Ubiquitous nucleosome crowding in the yeast genome

Răzvan V. Chereji[a,1] and Alexandre V. Morozov[a,b,2]

[a]Department of Physics and Astronomy and [b]BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854

Edited by Timothy J. Richmond, Swiss Federal Institute of Technology, Zurich, Switzerland, and approved February 25, 2014 (received for review November 7, 2013)

Nucleosomes may undergo a conformational change in which a stretch of DNA peels off the histone octamer surface as a result of thermal fluctuations or interactions with chromatin remodelers. Thus, neighboring nucleosomes may invade each other's territories by DNA unwrapping and translocation, or through initial assembly in partially wrapped states. A recent high-resolution map of distances between dyads of neighboring nucleosomes in *Saccharomyces cerevisiae* reveals that nucleosomes frequently overlap DNA territories of their neighbors. This conclusion is supported by lower-resolution maps of *S. cerevisiae* nucleosome lengths based on micrococcal nuclease digestion and paired-end sequencing. The average length of wrapped DNA follows a stereotypical pattern in genes and promoters, correlated with the well-known distribution of nucleosome occupancy: nucleosomal DNA tends to be shorter in promoters and longer in coding regions. To explain these observations, we have developed a biophysical model that uses a 10–11-bp periodic histone–DNA binding energy profile. The profile is based on the pattern of histone–DNA contacts in nucleosome crystal structures, as well as the idea of linker length discretization caused by higher-order chromatin structure. Our model is in agreement with the observed genome-wide distributions of interdyad distances, wrapped DNA lengths, and nucleosome occupancies. Furthermore, our approach explains in vitro measurements of the accessibility of nucleosome-covered target sites and nucleosome-induced cooperativity between DNA-binding factors. We rule out several alternative scenarios of histone–DNA interactions as inconsistent with the genomic data.

partially unwrapped nucleosomes | DNA accessibility | gene regulation

Eukaryotic genomes are organized into arrays of nucleosomes (1). Each nucleosome consists of a stretch of genomic DNA wrapped around a histone octamer core (2). The resulting complex of DNA with histones and other regulatory and structural proteins is called chromatin (1). Arrays of nucleosomes form 10-nm fibers that resemble beads on a string and, in turn, fold into higher-order structures (3). Depending on the organism and cell type, 75–90% of genomic DNA is packaged into nucleosomes (1). Because nucleosomal DNA is wrapped tightly around the histone octamer, its accessibility to various DNA-binding proteins, such as repair enzymes, transcription factors (TFs), polymerases, and recombinases, is suppressed. The question of how cellular functions are carried out on the chromatin template is one of the outstanding puzzles in eukaryotic biology.

Recently, nucleosome dyad positions and distances between dyads of neighboring nucleosomes were mapped genome-wide with high precision in *Saccharomyces cerevisiae* (4). The in vivo map was obtained by chemical modification of engineered histones, DNA backbone cleavage by hydroxyl radicals, and high-throughput sequencing (Fig. 1*A*). Although more precise than methods based on micrococcal nuclease (MNase) digestion, whose accuracy is affected by MNase sequence preferences and the possibility of DNA over- or underdigestion (5, 6), the map is subject to unknown hydroxyl radical cutting preferences for two alternate sites on each DNA strand, at −1 bp and +6 bp with respect to the dyad (4, 7). Thus, distances between neighboring dyads are only approximately equal to measured distances between hydroxyl cut sites.

Surprisingly, 38.7% of the distances between hydroxyl cleavage sites marking neighboring nucleosomes are less than 147 bp, indicating that nucleosomes frequently invade each other's territories (8). This is possible only if the DNA of at least one nucleosome in the pair is partially unwrapped (Fig. 1*A*). Furthermore, there are distinct 10–11-bp periodic oscillations in the histogram of DNA fragment lengths, consistent with the pattern of histone–DNA contacts in nucleosome crystal structures (2, 9, 10). The observed nucleosome crowding may be the result of fully wrapped nucleosomes being transiently unwrapped and translocated, or disassembled and reassembled, by thermal fluctuations and chromatin remodeling enzymes. Alternatively, chromatin initially may have been assembled with many nucleosomes in partially wrapped states.

We present a statistical mechanics framework that is in agreement with the observed crowding of genomic nucleosomes (4), as well as earlier experiments that probed differential accessibility of nucleosome-covered binding sites (11–14), and studied nucleosome-induced cooperativity between DNA-binding factors (15–17). Our model attributes short interdyad distances seen in the experiment to intrinsic energetics of histone–DNA interactions. It significantly extends previous work (18–21) by considering sequence-dependent formation of partially wrapped nucleosome arrays and by proposing a histone–DNA binding energy profile based on nucleosome crystal structures. Using our approach, we reproduce nucleosome occupancies and average lengths of wrapped DNA in the vicinity of transcription start sites (TSSs). We also predict sequence-specific free energies of nucleosome formation in the presence of nucleosome crowding, using paired-end high-throughput nucleosome maps based on MNase digestion as input.

Results

Histone–DNA Binding Energy. We model energetics of histone–DNA interactions by representing the total free energy $u_{S(N)}$ of a nucleosome with the DNA sequence $S(N)$ of length N as a sum of a sequence-independent term u_N^{SI} and a sequence-dependent

Significance

In eukaryotic cells, up to 90% of genomic DNA is occluded by nucleosomes, fundamental units of chromatin in which DNA is tightly bent around the surface of a histone octamer. We present evidence that many nucleosomes in a single-cell eukaryote, *Saccharomyces cerevisiae*, are wrapped only partially, allowing them to invade genomic territories of their neighbors. These findings significantly extend the canonical view of nucleosomes as separate "beads on a string" and raise the question of how cellular functions are carried out in this crowded environment.

Author contributions: R.V.C. and A.V.M. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]Present address: Program in Genomics of Differentiation, Eunice Kennedy Shriver National Institute for Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892.

[2]To whom correspondence should be addressed. E-mail: morozov@physics.rutgers.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1321001111/-/DCSupplemental.

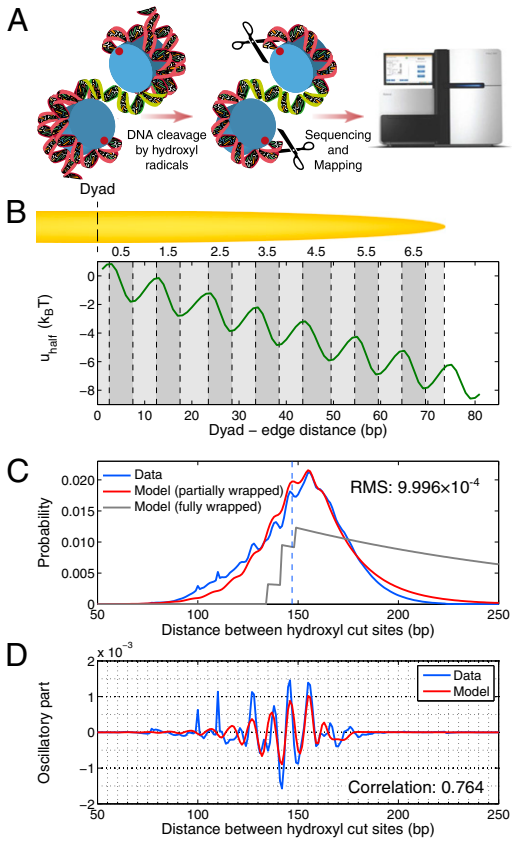

Fig. 1. Genome-wide distribution of distances between hydroxyl cut sites marking neighboring nucleosomes. (*A*) Overview of the chemical method for mapping nucleosome dyad positions and interdyad distances (4, 23). Mutant H4 histones (S47C) were modified by covalent attachment of a sulfhydryl-reactive, copper-chelating label to the cysteines. With the addition of copper and hydrogen peroxide, a localized cloud of hydroxyl radicals was produced that specifically cleaved the DNA backbone at sites symmetrically flanking nucleosome dyads. The cleavage products that correspond to DNA fragments linking neighboring nucleosomes were size-selected on an agarose gel, purified, sequenced using paired-end reads, and mapped to the *S. cerevisiae* genome. Note that the size-selection step likely causes depletion of very short and very long fragments from the sample. Each mapped pair of reads yields a measurement (biased by hydroxyl cleavage preferences) of the distance between dyads of neighboring nucleosomes positioned on the same chromosome. In the dinucleosome conformation shown, the interdyad distance is 100 bp (dyads are marked by red dots). Starting from the top dyad, 40 bp of DNA are wrapped around the histone octamer, followed by a 30-bp linker (green) and by 30 bp of DNA wrapped around the other octamer. (*B*) Histone–DNA binding/higher-order structure energy profile. The energy of a DNA segment $N = 2x + 1$ bp in length positioned symmetrically with respect to the dyad is given by $u_N^{SI} = 2u_{half}(x)$. The minima and maxima of the energy landscape are based on a crystal structure of the nucleosome core particle (9, 10). Dark gray bars show where the histone binding motifs interact with the DNA minor groove in the structure. Light gray bars show where the DNA major groove faces the histones. The energy profile was obtained by a polynomial fit (*SI Appendix, Model A*). (*C*) Normalized histogram of DNA fragment lengths from the high-resolution chemical map described in *A* (4) (blue) and from the models with partially (red) and fully (gray) wrapped DNA. In the latter, $a_{min} = a_{max} = 147$ bp. RMS, total root-mean-square deviation between the model and the data. (*D*) Oscillations in the observed (blue) and predicted (red) distributions of DNA fragment lengths, obtained by subtracting a smooth background from the data and the model with partially wrapped DNA in *C*. The smooth background was found by applying a Savitzky–Golay filter of polynomial order 3 with 31-bp length. Correlation refers to r_{osc}, the linear correlation coefficient between measured and predicted oscillations.

correction $u_{S(N)}^{SD}$ (*SI Appendix*). The sequence-independent term describes favorable interactions between histone side chains and the DNA phosphate backbone, as well as a free energy cost of bending DNA into the nucleosomal superhelix averaged over all sequences of length N in the genome. The sequence-dependent term represents deviation from this average due to the effects of a particular sequence $S(N)$ (22).

We introduce a simple model for u_N^{SI}, based on the 10–11-bp periodic pattern of histone contacts with the minor groove of the nucleosomal DNA (9, 10) (Fig. 1*B* and *SI Appendix, Model A*). As DNA is peeled off each contact patch, its free energy increases because hydrogen bonds and favorable electrostatic contacts between histone side chains and the DNA phosphate backbone are lost. However, once DNA breaks free from the contact patch, it may adopt multiple conformations, which allows it to increase its entropy and thus lower its total free energy. The favorable entropic term grows until the next contact patch is reached, completing one cycle in the oscillatory energy profile. The oscillations are superimposed on a straight line whose slope equals the average free energy lost when a DNA base pair is detached from the histone octamer surface.

We predict the distribution of interdyad distances using the sequence-independent model described above. We compute the conditional probability $P(c+d|c)$ of finding a nucleosome dyad at base pair $c+d$, given that the previous dyad is at base pair c (*SI Appendix*). Because interdyad distances cannot be used to distinguish whether nucleosomal DNA is wrapped symmetrically or asymmetrically with respect to the dyad, we assume the former for simplicity. The model is fit to the observed distribution of distances between hydroxyl cut sites (*SI Appendix*). The free parameters of the model include the amplitude of the oscillations, the slope of the free energy profile, and $a_{min(max)}$, the minimum (maximum) effective length of the nucleosome particle (*SI Appendix, Model A*). The minimum length of wrapped DNA is controlled by a_{min}, whereas a_{max} is allowed to exceed 147 bp to account for the effects of higher-order chromatin structure and linker histone deposition. We also fit the chemical potential of histone octamers and the relative frequency f of DNA cleavage by hydroxyl radicals at −1 bp with respect to the dyad (*SI Appendix*). We obtain $f = 0.5$, which implies that after averaging over hydroxyl cleavage preferences, interdyad distances are predicted to be 5 bp longer than the distances between hydroxyl cleavage sites marking neighboring nucleosomes (*SI Appendix*). With this correction, 25.8% of all interdyad distances are less than 147 bp. Our model reproduces both the overall shape and the fine oscillatory structure of the observed histogram of DNA fragment lengths (Fig. 1 *C* and *D*). In contrast, a model in which nucleosomes are always fully wrapped cannot fit the data (gray line in Fig. 1*C*).

Higher-Order Chromatin Structure and Linker Histone Energetics. The effective length of the particle found in the fit, $a_{max} = 163$ bp, is greater than 147 bp, the length of the DNA in the nucleosome core (10). Indeed, $a_{max} = 147$ bp is incompatible with the observed histogram of DNA fragment lengths (Fig. S3 *A* and *B*). The model is less sensitive to the value of a_{min}, because such short lengths of wrapped DNA are energetically unfavorable and therefore are not seen frequently in the data. The overall shape of the distribution is strongly affected by the slope of the energy profile in Fig. 1*B* (Fig. S3*C*). The fitted value of the slope yields 14.4 k_BT (where k_B is the Boltzmann constant, and T is the room temperature) for the average energy of a fully wrapped nucleosome.

Because a_{max} is greater than 147 bp, the energy profile in Fig. 1*B* describes both DNA attachment to the histone octamer (up to 73 bp from the dyad) and the effects of higher-order chromatin structure, including binding of Hho1p, the H1 linker histone of *S. cerevisiae* (24–26), or histone tails (27) to the DNA immediately outside the nucleosome core. Although Hho1p is less abundant in yeast than in higher eukaryotes, it is involved in higher-order chromatin organization, including chromatin compaction in stationary phase (26, 28). Relatively little is known

about the molecular mechanism of H1 binding, including the length and symmetry of its DNA footprint (24, 25). H1 binding and other factors that mediate chromatin folding into higher-order structures cause linker lengths to be discretized (1, 29). Linker length discretization may be described by a periodic decaying two-body effective potential between neighboring nucleosomes, with the first minimum ∼5 bp away from the nucleosome edge (1, 30, 31).

Based on these observations, we have constructed two models for the energy profile outside the nucleosome core region. The first model is a polynomial fit that extends the quasiperiodic profile of the histone–DNA binding energy through another cycle (Fig. 1*B* and *SI Appendix*, *Model A*). The depth and position of the first minimum outside the nucleosome core are additional free parameters. As may be seen in Fig. S4, our fit robustly predicts the first minimum to be positioned 5–6 bp outside the nucleosome core, in agreement with previous studies (1, 30, 31). The depth of this minimum is comparable to the depth of the other minima (Fig. S4 and *SI Appendix*, *Model A*).

The second model represents the energy profile outside the nucleosome core by a linear function (Fig. S5*A* and *SI Appendix*, *Model B*). The two free parameters are the slope and the range of the linear function, which are related to the H1–DNA interaction energy and the H1 footprint, respectively. This alternative scenario, although likely oversimplified, may be used to check the sensitivity of our results toward a particular energy profile outside the core region. We find that the linear profile fits the overall shape of the distribution of DNA fragment lengths less well than the oscillatory one (compare rms values in Fig. 1*C* and Fig. S5*B*), although the 10–11-bp periodic fine structure is reproduced in both cases (Fig. 1*D* and Fig. S5*C*). The optimal linear profile is 7 bp long, yielding a symmetric H1 footprint with two 7-bp half-sites (Fig. S5*D*) and the H1–DNA interaction energy of ≈ 5 k_BT (Fig. S5*E*).

Alternative Models of Histone–DNA Interactions. Next, we tested the sensitivity of our fits to the functional form of the histone–DNA binding energy. Although our primary model follows nucleosome crystal structures in creating a quasiperiodic energy profile with both 10- and 11-bp modes, strictly periodic 10- or 11-bp sinusoidal profiles yield nearly the same quality of the fit (Fig. S6 and *SI Appendix*, *Models C and D*). Because the initial phase of the oscillations is not determined by the crystal structure, it becomes another fitting parameter. The fitted initial phases in the 10- and 11-bp models make the periodic curves match the crystal structure further away from the dyad (Fig. S6*A*). The phases diverge closer to the dyad, where they are not as strongly constrained by the data. Rms deviation is less sensitive to the initial phase than r_{osc}, the linear correlation between predicted and observed oscillations in the histograms of DNA fragment lengths (Fig. S7).

Because the distribution of distances between hydroxyl cut sites has a prominent oscillatory component, it is not surprising that a linear model, in which histone–DNA binding energy per base pair is constant, does not fit the data equally well, although it does match its overall shape (Fig. S8*A* and *SI Appendix*, *Model E*). Less trivially, it was suggested on the basis of single-nucleosome unzipping experiments that nucleosome unwrapping proceeds with 5-bp periodicity because histones interact with each DNA strand separately where the DNA minor groove faces the histone octamer surface, creating two distinct contact "subpatches" (32). These single-molecule data were fit to a model with a stepwise free energy profile (33). Each step in the profile corresponds to breaking a point histone–DNA contact, and the steps occur every 5.25 bp on average. We do not find any evidence of 5-bp periodicity of histone–DNA interactions in the genomic data. Indeed, both 5-bp stepwise and 5-bp periodic sinusoidal profiles fit the data poorly, about as well as the linear model (Fig. S8 *B* and *C* and *SI Appendix*, *Models F and G*). Even the 10-bp stepwise energy profile, although clearly having the right periodicity, does not fit the data as well as the structure-based model (Fig. S8*D* and *SI Appendix*, *Model H*). These observations suggest that the picture of a gradual loss of favorable finite-range histone–DNA interactions followed by gain in DNA conformational entropy is closer to reality than abrupt disruption of short-range histone–DNA contacts. A direct comparison of single-molecule and genome-wide energy profiles unfortunately is obscured by the fact that the reported single-nucleosome unzipping experiments are specific to the 601-nucleosome–forming sequence (34), in contrast to our methodology, which provides the average, sequence-independent picture of histone–DNA energetics.

