Development and refinement of a coarse grained model for the dynamic representation of chromatin

Author: Martí Canet Vidal, mcanetvi9@ub.alumnes.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisors: Jaume Casademunt i Viader, jaume.casademunt@ub.edu
and Modesto Orozco López, modesto.orozco@irbbarcelona.org

Abstract: This project contributes to the development of a coarse-grained model at the MMB lab at IRB, designed to capture the sequence-dependent dynamics of DNA and its interactions with proteins using simplified Hamiltonians derived from atomistic potentials. A key objective was to assess its ability to reproduce the sequence-specific binding of transcription factors (TFs). To this end, a semi-stochastic algorithm was implemented to generate physically realistic initial configurations. Additionally, Free Energy Perturbation simulations were performed to quantify the impact of single-base mutations on TF binding affinity. Results highlight the differential energetic contribution of individual bases to protein-DNA interactions.
Keywords: Coarse-grained model, DNA sequence specificity, transcription factor binding, molecular dynamics, free energy perturbation, chromatin simulation.
SDGs: This work contributes to SDG 3 (Good Health and Well-Being), 4 (Quality Education) and 9 (Industry, Innovation and Infrastructure).

I. INTRODUCTION

Deoxyribonucleic acid (DNA) is an essential biomolecule for life, serving as the primary carrier of genetic information. It is structured as a double helix of complementary polynucleotides, molecules composed of nucleotides, *i.e.* a sugar-phosphate backbone joined to a nitrogenous base, covalently bonded along the backbone. These bases—adenine and guanine (purines), and thymine and cytosine (pyrimidines)—are paired complementarily between the two strands through hydrogen bonding. Each human cell contains approximately two meters of DNA, and therefore this polymer is known to be tightly compacted in a hierarchical manner. The double helix wraps around proteins known as histones, creating the nucleosomes, which then coil and fold to form the chromatin. The latter fibre is further compacted into loops and domains, ultimately forming a chromosome during cell division—the stage where DNA is maximally condensed.

Atomistic models to compute the dynamics and structure of DNA are very valuable tools for obtaining results at a resolution not currently achievable with experimental methods. Nevertheless, simulations at an atomic level are particularly computationally demanding, and consequently simpler models are often required. Coarse-grained (CG) models provide a simplified representation of complex systems, enabling more efficient simulations while preserving the chemical and physical properties as much as possible. To achieve this, sets of atoms are grouped into beads—also called interaction sites—which are generally point-like particles that interact under the influence of a suitable potential energy function that incorporates both bonded and non-bonded interactions.

II. CG MODEL

The primary objective of this project is to develop a coarse-grained model implemented within a Langevin–Brownian molecular dynamics framework, capable of reproducing the sequence-specific dynamics of DNA as described by accurate atomistic potentials. The model is further designed to incorporate proteins and to capture the binding free energy between transcription factors (TF) and their consensus DNA sequences (CS), as well as to model how this energy varies upon mutations. Consequently, it was necessary to construct interaction potentials that describe both the internal dynamics of the DNA double helix and the proteins, in addition to the sequence-dependent DNA–protein interactions.

On the one hand, proteins are represented using an Elastic Network Model (ENM) [2], in which each bead corresponds to the Cα atom of an amino acid and is connected to neighboring residues by harmonic springs. Additional beads representing the charged side chains of certain amino acids are included, as these are essential for accurately modeling electrostatic and Lennard-Jones interactions with DNA. The protein–protein interaction potential was recently developed based on the principle of minimal frustration. The elastic potential for each bead includes two contributions: a sequential term, E_{seq}, arising from interactions with adjacent residues in the amino acid sequence, and a spatial term, E_{cart}, accounting for interactions with nearby beads in three-dimensional space, regardless of their sequence separation. Let r denote the Cartesian distance and K the spring constant; the elastic energy between C_α beads is given by

$$E_{\text{elast}} = K(r - r_0)^2, \tag{1}$$

where the spring constants of each contribution are de-

fined as

$$\begin{cases} K_{\text{seq}} = \frac{C_{\text{seq}}^2}{s} & \text{if } s < 4 \\ K_{\text{cart}} = \left(\frac{C_{\text{cart}}}{r}\right)^6 & \text{if } s \geq 4 \;\&\; r < R \end{cases}, \tag{2}$$

with s denoting the sequential distance, $R = 15\,\text{Å}$, $C_{\text{seq}} = 20$ kcal·mol^{-1}·Å^{-2} and $C_{\text{cart}} = 5\,\text{Å}$. Side-chain interactions are also governed by the same elastic potential in Eq. (1) with $R = 9\,\text{Å}$ and

$$\begin{cases} K_{\text{seq}} = C_{\text{seq}}^2 & \text{if } s = 0 \\ K_{\text{cart}} = \left(\frac{C_{\text{cart}}}{r}\right)^6 & \text{if } s > 0 \;\&\; r < R \end{cases}, \tag{3}$$

in addition to a Lennard-Jones potential given by

$$E_{LJ} = 4\epsilon_{LJ}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], \tag{4}$$

where ϵ_{LJ} and σ are parameters specific to each amino acid and this interaction is only considered in the absence of an elastic interaction between the corresponding beads and if $r < 10\,\text{Å}$.

