Supporting Information for: Exploring the Free Energy Landscape of Nucleosomes

Bin Zhang,[†] Weihua Zheng,[†] Garegin A. Papoian,[‡] and Peter G. Wolynes[*,¶,†]

Department of Chemistry, and Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States, and Department of Physics and Astronomy, Rice University, Houston, TX 77005

E-mail: pwolynes@rice.edu

Contents

[*] To whom correspondence should be addressed

[†] Department of Chemistry, and Center for Theoretical Biological Physics, Rice University, Houston, TX 77005

[‡] Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States

[¶] Department of Physics and Astronomy, Rice University, Houston, TX 77005

Protein-DNA model

We combine two coarse-grained models that have been developed separately for protein and DNA molecules in order to study the free energy landscape of nucleosomes.

The Associative memory, Water mediated, Structure and Energy Model (AWSEM) is used to study protein molecules.[1] AWSEM is a coarse-grained predictive protein force field with each amino acid represented by three atoms corresponding to C_α, C_β and O. The potential energy function for this transferable force field adopts the following form

$$V_{\text{AWSEM}} = V_{\text{backbone}} + \lambda_{\text{c}} V_{\text{contact}} + \lambda_{\text{b}} V_{\text{burial}} + \lambda_{\text{h}} V_{\text{helical}} + \lambda_{\text{fm}} V_{\text{FM}}, \tag{S1}$$

which includes both physically motivated terms and bioinformatically inspired contributions. The backbone potential V_{backbone} restricts the polymer chain to protein-like conformations by taking into account the connectivity between adjacent residues, the orientation of side chains and the dihedral angle distribution of the backbone. V_{contact}, V_{burial}, and V_{helical} are each based on a different aspect of protein physics. V_{contact} describes tertiary contact potential among amino acids separated far apart in sequence, and makes use of two different forms to model short-range and long-range interactions respectively. The short-range form represents pair-wise direct contacts formed by spatially close amino acids; the long-range part, on the other hand, is modeled with a many body potential that can be further distinguished into either water mediated or protein mediated depending on the local density of amino acids. The burial term, V_{burial}, models the preference of an amino acid to be buried inside the protein or to be on the surface with a local density dependent energy term. V_{helical} explicitly models the hydrogen bonding between the carbonyl oxygen of residue i and the amide hydrogen of residue $i+4$. The strength of the interaction depends on the helical propensity of both residues participating in the interaction. The final fragment memory term V_{FM} takes into account many-body effects that are modulated by the local sequence by making use of structural information from solved crystal structures available in the PDB database. A detailed description for each term in Eq. S1 can be found in Ref. 1.

The strengths of the transferable interactions among amino acids defined in Eq. S1 are parameterized following the energy landscape theory prescription to maximize the ratio of folding temperature over glass transition temperature for a set of training proteins. All the λ in Eq. S1 were set to one during the statistical optimization. Further fine tuning of the contribution from each individual term was shown to improve the accuracy in predicting monomer structures and dimeric interfaces.[1,2] Here we choose $\lambda_{\mathrm{c}} = 0.75$, $\lambda_{\mathrm{b}} = 1.0$, $\lambda_{\mathrm{h}} = 0.5$ and $\lambda_{\mathrm{fm}} = 0.2$.

Recently, explicit electrostatic interactions modeled at the level of Debye-Hückel were introduced into the AWSEM force field.[3] Long-range electrostatic interactions were shown to be important for binding of protein dimers. As histone proteins contain a large number of charged residues, we model explicit electrostatic interactions as well with the following expression

$$V_{\mathrm{DH}} = K_{\mathrm{Elec}} \sum_{i<j} \frac{q_i q_j}{\varepsilon_r r_{ij}} e^{-r_{ij}/l_D}, \tag{S2}$$

where q_i and q_j are charges of residue i and j that are separated by a distance r_{ij}; charge is assigned to the C_β atom of each residue in AWSEM, with $q = 1$ for arginine and lysine and $q = -1$ for aspartate and glutamate. $K_{\mathrm{Elec}} = (4\pi\varepsilon_o)^{-1} = 332.24$ kcal Å/ mol. $\varepsilon_r = 78$ is the dielectric constant of the aqueous medium, and is chosen to capture the effect of the large amount of water molecules that are present in the interior of the histone core.[4] l_D refers to the Debye-Hückel screening length, and adopts a value of 9.6Å at a ionic concentration of 100 mM.

As discussed in the main text, combining the Debye-Hückel potential V_{DH} with the original AWSEM force field introduces a double counting effect for short-range electrostatic interactions. This double counting arises because the V_{contact} term itself already includes direct contact contributions from charged residues. Thus, to remedy this issue, we further use non-additive Gō potential defined as following to stabilize the native conformation of the histone octamer found in the crystal structure[5]

$$V_{\mathrm{Go}} = -\frac{1}{2} \sum_i |E_i|^p, \tag{S3}$$

where

$$E_i = \sum_j \varepsilon_{ij}(r_{ij}) \tag{S4}$$

and

$$\varepsilon_{ij}(r_{ij}) = -\left|\frac{1}{a}\right|^{1/p} \theta(r_c - r_{ij}^{\mathrm{Nat}})\gamma_{ij} \exp\left[-\frac{(r_{ij} - r_{ij}^{\mathrm{Nat}})^2}{2\sigma_{ij}^2}\right]. \tag{S5}$$

The indices i and j each run over all C_α and all C_β atoms, and r_{ij} is the distance between atoms i and j. p controls the degree of nonadditivity and is set to 2.5, a value that was shown to provide reasonable barriers for protein folding.[6] $r_c = 8.0$ Å sets the threshold for the range of neighboring sites to be included based on the distance in the native structure r_{ij}^{Nat} together with the step function $\theta(r_c - r_{ij}^{\mathrm{Nat}})$. The well width σ_{ij} and interaction strength γ_{ij} are given by the following

$$\sigma_{ij} = |i - j|^{0.15}\,\text{Å}, \tag{S6}$$

and

$$\gamma_{ij} = \begin{cases} 0.125 & |i-j| < 5 \\ 0.5 & \text{otherwise.} \end{cases} \tag{S7}$$

a is a normalization factor given by

$$a = \frac{1}{8N}\sum_i \left|\sum_j \gamma_{ij}\theta(r_c - r_{ij}^{\mathrm{Nat}})\right|^p, \tag{S8}$$

where N is the total number of residues.

The final potential energy function of the enhanced AWSEM force field therefore is

$$V_{\mathrm{AWSEM}}^{\dagger} = V_{\mathrm{AWSEM}} + V_{\mathrm{DH}} + \lambda_{\mathrm{g}} V_{\mathrm{Go}} \tag{S9}$$

We apply the Gō potential V_{Go} to the entire histone octamer, and r_{ij}^{Nat} in Eq. S5 therefore includes both intra and inter protein contacts. For simulations of the histone octamer and the nucleosome, we set $\lambda_{\mathrm{g}} = 0.02$. This strength is tuned such that the overall Gō potential is small in magnitude when compared with the physically motivated contact potential V_{contact}; it is also tuned to produce a free energy landscape whose global minimum does not locate at the octamer conformation $Q \sim 0.8$, as shown in Figure 1 of the main text (See the *Section: Fine tuning the strength of the Gō term* for details). For the tailless nucleosome, we set $\lambda_{\text{g}} = 0.0221$. This change is necessary to ensure that the pair-wise contact energies defined in Eq. S5 have the same value for intact and tailless nucleosomes; without this change, the energies will be different since the normalization factor a is different for the two systems due to its dependence on the number of residues as defined in Eq. S8.

We use the 3-Site-Per-Nucleotide model (version 3SPN.2C) developed by the de Pablo group for DNA molecules.[7,8] This model is parameterized following a top-down strategy by reproducing experimental data including base step energies, base stacking free energies and equilibrium values of bond lengths, bend angles and dihedral angles. The predictive power of this model was shown to go beyond these experimental inputs. In particular, 3SPN model accurate reproduces the persistence length of both single stranded and double stranded DNA at varying ionic strength and for different sequences.

The protein DNA interaction is described at a non-specific level. We model the electrostatic interaction between phosphate atoms of the DNA and C_β atoms of charged protein residues using the Debye-Hückel potential defined in Eq. S2. A charge of negative one is assigned to DNA phosphate atoms when calculating electrostatic interaction with protein residues; a value of -0.6 is assigned to phosphate phosphate electrostatic interactions as in the original DNA model. This discrepancy arises due to the implicit treatment of ions; in any case, we make this choice to produce an energetic barrier of DNA unwrapping that is consistent with experimental measurements. Further Lennard-Jones interactions are included among protein and DNA atoms to provide excluded volume effect and for weak non-specific attractions.

$$U_{\text{LJ}}(r) = \begin{cases} 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] - E_{\text{cut}} & r < r_c \\ 0 & r \geq r_c \end{cases} \tag{S10}$$

with $\varepsilon = 0.125$ kJ/mol, $\sigma = 5.7$Å $r_c = 2.5\sigma$ and $E_{\text{cut}} = 4\varepsilon[(\sigma/r_c)^{12} - (\sigma/r_c)^6]$. This weak attractive Lennard-Jones interaction is included to compensate for the oversimplified treatment of protein-DNA interaction. In particular, a uniform dielectric constant $\xi = 78$ is used in the Debye-Hückel potential, while it is well known that this is too large an estimate in the vicinity of the protein.[9,10]

System preparation

We built initial configurations of computer simulation using atomic coordinates provided in the crystal structure with PDB ID: 1KX5.[5] The DNA sequence for nucleosome simulations is also obtained from the PDB. To convert the atomistic structure into coarse-grained models, we used scripts provided by the AWSEM and the 3SPN.2C package, both of which are freely available online from (https://github.com/adavtyan/awsemmd) and (http://ime.uchicago.edu/de_pablo_lab/research/dna_folding_and_hybridization/3spn.2/) respectively.

We used two definitions for histone tails based on structural information and biochemistry of protease function respectively. For the first definition, which we term as PDB-tails, histone tails are identified as peptide fragments with disordered configurations that exhibit no prominent secondary structural motif as captured in the crystal structure. This definition identifies the following fragments as histone tails: residues 1-43 of H3, 1-24 of H4, 1-20 and 116-128 of H2A, and 1-35 of H2B. For the second definition, which we term as trypsin-tails, histone tails are identified as the peptide fragments removed by trypsin treatment of histone proteins. Trypsin catalyzes protein hydrolysis and cleaves peptide chains at specific sequences.[11] In particular, the fragments cleaved by trypsin include 1 to 27 of H3, 1-20 of H4, 1-13 of H2A, and 1-17 of H2B.

PDB-tails are used for structural characterization of histone octamer. For example, all the collective variables introduced in the main text that include native contacts exclude these PDB-tails. All the secondary structure biases in the AWSEM model, including the fragment memory term in Eq. S1, also exclude these tails. Trypsin-tails are used to build the tailless nucleosome to enable direct comparison with experimental studies.

Extended simulation details

The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) was used to conduct all the reported simulations. The simulations were performed at constant temperature using Nose-Hoover thermostat under non-periodic boundary conditions. We used a time step of 20 fs and saved the coordinates of the system every 1000 steps. All the collective variables presented in the main text, except the bound DNA base pairs, were calculated on the fly along the simulation at every 100 steps. We note the time step used here is larger than what has been traditionally used for AWSEM simulations, which is typically 2-5 fs. To use this large time step without affecting the simulation accuracy, we rescaled the mass of the atoms in the protein model to 240 a.m.u. As shown in Figure S11(A), probability distributions of the various non-bonded energy terms calculated using a timestep of 20 fs are indistinguishable with the ones from a timestep of 5 fs. Though there is a noticeable difference in the distribution of bonded energies at the two different time steps, we do not expect the interfacial nonbonded contacts, which are the focus of this study, to be impacted significantly by bonded energies. Figure S11(B) indeed demonstrates that the free energy profiles for the assembly of the histone core complex calculated using either the time step 20 fs or 5 fs are statistically equivalent.

As mentioned in the main text, the umbrella sampling technique was used to construct free energy profiles. In these simulations, an additional harmonic biasing potential was added to $V^{\dagger}_{\text{AWSEM}}$. In the umbrella sampling simulations conducted for histone proteins, we further included a half-harmonic wall potential to prevent proteins from diffusing too far away. This wall potential was applied to the radius of gyration of the histone proteins with the following definition

$$U = \begin{cases} \frac{K}{2}(R_{\text{protein}} - R_o)^2 & R \geq R_o \\ 0 & R < R_o, \end{cases} \tag{S11}$$

with $K = 5$ kcal/mol/Å^2 and $R_o = 50$ Å. The radius of gyration is defined as

$$R_{\text{protein}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{\text{com}})^2}. \tag{S12}$$

i in the above equation goes through all the C_α atoms of histone proteins excluding the PDB-tails. N is the total number of selected C$_\alpha$ atoms and $\mathbf{r}_{\text{com}}$ is the center of mass of all the selected atoms. This wall potential further constrains protein concentration to approximately 3 mM as estimated in the following.

$$c = \frac{1}{V} = \frac{1}{\frac{4}{3}\pi R^3} = \frac{10^{30}}{\frac{4}{3}\pi 50^3}\frac{1}{6.02\times 10^{23}}\ \text{mM} = 3.17\ \text{mM} \tag{S13}$$

Collective variable

To characterize the DNA unwrapping from histone proteins, in addition to the radius of gyration R_{DNA} introduced in the main text, we define a collective variable that measures the number of bound DNA base pairs to the protein core.