Extensive Crowding of Genomic Nucleosomes. Fig. 2*A*, in which genes are sorted by the intergenic length and aligned by the TSS, shows a canonical picture of nucleosomes depleted in promoters and well-positioned over coding regions (38). Interestingly, interdyad distances tend to be shorter in promoters (Fig. 2*B*). When averaged over all genes, the number of dyads at a given base pair and the average interdyad distance at that base pair are strongly correlated with each other (compare blue and red lines in Fig. 2*C*) and with the distribution of wrapped DNA lengths in an MNase-based assay, which mapped both nucleosomes and subnucleosome-size particles by paired-end sequencing (Fig. S9*A* and green line in Fig. 2*C*) (35). The observed behavior is the opposite of the initial expectation that nucleosome crowding should increase with occupancy but can be reproduced in a simple sequence-independent model in which nucleosomes phase off a potential barrier placed in the promoter region (Fig. S10) (39).

Crowded nucleosomes tend to have elevated histone turnover rates (37), both around TSSs and genome-wide (Fig. 2*C* and Fig. S9 *B* and *D*). Nucleosomes at loci enriched in transcription preinitiation complexes (PICs) (36) also are more crowded (Fig. 2*C* and Fig. S9*C*). Finally, interdyad distances tend to increase with the fraction of adenine/thymine (A/T) nucleotides (Fig. S9*E*). We note that it is misleading to equate internucleosome distances with peak-to-peak distances in the average profile of nucleosome dyad counts (blue line in Fig. 2*C*). The peak-to-peak distances are 164–165 bp, whereas the average distance between hydroxyl cut sites for the nucleosomes in the [TSS, TSS + 1,000] region is 149.6 bp (i.e., the average interdyad distance is 154.6 bp). Nucleosome crowding is prominent in many types of genomic functional regions (Fig. S11). Thus, nucleosome crowding and partial overlap of nucleosomal territories are more common than could be predicted by mapping single-nucleosome positions alone.

Accessibility of Nucleosomal DNA to Factor Binding. Partial unwrapping of nucleosomal DNA results in differential accessibility of factor binding sites with respect to their position inside the nucleosome: sites on the edges are more accessible than those closer to the dyad. In contrast, all-or-none nucleosome formation should not be sensitive to the binding site position—a nucleosome, once unfolded, liberates its entire DNA sequence. Polach and Widom (11) studied the differential accessibility of six restriction enzymes to their target sites. The sites were placed at various locations throughout the 5S rRNA nucleosome positioning sequence (Fig. 3*A*). A later study used the 601 sequence and an extended set of 11 restriction enzymes (Fig. 3*B*) (12). These studies measured equilibrium constants for site exposure, K_{eq}^{conf}, which are related to the probability that a site will be accessible for binding: $p_{open} = K_{eq}^{conf}/(1 + K_{eq}^{conf}) \approx K_{eq}^{conf}$ (20).

We use our crystal structure-based model of histone–DNA energetics (Fig. 1*B* and *SI Appendix*, *Model A*) to fit the data on site accessibility (11, 12). Here, the system consists of a single nucleosome, and asymmetric unwrapping of its DNA tails is allowed (i.e., wrapped DNA does not have to be centered around the dyad). We assume that a site becomes accessible for the enzyme only after an additional number of base pairs, d, have been unwrapped from the histone octamer surface (20). We also

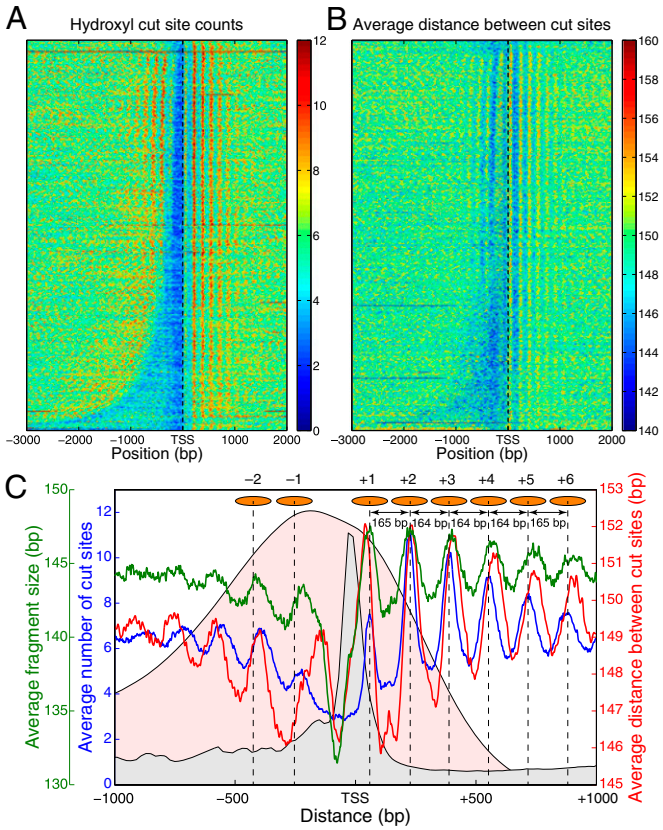

Fig. 2. Nucleosome crowding in the vicinity of transcription start sites. (*A*) Distribution of nucleosome dyad counts (4) near the TSS. Dyads are marked by the positions of hydroxyl cut sites. Four thousand seven hundred and sixty-three verified *S. cerevisiae* ORFs were aligned by their TSS and sorted by upstream intergenic lengths. Each horizontal line corresponds to one ORF. (*B*) Distribution of average distances between hydroxyl cut sites marking neighboring nucleosomes. ORFs are sorted as in *A*. In *A* and *B*, values at base pairs without mapped cut sites are obtained by interpolation, and heat maps are smoothed using a 2D Gaussian kernel with $\sigma = 3$ pixels. (*C*) Data in *A*, *B*, and Fig. S9 *A*–*C* averaged over all genes. Blue, nucleosome dyad counts; red, average distance between hydroxyl cut sites marking neighboring dyads; green, average length of DNA-bound particles mapped by MNase digestion (35) (see Fig. S9*A* for details). Curve with light gray background, combined occupancy of nine PICs (TATA-binding protein, TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, TFIIK, RNA polymerase II) (36); curve with light pink background, average histone turnover rate (37). The peaks in the dyad count profile (blue) are marked with orange ovals representing nucleosomes, and peak-to-peak distances are shown.

assume that once the dyad is unwrapped from either end, the entire nucleosome unfolds. Then the probability that a binding site is accessible is given by

$$p_{\text{open}}(x) = \begin{cases} 1 - \text{Occ}_{\text{nuc}}(x+d) & \text{for } x < x_{\text{d}} - d, \\ 1 - \text{Occ}_{\text{nuc}}(x_{\text{d}}) & \text{for } x_{\text{d}} - d \le x \le x_{\text{d}} + d, \\ 1 - \text{Occ}_{\text{nuc}}(x-d) & \text{for } x > x_{\text{d}} + d, \end{cases}$$

where $x \in [1, 147]$ bp, $x_{\text{d}} = 74$ bp is the position of the dyad, and the nucleosome occupancy is computed using Eq. **3**.

Besides d, the fitting parameters of the model are the overall slope of the binding energy, ε, and the histone chemical potential, μ (all other parameters are as in *SI Appendix*, *Model A*, except for $a_{max} = 147$ bp). For the 5S rRNA measurements (11), we obtain $\varepsilon^{5S} = -0.13$ k_BT/bp, $\mu = -17.5$ k_BT, and $d = 23$ bp. For the 601 measurements (12), $\varepsilon^{601} = -0.16$ k_BT/bp, $\mu = -16.4$ k_BT, and $d = 45$ bp. As expected, the nucleosome formation energy of the 601 sequence is $147 \times (\varepsilon^{5S} - \varepsilon^{601}) = 4.4$ k_BT more favorable than that of the 5S sequence, in agreement with the experimentally measured difference of 4.9 k_BT (40). The nucleosome formation energy of the 601 sequence is 24.1 k_BT, close to the 23.8 k_BT estimate made on the basis of 601 unzipping experiments (33). Interestingly, the 601 DNA has to unwrap more extensively past the binding site to allow access to restriction enzymes.

Overall, our model reproduces the observed differential accessibility of restriction enzyme binding sites with respect to the nucleosome dyad (Fig. 3). The only outliers are StyI and BfaI sites in the 601 series, which were not used in the fit. Because DNA unwrapping proceeds from nucleosome edges, these sites cannot be made more accessible than the PmlI site, which is located closer to the edge. It is possible that StyI and BfaI require less extensive unwrapping past their sites (smaller d) or have some affinity for nucleosome-wrapped DNA.

Nucleosome-Induced Cooperativity. If multiple binding sites reside within a single nucleosome, binding of one factor makes the other sites more accessible. This phenomenon is known as nucleosome-induced cooperativity (15, 16, 18). The cooperativity disappears in the absence of nucleosomes and reduces in extent with the distance between consecutive binding sites (15). Moreover, the cooperativity is not observed if the two sites are on the opposite sides of the nucleosome dyad (17).

We can use our model of histone–DNA energetics (*SI Appendix*, *Model A* with $a_{max} = 147$ bp) to capture all these aspects of nucleosome-induced cooperativity in a single-nucleosome model (Fig. 4). Specifically, for sites located more than 40 bp away from the dyad site, accessibility is strongly enhanced if DNA unwrapping is allowed (Fig. 4*A*). Interestingly, cooperativity between two TFs bound on the same side of the dyad is observed both with and without unwrapping (Fig. 4*B*). However, without unwrapping, it is impossible to show that binding on the opposite sides of the dyad is not cooperative, as observed in experiments (17) (Fig. 4*C*). Furthermore, the decrease of cooperativity with distance (15) cannot be reproduced (Fig. 4*D*). Thus, modeling partially wrapped nucleosomes is necessary for understanding how TFs and other DNA-binding proteins gain access to their nucleosome-covered sites.

Sequence Dependence of Nucleosome Energetics. We now focus on the sequence-dependent correction to the average free energy of nucleosome formation, $u^{\text{SD}}_{S(N)}$. We assume that $u^{\text{SD}}_{S(N)}$ depends only on the number of mono- and dinucleotides in the nucleosomal DNA (41, 42) (*SI Appendix*). We consider three in vivo nucleosome maps in *S. cerevisiae* based on paired-end sequencing (35, 43, 44) and one in vitro map in which nucleosomes were assembled on yeast genomic DNA and sequenced using single-end reads (45). In the latter case, we assume that all nucleosomes have a canonical length of 147 bp. All four maps used MNase digestion to isolate mononucleosomes. We compute the total free energy of nucleosome formation $u_{S(N)}$ using Eq. **4** and fit the sequence-dependent model to $u_{S(N)} - u^{\text{SI}}_N$, with the sequence-independent contribution u^{SI}_N obtained previously (Fig. 1*B* and *SI Appendix*, *Model A*). This procedure was adopted because resolution of MNase-based nucleosome maps is insufficient for predicting u^{SI}_N. The resulting energy parameters are listed in Table S9.

The number of partially wrapped nucleosome species may be as high as several thousand, depending on the minimum length of the DNA segment attached to the histones and the symmetry of its position with respect to the dyad. Because available read coverage levels (total number of reads divided by the genome length) are relatively low (Table S9), it is possible that robust predictions of $u^{\text{SD}}_{S(N)}$ and u^{SI}_N cannot be carried out using current datasets. To address this concern in the absence of high-resolution, high-coverage experimental data, we have tested our ability to predict nucleosome positioning and energetics by using a realistic model system (Fig. S12). We find that we can infer u^{SI}_N even at a coverage of 1 read per base pair (Fig. S12 *A* and *B*). Nucleosome occupancies and dyad positions are reproduced reasonably well starting with 10 reads per base pair, but at least

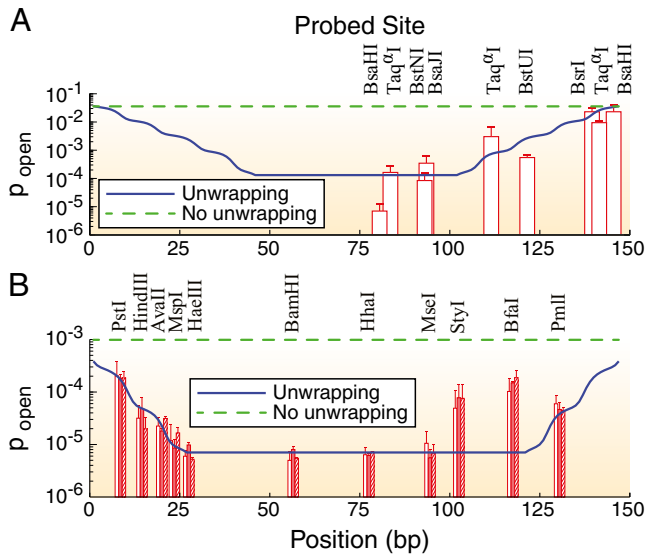

Fig. 3. Probability of binding site exposure within a nucleosome. The solid blue and dashed green lines represent model predictions with and without DNA unwrapping, respectively. In the latter case, $a_{min} = a_{max} = 147$ bp and all other parameters are as in the model with unwrapping. The dyad is fixed at base pair 74. (*A*) Restriction enzyme sites inserted into the 5S rRNA sequence at locations indicated by the centers of vertical red bars (11). (*B*) Restriction enzyme sites inserted into the 601 sequence at locations indicated by the centers of vertical red bars in the middle of each group (12). Each group of three bars corresponds to independent measurements in which the 601 sequence was flanked by different DNA sequences. In *A* and *B*, the height of each bar is the equilibrium constant for site exposure averaged over multiple experiments (error bars show standard deviations).

100 reads per base pair are required to recover the energy parameters (Fig. S12*C*).

Discussion and Conclusions

Recent MNase-based maps of nucleosome positions in yeast that use paired-end sequencing have identified numerous subnucleosome-size particles (35, 43, 44). However, shorter particles may not correspond to nucleosomes, and some DNA fragments may have been overdigested by MNase. These challenges were overcome in an experiment in which hydroxyl radicals were used to map nucleosome dyads in *S. cerevisiae* (4, 23). Used in conjunction with paired-end sequencing, this experiment provided information about both genomic dyad positions and the distances between dyads of neighboring nucleosomes. The histogram of distances between hydroxyl cut sites shows that DNA in many nucleosomes is wrapped only partially (Fig. 1*C*). Over regulatory and coding regions, the distribution of wrapped DNA lengths oscillates in phase with the nucleosome occupancy profile so that more wrapped nucleosomes also are more stable.

We have developed a statistical mechanics description of nucleosome arrays that allows neighboring nucleosomes to overlap each other's territories. We have shown that the prominent 10–11-bp periodicity in the distribution of interdyad distances (4) is consistent with a binding energy profile based on the pattern of histone–DNA contacts in nucleosome crystal structures. We could rule out several alternative scenarios, including stepwise and 5-bp periodic binding profiles. Furthermore, predicting the distribution of interdyad lengths required us to account for linker length discretization, commonly thought to be imposed by chromatin fiber formation (29). Our approach reproduces the stereotypical distribution of wrapped DNA lengths in the vicinity of TSSs, if potential barriers are placed at gene promoters. The barriers may be created in vivo by PICs (36) and other DNA-bound factors.

Our model yields estimates of nucleosome formation energies consistent with previous biophysical experiments. It also accounts for single-nucleosome observations, which showed that restriction enzymes and other factors can bind their nucleosome-occluded sites because of transient DNA unwrapping (11, 12, 15, 17). Moreover, nucleosome-covered binding sites closer to the edge of the nucleosome are more accessible to DNA-binding factors, and binding of the first factor enhances binding of subsequent factors on the same side of the dyad. The extent of nucleosome crowding in the yeast genome suggests that partially unwrapped nucleosomes should be considered in all future models of nucleosome positioning and chromatin energetics.

Materials and Methods

Here, we outline the exact theory for T molecular species simultaneously interacting with 1D DNA (details are provided in *SI Appendix*). The DNA-bound particles are subject to steric exclusion and also may be partially unwrapped. Let $u_t(k,l)$ $(t \in \{1,2,\ldots,T\})$ denote the binding energy of a particle of type t occupying base pairs k to l. The one-body distribution of such particles is given by

$$n_1^t(k,l) = \frac{1}{Z} Z^-(k) \langle t,k|z|t,l\rangle Z^+(l), \quad [1]$$

where Z is the grand canonical partition function and $\langle t,k|z|s,l\rangle = e^{\beta[\mu_s - u_s(k,l)]}\delta_{ts}$ ($\beta = 1/(k_B T)$ is the inverse temperature, μ_s is the chemical potential of particles of type s, and δ_{ts} is the Kronecker symbol). If the DNA length is L, vectors $|t,k\rangle$ are TL-dimensional with 1 at position $(t-1)L+k$ and

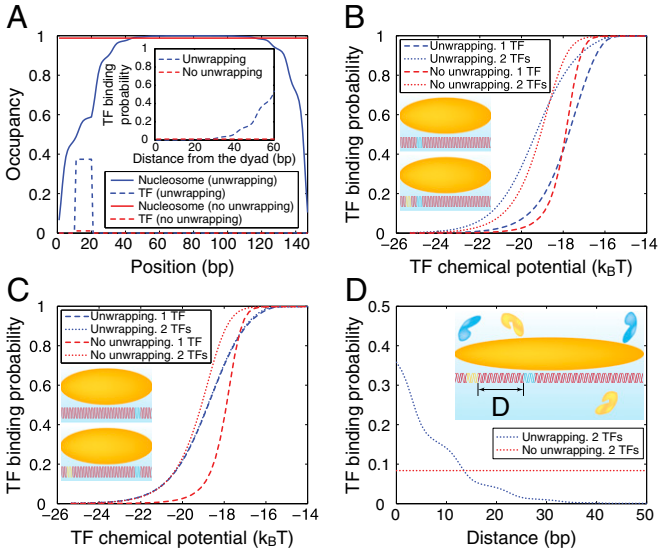

Fig. 4. Nucleosome-induced cooperativity between DNA-binding factors. (*A*) TF and nucleosome occupancy with and without unwrapping. The TF binding site occupies base pairs 11–20. (*Inset*) TF binding probability as a function of the distance between the nucleosome dyad and the proximal edge of the TF site, with and without unwrapping. (*B*) TF titration curves for one TF site vs. two TF sites located on the same side of the dyad. Site 1 occupies base pairs 11–20, site 2 occupies base pairs 31–40. (*Inset*) Binding site locations. (*C*) Same as *B*, but with the two TF binding sites located on the opposite sides of the dyad. Site 1 occupies base pairs 11–20, site 2 occupies base pairs 117–126. (*Inset*) Binding site locations. (*D*) Nucleosome-induced cooperativity as a function of the distance between two TF binding sites. The binding probability of the second TF is shown. Site 1 occupies base pairs 11–20, whereas the position of the second site is variable. (*Inset*) Definition of the distance between the two binding sites. In all panels, free energy of a fully wrapped nucleosome is $\ln(10^{-9})$ k_BT, histone chemical potential is $\ln(10^{-6})$ k_BT, and $\ln(10^{-6})$ k_BT to all other sites. TF chemical potential is $\ln(10^{-9})$ k_BT unless varied (19). Asymmetric unwrapping is allowed; in the model without unwrapping, $a_{min} = a_{max} = 147$ bp, and all other parameters are as in the model with unwrapping.

