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 adiuvate 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 approaches to predict the conformations of the chromatin in time are completelyThe aim of this project is to predict Hi-C matrices of contact by using molecular dynamics simulations

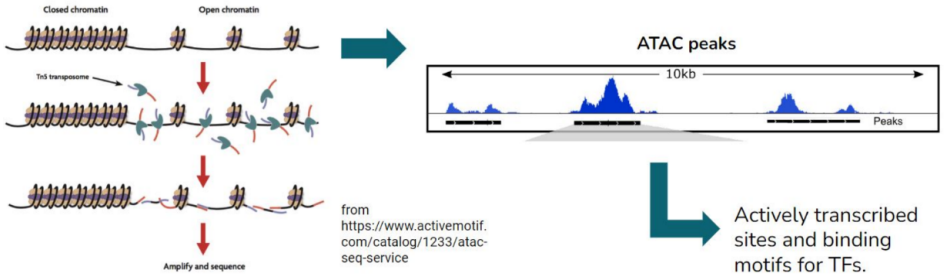
2 INTRODUCTION

2.1 Chromatin as the information center of a cell

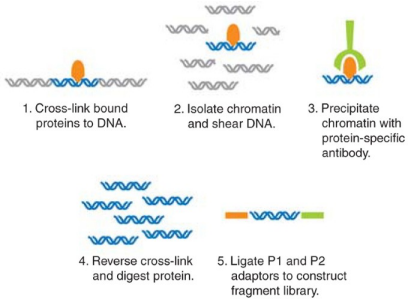
All the living organisms have, inside the nucleus, the largest portion of their DNA, which is the main molecule through which information is passed from the old generation to the daughter cells. Due to the extreme length of the chromosomes, a coordinated assembly of DNA, proteins and RNA, called chromatin, is generated in an ordered and functional manner[1]. The most important proteins used to reach this scope are histones, towards which DNA is wrapped around, forming the nucleosomes. To govern the functioning of the DNA, the histones and the DNA itself are subjected to a variety of modifications. Among those, methylation is the most important on involving the nucleic acid. This type of modification, 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]. On the other hand, CTCF is a protein conserved in eukaryotes and is ubiquitous in mammalians[4]. It contains a Zinc-finger which binds to DNA. The act of binding is performed in cooperation with cohesins, and causes the folding of the chromatin[4,5].

2.2 ATAC-sequencing and CTCF sequencing

ATAC-sequencing is a technology that allows for the identification of open chromatin regions[6,7]. In order to work, it requires the addition of TN-5, a hyper-active transposase. The latter is preloaded with sequencing adapters[7] 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[8]) 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 data used for this experiment, named in table 1, were obtained through a classical ChIP-sequencing, which is a method that combines chromatin immunoprecipitation with DNA sequencing to infer the possible binding sites of DNA-associated proteins (consult the ENCODE entry in table 1 for more information).



(a) Basic concept behind the ATAC-sequencing technique. The images were taken from the following link.

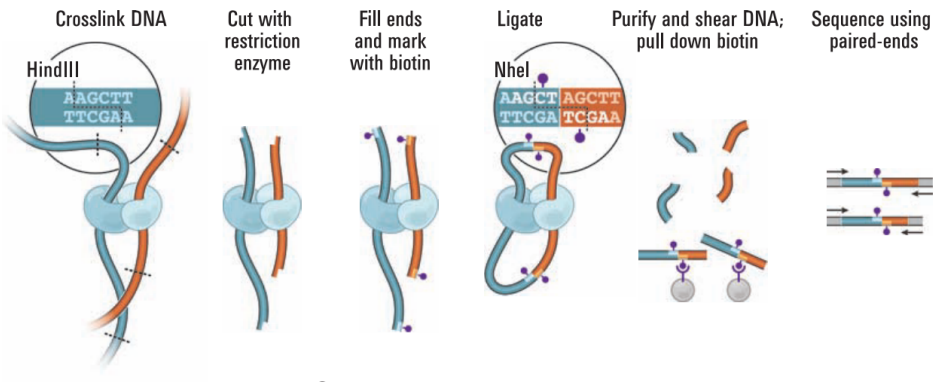


(b) Pipeline to perform a generic ChIP-sequencing analysis. The image was taken from the work of Anjali Shah[9].