For each base pair b, we determine whether it is bound to the protein core or not using the following measure. First, we calculate the number of protein atoms close to the given base pair,

$$C_b(\text{group1}, \text{group2}) = \sum_{i\in\text{group1}}\sum_{j\in\text{group2}}\frac{1-(|\mathbf{x}_i - \mathbf{x}_j|)/d_o)^n}{1-(|\mathbf{x}_i - \mathbf{x}_j|)/d_o)^m}. \tag{S14}$$

The switching function $(1-(|\mathbf{x}_i - \mathbf{x}_j|)/d_o)^n)(1-(|\mathbf{x}_i - \mathbf{x}_j|)/d_o)^m)$ approaches 1 for $d \ll d_o$, and is close to 0 for $d \gg d_o$. group1 includes the two sugar atoms of the base pair, and group2 includes all non-PDB-tail C$_\alpha$ atoms. We used $d_o = 10$Å, $n = 6$, and $m = 12$. Second, we introduce the following switching function to rescale $C_b(\text{group1}, \text{group2})$ into the range $[0, 1]$

$$\phi_b = 0.5(1 + \tanh[\sigma(\langle C_b\rangle - C_o)]), \tag{S15}$$

where the angle brackets indicate Boltzmann averaging. The total number of bound base pairs N_{bp} is then determined by summing ϕ_b over all the base pairs, $N_{\text{bp}} = \sum_b \phi_b$. $C_o = 1.5$ and $\sigma = 4.0$

are chosen such that in the crystal structure conformation, all the base pairs are bound in this measurement and $N_{\mathrm{bp}} \sim 147$.

Fine tuning the strength of the Gō term

As mentioned above and in the *Methods Section* of the main text, for most of our calculations, we made two modifications to the original AWSEM force field presented in Ref. 1 by including the non-additive Gō potential and the electrostatic potential among charged residues. To understand the effect of different potential energy terms on the stability of the histone core complex, we calculated the average energies as a function of the fraction of native contacts Q.

As shown in Figure S12, the AWSEM energy, which includes all the non-bonded potentials in Eq. S1, exhibits two minimums around $Q \approx 0.55$ and $Q \approx 0.75$ respectively. The energy landscape of AWSEM, i.e., the fully transferable force field in its original form without modifications, is therefore weakly funneled towards the octamer conformation. In other words, if used purely for structure prediction, as is mostly done with AWSEM, the force field will indeed do a reasonable job in predicting the octamer structure. The excellent performance of AWSEM in predicting dimer structures has indeed been demonstrated previously.[2] Not surprisingly, the average energy of the Gō potential, as shown in the middle panel, is funneled toward high Q values. By definition (see Eq. S5), the Gō potential favors conformations that resemble the octamer structure, i.e., the native state. Interestingly, the electrostatic potential energy shown in the bottom panel destabilizes the octamer conformation, and increases monotonically with Q.

Next, we evaluate the free energy cost for assembling the histone core complex as a function of the fraction of naive contacts Q with and without the Gō potential. To determine the free energy profile without the Gō term, we performed additional independent simulations of 5 million timesteps using the same protocol explained in the main text, i.e., with umbrella sampling and parallel tempering, using the potential energy function written in Eq. S9 with $\lambda_{\mathrm{g}} = 0$. As shown in Figure S13 (red line), the octamer conformation ($Q \sim 0.75$) is rather unstable when compared with the dissociated state. This is not unexpected given that electrostatic and entropic penalty associated with folding overcompensates the energetic stabilization shown in Figure S12. The free energy profile with the Gō term (yellow) is the same as the one shown in Figure 1 of the main text. Evidently, the Gō potential stabilizes the octamer conformation significantly.

Though Figure S13 argues for the significance of the Gō term, it leaves the question of its exact strength open. How can one choose the strength of the Gō term, i.e., λ_g in Eq. S9? Ideally, we would tune λ_g such that the melting temperature of the octamer complex predicted from the simulation matches experimental measurement. Unfortunately, such quantitative measurements of melting do not exist. We therefore use the following qualitative experimental observations from Ref. 12 and 13 to guide the choice of the parameter.

- The octamer conformation is stable at *high salt* (above 2 M NaCl) and *low temperature* ($4 < T < 20\ ^\circ C$).
- The octamer conformation is unstable at *high salt* (above 2 M NaCl) and *high temperature* ($T > 30\ ^\circ C$).
- The octamer conformation is unstable at *low salt* (below 2 M NaCl) and *low temperature* ($4 < T < 20\ ^\circ C$).

As shown in Figure S14, using $\lambda_g = 0.02$, our model with the fine-tuned Gō potential indeed correctly captures the various stability trends of the octamer conformation listed above. For example, in the low salt regime with an ionic concentration of 100 mM (part (A)), the octamer conformation is unstable at either high temperature $T = 300$ K or low temperature $T = 260$ K. On the other hand, in the high salt regime with an ionic concentration of 2 M (part (B)), though the octamer conformation is still unstable at high temperature $T = 300$ K, a free energy basin emerges at high Q value for low temperature $T = 260$ K and the octamer structure becomes stable. To determine the free energy profiles at the high-salt regime, we again performed additional independent simulations of 5 million timesteps using the same protocol explained in the main text, i.e., with umbrella sampling and parallel tempering, by setting the Debye-Hückel screening length l_D in Eq. S2 to 2.15 Å.

To characterize the impact of the Gō term on the mechanism of the histone core complex assembly, we further determined the fraction of native contacts for various protein-protein interfaces as a function of Q. The results presented in Figure S1 were obtained from analyzing simulations that were performed without including the Gō term, i.e., using the same data that were used to construct the red free energy profile shown in Figure S13. Comparing Figure S1(A) with Figure 2(A) of the main text, we find that the relative ordering of protein interface formation is qualitatively similar whether or not the Gō term is present. For example, in both cases, the interface (H3-H4)$_\alpha$:(H3-H4)$_\beta$ forms first, and the tetramer is always stable along the entire assembly process. Figure S1(B) further presents the probability distributions of various protein-protein interfaces at different simulation windows of the umbrella sampling. It is again clear that the tetramer (H3-H4)$_\alpha$:(H3-H4)$_\beta$ interface adopts larger Q values than do other interfaces in nearly all cases.

Convergence of the simulation

Convergence of the simulation is critical to draw any statistically significant conclusions. To evaluate convergence, we determined the statistical errors of equilibrium averages by calculating their variances from independent simulation blocks. In particular, we divided the simulated data into 5 non-overlapping blocks, and we repeated an independent analysis using the data from each individual block. The variance of the average determined from the 5 blocks is expected to decrease inversely with the simulation time of the block, and should approach zero for fully converged simulations. We plotted the squared root of the variance as error bars for equilibrium averages presented in the main text. In all cases, we find that the error bars are small compared to the features that are used to draw our conclusions.

Next, we examine the probability distributions of the various collective variables on which the harmonic biases were applied during umbrella sampling. As explained in the main text, the biasing collective variable for the histone core complex is the fraction of the native contacts Q; for the intact and the tailless nucleosome, the biasing collective variable is the radius of gyration of the DNA, R_{DNA}. Figures S15, S17 and S20 present the probability distributions of various collective variables from different simulation windows. The standard deviations of these distributions are drawn as shaded regions. Again, the errors are small compared to the topographical features of the probability distributions.

A more stringent and critical test for convergence is to investigate the probability distributions of other collective variables on which no biases were applied during simulation. This allows us to study potential *slow* variables of the system that present as challenges of equilibration during simulation without enhanced sampling. Only when the simulation time scale exceeds the relaxation time of these slow variables can we say that the simulation is converged. In Figures S16, S19 and S22, we plot the probability distribution of $Q_{\text{interface}}$ for various protein-protein interfaces from simulations conducted for the histone core complex, the intact nucleosome, and the tailless nucleosome respectively. The standard deviations of these distributions are again shown as shaded regions, and are small in magnitude compared to the important features of the probability distributions. In addition, since the nucleosome system has inherent symmetry, we can compare the probability distributions of different symmetric units to further validate the simulation convergence. For example, the two protein-protein interfaces, (H3-H4)$_\alpha$:(H2A-H2B)$_\alpha$ and (H3-H4)$_\beta$:(H2A-H2B)$_\beta$ are chemically identical, and the probability distributions of interfacial contacts for them should therefore be expected to be equivalent. Indeed, in most cases, the probability distributions, shown in red and yellow, for the two interfaces are statistically indistinguishable, thus arguing strongly for the convergence of the simulation.

For the intact and the tailless nucleosome, we further compare the probability distributions of the radius of gyration of the two segments of the DNA separated at the nucleosomal dyad. Again, the two DNA sequences are chemically identical, and the probability distributions are expected to be the same. As shown in Figures S18 and S21, the distributions are indeed statistically equivalent in most of the simulation windows.

Effect of individual histone tails on nucleosome stability

Comparing Figures 3 and 6 of the main text, we find a dramatic effect of histone tails on the stability of the nucleosome. In the simulations of the tailless nucleosome that are used to calculate Figure 6, however, all the histone tails from different proteins are removed, and the individual contribution from each tail is unclear. Here, we set out to provide a detailed characterization of the effect of each individual histone tail on nucleosome stability. To that end, we use the free energy perturbation method to calculate the free energy profiles of several mutated systems, in which the charged residues on individual histone tails have been neutralized. The free energy perturbation method allows us to calculate the free energy profile of mutated systems from the same data used to construct Figure 3 of the main text without performing any additional simulations.

The free energy of the mutated system can be calculated as

$$F_{\mathrm{mut}}(R_{\mathrm{DNA}}) = -k_B T \ln P_{\mathrm{mut}}(R_{\mathrm{DNA}}), \tag{S16}$$

where

$$P_{\mathrm{mut}}(R_{\mathrm{DNA}}) = \left\langle \delta(R_{\mathrm{DNA}} - R_{\mathrm{DNA}}(\mathbf{x})) e^{-(U_{\mathrm{mut}}(\mathbf{x}) - U_{\mathrm{wt}}(\mathbf{x}))} \right\rangle_{U_{\mathrm{wt}}}. \tag{S17}$$

$R_{\mathrm{DNA}}(\mathbf{x})$ is the radius of gyration of the DNA molecule as a function of the positions $\mathbf{x}$ of the system. $U_{\mathrm{wt}}(\mathbf{x})$ is the potential energy surface of the wild-type system (the intact nucleosome), and $U_{\mathrm{mut}}(\mathbf{x})$ is the potential energy surface for the mutated system. For $U_{\mathrm{mut}}(\mathbf{x})$, all the charges on the chosen histone tails are set to zero, and we effectively turned off all the electrostatic interactions of those tails with the rest of the system. The angle brackets indicate Boltzmann averaging over the configurations sampled using the wild-type potential energy surface. We use the trypsin-tails definition introduced in the *Section: System preparation* to select the set of residues belonging to a given histone tail to be consistent with the tailless nucleosome setup.

Figure S10 compares the free energy profiles of mutated nucleosomes with various neutralized histone tails (blue) to the one from the intact nucleosome (red). The most significant effect arises from neutralizing the two copies of the N-terminal tails from the histone protein H3, as shown in part (A). In contrast to the intact nucleosome with a single basin around $R_{\mathrm{DNA}} = 45$Å, an additional minimum now emerges at $R_{\mathrm{DNA}} = 60$ Å. The appearance of this additional basin is indeed consistent with the free energy profile shown in Figure 6 of the main text for the tailless nucleosome. Part (B) suggests that histone H4 tails have a similar effect in stabilizing the free energy basin at $R_{\mathrm{DNA}} = 60$ Å, though to a much lesser extent. The extent of involvement of histone tails from H2A and H2B in nucleosome stability is much more modest, though removing H2B tails appears to shift the basin at $R_{\mathrm{DNA}} = 45$ Å to larger values.

Periodicity of histone DNA contacts

Periodic contacts between histone proteins and the DNA have been reported.[5,14–16] For example, crystal structures of the nucleosome have revealed that, at the periodicity of 10-11 bp, the minor groove of the nucleosomal DNA will be inserted by an Arg residue. The strong electrostatic interactions formed between Arg residues and the DNA can in principle give rise to a non-uniform energy landscape with periodic minima located at the contacts formed by the Arg. Here we examine whether the coarse-grained protein-DNA model in this study captures such periodic effect.

We first note that the average number of DNA base pairs bound to histone proteins as a function of the radius of gyration of the DNA, shown as blue lines in Figures 3 and 6 of the main text, decreases in a step-wise manner. For example, in both figures, most of the DNA base pairs remain bound in the range $45 < R_{\mathrm{DNA}} < 47.5$ Å, then a sudden drop of ~ 10 bp occurs, followed by another plateau region $48 < R_{\mathrm{DNA}} < 52$ Å. The step-wise unwinding of the DNA from the nucleosome is even clearer at lower temperature. Figure S7 shows equivalent plots of Figure 3 and 6(A) of the main text, but at $T = 260$K, and now the plateau regions are highlighted by the red dashed lines as guides of the eye.

The step-wise unwrapping of the DNA would be consistent with a model that exhibits periodic contacts between histone proteins and the DNA. In this model, the energy landscape would exhibit periodic basins. While being trapped in a basin, the fraction of bound DNA base pairs will change little and will plateau. Only after overcoming the barriers between basins does the system experience a sudden unwinding of the DNA. As shown in Figure S7(C), the average contact energy between proteins and the DNA as a function of R_{DNA} does exhibit periodic basins of attraction at these regions where the bound DNA base pairs plateau.

Thermodynamics and Kinetics of DNA unwrapping

The free energy cost for unwrapping the outer layer of the nucleosomal DNA has been estimated in several studies to be around 7 to 10 kcal/mol.[17–19] To directly compare the estimated free energy cost from our simulation with these experimental measurements, we first need to have a precise definition of the state in which the outer layer has been unwound. To that end, we determined the free energy profile as a function of the pulling coordinate, which is defined as the end to end distance of the DNA molecule, i.e., the Euclidean distance between the base pair number 1 and number 147.

