Brewing COFFEE: A sequence-specific coarse-grained energy function for simulations of DNA-protein complexes.

Debayan Chakraborty,[*,†] Balaka Mondal,[†] and D. Thirumalai[*,†,‡]

†*Department of Chemistry, The University of Texas at Austin, 105 E 24th St, Stop A5300, Austin TX 78712, USA.*

‡*Department of Physics, The University of Texas at Austin, 2515 Speedway,Austin TX 78712, USA.*

E-mail: debayan.chakraborty@utexas.edu; dave.thirumalai@gmail.com

Abstract

DNA-protein interactions are pervasive in a number of biophysical processes ranging from transcription, gene expression, to chromosome folding. To describe the structural and dynamic properties underlying these processes accurately, it is important to create transferable computational models. Toward this end, we introduce **Co**arse grained **f**orce **f**ield for **e**nergy **e**stimation, COFFEE, a robust framework for simulating DNA-protein complexes. To brew COFFEE, we integrated the energy function in the Self-Organized Polymer model with Side Chains for proteins and the Three Interaction Site model for DNA in a modular fashion, without re-calibrating any of the parameters in the original force-fields. A unique feature of COFFEE is that it describes sequence-specific DNA-protein interactions using a statistical potential (SP) derived from a dataset of high-resolution crystal structures. The only parameter in COFFEE is the strength (λ_{DNAPRO}) of the DNA-protein contact potential. For an

optimal choice of λ_{DNAPRO}, the crystallographic B-factors for DNA-protein complexes, with varying sizes and topologies, are quantitatively reproduced. Without any further readjustments to the force-field parameters, COFFEE predicts the scattering profiles that are in *quantitative agreement* with SAXS experiments as well as chemical shifts that are consistent with NMR. We also show that COFFEE accurately describes the salt-induced unraveling of nucleosomes. Strikingly, our nucleosome simulations explain the destabilization effect of ARG to LYS mutations, which does not alter the balance of electrostatic interactions, but affects chemical interactions in subtle ways. The range of applications attests to the transferability of COFFEE, and we anticipate that it would be a promising framework for simulating DNA-protein complexes at the molecular length-scale.

Introduction

DNA-protein complexes play key roles within the cellular machinery, from orchestrating chromatin organization across different length-scales,[1–3] to regulating gene expression, replication, as well as DNA repair. Structural studies based on X-ray crystallography, NMR, and most recently cryo-EM techniques show that interactions between DNA and partner proteins could be non-specific, driven largely by electrostatic interactions and size and shape complementarity, or sequence-specific, requiring precise binding modules. Insights into the energetics of DNA-protein interactions have also come from gauging contact statistics within structural databases,[4] and other knowledge-based approaches.[5,6] However, simple rules for predicting the sequence-dependent changes in DNA-protein interactions have not emerged. Indeed, progressing towards a quantitative understanding of the molecular mechanisms governing the assembly of DNA-protein complexes is an important area of research.

Structural studies, although important, provide only a restricted view of the interactions driving the formation of DNA-protein complexes. The dynamics of DNA-protein binding could substantially modulate the efficacy of the recognition process, and tune the desired functionality. For instance, single molecule experiments[7,8] as well as computations based on minimal models[9–11] have revealed that many proteins including RNA polymerases and the tumor suppressor p53, may initially search for target sites on DNA via facilitated diffusion.[12]

Due to the limited spatial and temporal resolution, experiments alone cannot fully resolve the structural and dynamical details of DNA-protein assembly. Atomically detailed simulations,[13] based on different force-fields, have provided invaluable insights into various aspects of DNA-protein complexes, including the binding of transcription factors to cognate sites on DNA,[14–16] the spontaneous association of single-stranded DNA with SSB proteins,[17] and salt-dependent unwrapping in nucleosomes.[18,19] Despite these advances, all-atom simulations of DNA-protein complexes on biologically relevant time-scales are computationally intractable. In addition, force-field deficiencies often manifest themselves only over long simulation time-scales, and may require extensive reparametrization to yield the needed

accuracy for characterizing DNA-protein complexes.[20]

It has been shown repeatedly that it is advantageous to exploit a simplified or a 'coarse-grained' (CG) representation, especially for large systems such as protein-DNA complexes. By judiciously choosing a resolution that depends on the length and time-scales of interest, insights into a variety of problems could be obtained.[21] For example, using a single-bead polymer-type model, important predictions have been made regarding the force-induced unwrapping of nucleosomes,,[22] promoter melting by RNA polymerase,[23] and more recently the formation of condensates in low complexity RNA sequences.[24] In recent years, CG models of different resolutions have been introduced to simulate DNA-protein complexes. Combination of the AWESEM protein force-field[25] and the 3SPN model for DNA[26] was used to characterize the energy landscapes for Fis:DNA binding,[27] as well as nucleosome unwrapping.[28] Models of similar resolution have also been exploited by other groups to probe the molecular details of various DNA-protein assembly processes, such as DNA bending induced by architectural proteins,[29] sliding of ssDNA on protein surfaces,[30,31] and higher-order chromatin folding.[32,33] In addition, higher resolution CG force-fields such as SIRAH and MARTINI are successful in describing DNA-protein interfaces.[34,35] Several studies have also exploited a multiscale approach, in which certain regions of the biomolecule, particularly the DNA-protein interfaces are described in atomic detail, and the rest are represented at a CG level. Schulten and coworkers,[36] and subsequently others,[37] have adopted this scheme to probe the conformational dynamics of the lac repressor-DNA complex. However, for models with mixed granularity, the coupling between different resolutions could be non-trivial, which makes it challenging to obtain a balanced description of the energetics.[38]

During the last decade or so, we have introduced several CG models for understanding biomolecular folding and assembly. Among these, the Self-Organized Polymer model with side-chains (abbreviated as SOP-SC) has been remarkably successful in predicting the thermodynamics and kinetics of folding for a diverse range of protein sequences, at different denaturant, salt concentration, as well as pH.[39–45] For nucleic acids, the three interaction

site (TIS) model[46] provides a fine balance between accuracy and speed. Different variants of the TIS model have been exploited by us and others[26] to provide a quantitative description of ion-assisted RNA folding,[47–51] folding of G-quadruplexes,[52] and DNA mechanics and thermodynamics.[53] Encouraged by the wide range of applicability of our existing CG models, we develop **COFFEE** (**Co**arse-grained **F**orce **F**ield for **E**nergy **E**stimation), a hybrid potential that integrates the SOP-SC model for folded proteins, with the TIS model for DNA. It is worth emphasizing that while making COFFEE, we did not recalibrate the parameters in the SOP-SC or TIS energy functions, which individually provides quantitatively accurate description of the sequence-specific protein-protein and DNA-DNA interactions, respectively. A key ingredient of COFFEE is a knowledge based statistical potential (SP) used to describe the sequence-specific DNA-protein contacts.

We first show that the simulations based on COFFEE accurately reproduce the crystallographic B-factors for a variety of DNA-protein complexes with different sizes and topologies. In addition, COFFEE also predicts the scattering profiles in quantitative agreement with SAXS experiments, and chemical shifts that are consistent with NMR, *without any further fine-tuning.* As a further application of COFFEE, we probe the salt-induced unwrapping of a nucleosome. We show that various partially unwrapped nucleosome conformations are populated as the monovalent salt concentration is increased. Our observations recapitulate the key findings from previous simulations,[54,55] as well as SAXS experiments.[56,57] We also show that ARG to LYS mutations at defined superhelical locations destabilize the nucleosome structure, causing extensive unwrapping, in accord with experimental findings.[58,59]

The accuracy of our findings attests to the power of COFFEE. We anticipate that the method underlying COFFEE would be a promising framework for simulating large DNA-protein assemblies, particularly at the 10-100 nm length-scale, and would buttress the ongoing efforts[13] to transform the current state-of-the-art.

.

Methodology

COFFEE combines the Self Organized Polymer with Side-Chains (SOP-SC) model for proteins[39,60] with a Three Interaction Site (TIS) model for DNA.[46,53] In the SOP-SC model, each amino-acid residue is represented using two interaction sites, a backbone bead (BB) centered on the C_α atom and a side-chain (SC) bead placed on the center-of-mass of the side-chain. In the DNA model, each nucleotide is represented using three interaction sites, positioned on the center of mass of the phosphate (P), sugar (S), and base (B) groups. The coarse-grained energy function for the DNA-protein complex is:

$$U_{COFFEE} = U_{SOP-SC} + U_{TIS-DNA} + U_{DNA-PRO}, \tag{1}$$

where U_{SOP-SC} is the energy function corresponding to the SOP-SC model; $U_{TIS-DNA}$ is the energy function for the TIS-DNA model; and $U_{DNA-PRO}$ describes the specific as well as non-specific interactions between the DNA and protein molecules.

SOP-SC model for proteins: The SOP-SC model is optimized for studying single- as well as multi-domain protein folding[61] under a diverse set of conditions, and describes the thermodynamics and kinetics in quantitative agreement with experiments.[39,41–45] The SOP-SC energy function includes contributions from bonded (U_{FENE}), native (U_N), non-native (U_{NN}), as well electrostatic interactions (U_{ELE}^{PRO}), and is given by:

$$U_{SOP-SC} = U_{FENE} + U_N + U_{NN} + U_{ELE}^{PRO}. \tag{2}$$

TIS model for DNA: The TIS model for DNA provides a quantitatively accurate description of both single-stranded and double-stranded DNA, and has recently been used[53] to explore their sequence-dependent mechanical as well as thermodynamic properties. The TIS energy function includes contributions from bond (U_B), angular (U_A), stacking (U_S),

hydrogen-bonding (U_{HB}), excluded-volume (U_{EV}), and electrostatic (U_{ELE}^{DNA}) interactions:

$$U_{TIS} = U_B + U_A + U_S + U_{HB} + U_{EV} + U_{ELE}^{DNA}. \tag{3}$$

The detailed functional forms of the potentials in Eq. 2, and 3 are described in the Supporting Information (SI). The force-field parameters for the SOP-SC and TIS models are tabulated in the SI (Tables S1 to S7).

DNA-protein interactions: The non-bonded interactions between DNA and proteins, as described by the $U_{DNA-PRO}$, includes both electrostatic and non-electrostatic components,

$$U_{DNA-PRO} = U_{ELE} + U_{NOELE}. \tag{4}$$

The electrostatic interactions, U_{ELE}, between the phosphates on DNA and the amino-acid side-chains of the protein are described by the Debye-Hückel potential,

$$U_{ELE} = \sum_{i,j}{}' \frac{\lambda_{ele} q_i^{PRO} q_j^{DNA} \exp\left(-\kappa r_{ij}\right)}{\varepsilon_{DNAPRO} r_{ij}} \tag{5}$$

where q_i^{PRO} is the charge on the SC bead of amino-acid residue i (+1 for ARG and LYS, and -1 for ASP and GLU at physiological pH), and q_j^{DNA} denotes the renormalized charge (chosen to be 0.6 to account for counter-ion condensation,[62] as described previously.[53]) on the phosphate bead corresponding to nucleotide j; r_{ij} is the distance between the SC and the phosphate beads; κ denotes the inverse Debye length, and ε_{DNAPRO} is the dielectric constant. We assume that ε_{DNAPRO} to be temperature-independent, and set it equal to 78.0. Following previous studies,[28,63,64] we also include a scaling factor, λ_{ele}=1.67, to partially account for the lack of explicit counter-ions in the model.

