Supporting Information

Minary and Levitt 10.1073/pnas.1404475111

SI Materials and Methods

Generating ab Initio Nucleosome Energy or Occupancy Profiles. Given a genomic query sequence Q of length L_Q and a structural template T which consists of L_T base pairs ($L_T = 147$), we proceed as follows:

i) Generate the set $\Theta = \{S_1, S_2, \ldots, S_N\}$ of $N = L_Q - L_T + 1$ local sequences, where S_n is the continuous subsequence of Q from position n to position $(n + L_T - 1)$.

ii) Generate the set $\Omega = \{T_1, T_2, \ldots, T_N\}$ of query structures, where T_n is obtained by threading S_n onto the initial DNA template (termed T_0) using the algorithm defined in *Materials and Methods.*

iii) Cytosines may be methylated by conversion to 5-methyl-cytosine (5Me-C) on both template structures: the nucleosomal DNA [Protein Data Bank (PDB) ID code 1kx5] and linear B-DNA.

iv) The energy score E_n is calculated for each sequence on its template. First, each structure in Ω is minimized (conjugate gradient) (1) along Cartesian degrees of freedom to idealize local stereochemistry. This procedure is followed by conformational optimization using a maximal number of 10,000 natural move Monte Carlo (2) steps with temperature dropping from 500 to 5 K at an annealing rate of 0.996. Our rapid and local optimization pipeline preserves the overall shape of backbone without any restraint holding the DNA in a circle. All calculations have been performed using the Assisted Model Building with Energy Refinement (AMBER)99-bsc0 (3) force field with an implicit electrostatic solvent description (4), which uses a distance dependent dielectric of the form $\varepsilon(r) = D - \frac{D - D_0}{2}\left[(rS)^2 + 2rS + 2\right]\exp(-rS)$ with values $D_0 = 4$, $D = 80$, and $S = 0.4$ Å^{-1}. All interactions are subject to a cutoff, $r_{\mathrm{cut}} = 10.0$ Å and an adiabatic fourth order analytical switching function (5), $C_4(t) = \tau^5(70\tau^4 - 315\tau^3 + 540\tau^2 - 420\tau + 126)$ with $\tau = (r_{cut} - r)/\alpha$ and $\alpha = 1.0$ Å that smoothly turns off interactions in the interval $[r_{cut} - \alpha, r_{cut}]$. Furthermore, an implicit counterion effect is modeled by scaling the partial negative charge of the two most electronegative atoms (O1P and O2P) of the backbone so that each residue is neutralized; this scaling facilitates rapid optimization. The raw energy profile (the values of E_n for all n) is smoothed with a filter width of 147 bp (length of DNA on the nucleosome) using the direct form, transposed-implementation of the standard difference equation (6). All energies are reported in units of kilocalories per mole.

v) The raw energy score per base pair, $E(n) = E_n/L_T$ may be converted to a probability of occupancy $P(n) = (1/Z)\exp[-\beta E(n)]$ at temperature $1/\beta$ of 300 K, where Z is the Boltzmann normalization factor or partition function, $Z = \Sigma \exp(-\beta E(n))$. The raw probability profile [the values of $P(n)$ for all n] is smoothed using the procedure outlined in the previous step.

Our default initial template, T_0, is the nucleosomal crystal structure (PDB ID code 1kx5 from ref. 7). Further control calculations also use an ideal DNA superhelix with 80 bp per turn and a rise of 20.0 Å (8). To obtain the nucleosome formation energy, separate calculations are performed using as template an ideally straight right-handed B-DNA with 10.5 bp per turn. We refer to these three template structures as "1kx5," "ideal," and "linear," respectively. In all calculations we model all atoms including hydrogen atoms explicitly.

Fig. S1 illustrates and summarizes the various steps in our computational approach, which was performed using the software package Methodologies for Optimization and Sampling In Computational Studies (MOSAICS) (9) and associated scripts. The reproducibility of our calculations is facilitated by the online tutorial that is available at www.cs.ox.ac.uk/mosaics/nucleosome/nucleosome.html.