0 everywhere else. Similarly, the nearest-neighbor two-body distribution is given by

$$\overline{n}_2^{t,s}(i,j;k,l) = \frac{1}{Z} Z^-(i)\langle t,i|z|t,j\rangle \Theta(k-j)\langle s,k|z|s,l\rangle Z^+(l), \quad [2]$$

where Θ is the Heaviside step function. $Z^-(k)$ and $Z^+(k)$ are partial partition functions for the DNA segments [1,k] and (k,L], respectively. $Z^-(k)$, $Z^+(k)$, and Z can be computed recursively (*SI Appendix*). Using Eq. **1**, the particle occupancy is

$$\mathrm{Occ}_t(i) = \sum_{k=i-a^t_{\max}+1}^{i} \sum_{l=\max\left(i,k+a^t_{\min}-1\right)}^{k+a^t_{\max}-1} n_1^t(k,l), \quad [3]$$

where $a^t_{\min}$ ($a^t_{\max}$) is the minimum (maximum) length of DNA wrapped around a particle of type t.

The inverse problem of predicting DNA binding energies from one-particle distributions can also be solved. We obtain

$$\beta[u_t(i,j) - \mu_t] = -\ln\left[\frac{n_1^t(i,j)Z}{Z^-(i)Z^+(j)}\right], \quad [4]$$

where Z, Z^-, Z^+ are found recursively using only the one-particle distribution $n_1^t(i,j)$ as input (*SI Appendix*).

ACKNOWLEDGMENTS. We thank Leonid Mirny for insightful discussions. This research was supported by the National Institutes of Health (R01 HG004708 to A.V.M.). A.V.M. is an Alfred P. Sloan Research Fellow.

1. van Holde KE (1989) *Chromatin* (Springer, New York).
2. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389(6648):251–260.
3. Felsenfeld G, Groudine M (2003) Controlling the double helix. *Nature* 421(6921):448–453.
4. Brogaard K, Xi L, Wang JP, Widom J (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature* 486(7404):496–501.
5. Dingwall C, Lomonossoff GP, Laskey RA (1981) High sequence specificity of micrococcal nuclease. *Nucleic Acids Res* 9(12):2659–2673.
6. Chung HR, et al. (2010) The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One* 5(12):e15754.
7. Flaus A, Luger K, Tan S, Richmond TJ (1996) Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals. *Proc Natl Acad Sci USA* 93(4):1370–1375.
8. Engeholm M, et al. (2009) Nucleosomes can invade DNA territories occupied by their neighbors. *Nat Struct Mol Biol* 16(2):151–158.
9. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol* 319(5):1097–1113.
10. Richmond TJ, Davey CA (2003) The structure of DNA in the nucleosome core. *Nature* 423(6936):145–150.
11. Polach KJ, Widom J (1995) Mechanism of protein access to specific DNA sequences in chromatin: A dynamic equilibrium model for gene regulation. *J Mol Biol* 254(2):130–149.
12. Anderson JD, Thåström A, Widom J (2002) Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation. *Mol Cell Biol* 22(20):7147–7157.
13. Li G, Widom J (2004) Nucleosomes facilitate their own invasion. *Nat Struct Mol Biol* 11(8):763–769.
14. Tims HS, Gurunathan K, Levitus M, Widom J (2011) Dynamics of nucleosome invasion by DNA binding proteins. *J Mol Biol* 411(2):430–448.
15. Adams CC, Workman JL (1995) Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol Cell Biol* 15(3):1405–1421.
16. Miller JA, Widom J (2003) Collaborative competition mechanism for gene activation in vivo. *Mol Cell Biol* 23(5):1623–1632.
17. Moyle-Heyrman G, Tims HS, Widom J (2011) Structural constraints in collaborative competition of transcription factors against the nucleosome. *J Mol Biol* 412(4):634–646.
18. Mirny LA (2010) Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci USA* 107(52):22534–22539.
19. Teif VB, Ettig R, Rippe K (2010) A lattice model for transcription factor access to nucleosomal DNA. *Biophys J* 99(8):2597–2607.
20. Prinsen P, Schiessel H (2010) Nucleosome stability and accessibility of its DNA to proteins. *Biochimie* 92(12):1722–1728.
21. Möbius W, Osberg B, Tsankov AM, Rando OJ, Gerland U (2013) Toward a unified physical model of nucleosome patterns flanking transcription start sites. *Proc Natl Acad Sci USA* 110(14):5719–5724.
22. Morozov AV, et al. (2009) Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Res* 37(14):4707–4722.
23. Brogaard KR, Xi L, Wang JP, Widom J (2012) A chemical approach to mapping nucleosomes at base pair resolution in yeast. *Methods Enzymol* 513:315–334.
24. Zlatanova J, Seebart C, Tomschik M (2008) The linker-protein network: Control of nucleosomal DNA accessibility. *Trends Biochem Sci* 33(6):247–253.
25. Syed SH, et al. (2010) Single-base resolution mapping of H1-nucleosome interactions and 3D organization of the nucleosome. *Proc Natl Acad Sci USA* 107(21):9620–9625.
26. Georgieva M, Roguev A, Balashev K, Zlatanova J, Miloshev G (2012) Hho1p, the linker histone of Saccharomyces cerevisiae, is important for the proper chromatin organization in vivo. *Biochim Biophys Acta* 1819(5):366–374.
27. Moore SC, Ausió J (1997) Major role of the histones H3-H4 in the folding of the chromatin fiber. *Biochem Biophys Res Commun* 230(1):136–139.
28. Schäfer G, McEvoy CR, Patterton HG (2008) The Saccharomyces cerevisiae linker histone Hho1p is essential for chromatin compaction in stationary phase and is displaced by transcription. *Proc Natl Acad Sci USA* 105(39):14838–14843.
29. Widom J (1992) A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc Natl Acad Sci USA* 89(3):1095–1099.
30. Wang JP, et al. (2008) Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae. *PLOS Comput Biol* 4(9):e1000175.
31. Chereji RV, Tolkunov D, Locke G, Morozov AV (2011) Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions. *Phys Rev E Stat Nonlin Soft Matter Phys* 83(5 Pt 1):050903.
32. Hall MA, et al. (2009) High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nat Struct Mol Biol* 16(2):124–129.
33. Forties RA, et al. (2011) A quantitative model of nucleosome dynamics. *Nucleic Acids Res* 39(19):8306–8313.
34. Thåström A, et al. (1999) Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol* 288(2):213–229.
35. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S (2011) Epigenome characterization at single base-pair resolution. *Proc Natl Acad Sci USA* 108(45):18318–18323.
36. Rhee HS, Pugh BF (2012) Genome-wide structure and organization of eukaryotic preinitiation complexes. *Nature* 483(7389):295–301.
37. Dion MF, et al. (2007) Dynamics of replication-independent histone turnover in budding yeast. *Science* 315(5817):1405–1408.
38. Mavrich TN, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* 18(7):1073–1083.
39. Chereji RV, Morozov AV (2011) Statistical mechanics of nucleosomes constrained by higher-order chromatin structure. *J Stat Phys* 144(2):379–404.
40. Thåström A, Lowary PT, Widom J (2004) Measurement of histone-DNA interaction free energy in nucleosomes. *Methods* 33(1):33–44.
41. Locke G, Tolkunov D, Moqtaderi Z, Struhl K, Morozov AV (2010) High-throughput sequencing reveals a simple model of nucleosome energetics. *Proc Natl Acad Sci USA* 107(49):20998–21003.
42. Locke G, Haberman D, Johnson SM, Morozov AV (2013) Global remodeling of nucleosome positions in C. elegans. *BMC Genomics* 14:284.
43. Cole HA, Howard BH, Clark DJ (2011) The centromeric nucleosome of budding yeast is perfectly positioned and covers the entire centromere. *Proc Natl Acad Sci USA* 108(31):12687–12692.
44. Nagarajavel V, Iben JR, Howard BH, Maraia RJ, Clark DJ (2013) Global 'bootprinting' reveals the elastic architecture of the yeast TFIIIB-TFIIIC transcription complex in vivo. *Nucleic Acids Res* 41(17):8135–8143.
45. Kaplan N, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458(7236):362–366.

Supplementary Information for: Ubiquitous nucleosome crowding in the yeast genome

Răzvan V. Chereji Alexandre V. Morozov

Contents

1 Supplementary Methods **1**
1.1 Direct problem: Matrix solution . . . 1
1.1.1 Single particle type . . . 1
1.1.2 Multiple particle types . . . 4
1.2 Direct problem: Recursive solution for hard-core interactions . . . 5
1.2.1 General case . . . 5
1.2.2 Special case: Fixed DNA footprint . . . 6
1.3 Inverse problem: Recursive solution for hard-core interactions . . . 6
1.3.1 General case . . . 7
1.3.2 Special case: Fixed DNA footprint . . . 8
1.4 Sequence-specific nucleosome formation energies . . . 8
1.5 Parameter optimization . . . 9
1.6 Site-specific chemical cleavage bias . . . 11
1.7 Experimental data . . . 13

2 Supplementary Results **14**

3 Supplementary Figures **24**

4 Supplementary Tables **34**

1 Supplementary Methods

1.1 Direct problem: Matrix solution

1.1.1 Single particle type

We consider a problem of mutually interacting particles (one-dimensional rods) that can be reversibly adsorbed to a one-dimensional lattice of L sites [here, DNA base pairs (bps)]. In order to model nucleosomes with partially wrapped DNA, we allow the particles to cover a variable number of bps between $a_{\min}$ and $a_{\max}$. We assume that the particles cannot overlap while they are attached to the lattice. This is implemented using hard-core interactions between adjacent particles. There are also hard walls at the ends of the lattice so that particles are prevented from running off it. In addition, we allow generic two-body interactions between nearest-neighbor particles.

The attachment of a particle to the DNA modifies the total energy of the system in a sequence-specific manner. Physically, the binding energy may have contributions from DNA bending, electrostatic interactions, hydrogen bond formation, van der Waals contacts, etc. Thus a particle which covers bps $k, k+1, \ldots, l$ has a total one-body binding energy $u(k,l)$. Note that for pairs of coordinates (k,l) such

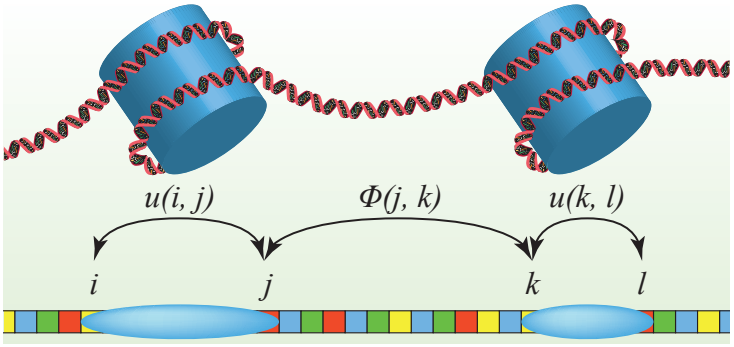

Figure S1: Schematic illustration of one-body and two-body potentials in a multi-nucleosome system. Nucleosomes may be partially wrapped, resulting in variable DNA footprints.

that $l-k+1 > a_{\max}$ or $l-k+1 < a_{\min}$, $u(k,l) = \infty$, because all particles must have the length between $a_{\min}$ and $a_{\max}$ bps. The theory presented below is valid for arbitrary binding energies $u(k,l)$.

Let $\Phi(j,k)$ be the two-body interaction between a pair of nearest-neighbor particles which cover base pairs $\ldots, j-1, j$ and $k, k+1, \ldots$ $(k > j)$ (Fig. S1). In the case of nucleosomes, such interactions may be used to account for the effects of higher-order chromatin structure [1]. Although we do not focus on two-body interactions in this work, they are included below for the sake of generality. We impose

$$\Phi(j,k) = \begin{cases} \infty & \text{if } k \le j, \\ V(k-j-1) & \text{if } k > j, \end{cases}$$

where $V(d)$ is an arbitrary interaction potential which depends only on the linear distance d between two neighboring particles.

For a fixed number of particles attached to the DNA, N, the canonical partition function is

$$Q_N = \sum_{\{i_n=1\ldots L\}_{n\in\{1,\ldots,2N\}}} e^{-\beta u(i_1,i_2)} e^{-\beta\Phi(i_2,i_3)} e^{-\beta u(i_3,i_4)} \ldots \times e^{-\beta u(i_{2N-3},i_{2N-2})} e^{-\beta\Phi(i_{2N-2},i_{2N-1})} e^{-\beta u(i_{2N-1},i_{2N})}, \quad (1)$$

where $\beta = 1/(k_B T)$ is the inverse temperature (k_B is the Boltzmann constant). Note that with our definitions of one-body energies, two-body interactions, and hard-wall boundary conditions, only legitimate configurations of non-overlapping particles will contribute to Eq. (1).

In order to simplify the notation, we introduce two $L \times L$ matrices:

$$\langle k|e|l\rangle = e^{-\beta u(k,l)},$$
$$\langle k|w|l\rangle = e^{-\beta\Phi(k,l)}.$$

Here $\langle k|M|l\rangle$ represents the element of matrix M in row k and column l. $|l\rangle$ is a column vector of dimension L with 1 at position l and 0 everywhere else, and $\langle k|$ is a row vector with 1 at position k and 0 otherwise.

Let $|J\rangle$ be a vector of dimension L with 1 at every position. Eq. (1) gives

$$Q_N = \begin{cases} \langle J|(ew)^{N-1}e|J\rangle & \text{if } N \ge 1, \\ 1 & \text{if } N = 0. \end{cases}$$

If the particles are allowed to attach and detach from the lattice, the system has a variable number

of particles, and the grand-canonical partition function is

$$\begin{aligned} Z &= \sum_{N=0}^{N_{\max}} e^{\beta N\mu} Q_N \\ &= 1 + \sum_{N=1}^{N_{\max}} \langle J|(zw)^{N-1}z|J\rangle \\ &= 1 + \sum_{M=0}^{\infty} \langle J|(zw)^{M}z|J\rangle \\ &= 1 + \langle J|(I-zw)^{-1}z|J\rangle, \end{aligned} \tag{2}$$

where μ is the chemical potential, $N_{\max}$ is the maximum number of particles that can fit on L bp, I is the identity matrix, and

$$\langle k|z|l\rangle = e^{\beta[\mu - u(k,l)]} \equiv \zeta(k,l).$$

Note that all particle configurations with $N > N_{\max}$ do not contribute to Z, allowing us to extend the upper limit from $N_{\max}$ to ∞. From the partition function, we can compute s-particle distribution functions (see the chapter by Stell in [2]):

$$\begin{aligned} n_1(k,l) &= \frac{\zeta(k,l)}{Z}\frac{\delta Z}{\delta\zeta(k,l)}, \\ n_2(i,j;k,l) &= \frac{\zeta(i,j)\zeta(k,l)}{Z}\frac{\delta^2 Z}{\delta\zeta(i,j)\delta\zeta(k,l)}, \end{aligned}$$

and in general

$$n_s(i_{1L},i_{1R};\ldots;i_{sL},i_{sR}) = \frac{\zeta(i_{1L},i_{1R})\ldots\zeta(i_{sL},i_{sR})}{Z}\frac{\delta^s Z}{\delta\zeta(i_{1L},i_{1R})\ldots\delta\zeta(i_{sL},i_{sR})}.$$

Using these relations, we obtain the one-particle distribution

$$n_1(k,l) = \frac{1}{Z}\langle J|(I-zw)^{-1}|k\rangle\langle k|z|l\rangle\langle l|(I-wz)^{-1}|J\rangle, \tag{3}$$

and the two-particle distribution

$$n_2(i,j;k,l) = \frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|z|j\rangle\langle j|w(I-zw)^{-1}|k\rangle\langle k|z|l\rangle\langle l|(I-wz)^{-1}|J\rangle. \tag{4}$$

In particular, the nearest-neighbor two-particle distribution is given by

$$\overline{n}_2(i,j;k,l) = \frac{1}{Z}\langle J|(I-zw)^{-1}|i\rangle\langle i|z|j\rangle\langle j|w|k\rangle\langle k|z|l\rangle\langle l|(I-wz)^{-1}|J\rangle. \tag{5}$$

Eqs. (3) and (4) allow an obvious interpretation. To find the probability of starting a particle covering positions from k to l [Eq. (3)], we have to add statistical weights of all the configurations that contain a particle at that position, and divide the resulting sum by the partition function. Similarly, in order to find the probability of a pair of particles, one covering positions i to j and the other covering positions k to l [Eq. (4)], we need to sum statistical weights of all the configurations containing that pair of particles, and divide by the partition function.