Figure S23(A) presents the calculated free energy profile from the same data used to constructed Figure 3 of the main text. Interestingly, we find see that the free energy profile of the intact nucleosome (part (A)) can be partitioned into two regimes at approximately 210 Å. The two regimes differ significantly in slope, and therefore the force needed to unwind the DNA. The existence of two regimes that differ in pulling force has indeed been observed experimentally, and the transition point between the two has been used to define the state with unwound outer layer. Using this experimentally motivated definition, we find that the free energy cost of unwinding the outer layer of the DNA from our simulation to be 8 kcal/mol, which is in good agreement with the experimental measurements.

For the tailless nucleosome, we find that the free energy profile exhibits a minimum located around 210 Å, as shown in part (B). This free energy basin further supports our choice to define the state for the unwound outer layer DNA. Though there does not exist an experimental estimate for the free energy cost of unwinding the tailless nucleosome, several qualitative observations suggest that the barrier should be lower in comparison to that from the intact nucleosome. For example, the binding affinity between histone proteins and the DNA was found to be smaller than the one for the intact nucleosome; the kinetic rate for unwinding the outer layer DNA was also found to be faster for the tailless nucleosome. The trend predicted from our simulations is therefore consistent with these qualitative observations.

Estimating the kinetic rate for DNA unwrapping directly from our simulation is challenging, especially given the challenge in mapping the time scale of coarse-grained models into real time units. However, if we were to use a diffusion constant of $D = 5500$ bp^2/s estimated from single molecule pulling experiments,[20] we find the rate for unwrapping the outer layer DNA for the intact nucleosome to be approximately 3.6×10^{-4}s^{-1}, which is in good agreement with reported rate 0.00038 s^{-1} from single molecule pulling experiments.[18] The rate is calculated using the expression $1/k = t = \frac{\delta l^2}{D} \exp(\Delta G/k_B T)$, with $\delta l = 5$ bp being the characteristic fluctuation in the free energy minimum. We note the diffusion constant is much smaller than the value expected from hydrodynamics analyses, suggesting ruggedness of the energy landscape.

References

(1) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. *J Phys Chem B* **2012**, *116*, 8494–8503.

(2) Zheng, W.; Schafer, N. P.; Davtyan, A.; Papoian, G. A.; Wolynes, P. G. *Proc Natl Acad Sci U S A* **2012**, *109*, 19244–19249.

(3) Tsai, M.-Y.; Zheng, W.; Balamurugan, D.; Schafer, N. P.; Kim, B. L.; Cheung, M. S.; Wolynes, P. G. *Protein Sci* **2016**, *25*, 255–269.

(4) Materese, C. K.; Savelyev, A.; Papoian, G. A. *J Am Chem Soc* **2009**, *131*, 15005–15013.

(5) Davey, C. A.; Sargent, D. F.; Luger, K.; Maeder, A. W.; Richmond, T. J. *J Mol Biol* **2002**, *319*, 1097–1113.

(6) Eastwood, M. P.; Wolynes, P. G. *J Chem Phys* **2001**, *114*, 4702–4716.

(7) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; de Pablo, J. J. *J Chem Phys* **2013**, *139*, 144903.

(8) Freeman, G. S.; Hinckley, D. M.; Lequieu, J. P.; Whitmer, J. K.; de Pablo, J. J. *J Chem Phys* **2014**, *141*, 165103.

(9) Freeman, G. S.; Lequieu, J. P.; Hinckley, D. M.; Whitmer, J. K.; de Pablo, J. J. *Phys Rev Lett* **2014**, *113*, 168101.

(10) Pitera, J. W.; Falta, M.; van Gunsteren, W. F. *Biophys J* **2001**, *80*, 2546–2555.

(11) Brower-Toland, B.; Wacker, D. A.; Fulbright, R. M.; Lis, J. T.; Kraus, W. L.; Wang, M. D. *J Mol Biol* **2005**, *346*, 135–146.

(12) Eickbush, T. H.; Moudrianakis, E. N. *Biochemistry* **1978**, *17*, 4955–4964.

(13) Ruiz-Carrillo, A.; Jorcano, J. L. *Biochemistry* **1979**, *18*, 760–768.

(14) Luger, K.; Mader, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. *Nature* **1997**, *389*, 251–260.

(15) Hall, M. A.; Shundrovsky, A.; Bai, L.; Fulbright, R. M.; Lis, J. T.; Wang, M. D. *Nat Struct Mol Biol* **2009**, *16*, 124–129.

(16) Chereji, R. V.; Morozov, A. V. *Proc Natl Acad Sci U S A* **2014**, *111*, 5236–5241.

(17) Brower-Toland, B. D.; Smith, C. L.; Yeh, R. C.; Lis, J. T.; Peterson, C. L.; Wang, M. D. *Proc Natl Acad Sci USA* **2002**, *99*, 1960–1965.

(18) Mihardja, S.; Spakowitz, A. J.; Zhang, Y.; Bustamante, C. *Proc Natl Acad Sci USA* **2006**, *103*, 15871–15876.

(19) Kruithof, M.; van Noort, J. *Biophys J* **2009**, *96*, 3708–3715.

(20) Mochrie, S. G.; Mack, A. H.; Schlingman, D. J.; Collins, R.; Kamenetska, M.; Regan, L. *Phys Rev E Stat Nonlin Soft Matter Phys* **2013**, *87*, 012710.

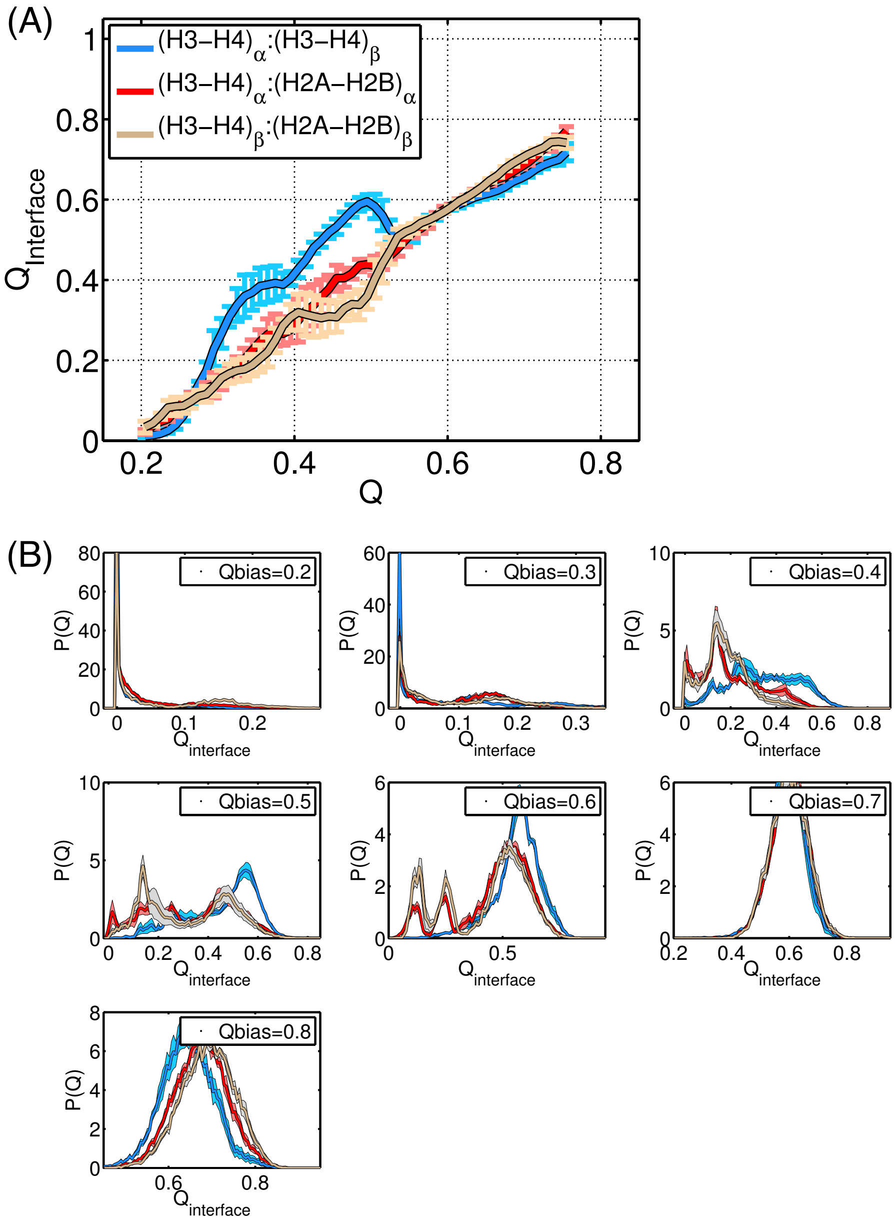

Figure S1: Mechanism for the assembly of the histone core complex without the Gō term. (A) Average fraction of native contacts for various protein-protein interfaces as a function of the global *Q*. See Figure 2(A) of the main text for a comparison. (B) Probability distributions of various protein-protein interfacial contacts at different simulation windows in which the global *Q* is biased to the target value shown in the legends. The color code is the same as in part (A). See SI *Section: Fine tuning the strength of the Gō term* for details.

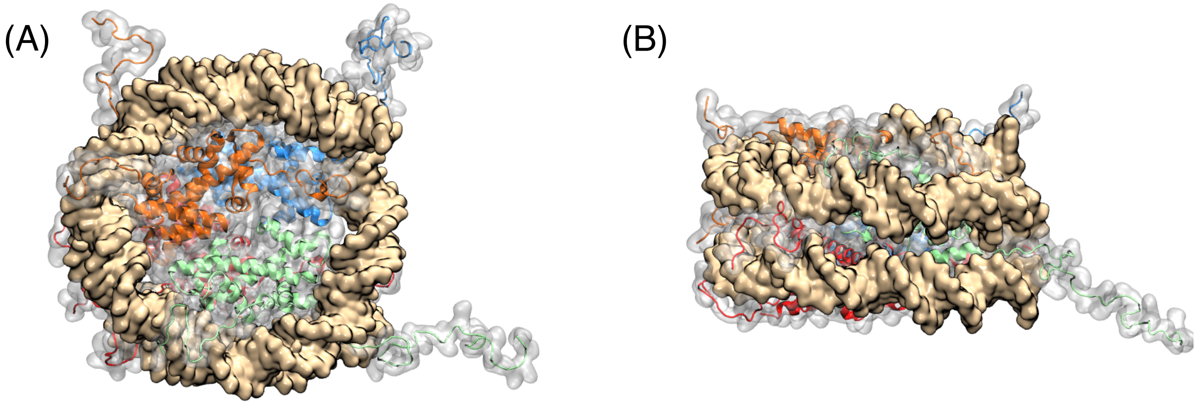

Figure S2: Structure representation of the nucleosome using atomistic coordinates recorded in the crystal structure with PDB ID: 1KX5 viewed from the top (A) and from the side (B). The DNA is shown in the Surface representation with yellow color. Proteins are drawn with the NewCartoon representation, with the two H3-H4 dimers colored in blue and green, and the two H2A-H2B dimers in red and orange. Images were produced using the software Visual Molecular Dynamics .

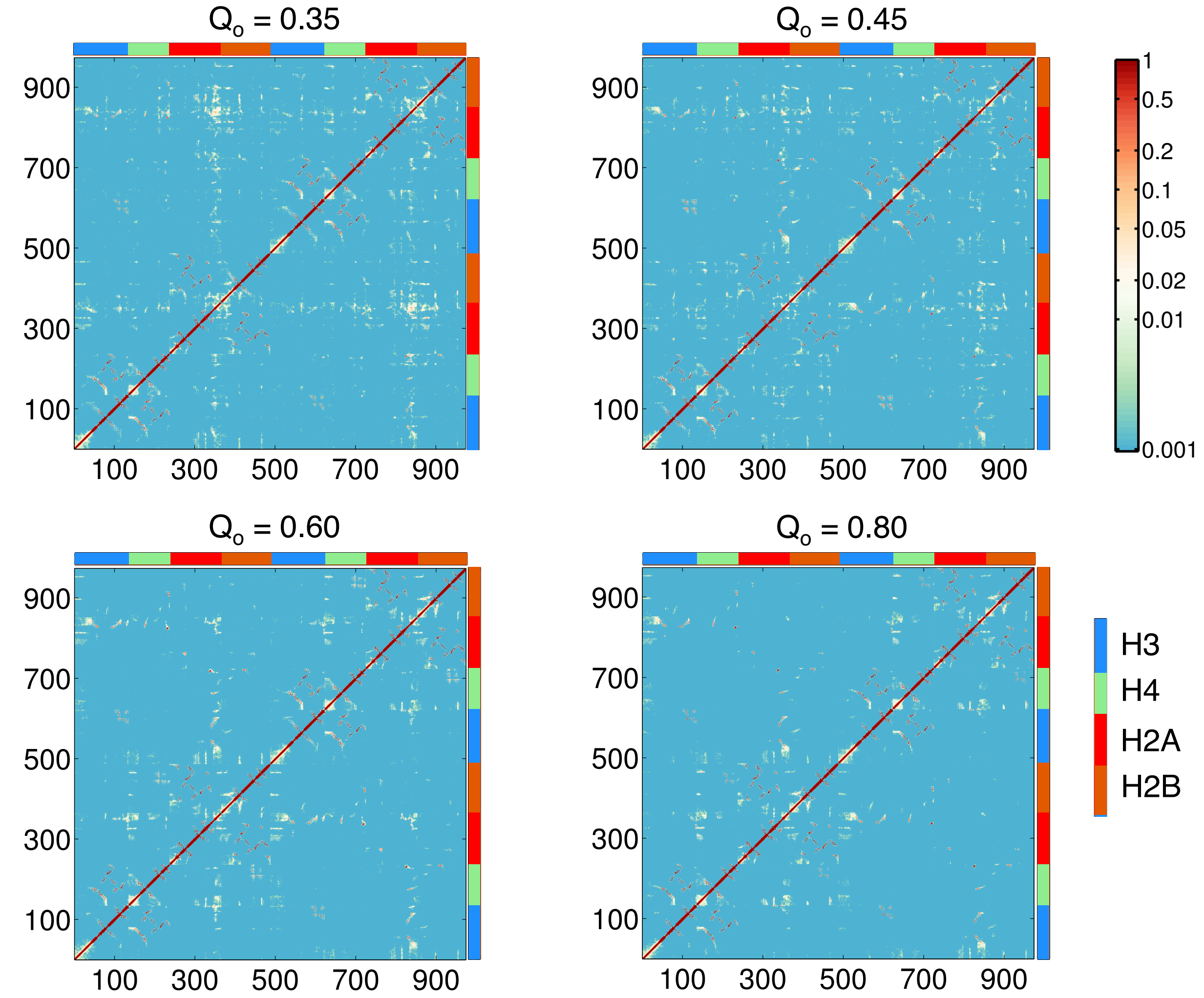

Figure S3: Average contact maps between histone protein residues calculated using ensembles of protein conformations that adopt Q values within the range $[Q_o - 0.05, Q_o + 0.05]$, with the value of Q_o indicated in the titles. A pair of residues is considered in contact if the distance between the two corresponding C_α atoms is less than 8.5 Å. For each map, 1000 configurations were selected from the data used to construct the free energy profile shown in Figure 1 of the main text. Contact probabilities are shown in log scale. The color bars next to the contact maps indicate the boundary of each protein monomer. From these contact maps, we find that at low Q values ($Q = 0.35$ and 0.45), the contacts between H2A-H2B dimers and the (H3-H4)$_2$ tetramer are rather non-specific and diffusive. For example, each H2A-H2B dimer interacts with both copies of H3-H4 dimers that form the tetramer. We note that the probability for most of the contacts formed between H2A-H2B dimers and the (H3-H4)$_2$ tetramer is rather small at low Q values, suggesting that these contacts are transient and short-lived. These non-specific contacts begin to disappear at high Q values as more stable interactions form.