On the other hand, the DNA Model consists of a single bead per nucleotide positioned at the C1' atom, covalently linked to a charged, mass-less dummy particle. These virtual particles are used to define constraints without requiring explicit integration of their trajectories—they are instead reconstructed based on the configuration of the C1' beads, thereby reducing computational cost. The dummy particle is placed at atom N6 for adenine, O4 for thymine, N4 for cytosine and N6 for guanine. To further restrict large deformations or elongations of the double helix, several interparticle distances between each dummy particle and neighboring C1' were fixed during the reconstruction of the virtual particles. Each C1' bead carries a negative charge $q = -0.5e$, while the dummy particles are assigned $q = +0.5e$ for adenine and cytosine and $q = -0.5e$ for thymine and guanine, where e is the elementary charge.

The Hamiltonian for the C1' beads includes a sequence-dependent (bonded) term and a long-range (non-bonded) term,

$$E = E_{\text{seq}} + E_{\text{remote}}. \tag{5}$$

The sequential-dependent contribution, E_{seq}, follows the formulation proposed in [8]. Each C1' bead in the Watson strand (denoted i) interacts with its immediate neighbours ($i \pm 1$), its paired bead in the Crick strand (j), five upstream and five downstream beads in the complementary strand (i.e., $j \pm 1$ to $j \pm 5$) and the second-nearest neighbours ($i \pm 2$) through angle-dependent interactions. This term is composed of two contributions as follows,

$$E_{\text{seq}} = E_{\text{seq-4mer}} + E_{\text{seq-distant}}, \tag{6}$$

where the first component is a tetramer-specific term and the second accounts for long-range interactions within the 11-bead window. As detailed in [4], both *intrastrand* interactions—*stacking* (i.e. $i : i+1$, $j : j-1$) and *angle* ($i : i+1 : i+2$, $j : j-1 : j-2$)—and *interstrand* interactions—*pairing* ($i : j$) and *fan* ($i : j \pm 1$ to $i : j \pm 5$)—are modeled via polynomial expansions truncated to the fourth order. Therefore, sequence-dependent interactions are

$$E_{\text{stacking,pairing,fan}} = \sum_{a=2}^{4} K_a(\ell - \ell_0)^a, \tag{7}$$

$$E_{\text{angle}} = \sum_{a=2}^{4} K_a(\alpha - \alpha_0)^a, \tag{8}$$

where K_a are force constants and ℓ_0 and α_0 are the equilibrium distance and angle between beads respectively. The total tetramer-specific sequence energy is then

$$E_{\text{seq-4mer}} = \sum_{\text{Tetramer}} \left(\sum^{2} (E_{\text{stacking}} + E_{\text{pairing}}) + \sum^{4} E_{\text{angle}} + \sum^{8} E_{\text{fan}} \right), \tag{9}$$

and the longer-range fan interactions are given by

$$E_{\text{seq-distant}} = \sum_{\text{index}=\pm 4,\pm 5} \left(\sum_{i} E_{\text{fan}} \right). \tag{10}$$

The potential function for the remote term includes a Lennard-Jones component, which captures short-range excluded volume and Van der Waals interactions, and a Debye-Hückel component, which accounts for screened electrostatic interactions in solution,

$$E_{\text{remote}} = E_{LJ} + E_{DH}. \tag{11}$$

These interactions are not considered between particles located within 5 base-pairs (bp) of each other on either strand to avoid double counting with the bonded terms. Let k denote the Debye screening constant and $\varepsilon = \varepsilon_r \varepsilon_0$ the permittivity of water. The parameters ϵ_{LJ} and σ, specific to each bead type, describe the interaction strength and the effective bead size, respectively. Then the components of the remote potential are given by Eq. (4) and

$$E_{DH} = \frac{q_1 q_2}{\varepsilon_r \varepsilon_0 r} e^{-kr}. \tag{12}$$

The DNA and the protein also interact according to a force field properly fitted using Machine Learning (ML) techniques, to guarantee that free energy of binding estimates derived from funnel-metadynamics are consistent with the experimental free energy values obtained from high-throughput SELEX experiments [6, 7]. This interaction potential also includes both Lennard-Jones and Debye-Hückel terms, as defined in Eqs. (4) and (12).