With the one-particle distribution, we can define the occupancy of a bp i as the probability of finding that bp in contact with any particle. In other words, we need to sum the probabilities of all configurations in which particles cover bp i:

$$\mathrm{Occ}(i) = \sum_{k=i-a_{\max}+1}^{i} \sum_{l=\max(i,k+a_{\min})}^{k+a_{\max}-1} n_1(k,l).$$

Note that $1 - \mathrm{Occ}(i)$ is the probability that bp i is not covered by any particles.

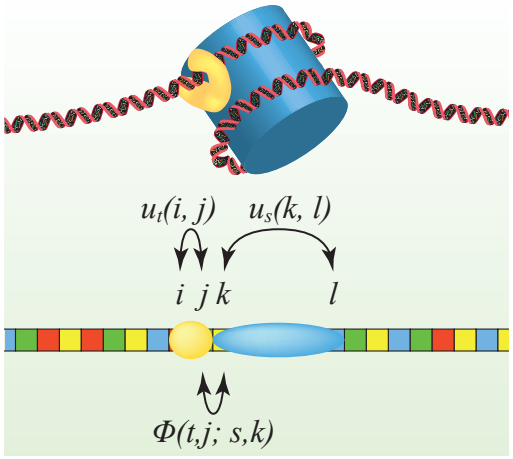

Figure S2: Schematic illustration of one-body and two-body potentials in a system with multiple particle types. In principle, the model allows all particles to adopt partially wrapped states. In practice, nucleosomes are allowed to be partially wrapped, but transcription factors (TFs) and other DNA-binding proteins have fixed DNA footprints.

1.1.2 Multiple particle types

The above formalism can be easily extended to the case in which T types of particles can attach to the one-dimensional lattice. Let the binding energy of a particle of type $t \in \{1, \ldots, T\}$ that covers bps i to j on the lattice be $u_t(i, j)$. The interaction between a particle of type t ending at position k, and the next particle of type s starting at position l will be denoted by $\Phi(t, k; s, l)$ (Fig. S2). Each particle of type t, when attached to the DNA, is in contact with a number of bps between $a^t_{\min}$ and $a^t_{\max}$. Thus $u_t(i, j) = \infty$ if i and j do not satisfy the constraints $a^t_{\min} \leq j - i + 1 \leq a^t_{\max}$. Also, $\Phi(t, i; s, j) = \infty$ for $j \leq i$ since the particles cannot overlap. With this notation, the grand-canonical partition function becomes

$$Z = \sum_{\text{all states}} e^{-\beta[u_{t1}(i_{1L}, i_{1R}) - \mu_{t_1}]} e^{-\beta\Phi(t_1, i_{1R}; t_2, i_{2L})} e^{-\beta[u_{t_2}(i_{2L}, i_{2R}) - \mu_{t_2}]} \ldots, \tag{6}$$

where μ_t is the chemical potential of the particles of type t. The sum is over all configurations, which can have variable numbers of particles of any type.

Using the matrix notation,

$$\langle t, k|z|s, l\rangle = e^{-\beta[u_t(k,l) - \mu_t]}\delta_{ts} \equiv \zeta_t(k, l)\delta_{ts},$$
$$\langle t, k|w|s, l\rangle = e^{-\beta\Phi(t,k;s,l)},$$

where δ_{ts} is the Kronecker delta symbol, the partition function can be written as

$$Z = \sum_{\text{all states}} \langle t_1, i_{1L}|z|t_1, i_{1R}\rangle\langle t_1, i_{1R}|w|t_2, i_{2L}\rangle\langle t_2, i_{2L}|z|t_2, i_{2R}\rangle \ldots$$

Each vector $|t, i\rangle$ has dimension TL and only one non-zero element, set to 1 for normalization. For example, $|1, i\rangle$ vectors have a 1 at position i, $|2, i\rangle$ vectors have a 1 at position $L + i$, etc. As above, we denote by $|J\rangle$ a vector in which all TL elements are equal to 1, to obtain

$$Z = 1 + \langle J|(I - zw)^{-1}z|J\rangle,$$

equivalent to Eq. (2).

Similarly to the previous case of a single particle type, we can compute the one-particle density

$$\begin{aligned} n_1^t(k,l) &= \frac{\zeta_t(k,l)}{Z}\frac{\delta Z}{\delta \zeta_t(k,l)} \\ &= \frac{1}{Z}\langle J|(I-zw)^{-1}|t,k\rangle\langle t,k|z|t,l\rangle\langle t,l|(I-wz)^{-1}|J\rangle. \end{aligned} \tag{7}$$

We can also obtain the two-particle density

$$\begin{aligned} n_2^{t,s}(i,j;k,l) &= \frac{1}{Z}\langle J|(I-zw)^{-1}|t,i\rangle\langle t,i|z|t,j\rangle \\ &\quad \times\langle t,j|w(I-zw)^{-1}|s,k\rangle\langle s,k|z|s,l\rangle\langle s,l|(I-wz)^{-1}|J\rangle \end{aligned} \tag{8}$$

and the nearest-neighbor two-particle density

$$\begin{aligned} \overline{n}_2^{t,s}(i,j;k,l) &= \frac{1}{Z}\langle J|(I-zw)^{-1}|t,i\rangle\langle t,i|z|t,j\rangle \\ &\quad \times\langle t,j|w|s,k\rangle\langle s,k|z|s,l\rangle\langle s,l|(I-wz)^{-1}|J\rangle. \end{aligned} \tag{9}$$

Eqs. (8) and (9) give the joint probability that a particle of type t covers bps i to j, while a second particle of type s covers bps k to l.

Using Eq. (7) we can compute occupancy for each type of particles t and for each bp i:

$$\mathrm{Occ}_t(i) = \sum_{k=i-a^t_{\max}+1}^{i}\ \sum_{l=\max(i,k+a^t_{\min})}^{k+a^t_{\max}-1} n_1^t(k,l).$$

In the following sections we will focus on the one-particle density function.

1.2 Direct problem: Recursive solution for hard-core interactions

A straightforward application of Eqs. (7) and (8) entails computationally intensive matrix manipulations. Fortunately, for particles that interact only through hard-core repulsion rather than long-range two-body interactions, the one-particle distribution can be computed recursively and therefore much more efficiently.

1.2.1 General case

With multiple particle types, Eq. (7) can be rewritten as

$$n_1^t(i,j) = \frac{1}{Z}Z^-(i)\langle t,i|z|t,j\rangle Z^+(j), \tag{10}$$

where $Z^-(i)$ and $Z^+(j)$ are the partition functions for the domains $[1,i)$ and $(j,L]$, respectively. Note that in the case of hard-core interactions alone, $Z^-(i)$ and $Z^+(j)$ do not depend on the type of the particle occupying positions i through j.

In the case of steric exclusion alone, these partial partition functions satisfy the following recursion relations:

$$Z^-(i) = Z^-(i-1) + \sum_s \sum_{i-a^s_{\max}\le j\le i-a^s_{\min}} Z^-(j)\langle s,j|z|s,i-1\rangle, \tag{11}$$

and

$$Z^+(i) = Z^+(i+1) + \sum_s \sum_{i+a^s_{\min}\le j\le i+a^s_{\max}} \langle s,i+1|z|s,j\rangle Z^+(j). \tag{12}$$

Here each particle type s has two characteristic lengths, corresponding to its minimum and maximum DNA footprints, respectively: $a^s_{\min}$ and $a^s_{\max}$. The boundary conditions are $Z^-(1)=1$ and $Z^+(L)=1$.

The full partition function is given by $Z = Z^-(L+1) = Z^+(0)$. Note that all unphysical terms for which bound particles run off the lattice automatically vanish from Eqs. (11) and (12). To avoid numeric instabilities, the recursion should be done in log space. Let

$$F(i) = \ln Z^-(i),$$
$$R(i) = \ln Z^+(i).$$

With this notation, Eqs. (11) and (12) become

$$F(i) = F(i-1) + \ln\Big\{1 + \sum_s \sum_{i-a^s_{\max} \le j \le i-a^s_{\min}} e^{F(j)-F(i-1)+\beta[\mu_s - u_s(j,i-1)]}\Big\}, \tag{13}$$
$$R(i) = R(i+1) + \ln\Big\{1 + \sum_s \sum_{i+a^s_{\min} \le j \le i+a^s_{\max}} e^{R(j)-R(i+1)+\beta[\mu_s - u_s(i+1,j)]}\Big\},$$

with the boundary conditions $F(1) = R(L) = 0$.

Then the one-particle distribution function is

$$n_1^t(i,j) = e^{F(i)+R(j)-\ln Z+\beta[\mu_t - u_t(i,j)]}, \tag{14}$$

where $\ln Z = F(L+1) = R(0)$.

The two-particle distribution given by Eq. (8) can be computed in a similar way. The only new ingredient is the partition function for the box with walls at two arbitrary positions, $Z(t,j,s,k) \equiv \langle t,j|w(I-zw)^{-1}|s,k\rangle$. This partition function can be computed recursively, exactly as the partial partition functions $Z^\pm$ discussed above. Finally, in the specific case of the nearest-neighbor two-particle distribution, Eq. (9) becomes

$$\overline{n}_2^{t,s}(i,j;k,l) = \frac{1}{Z} Z^-(i)\langle t,i|z|t,j\rangle\Theta(k-j)\langle s,k|z|s,l\rangle Z^+(l), \tag{15}$$

where Θ is the Heaviside step function.

1.2.2 Special case: Fixed DNA footprint

The special case in which all particles are fully attached to their DNA sites (i.e., DNA is completely wrapped) can be easily obtained from our general formalism. Indeed, in this case we restrict $a^s_{\min} = a^s_{\max} = a^s$ in Eq. (13), obtaining

$$F(i) = F(i-1) + \ln\Big\{1 + \sum_s e^{F(i-a^s)-F(i-1)+\beta[\mu_s - u_s(i-a^s,i-1)]}\Big\}, \tag{16}$$
$$R(i) = R(i+1) + \ln\Big\{1 + \sum_s e^{R(i+a^s)-R(i+1)+\beta[\mu_s - u_s(i+1,i+a^s)]}\Big\}.$$

As before, the boundary conditions are $F(1) = R(L) = 0$.

The one-particle distribution is given by

$$n_1^t(i,i+a^t-1) = e^{F(i)+R(i+a^t-1)-\ln Z+\beta\left[\mu_t - u_t(i,i+a^t-1)\right]}. \tag{17}$$

1.3 Inverse problem: Recursive solution for hard-core interactions

In the previous section, we have solved the direct problem: given the binding energies for all particle types, we compute s-particle distributions. However, typically it is particle distributions that are observed experimentally, and the energetics of particle-DNA interactions need to be inferred. This inverse problem

can be solved recursively for the case of systems with multiple particle types, partial wrapping (variable DNA footprints), and steric exclusion. The recursive solution is efficient enough to be employed on the genome-wide scale. Here we focus on one-particle distributions and one-body energies; the exact matrix formulation of the inverse problem for a single particle type with the two-particle distribution and the two-body potential is available in Ref. [1].

1.3.1 General case

Using Eqs. (7), (11) and (12), we obtain:

$$\begin{aligned} Z^-(i) &= Z^-(i-1)\left[1+\sum_{\substack{t,\\ i-a^t_{\max}\le j\le i-a^t_{\min}}} \frac{Z}{Z^-(i-1)Z^+(i-1)} n_1^t(j,i-1)\right] \\ &= Z^-(i-1)\left[1+\frac{N^R(i-1)}{\xi(i-1)}\right], \end{aligned} \tag{18}$$

where $N^R(i)=\sum_t \sum_{i-a^t_{\max}+1\le j\le i-a^t_{\min}+1} n_1^t(j,i)$ represents the probability of finding a particle of any type with the right edge at bp i, and $\xi(i)=Z^-(i)Z^+(i)/Z$.

Z^+ satisfies a similar recursive relation:

$$Z^+(i) = Z^+(i+1)\left[1+\frac{N^L(i+1)}{\xi(i+1)}\right], \tag{19}$$

where $N^L(i)$ is the probability of finding a particle of any type with the left edge at bp i:

$$N^L(i) = \sum_t \sum_{i+a^t_{\min}-1\le j\le i+a^t_{\max}-1} n_1^t(i,j). \tag{20}$$

The quantity $\xi(i)$ satisfies

$$\begin{aligned} \xi(i+1)-\xi(i) =& \frac{1}{Z}\left[Z^-(i+1)Z^+(i+1)-Z^-(i)Z^+(i)\right] \\ =& \frac{1}{Z}\Big\{Z^-(i+1)\left[Z^+(i+1)-Z^+(i)\right] \\ &+ Z^+(i)\left[Z^-(i+1)-Z^-(i)\right]\Big\} \\ =& N^R(i)-N^L(i+1), \end{aligned}$$

so that

$$\xi(i) = 1+\sum_{k=0}^{i-1}\left[N^R(k)-N^L(k+1)\right], \tag{21}$$

where the initial condition $\xi(0)=1$ has been used.

After we compute both Z^- and Z^+ in this way, the total partition function is given by $Z=Z^-(L+1)=Z^+(0)$ as before, and the binding energy for any particle attached to the DNA is given by

$$\beta\left[u_t(i,j)-\mu_t\right] = -\ln\left[n_1^t(i,j)\frac{Z}{Z^-(i)Z^+(j)}\right]. \tag{22}$$

1.3.2 Special case: Fixed DNA footprint

In the case of the all-or-none binding, all matrix elements $\langle i|n_1^t|j\rangle$ vanish unless $j = i + a^t - 1$, where a^t is the length of the binding site for the particle of type t. Thus

$$N^L(i) = \sum_t n_1^t(i, i + a^t - 1),$$
$$N^R(i) = \sum_t n_1^t(i - a^t + 1, i).$$

Using these expressions, we can employ Eqs. (18), (19) and (21) to compute Z^+ and Z^- in log space. Finally, Eq. (22) can be used to compute the binding energies.

If all particles are of the same type, the quantity ξ can be simplified further:

$$\xi(i) = 1 - \sum_{k=i-a+1}^{i} N^L(k) = 1 - \mathrm{Occ}(i),$$

where $\mathrm{Occ}(i)$ is the probability that bp i is covered by a particle. Thus in this limit $\xi(i)$ is simply the probability that bp i is not bound by any particles.

The recursion relations for Z^- and Z^+ become

$$Z^-(i+1) = Z^-(i)\left[1 + \frac{N^L(i-a+1)}{1 - \mathrm{Occ}(i)}\right],$$
$$Z^+(i) = Z^+(i+1)\left[1 + \frac{N^L(i+1)}{1 - \mathrm{Occ}(i+1)}\right].$$

These expressions are equivalent to those previously obtained in Ref. [3].

1.4 Sequence-specific nucleosome formation energies

The binding energy of a nucleosome is the sum of two components. One is favorable interactions between the negatively charged DNA wrapped around the positively charged histone octamer. The other is the elastic energy required to bend the DNA polymer into the left-handed nucleosomal superhelix. The total energy of the nucleosome formation is likely to be negative for almost any DNA sequence.

We model the total nucleosome formation energy $u_{S(N)}$ as a function of the DNA sequence $S(N) = S_1 S_2 \cdots S_N$ wrapped around the histone octamer, where S_i is the nucleotide at position i in a given DNA sequence, and N is its length, which can be different from 147 bp. In general, $u_{S(N)}$ varies among different DNA sequences of length N because the free energy cost of DNA bending is sequence-dependent. Let us denote the average energy of all genomic sequences of length N by $\langle u_{S(N)}\rangle$, and the deviation of the energy of a given nucleosome from this average by $\delta u_{S(N)} = u_{S(N)} - \langle u_{S(N)}\rangle$. Then

$$u_{S(N)} = \langle u_{S(N)}\rangle + \delta u_{S(N)} \equiv u_N^{\mathrm{SI}} + u_{S(N)}^{\mathrm{SD}},$$

where $u_N^{\mathrm{SI}} = \langle u_{S(N)}\rangle$ and $u_{S(N)}^{\mathrm{SD}} = \delta u_{S(N)}$ represent the sequence-independent and sequence-dependent contributions to the binding energy, respectively. Note that by definition

$$\langle u_{S(N)}^{\mathrm{SD}}\rangle = 0, \tag{23}$$

where the average is over all distinct genomic sequences of length N.

We assume that the sequence-dependent part of the total nucleosome formation energy depends only on the mono- and dinucleotide counts in the nucleosomal DNA [3]:

$$u_{S(N)}^{SD} = \sum_{i=1}^{N} \epsilon_{S_i} + \sum_{i=1}^{N-1} \epsilon_{S_i S_{i+1}},$$

where ϵ_{S_i} and $\epsilon_{S_iS_{i+1}}$ are the contributions of the mononucleotide S_i and the dinucleotide S_iS_{i+1}, respectively. Because for each sequence $S(N)$, we also include its reverse complement $\widetilde{S}(N)$ in the modeling, $u^{SD}_{S(N)}$ is a function of 12 unique parameters: $\epsilon_{A/T}$, $\epsilon_{C/G}$, $\epsilon_{AA/TT}$, $\epsilon_{AC/GT}$, $\epsilon_{AG/CT}$, $\epsilon_{AT/AT}$, $\epsilon_{CA/TG}$, $\epsilon_{CC/GG}$, $\epsilon_{CG/CG}$, $\epsilon_{GA/TC}$, $\epsilon_{GC/GC}$, and $\epsilon_{TA/TA}$.