Generating Sequence Motif Distributions for Local Nucleosome Sequences. To generate the probability of occurrence of motifs in local sequences, S (*Generating ab Initio Nucleosome Energy or Occupancy Profiles*), along a query genomic sequence, Q, our algorithm works as follows:

i) Given Q, generate a set of N local sequences, $\Theta = \{S_1, S_2, \ldots, S_N\}$, where S_n is the 147-nt continuous subsequence of Q starting at position n and representing a putative nucleosome position.

ii) For a particular sequence motif H, calculate $H(n)$, the number of times it appears in each local sequence S_n.

iii) The sequence motif probability distribution, $P_H(n)$ is defined as $P_H(n) = H(n)/Z$, where $Z = \Sigma H(n)$ is a normalization factor chosen to ensure that the probabilities sum to 1.

SI Results and Discussion

In Silico Prediction of Occupancy Profiles on the Genomic Scale. The method outlined in Fig. S1 was used for predicting the nucleosome occupancy profile along the 187,000 to 207,000-bp region of chromosome 14 in yeast (10, 11). This DNA sequence was chosen because its occupancy profile has been carefully investigated by experimental methods both in vitro and in vivo (11, 12). Fig. S2 compares the predicted nucleosome-positioning distribution (ab initio) along this 20,000-bp region with occupancy profiles obtained from both in vitro and in vivo experiments. These calculations were performed using the AMBER99-bsc0 force field (3), an implicit electrostatic solvent description (4), and a DNA template from the crystal structure (7) with PDB ID code 1kx5. The relatively simple description of the solvent is not only in good agreement with experimental data (4) but also needed here so as to realize the computational advantage of collective moves (2). Fig. S2 also presents the computed profile referred to as "ab initio$|_R$" that depends on nucleosome formation energy and has been calculated with respect to linear DNA (ideally straight right-handed B-DNA with 10.5 bp per turn). It is clear from Fig. S2 that both the ab initio and ab initio$|_R$ profiles capture many features of the in vitro profile whereas they have less resemblance to the in vivo profile.

Sequence Statistics, ab Initio, and Experimental Predictions. Fig. S4*A* compares the sequence position-dependent correlations between the in vitro occupancy and three other profiles based on ab initio prediction, Segal model (12) (in vitro reconstruction), and in vivo occupancies, respectively. Given that the Segal model is trained on the in vitro data, their very strong correlation [correlation coefficient (CC) = 0.8185] is expected, however it is interesting to notice that Segal model correlates even more with GC percentage (Fig. S4*B*). It is clear from the figure that correlation between the ab initio and in vitro (CC = 0.6155) is significantly stronger than correlation between the two experimental profiles, in vitro and in vivo (CC = 0.5065). Given, that the correlations between in vitro and in vivo are dramatically below the rest of the curves it suggest that the in vivo occupancy is affected by many additional factors only present in the cell.

Fig. S4*B* shows the correlations between GC percentage and ab initio nucleosome occupancies, GC percentage and the Segal model, and CG percentage and in vitro predictions as a function of sequence position. In general, the correlation of GC percentage with in vitro occupancy is similar to that with ab initio occupancy. Furthermore, Fig. S4*B* clearly identifies the correlation between the Segal model (12) and GC percentage as the strongest. This observation is also supported with the largest overall correlation coefficient $CC_{\text{GC-Segal}} = 0.8779$ compared with $CC_{\text{GC–ab initio}} = 0.7431$ and $CC_{\text{GC–in vitro}} = 0.6843$.

It is interesting to mention that the most widely used sequence knowledge-based model (12), which has been trained on the in vitro data, is more correlated with GC percentage (CC = 0.8779) than with experiment (CC = 0.8185) and the dominance of the GC effect was reported in several studies (13). Thus, as our calculations further verifies this effect, the in vitro experimental profile is primarily driven by the GC percentage to a strong extent signified by a correlation coefficient CC = 0.6843. To arrive to a final nucleosome occupancy profile, the basic GC effect is augmented with more delicate sequence- or structure-based factors, which may include the distribution of certain motifs along the sequence and/or positions with varying DNA groove width (12, 14–17).

Sensitivity to Counterion Treatment, Force Fields, Solvent, and Template Structures. Fig. S6 shows that our default implicit counterion treatment produces sufficiently identical results to the explicit case. In the explicit counterion treatment, a single sodium counterion with a positive charge and the standard atom type NA (sodium) of the AMBER-99bsc0 (3) force fields has been initially randomly placed close to each nucleotide along the backbone of the DNA superhelix. The position of each counterion is updated at each step of the simulated annealing Monte Carlo trajectory; a new position is drawn from a 3D normal distribution centered on the current position.