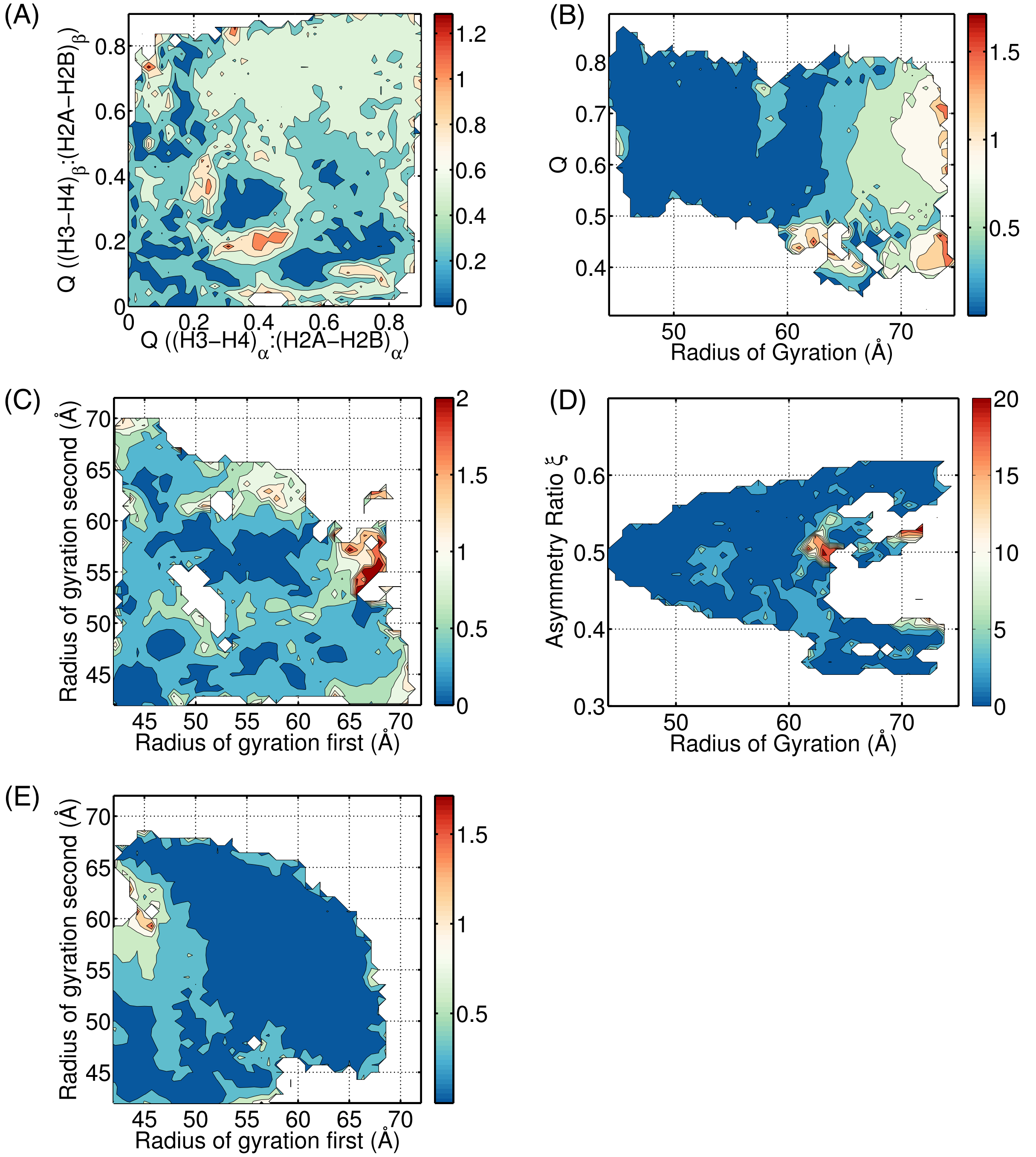

Figure S4: Standard deviation of the two-dimensional figures shown in the main text, with part A for Figure 2(B), part B for Figure 4(A), part C for Figure 5(B), part D for Figure 5(C) and part E for Figure 6(C). These variances are generally small compared to the main features of the figures shown in the main text.

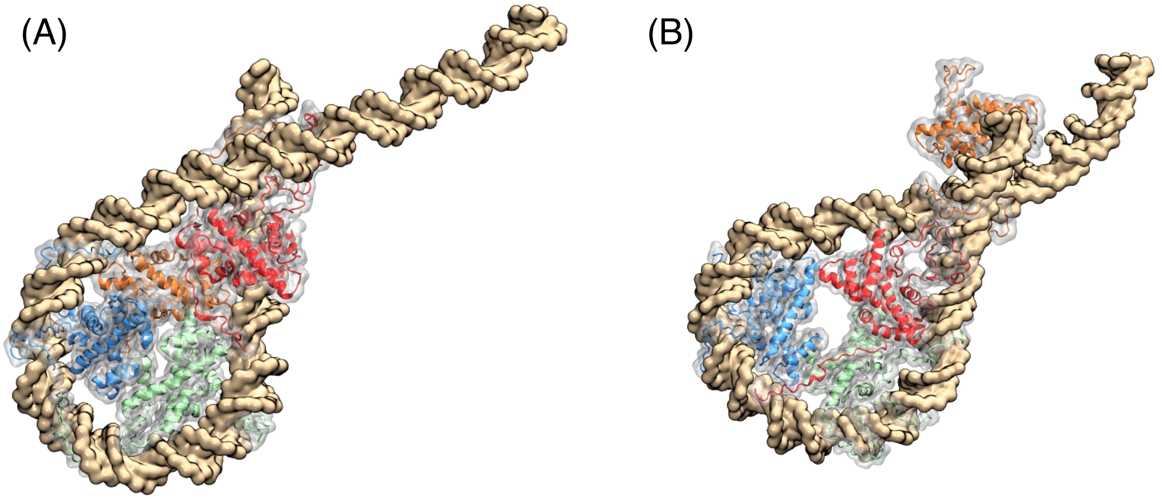

Figure S5: Example configurations of the nucleosome at large R_{DNA} in which the histone core complex has a Q value of approximately 0.35. The coloring scheme is the same as the one shown in Figure S2.

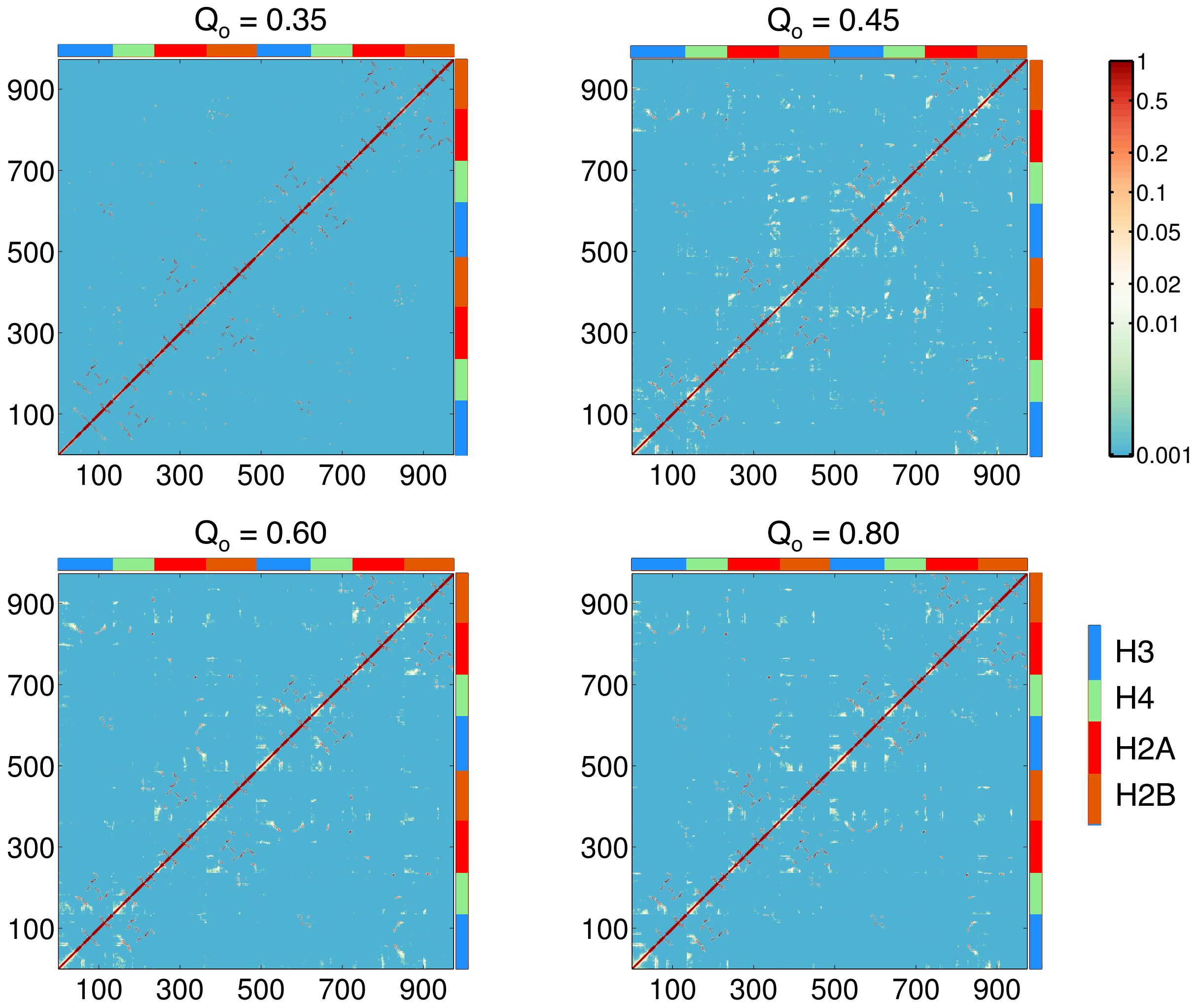

Figure S6: Average contact maps between histone protein residues calculated using ensembles of protein conformations that adopt Q values within the range $[Q_o - 0.05, Q_o + 0.05]$, with the value of Q_o indicated in the titles. A pair of residues is considered in contact if the distance between the two corresponding C$_\alpha$ atoms is less than 8.5 Å. For all the three maps with $Q_o = 0.45, 0.60, 0.80$, 1000 configurations were selected from the data used to construct the free energy profile shown in Figure 3 of the main text. For the map with $Q_o = 0.35$, only 12 frames were found, and thus the estimated contact probabilities are less accurate. Overall, these contact maps exhibit similar patterns as the ones shown in Figure S3 for the histone core complex simulated without the presence of the DNA.

