PLACE HOLDER

Maurizio Gilioli

Contents

1 ABSTRACT

2 INTRODUCTION

2.1 Chromatin as the information center of a cell

[1] All the living organisms possess DNA, which is the main molecule through which information are passed from a generation to another. Chromatin is contained inside the nucleus in an ordered manner[1] . DNA is wrapped around histones forming nucleosomes. Throughout the report, I will call the nucleomes beads, which is a term that underlines the spherical shape of the DNA-histone complex. DNA and histones are subjected to different modifications; among those, methylation is the most important modification involving DNA. Methylation in mammals occurs in specific sites of the genome, called CpGs, where a cyto- sine is connected directly to a guanine. Methy- lations of regulatory elements have been im- plicated in determining cell identity and chro- matin structure[?,2] . 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.

2.2 ATAC-sequencing and CTCF sequencing

ATAC-sequencing is a technology that allows for the identification of open- chromatin regions[4,5]. In order to work, it requires the addition of TN-5, a hyper-active transposase. The latter is preloaded with sequencing adapters[5] 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 (in our case MACS-2[6]) 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 enrich- ment of reads is possible only in accessible regions, which are generally also the most active ones and with available sites for transcription factors binding. CTCF data, named in chapter ..., were obtained through a classical CHIP-sequencing.

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.

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

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[7] . The way it learns chromatin states signatures by using a multivariate hidden Markov model: In each genomic position, it returns the most probable chromatinic state (segments) 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 important portions of the genome (CPG islands, exons, genes). In the case of my study, ChromHMM was used to

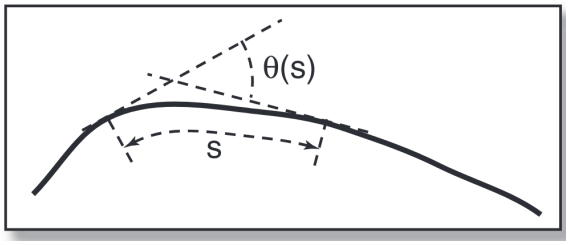

Figure 1. examplecaption

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. The scope of the report is also to gather opinions and useful feedback to improve the thesis work, which will continue for another two months.

4 METHODS

All the simulations were performed making use of the simulation software LAMMPS[?,8]

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 about the Origami simulation tool[9] were used for the project (see table 1).

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 also of the Origami[9] model. The written entries are for the ENCODE database[10].

The analyzed region was ANPEP in IMR90, taking into account the bases between the 89'000'000th and the 91'000'000.

4.2 The Model

The objective was to create a polymer model with a coarse-graining resolution of 5000 bp. The region of interest was ANPEP, it had a length of 2000000 bp and was considered in human genomes. The model creation was done by using code written by Marco Di Stefano. The total Genome length was obtained from the UCSC genome browser[11] . All the simulations were performed making use of periodic boundaries (chapter

All the simulations were performed making use of the *run_lammps* function inserted in the TADPHYS package. The *bond_style* that was set for LAMMPS[12] was the the *fene* bond style whose potential is written in equation 4.

The FENE potential is a finite extensible nonlinear elastic potential and is generally used for polymer models. The first term is attractive, the second Lennard-Jones (LJ) term is repulsive. The first term extends to R_0, the maximum extent of the bond. The second term is cutoff at $2^{1/6}\sigma$, the minimum of the LJ potential[12] . The K is an $^{\text{energy}}/_{\text{distance}^2}$ 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: 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.

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

4.2.1 The computation of the parameters for Coarse Graining

Both parameters for the CG model (table ??) and the CG model 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 (ν_{CG}).

Property	Formula	Value
c	*const.*	19
ν_{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)	$\frac{\text{DNAcontent}}{\nu_{FS}} * \text{ncopies}$	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)	$\frac{\rho_{FS} \cdot b_{FS}}{\nu_{FS} \cdot \text{lk}_{FS}}$	$1.2e-05$ 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)	$(lk_{FS} \cdot b_{FS})/\nu_{FS}$	2.5 nm^2/bp
Dlk_{FS} (DNA content of a Kuhn length FS)	$(lk_{FS} \cdot \nu_{FS})/b_{FS}$	1000 bp

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

Property	Formula	Value
c	*const.*	19
ν_{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)	$\frac{\text{DNAcontent}}{\nu_{CG}} * \text{ncopies}$	1200 mon.
side$_{CG}$ (size of the cubic simulation box)	$\frac{(N_{CG} \cdot \nu_{CG}/\rho_{CG})^{1/3}}{b_{CG}}$	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)	$\frac{\rho_{CG} \cdot b_{CG}}{\nu_{CG} \cdot \text{lk}_{CG}}$	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

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.

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. persistence length (PL) regulates the potential produced by angle interaction. The optimal pressure was attested at 0.192

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

5 RESULTS

6 DISCUSSION AND CONCLUSIONS

7 CONCLUSIONS

8 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

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. 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).
5. 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).
6. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137, DOI: 10.1186/gb-2008-9-9-r137 (2008).
7. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**, 2478–2492, DOI: 10.1038/nprot.2017.124 (2017).
8. Steve Plimpton, Aidan Thompson, Stan Moore, Axel Kohlmeyer & Richard Berger. LAMMPS. https://www.lammps.org/#gsc.tab=0.
9. 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).
10. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, DOI: 10.1038/nature11247 (2012).
11. UCSC Genome Browser Home. https://genome.ucsc.edu/.
12. 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).