The sequence-dependent energy of a histone octamer attached to bps i through j (such that $N = j - i + 1 \leq 147$ bp and $S(N) = S_iS_{i+1}\dots S_j$) is thus given by

$$u^{SD}(i,j) = \sum_{k=i}^{j} \epsilon_{S_k} + \sum_{k=i}^{j-1} \epsilon_{S_kS_{k+1}},$$

where ϵ_{S_k} is the energy contribution of the mononucleotide pair $S_k/\widetilde{S}_k$, and $\epsilon_{S_kS_{k+1}}$ is the energy contribution of the dinucleotide pair $S_kS_{k+1}/\widetilde{S}_{k+1}\widetilde{S}_k$ ($\widetilde{S}_k$ is a nucleotide complementary to S_k).

The 12 parameters which describe nucleosome formation energies are not all independent. Eq. (23) implies that

$$\langle \sum_{k=i}^{j} \epsilon_{S_k} + \sum_{k=i}^{j-1} \epsilon_{S_kS_{k+1}} \rangle = 0,$$

which is equivalent to

$$(j-i+1)\langle \epsilon_{S_k} \rangle + (j-i)\langle \epsilon_{S_kS_{k+1}} \rangle = 0.$$

This has to be true for all sequence lengths N, so that

$$\begin{cases} \langle \epsilon_{S_k} \rangle = 0, \\ \langle \epsilon_{S_kS_{k+1}} \rangle = 0, \end{cases}$$

or, equivalently,

$$\begin{cases} \epsilon_{A/T} f(A/T) + \epsilon_{C/G} f(C/G) = 0, \\ \epsilon_{AA/TT} f(AA/TT) + \epsilon_{AC/GT} f(AC/GT) + \ldots + \epsilon_{TA/TA} f(TA/TA) = 0, \end{cases} \tag{24}$$

where $f(S_i/\widetilde{S}_i)$ and $f(S_iS_{i+1}/\widetilde{S}_{i+1}\widetilde{S}_i)$ represent the genomic frequencies of mononucleotide pair $S_i/\widetilde{S}_i$ and dinucleotide pair $S_iS_{i+1}/\widetilde{S}_{i+1}\widetilde{S}_i$, respectively. *S. cerevisiae* mono- and dinucleotide genomic frequencies are:

Sequence	Frequency
A/T	0.6170
C/G	0.3830
AA/TT	0.2161
AC/GT	0.1054
AG/CT	0.1168
AT/AT	0.0894
CA/TG	0.1297
CC/GG	0.0779
CG/CG	0.0294
GA/TC	0.1247
GC/GC	0.0375
TA/TA	0.0733

1.5 Parameter optimization

Parameter inference for the total free energy of nucleosome formation is done in two steps.

In the first step, we estimate the sequence-independent part of the nucleosome formation energy, $u^{\mathrm{SI}}(i,j)$. Because $\langle u^{\mathrm{SD}}(i,j) \rangle = 0$, in the first approximation we can neglect the contribution of the

sequence-dependent part. We test eight different models of sequence-independent energies (SI Results, Models A-H). In each case, we predict a nucleosome distribution $n_1^{\rm nuc}(i,j)$, which gives both the distributions of nucleosome footprint sizes and inter-dyad distances. We find the optimal set of parameters, for all models, by minimizing the error between the histograms of lengths predicted by the model and that observed in the experiments. In the case of data from Brogaard et al. [4], the paired-end DNA fragments yield the histogram of inter-dyad distances, whereas in the case of an MNase-seq experiment, the paired-end reads give the histogram of MNase-protected nucleosomal footprints. Both of these histograms can be used to fit the parameters of Models A-H (SI Results).

In particular, with the Brogaard et al. data [4] we predict the observed distribution of inter-dyad distances by computing the conditional probability of a nucleosome with the dyad at bp $c+d$, given that the adjacent upstream nucleosome has its dyad at bp c:

$$P(c+d|c) = \frac{N_2(c,c+d)}{N_1(c)}. \tag{25}$$

The probability distributions of the nucleosome centers can be computed using Eqs. (10) and (15) for a single particle type:

$$N_1(c) = \sum_{\Delta_1} n_1^{\rm nuc}(c-\Delta_1, c+\Delta_1),$$
$$N_2(c,c+d) = \sum_{\Delta_1,\Delta_2} \overline{n}_2^{\rm nuc,nuc}(c-\Delta_1, c+\Delta_1; c+d-\Delta_2, c+d+\Delta_2).$$

Here $2\Delta_{1,2}+1$ are lengths of the particles centered at bp c and $c+d$, respectively. To estimate the conditional probability $P(c+d|c)$, we use $c = 5$ kbp and a box of length $L = 10$ kbp, so that the boundaries of the box are far away. We neglect nucleosome sequence specificity, convolve $P(c+d|c)$ with a kernel to account for site-specific chemical cleavage bias (Section 1.6), and fit model parameters to reproduce the observed distribution of inter-dyad distances.

Specifically, we employ the genetic algorithm optimization function `ga` from the MATLAB Global Optimization toolbox to minimize the objective function

$$\text{O.F.} = \begin{cases} RMS & \text{if } RMS \geq 10^{-3}, \\ RMS - r_{\rm osc} \simeq -r_{\rm osc} & \text{if } RMS < 10^{-3}, \end{cases}$$

where RMS is the root-mean-square deviation between predicted and observed inter-dyad distributions, and $r_{\rm osc}$ is the linear correlation between observed and predicted oscillations after the smooth background has been subtracted from the inter-dyad distributions, as in Fig. 1D. In this way, the parameters are initially optimized to capture the overall shape of the histogram. Once RMS decreases below a threshold of 10^{-3}, the objective function is replaced by $r_{\rm osc}$, and the fine oscillatory structure of the histogram is fitted. The optimized parameters for all models are given in SI Results.

In the second step of the optimization procedure, we compute the sequence-dependent part of the nucleosome formation energy, $u^{\rm SD}(i,j)$, by subtracting the sequence-independent part $u^{\rm SI}(i,j)$ from the total energy $u(i,j)$ given by Eq. (22). Thus we obtain the following system of equations:

$$u^{\rm SD}(i,j) - \mu = \sum_{k=i}^{j} \epsilon_{S_k} + \sum_{k=i}^{j-1} \epsilon_{S_k S_{k+1}} - \mu$$
$$= \begin{pmatrix} m_{A/T} & m_{C/G} & m_{AA/TT} & \cdots & m_{TA/TA} & -1 \end{pmatrix} \begin{pmatrix} \epsilon_{A/T} \\ \vdots \\ \epsilon_{TA/TA} \\ \mu \end{pmatrix}, \tag{26}$$

where μ is the chemical potential and $m_{X/\widetilde{X}}$, $m_{XY/\widetilde{Y}\widetilde{X}}$ are the counts of mono- and dinucleotide pairs $X/\widetilde{X}$ and $XY/\widetilde{Y}\widetilde{X}$ in the sequence, respectively.

Using all possible combinations of pairs (i, j) where a nucleosome can form, we obtain a large number P of equations of the type

$$E - \mu \;=\; M \;\begin{pmatrix} \epsilon \\ \mu \end{pmatrix}. \tag{27}$$

Here, $E - \mu$ is a column vector of dimension P, where each row contains one $u^{\mathrm{SD}}(i, j) - \mu$ element from Eq. (26). $\begin{pmatrix} \epsilon \\ \mu \end{pmatrix}$ is the column vector from Eq. (26), and M is a $P \times 13$ matrix with mono- and dinucleotide counts and -1's in the last column. Using Eq. (27), we derive the energy parameters ϵ and μ by a least-squares fit.

Because in every DNA sequence the number of mononucleotides is equal to the length of the sequence, and the number of dinucleotides is equal to the length of the sequence minus 1, the columns of the matrix M are not linearly independent. Indeed, the column vector

$$|V\rangle = \begin{pmatrix} 1 \\ 1 \\ -1 \\ \vdots \\ -1 \\ 1 \end{pmatrix}$$

is the only linearly independent vector from the kernel of M: $M|V\rangle = 0$, i.e. the kernel of M is spanned by $|V\rangle$. Thus the rank of matrix M is 12, which is greater than the number of independent parameters, 11 (we have 10 independent ϵ parameters since 2 out of 12 are fixed by Eq. (23)], and μ). This means that a least-squares fit with 2 constrains given by Eq. (23) will result in a unique set of parameters. The constrained linear least-squares problem was solved in MATLAB using function `lsqlin` from the Optimization toolbox.

1.6 Site-specific chemical cleavage bias

Hydroxyl radicals that cleave DNA near the nucleosome dyad have two preferred cutting sites, at positions -1 bp and +6 bp with respect to the dyad, with the 5' to 3' direction defined as positive [4, 5]. Subsequent treatment with Klenow polymerase removes 3' overhangs and fills in 5' overhangs to create blunted DNA fragments. Thus after the Klenow reaction, out of four possible cleavage sites corresponding to each double-stranded DNA fragment, only positions of the two cuts at the 5' end of each strand are retained. If the DNA is cut near each dyad at the two positions with frequencies f and $1 - f$, respectively, the left edge of the DNA fragment will have the cut at position

$$x^L_{\mathrm{cut}} = x^L_{\mathrm{dyad}} + b_L,$$

where x^L_{dyad} is the genomic position of the left dyad, and the offset b_L is

$$b_L = \begin{cases} -1 \\ 6 \end{cases} \quad \text{with probability} \quad \begin{cases} f \\ 1 - f \end{cases} \tag{28}$$

Similarly, the right cleavage site will be at position

$$x^R_{\mathrm{cut}} = x^R_{\mathrm{dyad}} + b_R,$$

where x^R_{dyad} is the genomic position of the right dyad, and the offset b_R is

$$b_R = \begin{cases} +1 \\ -6 \end{cases} \quad \text{with probability} \quad \begin{cases} f \\ 1 - f \end{cases} \tag{29}$$

Since the two cleavage events and the corresponding offsets are independent, the distance between two consecutive cuts is given by

$$\begin{aligned} d_{\text{cuts}} &= x^R_{\text{cut}} - x^L_{\text{cut}} \\ &= (x^R_{\text{dyad}} - x^L_{\text{dyad}}) + (b^R - b^L) \\ &= d_{\text{dyads}} + b_d, \end{aligned}$$

where d_{dyads} is the distance between two neighboring dyads, and the bias b_d is

$$b_d = \begin{cases} -12 \\ -5 \\ 2 \end{cases} \quad \text{with probability} \quad \begin{cases} (1-f)^2 \\ 2f(1-f) \\ f^2 \end{cases}$$

The probability distribution of the distance between the cleavage sites, d_{cuts}, is given by

$$P(d_{\text{cuts}} = z) = \sum_{x=-\infty}^{\infty} P(b_d = x) \cdot P(d_{\text{dyads}} = z - x),$$

which is the convolution of the inter-dyad distance probability, obtained from Eq. (25), with the kernel

$$F(x) \equiv P(b_d = x) = \begin{cases} (1-f)^2 & \text{for } x = -12, \\ 2f(1-f) & \text{for } x = -5, \\ f^2 & \text{for } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

The convolution gives the predicted distribution of the distance between neighboring cleavage sites, which is then compared with the observed distribution of paired-end DNA fragment lengths.

Note that each cleavage event gives rise to two DNA fragments, one to the left and one to the right of the cleavage site. Therefore, with high enough read coverage each dyad will generate an equal number of right and left DNA fragment ends. Using Eqs. (28) and (29), we obtain for every genomic bp:

$$\begin{aligned} \langle x_{\text{cut}} \rangle &= \frac{1}{2}\langle x^L_{\text{cut}} \rangle + \frac{1}{2}\langle x^R_{\text{cut}} \rangle \\ &= \frac{1}{2}(x_{\text{dyad}} + \langle b_L \rangle) + \frac{1}{2}(x_{\text{dyad}} + \langle b_R \rangle) \\ &= x_{\text{dyad}} + \frac{-f + 6(1-f)}{2} + \frac{f - 6(1-f)}{2} \\ &= x_{\text{dyad}}. \end{aligned}$$

Thus positions of observed hydroxyl cleavage sites averaged over cut frequencies coincide with dyad positions. In contrast,

$$\begin{aligned} \langle d_{\text{cuts}} \rangle &= d_{\text{dyads}} + \langle b_d \rangle \\ &= d_{\text{dyads}} + \left[-12(1-f)^2 - 10f(1-f) + 2f^2\right] \\ &= d_{\text{dyads}} + (14f - 12), \end{aligned}$$

so that the average distance between neighboring cleavage sites averaged over cut frequencies is shifted with respect to the inter-dyad distance by

$$\Delta = 14f - 12.$$

For $f \in [0.5, 1]$, $\Delta \in [-5, 2]$, meaning that the average fragment size differs from the inter-dyad distance by at most 5 bp. In particular, with $f = 0.5$ the inter-dyad distances are 5 bp longer that the observed distances between hydroxyl cut sites marking neighboring nucleosomes.

1.7 Experimental data

All datasets used in this study are available in the GEO database: `http://www.ncbi.nlm.nih.gov/gds`. For the study of distances between neighboring nucleosomal dyads, we used paired-end sequencing data from Brogaard et al. [4]. We took the data from the experiment with the smallest time delay between introduction of hydrogen peroxide and quenching the chemical mapping reaction (1.5 min as opposed to 20 min in the other paired-end dataset provided by Brogaard et al.). This helps to minimize, as much as possible, the effect of hydroxyl radical diffusion which might lead to random DNA cuts at non-dyad positions. The GEO accession number for the 1.5 min experiment is GSM880651. There are 40,788,544 mapped DNA reads in the 1.5 min dataset, resulting in the average genome-wide coverage of 6.76 dyads/bp (note that each read contains information about two neighboring dyads).

For the *in vitro* nucleosome map we combined two biological replicates from Kaplan et al. [6] (GSM351491). For *in vivo* nucleosome maps, we used paired-end reads from three studies: Henikoff et al. [7] (GSM756481), Cole et al. [8] (combination of two replicates: GSM651368 and GSM651369) and Nagarajavel et al. [9] (combination of two replicates: GSM1087269 and GSM1087270). All paired-end reads which yield DNA fragments of < 100 bp were excluded from the three paired-end datasets because they are likely to correspond to particles other than nucleosomes. We also excluded DNA fragments with > 147 bp from these datasets because longer fragments are more likely to come from underdigestion by MNase, and because the sequence-dependent energy in Eq. (26) is assumed to be valid only for DNA lengths of ≤ 147 bp. Because sequence reads that come from repetitive regions of the yeast genome cannot be mapped uniquely, we mark all regions annotated as repeat region, long terminal repeat (LTR), or ribosomal RNA (rRNA), and exclude them from nucleosome dyad counts that serve as input in predicting nucleosome energies. However, these regions were used in estimating average inter-dyad distances in Fig. S11. The genome annotations were obtained from the Saccharomyces Genome Database (`http://www.yeastgenome.org`).

2 Supplementary Results

Model A: Crystal structure augmented with an additional well. The binding energy of a particle of length $a = 1 + x_1 + x_2$ (1 bp for the dyad, and x_1 and x_2 for the extra number of bps in contact with the histone octamer on each side of the dyad) is given by $u = u_{\text{half}}(x_1) + u_{\text{half}}(x_2)$, with

$$u_{\text{half}}(x) = \texttt{interp1}(...) - \frac{E_b}{147}x,$$

where E_b is the binding energy of a fully wrapped particle in the absence of 10-11 bp oscillations. The oscillations are based on the the positions of the histone-DNA contacts in the crystal structure [10]. The MATLAB function `interp1`(...) was used to generate the oscillatory pattern by piecewise cubic Hermite interpolation between the following data points:

x (Position)	f(x) (Energy)
-1	-A
3	A
7	-A
13	A
17	-A
24	A
28	-A
34	A
38	-A
44	A
49	-A
55	A
59	-A
65	A
69	-A
75	A
p	-d
85	A

The oscillatory pattern was superimposed onto a line with the slope of $-E_b/147$.

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	3 bp
E_b	14.39 k_BT
μ	-14.51 k_BT
A	1.13 k_BT
f	0.51
p	79 bp
d	0.86 k_BT

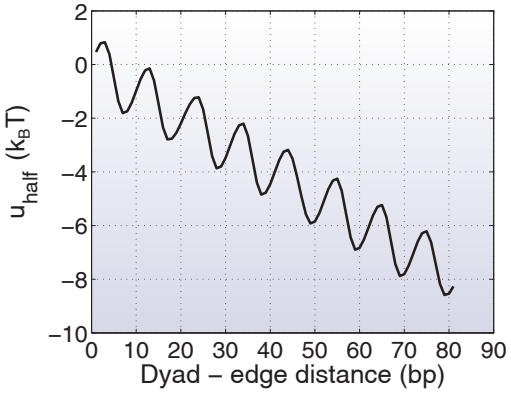

Table S1: Fitted parameters for Model A. $a_{\max}$ and $a_{\min}$ are the maximum and minimum lengths of the nucleosome particle, μ is the histone octamer chemical potential, A is the amplitude of the oscillations, and f is the hydroxyl radical cutting frequency. p and d are the position and the depth of the first minimum outside of the nucleosome core particle, respectively. E_b is the binding energy of a fully wrapped particle in the absence of 10-11 bp oscillations.

Fit residuals:

$$RMS = 9.996 \times 10^{-4}$$
$$r_{\text{osc}} = 0.764$$
$$RMS_{\text{osc}} = 1.866 \times 10^{-4}$$

RMS is the root-mean-square error of the predicted distribution of distances between hydroxyl cut sites marking neighboring nucleosomes. r_{osc} is the linear correlation between the oscillatory parts of the measured and predicted distributions of DNA fragment lengths. The oscillatory part was obtained by subtracting the smooth background from the full distribution. Smoothing was done by applying a Savitzky-Golay smoothing filter [also known as least-squares, or DISPO (Digital Smoothing Polynomial) filter] of polynomial order 3 and length 31 bp. RMS_{osc} is the root-mean-square error of the oscillatory part of the predicted distribution of distances between hydroxyl cut sites.

