
–> capture known classes of genomic elements
Multivariate HMM for combinatorial presence absence of multiple marks
mark emission probabilities are assumed to be independent.
"The model parameters are learned from the data de novo on the basis of an unsupervised machine-learning procedure that iteratively attempts to maximize the model fit to the data" (Ernst and Kellis, 2017, p. 2478)
partitions the genome at 200 nucleotide intervalds which roughly correspond to the resolution fo a nucleosome
determines the presence or absence of each mark.

15. Chilled House Vibes. Learning Chromatin States from ChIP-seq Data: ChromHMM - Luca Pinello (2015). CHROMHMM
tracks
chromatin states correspnding to combination of chromatin states
Workflow
1- get CHIP sq raw reads for different histone mod
2- align reads to a reference genome
from the fastq file of an illumina machien (Bowtie or BWA)
3- Convert aligned reads to bed format
for each bam file you convert in bed file
bedtools bamtobed -i cell1_mark1.bam > (same folder)/file.bed
4- Create binned and binarized tracks
decide the resolution, default 200 bps
when bin with strong signal –> 1, 0 otherwise
es:
java -mx1200M (allocate memory) -jar ChrommHMM.jar BinarizeBed -b 200 CHROMSIZES/... (sizes of the chromosomes) cellmarkfiletable.txt (one row for each bed file) SAMPLEDATA (output folder)
5- train the model
LearnModel
–> Model + Segmentation
n of chromatin states and genome of reference
to be decided, you have to play with it
6- infer the state
html page
1- Model learned
emission parameters
transition parameters: go from state 1 to 2 exc.
2- Enriched functional categories
3- bed files to see the segmentation
upload on IGV
7- interpretation.

16. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods (San Diego, Calif.)* **72**, 65–75, DOI: 10.1016/j.ymeth.2014.10.031 (2015). HI-C
should be present in chromosomal domains
"Chromosomal domains, i.e., regions displaying preferential contacts within themselves rather than with their flanking regions, have been called with different names (topologically associating domains or TADs, chromosome interacting domains, micro-domains, etc.)" (Matthey-Doret et al., 2022, p. 2)
significance yet to be understood
chimeras are formed, which are composed of fragment of DNA ligated together
all genome wide characterization of interactions
(Wikipedia)
–> messages
1- "Interactions of loci located in different nuclei are less frequent than those in the same nucleus" (Lajoie et al., 2015, p. 15)
2- "Interactions of loci located on different chromosomes are less frequent than those in the same chromosome." (Lajoie et al., 2015, p. 15)
3- "Interactions of loci located far from each other along a chromosome are less frequent than loci that are near each other" (Lajoie et al., 2015, p. 15)
DATA RESOLUTION
6 bp cutting enzyme –> 10^6 fragments
Normally the fragments have a length of 200-500 bps
aggregate in bins
specific for the task
important info: coverage (related to library complexity) –> m aximum of quality (saturation curve)

quality measured through unique chimeric moelcules

READ MAPPING

standard aligner, not the paired-end mode. “One straightforward solution is to map each side of the paired end read separately/independently using a standard alignment procedure” (Lajoie et al., 2015, p. 4)

there are iterative strategies

FRAGMENT ASSIGNMENT

“The mapped read is assigned to a single restriction fragment according to its 5’ mapped position.” (Lajoie et al., 2015, p. 5)

“Mapped read positions should fall close to a restriction site, and no further than the maximal molecule length away.” (Lajoie et al., 2015, p. 5)

FRAGMENT LEVEL FILTERING

“If the read pair maps to the same restriction fragment, it can represent either an un-ligated fragment (“dangling end”) or a ligated, circularized fragment (“self-circle”).” (Lajoie et al., 2015, p. 5)

–> remove, non informative

Remove PCR duplicates

identify 5’ exact same alignment, same paired end sequencing

Filter undigested restriction sites (reads mapped to same strand, small distance between positions)

