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

Maurizio Gilioli

Contents

1 ABSTRACT

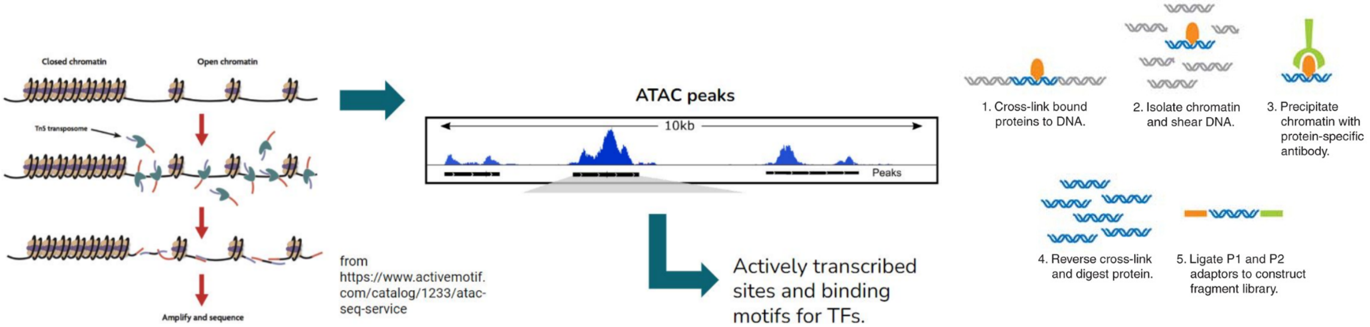
2 INTRODUCTION

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]. On the other hand, CTCF is a protein conserved in eukariotes and is ubiquitous in mammalians[3]. It contains a Zync-finger which binds to DNA. The act of binding is performed in cooperation with cohesins, and causes the folding of the chromatin[3,4].

2.2 ATAC-sequencing and CTCF sequencing

ATAC-sequencing is a technology that allows for the identification of open chromatin regions[5,6]. In order to work, it requires the addition of TN-5, a hyper-active transposase. The latter is preloaded with sequencing adapters[6] 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[7]) 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).


Closed chromatin
Open chromatin
Tn5 transposome
Amplify and sequence
ATAC peaks
10kb
Peaks
from
https://www.activemotif.
com/catalog/1233/atac-
seq-service
Actively transcribed
sites and binding
motifs for TFs.
1. Cross-link bound proteins to DNA.
2. Isolate chromatin and shear DNA.
3. Precipitate chromatin with protein-specific antibody.
4. Reverse cross-link and digest protein.
5. Ligate P1 and P2 adaptors to construct fragment library.

(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[8].

Figure 1

2.3 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[?] .

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[?] .

2.3.1 *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.

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

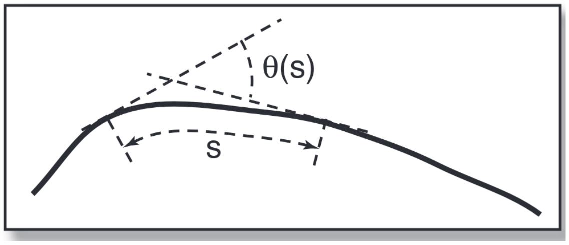


Figure 2. Image taken from Grosberg *et al.* 2011[?]. Angle formed between the extremes of a contour length.

2.4 *ChromHMM* allows the sequential characterization of DNA regions

ChromHMM is a tool which helps in the annotation of genomic DNA by using epigenomic information[9]. 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)[9,10].

The package works through two functions in particular, which are the following[9]:

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*[11, 12]) 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[?, 13], 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[12], 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.

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[12] model. The written entries can be found in the ENCODE database[14].

The ATAC-sequencing data were produced by following the standard ENCODE procedure[15]. In particular, the processes of read trimming, alignment, and filtering were performed making use of the *Bowtie 2*, *Samtools*, *Sambamba*, *Picard* and *cutadapt* sofwares[?]. An explanation of the processes could be found in the work made by Feng Y. and colleagues in 2020[16].

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[17]. To sum up, at first the reference genome was indexed with the *BWA*, then, the alignments between the reads and the reference genome (hg38[18]) 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[19].

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[20]. Secondly, the dimension of the nucleus and the nucleolus of the IMR90 cell had to be specified. By consulting some accessible literature[21–23], 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 **??**.