Model B: Crystal structure augmented with a linear function. Same as Model A for $x \in [1, 73]$, followed by a linear function:

$$u_{\text{half}}(x) = \begin{cases} \texttt{interp1}(...) - \frac{E_b}{147} x & \text{for } x \in [1, 73], \\ \texttt{interp1}(...) - 73\frac{E_b}{147} - \frac{\Delta E}{\Delta X}(x - 73) & \text{for } x \in [74, 73 + \Delta X], \end{cases}$$

where $\Delta E/\Delta X$ is the slope of the linear function (i.e., ΔE is the energy difference between the first and last points of the linear function and ΔX is the cardinality of the range of the linear function).

Parameter	Value
$a_{\min}$	27 bp
E_b	14.66 k_BT
μ	-15.04 k_BT
A	1.28 k_BT
f	0.50
ΔE	-2.47 k_BT
ΔX	7 bp

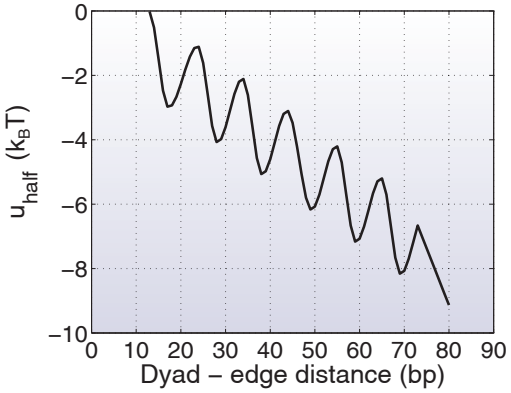

Table S2: Fitted parameters for model B. All parameters are as in Model A, except for ΔE and ΔX which are defined above.

Fit residuals:

$RMS = 0.0012$ (RMS cannot decrease below 10^{-3} for this model)
$r_{\text{osc}} = 0.769$
$RMS_{\text{osc}} = 1.841 \times 10^{-4}$

All residuals are defined as in Model A.

Model C: 10-bp oscillations superimposed onto a linear function.

$$u_{\text{half}}(x) = -A\cos\left(\frac{2\pi}{10}(x - x_0)\right) - \frac{E_b}{147}x,$$

where A is the amplitude of the oscillations, x_0 determines the phase of the oscillations, and E_b is the binding energy of a fully wrapped particle in the absence of the oscillations.

Parameter	Value
$a_{\max}$	165 bp
$a_{\min}$	3 bp
E_b	14.43 k_BT
μ	-13.99 k_BT
A	1.06 k_BT
x_0	79 bp
f	0.50

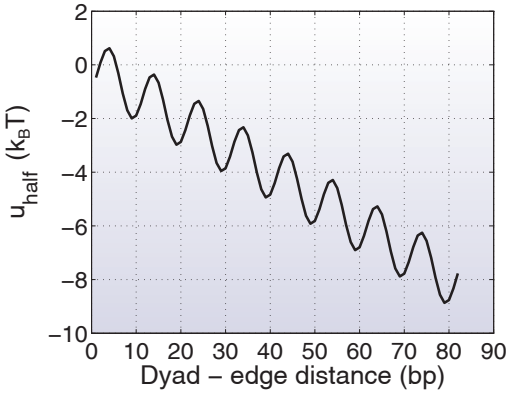

Table S3: Fitted parameters for Model C. All parameters are as in Model A, except for x_0 which is defined above.

Fit residuals:

$$RMS = 9.986 \times 10^{-4}$$
$$r_{\text{osc}} = 0.709$$
$$RMS_{\text{osc}} = 2.020 \times 10^{-4}$$

All residuals are defined as in Model A.

Model D: 11-bp oscillations superimposed onto a linear function.

$$u_{\mathrm{half}}(x) = -A\cos\left(\frac{2\pi}{11}(x - x_0)\right) - \frac{E_b}{147}x,$$

where A, x_0 and E_b have the same meaning as in Model C.

Parameter	Value
$a_{\max}$	161 bp
$a_{\min}$	25 bp
E_b	13.99 k_BT
μ	-14.30 k_BT
A	1.03 k_BT
x_0	80 bp
f	0.52

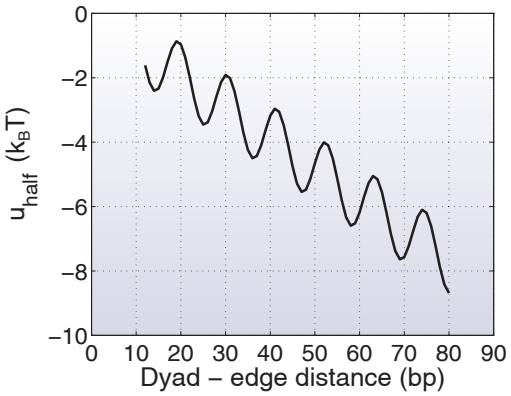

Table S4: Fitted parameters for Model D. All parameters are as in Models A and C.

Fit residuals:

$$RMS = 9.912 \times 10^{-4}$$
$$r_{\mathrm{osc}} = 0.689$$
$$RMS_{\mathrm{osc}} = 2.074 \times 10^{-4}$$

All residuals are defined as in Model A.

Model E: Uniform energy profile.

$$u_{\text{half}}(x) = -\frac{E_b}{147}x,$$

where E_b is the binding energy of a fully wrapped nucleosome.

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	35 bp
E_b	13.40 k_BT
μ	-13.14 k_BT
f	0.58

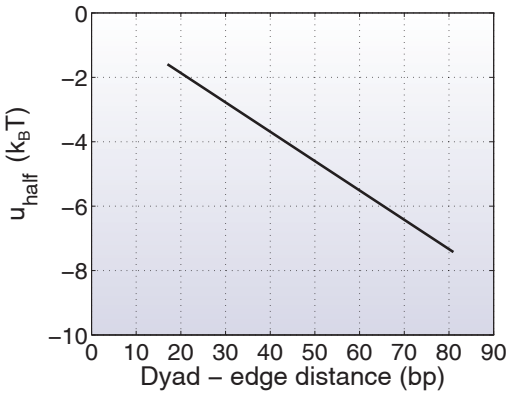

Table S5: Fitted parameters for Model E. All parameters are as in Model A.

Fit residuals:

$$RMS = 9.973 \times 10^{-4}$$
$$r_{\text{osc}} = 0.275$$
$$RMS_{\text{osc}} = 2.751 \times 10^{-4}$$

All residuals are defined as in Model A.

Model F: 5-bp oscillations superimposed onto a linear function.

$$u_{\mathrm{half}}(x) = -A\cos\left(\frac{2\pi}{5}(x - x_0)\right) - \frac{E_b}{147}x,$$

where A, x_0 and E_b have the same meaning as in Model C.

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	39 bp
E_b	13.50 k_BT
μ	-16.13 k_BT
A	2.36 k_BT
x_0	74 bp
f	0.63

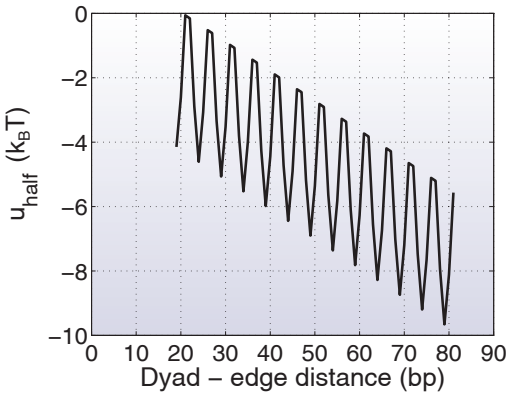

Table S6: Fitted parameters for Model F. All parameters are as in Models A and C.

Fit residuals:

$$RMS = 9.898 \times 10^{-4}$$
$$r_{\mathrm{osc}} = 0.206$$
$$RMS_{\mathrm{osc}} = 3.055 \times 10^{-4}$$

All residuals are defined as in Model A.

Model G: 5-bp stepwise energy profile.

$$u_{\mathrm{half}}(x) = -E_{\mathrm{step}}\ \texttt{ceil}\left(\frac{x - x_0}{5}\right),$$

where E_{step} is the amount of energy lost in each step, and x_0 determines the phase of the stepwise profile.

Parameter	Value
$a_{\max}$	163 bp
$a_{\min}$	39 bp
E_{step}	0.48 k_BT
μ	-12.83 k_BT
x_0	2 bp
f	0.63

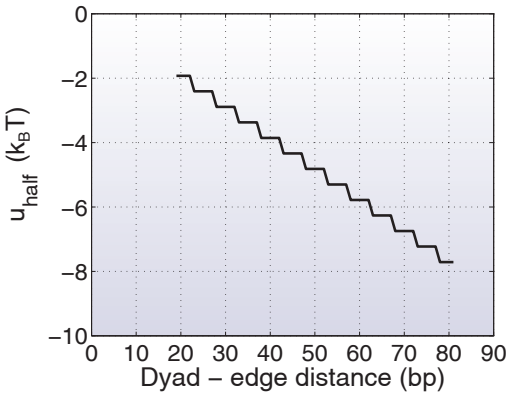

Table S7: Fitted parameters for Model G. All parameters are as in Model A, except for E_{step} and x_0 defined above.

Fit residuals:

$$RMS = 9.904 \times 10^{-4}$$
$$r_{\mathrm{osc}} = 0.283$$
$$RMS_{\mathrm{osc}} = 2.742 \times 10^{-4}$$

All residuals are defined as in Model A.

Model H: 10-bp stepwise energy profile.

$$u_{\mathrm{half}}(x) = -E_{\mathrm{step}}\ \texttt{ceil}\left(\frac{x - x_0}{10}\right),$$

where E_{step} is the amount of energy lost in each step, and x_0 determines the phase of the stepwise profile.

Parameter	Value
$a_{\max}$	169 bp
$a_{\min}$	3 bp
E_{step}	1.16 k_BT
μ	-12.04 k_BT
x_0	3 bp
f	0.62

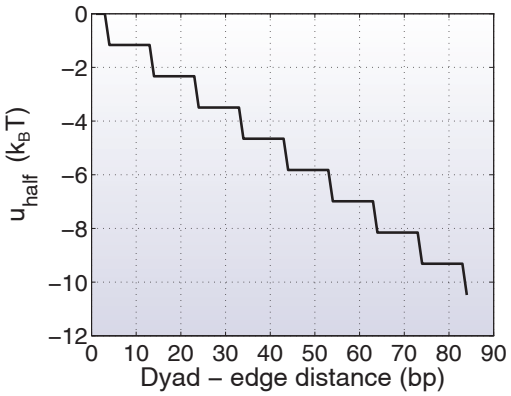

Table S8: Fitted parameters for Model E. All parameters are as in Models A and F.

Fit residuals:

$$RMS = 9.790 \times 10^{-4}$$
$$r_{\mathrm{osc}} = 0.545$$
$$RMS_{\mathrm{osc}} = 2.457 \times 10^{-4}$$

All residuals are defined as in Model A.

Supplementary References

[1] Chereji RV, Tolkunov D, Locke G, Morozov AV (2011) Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions. *Phys. Rev. E* 83:050903.

[2] Frisch HL, Lebowitz JL (1964) *The Equilibrium Theory of Classical Fluids: A Lecture Note.* (Benjamin).

[3] Locke G, Tolkunov D, Moqtaderi Z, Struhl K, Morozov AV (2010) High-throughput sequencing reveals a simple model of nucleosome energetics. *Proc. Natl. Acad. Sci. USA* 107:20998–21003.

[4] Brogaard K, Xi L, Wang JP, Widom J (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature* 486:496–501.

[5] Flaus A, Luger K, Tan S, Richmond TJ (1996) Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals. *Proc. Natl. Acad. Sci. U.S.A.* 93:1370–1375.

[6] Kaplan N et al. (2008) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458:362–366.

[7] Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S (2011) Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. USA* 108:18318–18323.

[8] Cole HA, Howard BH, Clark DJ (2011) The centromeric nucleosome of budding yeast is perfectly positioned and covers the entire centromere. *Proc. Natl. Acad. Sci. USA* 108:12687–12692.

[9] Nagarajavel V, Iben JR, Howard BH, Maraia RJ, Clark DJ (2013) Global 'bootprinting' reveals the elastic architecture of the yeast TFIIIB–TFIIIC transcription complex in vivo. *Nucleic Acids Res.* 41:8135–8143.

[10] Richmond TJ, Davey CA (2003) The structure of DNA in the nucleosome core. *Nature* 423:145–150.

[11] Dion MF et al. (2007) Dynamics of replication-independent histone turnover in budding yeast. *Science* 315:1405–1408.

[12] Rhee HS, Pugh BF (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* 483:295–301.

3 Supplementary Figures

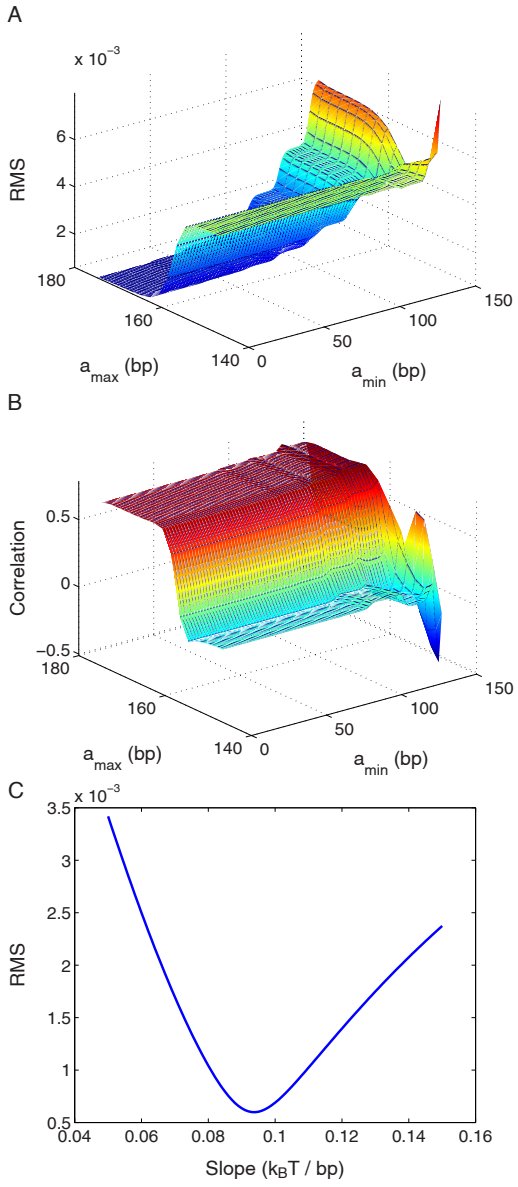

Figure S3: **Sensitivity of the predicted distribution of distances between hydroxyl cut sites marking neighboring nucleosomes to the parameters of the histone-DNA binding energy profile based on nucleosome crystal structures.** (A) Root-mean-square error (RMS) of the distribution of DNA fragment lengths predicted using the model in Fig. 1B, as a function of $a_{\min}$ and $a_{\max}$. (B) The linear correlation coefficient between oscillations in the predicted and observed distributions ($r_{\rm osc}$), as a function of $a_{\min}$ and $a_{\max}$. The oscillations were obtained by subtracting the smooth background from the distributions of DNA fragment lengths, as described in the Fig. 1 caption. (C) Variation of the RMS with the slope of the energy profile in Fig. 1B. In all panels, model parameters that were not varied were kept fixed at their best-fit values (SI Results, Model A).

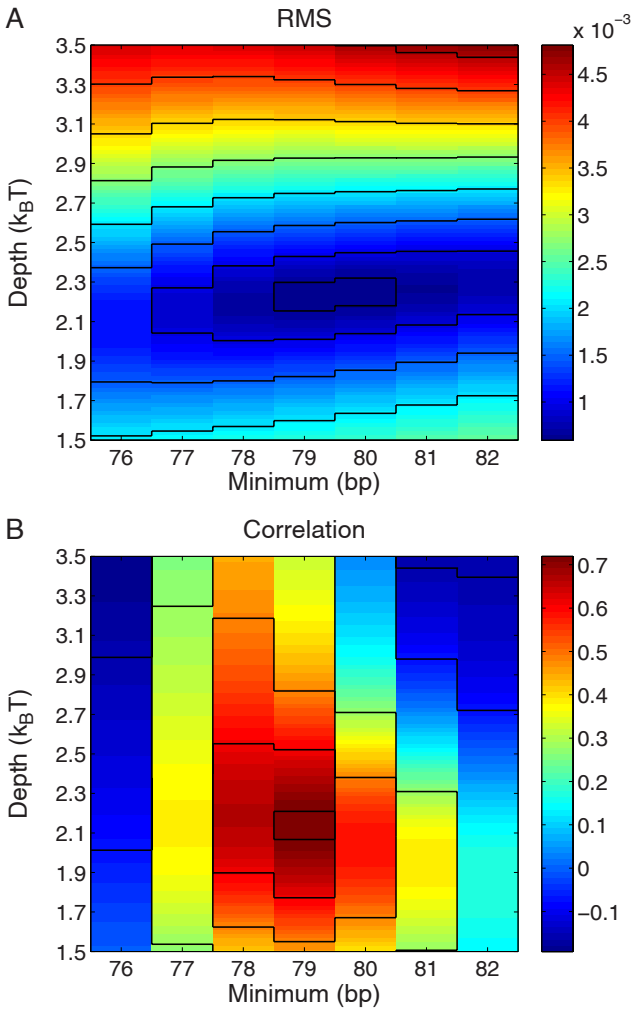

Figure S4: **Sensitivity of the predicted distribution of distances between hydroxyl cut sites marking neighboring nucleosomes to model parameters describing higher-order chromatin structure.** The histone-DNA binding energy profile is based on nucleosome crystal structures (SI Results, Model A). (A) Root-mean-square error (RMS) of the predicted distribution of DNA fragment length, as a function of the position and the depth of the first minimum outside of the nucleosome core (Fig. 1B). The depth of the first minimum is computed with respect to $u_{\rm half}(x = 73$ bp). (B) The linear correlation coefficient between oscillations in the predicted and observed distributions of DNA fragment lengths ($r_{\rm osc}$), as a function of the position and the depth of the first minimum outside of the nucleosome core (Fig. 1B). The oscillations were obtained by subtracting the smooth background from the distributions of DNA fragment lengths, as described in the Fig. 1 caption. In both panels, model parameters that were not varied were kept fixed at their best-fit values (SI Results, Model A).