III. STRUCTURE INITIALIZATION

To study chromatin dynamics, the model was implemented within a Langevin Dynamics code, integrated via the velocity Verlet algorithm at a temperature of 300 K with a 0.032 ps timestep. 50,000 frames were saved, sampling every 625 steps (20 ps), to ensure stability. Brownian stochastic forces were introduced using the Box-Müller transformation. However, to initialize simulations, a realistic configuration of the DNA-protein system was needed.

Given a DNA sequence, nucleosome positions (dyads) were estimated using ML-based predictions, assuming a 147-bp nucleosomal wrap around histones. The sequence was segmented into rigid nucleosomal DNA and flexible linkers, whose length depended on dyad spacing. Each segment was built analytically and aligned through isometric transformations, connecting the endpoints to form a full polymer in a semi-stochastic ordered fashion.

However, random placement often led to structural overlaps. To avoid this, a sequential placement strategy was used, starting with the first linker and then placing each nucleosome, along with its following linker, in succession. Minimum distances were set to 23 Å between linkers, 65 Å between linkers and nucleosomes, and 105 Å between nucleosome cores. Inter-fragment distances after each addition were checked, and if criteria weren't met, the latest nucleosome-linker pair was rotated until all conditions were satisfied (see Appendix C for further details).

To increase spatial freedom, three orthonormal axes were defined: (1) the vector from nucleosomal start to histone center, (2) the penultimate linker's direction, and (3) their cross product. The end of the last segment was centered at the origin, and up to 1337 unique rotation combinations of small-angle rotations ($\pm 5°$) around the three axes were tested. Upon success, the next unit was added.

In some cases with very short linkers (closely spaced dyads), no solution was found. To handle these, an additional constraint ensured enough space was left for the next nucleosome. After checking current overlaps, an extra loop verified that the end of the last linker maintained a safe distance from all existing fragments, thereby guaranteeing enough space for the next nucleosome. Otherwise, additional rotations were reapplied. In extreme cases, further freedom was introduced. First, the last three elements—two linkers and a nucleosome—were rotated around the second to last linker's start to avoid overlapping; then, the last linker was independently rotated around its own base until all conditions were met.

The algorithm could have been improved by accounting for the ellipsoidal shape of histones, but this would have relaxed spatial constraints, hindering convergence and increasing computational cost without clear benefits.

IV. DNA-TF BINDING FREE ENERGY

Because the interactions described are sequence-dependent, it is of great interest to quantify how the binding free energy between a transcription factor (TF) and a DNA segment changes upon mutation. Each TF is associated with a canonical or consensus sequence (CS)—the segment to which it binds with highest affinity. The consensus sequence is typically determined by identifying the most frequent nucleotide at each position.

This energy variation can be calculated using a thermodynamic cycle (see Appendix D), where each vertex of the square corresponds to one of the four possible states: consensus or the mutated sequences, either free or bound to the TF. Denoting the binding free energies of the consensus and mutated (random, RS) sequences as $\Delta G^{\mathrm{CS}}_{\mathrm{bind}}$ and $\Delta G^{\mathrm{RS}}_{\mathrm{bind}}$, and the free energy changes upon mutation in the free and bound states as $\Delta G^{\mathrm{free}}_{\mathrm{mut}}$ and $\Delta G^{\mathrm{bound}}_{\mathrm{mut}}$, the relative binding free energy can be computed without directly calculating the absolute value as

$$\Delta G^{\mathrm{RS}}_{\mathrm{bind}} - \Delta G^{\mathrm{CS}}_{\mathrm{bind}} = \Delta G^{\mathrm{bound}}_{\mathrm{mut}} - \Delta G^{\mathrm{free}}_{\mathrm{mut}}. \tag{13}$$

This calculation assumes that the single base-pair mutation is a microscopically reversible process. Consequently, the entire consensus sequence can be mutated into a random sequence by sequentially applying single-base mutations, each of which contributes additively to the overall free energy change.

A. Free Energy Perturbation

Results showed that directly switching a base pair was energetically too aggressive. To overcome this issue, the transformation between two bases was divided into smaller steps using a coupling parameter $\lambda \in [0, 1]$, where $\lambda = 0$ corresponds to the CS and $\lambda = 1$ to the RS. The energy at each intermediate state was computed as a linear interpolation between the two endpoints,

$$E(\lambda) = \lambda \cdot E_B + (1 - \lambda) E_A. \tag{14}$$

While the endpoints are still physically meaningful configurations, the intermediate values of λ describe computationally constructed chimeric systems containing features of both the consensus and random sequences. This forms what is known as an alchemical path. According to Free Energy Perturbation (FEP) theory [1], the free energy difference between the initial and final states is then

$$\langle G \rangle_B - \langle G \rangle_A = \sum_{\lambda=0}^{1} k_B T \ln \left\langle E^{(E_{\lambda+d\lambda} - E_\lambda)/k_B T} \right\rangle_\lambda . \tag{15}$$

Perturbative calculations were performed for five transcription factors—4iri, pou1f1, foxg1, nr4a2 and lmx1a—by randomly generating 10 mutated sequences for each.