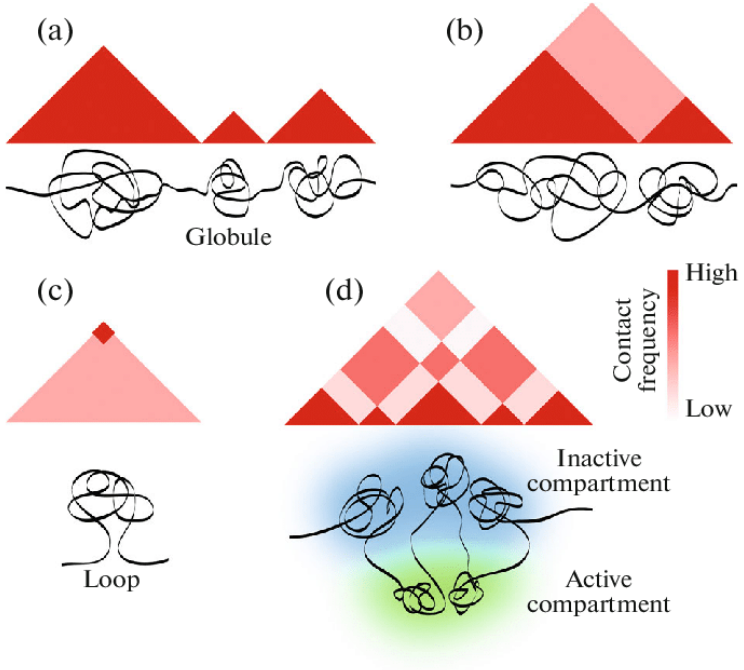
Figure 1. The ATAC and CTCF procedures explanations.

2.3 Hi-C matrices

Hi-C maps are useful tools to detect the interactions occuring inside a genome in analysis. Indeed, it allows to gain insights about the stuctural disposition of chromatin domains, loops and regions.[10]. The experimental procedure is described in figure 2a. Each position, written as a touple (i, j) in a Hi-C matrix, such as the one in figure ... represent the number of contacts between the position i^{th} and j^{th}. 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 we want to recover from the data[10].



(a) Image taken from the following link[11], representing the Hi-C sequencing technique in a schematized way.



(b) Figure representing the contact typologies described in this chapter. The image was taken from the work of Razin and colleagues[12].

Figure 2

The following list of patterns can be found by inspecting an Hi-C matrix[10,13].

1. **Cis/trans interaction ratio**: Those interactions manifest themselves in square blocks. 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 viewed specificity could be a direct consequence of the presence of genomic territories. The ratio between *cis/trans* interactions could be indicative of the quality of the obtained data.

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 distances. Several studies tried to predict this interesting behavior. In particular, it was found that in yeast the probability of interaction could be described with the following equation[10]:

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

3. **Genomic compartments**: Genomic compartments have been found to be correlated with chromatin state, involving DNA accessibility, gene density, replication timing, GC content and histone marks[11]. The compartments can pertain to two categories, A and B, and are found by making a principal components analysis with the matrix generated with Pearson Correlation coefficients (a formula for it is found in chapter 4.6). In general A-type compartments are defined as the euchromatic gene-dense regions, while B compartments are defined as gene-poor heterochromatic regions. The blocks are usually 1-10 Mb long[10]. The positions where they are found depend on the cell type and biological conditions[10].

4. **Topological domains**: Also called TADs, they are regions with sub-Mb length, and can be visually found in Hi-C matrices as larger squared boxes. It is hypothesized that TADs specify elementary regulatory micro-environments in which promoters interact with local enhancers, and, also, some proteins like cohesins and CTCF tend to interact with the genome at the boundaries of the TADs[10].

5. **Point interactions**: Those are interactions occuring among small regions, and involve sequences of a few kb length. Biologically speaking, those points could indicate for example the interaction between enhancers and promoters. When considering point connections, the found value should be compared to the expected number of interactions, and the significance should be computed[10].

2.4 Chromatin Coarse-Graining, chromatin as a polymeric fluid

The Young's modulus (E) of a chain is the extent to which a solid material (or a polymeric fluid, in this case) can be deformed As Robert Hook noticed, the following is valid

$$\sigma = E\frac{\Delta l}{l} \tag{1}$$

Where l represents the length of the chain and Δl the deformation σ[14]

The entanglement length (N_e) corresponds to the Young's modulus that is experimentally found in the plateau region where a force starts to produce irreversible deformations in a chain.

It is also defined as the "the average number of monomer units along the chain between two nearest effective cross-links." and is related to the ability of chains to form knots between each other[14] .

2.4.1 Persistence length of a polymer chain

The persistence length (PL) of a polymer represents the degree of bendability of the chain. During this project, the persistence length is shown in equation 6.

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

With the idea of following a "journey" on the polymer chain, the average angle that you obtain at a contour length s is as in equation 3. In general, the lower is the contour length analyzed with respect to the persistance length, the higher is the probability of having a low degree angle ($\cos\theta \sim 1$). On the contrary, by analyzing larger lengths, it is possible to obtain a wider range of angles. The concept of persistance length is used in the polymer model to compute the angle bending potentials, which is calculated as in equation 7.

$$\langle \cos\theta(s) \rangle = \exp\left(-\frac{s}{l}\right) \tag{3}$$

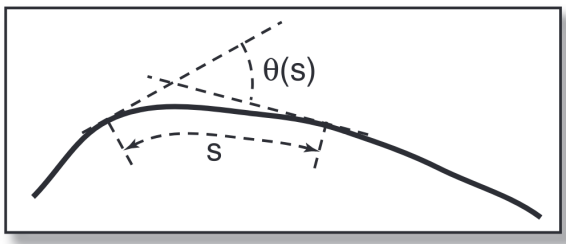


Figure 3. Image taken from Grosberg *et al.* 2011[14]. Angle formed between the extremes of a contour length.