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[13,24]. 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 $2^{1/6}\sigma$, were the value found is the minimum of the LJ potential[24]. 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 for the FENE interactions:

  1. 30.0 N: Maximum force the bond can withstand. It represents the stiffness or strength of the bond.
  2. 1.5: Maximum extension of the bond. This is the maximum distance at which the bond can be stretched.
  3. 1.0: Equilibrium bond length. This is the ideal or equilibrium distance between the bonded particles.
  4. 1.0: Bond force constant. It affects how quickly the potential energy increases as the bond length deviates from the equilibrium length.

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

  $$\begin{aligned} E_{LJ} &= 4\varepsilon\left[\left(\frac{r_0}{r}\right)^{12} - \left(\frac{r_0}{r}\right)^{6}\right] && r < r_c \\ &= 4\varepsilon\left[\left(\frac{2^{1/6}\sigma}{r}\right)^{12} - \left(\frac{2^{1/6}\sigma}{r}\right)^{6}\right] && r < r_c \end{aligned} \tag{5}$$

  Three parameters are set:

  1. ε (energy unit): 1.0, Note that σ is defined in the LJ formula as the zero-crossing distance for the potential, not as the energy minimum at $r_0 = 2^{1/6}\sigma$.
  2. σ (distance unit): 1.0
  3. LJ cutoff (distance unit): 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.

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

  The angle potentials are given to LAMMPS and calculated setting the K parameter in equation 7. The larger is the value of K, the larger is in general the potential associated to the angles.

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

. All the simulations were performed making use of periodic boundaries (chapter
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 ??) and the CG model (table ??) 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 ??). The thickness of a bead (of a nucleosome) was taken as equal to 10 nm, while instead the default Kuhn length was set to 50 nms. The genome densities ($\rho_{FS} = \rho_{CG} = 0.012$ bp$/$nm^3) are imposed to be the same for the FS and the CG model.

. The beads and bonds have all the same length, consequently, the contour length is exactly the product between the number of beads and the size of the beads in the FS and in the CG model. The DNA content was 2000000 bp. To find the parameters for the CG model, the DNA content of the Kuhn segments (Dlk_{CG}) were tuned to match the desired value of DNA content in CG beads (v_{CG}).

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
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

Generation of the initial conformation rosettes: Once found the coarse graining parameters, rosettes for 100 replicates were built, with a radius of 12.0 nm inside a cubic box of 300 nm. The particle radius 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.

Finding the optimal pressure: Once the rosettes were made, a decompaction was performed starting from the compacted rosettes. A range of values of pressure was tested from 0.1 to 1 with steps every 0.1. 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 was long 1000 steps with a duration of 0.001 ps. and the final structure was taken as the decompressed one. The optimal pressure was attested at 0.192

Decompaction and relaxation: Once the optimal value for the pressure was found (0.192) for each replicate, other two simulations respectively $5,000,000$ and $25,000,000$ steps long were performed (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 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, the collection was accomplished by capturing one frame for every 50,000 steps.

Computing matrices of contact: Once defined the step at which the simulations were considered to be at the equilibrium, some dictionaries produced with *ChromHMM*[9, 10] , by using CTCF and ATAC-seq data, were used to define the identity of the beads. During the next steps, we will try to select the best attraction parameters; the stepwise algorithm 3 is being currently tested. All the beads were considered to have a radius of 0.5 long. To quantify attraction potentials, a new Lennard-Jones potential was added (as defined in equation 5). In particular, the cutoff distance r_0 was set to be equal to:

$$r_0 = r_{\text{cutoff}} = \sigma * 2.5$$

Once the interaction parameters are set, other 1 second long simulations are performed with the intention of generating contact maps. Only the interactions occurring intra-chain were considered.

4.4 Trajectories analysis

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

1. ***RMSD***: The Root Mean Square Deviation is calculated by using the *MDanalysis* package[?] 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 their specific replicate). This type of alignment was done making use of the *AlignTraj* function[?] . In general, the smaller is the difference between two structures, the lower is the value of RMSD. The results are written in graph 6a and 6b.

2. R_g: The Radius of Gyration computed by *MDanalysis* was calculated as written in equation 9. 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[?, ?] . The results are written in graph 7a and 7b.

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

3. ***Autocorrelation function:*** The autocorrelation function can be written as shown in equation 10[25] . The results are shown in graphs 8a and 8b.

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

   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 4.5.1).

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

4.5.1 the SCC metric

The SCC metric is described in the paper written by Yang and colleagues in 2017[26,27] . It is a Stratum Adjusted Correlation Coefficient and 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[26,27] .