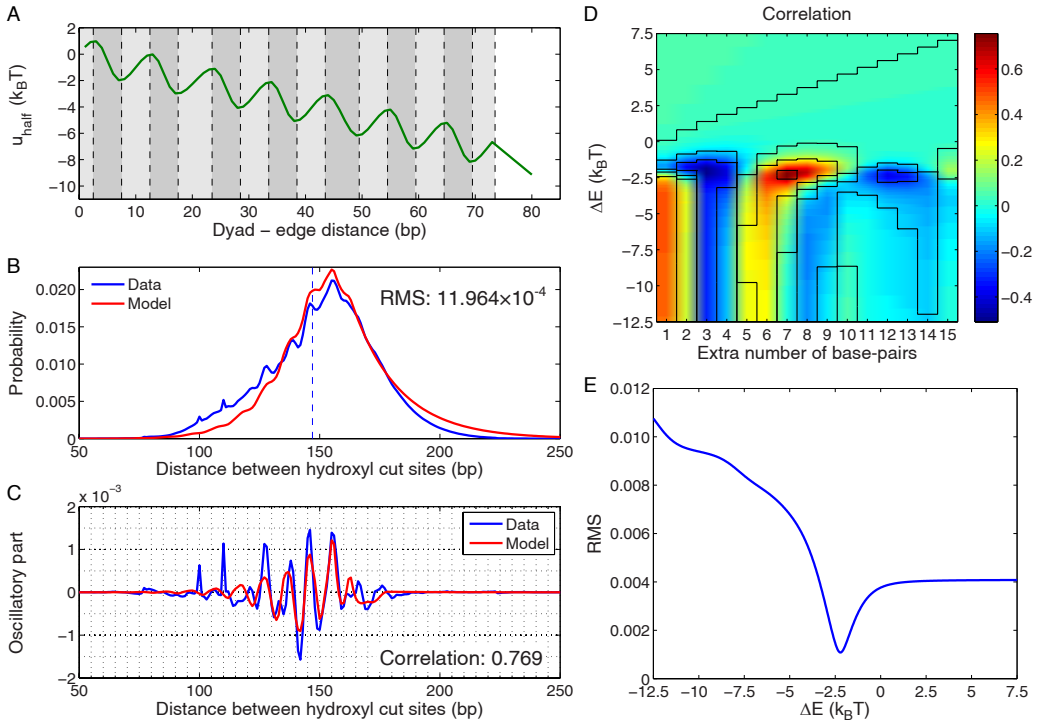

Figure S5: **Crystal structure-based model augmented by a linear potential outside of the nucleosome core.** (A) The energy profile fitted to reproduce the distribution of DNA fragment lengths shown in (B). All fitting parameters are listed in SI Results, Model B. If the DNA is symmetrically wrapped with respect to the nucleosome dyad, the energy of a nucleosome which covers $2x+1$ bps is given by $2u_{\rm half}(x)$. (B) The distribution of distances between hydroxyl cut sites marking neighboring nucleosomes observed in a high-resolution nucleosome map [4] (blue line), and predicted using Model B in SI Results (red line). RMS - root-mean-square deviation between the model and the data. Note that in this model RMS below 10^{-3} could not be achieved, and thus optimization was switched to maximize the correlation coefficient $r_{\rm osc}$ once RMS reached 1.2×10^{-3} (see SI Methods for details). (C) Oscillations in the observed (blue line) and predicted (red line) distributions of distances between hydroxyl cut sites. The oscillations were obtained by subtracting the smooth background from the data and the model in (B), as described in the Fig. 1 caption. Correlation refers to $r_{\rm osc}$, the linear correlation coefficient between measured and predicted oscillations. (D) Heatmap with superimposed contour lines of the $r_{\rm osc}$ dependence on the two parameters of the linear potential outside of the nucleosome core: $\Delta x = x_{\rm last} - 73$ bp and $\Delta E = u_{\rm half}(x_{\rm last}) - u_{\rm half}(73)$, where $[1, x_{\rm last}]$ is the range of the energy profile (SI Results, Model B). Note that the best fit corresponds to $\Delta x = 7$ bp. (E) The dependence of the RMS on ΔE for the best-fit value of $\Delta x = 7$ bp. All parameters not explicitly varied in (D) and (E) were kept fixed at their best-fit values (SI Results, Model B).

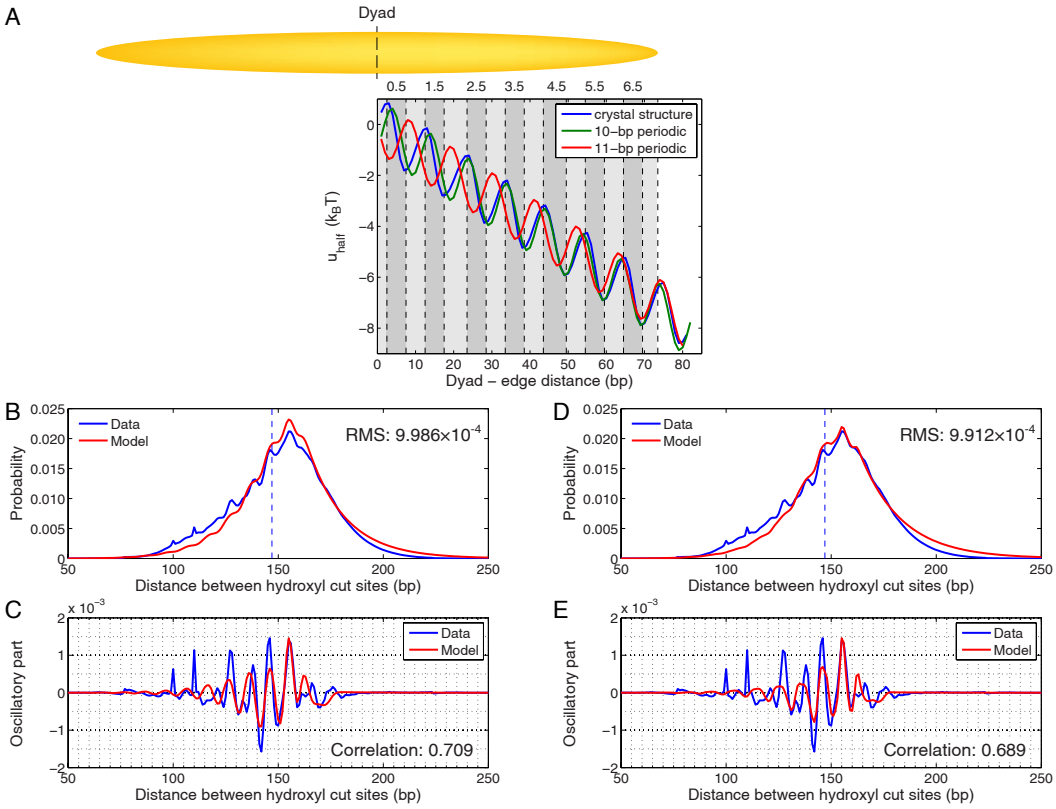

Figure S6: **Strictly periodic models of histone-DNA binding and effects of higher-order structure.** (A) If the DNA is symmetrically wrapped with respect to the nucleosome dyad, the energy of a nucleosome that covers $2x+1$ bps is given by $2u_{\mathrm{half}}(x)$. The minima and maxima of the energy landscape are either based on the crystal structures of the nucleosome core particle as in Figure 1 (blue), or else are 10 (green) and 11 (red) bp-periodic oscillations with fitted initial phase (SI Results, Models C and D). Dark gray bars show where the histone binding motifs interact with the DNA minor groove. Light gray bars indicate where the DNA major groove faces the histones. (B) The distribution of distances between hydroxyl cut sites marking neighboring nucleosomes, from a high-resolution nucleosome map [4] (blue line), and from the 10 bp-periodic model (red line). All model parameters are listed in SI Results, Model C. RMS - root-mean-square deviation between the model and the data. (C) Oscillations in the observed (blue line) and predicted (red line) distributions of distances between hydroxyl cut sites. The oscillations were obtained by subtracting a smooth background from the data and the model in (B), as described in the caption of Fig. 1. Correlation refers to r_{osc}, the linear correlation coefficient between measured and predicted oscillations. (D) Same as (B), for the 11 bp-periodic model. All model parameters are listed in SI Results, Model D. (E) Same as (C), for the 11 bp-periodic model.

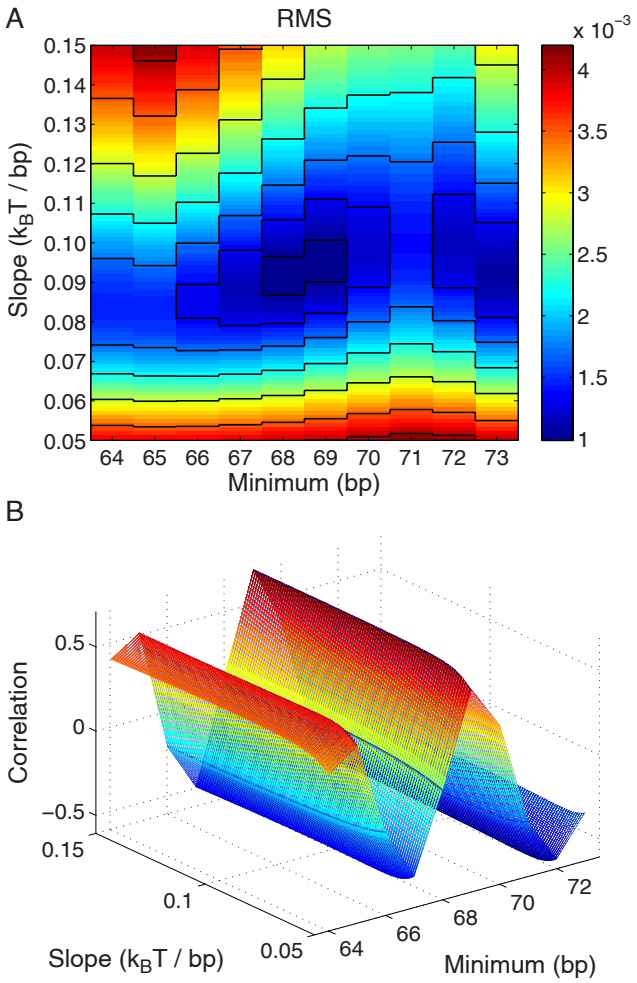

Figure S7: **Sensitivity of the predicted distribution of distances between hydroxyl cut sites marking neighboring nucleosomes to the parameters of the 10 bp-periodic model.** (A) Heatmap with superimposed contour lines of the RMS dependence on the slope of the energy profile and the position of the last minimum within the nucleosome core. RMS - root-mean-square deviation between the model and the data. (B) The linear correlation coefficient r_{osc} between oscillations in the predicted and observed distributions of distances between hydroxyl cut sites, as a function of the overall slope of the energy profile and the position of the last minimum within the nucleosome core. All parameters not explicitly varied were kept fixed at their best-fit values (SI Results, Model C).

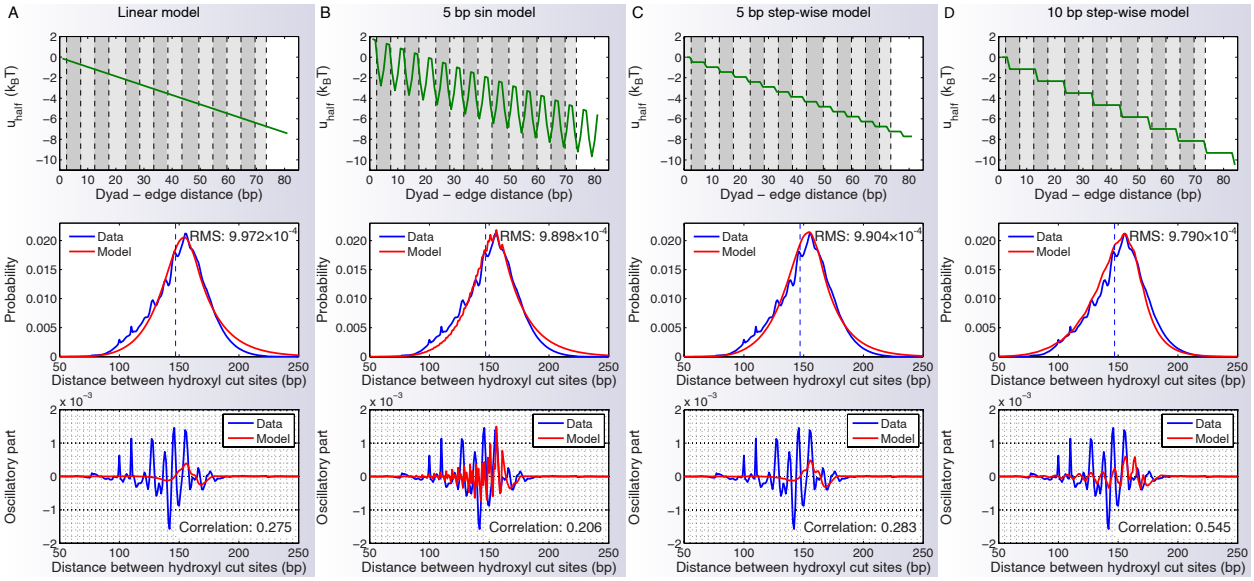

Figure S8: **Alternative models of histone-DNA binding energies and effects of higher-order structure.** (A) Linear model (SI Results, Model E). (B) 5-bp periodic model (SI Results, Model F). (C) 5-bp step-wise model (SI Results, Model G). (D) 10-bp step-wise model (SI Results, Model H). In each column, the upper panel shows the energy profile (as in Fig. 1B), the middle panel shows the comparison of experimental and predicted distance distributions (as in Fig. 1C), and the lower panel shows observed and predicted oscillations in the distribution of distances between hydroxyl cut sites marking neighboring nucleosomes (as in Fig. 1D).

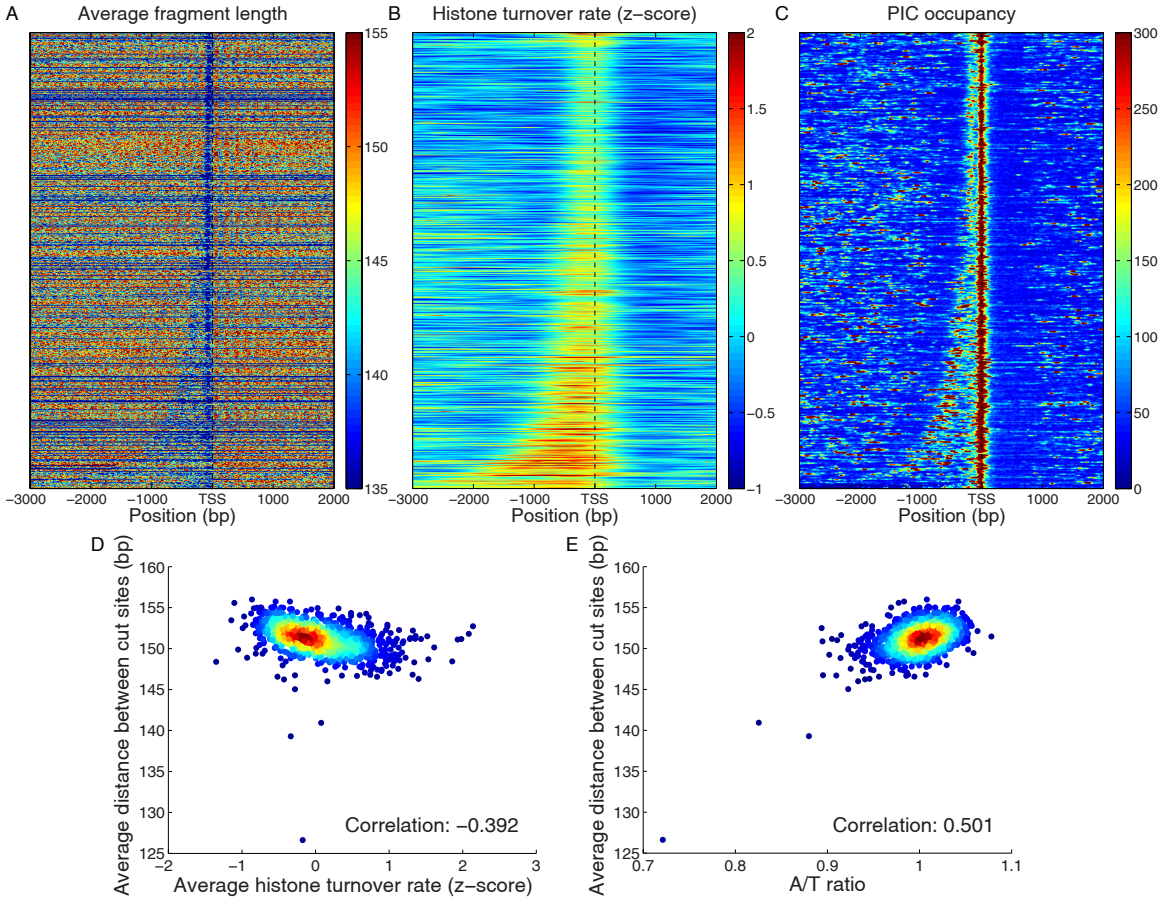

Figure S9: **Genome-wide distributions of nucleosome lengths, inter-dyad distances, histone turnover rates, and transcription pre-initiation complexes.** (A) Distribution of average lengths of DNA-bound particles mapped by MNase digestion [7] in the vicinity of TSS. We considered particles with sizes between 80 and 200 bp and assigned particle lengths to the mid-point of each particle. Values for bps without dyads were obtained by interpolation. (B) Distribution of histone turnover rates [11] in the vicinity of TSS. (C) Distribution of the combined occupancy of 9 transcription pre-initiation complexes (PICs) [12] in the vicinity of TSS. PIC occupancies provided at 20 bp interval in Ref. [12] were interpolated. In panels (A)-(C), gene order is as in Figure 1B, and the heatmaps were smoothed using a 2D Gaussian kernel with $\sigma = 3$ pixels. (D) Correlation of DNA fragment lengths between hydroxyl cut sites corresponding to neighboring nucleosomes, and histone turnover rates. Both quantities are averaged over 10 kbp windows tiling the yeast genome. (E) Correlation of average DNA fragment lengths between hydroxyl cut sites corresponding to neighboring nucleosomes, and the A/T ratio in 10 kbp windows tiling the yeast genome. The A/T ratio is the fraction of A/T nucleotides in the 10 kbp window, divided by the genome-wide A/T fraction. Correlation in (D) and (E) refers to the linear correlation coefficient.