2.5 *ChromHMM* allows the sequential characterization of DNA regions

ChromHMM is a tool which helps in the annotation of genomic DNA by using epigenomic information[15]. 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 other useful information, such as the emission/transition parameters of the states, the abundance of the states at the TSS (Transcriptional Starting Site), at the TES (Transcriptional Ending site), and other relevant portions of the genome (CPG islands, exons, genes)[15,16].

The package works through two functions in particular, which are the following[15]:

1. ***BinarizeBam***: it converts a set of *.bam* files of aligned reads into binarized data files in a specified output directory, which can then be used as input to the *LearnModel* function. When using this command, it has to be specified the bin size, that in the case of this project was set to be 200 bps.

2. ***LearnModel***: it takes a set of binarized data files, learns chromatin state models, and by default produces all the data already mentioned. Additionally, a webpage is created with links to all the files and images created.

The results obtained are shown in chapter 5.1.

3 Aim of the project

The project is part of the thesis, whose aim is to predict matrices of contact of chromatin through the results obtained with molecular coarse-grained simulations of 2 Mb portions of the chromatin. Our focus was in particular placed on a region including and sorrounding the ANPEP *locus*, starting at position $89,000,000$, and finish at the $91,000,000^{\text{th}}$ base. Also, a comparison between this modelling approach and others currently available (such as *Hip-Hop* and *cOrigami*[17,18]) would give us a better idea about the potential application of this approach.
The scope of this work 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[19,20], and some already made codes of Marco di Stefano, researcher at IGH-CNRS (France)

4.1 Data used for the project

The data of CTCF and ATAC for the IMR90 cell line, included in the paper written by Jimin and colleagues in 2023[18], were used for the project (see table 1). It was decided to use the same data of *Origami* to allow a better comparison between its predictions and those produced by our modelling. Both the two replicates, included in the listed ENCODE entries, were considered.

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

Table 1. Table referring to the data used for the analysis. All of them were used for the training of the Origami[18] model. The written entries can be found in the ENCODE database[21].

The ATAC-sequencing data were produced by following the standard ENCODE procedure[22]. In particular, the processes of read trimming, alignment, and filtering were performed making use of the *Bowtie 2*, *Samtools*, *Sambamba*, *Picard* and *cutadapt* softwares[23]. An explanation of the processes could be found in the work made by Feng Y. and colleagues in 2020[24].

When it comes to the ChIP-seq data, again, the standard procedure of ENCODE was used to produce the online available data. An overview can be found at the following link[25]. To sum up, at first the reference genome was indexed with the *BWA*, then, the alignments between the reads and the reference genome (hg38[26]) 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.

4.2 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 the chromosome. 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 making use of the IGV visualization software[27].