The non-electrostatic component, U_{NOELE}, includes both native (U_N) and non-native (U_{NN}) contributions,

$$U_{NOELE} = \lambda_{DNAPRO} U_N + U_{NN}. \tag{6}$$

In Eq. 6, the adjustable parameter λ_{DNAPRO} tunes the strength of the native interactions in the DNA-protein complex. A contact is presumed to be native if the distance between a DNA and a protein bead in the reference structure is less than 12 Å. The native interactions among the DNA and protein beads are described in terms of Lennard-Jones type potentials:

$$\begin{aligned} U_N &= \sum_{i=1}^{N_{PBB}} \epsilon_{PBB} \left[\left(\frac{r^0_{i,ref}}{r_i} \right)^{12} - 2 \left(\frac{r^0_{i,ref}}{r_i} \right)^{6} \right] + \sum_{i=1}^{N_{SBB}} \epsilon_{SBB} \left[\left(\frac{r^0_{i,ref}}{r_i} \right)^{12} - 2 \left(\frac{r^0_{i,ref}}{r_i} \right)^{6} \right] \\ &+ \sum_{i=1}^{N_{BBB}} \epsilon_{BBB} \left[\left(\frac{r^0_{i,ref}}{r_i} \right)^{12} - 2 \left(\frac{r^0_{i,ref}}{r_i} \right)^{6} \right] + \sum_{i=1}^{N_{BSC}} \epsilon_{BSC} \left[\left(\frac{r^0_{i,ref}}{r_i} \right)^{12} - 2 \left(\frac{r^0_{i,ref}}{r_i} \right)^{6} \right] \\ &+ \sum_{i=1}^{N_{SSC}} \epsilon_{SSC} \left[\left(\frac{r^0_{i,ref}}{r_i} \right)^{12} - 2 \left(\frac{r^0_{i,ref}}{r_i} \right)^{6} \right] \end{aligned} \tag{7}$$

The first term in Eq. 7 accounts for the native interactions between the phosphates (P) of DNA and the protein backbone (BB), with ϵ_{PBB} being the interaction energy scale. The second and the third terms denote the pairwise interactions between the protein backbone (BB) with the sugars (S) and nucleobases (B) of DNA. The strengths of these non-bonded interactions are denoted by ϵ_{SBB} and ϵ_{BBB}, respectively. Based on the distance cutoff, native contacts can also be defined between the protein side-chains (SC) and the DNA nucleobases (B), as well as sugars (S). The corresponding interaction energies are denoted by ϵ_{SSC} and ϵ_{BSC}. Following previous works,[65,66] we assume that the interactions between the DNA phosphate backbone and the amino-acid side-chains is purely electrostatic in nature. In all cases, $r^0_{i,ref}$ denotes the distance between the ith pair of DNA and protein beads in the coarse-grained representation of the reference (experimental) structure.

The energy scales for the kth pair of native interactions, depend on the protein and DNA sequences, and can be generically represented as, $\epsilon_{XY} = \left| \epsilon_k^{XY} - d^{XY} \right|$, where $X \in [P, B, S]$ denotes the type of DNA bead, and $Y \in [BB, SC]$ denotes the type of protein-bead. Here, ϵ^{XY} is related to the 'free energy' cost, ΔG^{XY}, of forming a contact between residues X

and Y, and d^{XY} is an offset parameter, which ensures that ϵ_{XY} values are positive. The ϵ_k^{XY} values for the different amino-acid-nucleotide pairs were derived from the statistics of DNA-protein contacts, as described below.

	GUA	ADE	CYT	THY
ALA	0.75	0.71	0.70	0.78
GLY	1.07	1.02	1.04	1.01
LEU	0.00	0.03	0.01	0.24
ILE	0.32	0.38	0.43	0.40
TYR	0.86	0.74	0.90	0.85
TRP	0.21	0.20	0.46	0.39
LYS	1.13	1.09	1.22	1.12
PRO	0.63	0.64	0.58	0.69
GLU	0.69	0.42	0.66	0.43
HIS	1.21	1.13	1.18	1.17
ASP	1.00	0.59	0.98	0.53
ARG	1.52	1.38	1.49	1.41
GLN	1.09	1.06	1.03	1.08
ASN	1.22	1.20	1.21	1.30
CYS	0.56	0.50	0.68	0.44
SER	1.27	1.11	1.24	1.14
MET	0.51	0.62	0.63	0.72
VAL	0.63	0.61	0.55	0.64
THR	1.09	1.07	1.08	1.15
PHE	0.64	0.50	0.54	0.62

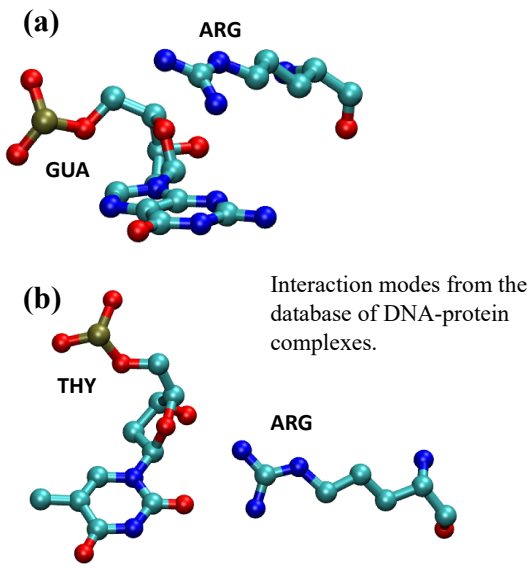


Figure 1: Left: A matrix denoting the effective energy scales (in kcal/mol) derived from the statistics of DNA-protein contacts for nucleobases (B) and amino-acid side-chains (SC). Each cell of the matrix is color-coded in accordance with the strength of the DNA-protein contact, with intense colors denoting stronger interactions. Right: Illustrative examples of contacts involving (a) GUA and ARG; (b) THY and ARG.

Statistical potential for native DNA-protein contacts: The development of statistical potentials (SPs) based on residue-residue contacts found within structures deposited in the PDB database, was pioneered by Tanaka and Scheraga[67,68] in their studies on protein folding. Subsequently, various improvements were suggested by Miyazawa and Jernigan,[69] and others.[70–72] The concept of SPs was also extended to RNA folding, and used for RNA secondary structure prediction.[73] To generate the contact statistics, we consider the non-redundant set of DNA-protein complexes available in the Protein-DNA Interface Database.[74] The PDB IDs of the complexes are listed in the Supporting Information (Table S8). The amino-acid-nucleotide contact maps were computed after coarse-graining the atomistic structure of each DNA-protein complex.

The generation of SPs for DNA-protein contacts involves the following steps. (i) Let N_{ij}^{XY} be the total number of amino-acids of type i in contact with nucleotide of type j within the non-redundant set of PDB structures. The superscript $X \in [P, B, S]$ denotes type of the DNA bead, and $Y \in [BB, SC]$ denotes the type of protein bead. For a contact to exist between X and Y, the distance of separation must be less than or equal to R_c^{DNAPRO}. (ii) The probability, P_{ij}^{XY}, that a nucleotide i is in contact with an amino-acid j is given by: $P_{ij}^{XY} = N_{ij}^{XY} / \sum_{i,j} N_{ij}^{XY}$. (iii) In the computation of SPs, the choice of the reference state is important.[73] For simplicity, we assume the random occurrence of a pair ij to be the reference state. The probability that a pair ij would be in contact by random chance within the ensemble of PDB sequences is given by: $P_R^{XY} = P_i^X P_j^Y$. (iv) Assuming that the ensemble of structures deposited within the PDB database are at quasi-equilibrium, and that Boltzmann statistics applies, the effective contact 'free energy' (or equivalently the SP), ΔG_{ij}^{XY}, between nucleotide i and amino-acid j is given by:

$$\Delta G_{ij}^{XY} = -k_B T \ln \frac{P_{ij}^{XY}}{P_R^{XY}} \tag{8}$$

where k_B is Boltzmann's constant, and T is the effective temperature. The interaction energy ϵ_k^{XY} for the kth pair (between nucleotide i and amino-acid j) is simply $-\Delta G_{ij}^{XY}$. Despite well-known limitations of this approach,[75,76] SPs have been used successfully in applications related to biomolecular structure prediction,[67,73] thermodynamics and dynamics of protein folding,[41,43] and more recently in simulations of chromosome organization.[77]

The ϵ_k^{XY} values estimated from the contact statistics are tabulated in the Supporting Information (Table S9 to S13). As an illustrative example, we show the variations in the ϵ_k^{BSC} values for contacts between different nucleobases (B) and amino-acid side-chains (SC) in Figure 1. We find that Guanine-Arginine contacts are the most favorable, whereas Guanine-Leucine contacts are least favored, in agreement with a previous statistical survey of DNA-protein interactions.[78]

The non-native interactions, ($r_i > R_c^{DNAPRO}$) between DNA and protein are taken to be purely repulsive. The interaction potential, U_{NN} is given by,

$$\begin{aligned} U_{NN} &= \sum_{i=1}^{N_{PBB}} \epsilon_{rep} \left[\left(\frac{\sigma_{i,PBB}}{r_i} \right)^6 \right] + \sum_{i=1}^{N_{SBB}} \epsilon_{rep} \left[\left(\frac{\sigma_{i,SBB}}{r_i} \right)^6 \right] \\ &+ \sum_{i=1}^{N_{BBB}} \epsilon_{rep} \left[\left(\frac{\sigma_{i,BBB}}{r_i} \right)^6 \right] + \sum_{i=1}^{N_{BSC}} \epsilon_{rep} \left[\left(\frac{\sigma_{i,BSC}}{r_i} \right)^6 \right] \\ &+ \sum_{i=1}^{N_{SSC}} \epsilon_{rep} \left[\left(\frac{\sigma_{i,SSC}}{r_i} \right)^6 \right] \end{aligned} \tag{9}$$

In the above equation, $\sigma_{i,XY} = 0.5\left(\sigma_j^X + \sigma_k^Y\right)$ is the van der Waals radius for the ith nucleotide-amino-acid pair; σ_j^X and σ_k^Y denote the radii of DNA bead j of type X, and protein bead k of type Y, respectively. The value of ϵ_{rep} is set to 1 kcal/mol. The van der Waals radii for the different DNA and protein beads are tabulated in the Supporting Information (Tables S2 and S5).

Calibrating the DNA-protein contact potential: The only free parameter in the COFFEE is λ_{DNAPRO} (Eq. 6), which tunes the overall strength of the native interactions. To calibrate the contact potential, we initiated simulations from a high-resolution crystal structure of the human TBP core domain complexed with DNA (PDB ID: 1CDW),[79] for different values of λ_{DNAPRO}. Beyond $\lambda_{DNAPRO} \approx 0.15$, there is no significant change in the root-mean-squared (RMSE) between the B factors reported in the crystal structure, and those calculated from our simulations (Fig. 2(a) and 2(c)). For $\lambda_{DNAPRO} = 0.15$, the residue-specific B factors estimated from simulations are also in good agreement with experimental values (Fig. 2(b) and 2(d)). However, for some residues, the fluctuations are enhanced, which is expected because our simulations do not strictly model the crystal environment.

Simulations: To efficiently sample the conformational space of DNA-protein complexes, we carried out Langevin Dynamics (LD) simulations in the underdamped regime. The equation of motion for bead i is, $m_i \ddot{\boldsymbol{r}}_i = -\gamma \dot{\boldsymbol{r}}_i + \boldsymbol{F}_i + \boldsymbol{g}_i$, where m_i denotes the mass of the bead, γ

denotes the frictional drag coefficient, $\boldsymbol{F}_i$ denotes the conservative force that acts on bead i as a result of interactions with other beads; $\boldsymbol{g}_i$ is a Gaussian random force, which satisfies $\langle \boldsymbol{f}_i(t)\boldsymbol{f}_j(t')\rangle = 6k_BT\gamma\delta_{ij}\delta(t-t')$. Each simulation was carried out at 298 K for 10^8 steps. Simulations were initiated from twenty different random number seeds to obtain meaningful statistics for any observable. In addition to a CPU implementation within our in-house code, we also ported COFFEE to OpenMM,[80] and exploited its optimal performance on GPUs to accelerate some of the nucleosome simulations.