BINNING

various interval sizes –> smooth noise

bins assigned through midpoint coordinates of fragments

typically 40K to 1MB length

BIN-LEVEL FILTERING

“it is advised to remove any bins (rows/columns) from the dataset that have either very noisy or too low of a signal.” (Lajoie et al., 2015, p. 6). “normally found in genomic regions with low mappability or high repeat content, such as around telomeres and centromeres.” (Lajoie et al., 2015, p. 6)

compare signals from single bin to the signal of all bins

BIASES from (“MCB 182 Lecture 10.4 - Chromatin conformation capture (Hi-C) assays”, 2020)

matrix to represent the number of interactions

normalized with the expected read counts matrix

take the matrix and compute the correlation matrix

correlation when there are a lot of neighbors in the loci

(maintain a similar behavior)

blocks of rows with similar behavior

–> PCA to find the similars (PC1 takes the most of the differences)

there is a bias due to the GC content

bias to the fact that you have to ligate biotine

BALANCING

Oother than the direct correction of the biases listed, there is another algoritghm that can be used

“Sinkhorn-Knoppbalancing algorithm” (Lajoie et al., 2015, p. 7)

“since we are interrogating the entire interaction space in an unbiased manner, each fragment/bin should be observed approximately the same number of timesin the experiment” (Lajoie et al., 2015, p. 7)

–> equilizing the sum ove the rows and the cols

INTERPRETATION OF SIGNALS

differentiate signal from noise, interpret patterns

defects

1- each observation involve a population of cells, not single cells. if a structure has a low HiC signal, it could simply mean that the cells have different structures and are not consistent.

2- genomic compartments found through PCA –> difficult to compare diffent methods results

3- frequencies but not distances

4-“These patterns vary in scale, from genome-wide patterns to point interactions between loci, and in their ubiquity, from constant between different species to condition-specific. Due to the speculative nature of biological interpretation of interaction patterns and the aforementioned complication” (Lajoie et al., 2015, p. 8)

POSSIBLE PATTERNS

CIS-TRANS INTERACTION RATIO

genome-level patterns, square blocks. Higher interaction freq on average of pairs of loci in the same chromosome (cis) than loci which reside on different chromosomes (trans)

due to chromosome territories

way to evaluate qual data

“a noisier experiment will result in a lower ratio between cis and trans interactions.” (Lajoie et al., 2015, p. 9)

“cis/trans ratio in high quality experiments are in the range 40-60 for the human genome.” (Lajoie et al., 2015, p. 9)

(Lajoie et al., 2015, p. 25)

DISTANCE DEPENDENT INTERACTION FREQUENCIES

"In the interaction matrix this pattern appears as a gradual decrease of interaction frequency the further one moves away from the diagonal (decay)." (Lajoie et al., 2015, p. 9)

some physical models can predict the structural variability inprinted in the HIC essay

"pinteraction(x,y)=Z*dist(x,y)-1.5. In fact, this specific decay matches the distance-dependent interaction frequency observed in yeast" (Lajoie et al., 2015, p. 9)

notice that sev. polymer models can produce the same decay func

GENOMIC COMPARTMENTS

position-specific

"Genomic compartments have been found to be correlated with chromatin state, including DNA accessibility, gene density, replication timing, GC content and histone marks (Lieberman-Aiden et al. 2009)" (Lajoie et al., 2015, p. 11)

alternating blocks 1-10 MB

two possible types: A B

"A-type compartments are defined as the euchromatic gene-dense regions while B compartments are defined as gene-poor heterochromatic region" (Lajoie et al., 2015, p. 11)

high plasticity, different in cell-types and biological condition

TOPOLOGICAL DOMAINS (TADs)

sub-Mb distance

small blocks appearing at the diagonl

"it is hypothesized that TADs specify elementary regulatory micro-environments in which promoters interact with local enhancers(" (Lajoie et al., 2015, p. 11)

several methods to find the TADs

"some genomic features such as CTCF and cohesin binding have been shown to be enriched at TAD boundaries" (Lajoie et al., 2015, p. 12)

POINT INTERACTIONS

"Given sufficient resolution, we expect such point interactions to appear as a local enrichment in contact probability." (Lajoie et al., 2015, p. 12)

–> define background model with interaction patterns but also TADs

significant points corrected for multiple testing

could be not easy to distinguish noise

careful interpretation of biological significance

POLYMER MODELLING

problems

"chromatin physics is limited and chromatin structure is much less constrained than protein structure." (Lajoie et al., 2015, p. 13)

chromatin fibers very long

very variable (highly stochastic structure)

two possible ways

consensus structure: a single that is consistent with 3D matrix

–> assume frequency interaction - distance correlation

ensemble

"ypically try to create a set of structures such that either the average distances or the contact probability between every two loci are consistent with the observed interaction frequencies" (Lajoie et al., 2015, p. 14)

–> difficult interpretation.