All mutations maintained the same length as their corresponding canonical sequences, and each base-pair substitution was divided into 20 intermediate λ-states along the alchemical path.

Since the binding sites of these proteins are relatively short—ranging from 12 to 16 base pairs—border effects became significant. Dummy particles were consequently not properly reconstructed at the ends of the double helix, leading to potential simulation instabilities. To mitigate this, dummy particles were temporarily deactivated.

B. Results and analysis

FEP calculations only converge when the difference between two states is sufficiently small. If the system is in thermal equilibrium, the transition probability from the CS to the RS is related to its reverse process by the Boltzmann factor. As a result, both ergodicity and detailed balance are satisfied, and relation $\Delta G(A \to B) = -\Delta G(B \to A)$ becomes a necessary and sufficient condition for microscopic reversibility. However, the free-energy maps obtained from our simulations—plots of ΔG as a function of λ—exhibited a noticeable hysteresis error. Although this might suggest a significant discrepancy between the forward and reverse processes, the error appeared to grow linearly with the number of alchemical steps, indicating a systematic source, likely due to integration inaccuracies in the molecular trajectories. Since this error appears equally in both the free and bound states, it cancels out when computing the binding free energy difference, making it negligible for our purposes.

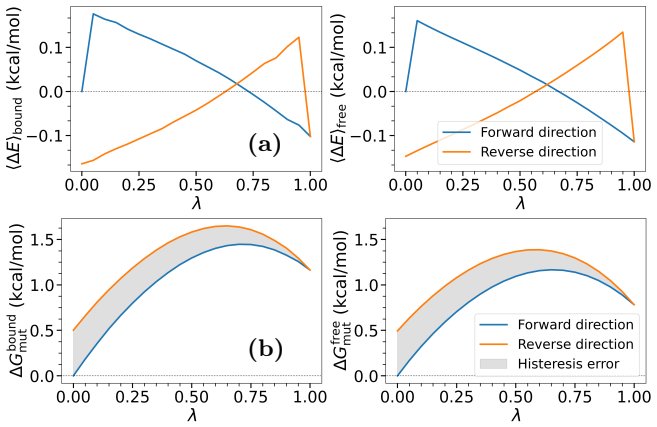


FIG. 1: Energy landscape (a) and free-energy map (b), for foxg1 with mutation T7>C, in the bound (left) and free (right) states.

All generated plots (see Fig. 1) showed an initial increase in free energy as the system transitioned from the CS at $\lambda = 0$, reaching a peak in one of the intermediate chimeric states, and then decreasing back towards $\lambda = 1$. This confirms that the physical endpoints correspond to stable energy minima. Additionally, the total energy averaged over all frames increased sharply when transitioning out of a physical state, then gradually decreased as the opposite state was approached.

While it may seem that the canonical sequence corresponds to the configuration with minimal binding energy for a given TF, this is not necessarily the case, as it is determined without accounting for collective effects or base-pair correlations. As a result, certain mutations may lower the free energy, others may raise it, and in some cases multiple mutations can even cancel each other out, resulting in little or no net change in binding affinity.

Due to the initial deactivation of dummy particles, interactions between DNA and TF were only partially captured, leading to reduced sequence specificity and smaller than expected binding energy differences. After improving the reconstruction of these virtual particles, I was able to run some functional simulations including them—though not exhaustively, due to time constraints. The qualitative behaviour remained consistent with previous observations, but this time the energy variations aligned more closely with theoretical expectations.