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

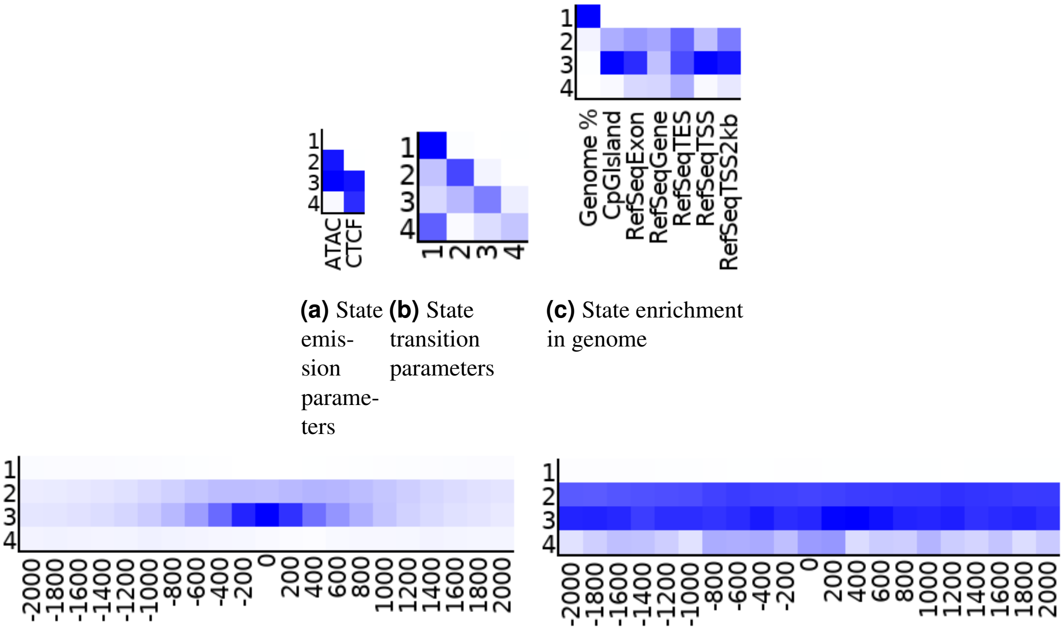
$$\rho_s = \sum_{k=1}^{K} \left(\frac{N_k r_{2k}}{\sum_{k=1}^{K} N_k r_{2k}} \right) \rho_k \tag{11}$$

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[18] . Results are shown in image 3; 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 3a. 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 3b)
3. State 2 (with ATAC) and 3 (with ATAC and CTCF) tend to localize in CpG islands, exons and Transcriptional Starting Sites (figures 3c, 3d, 3e)