17. Di Stefano, M., Paulsen, J., Lien, T. G., Hovig, E. & Micheletti, C. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci. Reports* **6**, 35985, DOI: 10.1038/srep35985 (2016).

18. Halverson, J. D., Smrek, J., Kremer, K. & Grosberg, A. Y. From a melt of rings to chromosome territories: The role of topological constraints in genome folding. *Reports on Prog. Physics. Phys. Soc. (Great Britain)* **77**, 022601, DOI: 10.1088/0034-4885/77/2/022601 (2014).

19. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Sci. (New York, N.Y.)* **326**, 289–293, DOI: 10.1126/science.1181369 (2009).

20. Razin, S., Ulianov, S. & Gavrilov, A. 3D Genomics. *Mol. Biol.* **53**, 802–812, DOI: 10.1134/S0026893319060153 (2019).

21. Kremer, K. & Grest, G. S. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *The J. Chem. Phys.* **92**, 5057–5086, DOI: 10.1063/1.458541 (1990).

22. Grosberg, A. J., Chochlov, A. R. & de Gennes, P.-G. *Giant Molecules: Here, There, and Everywhere* (World Scientific, New Jersey, 2011), 2. ed edn.

Dynamics of polymeric fluids

viscous liquid, made of entangled polymer chains. (HIGH viscosity - internal friction larger than in water)

(Grosberg et al., 2011, p. 256) "can we calculate the force f which must be exerted on the upper plate to maintain its motion with speed v and, therefore, to maintain a simple shear flow of the liquid" (Grosberg et al., 2011, p. 256). OBV it depends on the area of the plate, the viscosity, the speed that you want to reach. Also the more are the strata between the plate and the bottom, the lower is the necessary force to exert. nu is coefficient for viscosity.

Viscoelasticity

Universal property

"Depending on how rapidly the external force changes, polymeric fluids can behave either like normal, low molecular weight liquids, albeit very viscous, or like elastic solids" (Grosberg et al., 2011, p. 257). When slow force, viscous response, elastic response if very fast change in force

Reptation Model

"How can it flow? Obviously, a certain chain, if it wants to move, has to slither along a little wiggly corridor inside the bunch, undoing the knots on its way. This sort of picture inspired the theory of reptations (named after the snaky motion of reptile" (Grosberg et al., 2011, p. 258)

Considering a case with fixed obstacles and obstacles that are step by step removed,“The chain will snake out of the “frozen” tube much sooner than it takes the constraints to decay” (Grosberg et al., 2011, p. 259)

-> the motion in fixed obstacles si the main mechanism for hte dynamics of a highly entangled chain

“The Mathematics of a Simple Polymer Coil” (Grosberg et al., 2011, p. 112)

The motion of a random chain could be sinthetized with a random motion. “Since a Brownian particle moves due to collisions with molecules, its path breaks into a sequence of many very short flights and turns. In this sense, a Brownian trajectory is pretty similar to the shape of the polymer chains” (Grosberg et al., 2011, p. 115). An example is a man that follows a random path, taking afterwards that path

difference random and Brownian path

(Grosberg et al., 2011, p. 115)

In the latter case, R is the root mean square displacement frownian

lR can be also rewritten as

(Grosberg et al., 2011, p. 116)

where L is the contour length and l is the analog for the polimer of the Einstein-Smoluchowski result

This type of correlation can be actually found and demostrated as in “Derivation of the “Square Root” Law” (Grosberg et al., 2011, p. 118)

In a polimer

“Interactions between such monomers during their collisions are known as volume interactions. They occur in the bulk of a polymer coil, or inside its “volume” — in contrast to “linear” interactions that hold together the neighboring monomers along the chain” (Grosberg et al., 2011, p. 117)

A random coil generated by a DNA helix would generate an R hich is

(Grosberg et al., 2011, p. 117)

–> not sufficient, more interactions are needed to restrict ulteriorly the volume

persistence length governs the flexibility of a polymer

comes from the Kuhn length, set to be equal to 100 nm in the DNA and indicated with leff in the following equation

(Grosberg et al., 2011, p. 121)

the kuhn length is a sort of memory of the structure along a chain

s << l –> costheta gives 1

s >> l –> costheta gives 0

The kuhn length is twice the persistence length.