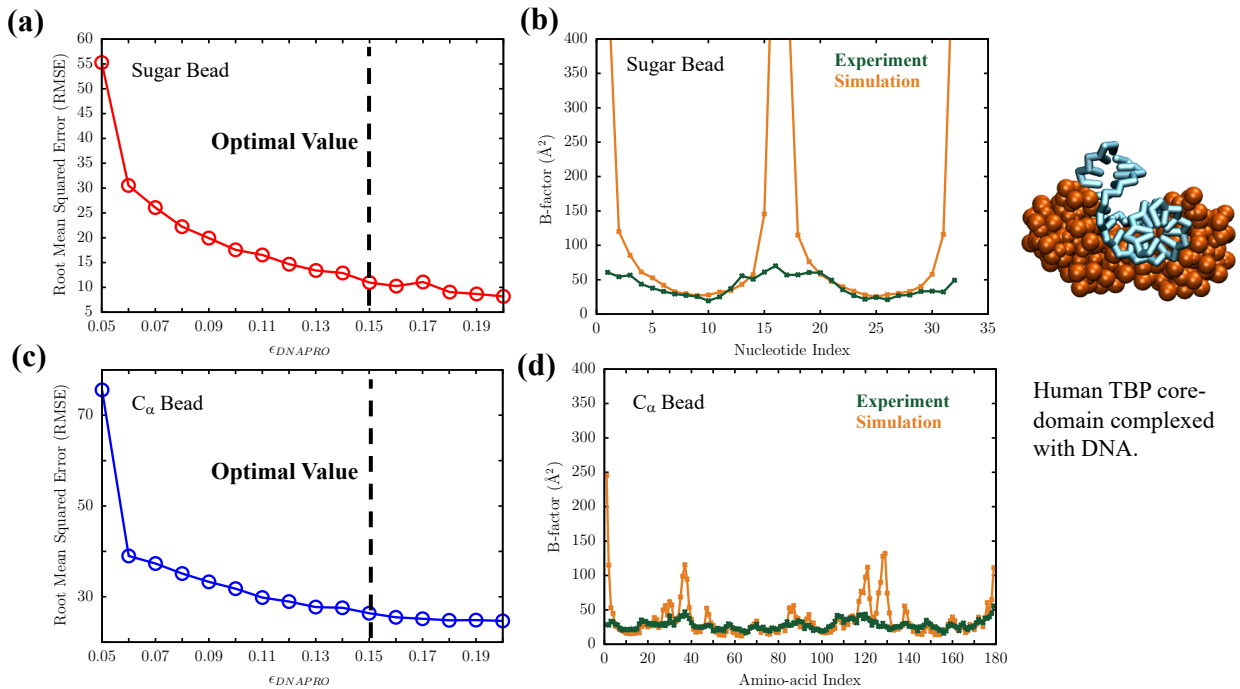


Figure 2: Calibration of the DNA-protein contact potential using a high-resolution crystal structure of the human TBP-core domain in complex with DNA. The DNA duplex is shown in cyan, and the TBP domain is rendered in brown using a space-filling representation. (a) The variation of the Root Mean Squared Error (RMSE) between experimental and simulated B-factors for the sugar bead. (b) Comparison between the crystallographic B-factors (green) and those estimated from COFFEE simulations (orange) for the sugar groups.(c) The variation of the RMSE between experimental and simulated B-factors for the C_α bead. In (a) and (c), the dashed line indicates the optimal value of $\epsilon_{DNAPRO} = 0.15$. (d) Comparison between the crystallographic C_α B-factors (green) and those estimated from simulations (orange).

SAXS profiles: The SAXS profiles for the DNA-protein complexes were computed from

the trajectories using the Debye formula,

$$I(q) = \sum_{j=1}^{N_t} \sum_{i=1}^{N_t} f_i(q) f_j(q) \frac{\sin(q r_{ij})}{q r_{ij}}, \tag{10}$$

where $f_i(q)$ and $f_j(q)$ are the q-dependent coarse-grained form factors for beads i and j, respectively. The residue-specific form factors for the protein backbone and different amino-acid side-chains were taken from Table S2 of Tong *et al.*[81] The nucleotide-specific coarse-grained form factors were derived using the independent bead approximation[82] from a database of high-resolution DNA crystal structures. Further details of this procedure are included in the Supporting Information.

Calculation of Chemical Shifts: The C_α chemical shifts were calculated from the trajectory using the LARMOR-C_α formalism.[83] This method exploits a number of geometrical features based on C_α-C_α distances, and is trained using a random forest classifier on the RefDB database. As shown previously,[83] LARMOR-C_α predicts chemical shifts in quantitative agreement with experiment, for a number of folded proteins.

Calculation of the number of unwrapped base-pairs: To probe the extent of nucleosome unwrapping at different salt concentrations, we use an order parameter introduced previously.[27] For each base pair b, we determined if it is bound to the histone core using:

$$C_b(g1, g2) = \sum_{i \in g1} \sum_{j \in g2} \frac{1 - ((|x_i - x_j|)/d_0)^n}{1 - ((|x_i - x_j|)/d_0)^m}, \tag{11}$$

where the switching function $1 - ((|x_i - x_j|)/d_0)^n / 1 - ((|x_i - x_j|)/d_0)^m$ is bound between 0 and 1; group $g1$ includes the sugar beads of the base-pairs, and $g2$ includes the C_α atoms of the histone core proteins. Following a previous study,[27] we set $d_0 = 10\,\text{Å}$, $n = 6$ and $m = 12$. C_b is rescaled to lie between 0 and 1 using another switching function,[27]

$$\phi_b = 0.5(1 + \tanh\left[\sigma\left(\langle C_b \rangle - C_0\right)\right], \tag{12}$$

where the angular brackets denote thermal averaging, $\sigma = 4.0$ and $C_0 = 1.5$.

The total number of unwrapped base pairs is given by:

$$\phi_{UBP} = N_b - \sum_{i=1}^{N_b} \phi_b \tag{13}$$

where N_b is the total number of base-pairs.

Results and Discussion

COFFEE reproduces the crystallographic B-factors for diverse DNA-protein complexes: To assess the transferability of COFFEE, we simulated six DNA-protein complexes with different sizes, topologies, and number of subunits. In all cases, the simulations were initiated from the crystal-structures deposited in the PDB database (see below).

Yeast TBP/TATA-box complex: In the crystal structure (PDB ID: 1YTB), a sharp bend near the major groove of the hairpin facilitates the binding between the TATA-box and the TATA-box binding protein (TBP).[84] The DNA-protein interface in this complex is primarily stabilized by hydrophobic contacts, which remain stable during the course of our simulations. As shown in Fig. 2(a), crystallographic B-factors for the sugar moieties, and the C_α beads of amino-acid residue are accurately reproduced by COFFEE. In the simulations, the DNA ends are unconstrained, and often fray along the trajectory, resulting in relatively high B-factors for the terminal residues. Some of the loop residues connecting different β-strands of TBP exhibit enhanced fluctuations, and are associated with higher B-factors (residues $\approx$ 34-40 and 126-130). This is likely due to the absence of crystal packing forces, which stabilize the crystal structure, resulting in suppressed motion.

TFIIB/TBP/TATA-element ternary complex: The high-resolution crystal structure (PDB ID: 1VOL) of the complex formed between human transcription factor IIB (TFIIB), TBP and the TATA element provides structural insights into the early steps of transcription initiation.[85] In TFIIB, a two-domain α-helical protein establishes contacts with the TBP

and the TATA element through extensive protein-protein and DNA-protein interactions. The B-factors of the TATA element are well reproduced by our simulations (Fig. 2(b)). The ordered secondary structure elements (helical motifs and β-strands) of the TFIIB and TBP domain are stable throughout the trajectory, with the C_α fluctuations that are comparable to the crystallographic B-factors. On the other hand, disordered loops connecting the different ordered segments, as well some of the terminal residues exhibit enhanced fluctuations. This is reflected in the relatively high B-factors in multiple stretches of residues (for example, $\approx$ 189-212 and 238-250).

Brf2-TBP complex bound to DNA: The Brf2-TBP complex bound to its natural U6 promoter (PDB ID: 4ROC) provides a molecular view of transcriptionally active pre-initiation complexes.[86] The DNA-TBP binding interface is similar to that in the TFIIB/TBP/TATA-element complex, and many studies have linked this striking resemblance to a common architectural design of the core in different initiation complexes of the transcription machinery.[87] Just as in the previous two examples, the B-factors of the DNA strands are quantitatively reproduced. In this ternary complex, the different α-helices within the cyclin repeats of Brf2 are connected by flexible loops. Several residues within these loops ($\approx$ 115-132 and $\approx$ 213-226) are associated with high B-factors. The disordered linker connecting the TBP anchor domain and a Brf2-specific small helical motif (termed as the "molecular pin") remains highly dynamic during the simulations, with a long stretch of residues ($\approx$ 250 to 303) exhibiting significant deviations from the crystal structure.

Engineered I-CreI derivative in complex with DNA from XPC gene: In this engineered complex (PDB ID: 2VBJ), the DNA-protein interface is formed between a heterodimeric derivative of I-CreI (termed as Amel3-Amel4) and 24 bp DNA from the XPC gene.[88] I-CreI is a member of the homing endonuclease family, and because of its high specificity, it can selectively cleave DNA sequences in complex genomes.[88,89] The B-factors of the sugar moieties computed from our simulations are somewhat smaller than those reported for the crystal structure, and do not exhibit position-dependent variations (Fig. 3(d)). This

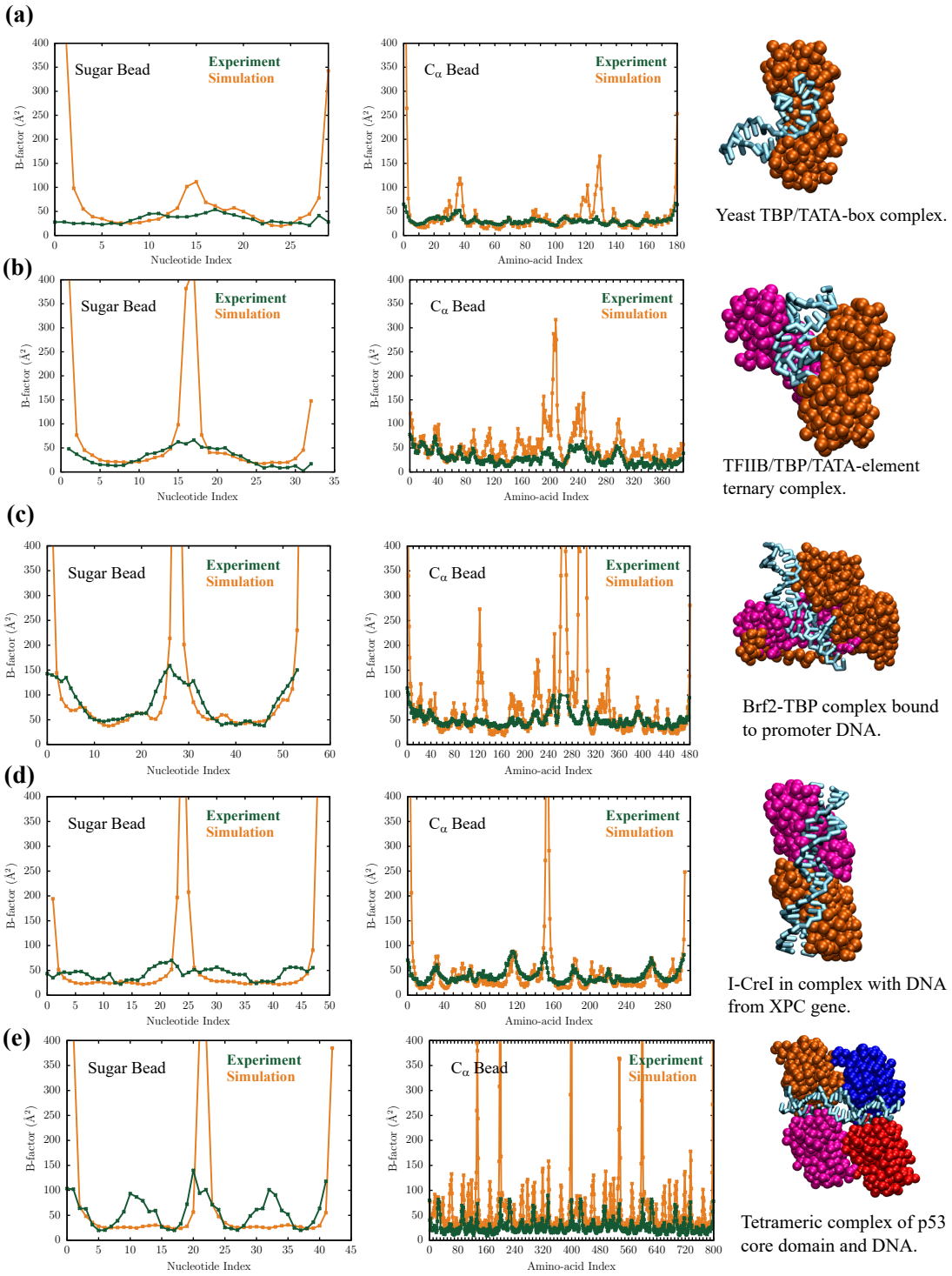


Figure 3: Comparison of the crystallographic B-factors (green) with estimates from COFFEE simulations (orange) for (a) Yeast TBP/TATA-box complex (PDB ID: 1YTB). (b) TFIIB/TBP/TATA element ternary complex (PDB ID: 1VOL). (c) Brf2-TBP complex bound to DNA (PDB ID: 4ROC). (d) I-CreI in complex with DNA from the XPC gene (PDB ID: 2VBJ). (e) Tetrameric complex of p53 with DNA (PDB ID: 4HJE). A coarse-grained representation of each complex is shown on the right. The DNA molecule is shown in cyan, and the individual protein subunits are rendered in different colors using a surface-filling representation.

could imply that our coarse-grained simulations do not fully capture the nuanced features of DNA recognition that make I-CreI and its derivatives highly specific scaffolds for genome manipulation. The Amel3-Amel4 heterodimer consists of different secondary structure motifs (α-helices and β-strands), which are connected by short loops. As is evident from Fig. 2(d), the crystallographic B-factors of the ordered elements are quantitatively reproduced in our simulations. Unlike the other examples that we have considered, in this complex close-range contacts at the DNA-protein interface also involve many residues within the loops. Hence, it is not surprising that loop residues only exhibit minimal fluctuations along the trajectory (Fig 2(d)), and stay close to their crystallographic coordinates.