(a) State emission parameters

(b) State transition parameters

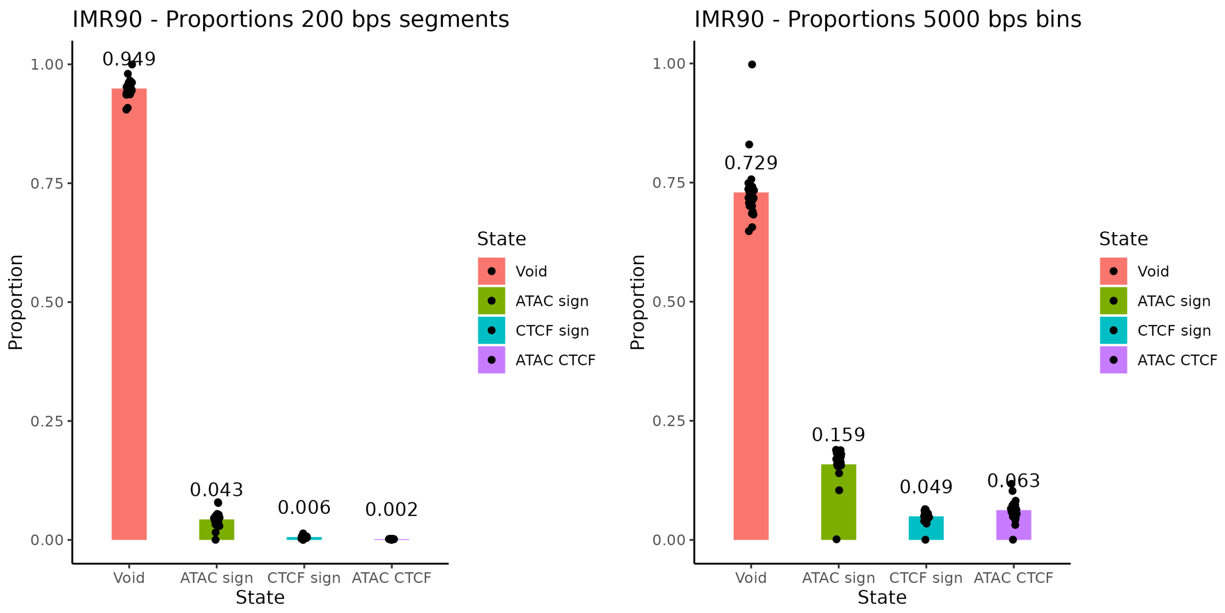
(c) State enrichment in genome

(d) TSS state population

(e) TES state population

Figure 3. Results from ChromHMM for the IMR90 replicates

The proportions represented in image 4b 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 4. Proportions found in 200 bp segments and in 5000 bins.

The following image 5 was created by using IGV, a visualization tool[?, ?] . 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 enoughly signaled, 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.

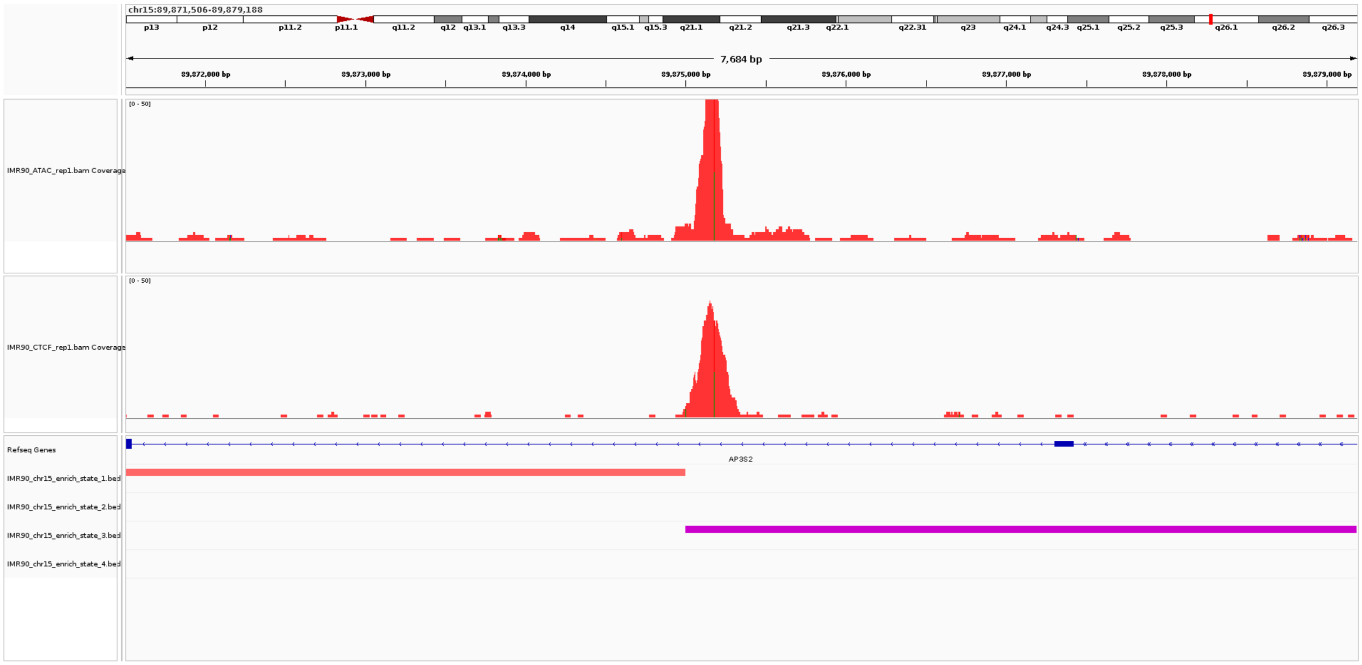
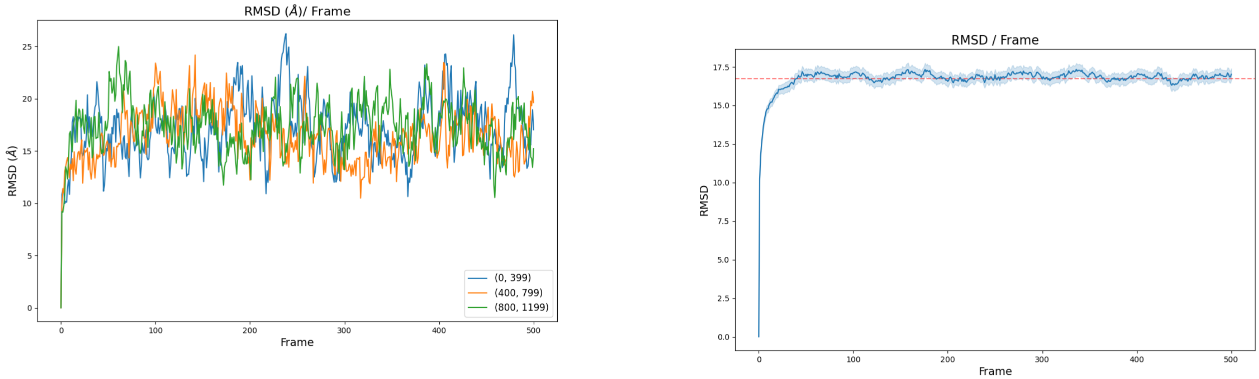