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                                          */
```

4.3 The Model

The model creation was done by using and modifying scripts and code written by Marco Di Stefano. The code included in the TADphys unpublished package. The objective of the modelling process was to create a polymer model with a coarse-graining 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 specified, and was obtained by consulting the UCSC genome browser[28]. Secondly, the dimension of the nucleus and the nucleolus of the IMR90 cell had to be specified. By consulting some accessible literature[29–31], I came to the conclusion that they have respectively a dimension of 520 µm and 100 µm.

With the aim of being able to make statistics out of the generated data, it was decided to make 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

All the simulations were performed making use of the *run_lammps* function included in the TADphys package. Three types of potential energies were used to regulate the interactions between beads, and those are presented in the list below:

- **FENE potential:** The FENE potential is a finite extensible nonlinear elastic potential energy and is generally used for polymer models[19,20]. The first term in the equation is attractive, whilst the second term 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 found is the minimum of the LJ potential[20]. Indeed, in that position, $V_{\mathrm{LJ}} = -\varepsilon$. The K presented in the equation 4 is a $\frac{\text{energy}}{\text{distance}^2}$ measure.

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

  The following specifications were made to include the FENE interactions:

  1. $K = 30.0$: It weights the contributions deriving from the attractive part of the equation. It represents the stiffness or strength of the bond.
  2. $R_0(\text{distance unit}) = 1.5$: Maximum extension of the bond. This is the maximum distance at which the bond can be stretched.
  3. $\varepsilon(\text{energy unit}) = 1.0$: This term scales the LJ potential contribution.
  4. $\sigma(\text{distance unit}) = 1.0$: Equilibrium bond length for the LJ potential. This is 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, written as in equation 5.

$$E_{LJ} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \qquad r < r_c \tag{5}$$

  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 are directly correlated with the persistence length (PL), which is calculated as in equation 6. All the quantities are presented in the table 3. The angle related style and values were set with the *angle_style* and *angle_value* of LAMMPS[20].

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

  The angle potentials are given to LAMMPS and calculated by setting the K parameter in equation 7. The larger is the value of K, the larger is the potential generated after the bending of the chain with a specific angle θ.

$$E = K[1 + \cos(\theta)] \tag{7}$$

The steps performed will be described in the following paragraphs:

The computation of the parameters for Coarse Graining: Both parameters for the fine-scale model (table 2) and the CG model (table 3) were calculated. The number of bp wrapping around a bead in the FS method was considered to be 150, while instead the linker portion was considered 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 ($\rho_{FS} = \rho_{CG} = 0.012\ \mathrm{bp/nm^3}$) are 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 human nuclei[32]. The beads and bonds have all the same length in the FS and the CG models, consequently, the contour length (which represents the maximal length of the polymer chain) is exactly the product between the number of beads and their size, in the FS and in the CG models. The DNA content of the analyzed region was set to be equal to $2,000,000$ bp. 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 (v_{CG}).