Tetrameric complex of p53 with DNA: The crystal structure of the tetrameric core domain of p53 bound to a 21 bp response element (RE) from the BAX promoter gene (PDB ID: 4HJE) is somewhat unusual because unlike other p53-RE complexes,[90] a single bp spacer (G11-C32) is embedded within the middle of the DNA sequence.[91] The spacer induces local distortions within the DNA duplex, resulting in enhanced B-factors for the adjacent basepairs (Fig 3(e), green profile). Our simulations, however, do not recapitulate this trend (Fig 3(e), orange profile), and the B-factors of all nucleotides are practically similar. Chen and coworkers[91] argued that the enhanced fluctuations of nucleotides near the spacer could reflect deviations from the canonical Watson-Crick geometry, or transient base-flipping. These transitions are rather slow, occurring on the millisecond time-scale,.[92,93] Our simulations primarily probe fluctuations around the native basin, and are unlikely to capture such rare events. The residue-wise B-factor of the p53 core are predicted to be consistently higher than experiment (Fig 2(e)). Similar to previous examples, loop residues connecting different secondary structure elements (for example, residues $\approx$ 197-201), such as β-strands and α-helices, as well as terminal residues ($\approx$ 397-400) exhibit the highest fluctuations.

Previous works[94,95] suggest that a rigorous comparison to experimental B-factors entails a precise description of the crystal lattice as well as the buffer. Although these crystallization conditions were not explicitly taken into our simulation setup, the B-factor estimates are in

near quantitative agreement with those reported in the crystal structures, for most, but not all, of the examples. This is encouraging, and attests to the efficacy of our brewing protocol, which combines two independently developed CG models for proteins and DNA, with a DNA-protein interaction potential derived from contact-statistics.

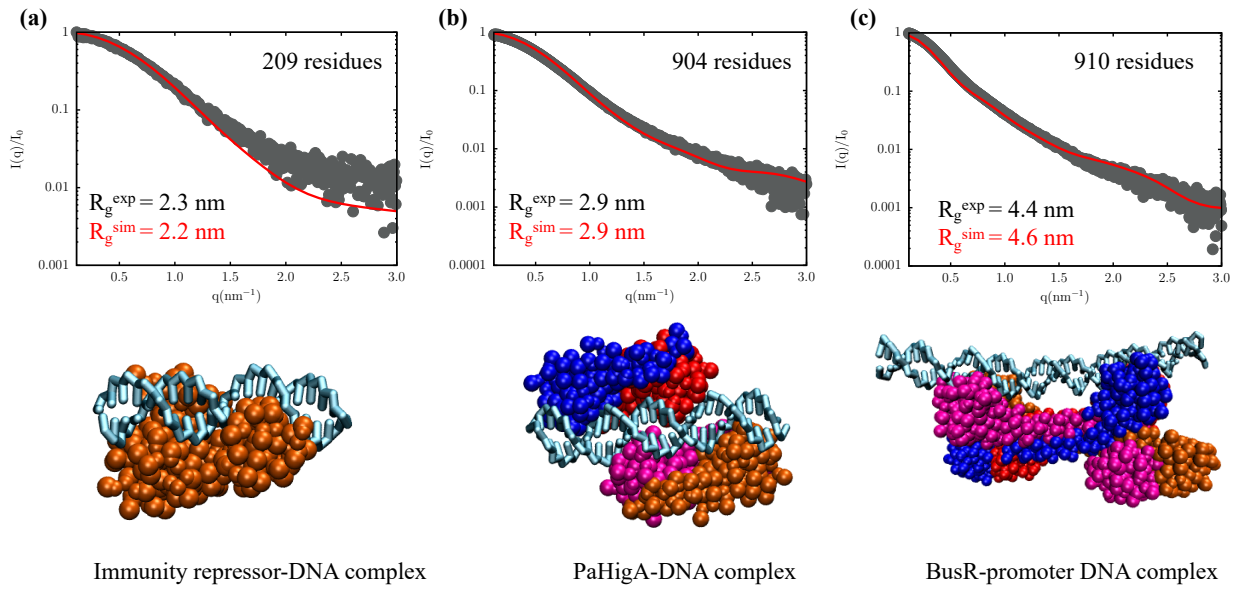


Figure 4: Structure factor calculated from simulations (red curve) and from SAXS experiments (grey points) for (a) Immunity repressor-DNA complex (PDB ID: 7R6R). (b) PaHigA-DNA complex (PDB ID: 6JPI) (c) BusR-promoter DNA complex (PDB ID: 7B5Y). Coarse-grained representations of the DNA-protein complexes are shown below the scattering profiles. In the snapshots, the DNA molecule is rendered in cyan, and the protein subunits are in different colors using the space-filling representation. For each complex, the simulations were carried out at different ionic strengths (0.5 M for (a), 0.3 M for (b) and 0.1 M for (c)) to mimic the solution conditions in the scattering experiments. In (a)-(c), R_g^{sim} and R_g^{exp} denote the radius of gyration estimated from simulations, and from a Guinier analysis of the experimental SAXS profiles, respectively. For all three complexes, these values are in excellent agreement with experiments.

Calculated scattering profiles are in quantitative agreement with SAXS experiments: Different variants of small-angle X-ray scattering (SAXS) are routinely used to probe the global dimensions of DNA-protein complexes, as well as characterize their structures at low-resolution.[96,97] We simulated three DNA-protein complexes of different sizes and topologies: the immunity repressor-DNA complex (209 residues) in which the DNA-protein binding is asymmetric and is mediated by two independent domains,[98] the PaHigA-DNA complex

(904 residues) in which helix-turn-helix motifs from the protein dimers insert into the DNA major groove to form the binding interface,[99] and the BusR-promoter (910 residues) complex where the binding to a 22 bp DNA duplex is mediated by a coiled-coil tetramer.[100] As shown in Fig. 4, the computed scattering profiles, $I(q)$, as a function of the scattering vector q are in quantitative agreement with those reported by SAXS experiments. The scattering curves in the Kratky representation ($q^2 I(q)/I(0)$ vs q) are included in the Supporting Information. In particular, the agreement is remarkable in the low q regime ($q \leqslant 1.3$ nm^{-1}), suggesting that COFFEE accurately reproduces the global dimensions of the DNA-protein complexes. The deviations from the experimental SAXS curves are rather modest even at high q values, which shows that our model accurately captures the structural details even at small length-scales. The radii of gyration calculated from simulations (R_g^{sim}s) are practically indistinguishable from the values (R_g^{exp}) reported from a Guinier analysis of the experimental SAXS profiles (Fig 4). This is striking because we did not tweak any parameter in COFFEE to obtain the reported accuracy of $I(q)$ in Fig 4.

Comparison with experimental chemical shifts: In contrast to SAXS, which is useful for elucidating the global dimensions of biomolecules, NMR provides structural and dynamical information at the atomic scale.[102] To evaluate if COFFEE faithfully captures the conformational fluctuations within a NMR-derived structural ensemble. We considered the homodimeric Lac Repressor DNA binding domain (DBD) bound to its natural operator, O1 DNA.[101] In this complex, helix-turn-helix (HTH) motifs within the DBD bind specifically to the major groove of the operator DNA. A small residue fragment adjacent to the DBD, known as the helical hinge domain (residues $\approx$ 50-58) inserts into the DNA minor groove, lending additional stability. Given its small size, this complex has been extensively studied using molecular dynamics simulations[37,103–105] to infer details of DNA-protein binding specificity in the context of transcription regulation.

We initiated simulations at 315 K using the coordinates of the NMR-derived structure deposited in the PDB database (PDB ID: 1L1M).[101] The residue-wise C_α chemical shifts

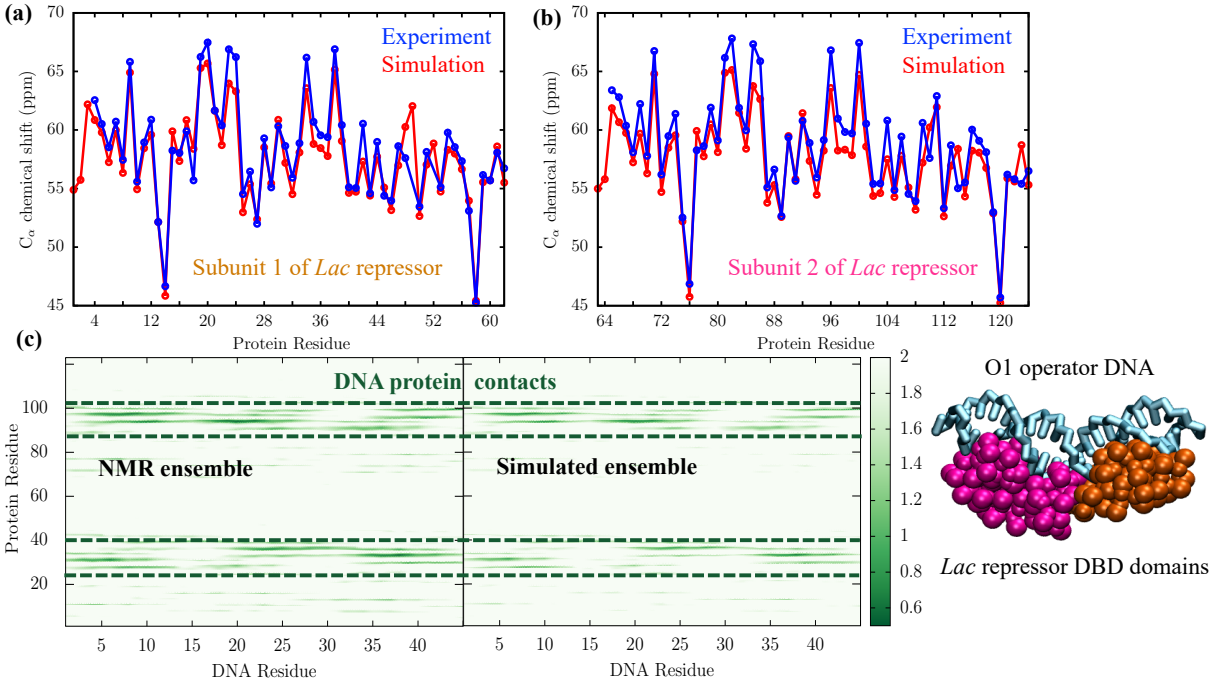


Figure 5: (a) and (b) Comparison of the experimental C_α chemical shifts (blue open circles)[101] with those computed from simulations using the LARMOR-C_α formalism (red filled circles) for the two subunits of Lac repressor-DNA binding domain (DBD) when bound to its natural operator, O1. (c) DNA-protein contact maps computed from the NMR ensemble (left) and simulations (right). Intense colors in certain regions of the contact map (demarcated by dashed lines) indicate that residues are in close contact. A coarse-grained representation of the DNA-protein complex (PDB ID: 1L1M) is also shown. The O1 operator DNA rendered in cyan. The Lac repressor is rendered using a space-filling representation, with Subunit 1 in orange, and Subunit 2 in magenta.

for the two independent subunits of the DBD domain determined using the LARMOR-C_α formalism are shown in Fig. 5(a) and 5(b). As is evident, our simulations accurately reproduce the experimental chemical shifts.[101] However, we do observe $\approx$ 2-3 ppm downshift for some residues within the HTH motifs, indicating a marginal loss of α-helical character along the trajectory. It is important to note that the DNA-protein contact potential was calibrated to reproduce the B-factors reported for a low-temperature crystal structure (Fig. 2), and without any further readjustments to the force-field parameters, we obtain quantitative agreement with NMR chemical shifts (recorded at 315 K) for a completely unrelated DNA-protein complex.