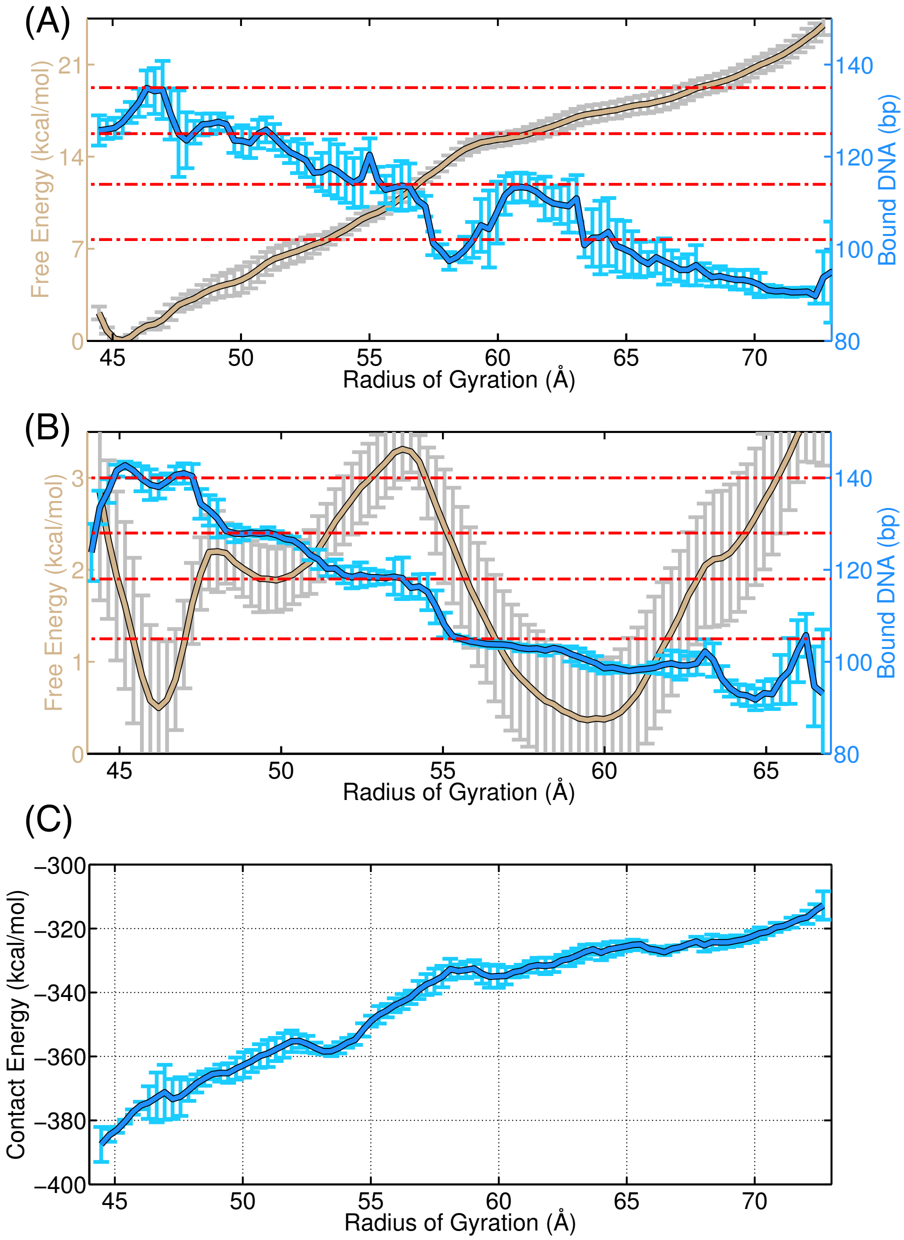

Figure S7: Step-wise unwinding of the DNA molecule from histone proteins. In both parts (A) and (B), the yellow line corresponds to the free energy profile as a function of the DNA radius of gyration R_{DNA}, while the blue line measures the average number of DNA base pairs bound to histone proteins. These two figures are similar to the ones shown in Figure 3 and Figure 6 of the main text but for $T = 260\mathrm{K}$ instead of $T = 300\mathrm{K}$. The red lines highlight the regions of R_{DNA} in which the bound DNA base pairs exhibit minimal fluctuations. (C) The average contact energy between histone proteins and the DNA molecule as a function of R_{DNA}. The contact energy exhibit clear basins in the regions in which the bound DNA base pairs plateau. See SI *Section: Periodicity of histone DNA contacts* for details.

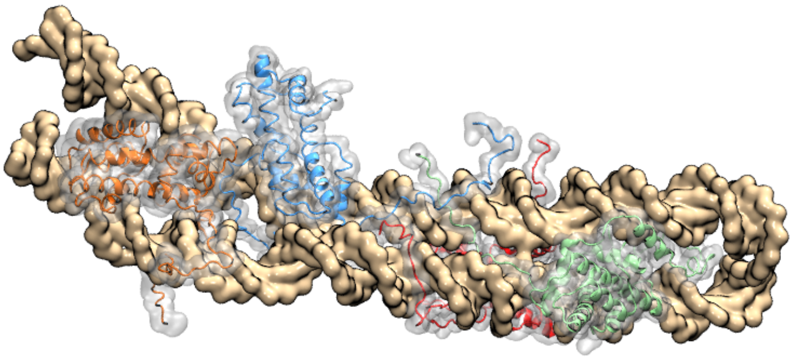

Figure S8: An example nucleosome configuration with $\xi \sim 0.5$ in which the DNA is seen to fold back onto itself while being screened by a completely dissociated histone protein core. These configurations have low electrostatic energy, and contribute to the blue region for R_{DNA} between 62 and 70 Å in Figure 5(C) of the main text. These configurations have a relatively high free energy cost despite their low electrostatic energy and only appear occasionally in the umbrella sampled simulations.

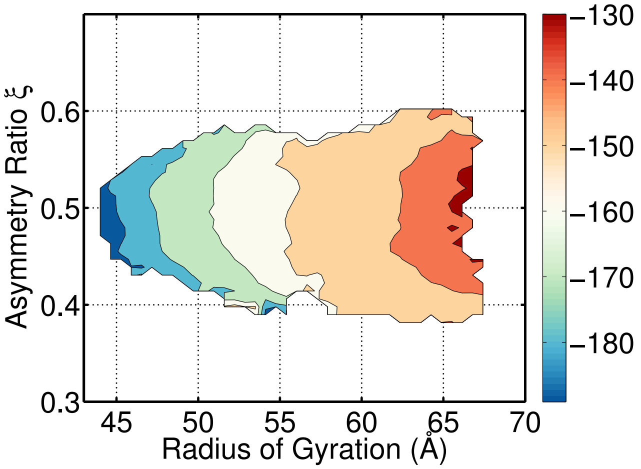

Figure S9: Average protein DNA electrostatic interactions as a function of the DNA radius of gyration R_{DNA} and the asymmetry measure ξ for the tailless nucleosome. Energies in kcal/mol.

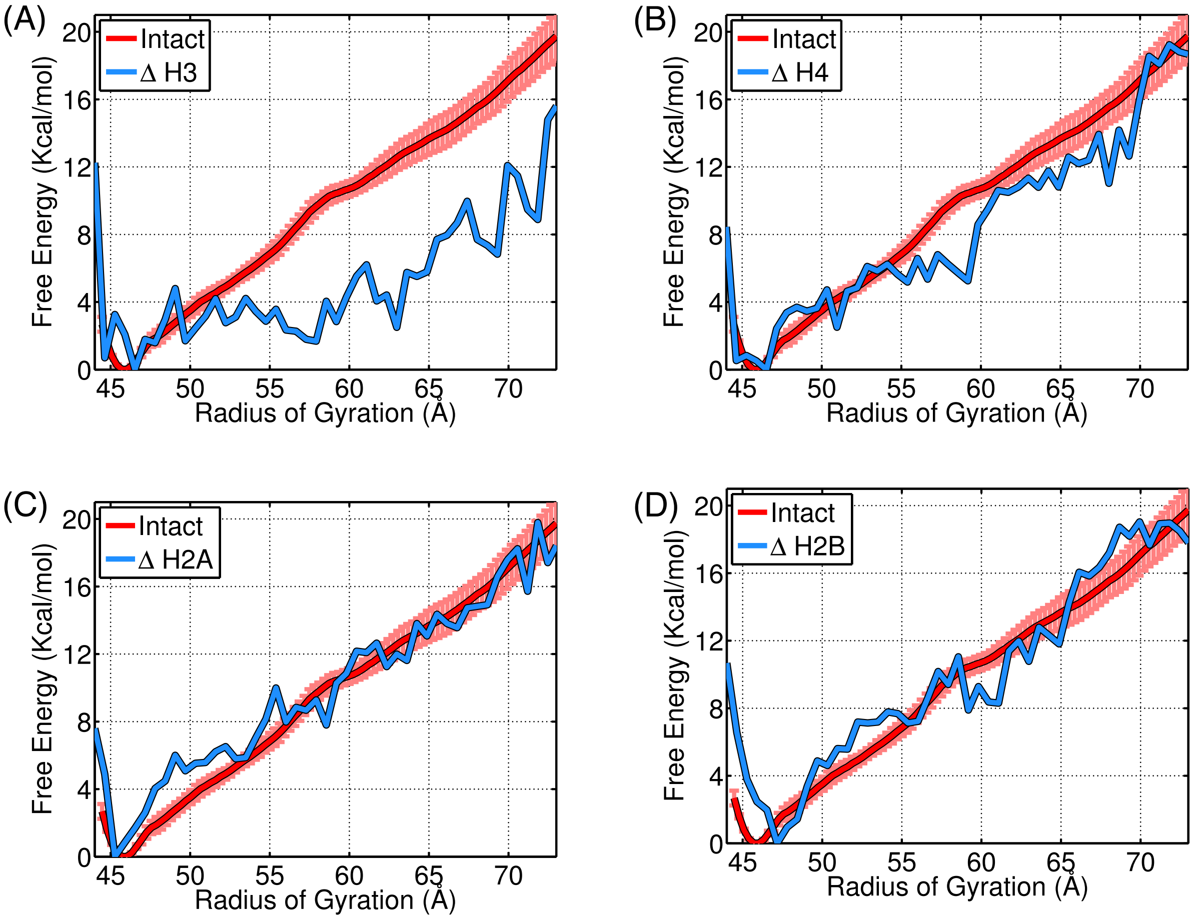

Figure S10: Effect of the various histone tails on the stability of the nucleosome. In each figure, the free energy profile of the intact nucleosome is identical to the one presented in Figure 3 of the main text. The blue lines are the free energy profiles for a system in which the charges on the histone H3 tails (A), H4 tails (B), H2A tails (C), and H2B tails (D) are removed. These free energy profiles for individual histone tails were determined using a free energy perturbation method from the same data used to construct the red curve. See SI *Section: Effect of individual histone tails on nucleosome stability* for details.

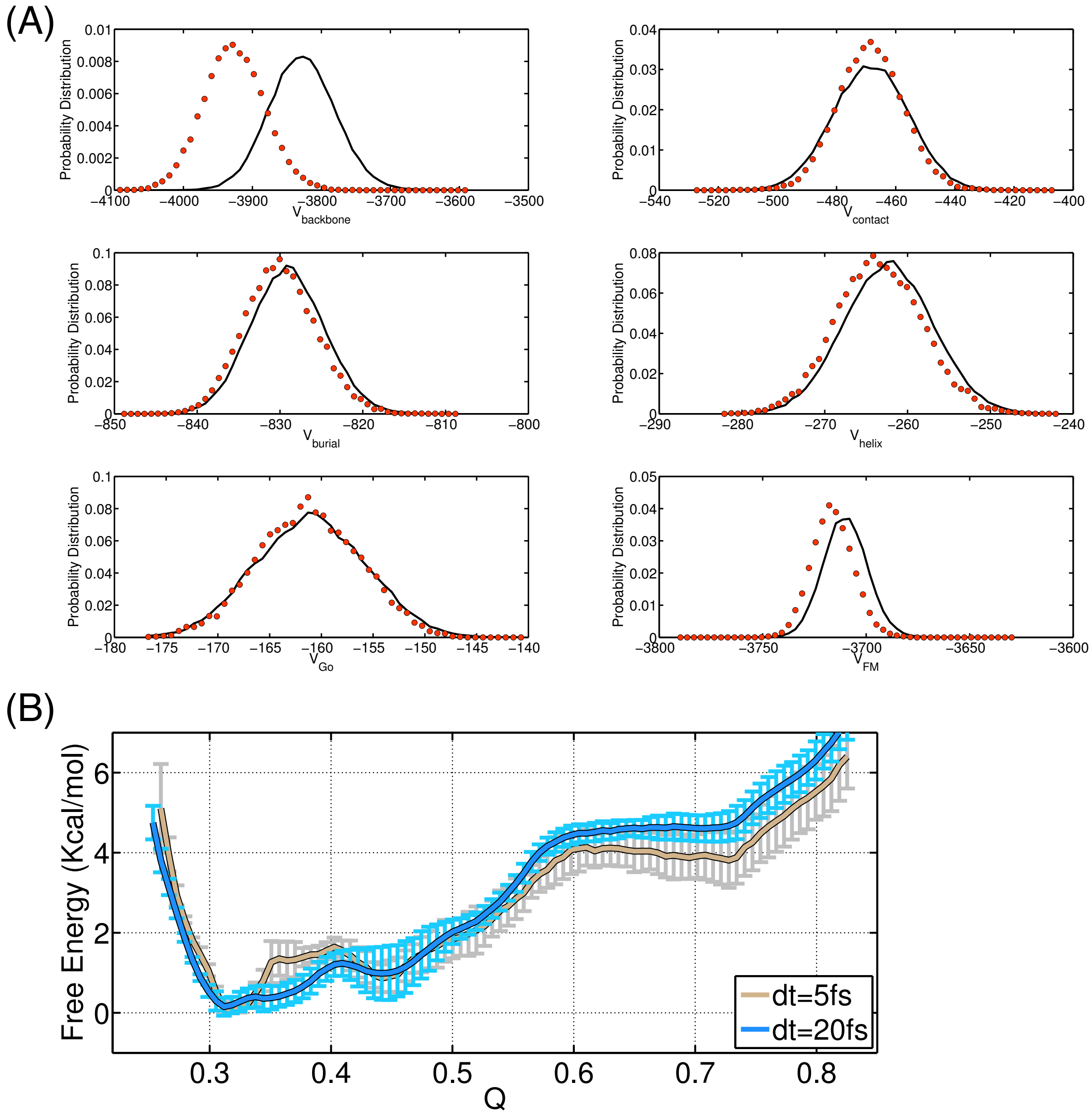

Figure S11: (A) Probability distribution of various energy terms calculated from simulations performed for the intact nucleosome with a time step of 20 fs (black) and with a time step of 5 fs (red). (B) Comparison of the free energy profiles calculated with a time step of 5 fs (yellow) and 20 fs (blue) for the assembly of the histone core complex at the ionic concentration of 2 mM (high salt condition).