Generation of the initial conformation rosettes: Once found the coarse graining parameters, rosettes for all the replicates were built, with a radius of 12.0 nm inside a cubic box whose edge length is 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 you make the calculation $2,000,000/5,000 = 400$.

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 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. 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 $1 * 10^{-6}$, a maximal number of iterations and evaluations of 100000 steps. The process for finding the optimal pressure parameter was long 1000 steps with a duration of 0.001 ps. The optimal pressure was attested 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 (not after a minimization phase). 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 5000 steps. At the end, the trajectories were collected and analyzed by computing the *RMSD*, *Rg* and the autocorrelation function as described in chapter 4.4. For the sake of simplicity, saving memory and computational cost, 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*[15,16] , whose input were CTCF and ATAC-seq datasets (chapter 4.1), were used to define the identity of the beads. Then, the best attraction parameters were selected in the iterative manner described by the 3 algorithm. To add the attraction potentials between the beads, new Lennard-Jones potentials were added (in the form expressed in equation 5). In particular, the cutoff distance of the potential r_0 was set to be equal to: $r_0 = r_{\mathrm{cutoff}} = \sigma * 2.5$, where σ corresponds to the sum of the *radii* of the interacting particles (set in all the cases to 0.5 nm). The trial value of the attraction parameter associated to the interaction was assigned to the ε variable, present in the LJ potential.

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 is established whenever two coarse-grained particles are found at a distance lower then 100 nm.

4.4 Trajectories analysis

The analysis of the trajectories was done by taking together and considering as independent the results obtained with each single chain of each replicate. For this reason, it was decided to put together all the results and produce the collective plots represented in figures 8b, 9b and 10b.

Three types of analysis were performed (all the following terms are explained in the Glossary):

1. ***RMSD***: The Root Mean Square Deviation (RMSD) is calculated by using the *MDanalysis* package[33] (*rmsd* in *MDAnalysis.analysis.rms*) and is calculated as follows:

$$RMSD = \rho(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i \left(\vec{x}_i(t) - \vec{x}_i^{REF}\right)^2} \tag{8}$$

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[33] . 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 8a and 8b.

2. R_g: The Radius of Gyration was calculated as written in equation 9 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[33,34] . The results are written in graph 9a and 9b.

$$R_g = \sqrt{\frac{\sum_i m_i \vec{r}_i^{\,2}}{\sum_i m_i}} \tag{9}$$

3. ***Autocorrelation function:*** The autocorrelation function can be written as shown in equation 10[35]. The results are shown in graphs 10a and 10b.

$$r_k = \frac{C_k}{C_0} = \frac{\frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})(A_{t+k} - \bar{A})}{\frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})^2} \tag{10}$$

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 4.3 about the simulations, the attraction potentials are sequentially added to the model in order to improve the predictions. The SCC metric 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 2 and 3. 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, I could 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 2. If that assumption is not accepted, then either you test all the possible configurations, either you find a better way to test the generated models. The algorithm 3 is a stepwise solution which allows to solve partially the problem at the cost of more simulations to perform.