The binding of the Lac Repressor to O1 is asymmetric, with the individual DBD subunits adopting different orientations within the complex.[101] As a result, the pattern of DNA-protein contacts established by the two subunits are distinct. The observed asymmetry is visible from the DNA-protein interaction maps shown in Fig. 5(c), particularly within the dashed lines, which demarcate the interactions of HTH motif within each DBD with the major grooves of the operator. In both the NMR and the simulated interaction maps (Fig. 5(c)), the intensity of certain pixels (which denote the contact distance between a DNA and a protein residue) is clearly different for equivalent positions on the two major grooves. Strikingly, the simulated ensemble retains the key residue-residue contacts resolved using NMR. The reduced pixel intensity in certain regions of the map, however, imply that some DNA-protein interactions (present in the NMR ensemble) are transiently broken along the simulation trajectory. This is not entirely unexpected given the dynamic nature of the Lac-DNA complex, and its ability to exploit multiple binding modes.[103]

In the previous sections, we illustrated how COFFEE accurately reproduces different experimental observables, such as crystallographic B-factors, SAXS profiles, and C_α chemical shifts for diverse DNA-protein complexes. Of course, in these examples, the conformational ensembles are rather restricted and primarily include small fluctuations around the native basin. To test whether our model can describe large-scale conformational transitions with

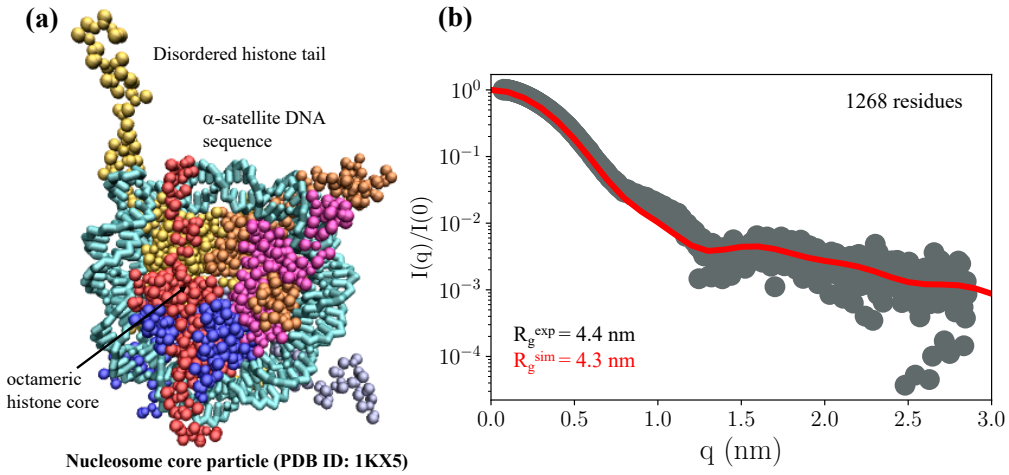


Figure 6: (a) A coarse-grained representation of the nucleosome core particle consisting of the α-satellite DNA sequence (PDB ID: 1KX5).[106] DNA is rendered in cyan, and the histone proteins, including the disordered tails, are shown in different colors using a space-filling representation. (b) Structure factor calculated from simulations (red) and SAXS experiments (gray points) at a monovalent salt concentration of 0.2 M. The experimental profile[107] corresponds to the 601 Widom sequence, which has the same number of basepairs, but a different composition than the α-satellite DNA sequence considered here. The R_g^{exp} and R_g^{sim} are the radii of gyration calculated from simulations, and from a Guinier analysis of the experimental SAXS profiles, respectively.

the same level of accuracy, we probe the salt-dependent unwrapping in nucleosome. These results are described in the following sections.

COFFEE reproduces the experimental scattering curve for the nucleosome: In eukaryotic cells, genomic DNA is highly compacted in a hierarchical fashion, and packaged into a micron sized nucleus. At the nanometer scale, packaging occurs through the formation of nucleoprotein complex known as chromatin. The nucleosome core particle (or simply the nucleosome) is the basic repeating unit. In a nucleosome, ≈ 147 bp of double-stranded DNA wraps around an octameric core of histone proteins to form a left-handed superhelix consisting of ~ 1.65 turns (Fig. 6(a)).[106,108] A canonical histone core consists of a H3-H4 tetramer, and two H2A-H2B dimers. Each histone protein consists of three α-helices connected by intervening loops.[106,108] The N-termini of each histone consists of disordered tails, while H2A also has a tail at its C-terminus. The histone tails are hotspots for post-translational modifications (PTMs), and often mediate inter-nucleosome interactions within chromatin.[108]

We simulated a nucleosome core particle consisting of 147 bp of palindromic dsDNA derived from the human α-satellite sequence repeat. The DNA helix is wrapped around *Xenopus laevis* core histones. The initial coordinates for this nucleosome sequence were taken from a high resolution crystal structure reported by Richmond and coworkers (Fig. 6(a)).[106] We compare the simulated scattering curve at 0.2 M with the experimental SAXS profile corresponding to a nucleosome reconstituted from the Widom 601 sequence.[107] Although the sequence composition of the α-satellite sequence differs greatly from Widom 601, the simulated scattering curve is almost superimposable on the experimental profile (Fig. 6(b)). Below $q \approx 0.8$ nm (the Guinier regime) the agreement between the simulated and experimental profiles is particularly impressive (Fig. 6(b)). In addition, the predicted radius of gyration (R_g^{sim}) almost coincides with the experimental estimate (R_g^{exp}), despite the differences in the DNA sequence. The minor deviations at $q \approx$ 0.8-1.0 nm could stem from the weaker positioning of the α-satellite DNA sequence (compared to the Widom 601 construct), which increases deformability at intermediate length-scales. Our observation suggests that at low salt concentration, where the nucleosome remains mostly in the wrapped configuration, the global dimension (as determined by R_g) is insensitive to the variations in the DNA sequence. Sequence-specific features probably manifest themselves at shorter length-scales.

Exploring salt-dependent nucleosome unwrapping using COFFEE: To probe nucleosome stability at different monovalent salt concentrations, we calculated $P(\phi_{UBP})$, the distributions of the number of unwrapped basepairs (see Eq. (11)-(13) for details). At $\approx$ 0.1 M, $P(\phi_{UBP})$ is extremely narrow and is centered around zero (Fig. 7), suggesting that the nucleosome prefers to be in the fully wrapped conformation, exhibiting only local conformational fluctuations. As the salt concentration is increased, electrostatic interactions between DNA and the histone core are weakened, and the population shifts towards partially unwrapped states. For instance, at $\approx$ 0.5 M, about 25 bp are detached from the histone core (Fig. 7)). The distributions also become progressively broader, suggesting that a wide array of conformations are accessible. At intermediate salt concentrations, $P(\phi_{UBP})$ curves feature

multiple peaks. These correspond to metastable states exhibiting different extent of unwrapping. We find that the ensemble of partially unwrapped states is diverse, characterized by both asymmetric (snapshot A, Fig. 7), as well as symmetric detachment (snapshot S, Fig. 7) of DNA from the histone core. Even at a very high salt concentration ($\approx 1.0\,\mathrm{M}$) , DNA does not fully detach from the histone core. Indeed, in some trajectories, the free DNA partially folds back and interacts with the disordered histone tails. There is no evidence of histone loss, although experiments[57] suggest that H2A-H2B dimers may dissociate at $\approx$ 1-1.5 M, leading to the formation of sub-nucleosomal particles.

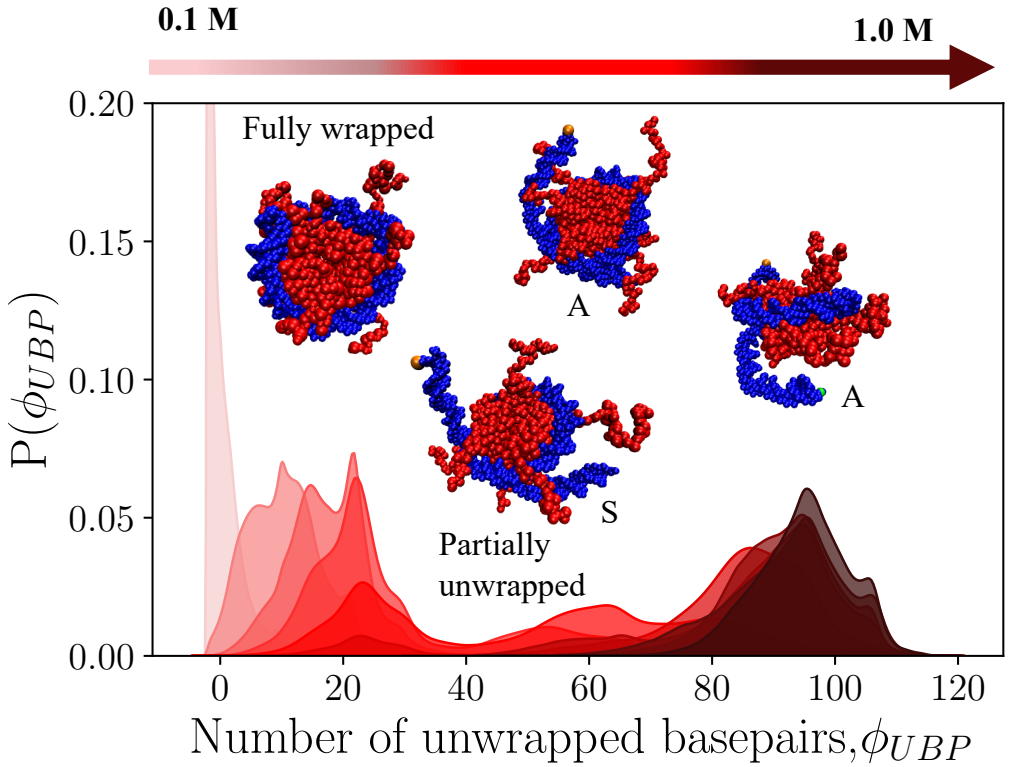


Figure 7: The probability distributions of ϕ_{UBP}, the number of unwrapped base-pairs at different salt concentrations. The nucleosome core particle consists of the human α-satellite DNA sequence (shown in blue) wrapped around an octamer of *Xenopus laevis* histones (shown in red). At low salt-concentrations, the nucleosome is in the fully wrapped configuration. Various partially unwrapped states (asymmetric (A), and symmetric (S)) are populated as the salt concentration is increased. Some representative snapshots are shown superimposed on the graph.

The stabilization of different unwrapped states with increase in salt concentration suggests that nucleosome disassembly is likely to proceed via multiple pathways, with the finer

details of the mechanism being dependent on the sequence as well as other external factors. Recent SAXS experiments, as well as ensemble optimization techniques,[56,57] which generate pools of structures that are compatible with experimental data, already hint at such a scenario.

Deletion of histone tails destabilizes the nucleosome: Within the nucleosome, the disordered histone tails often function as gatekeepers, preventing unwarranted sliding or unwrapping. They are rich in positively charged amino acids (LYS and ARG), which form specific interactions with the negatively charged phosphate backbone. Truncation or deletion of histone tails destablize the nucleosome,[109,110] affecting not only nucleosome repositioning, but also higher order chromatin organization.

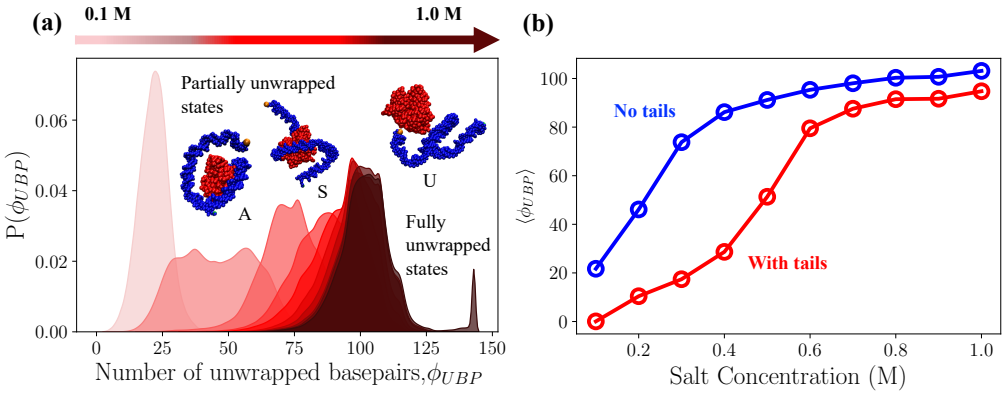


Figure 8: The effect of tail removal for a nucleosome consisting of 147 bp of human α-satellite DNA wrapped around an octameric core of *Xenopus laveis* histones. (a) The probability distributions of ϕ_{UBP}, the number of unwrapped base-pairs at different salt concentrations. DNA is rendered in blue, and the histone core in red. (b) The variation of the average number of unwrapped base-pairs, $\langle\phi_{UBP}\rangle$, with salt concentration for a nucleosome with tails (red) and without tails (blue).