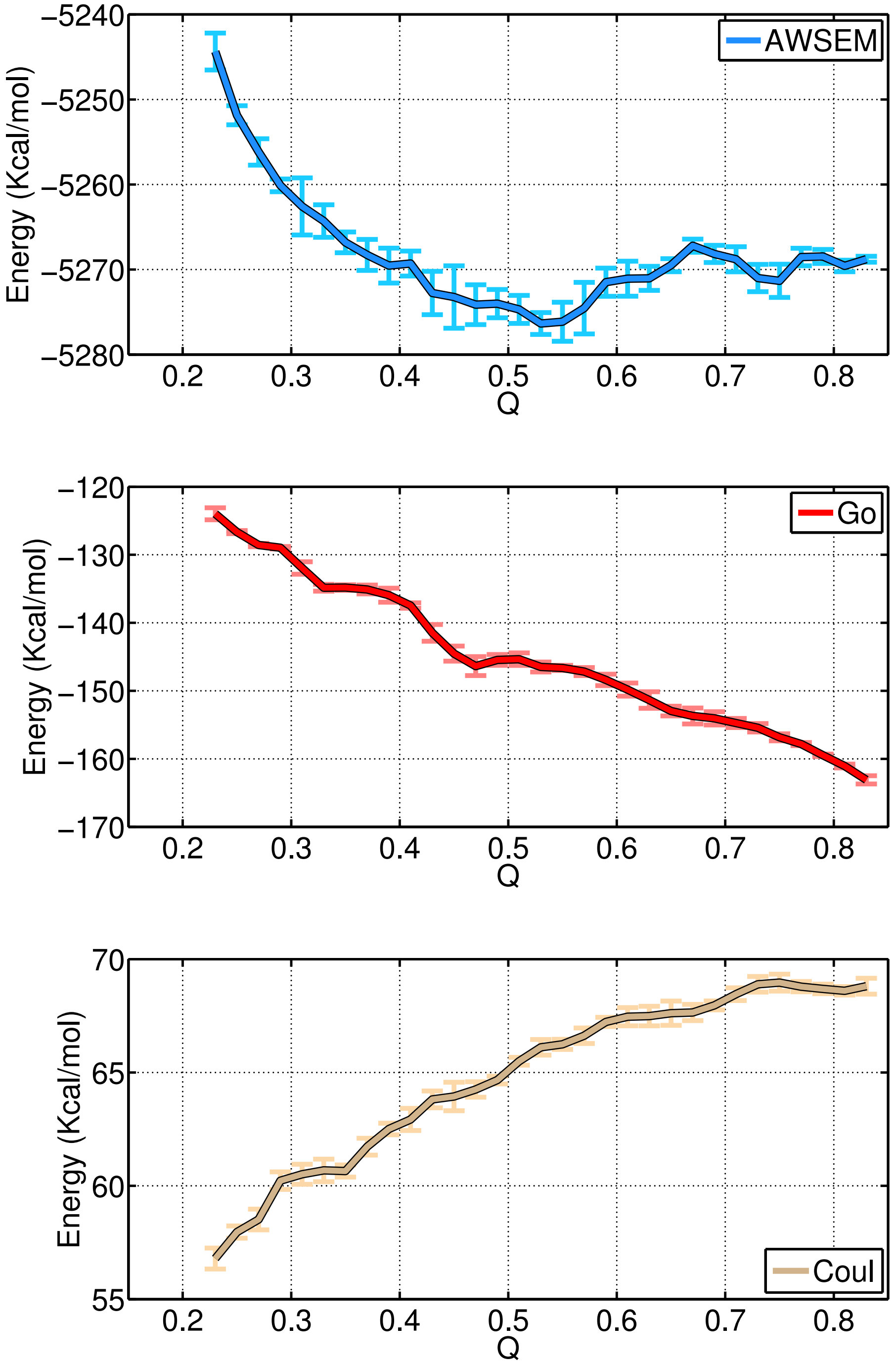

Figure S12: Average energies as a function of the fraction of native contacts Q for the non-bonded AWSEM potential (*top panel*), the Gō potential (*middle panel*) and the electrostatic potential (*bottom panel*). These energies were calculated from the same data used to determine the free energy profile for the assembly of the histone core complex shown in Figure 1 of the main text.

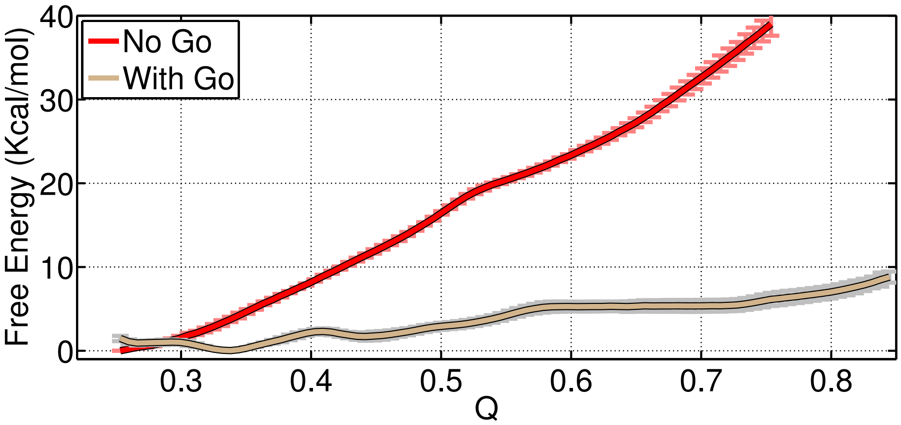

Figure S13: Comparison of the free energy profiles calculated using a potential energy function that includes the Gō term (yellow) and the one that does not (red). The yellow curve is identical to the one presented in Figure 1 of the main text. See text for details.

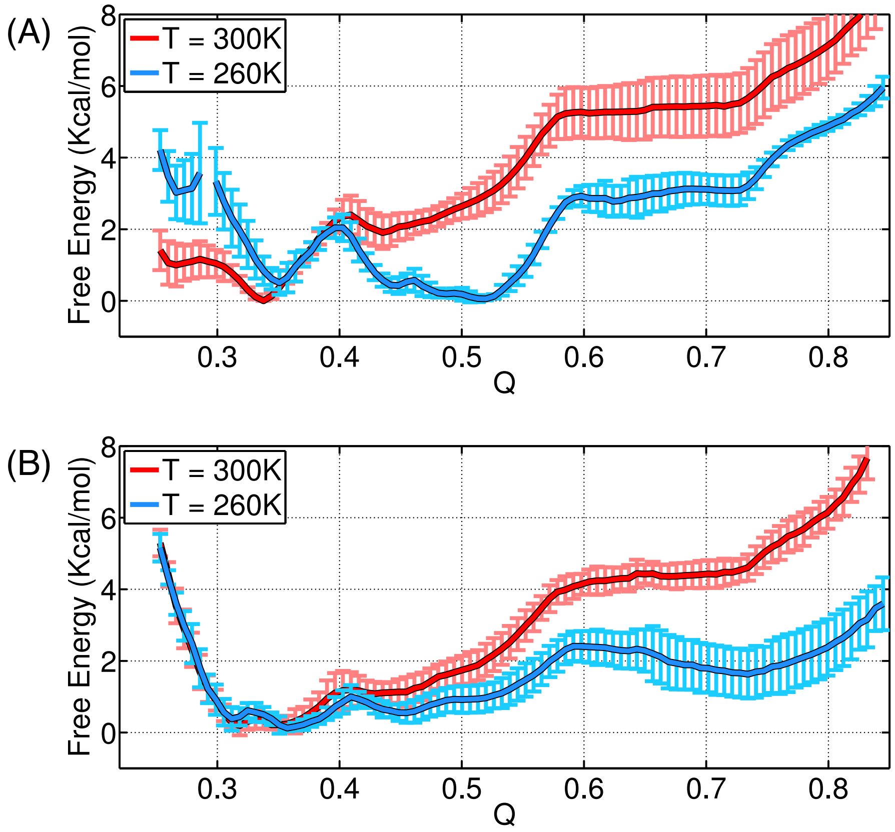

Figure S14: Free energy profiles at $T = 300$K (red) and $T = 260$K (blue) as a function of the fraction of native contacts Q calculated from simulations performed at a salt concentration of 100 mM (A) and 2 M (B). See text for details.

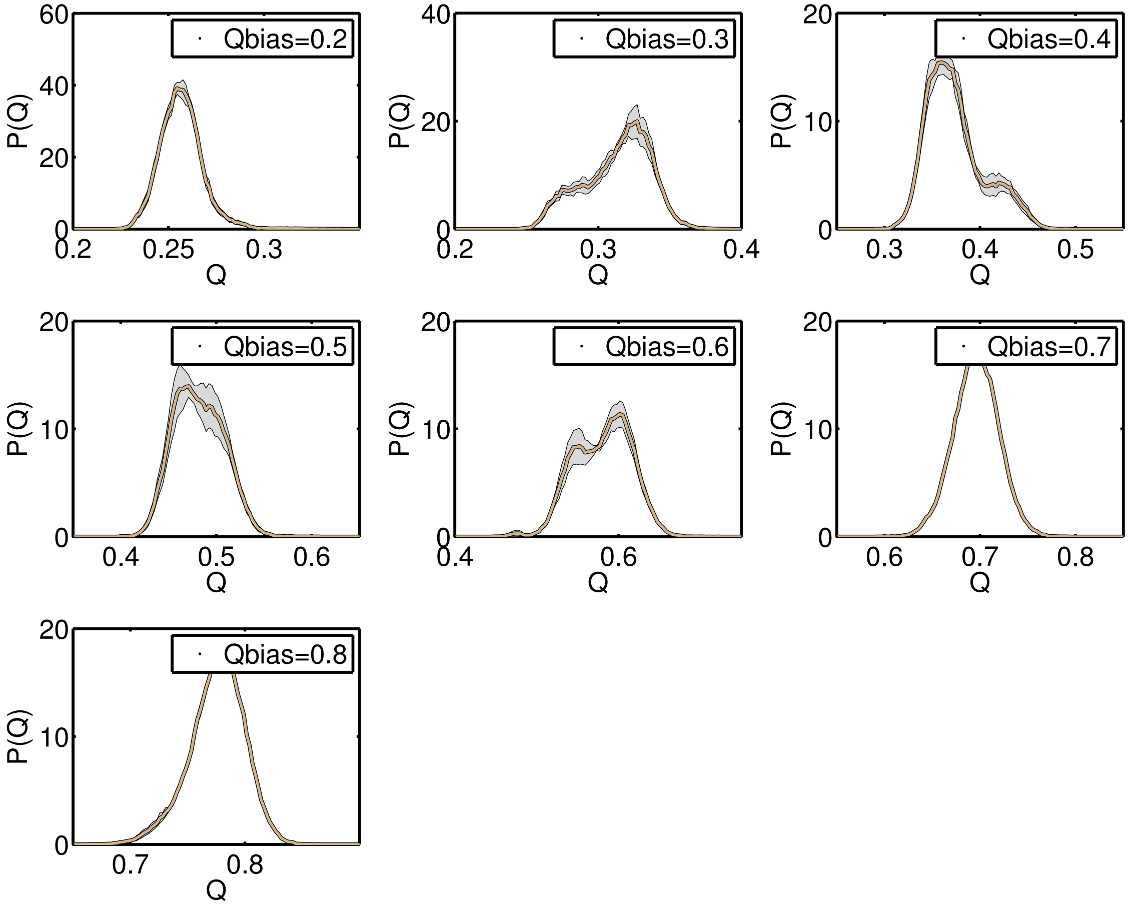

Figure S15: Probability distributions of the fraction of native contacts Q at different simulation windows in which the global Q is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the histone core complex shown in Figure 1 of the main text. See text for details.

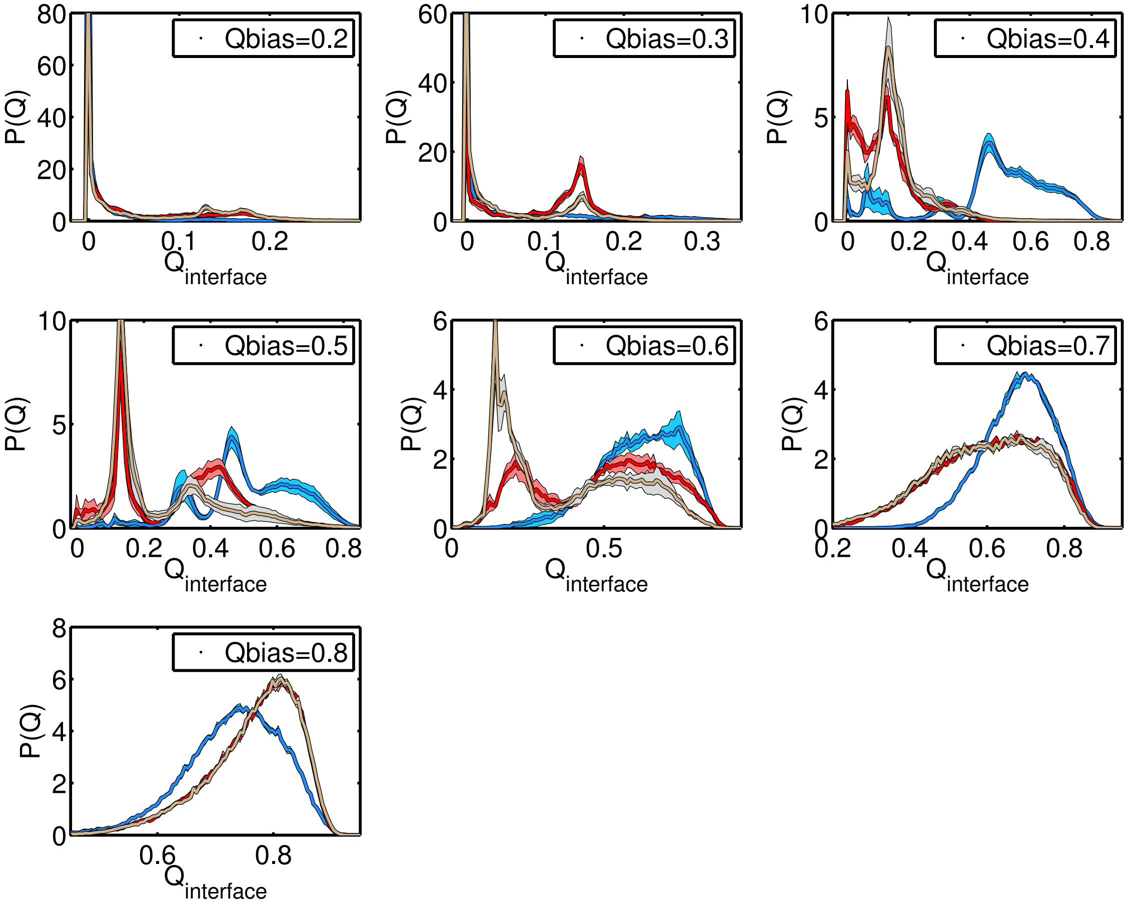

Figure S16: Probability distributions of various protein-protein interfacial contacts at different simulation windows in which the global Q is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the histone core complex shown in Figure 1 of the main text. The coloring scheme is the same as in Figure 2(A) of the main text, with the interface (H3-H4)$_\alpha$:(H3-H4)$_\alpha$ in blue, (H3-H4)$_\alpha$:(H2A-H2B)$_\alpha$ in red, and (H3-H4)$_\beta$:(H2A-H2B)$_\beta$ in yellow. See text for details.