Algorithm 2: Greedy matrix comparison

Result: Best performing greedy model
forall *attraction parameter* **do**
    Construct models by adding the most present attraction parameter to the $(n-1)^{\text{th}}$ step configuration ;
    Compute SCC of the model with respect to the reference ranging among a list of possible values;
    Select the value of the parameter which gives the best results;
    Add that attraction with that coefficient;
return *Best greedy model*;

Algorithm 3: Step-wise process for matrix comparison

Result: Best model
$n = 0$;
Continue = False;
while *Continue == True* **do**
  Continue = False;
  Construct models by adding the most present attraction parameter to the $(n-1)^{\text{th}}$ step configuration ;
  Compute SCC of the model with respect to the reference ranging among a list of possible values;
  Select the value of the parameter which gives the best results;
  if *addition gives better results* **then**
    Add that attraction with that coefficient;
    Continue = True
  forall *state in model* **do**
    Remove that state from the model;
    Vary the value associated to the state with the highest frequency among those remaining;
    Compute SCC of each model with respec to the reference ranging among a list of possible values
  Select the reduced model which gives the best results;
  if *removal gives better results* **then**
    Perform the reduction;
    Continue = True
  $n = n+1$
return *Best model*;
`/* n is the step number */`

Another possible way to compare the matrices 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 cases

4.6 the Stratum Adjusted Correlation Coefficient (SCC) metric

The SCC metric is described in the paper written by Yang and colleagues in 2017[36,37]. It can quantify the similarity between an Hi-C matrix and another. In general, the most common techniques to use in these situations is 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) and A/B compartments) and distance dependence. 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[36,37] .

The SCC metric could be seen as a weighted Pearson coefficient, as written in equation 11.

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;
$r_{2k} = \sqrt{\text{var}(X_k) \cdot \text{var}(Y_k)}$	$k \in K$	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

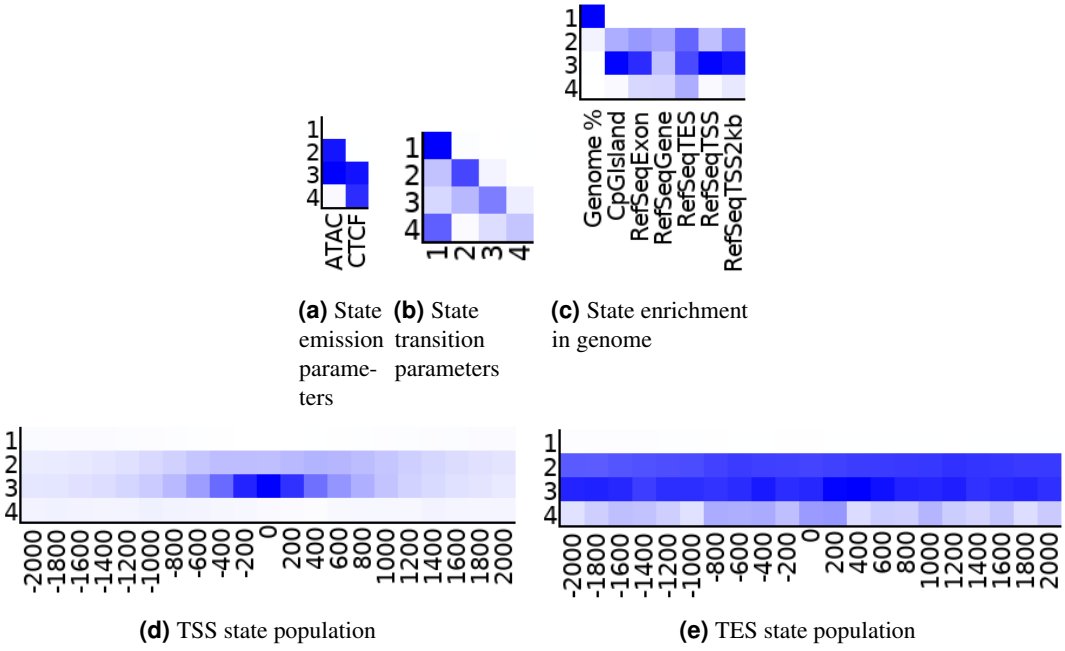
$$\rho_s = \sum_{k=1}^{K} \left(\frac{N_k r_{2k}}{\sum_{k=1}^{K} N_k r_{2k}} \right) \rho_k \tag{11}$$

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[26] . Results are shown in image 4; by taking a look to its subfigures, the following considerations could be done:

1. Clear absence or presence of ATAC and CTCF signals could be detected in figure 4a. 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 4b)
3. State 2 (with ATAC) and 3 (with ATAC and CTCF) tend to localize in CpG islands, exons and Transcriptional Starting Sites (figures 4c, 4d, 4e)

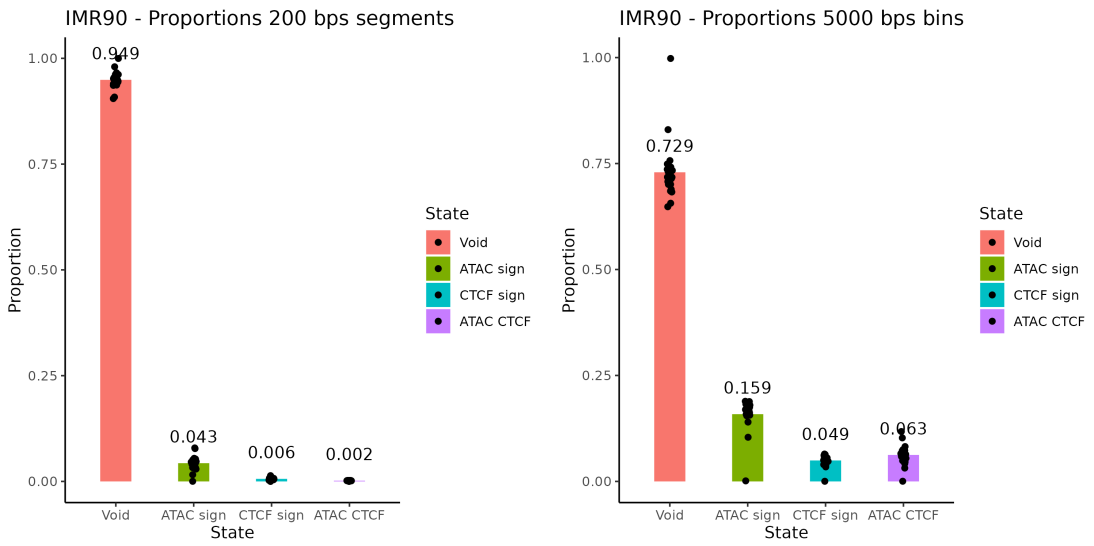


(a) State emission parameters **(b)** State transition parameters **(c)** State enrichment in genome

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

Figure 4. Results from *ChromHMM* for the IMR90 replicates

The proportions represented in image 5b were obtained. In general, the proportions calculated taking into consideration the segments are much lower with respect to those calculated with the bins. The reason is that the first state is in general much more present than the other three. The ratios are larger with 5 kbp bins as in fact whenever there are a few occurrences of the rarely present states, those are assigned with a great probability and convert a great number of segments which are not anymore assigned to state 1.