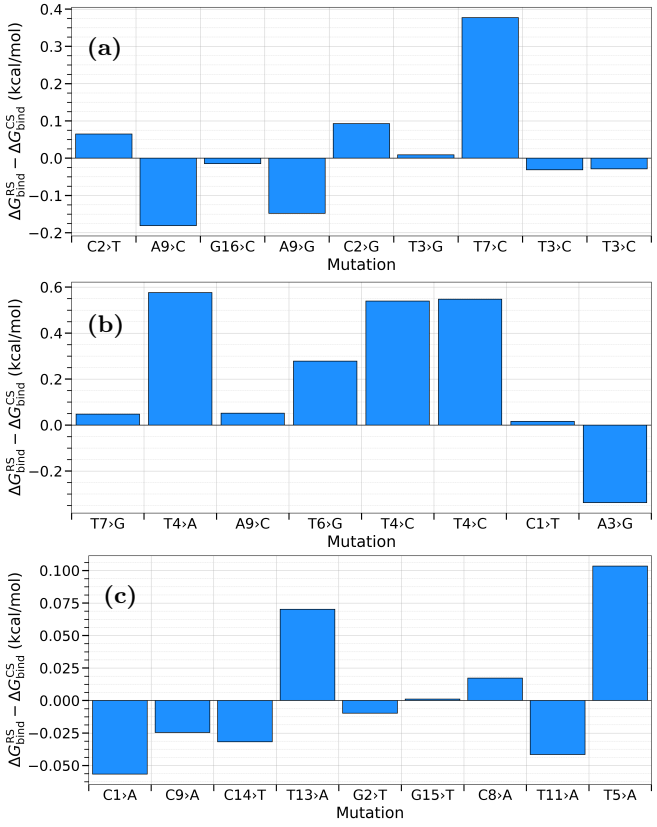


FIG. 2: Change in binding free energy ($\Delta\Delta G$) after mutations for different transcription factors: (a) foxg1, (b) lmx1a, and (c) nr4a2.

As shown in Fig. 2, certain bases—like T4 in the CS of lmx1a—play a significant role in binding affinity, with mutations at these sites leading to substantial changes in free energy. In contrast, other positions—such as T3 in

the foxg1 CS—are less critical for stable protein-DNA interactions. To verify the reliability of the simulations, the distance between the TF and DNA was monitored to assess the duration for which the complex remained bound. Interestingly, some simulations revealed two distinct stable binding distances between the centers of mass. However, when measuring the minimum distance between the two structures, the system consistently equilibrated within a narrow range (see Fig. 3). This demonstrates the ability of the double helix to flip while maintaining strong binding affinity in two distinct orientations.

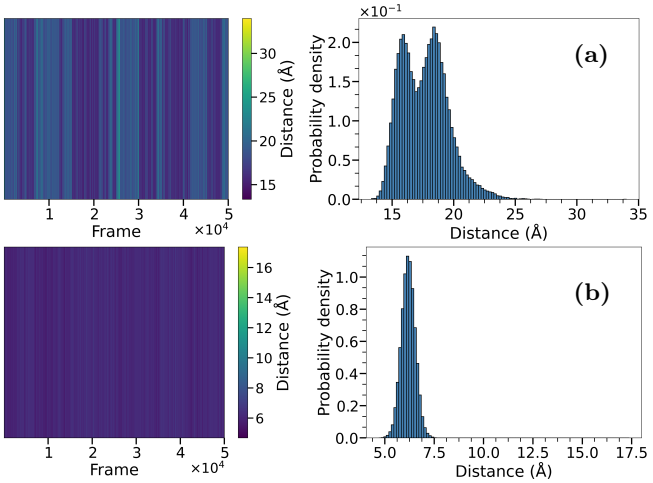


FIG. 3: Color maps and histograms with the distances between centers of mass (a) and closest beads (b) of nr4a2 and its CS with mutation T5>A and $\lambda = 0.6$.

V. CONCLUSIONS

During my time at the MMB lab at IRB, I worked with a coarse-grained model to dynamically represent chromatin. First, I developed a semi-stochastic algorithm to generate physically realistic initial configurations of DNA-protein assemblies. Second, I ran molecular dynamics simulations and applied Free Energy Perturbation theory to investigate the sequence specificity of transcription factors (TF) binding and to quantify the variation on binding free energy following mutations to canonical DNA sequences.

The results demonstrated that certain nucleotides contribute more significantly to TF binding affinity, with mutations at these sites producing larger changes in free energy—up to approximately 0.6 kcal/mol in some cases. Importantly, the TF-DNA complex remained stably bound throughout FEP simulations, validating the proper functioning of the model. Additionally, the DNA showed the ability to flip orientation relative to the protein while maintaining strong interaction. However, due to time constraints and technical challenges related to the reconstruction of dummy particles—beads representing nitrogenous bases in DNA—the number of completed simulations was limited. Future work should aim to expand this analysis to a broader set of transcription factors and mutation patterns. It would also be of interest to simulate competitive binding scenarios, such as multiple TFs interacting with a single binding site, or a single TF encountering multiple potential targets.

Acknowledgments

I sincerely thank Lucía, my colleague David Farré-Gil, my family, the MMB lab at IRB Barcelona and my two advisors, Modesto Orozco and Jaume Casademunt, without whose support this project would not have been possible.

[1] Christopher J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, (John Wiley & Sons, England, 1961).