The effects of histone tail deletion on nucleosome stability are accurately captured by our model. As shown in Fig. 8(a), at low salt concentrations (≈ 0.1 M), the tailless nucleosome preferentially populates partially unwrapped states. The average number of unwrapped base pairs, $\langle\phi_{UBP}\rangle$, is also consistently higher at all salt concentrations for the tailless nucleosome (Fig. 8(b)). We find that the ϕ_{UBP} distributions at intermediate salt concentrations are typically broader than those depicted in Fig. 7. This trend implies that the tailless nucleosome is

indeed more pliant, and can readily switch between alternate conformations. The unwrapped states exhibit both symmetric, as well as asymmetric DNA detachments (snapshots A and S, Fig. 8(a)). Interestingly, at ≈ 1.0,M, we observe a small peak in $P(\phi_{UBP})$, corresponding to the fully unwrapped state (snapshot U, Fig. 8(a)). Even after completely disassembly, the histone core remains completely intact, consistent with the pioneering studies of Kornberg[111] and Moudrianakis,[112] which suggest that the octamer readily breaks down only at low salt concentrations.

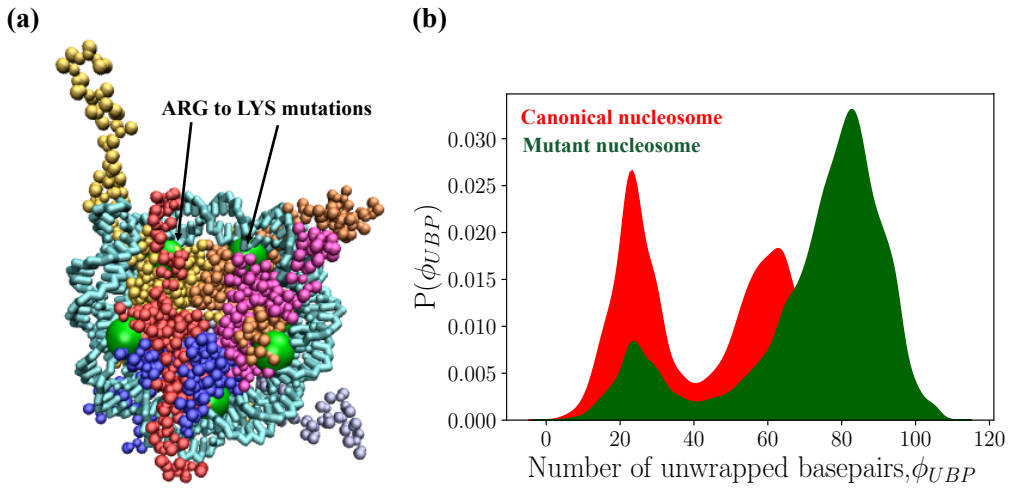


Figure 9: (a) A coarse-grained representation of the mutant nucleosome. The coloring scheme is the same as in Fig. 6(a). The green spheres denote the superhelical locations where Arginines are replaced by Lysines. (b) Probability distributions of ϕ_{UBP}, the number of unwrapped base-pairs at a salt concentration of 0.5 M, for the canonical (red) and the mutant nucleosome (green).

Mutations at superhelical locations enhance nucleosome flexibility: Besides electrostatic complementarity, the nucleosome complex is also stabilized by non-covalent interactions of chemical nature, which provide additional finesse during genome organization.[106] Among these interactions, Arginine-phosphate salt-bridges at specific superhelical locations where the nucleosomal DNA makes contacts with the histone core, are the most critical.[113,114] Indeed, disruption of these 'special' contacts are associated with SIN mutations in yeast, and are known to enhance nucleosome accessibility.[115]

To probe if COFFEE is sensitive to chemical perturbations, we simulated a mutant nucle-

osome sequence, where the Arginines at eight superhelical locations are replaced with Lysines (Fig. 9(a)). Both ARG and LYS have a charge of +1 at neutral pH, and hence changes in stability cannot be explained by electrostatic interactions alone. The ϕ_{UBP} distributions at 0.5 M for the canonical and the mutant nucleosome are shown in Fig. 9(b). Both sequences exhibit a bimodal distribution, suggesting the presence of at least two metastable states. For the canonical nucleosome, a partially unwrapped structure with $\phi_{UBP} \approx 30$ is the major state, while structures exhibiting more extensive unwrapping ($\phi_{UBP} \approx 60$) are less populated. As is evident from Fig. 9(b), the population shifts in favor of highly unwrapped states ($\phi_{UBP} \approx 80$) in the ARG→LYS mutant nucleosome, with partially unwrapped configurations ($\phi_{UBP} \approx 30$, similar to those found in the canonical nucleosome) being substantially destabilized. Hence, mutating ARG to LYS substantially weakens DNA-protein contacts and enhances nucleosome flexibility. This effect becomes particularly important in the context of CENP-A nucleosomes, in which ARG to LYS replacement facilitates proteolysis, thereby preventing promiscuous assembly, and providing a clearance mechanism from euchromatin regions.[58,59] The ARG to LYS mutation clearly does not alter the net charge of the nucleosome complex, which implies that there is at best only a minor change in the electrostatic interactions. To explain the reduced stability in the LYS mutant requires accounting for specific chemical effects, which are not considered in many CG models. In instances where sequence specificity is important, brewing COFFEE would be ideal method.

Concluding Remarks

In order to move towards a quantitative description of DNA-protein complexes, which are major components of the cellular machinery, we developed, COFFEE, a transferable coarse-grained (CG) model. Simulations based on COFFEE for a number of DNA-protein complexes demonstrate that it is a robust computational framework that takes into account the sequence-specific chemistry explicitly. The novel feature of COFFEE is that it de-

scribes DNA-protein binding using statistical potential (SP) derived from a database of high-resolution structures. Incorporation of the SP into the previously introduced TIS model for DNA[53] and the SOP-SC model for proteins[42] results in COFFEE having a only single adjustable parameter (see below). Applications to a variety of DNA-protein complexes, including the nucleosome, attests to the accuracy and transferability of COFFEE.

Force-field parameters and calibration: To brew COFFEE, we use the TIS model for DNA[53] and the SOP-SC potential[39,60] for folded proteins as the key ingredients. The TIS-DNA model,[53] developed using a 'top-down' approach, includes sequence-dependent base-pairing and base-stacking interactions, which were calibrated to reproduce the thermodynamics of hairpin melting, and dimer stacking. The electrostatic interactions are described implicitly using a Debye-Hückel potential. Importantly, the TIS-DNA quantitatively reproduce many experimentally determined observables, as shown previously.[53] The SOP-SC model for folded proteins, based on a similar conceptual framework, has been used with considerable success in studying protein folding in the presence of denaturants and pH.[40,41,43] A key ingredient of the SOP-SC energy function is a knowledge-based potential,[72] which encodes the sequence specific interaction between amino acids. These unique features of the TIS-DNA and SOP-SC force-fields, make them ideal for integration into COFFEE.

The TIS-DNA model has two adjustable parameters, which modulate the strength of the base-stacking and hydrogen-bonding interactions. The SOP-SC model has three adjustable energy scales that were adjusted to reproduce the melting temperatures of globular proteins.[39,42] Here, we used the previously determined optimal values for the TIS-DNA,[53] which reproduce the sequence-dependent melting temperatures of DNA hairpins. For the SOP-SC force-field, we adopted the parameter set that was used to predict the effect of pH on the folding thermodynamics and kinetics of ubiquitin.[42] We emphasize that TIS-DNA and SOP-SC have been combined in a modular fashion ("as is"), without requiring any re-optimization of the individual force-fields.

COFFEE is calibrated by adjusting a single parameter (λ_{DNAPRO}), describing the strength of

the non-covalent DNA-protein interactions. Strikingly, without any additional adjustments to the other force-field parameters, COFFEE faithfully reproduces the crystallographic B-factors for DNA-protein complexes of diverse shapes and sizes. Furthermore, COFFEE reproduces quantitatively the scattering profiles, as well as chemical shifts in quantitative agreement with experiments. The accuracy of COFFEE with (λ_{DNAPRO} being the only parameter) makes it an attractive transferable model for simulating arbitrary DNA-protein complexes.

Conformational ensembles of nucleosomes are consistent with experiments: As a key application of COFFEE, we probed the salt-dependent conformational changes of a nucleosome core particle consisting of the α-satellite DNA sequence wrapped around *Xenopus laevis* core histones. The simulations *quantitatively* reproduce the dimensions and the scattering curve reported at 0.2 M.[107] We also show that diverse metastable states, exhibiting different extent of DNA detachment, are populated during salt-induced unwrapping, in accord with recent findings in time-resolved SAXS experiments.[57] Interestingly, the nucleosome becomes more flexible when Arginines at certain super-helical locations are mutated to Lysines, which does not alter the electrostatic interactions. Such destabilization arise due to subtle chemical effects, which are unlikely to be captured without explicitly taking sequence effects into account. The changes in stability due to ARG→LYS mutations cannot be explained based on electrostatic interactions alone. These subtle effects are accurately described by the SP developed from contact statistics, which is an integral part of COFFEE.

Resolving sequence effects in nucleosomes with COFFEE: DNA and protein sequences dictate the conformational dynamics of nucleosomes,[116–120] as well as higher-order chromatin organization.[121] A remarkable experiment by Ha and coworkers,[120] combining optical tweezers and FRET, brought the role of sequence into the spotlight. The authors showed that the unwrapping direction could be controlled by systematically varying the local DNA sequence within the outer and inner wraps of a nucleosome. In a biological context, sequence-encoded plasticity is important for fulfilling key regulatory roles.[116] For instance,

strongly positioned sequences (akin to the Widom 601 sequence) are particularly enriched near intergenic and coding regions, where maintaining genome integrity is critical. On the other hand, weaker sequences, which unwrap easily, are abundant in highly transcribed regions.[116]

Despite these important insights, a microscopic picture of how sequence-encoded interactions predispose the nucleosome for spontaneous gaping/unwrapping, or invasion by chromatin remodellers is missing. The dust has also not completely settled on unidirectional unwrapping.[120] Whether it represents an universally preferred mode of DNA detachment in nucleosomes continues to be debated.[122] Given that COFFEE blends robust models for describing sequence-dependent properties of DNA and proteins with a knowledge-based statistical potential for DNA-protein interactions, we anticipate that it would be a suitable framework for addressing these critical questions.

Supporting Information Available

A description the SOP-SC and TIS-DNA force-fields; derivation of coarse-grained form factors using the isolated bead approximation (IBA); tabulated values of the SOP-SC and TIS-DNA force-field parameters; PDB IDs of complexes used for deriving the statistical potential (SP) for native DNA-protein contacts; Matrices denoting the effective energy scales for DNA-protein contacts; PDB IDs of DNA sequences used for deriving coarse-grained form factors; tabulated values of the coarse-grained form factors. Scripts and relevant input files for carrying out nucleosome simulations using the COFFEE framework within OpenMM are available from `https://github.com/balaka92/Nucleosome`.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Acknowledgement

We are grateful to Hung T. Nguyen and Sucheol Shin for fruitful discussions. We acknowledge the Texas Advanced Computing Center (TACC) for providing the necessary computing resources. Our work was supported by grants from the National Institutes of Health (GM-107703), National Science Foundation (CHE 2320256), as well as a grant from the Welch Foundation (F-0019) administered through the Collie-Welch Regents Chair.

The authors declare no competing interests.

References

(1) Luger, K.; Dechassa, M. L.; Tremethick, D. J. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell. Biol.* **2012**, *13*, 436–447.

(2) McGinty, R. K.; Tan, S. Nucleosome structure and function. *Chem. Rev.* **2015**, *115*, 2255–2273.

(3) Kim, K. Potential roles of condensin in genome organization and beyond in fission yeast. *J. Microbiol.* **2021**, *59*, 449–459.

(4) Donald, J. E.; Chen, W. W.; Shakhnovich, E. Energetics of protein-DNA interactions. *Nucleic Acids Res.* **2007**, *35*, 1039–1047.

(5) Gao, M.; Skolnick, J. DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.* **2008**, *36*, 3978–3992.

(6) Xu, B.; Yang, Y.; Liang, H.; Zhou, Y. An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription factor binding profiles. *Proteins* **2009**, *76*, 718–730.

(7) Yang, Y. M.; Austin, R. H.; Cox, E. C. Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys. Rev. Lett. 97*, 048302.

(8) Subekti, D. R. G.; Murata, A.; Itoh, Y.; Takahashi, S.; Kamagata, K. Transient binding and jumping dynamics of p53 along DNA revealed by sub-millisecond resolved single-molecule flourescence tracking. *Sci. Rep.* **2020**, *10*, 13697.