(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. Each dot represents the proportion relative to a chromosome.

Figure 5. Proportions found in 200 bp segments and in 5000 bins.

The following image 6 was created by using IGV, a visualization tool[27,38] . 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* signals the presence of the 4th state, the relative bin is assigned to it. What happens is that, when the fourth state is found, if the 3th (with both ATAC and CTCF) is not signaled enough, the information about the presence of ATAC peaks is 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.

Figure 6. 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

5.2 Results obtained while defining the models

The results in this section will be inserted by following the paragraph order written in section 4.3.

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

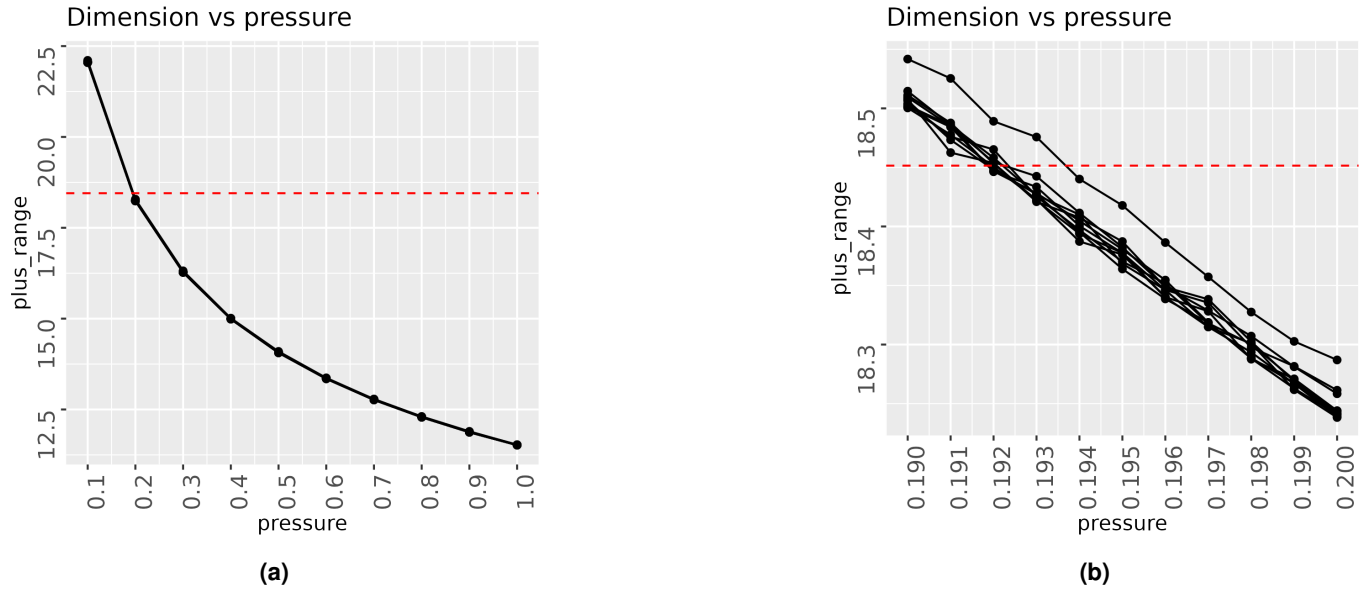
Property	Formula	Value
c	*const.*	19
v_{FS} (DNA content of a monomer in b.)	*const.*	150+50 bp = 200 bp
b_{FS} (Diameter of a bead in nm)	*const.*	10 nm
lk_{FS} (Kuhn length of the chain in FS)	*const.*	50 nm
ρ_{FS} (Genome density)	*const.*	0.012 bp/nm^3[32]
N_{FS} (Number of monomers to represent 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 7. To perform this step, just 5 replicates of the 100 total replicates were used for simplicity.

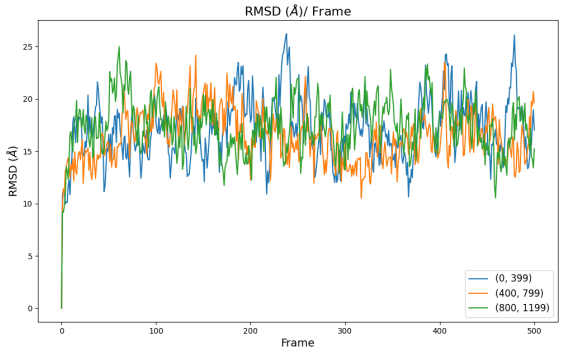


(a) **(b)**

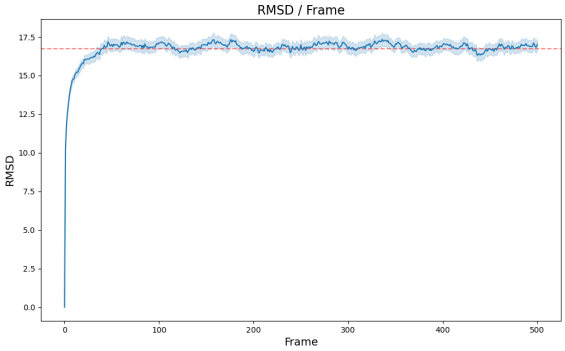
Figure 7. The side estimates obtained for different pressure values are represented in the two graphs. As said in the Methods chapter, the range of trial values for the pressure was shortened by adding more decimals to the quantity. For example, at first the values between 0.1 and 0.9, with a difference 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 8a and 8b. The R_g results are instead plotted in images 9a and 9b. Finally the autocorrelation function graphs are reported in 10a and 10b. Although the RMSD seems to be less variable and 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^{th} frame, which corresponds to the $50 * 50000 = 2,500,000$ step. I insist in saying that, right now, there are no interactions between the beads of the polymer. As a consequence, the equilibrium that we think we obtained is due to the reaching of the maximal and most stable extension of the chain.

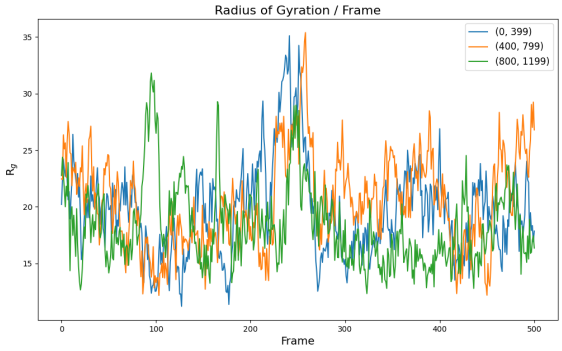


(a) Graph representing the **RMSD** of the chains pertaining to the first replicate. As it is possible to see from the legend, the first, the second and the third chain are represented respectively in blue orange and green.

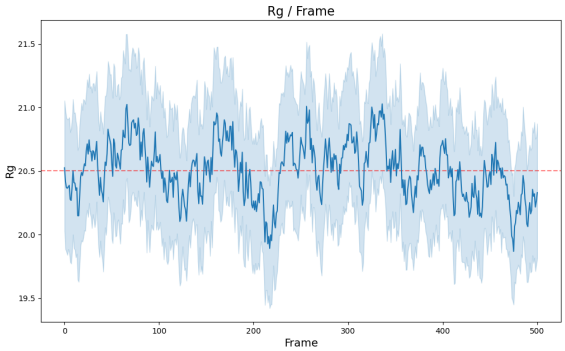


(b) Figure representing the collective behaviour of all the chains of all the 100 replicates. It is possible to observe a plateau at approximately $50 * 50000$ steps. The red dashed line represents the average value.

Figure 8. RMSD profiles

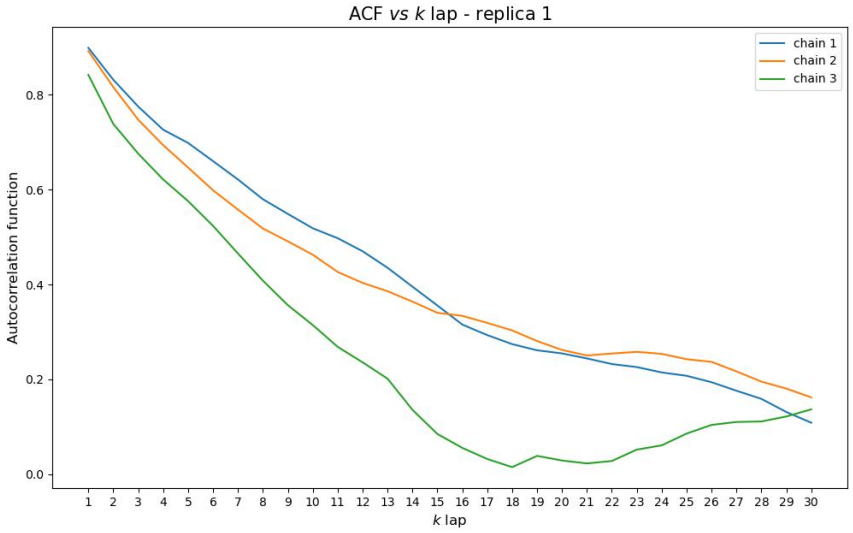


(a) Graph representing the R_g of the chains pertaining to the first replicate. As it is possible to see from the legend, the first, the second and the third chain are represented respectively in blue orange and green.

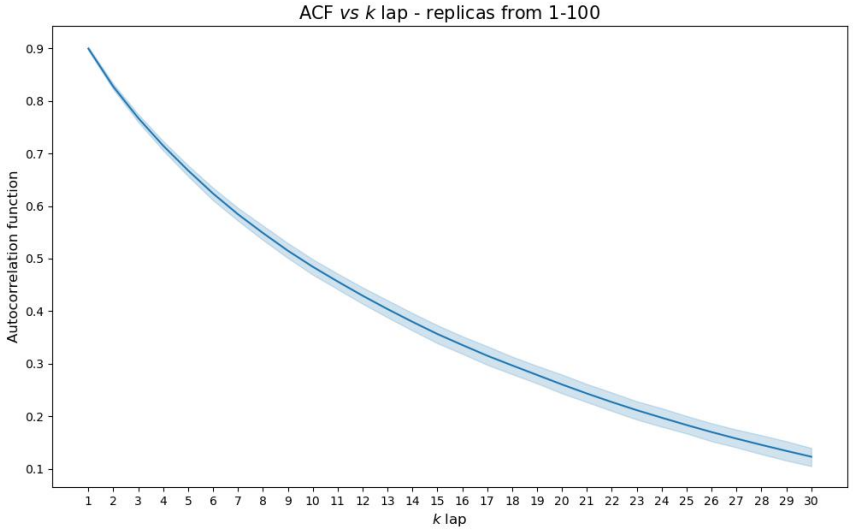


(b) Figure representing the collective behavior of all the chains of all the 100 replicates. The red dashed line represents the average value.

Figure 9. 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 10. Autocorrelation function results

5.4 Results from model selection

6 CONCLUSIONS