23. Buckle, A., Brackley, C. A., Boyle, S., Marenduzzo, D. & Gilbert, N. Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci. *Mol. Cell* **72**, 786–797.e11, DOI: 10.1016/j.molcel.2018.09.016 (2018). H3K27ac epienetic marks + chromatin accesibility (ATAC) + structural anchors (CTCF)

INTRODUCTION

INTERACTION BETWEEN NUCLEOSOMES CHROMAIN BINDING PROTEINS AND STRUCTURAL COMPONENTS

disrupted at transcriptional hotspots, compact in other ocations

previoudly developed TF model

“complexes of TFs and polymerase form enhancer-promoter interactions to organize active regions, while PRC (polycomb repressor complex) or HP1 proteins might arrange inactive and repressed regions” (Buckle et al., 2018, p. 786)

used in combination with LE (loop extrusion) –> not very good results

PAX6

“Pax6 is surrounded by constitutively expressed genes and multiple enhancers, providing a paradigm for complex genetic interactions” (Buckle et al., 2018, p. 786)

(Buckle et al., 2018, p. 787)

RESULTS

PREDICT LOCUS FOLDING ONLY IN SOME CELL LINES.

24. Tan, J. *et al.* Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat. Biotechnol.* **41**, 1140–1150, DOI: 10.1038/s41587-022-01612-8 (2023).

25. Steve Plimpton, Aidan Thompson, Stan Moore, Axel Kohlmeyer & Richard Berger. LAMMPS. https://www.lammps.org/#gsc.tab=0.

26. Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171, DOI: 10.1016/j.cpc.2021.108171 (2022).

27. ATAC-seq (unreplicated) – ENCODE. https://www.encodeproject.org/pipelines/ENCPL344QWT/.

28. Michael Cherry, Jason Buenrostro, Alicia Schep & Will Greenleaf. ATACSeq Pipeline. https://docs.google.com/document/d/1f0Cm4vRyDQDu0bMehHD7P7KOMxTOP-HiNoIvL1VcBt8/edit?usp=sharing&usp=embed_facebook.

29. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: A hitchhiker’s guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22, DOI: 10.1186/s13059-020-1929-3 (2020).

30. Transcription Factor ChIP-seq Data Standards and Processing Pipeline – ENCODE. https://www.encodeproject.org/chip-seq/transcription_factor/.

31. Homo sapiens genome assembly GRCh38 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.

32. UCSC Genome Browser Home. https://genome.ucsc.edu/.

33. Ehler, E., Babiychuk, E. & Draeger, A. Human foetal lung (IMR-90) cells: Myofibroblasts with smooth muscle-like contractile properties. *Cell Motil.* **34**, 288–298, DOI: 10.1002/(SICI)1097-0169(1996)34:4<288::AID-CM4>3.0.CO;2-4 (1996).

34. Ingram, S. P. *et al.* Hi-C implementation of genome structure for in silico models of radiation-induced DNA damage. *PLOS Comput. Biol.* **16**, e1008476, DOI: 10.1371/journal.pcbi.1008476 (2020).

35. Maiser, A. *et al.* Super-resolution in situ analysis of active ribosomal DNA chromatin organization in the nucleolus. *Sci. Reports* **10**, 7462, DOI: 10.1038/s41598-020-64589-x (2020). Nucleolar dimension around 100 microm^3.

36. Conad Corso Lodi 130, 20139 Milano (MI) | Conad. https://www.conad.it/ricerca-negozi.

37. Golkaram, M., Jang, J., Hellander, S., Kosik, K. S. & Petzold, L. R. The Role of Chromatin Density in Cell Population Heterogeneity during Stem Cell Differentiation. *Sci. Reports* **7**, 13307, DOI: 10.1038/s41598-017-13731-3 (2017).

38. Robinson, J. T., Thorvaldsdóttir, H., Turner, D. & Mesirov, J. P. Igv.js: An embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Preprint, Bioinformatics (2020). DOI: 10.1101/2020.05.03.075499.

39. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Python in Science Conference*, 98–105, DOI: 10.25080/Majora-629e541a-00e (Austin, Texas, 2016).

40. Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts (Oxford Univ. Press, Oxford, 2015), reprinted (with corr.) edn.

41. Suma, A., Di Stefano, M. & Micheletti, C. Electric-Field-Driven Trapping of Polyelectrolytes in Needle-like Backfolded States. *Macromolecules* **51**, 4462–4470, DOI: 10.1021/acs.macromol.8b00019 (2018).

42. Lin, D., Sanders, J. & Noble, W. S. HiCRep.py: Fast comparison of Hi-C contact matrices in Python. *Bioinformatics* **37**, 2996–2997, DOI: 10.1093/bioinformatics/btab097 (2021).

43. Yang, T. *et al.* HiCRep: Assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949, DOI: 10.1101/gr.220640.117 (2017). Methods to confront Hi-C maps

Pearson and Spearman

ignore spatial features

SCC

(Yang et al., 2017, p. 1941)

quantifies differences better

". However, Hi-C data have certain unique characteristics, including domain structures (such as topological association domain [TAD] and A/B compartments) and distance dependence, which refers to the fact that the chromatin interaction frequencies between two genomic loci, on average, decrease substantially as their genomic distance increases." (Yang et al., 2017, p. 1939)

The smoothing process is very important: "the SCC analysis on unsmoothed data no longer recapitulates the expected relationships among cell lineages, indicating that the smoothing stage is an indispensable component of HiCRep" (Yang et al., 2017, p. 1942). Also, "our smoothing procedure improves the performance of Pearson- and Spearman-based approaches (Supplemental Fig. S3C,D), confirming its effectiveness." (Yang et al., 2017, p. 1942)

patterns

"One prominent pattern is the strong decay of interaction frequency as genomic distance increases between interaction loci" (Yang et al., 2017, p. 1940)

(Yang et al., 2017, p. 1940) dependence on genomic distance

Pearson correlation coefficient show little difference between HiC of different samples –> no distinguish real biological replicates from unrelated

"Another important pattern of Hi-C data is the domain structure in contact maps." (Yang et al., 2017, p. 1940)

stable across cell types

Pearson and Spearman do not take into consideration domains, they just consider points in the matrix

"As a result, two biological replicates that have highly similar domain structures may have a low Spearman correlation coefficient; conversely, a sample may have a higher Spearman correlation with an unrelated sample than with its biological replicates when the stochastic variation is high." (Yang et al., 2017, p. 1940)

Tests

gives better results when confronting pseudo-replicates, replicates and different cell lines

(Yang et al., 2017, p. 1942) SCC allows the separation of different cells

increases monotonicaly when increasing sequencing at low coverages, remains fixed at a certain value of coverage, reflecting saturation

Sequencing depth when increased increases the reproducibility up to a certain plateau (Yang et al., 2017, p. 1945)

(Yang et al., 2017, p. 1944) the scc monotonically decreases with the sequencing depth, contrarily to other matrices

Allows good anallysis of cell lineages, contrarily to Spearman and pearson

stages of program

1- "The first stage is smoothing the raw contact matrix in order to reduce local noise in the contact map and to make domain structures more visible. The smoothing is accomplished by applying a 2D mean filter, which replaces the read count of each contact in the contact map with the average counts of all contacts in its neighborhood" (Yang et al., 2017, p. 1941)

(Yang et al., 2017, p. 1946)

"o handle this issue, we first smoothed the contact map before assessing reproducibility. Although smoothing will reduce the individual spatial resolution, it can improve the contiguity of the regions with elevated interaction," (Yang et al., 2017, p. 1946)

span size h

tuning, too small –> no enhance domain boundaries; too big –> boundaries blurry

The value for the neighborhood in a matrix of 10 000 resolution is preferentially 20

2- "we stratified the smoothed chromatin interactions according to their genomic distance, and then we applied a novel stratum-adjusted correlation coefficient statistic (SCC)" (Yang et al., 2017, p. 1941)

Pearson coeff are aggregated in a stratum-specific way THORUGH a weighted average, weights from CMH

Cochran-Mantel-Haenszel statistic M^2.

44. Robinson, J. T. Integrative genomics viewer. *c o rresp o n de n ce* **29**, 3 (2011).