When predicting nucleosome occupancy along sequence regions, it is essential to understand how these calculations depend on the force fields. Here, we choose a short 2,000-bp sequence region (chromosome 14, 187,000–189,000) that has a characteristic occupancy profile according to in vitro experiments (12). Based on the nucleosome DNA crystal structure 1kx5 (7) (Fig. 1), we obtain nucleosome occupancy profiles (Fig. S2) for three different force fields—AMBER99 (18), AMBER99-bsc0 (3) (which is used as our standard), and Chemistry at Harvard Macromolecular Mechanics (CHARMM)27 (19)—and the same implicit description of solvent (4). It is clear from Fig. S7*A* that the main features of the in silico occupancy profile are not force-field dependent. Fig. S7*A* also indicates that the fit between the ab initio and in vitro profile depends on position along the DNA with predicted regions that are medium (187,000–187,500), good (187,500–188,500), and bad (188,500–189,000). In fact, the fit is unusually bad in the region 188,500–189,000 and Fig. 2*A* gives a position-dependent correlation coefficient of 0.0 between in vitro and predicted nucleosome occupancy at 189,000 (2,000-bp window from 188,000 to 190,000).

Fig. S7*B* compares nucleosome occupancy profiles computed in vacuo and the one obtained using our standard solvation model (4). It is clear that removing the effect of solvent has a significant impact on the computed occupancy profiles; sequence regions with low occupancy become less populated whereas regions with high occupancy become more populated. For example, of the three most characteristic peaks of the profile with solvent, only the most dominant is present in the profile without solvent (sequence position 188,250). It is also clear from Fig. S7*B* that removing the solvent effect can be compensated by an increase in temperature so that the solvent-free profile obtained for high temperatures resemble the standard 300-K profile with solvent effect. Given that our solvent treatment only affects the electrostatic terms in the energy function, this indicates that electrostatic contribution is the dominant energy term guiding nucleosome positioning.

Another concern about our ab initio protocol is the dependence on the initial structure of the DNA template. We investigate this phenomenon by generating profiles for two distinct structural templates: nucleosome crystal structure 1kx5, which is used as our standard, and an ideal DNA superhelix (8). Fig. S7*C* clearly shows that the main features of the ab initio occupancy predictions do not depend on the choice of the initial template. It is particularly interesting that even in the case of the ideal template, good correlation with the experiment is found. On the other hand, Fig. S7*C* indicates some subtle differences between profiles generated with the experimental template and that obtained using the ideal template. Note that these subtle differences can affect accuracy when predicting binding locations along strong positioning sequences.

The Effect of Changing the Threading Protocols. Does our method of threading affect the calculated occupancy profile? Given our standard force field (AMBER-99bsc0) and structural template (1kx5) one can thread the genomic sequence onto strand A starting at the 3′ end while at the same time strand B accommodates the complementary sequence. Alternatively one could start threading strand A at the 5′ end. In addition, one could exchange the genomic and its complementary sequence between strands. Fig. S7*D* illustrates that the way of threading has little effect on the predicted occupancy profiles. Clearly, the choice made in defining our standard threading protocol, which starts threading the genomic sequence to the 3′ end of strand A, is of no consequence. Note that this claim only holds for an ideal superhelical template such as that described in ref. 7 but may not hold for an asymmetric template, such as the nucleosomal crystal structures that include 601 DNA.

Toward Fragment-Based Calculations. Fig. S8 illustrates the effect of using only short-range (<7 Å) interactions on the nucleosome-formation energy profile for the first 1,000 sequence positions in Fig. 5*A*. It is clear from Fig. S8 that using short-range interactions, we can also predict some nucleosome positions, however this approach can lead to the identification of false minima that predict the presence of nucleosomes outside nucleosome-positioning sequences. This sort of short-range calculation mimics only one aspect of fragment-based methods in that they do not take into account all of the interactions we explicitly include in our approach (e.g., neighboring double-helix interactions). The use of fragment-based methods is further exposed to the lack-of-fragment-boundary problem during conformational minimization because the terminal nucleotides of fragments do not experience the presence of adjacent nucleotides. These approximations can lead to significant deviations from DNA geometry. Because we model the entire nucleosomal DNA, we are not exposed to the above dangling-boundary effects.