To conclude, several simulations of 2 Mb chromatinic regions containing the ANPEP locus were performed. The segment was considered as composed by beads with fixed dimension (chapter 4.3), which were assigned to one of the state presented in the *ChromHMM* results (chapters and 2.5 and 4.2) 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 make the comparison 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 of SCC and Spierman correlation coefficient, and to understand which of them is better in what cases.

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 interesing simulation softwares, such as *Origami* and *Hip-Hop*[17,18].

7 Glossary

Bead	The complex formed by the DNA and the histone proteins
TSS	Transcriptional Starting Site
TES	Transcriptional Ending site
FS	Fine Scale
CG	Coarse Graining
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 Med* **47**, e166–e166, DOI: 10.1038/emm.2015.33 (2015).
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. 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).
7. 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).
8. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137, DOI: 10.1186/gb-2008-9-9-r137 (2008).
9. Shah, A. Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system. *Nat Methods* **6**, ii–iii, DOI: 10.1038/nmeth.f.247 (2009).
10. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods* **72**, 65–75, DOI: 10.1016/j.ymeth.2014.10.031 (2015).
11. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* **326**, 289–293, DOI: 10.1126/science.1181369 (2009).
12. Razin, S., Ulianov, S. & Gavrilov, A. 3D Genomics. *Mol. Biol.* **53**, 802–812, DOI: 10.1134/S0026893319060153 (2019).
13. 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 Rep* **6**, 35985, DOI: 10.1038/srep35985 (2016).
14. Grosberg, A. J., Chochlov, A. R. & de Gennes, P.-G. *Giant Molecules: Here, There, and Everywhere* (World Scientific, New Jersey, 2011), 2. ed edn.
15. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**, 2478–2492, DOI: 10.1038/nprot.2017.124 (2017).
16. Chilled House Vibes. Learning Chromatin States from ChIP-seq Data: ChromHMM - Luca Pinello (2015).
17. 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).
18. 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).
19. Steve Plimpton, Aidan Thompson, Stan Moore, Axel Kohlmeyer & Richard Berger. LAMMPS. https://www.lammps.org/#gsc.tab=0.
20. 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).
21. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, DOI: 10.1038/nature11247 (2012).
22. ATAC-seq (unreplicated) – ENCODE. https://www.encodeproject.org/pipelines/ENCPL344QWT/.
23. Michael Cherry, Jason Buenrostro, Alicia Schep & Will Greenleaf. ATACSeq Pipeline. https://docs.google.com/document/d/1f0Cm4vRyDQDu0bMehHD7P7KOMxTOP-HiNoIvL1VcBt8/edit?usp=sharing&usp=embed_facebook.
24. 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).
25. Transcription Factor ChIP-seq Data Standards and Processing Pipeline – ENCODE. https://www.encodeproject.org/chip-seq/transcription_factor/.
26. Homo sapiens genome assembly GRCh38 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.

27. 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.

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

29. 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).

30. 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).

31. Maiser, A. *et al.* Super-resolution in situ analysis of active ribosomal DNA chromatin organization in the nucleolus. *Sci Rep* **10**, 7462, DOI: 10.1038/s41598-020-64589-x (2020).

32. 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 Rep* **7**, 13307, DOI: 10.1038/s41598-017-13731-3 (2017).

33. 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).

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

35. 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).

36. 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).

37. 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).

38. Robinson, J. T. Integrative genomics viewer. *C O Rresp O N N Ce* **29**, 3 (2011).