[2] Orellana, L., Rueda, M., Ferrer-Costa, C., Lopez-Blanco, J. R., Chacón, P., and Orozco, M. "Approaching Elastic Network Models to Molecular Dynamics Flexibility". Journal of Chemical Theory and Computation **6**: 2910–2923 (2010).

[3] Gelpí, J. L., Kalko, S. G., Barril, X., Cirera, J., de la Cruz, X., Luque, F. J., and Orozco, M. "Classical Molecular Interaction Potentials: Improved Setup Procedure in Molecular Dynamics Simulations of Proteins". Proteins: Structure, Function, and Bioinformatics **45**: 428–437 (2001).

[4] Farré-Gil, D., Arcon, J. P., Laughton, C. A., and Orozco, M. "CGeNArate: a sequence-dependent coarse-grained model of DNA for accurate atomistic MD simulations of kb-long duplexes". Nucleic Acids Research **52**: 6791–6801 (2024).

[5] Barissi, S., Sala, A., Wieczór, M., Battistini, F., and Orozco, M. "DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors". Nucleic Acids Research **50**: 9105–9114 (2022).

[6] Pantier, R., Chhatbar, K., Alston, G., Lee, H. Y., and Bird, A. "High-throughput sequencing SELEX for the determination of DNA-binding protein specificities in vitro". STAR Protocols **3**: 101490 (2022).

[7] Ogawa, N. and Biggin, M.D. "High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro". In: Methods in Molecular Biology **786**: 51–63 (2012).

[8] Savelyev, A. and Papoian, G.A. "Chemically accurate coarse graining of double-stranded DNA". Proceedings of the National Academy of Sciences **107**: 20340–20345 (2010).

Desenvolupament i refinament d'un model de gra fi per a la representació dinàmica de la cromatina.

Author: Martí Canet Vidal, mcanetvi9@ub.alumnes.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisors: Jaume Casademunt i Viader, jaume.casademunt@ub.edu
and Modesto Orozco López, modesto.orozco@irbbarcelona.org

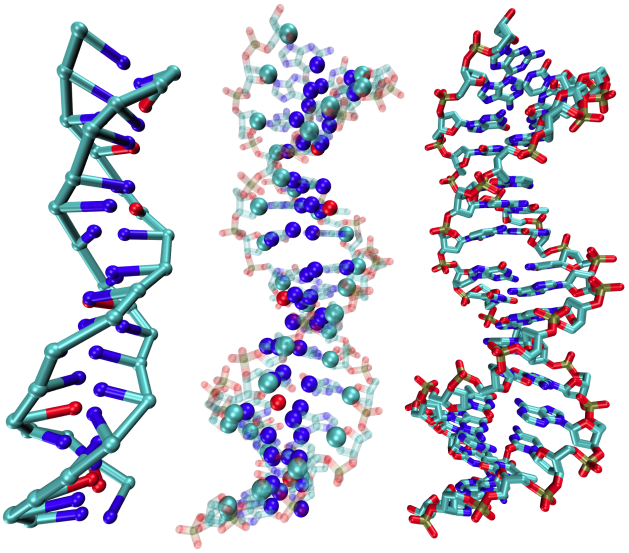
Resum: Aquest treball contribueix en el desenvolupament i millora d'un model de gra fi (*coarse-grained*) per descriure la dinàmica de la cromatina d'una manera computacionalment eficient, mantenint alhora la dependència amb la seqüència d'ADN. El model ha estat recentment estès per incloure la interacció amb proteïnes, com ara factors de transcripció. En aquest projecte, s'ha dissenyat un algoritme semi-estocàstic per generar configuracions inicials físiques realistes, i s'han dut a terme simulacions de dinàmica molecular amb mutacions puntuals per calcular variacions en l'energia lliure d'unió mitjançant un mètode de pertorbacions. Els resultats mostren que algunes bases són més rellevants que d'altres per a l'afinitat entre ADN i proteïna, evidenciant la capacitat del model per capturar efectes específics de seqüència.
Paraules clau: Model de gra fi, especificitat de seqüència, unió ADN-proteïna, dinàmica molecular, energia lliure d'unió, FEP.
ODSs: Aquest TFG està relacionat amb els ODS 3 (Salut i benestar), 4 (Educació de qualitat) i 9 (Indústria, innovació i infraestructures).