(9) Bagchi, B.; Blainley, P. C.; Xie, S. Diffusion constant of a nonspecifically bound protein undergoing curvilinear motion along DNA. *J. Phys. Chem. B* **2008**, *112*, 6282–6284.

(10) Lomholt, M. A.; van den Broek, B.; Kalisch, S. M. J.; Wuite, G. J. L.; Metzler, R. Faciliated diffusion with DNA coiling. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8204–8208.

(11) Bauer, M.; Metzler, R. Generalized Faciliated Diffusion Model for DNA-Binding Proteins with Search and Recognition Sites. *Biophys. J.* **2012**, *102*, 2321–2330.

(12) Berg, O. G.; Winter, R. B.; von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and Theory. *Biochemistry* **1981**, *20*, 6969–6948.

(13) Yoo, J.; Winogradoff, D.; Aksimentiev, A. Molecular dynamics simulations of DNA-DNA and DNA-protein interactions. *Curr. Opin. Struc. Biol.* **2020**, *64*, 88–96.

(14) Etheve, L.; Martin, J.; Lavery, R. Dynamics and recognition within a protein-DNA complex: a molecular dynamics study of SKN-1/DNA interaction. *Nucleic Acid Res.* **2016**, *44*, 1440–1448.

(15) Etheve, L.; Martin, J.; Lavery, R. Protein-DNA interfaces: a molecular dynamic analysis of time-dependent recognition processes for transcription factors. *Nucleic Acid Res.* **2016**, *44*, 9990–10002.

(16) Liao, Q.; Luking, M.; Kruger, D. M.; Deindl, S.; Elf, J.; Kasson, P. M.; Kamerlin, S. C. L. Long Time-Scale Atomistic Simulations of the Structure and Dynamics of Transcription Factor-DNA Recognition. *J. Phys. Chem. B* **2019**, *123*, 3576–3590.

(17) Maffeo, C.; Aksimentiev, A. Molecular mechanism of DNA association with single-stranded DNA binding protein. *Nucleic Acid Res.* **2017**, *45*, 12125–12139.

(18) Winogradoff, D.; Aksimentiev, A. Molecular Mechanism of Spontaneous Nucleosome Unraveling. *J. Mol. Biol.* **2019**, *431*, 323–335.

(19) Armeev, G. A.; Kniazeva, A. S.; Komarova, G. A.; Kirpichinikov, M. P.; Shaytan, A. K. Histone dynamics mediate DNA unwrapping and sliding in nucleosomes. *Nat. Commun.* **2021**, *12*, 2387.

(20) Tucker, M. R.; Piana, S.; Tan, D.; LeVine, M. V.; Shaw, D. E. Development of Force Field Parameters for the Simulation of Single and Double-Stranded DNA Molecules and DNA-Protein Complexes. *J. Phys. Chem. B* **2022**, *126*, 4442–4457.

(21) Hyeon, C.; Thirumalai, D. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.* **2011**, *2*, 487.

(22) Reddy, G.; Thirumalai, D. Asymmetry in histone rotation in forced unwrapping and force quench rewrapping in a nucleosome. *Nucleic Acid Res.* **2021**, *49*, 4907–4918.

(23) Chen, J.; Darst, S. A.; Thirumalai, D. Promoter melting triggered by bacterial RNA polymerase occurs in three steps. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 12523–12528.

(24) Nguyen, H. T.; Hori, N.; Thirumalai, D. Condensates in RNA repeat sequences are heterogeneously organized and exhibit reptation dynamics. *Nat. Chem.* **2022**, *14*, 775–785.

(25) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* **2012**, *116*, 8494–8503.

(26) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; de Pablo, J. J. An experimentally-informed coarse-grained 3-site-per nucleotide model of DNA: Structure, thermodynamics and dynamics of hybridization. *J. Chem. Phys.* **2013**, *139*, 144903.

(27) Tsai, M. Y.; Zhang, B.; Zheng, W.; Wolynes, P. Molecular Mechanism of Faciliated Dissociation of Fis Protein from DNA. *J. Am. Chem. Soc.* **2016**, *138*, 13497–13500.

(28) Zhang, B.; Zheng, W.; Papoian, G. A.; Wolynes, P. G. Exploring the Free Energy Landscape of Nucleosomes. *J. Am. Chem. Soc.* **2016**, *138*, 8126–8133.

(29) Tan, C.; Terakawa, T.; Takada, S. Dynamic coupling among protein binding, sliding and DNA bending revealed by molecular dynamics. *J. Am. Chem. Soc.* **2016**, *138*, 8512–8522.

(30) Mishra, G.; Levy, Y. Molecular determinants of the interactions between proteins and ssDNA Molecular determinants of the interactions between proteins and ssDNA. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 5033–5038.

(31) Mishra, G.; Bigman, L. S.; Levy, Y. ssDNA diffuses along Replication protein A via a reptation mechanism. *Nucleic Acid Res.* **2020**, *48*, 1701–1714.

(32) Ding, X.; Lin, X.; Zhang, B. Stability and folding pathways of tetra-nucleosome from six-dimensional free energy surface. *Nat. Commun.* **2021**, *12*, 1091.

(33) Alvarado, W.; Moller, J.; Ferguson, A. L.; de Pablo, J. J. Tetranucleosome interactions drive chromatin folding. *ACS Cent. Sci.* **2021**, *7*, 1019–1027.

(34) Brandner, A.; Schuller, A.; Melo, F.; Pantano, S. Exploring DNA dynamics within oligonucleosomes with coarse-grained simulations: SIRAH force field extension for protein-DNA complexes. *Biochem. Biophys. Res. Commun.* **2018**, *498*, 319–326.

(35) Honorato, R.; Roel-Touris, J.; Bonvin, A. M. J. J. MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing. *Front. Mol. Biosci.* **2019**, *6*, 102.

(36) Villa, E.; Balaeff, A.; Schulten, K. Structural dynamics of the lac repressor-DNA complex revealed by multiscale simulation. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6783–6788.

(37) Machado, M. R.; Pantano, S. Exploring LacI-DNA Dynamics by Multiscale simulations using the SIRAH force field. *J. Chem. Theory Comput.* **2015**, *11*, 5012–5023.

(38) Wassenaar, T. A.; Ingolfsson, H. I.; Prieb, M.; Marrink, S. J.; Schafer, L. V. Mixing MARTINI: Electrostatic coupling in hybrid atomistic-coarse-grained biomolecular simulations. *J. Phys. Chem. B* **2013**, *117*, 3516–3530.

(39) Liu, Z.; Reddy, G.; O'Brien, E. P.; Thirumalai, D. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 7787–7792.

(40) O Brien, E. P.; Brooks, B. R.; Thirumalai, D. Effects of pH on Proteins: Predictions for Ensemble and Single-Molecule Pulling Experiments. *J. Am. Chem. Soc.* **2012**, *134*, 979–987.

(41) Reddy, G.; Liu, Z.; Thirumalai, D. Denaturant-dependent folding of GFP. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17832–17838.

(42) Reddy, G.; Thirumalai, D. Dissecting Ubiquitin Folding Using the Self-Organized Polymer Model. *J. Phys. Chem. B* **2015**, *34*, 11358–11370.

(43) Reddy, G.; Thirumalai, D. Collapse Precedes Folding in Denaturant-Dependent Assembly of Ubiquitin. *J. Phys. Chem. B* **2017**, *121*, 995–1009.

(44) Maity, H.; Muttathukattil, A. N.; Reddy, G. Salt effects on Protein Folding Thermodynamics. *J. Phys. Chem. Lett.* **2018**, *9*, 5063–5070.

(45) Liu, Z.; Thirumalai, D. Cooperativity and Folding Kinetics in a Multidomain Protein with Interwoven Chain Topology. *ACS Cent. Sci.* **2022**, *8*, 763–774.

(46) Hyeon, C.; Thirumalai, D. Mechanical unfolding of RNA hairpins. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6789–6794.

(47) Denesyuk, N. A.; Thirumalai, D. How do metal ions direct ribozyme folding? *Nat. Chem.* **2015**, *7*, 793–801.

(48) Hori, N.; Denesyuk, N. A.; Thirumalai, D. Ion Condensation onto Ribozyme is Site Specific and Fold Dependent. *Biophys. J.* **2019**, *116*, 2400–2410.

(49) Nguyen, H. T.; Hori, N.; Thirumalai, D. Theory and simulations for RNA folding in mixtures of monovalent and divalent ions. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 21022–21030.

(50) Hori, N.; Denesyuk, N. A.; Thirumalai, D. Shape changes and cooperativity in the folding of the central domain of the 16S ribosomal RNA. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2020837118.

(51) Kumar, S.; Reddy, G. TPP Riboswitch populates Holo-Form-like structure even in the absence of cognate ligand at high Mg^{2+} concentration. *J. Phys. Chem. B* **2022**, *126*, 2369–2381.

(52) Ugrina, M.; Burkhart, I.; Muller, D.; Schwalbe, H.; Schwierz, N. RNA G-quadruplex folding is a multi-pathway process with a variety of short-lived intermediates. *biorxiv* **2022**,

(53) Chakraborty, D.; Hori, N.; Thirumalai, D. Sequence-Dependent Three Interaction Site Model for Single- and Double-Stranded DNA. *J. Chem. Theory Comput.* **2018**, *14*, 3763–3779.

(54) Kenzaki, H.; Takada, S. Partial unwrapping and histone tail dynamics in nucleosome revealed by coarse-grained molecular simulations. *PLoS Comput. Biol.* **2015**, *11*, e100443.

(55) Kono, H.; Sakuraba, S.; Ishida, H. Free energy profiles for unwrapping the outer superhelical turn of nucleosomal DNA. *PLoS Comput. Biol.* **2018**, *14*, e1006024.

(56) Chen, Y.; Tokuda, J. M.; Topping, T.; Sutton, J. L.; Meisburger, S. P.; Pabit, S. A.;

Gloss, L. M.; Pollack, L. Revealing transient structures of nucleosomes as DNA unwinds. *Nucleic Acid Res.* **2014**, *42*, 8767–8776.

(57) Chen, Y.; Tokuda, J. M.; Topping, T.; Meisburger, S. P.; Pabit, S. A.; Gloss, L. M.; Pollack, L. Asymmetric unwrapping of nucleosomal DNA propagates asymmetric opening and dissociation of the histone core. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 334–339.

(58) Conde e Silva, N.; Black, B. E.; Sivolob, A.; Filipski, J.; Cleveland, D. W.; Prunell, A. CENP-A containing Nucleosomes: Easier Disassembly versus Exclusive Centromeric Localization. *J. Mol. Biol.* **2007**, *370*, 555–573.

(59) Panchenko, T.; Sorensen, T. C.; Woodcock, C. L.; Kan, Z. Y.; Wood, S.; Resch, M. G.; Luger, K.; Englander, S. W.; Hansen, J. C.; Black, B. E. Replacement of histone H3 with CENP-A directsglobal nucleosome array condensation andloosening of nucleosome superhelical termini. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 16588–16593.

(60) Klimov, D. K.; Thirumalai, D. Mechanisms and kinetics of β-hairpin formation. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 2544–2549.

(61) Liu, Z.; Thirumalai, D. Cooperativity and Folding Kinetics in a Multidomain Protein with Interwoven Chain Topology. *ACS Cent. Sci.* **2022**, *8*, 763–774.

(62) Manning, G. S. Limiting Laws and Counterion Condensation in Polyelectrolyte Solutions I. Colligative Properties. *J. Chem. Phys.* **1969**, *51*, 924–933.

(63) Freeman, G. S.; Lequieu, J. P.; Hinckley, D. M.; Whitmer, J. K.; de Pablo, J. J. DNA shape dominates sequence affinity in nucleosome formation. *Phys. Rev. Lett.* **2014**, *113*, 168101.

(64) Lequieu, J.; Cordoba, A.; Schwartz, D. C.; de Pablo, J. J. Tension-Dependent Free Energies of Nucleosome Unwrapping. *ACS Cent. Sci.* **2016**, *2*, 660–666.

(65) Kenzaki, H.; Takada, S. Partial Unwrapping and Histone Tail Dynamics in Nucleosome Revealed by Coarse-Grained Molecular Simulations. *PLoS Comput. Biol.* **2015**, *11*, 1004443.

(66) Niina, T.; Brandani, G. B.; Tan, C.; Takada, S. Sequence-dependent nucleosome sliding in rotation-coupled and uncoupled modes revealed by molecular simulations. *PLoS Comput. Biol.* **2017**, *13*, 1005880.

(67) Tanaka, S.; Scheraga, H. A. Model of protein folding: inclusion of short- medium- and long-range interactions. *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 3802–3806.