Contact Density in the Nucleosome Core Particle. We investigated the contact density in the reconstructed nucleosome core particle that is composed of our optimized nucleosomal DNA and the histone core proteins that we obtain from PDB structure 1kx5. After adding all hydrogens to histone core proteins (we always model DNA hydrogen atoms explicitly), we count the contacts made by all nucleotide base atoms with the remaining structure in 0.1-Å bins in the interval from 1.5 and 7 Å. In addition, we distinguish the contacts that occur within the DNA strands from the contacts that occur between them and the last types of contacts occurring between any of the DNA strands and the histone core proteins. Fig. S9*A* shows a typical contact density profile for a reconstructed nucleosome core particle. It is very clear from Fig. S9*A* that the contacts within and between the DNA strands dominate the contacts that occur between the DNA and histone proteins. Note that it has been also pointed out (20) that those limited numbers (compared with the number of DNA-DNA contacts) of DNA–protein contacts are nonspecific, therefore further supporting our finding that the physics-based factors (tested in the in vitro experiments) guiding preferential binding toward positioning sequences mainly arise from the properties of the nucleosomal DNA.

The Methylation Effect on Contact Density. Fig. S9*B* illustrates the effect of methylation on the contact density profiles shown in Fig. S9*A*. Although close inspection identifies very minor effects on the contact density profile within the DNA strands, Fig. S9*B* makes it clear that no new protein-DNA contacts form due to the methylation. Focusing on the contacts made by the cytosine-type base atoms (both in pure and 5Me-C DNA seen in Fig. S9*C*) shows that new contacts due to methylation occur mainly within DNA strands, suggesting a primary effect on either base–base stacking or the interaction between base and backbone atoms of the same strand. Fig. S9*D* demonstrates that the contacts of thymine bases are less directly affected by methylating of all cytosine residues.

Comparison of the sequence dependence and the methylation effect on the contact density is presented in Fig. S10. Fig. S10*A* shows the contact density between DNA strands and demonstrates that the methylation effect is not comparable in magnitude to the sequence effect for contacts that are within ∼5.5 Å. For more distant contacts the effect of methylation becomes as noticeable as the effect of changing the sequence. The effect of methylation is as significant as the sequence effect for all contacts within DNA strands. This finding further demonstrates that methylation have more influence on the packing of nucleotides within one DNA strand rather than on the contacts between individual DNA strands.

Reversibility of the Calculation. Using the entire set of optimized structures representing all nucleosome locations used in Fig. 5*A*, we performed calculations to demonstrate the reversibility of our approach. To do this we used the above set of structures to obtain an ensemble of new templates. Fig. S11*A* presents the distribution of the energies we obtain for the native sequence (from 1kx5 of ref. 7) when it is threaded and optimized on the above ensemble of nucleosomal templates. Compared with the energy distribution of different sequences on the same template structure (1kx5), we clearly show that the energy distribution of the native sequence on an ensemble of different templates is narrow. In addition, the rmsd distribution shown in Fig. S11*B* illustrates that we can recover the native structure within 1.5-Å resolution, still within the original resolution of the nucleosome core particle. Note, that for this test, we found it sufficient to study the energy of the nucleosome instead of studying the nucleosome formation energy, that is the energy difference between the nucleosomal and linear forms.

Minor Grooves and Helical Parameters of the DNA. As it is indicated in Fig. S11, our procedure when applied for a local sequence (representing a nucleosome position) does not change the template structure beyond its own resolution (1.9 Å) because only local conformational optimization is used so that we could afford to study many nucleosomal systems (over ∼100,000 in this work). These findings are in accordance with the conservation of DNA minor groove variations shown in Fig. S12*A* and the DNA helical parameters depicted in Fig. S12*B*. These conserved minor groove variations carry information about the presence of the histone core and only modulate the occupancy profiles, whose coarse shape is determined by global superhelix geometry. This finding is in agreement with observations in ref. 21.

The simulated linear DNA remains ideally straight as in the initial structure. For the same sequence position used in Fig. S12, this observation can be quantified by the 0.384-Å C4′-rmsd deviation between the backbones of the aligned simulated and initial structures, where the latter was built to be ideally straight along the helical axis. Next, we calculated the angle between the helical axis for the initial and the best-fit helical axis for the simulated structures. This angle was found to be 0.001 radian (0.081°). Based on this observation, the deviation from the ideally straight geometry is on the order of 10^{-3} radian or 0.1°. Another measure of the deviation from ideally straight geometry is the modulation in minor groove width, which was found to be on the order of 0.1 Å and all minor groove values are in the interval [11.7 Å, 11.9 Å]. Finally, the preservation of the ideal DNA helical parameters is further demonstrated by the 3.35-Å axial rise per residue with a SD of 0.07 Å found in the simulated structures. Note that the magnitude of these deviations (∼0.1 Å) from the ideal straight geometry is at least an order of magnitude smaller than the resolution (1.9 Å) of the nucleosomal DNA template (7) structure.