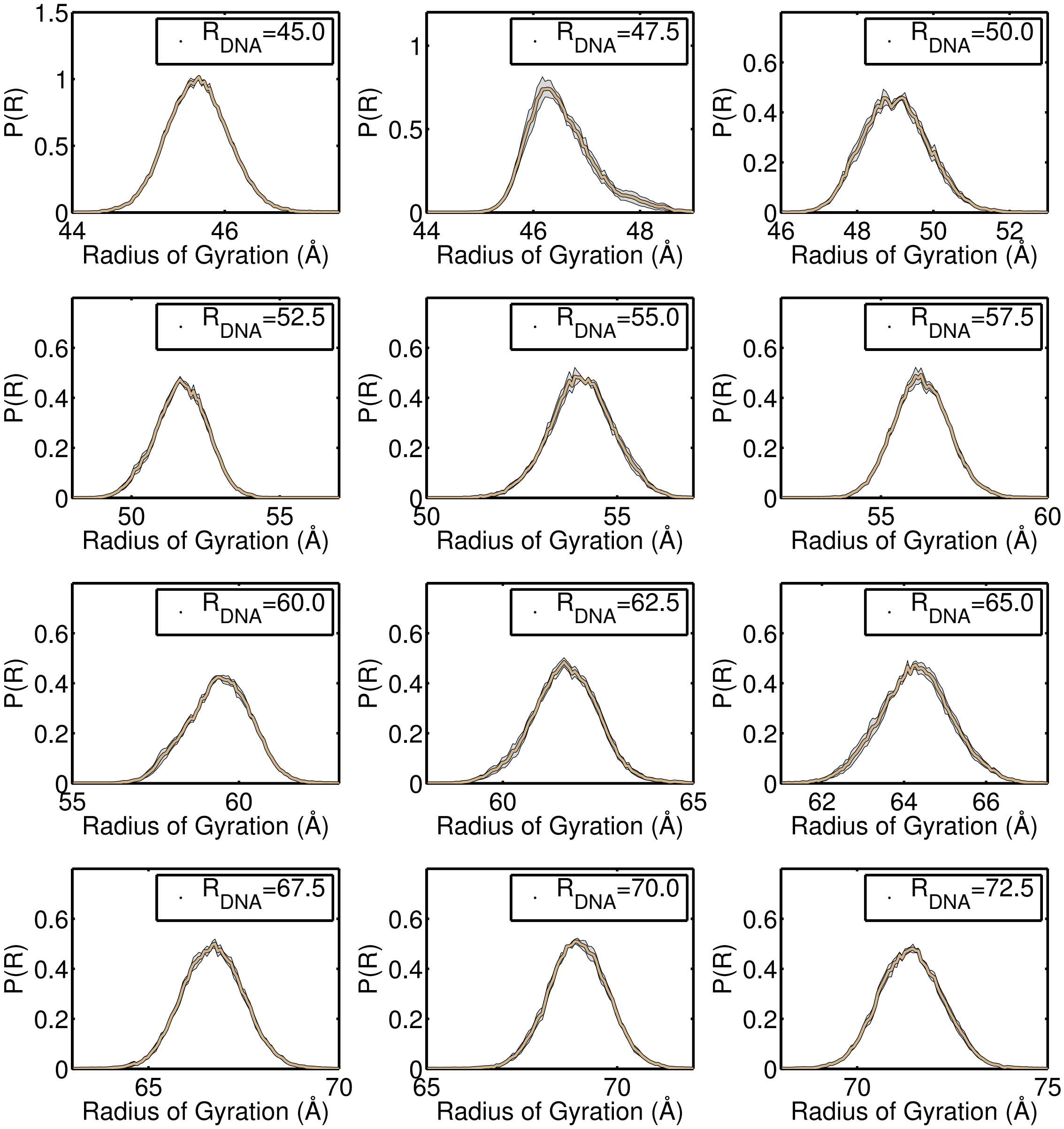

Figure S17: Probability distributions of the radius of gyration of the DNA, R_{DNA}, at different simulation windows in which the R_{DNA} is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the intact nucleosome shown in Figure 3 of the main text. See text for details.

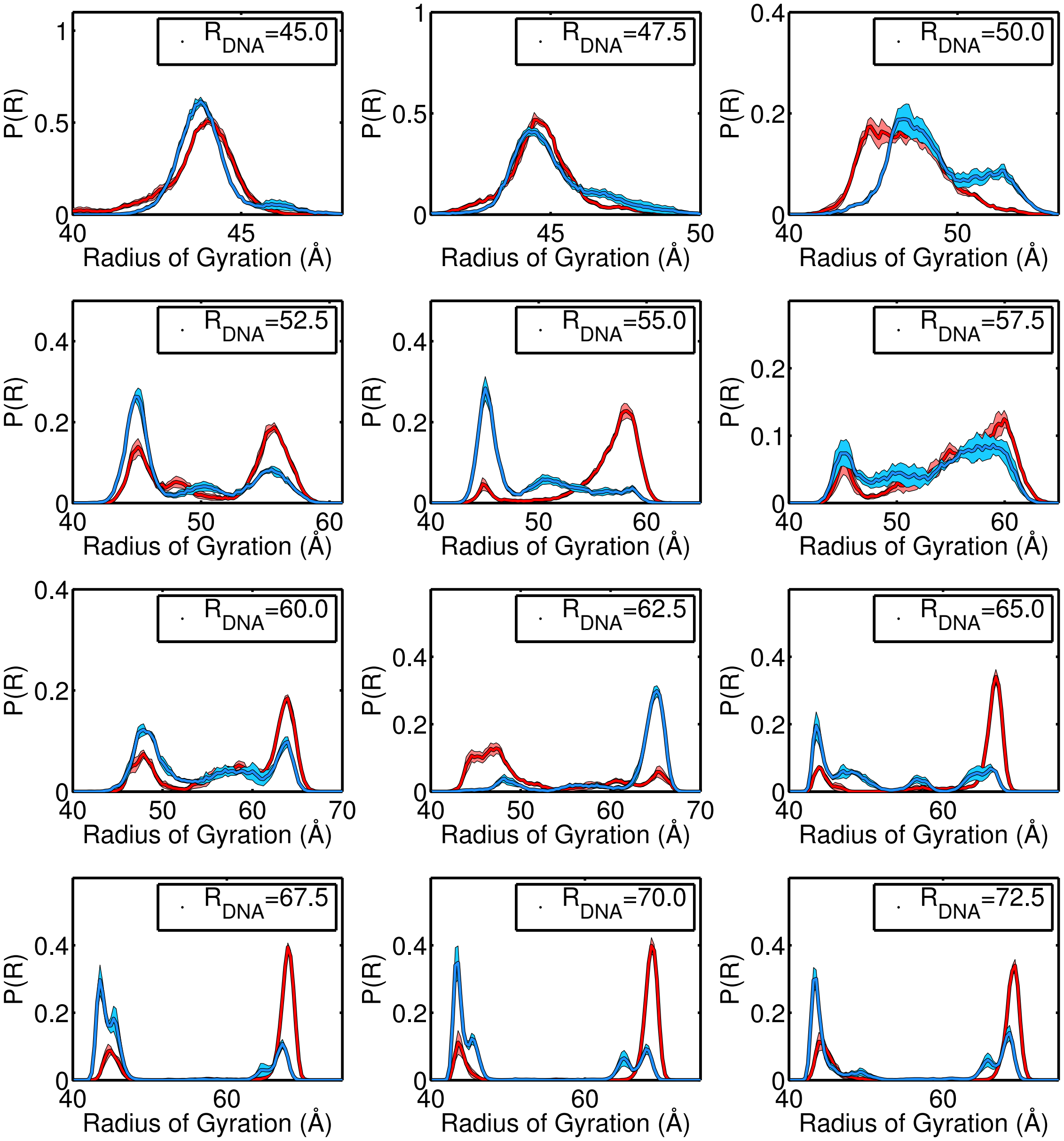

Figure S18: Probability distributions of the radius of gyrations of the two DNA segments separated at the nucleosome dyad at different simulation windows in which the R_{DNA} is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the intact nucleosome shown in Figure 3 of the main text. See text for details.

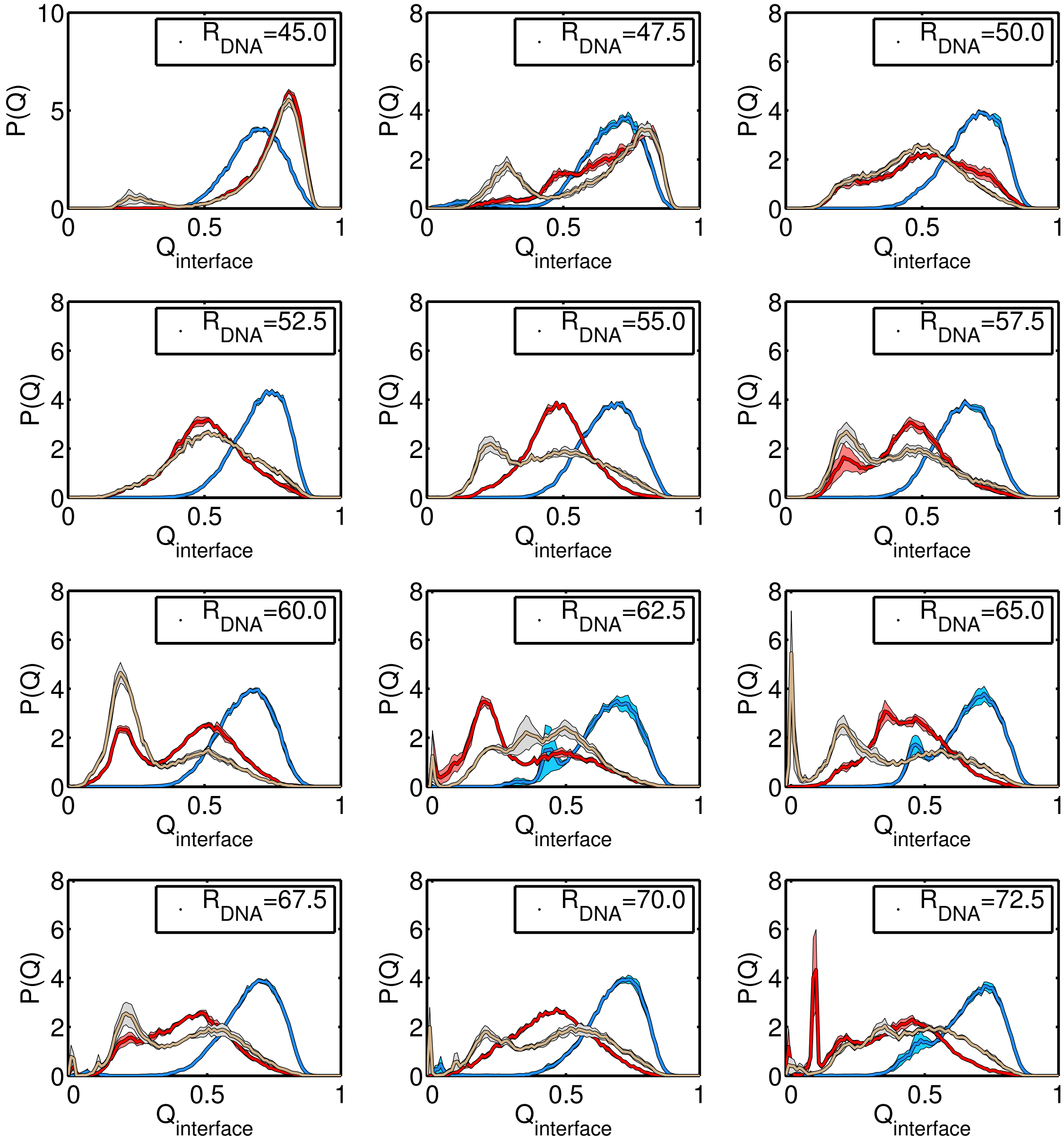

Figure S19: Probability distributions of the various protein-protein interfacial contacts at different simulation windows in which the R_{DNA} is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the intact nucleosome shown in Figure 3 of the main text. The coloring scheme is the same as in Figure 4(B) of the main text, with the interface (H3-H4)$_\alpha$:(H3-H4)$_\alpha$ in blue, (H3-H4)$_\alpha$:(H2A-H2B)$_\alpha$ in red, and (H3-H4)$_\beta$:(H2A-H2B)$_\beta$ in yellow. See text for details.