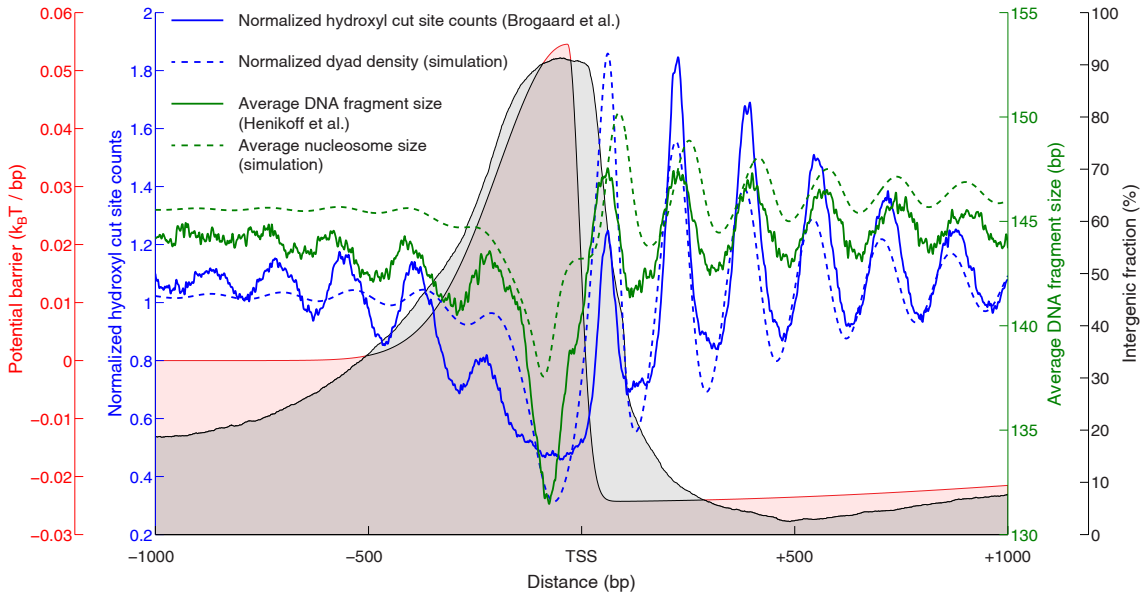

Figure S10: **Modeling distributions of nucleosome lengths and dyad positions in the vicinity of TSS.** We align all yeast genes by their TSS as in Fig. 2 and for each bp compute the fraction of times that bp is found in an intergenic region, as opposed to the ORF of a neighboring gene (grey background curve with black border). The intergenic fraction has an asymmetric shape with a maximum at about 50 bp upstream of the TSS. We use this shape as a guide for constructing an energy barrier for *in vivo* histone deposition (pink background curve with red border). The barrier is composed of three half-Gaussians: $B(x) = H\exp[-\frac{(x-c)^2}{2\sigma_1^2}] \quad (x \le c), \quad (H+D)(\exp[-\frac{(x-c)^2}{2\sigma_2^2}] - 1) - D\exp[-\frac{(x-c)^2}{2\sigma_3^2}] \quad (x > c)$. The free parameters of the barrier are fit to maximize the sum of two correlations: between observed and predicted normalized dyad counts [4] (solid and dashed blue lines, respectively), and between observed and predicted average nucleosome DNA lengths [7] (solid and dashed green lines, respectively). Normalized dyad counts (approximated by the positions of hydroxyl radical cut sites) are computed as the total number of dyads at a given bp for all genes, divided by the average of this quantity in a [-1000,1000] bp window around the TSS. Average DNA lengths are computed for all nucleosomes with a midpoint at a given bp, for all genes (if the midpoint falls in between two bps, the one on the left is used). The fitted parameters are: $H = 0.0545$ k$_B$T, $D = 0.0243$ k$_B$T, $c = x_{\mathrm{TSS}} - 32$ bp, $\sigma_1 = 162.7$ bp, $\sigma_2 = 28.0$ bp, $\sigma_3 = 2090.9$ bp, where x_{TSS} is the absolute position of the TSS in the box, c is the center of the 3 Gaussians, H is the height of the first Gaussian, D is the depth of the third Gaussian, and $\sigma_1, \sigma_2, \sigma_3$ are the standard deviations of the three Gaussian distributions. The simulations were done in a 15 kbp box with the barrier placed at its center to eliminate the boundary effects. DNA was assumed to be wrapped symmetrically with respect to the dyad, and the nucleosome structure-based energy profile (SI Results, Model A) was used. The total free energy $u_{\mathrm{nuc}}(k,l)$ of a nucleosome occupying bps $k,\ldots,l$ is a sum of $u_{\mathrm{nuc}}^{\mathrm{SI}}$ and $u_{\mathrm{nuc}}^{\mathrm{barrier}} = \sum_{j=k}^{l} \epsilon_j$, where ϵ_j is the value of the barrier at bp j. Note that the chemical mapping data underestimates the number of -1 and +1 nucleosomes due to gel selection bias [4], and that on average hydroxyl cut sites coincide with dyad positions (cf. Section 1.6).

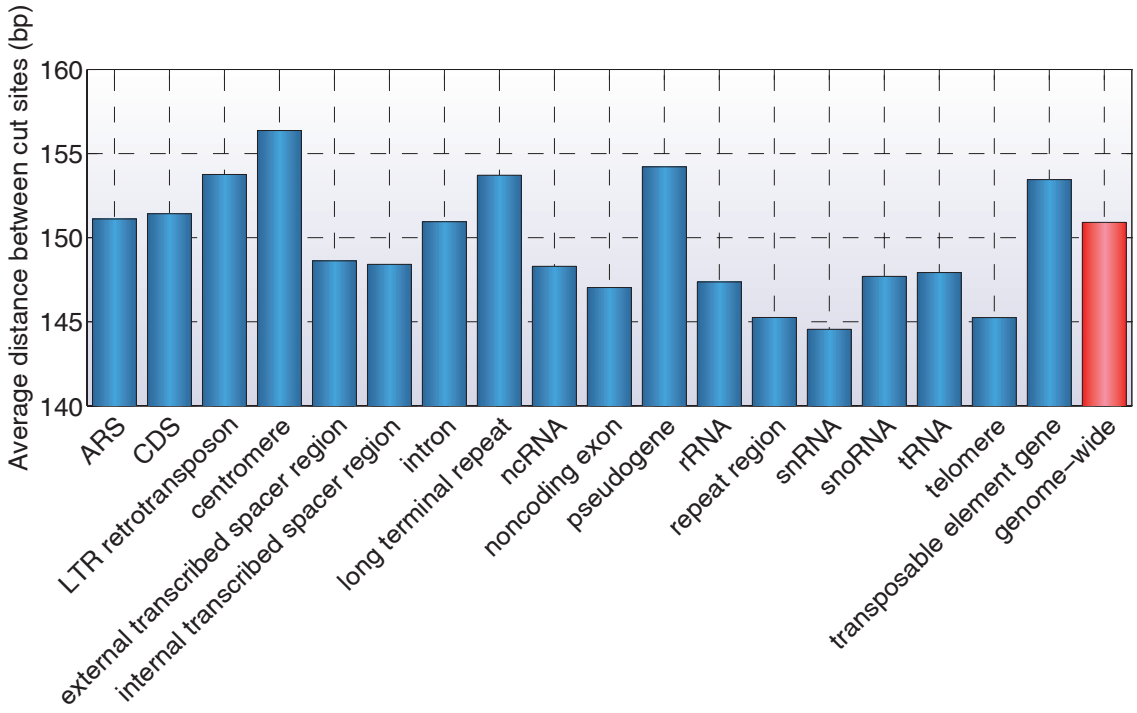

Figure S11: **Distances between hydroxyl cut sites corresponding to neighboring nucleosomes averaged over genomic functional elements.** ARS - autonomously replicating sequence, CDS - coding DNA sequence, LTR retrotransposon - long terminal repeat retrotransposon, ncRNA - non-coding RNA, rRNA - ribosomal RNA, snRNA - small nuclear RNA, snoRNA - small nucleolar RNA, tRNA - transfer RNA. Note that according to Section 1.6 and using $f = 0.5$ for the cut frequency (a value typical of those found in our fits, cf. SI Results), the average distance between cut sites shown in the bar plot above is 5 bp less than the average inter-dyad distance.

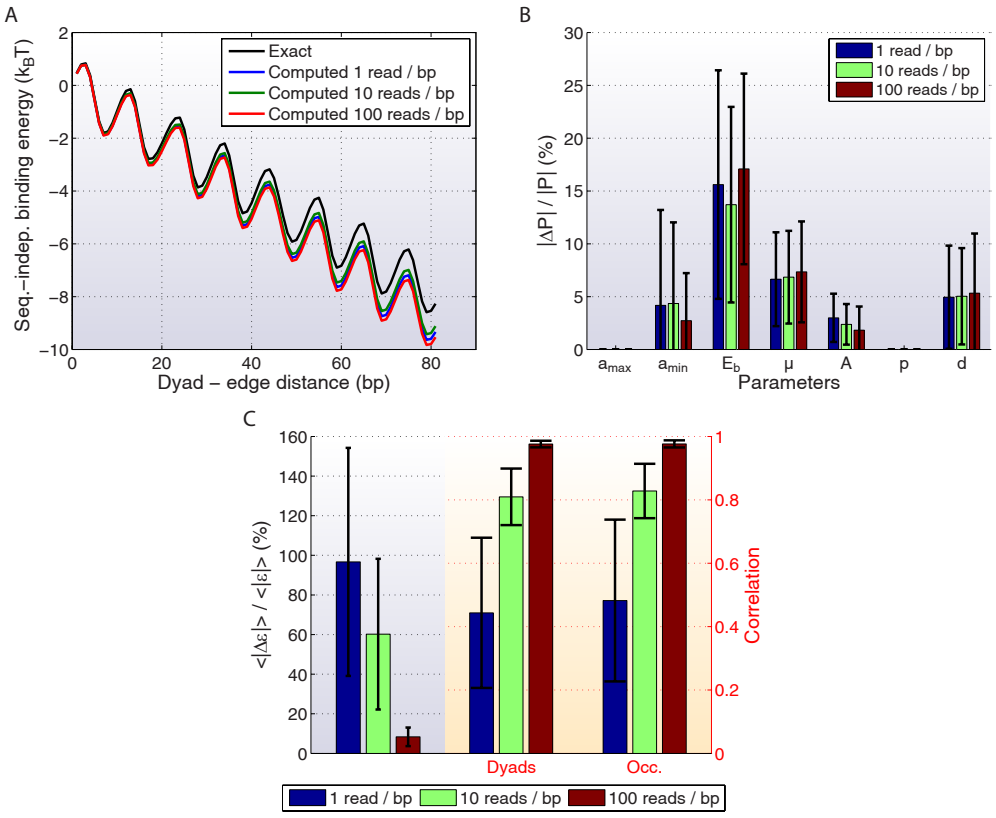

Figure S12: **Inference of the histone-DNA binding/higher-order structure energy profile and sequence-specific nucleosome formation in a model system.** We assume that the sequence-dependent correction to the average free energy of nucleosome formation, $u^{\mathrm{SD}}_{S(N)}$, is given by the energy parameters inferred from the Henikoff et al. dataset [7] (Table S9), and the sequence-independent contribution, u^{SI}_N, is defined as in SI Results, Model A. Using Eq. [1] in Materials and Methods with $\mu = -13$ $\mathrm{k_B T}$, we compute the exact nucleosome distribution $n^{\mathrm{nuc}}_1(k,l)$ for the *S.cerevisiae* chromosome I with $L = 230,218$ bp. We sample paired-end nucleosomal reads $[k,l]$ from $n^{\mathrm{nuc}}_1(k,l)$ until a desired level of read coverage is reached. In total, $M \times L$ reads ($M \in \{1, 10, 100\}$ is the level of read coverage per bp) are randomly sampled from the exact distribution n^{nuc}_1. From this finite sample, we construct a chromosome-wide histogram of nucleosome DNA lengths $P(N)$, and fit the parameters of u^{SI}_N by using a genetic algorithm optimization function `ga` from the MATLAB Global Optimization toolbox to minimize the root-mean-square error between the predicted distribution of nucleosome lengths generated with u^{SI}_N, and $P(N)$. Next, we use Eq. [4] in Materials and Methods to infer the total free energy of nucleosome formation $u_{S(N)}$ from the same sample, and fit the sequence-dependent model to $u_{S(N)} - u^{\mathrm{SI}}_N$, assuming that the dyad is at the mid-point of each particle. Finally, using the sum of predicted u^{SI}_N and $u^{\mathrm{SD}}_{S(N)}$, we compute the approximate nucleosome distribution $\tilde{n}^{\mathrm{nuc}}_1$, which is then compared with the exact profile n^{nuc}_1. The difference between $\tilde{n}^{\mathrm{nuc}}_1$ and n^{nuc}_1 is due to limited sampling of sequence reads. (A) Exact energy profile vs. profiles predicted at three levels of read coverage. Note that the overall slope of the potential is slightly overpredicted, likely because the histogram of particle lengths is affected by well-positioned nucleosomes with negative formation energies. The average of these energies may bias the slope. (B) Relative errors between predicted and exact (SI Results, Model A) parameters of the energy profile, and predicted and exact chemical potential μ. P denotes any of the parameters on the horizontal axis. (C) Relative errors between predicted and exact energy parameters (Table S9, Henikoff et al. dataset [7]) (light blue background). Linear correlations between predicted and exact distributions of dyad positions and nucleosome occupancy (light pink background). The height of each bar in (B) and (C) represents the mean relative error for the corresponding parameter or the mean correlation coefficient, averaged over 100 random sampling simulations. The uncertainty intervals represent standard deviations.

4 Supplementary Tables

Table S9: **Table of sequence-dependent energy parameters inferred from large-scale maps of *S.cerevisiae* nucleosomes obtained by MNase digestion and high-throughput sequencing.** Energy parameters were predicted using three *in vivo* nucleosome maps based on paired-end reads (Henikoff et al. [7], Nagarajavel et al. [9], and Cole et al. [8]), and one *in vitro* nucleosome map based on single-end reads (Kaplan et al. [6]) (SI Results, Section 1.7). In the Kaplan et al. map, each single-end sequence read was extended to the canonical nucleosome length of 147 bp. The dyad positions are assumed to be at the midpoint of each genomic DNA fragment defined by either a mate pair or a single 147 bp-long read (if the midpoint falls in between two bps, the one on the left is used). For each nucleosome map, we obtain an estimate of the nucleosome distribution $n_1^{\mathrm{nuc}}(i,j)$ by normalizing raw read counts (the number of nucleosomes of any length that start at a given bp) so that the maximum nucleosome occupancy is 1.0 for each chromosome. We compute the total nucleosome formation energy $u_{\mathrm{nuc}}(i,j)$ from $n_1^{\mathrm{nuc}}(i,j)$ using Eq. (22) (all bps i for which $n_1^{\mathrm{nuc}}(i,j) = 0$ are excluded), subtract the sequence-independent part $u_{\mathrm{nuc}}^{\mathrm{SI}}(i,j)$ predicted using Brogaard et al. data [4] (SI Results, Model A) from it, and fit the sequence-dependent model $u_{\mathrm{nuc}}^{\mathrm{SD}}(i,j)$ to the resulting difference as described in SI Results, Sections 1.4 and 1.5. N_{rbp}: number of reads per bp in each dataset.

	Kaplan et al.	Henikoff et al.	Nagarajavel et al.	Cole et al.
$\epsilon_{A/T}$	-0.180	-0.081	-0.200	-0.195
$\epsilon_{C/G}$	0.290	0.130	0.322	0.314
$\epsilon_{AA/TT}$	0.221	0.069	0.210	0.210
$\epsilon_{AC/GT}$	-0.076	-0.031	-0.055	0.017
$\epsilon_{AG/CT}$	-0.068	-0.010	-0.123	-0.130
$\epsilon_{AT/AT}$	0.201	0.089	0.141	0.198
$\epsilon_{CA/TG}$	-0.092	-0.003	-0.085	-0.166
$\epsilon_{CC/GG}$	-0.305	-0.138	-0.324	-0.306
$\epsilon_{CG/CG}$	-0.319	-0.169	-0.396	-0.462
$\epsilon_{GA/TC}$	-0.054	-0.023	0.013	0.016
$\epsilon_{GC/GC}$	-0.315	-0.162	-0.253	-0.148
$\epsilon_{TA/TA}$	0.189	0.090	0.246	0.173
N_{rbp}	1.02	5.53	0.81	2.71