Figure 5. IGV snapshot of a portion of the ANPEP region analyzed. The first track reports the alignment results obtained from the ATAC data, the second the CTCF data

5.2 Trajectory analysis results

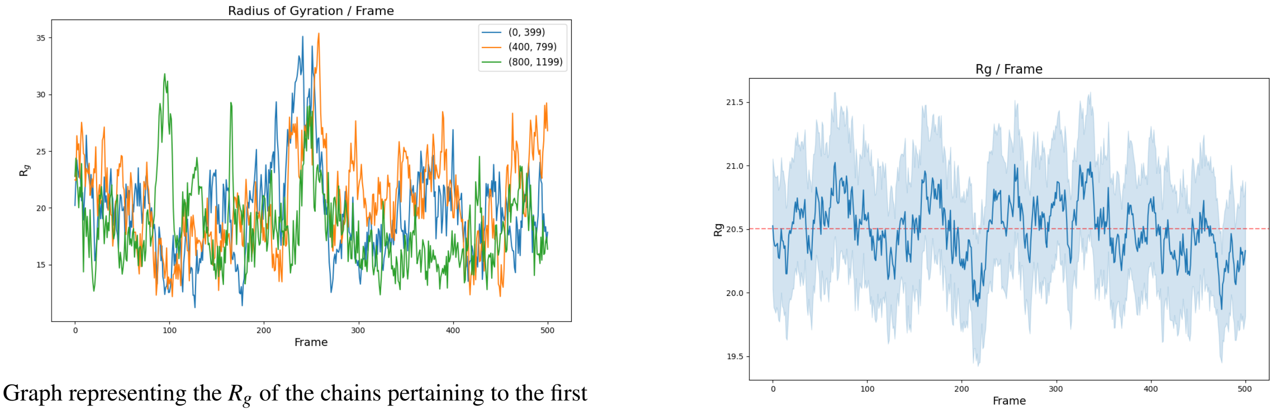
Results from the RMSD analysis are reported in figure 6a and 6b. The R_g results are instead plotted in images 7a and 7b. Finally the autocorrelation function graphs are reported in 8a and 8b. Although the RMSD seems to be less variable and the R_g profile, in reality the interval is approximately large the same. By looking to the RMSD profile, it can be argued that the final distension is obtained at the 50th 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.