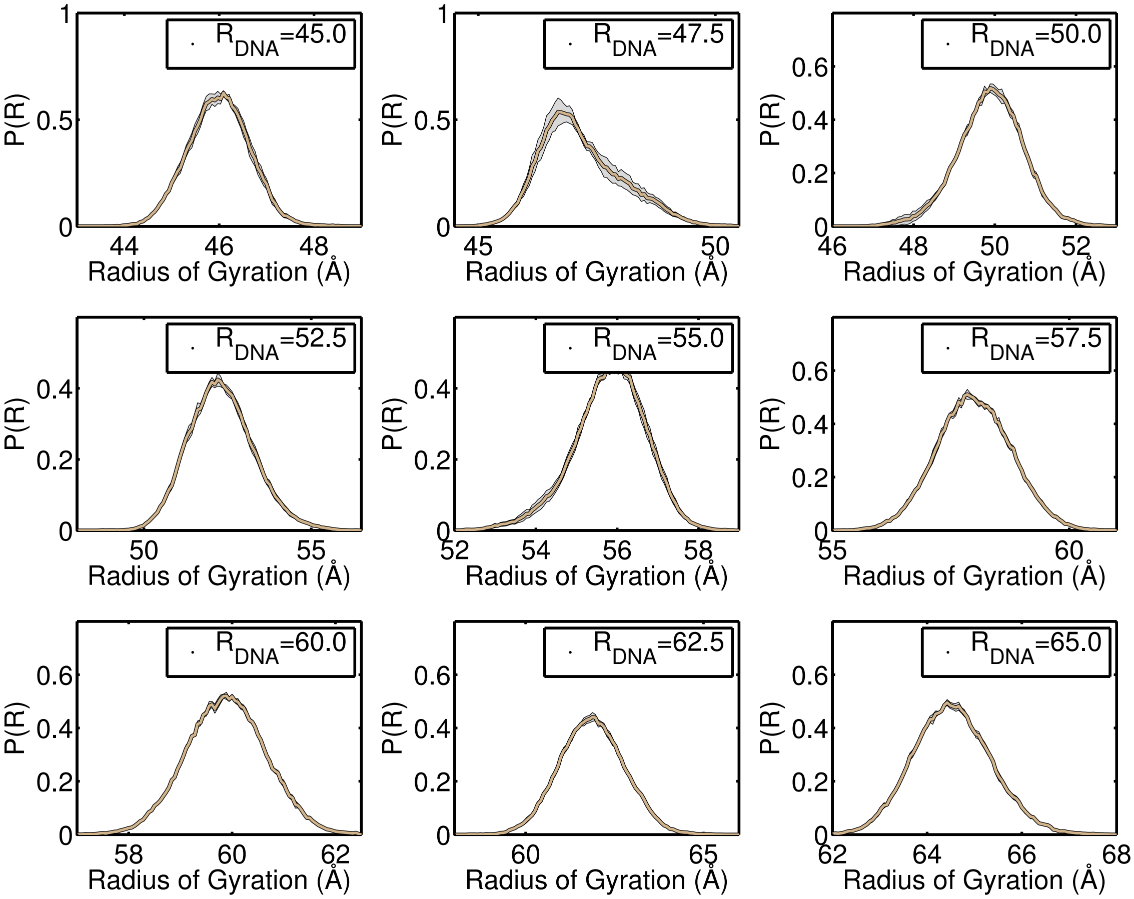

Figure S20: Probability distributions of the radius of gyration of the DNA, R_{DNA}, at different simulation windows in which the R_{DNA} is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the tailless nucleosome shown in Figure 6(A) of the main text. See text for details.

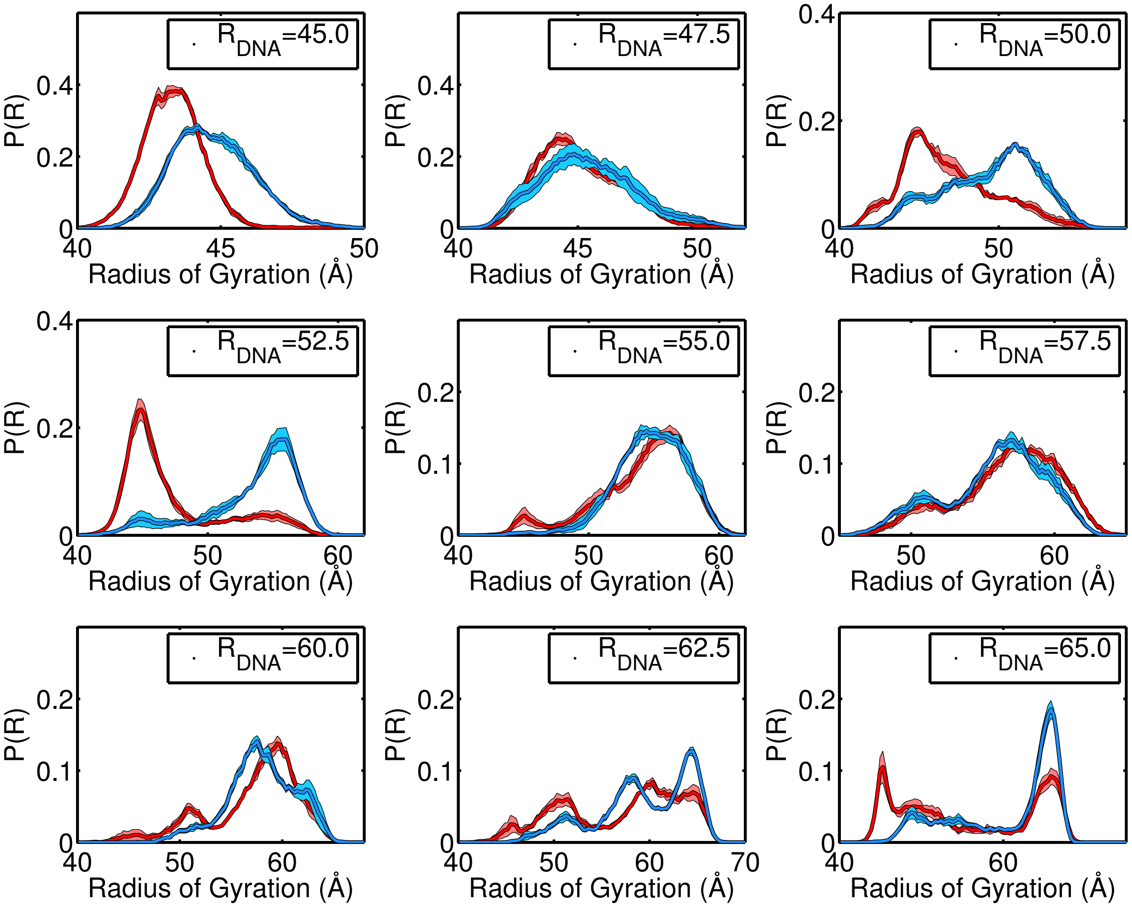

Figure S21: Probability distributions of the radius of gyrations of the two DNA segments separated at the nucleosome dyad at different simulation windows in which the R_{DNA} is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the tailless nucleosome shown in Figure 6(A) of the main text. See text for details.

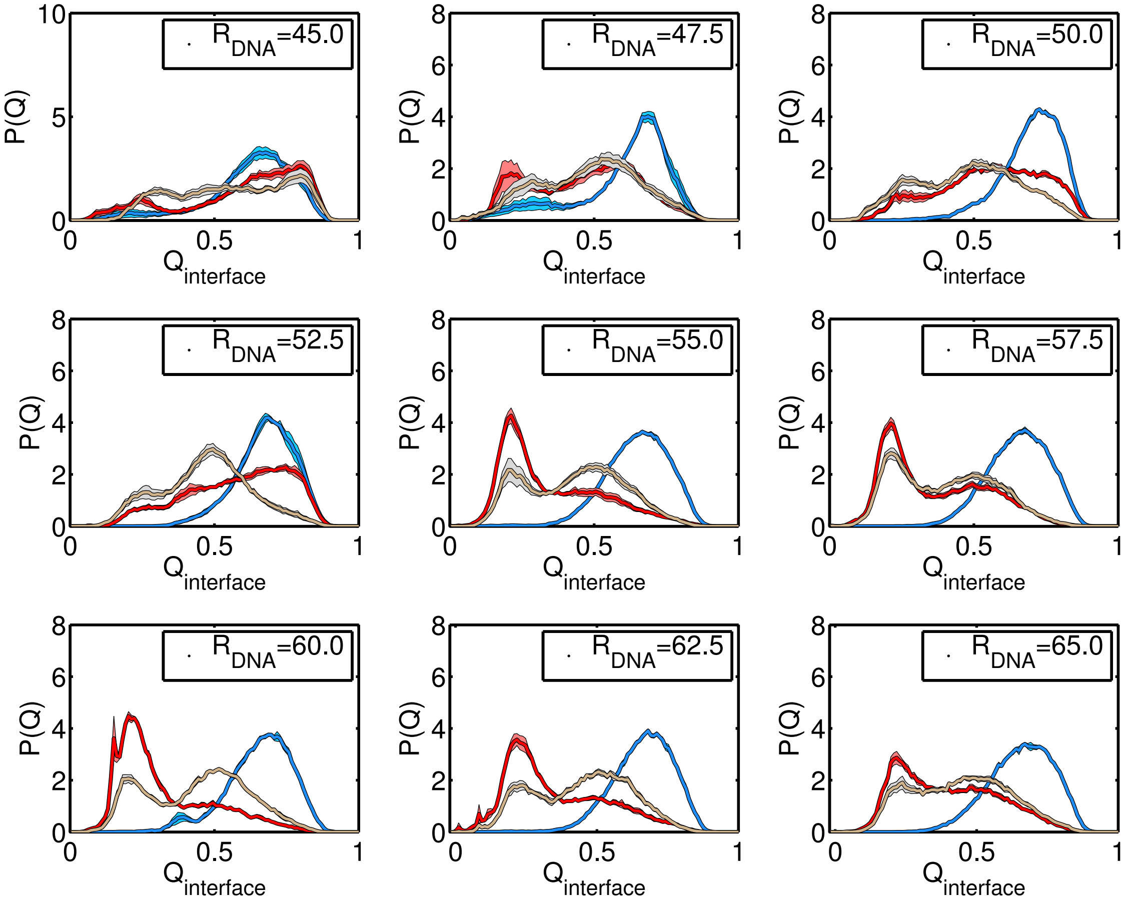

Figure S22: Probability distributions of the various protein-protein interfacial contacts at different simulation windows in which the R_{DNA} is biased to the target value shown in the legends. These plots were generated from the same trajectories used to determine the free energy profile for the tailless nucleosome shown in Figure 6(A) of the main text. The coloring scheme is the same as in Figure 6(B) of the main text, with the interface $(\text{H3-H4})_\alpha$:$(\text{H3-H4})_\alpha$ in blue, $(\text{H3-H4})_\alpha$:$(\text{H2A-H2B})_\alpha$ in red, and $(\text{H3-H4})_\beta$:$(\text{H2A-H2B})_\beta$ in yellow. See text for details.

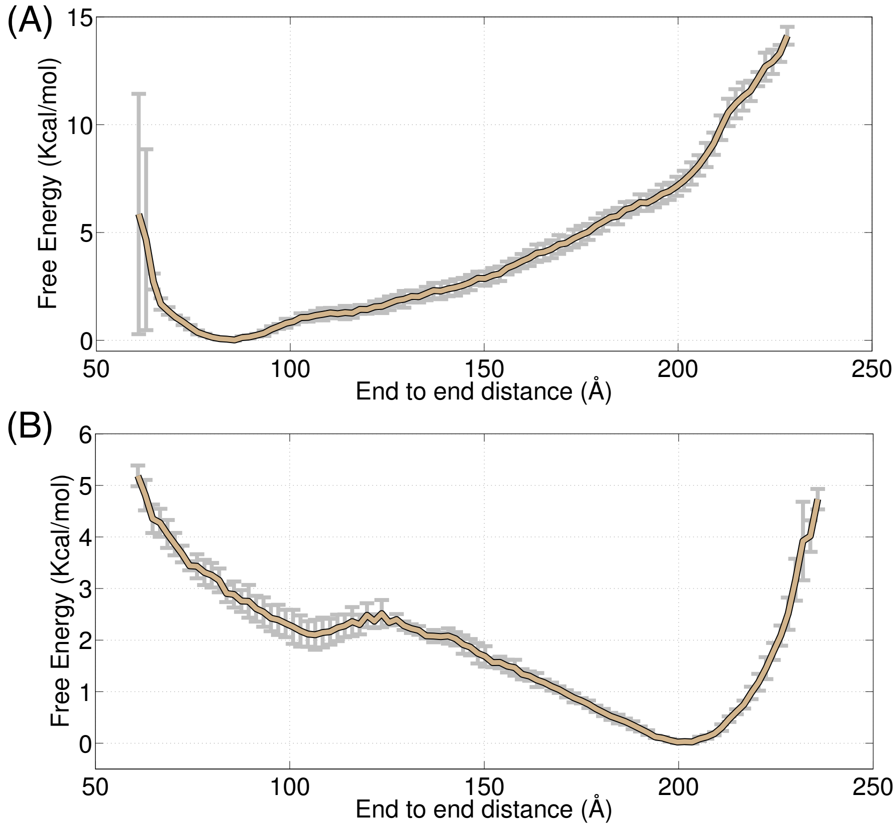

Figure S23: Free energy profiles as a function of the end-to-end distance of the DNA molecule for the intact (A) and the tailless (B) nucleosome. Part (A) and (B) were calculated from the same data used to construct Figure 3 and Figure 6(A) of the main text respectively. See SI *Section: Thermodynamics and Kinetics of DNA unwrapping* for details.