Objectius de Desenvolupament Sostenible (ODSs o SDGs)

1. Fi de la es desigualtats		10. Reducció de les desigualtats	
2. Fam zero		11. Ciutats i comunitats sostenibles	
3. Salut i benestar	X	12. Consum i producció responsables	
4. Educació de qualitat	X	13. Acció climàtica	
5. Igualtat de gènere		14. Vida submarina	
6. Aigua neta i sanejament		15. Vida terrestre	
7. Energia neta i sostenible		16. Pau, justícia i institucions sòlides	
8. Treball digne i creixement econòmic		17. Aliança pels objectius	
9. Indústria, innovació, infraestructures	X		

El contingut d'aquest TFG, desenvolupat en el marc d'un grau universitari de Física, es relaciona amb l'ODS 4, en particular amb la fita 4.4, ja que contribueix a l'educació científica a nivell superior mitjançant l'adquisició de compotències en modelització computacional. També es vincula amb l'ODS 3, fita 3.b, perquè promou la recerca i el desenvolupament de tecnologies relacionades amb la salut, i amb l'ODS 9, fita 9.5, pel seu impuls a la recerca científica i la innovació en bioinformàtica estructural.

GRAPHICAL ABSTRACT



Appendix A: CG MODEL OUTLINE

The following figure provides a schematic representation of the sequential interactions considered in the CG model developed in this work. It complements the description in Section II, visually distinguishing long-range and tetrameric interactions.

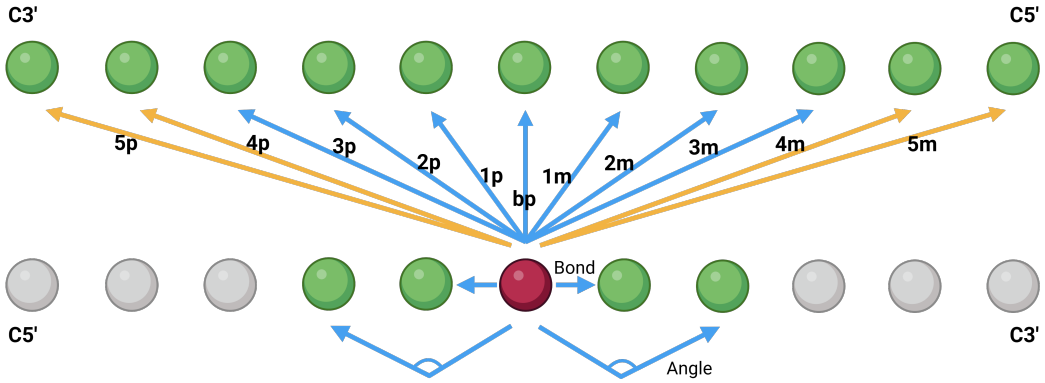


FIG. 4: Outline of the sequential terms used in the CG model. Yellow arrows represent long-range terms; blue arrows correspond to tetramer terms.

Appendix B: FORCE FIELD PARAMETERS AND BEADS PROPERTIES

As described in Section II, the model considers electrostatic interactions between DNA and protein beads. Therefore it is important to highlight the amino acids with electrically charged side chains (SC), as summarized in Table I.

TABLE I: Charge of the side chain bead for each amino acid, in units of the elementary charge, e.

Aminoacid	SC bead charge (e)
Arg	+1.0
His	+1.0
Lys	+1.0
Asp	−1.0
Glu	−1.0

Letting z_i and n_i denote the valence and number density of ion species i, respectively, and k_B the Boltzmann constant, the screening constant k from Eq. 12 is defined as the inverse of the Debye length,

$$\lambda_D = \frac{1}{k} = \sqrt{\frac{\varepsilon_0 \varepsilon_r k_B T}{\sum_i n_i z_i^2 e^2}}, \tag{B1}$$

which represents the characteristic distance over which electric potentials are screened in an ionic solution.

All amino acids are represented by a backbone bead, while certain amino acids also include a side chain bead with distinct fitted physical parameters. In addition to Debye–Hückel electrostatics, the DNA–protein force field incorporates Lennard-Jones interactions with well depths ϵ_{LJ} specific to each bead type, as shown in Table II.

TABLE II: Fitted values of ϵ_{LJ} for each bead.

Bead	C1'	N1, N4, N6	O4	Cα	SC
ϵ_{LJ} (kcal/mol)	0.12	0.04	0.03	0.01	0.01

The effective distances σ, at which the attractive and repulsive Lennard-Jones forces cancel, are presented for each amino acid and its side chain in Table III.