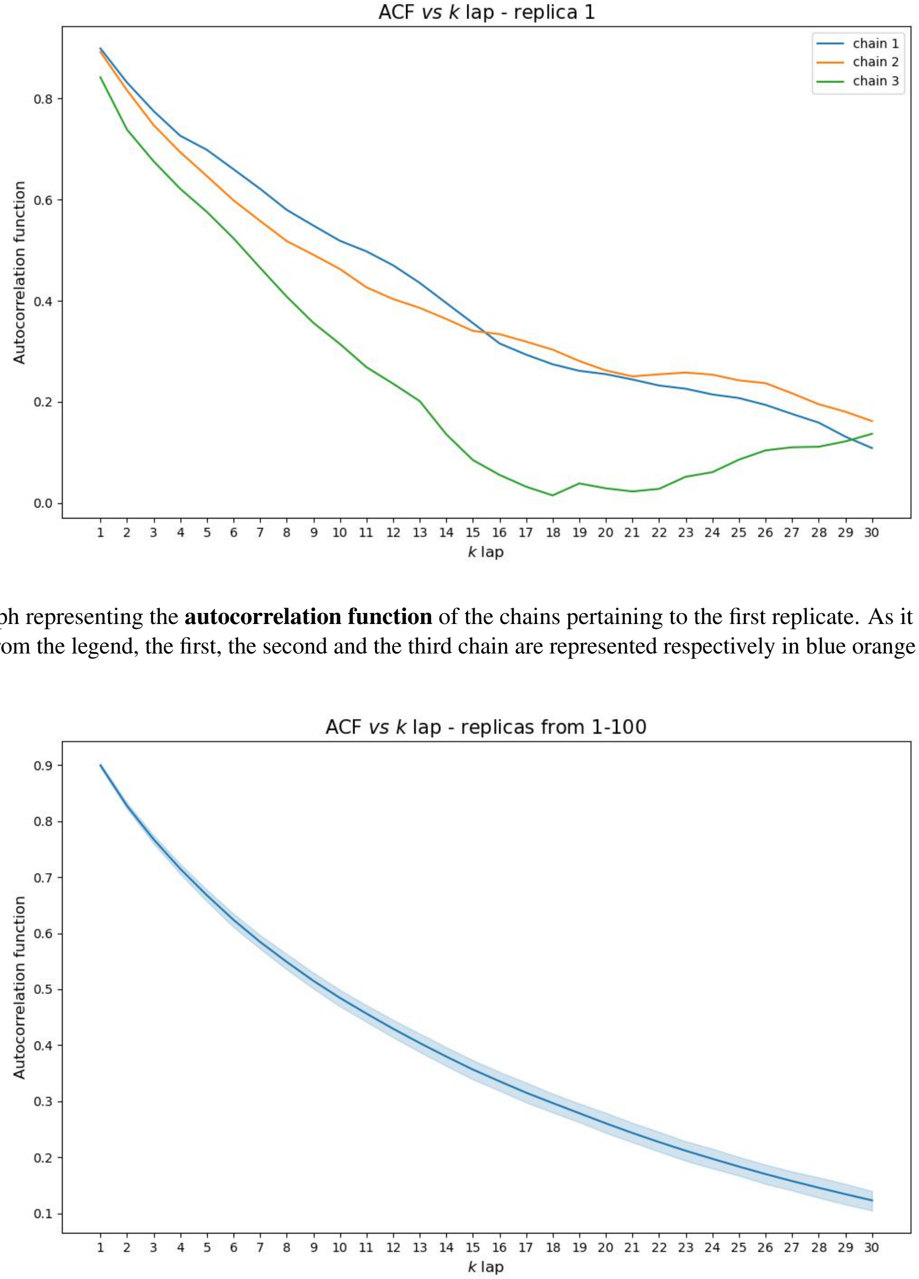
Figure 6. RMSD profiles



(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 7. 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 8. Autocorrelation function results

5.2.1 PCA analysis

5.3 Results from model selection

6 CONCLUSIONS

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. Additionally, the results obtained with the model could be compared to those resulting from other very interesing simulation softwares, such as *Origami* and *Hip-Hop*[11,12].

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

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. 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).
3. 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).
4. 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).
5. 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).
6. 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).
7. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137, DOI: 10.1186/gb-2008-9-9-r137 (2008).
8. Shah, A. Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system. *Nat Methods* **6**, ii–iii, DOI: 10.1038/nmeth.f.247 (2009).
9. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**, 2478–2492, DOI: 10.1038/nprot.2017.124 (2017).
10. Chilled House Vibes. Learning Chromatin States from ChIP-seq Data: ChromHMM - Luca Pinello (2015).
11. 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).
12. 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).
13. Steve Plimpton, Aidan Thompson, Stan Moore, Axel Kohlmeyer & Richard Berger. LAMMPS. https://www.lammps.org/#gsc.tab=0.
14. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, DOI: 10.1038/nature11247 (2012).
15. ATAC-seq (unreplicated) – ENCODE. https://www.encodeproject.org/pipelines/ENCPL344QWT/.
16. 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).
17. Transcription Factor ChIP-seq Data Standards and Processing Pipeline – ENCODE. https://www.encodeproject.org/chip-seq/transcription_factor/.
18. Homo sapiens genome assembly GRCh38 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.
19. 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.
20. UCSC Genome Browser Home. https://genome.ucsc.edu/.
21. 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).
22. 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).
23. 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).
24. 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).
25. 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).
26. 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).
27. 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).