(68) Tanaka, S.; Scheraga, H. A. Medium and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **1976**, *9*, 945–950.

(69) Miyazawa, S.; Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and unfavorable high packing density term for simulation and threading. *J. Mol. Biol.* **1996**, *256*, 623–644.

(70) Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **1990**, *213*, 859–883.

(71) Skolnick, J.; Jaroszewski, L.; Kolinski, A.; Godzik, A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximiation correct? *Protein Sci.* **1997**, *6*, 676–688.

(72) Betancourt, M. R.; Thirumalai, D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **1999**, *8*, 361–369.

(73) Dima, R. I.; Hyeon, C.; Thirumalai, D. Extracting Stacking Interaction Parameters for RNA from the Data Set of Native Structures. *J. Mol. Biol.* **2005**, *347*, 53–69.

(74) Norambuena, T.; Melo, F. The Protein-DNA Interface database. *BMC Bioinformatics* **2010**, *11*, 262.

(75) Godzik, A.; Kolinski, A.; Skolnick, J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* **1995**, *4*, 2107–2117.

(76) Thomas, D. P.; Dill, K. A. An iterative method for extracting energy-like quantities from protein structures. *J. Mol. Biol.* **1996**, *257*, 457–469.

(77) Shin, S.; Shi, G.; Thirumalai, D. A method for extracting effective interactions from Hi-C data with applications to interphase chromosomes and inverted nuclei. *biorxiv* **2023**,

(78) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino acid-base interactions: A three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **2001**, *29*, 2860–2874.

(79) Nikolov, D. B.; Chen, H.; Halay, E. D.; Hoffman, A.; Roeder, R. G.; Burley, S. K. Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 4862–4867.

(80) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PloS Comp. Biol.* **2017**, *13*, e1005659.

(81) Tong, D.; Yang, S.; Lu, L. Accurate optimization of amino acid form factors for computing small-angle X-ray scattering intensity of atomistic protein structures. *J. Appl. Crystallogr.* **2016**, *49*, 1148–1161.

(82) Yang, S.; Park, S.; Makowski, L.; Roux, B. A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes. *Biophys. J.* **2009**, *96*, 4449–4463.

(83) Frank, A. T.; Law, S. M.; Ahlstrom, L. S.; Brooks III, C. L. Predicting Protein Backbone Chemical Shifts From Cα Coordinates: Extracting High Resolution Experimental Observables from Low Resolution Models. *J. Chem Theory Comput.* **2015**, *11*, 325–331.

(84) Kim, Y.; Geiger, J. H.; Hahn, S.; Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **1993**, *365*, 512–520.

(85) Nikolov, D. B.; Chen, H.; Halay, E. D.; Usheva, A. A.; Hisatake, K.; Lee, D. K.; Roeder, R. G.; Burley, S. K. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* **1995**, *377*, 119–128.

(86) Gouge, J.; Satia, K.; Guthertz, N.; Widya, M.; Thompson, A. J.; Cousin, P.; Dergai, O.; Hernandez, N.; Vannini, A. Redox Signaling by the RNA Polymerase III TFIIB-Related Factor Brf2. *Cell* **2015**, *163*, 1375–1387.

(87) Vannini, A.; Cramer, P. Conservation between the RNA Polymerase I, II and III Transcription Initiation Machineries. *Mol. Cell* **2012**, *45*, 439–446.

(88) Redondo, P.; Prieto, J.; Munoz, I. G.; Alibes, A.; Stricher, F.; Serrano, L.; Cabaniols, J. P.; Daboussi, F.; Arnould, S.; Perez, C.; Duchateau, P.; Paques, F.; Blanco, F. J.; Montoya, G. Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature* **2008**, *456*, 107–111.

(89) Rosen, L. E.; Morrison, H. A.; Masri, S.; Brown, M. J.; Springstubb, B.; Sussman, D.; Stoddard, B. L.; Seligman, L. M. Homing endonuclease I-CreI derivatives with novel DNA target specificities. *Nucleic Acid Res.* **2006**, *34*, 4791–4800.

(90) Chen, Y.; Dey, R.; Chen, L. Crystal structure of the p53core domain bound to a full consensus site as a self-assembled tetramer. *Structure* **2010**, *18*, 246–256.

(91) Chen, Y.; Zhang, X.; Machado, A. C. D.; Ding, Y.; Chen, Z.; Qin, P. Z.; Rohs, R.; Chen, L. Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accomodate base-pair insertion. *Nucleic Acid Res.* **2013**, *41*, 8368–8376.

(92) Gueron, M.; Kochoyan, M.; Leroy, J. L. A single mode of DNA base-pair opening drives imino proton exchange. *Nature* **1987**, *328*, 89–92.

(93) Dornberger, U.; Leijon, M.; Fritzsche, H. High base pair opening rates in tracts of GC base pairs. *J. Biol. Chem.* **1999**, *274*, 6957–6962.

(94) Cerutti, D. S.; Le Trong, I.; Stenkamp, R. E.; Lybrand, T. P. Simulations of a Protein Crystal: Explicit Treatment of Crystallization Conditions Links Theory and Experiment in the StreptavidinBiotin Complex. *Biochemistry* **2008**, *47*, 12065–12077.

(95) Cerutti, D. S.; Freddolino, P. L.; Duke Jr, R. E.; Case, D. A. Simulations of a Protein Crystal with a High Resolution X-ray Structure: Evaluation of Force Fields and Water Models. *J. Phys. Chem. B* **2010**, *114*, 12811–12824.

(96) Hura, G. L.; Budworth, H.; Dyer, K. N.; Rambo, R. P.; Hammel, M.; McMurray, C. T.; Tainer, J. A. Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. *Nat. Meth.* **2013**, *10*, 453–454.

(97) Brosey, C. A.; Tainer, J. A. Evolving SAXS versatility: solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology. *Curr. Opin. Struc. Biol.* **2019**, *58*, 197–213.

(98) McGinnis, R. J.; Brambley, C. A.; Stamey, B.; Green, W. C.; Gragg, K. N.; Cafferty, E. R.; Terwilliger, T. C.; Hammel, M.; Hollis, T. J.; Miller, J. M.; Gainey, M. D.;

Wallen, J. R. A monomeric mycobacteriophage immunity repressor utilizes two domains to recognize an asymmetric DNA sequence. *Nat. Commun.* **2022**, *13*, 4105.

(99) Liu, Y.; Gao, Z.; Liu, G.; Geng, Z.; Dong, Y.; Zhang, H. Structural Insights into the Transcriptional Regulation of HigBA Toxin-Antitoxin System by Antitoxin HigA in Pseudomonas aeruginosa. *Front. Microbiol.* **2019**, *10*, 3158.

(100) Bandera, A. M.; Bartho, J.; Lammens, K.; Drexler, D. J.; Kleinschwarzer, J.; Hopfner, K. P.; Witte, G. BusR senses bipartite DNA binding motifs by a unique molecular ruler architecture. *Nucleic Acid Res.* **2021**, *49*, 10166–10177.

(101) Kalodimos, C. G.; Bonvin, A. M.; Salinas, R. K.; Wechselberger, R.; Boelens, R.; Kaptein, R. Plasticity in protein-DNA recognition: lac repressor interacts with its natural operator O1 through alternative conformations of its DNA-binding domain. *EMBO J.* **2002**, *21*, 2866–2876.

(102) Mertens, H. D. T.; Svergun, D. I. Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics. *Arch. Biochem. Biophys.* **2017**, *628*, 33–41.

(103) Furini, S.; Barbini, P.; Domene, C. DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and non-specific DNA sequence. *Nucleic Acid Res.* **2012**, *41*, 3693–3972.

(104) Liao, Q.; Luking, M.; Kruger, D. M.; Deindl, S.; Elf, J.; Kasson, P. M.; Caroline Lynn Kamerlin, S. Long Time-Scale Atomistic Simulations of the Structure and Dynamics of Transcription Factor-DNA Recognition. *J. Phys. Chem. B* **2019**, *123*, 3576–3590.

(105) Seckfort, D.; Montogomery Pettitt, B. Price of disorder in the Lac Repressor hinge helix. *Biopolymers* **2019**, *110*, e23239.

(106) Davey, C. A.; Sargent, D. F.; Luger, K.; Maeder, A. W.; Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 AA resolution. *J. Mol. Biol.* **2002**, *319*, 1097–1113.

(107) Marabelli, C.; Marrocco, B.; Piotto, S.; Chittori, S.; Picaud, S.; Marchese, S.; Ciossani, G.; Forneris, F.; Filippakopoulos, P.; Schoehn, G.; Rhodes, D.; Subramaniam, S.; Mattevi, A. *Cell Rep.* **2019**, *27*, 387–399.

(108) McGinty, R. K.; Tan, S. Nucleosome Structure and Function. *Chem. Rev.* **2015**, *115*, 2255–2273.

(109) Ferreira, H.; Somers, J.; Webster, R.; Flaus, A.; Owen-Hughes, T. Histone tails and the h3 alphan helix regulate nucleosome mobility and stability. *Mol. Cell. Biol.* **2007**, *27*, 4037–4048.

(110) Biswas, M.; Voltz, K.; Smith, J. C.; Langowski, J. Role of Histone Tails in Structural Stability of the Nucleosome. *PloS Comput. Biol.* **2011**, *7*, e1002279.

(111) Thomas, J. O.; Kornberg, R. D. A octamer of histones in chromatin and free in solution. *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 2626–2630.

(112) Eickbush, T. H.; Moudrianakis, E. N. The histone core complex: an octamer assembled by two sets of protein-protein interactions. *Biochemistry* **1978**, *17*, 4955–4964.

(113) Rohs, R.; West, S. M.; Sosinsky, A.; Liu, P.; Mann, R. S.; Honig, B. The role of DNA shape in protein-DNA recognition. *Nature* **2009**, *461*, 1248–1253.

(114) Yusufaly, T. I.; Li, Y.; Singh, G.; Olson, W. K. Arginine-phosphate salt bridges between histones and DNA: Intermolecular actuators that control nucleosome architecture. *J. Chem. Phys.* **2014**, *141*, 165102.

(115) Flaus, A.; Rencurel, C.; Ferreira, H.; Wiechens, N.; Owen-Hughes, T. Sin mutations alter inherent nucleosome mobility. *EMBO J.* **2004**, *23*, 343–353.

(116) Segal, E.; Fondufe-Mittendorf, Y.; Chen, L.; Thastrom, A.; Field, Y.; Moore, I. K.; Wang, J. P. Z.; Widom, J. A genomic code for nucleosome positioning. *Nature* **2006**, *442*, 772–778.

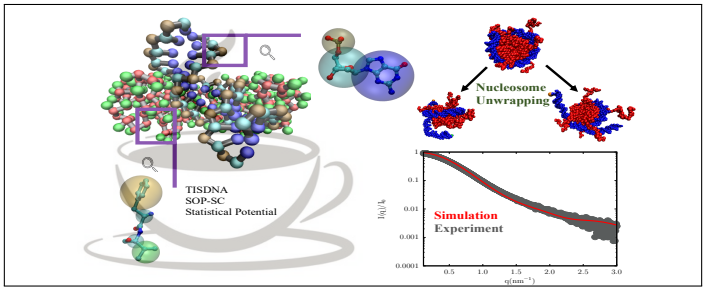
(117) Eslami-Mossallam, B.; Schiessel, H.; van Noort, J. Nucleosome dynamics: Sequence matters. *Adv. Colloid Interface Sci.* **2016**, *232*, 101–113.

(118) Culkin, J.; de Bruin, L.; Tompitak, M.; Phillips, R.; Schiessel, H. The role of DNA sequence in nucleosome breathing. *Eur. Phys. J. E. Soft Matter* **2017**, *30*, 106.

(119) Lequieu, J.; Schwartz, D. C.; de Pablo, J. J. In silico evidence for sequence-dependent nucleosome sliding. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E9197–E9205.

(120) Ngo, T. M.; Zhang, Q.; Zhou, R.; Yodh, J. G.; Ha, T. J. Asymmetric unwrapping of nucleosomes under tension directed by local DNA flexibility. *Cell* **2015**, *160*, 1135–1144.

(121) Ordu, O.; Lusser, A.; Dekker, N. H. DNA Sequence is a Major Determinant of Tetrasome Dynamics. *Biophys. J.* **2019**, *117*, 2217–2227.

(122) Zhao, D.; Le, J. V.; Darcy, M. A.; Crocker, K.; Poirier, M. G.; Castro, C.; Bundschuh, R. Quantitative Modeling of Nucleosome unwrapping from Both Ends. *Biophys J.* **2019**, *117*, 2204–2216.

Graphical TOC Entry