Appendix C: MATHEMATICAL TREATMENT IN INITIAL CONFIGURATION GENERATION

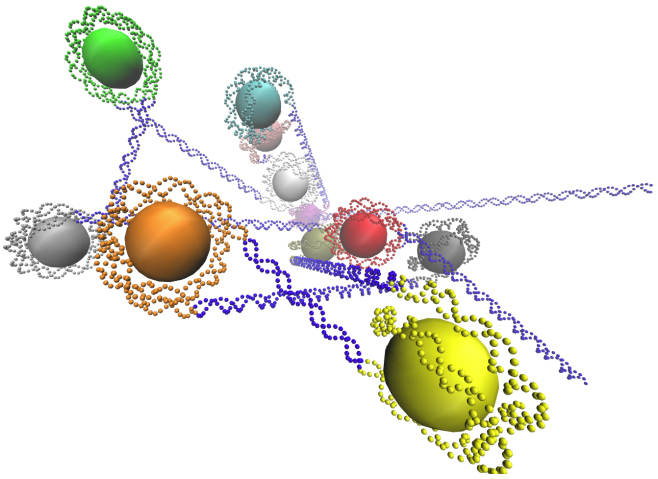


FIG. 5: Overview of a generated initial chromatin structure. Linkers are shown in dark blue; the coloured structures correspond to nucleosomes.

As described in Section III, to avoid overlapping structures during initial configuration generation, the most

recently DNA fragments and associated histone were rotated after a strategic translation to recenter the reference frame. Because realistic systems contain a large number of beads, computing all pairwise distances at each generation step would be computationally expensive. To reduce this burden, a simplified representation was used.

Each nucleosome was represented by a single bead located at the center of mass, corresponding to the histone core. DNA linkers, on the other hand, were represented by two virtual particles. Instead of modelling the double helix, a regression line was fitted through all C1' atoms of the linkers, and the projections of its two endpoints were used to define the linker's position. As a result, structural overlaps were assessed by comparing segments and single points, significantly reducing computational cost.

The minimum distances were established based on the approximate radius of the DNA helix, 11 Å, and the overall radius of a nucleosome—including both the histone core and its surrounding DNA—about 57 Å.

Once the three orthonormal axes were defined, all possible 7-element combinations with repetition of the axes were computed. Since rotations are non-commutative and can proceed in either direction, both the order of the axes and their orientations were considered. To avoid unnecessary computations, redundant operations—such as a clockwise rotation followed by its inverse—were excluded, reducing the total to the 1,337 unique combinations mentioned.

Given a 3D array A—where the first dimension indexes DNA fragments, the second indexes beads within each strand, and the third holds the Cartesian coordinates—the translation of all beads in a fragment along a direction B is given by

$$A'[:][:] = A[:][:] + B, \tag{C1}$$

where $A[:][:]$ and $A'[:][:]$ represent the original and translated coordinates, and B is the vector needed to bring the bead of interest to the origin. Let $\hat{k}$ be the unit vector defining the axis of rotation and θ the angle of rotation. The rotation matrix R is given by

$$R = \mathbb{1} \cdot \cos\theta + \sin\theta \cdot [\hat{k}]_\times + (1 - \cos\theta)(\hat{k} \otimes \hat{k}), \tag{C2}$$

where $\mathbb{1}$ is the identity matrix, $\hat{k} \otimes \hat{k}$ is the outer product and $[\hat{k}]_\times$ is the skew-symmetric cross-product matrix of $\hat{k}$,

$$[\hat{k}]_\times = \begin{pmatrix} 0 & -k_z & k_y \\ k_z & 0 & -k_x \\ -k_y & k_x & 0 \end{pmatrix}. \tag{C3}$$

Appendix D: FREE ENERGY CYCLE

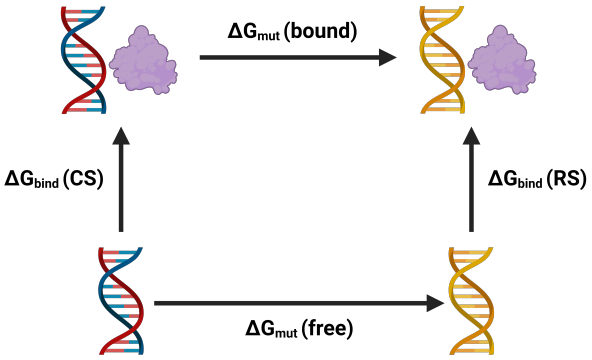


FIG. 6: Thermodynamic cycle used to compute the change in binding free energy upon mutation. The top row represents bound states, while the bottom row shows the unbound (free) states. The left column corresponds to the consensus sequence, and the right column to the mutated or random sequence.

TABLE III: Fitted values of σ for each amino acid backbone (BB σ) and side chain (SC σ) beads. Amino acids without side chains are marked with (–).

Amino acid	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
BB σ (Å)	4.82	7.54	9.06	4.68	5.84	5.58	4.72	4.84	6.90	9.16	9.24	5.56	12.16	7.32	6.44	5.50	6.60	7.02	6.68	6.78
SC σ (Å)	(-)	4.06	(-)	2.58	(-)	2.54	4.10	(-)	4.26	(-)	(-)	3.80	(-)	4.20	(-)	(-)	(-)	7.62	4.84	(-)