Therefore, sequence effects on the helical parameters of linear B-DNA have no noticeable consequence in our calculation. In addition, note that both nucleosome occupancy (Fig. S2*B*) and the methylation effect (Fig. 3 *B–D*) are arising from the nucleosomal structure, whereas the energy of the linear structure is mainly used to eliminate dependence on trivial factors such as the number of hydrogen bonds between strands. Nevertheless, the initial B-DNA template was constructed from an ideally straight reference polyadenine sequence.

Computational Requirements. The computational requirements of the physics-based approach depends on the details of the underlying protocols such as the total number of energy evaluations needed to obtain a score for a particular sequence position. Given that matching a sequence to its template requires only a local conformational search, the method of energy evaluation can be customized to reduce the cost of energy evaluation. Furthermore, the independence of each sequence calculation enables perfect parallel processing in which different subsequences are given to each of many processors running independently. It may not be necessary to slide the nucleosome along the DNA one nucleotide at a time. The optimal sliding step size consisting of several nucleotides can be determined based on deviation from profiles obtained with a single step size.

Limitations of Our Approach. The major limitation of the present study is its speed compared with sequence-dependent knowledge-based approaches. Thus, the length of genomic sequences that can be predicted is smaller with the present ab initio approach. Another limitation of the approach is the accuracy of the underlying force field, a limitation shared by all molecular modeling studies. Although our method is unique in its ability to bring consistent improvement into this rapidly developing field in parallel with our advancing knowledge of basic atomistic interactions, the present state of the art force fields are currently unable to provide prediction accuracy at base-pair resolution and predict experimentally known nucleosome-positioning propensities. By specifically developing force fields for this particular question and custom-designing structural optimization protocols to better capture the effect of sequence on the nucleosome template, we hope that we can overcome the above limitations in the next few years. Thus, the current work, to our knowledge, only represents the first step in a series of incremental advances.

Further Prospects. Cost-effective knowledge-based sequence-dependent methods could be combined with our ab initio approach leading to hybrid protocols. One possibility would be to use our approach to refine occupancy profiles based on sequence statistics. Another possibility would be to use our occupancy profiles for methylated DNA as a training set for sequence statistics-based methods (12, 16) so that the effect of methylation on the sequence specificity of nucleosome positioning could be evaluated genome-wide.

Because our ab initio approach is expected to predict the sequence preference for arbitrary DNA templates, it can be used to model nucleosome translocation mechanisms. This could be achieved with a series of DNA templates each representing an intermediate state along a proposed nucleosome translocation reaction coordinate. The sequence preferences of these intermediates could help to support or undermine a proposed nucleosome translocation mechanism predicted for a certain sequence region. Thus, we may be able to investigate whether nucleosome translocation is dictated by the same mechanism throughout the whole genome or whether there are local preferences, each triggered by local sequence patterns.

1. Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. *J Res Natl Bur Stand* 49(6):409–436.
2. Minary P, Levitt M (2010) Conformational optimization with natural degrees of freedom: A novel stochastic chain closure algorithm. *J Comput Biol* 17(8):993–1010.
3. Pérez A, et al. (2007) Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys J* 92(11):3817–3829.
4. Hingerty B, Richie RH, Ferrel TL (1985) Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopolymers* 24(3):427–439.
5. Watanabe M, Reinhardt WP (1990) Direct dynamical calculation of entropy and free energy by adiabatic switching. *Phys Rev Lett* 65(26):3301–3304.
6. Oppenheim AV, Schafer RW (1989) *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ), pp 311–312.
7. 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.
8. Levitt M (1978) How many base-pairs per turn does DNA have in solution and in chromatin? Some theoretical calculations. *Proc Natl Acad Sci USA* 75(2):640–644.
9. Minary P (2007) Methodologies for Optimization and SAmpling In Computational Studies (MOSAICS), Version 3.9. Available at www.cs.ox.ac.uk/mosaics. Accessed March 7, 2014.
10. Cherry JM, et al. (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26(1):73–79.
11. Yuan GC, et al. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* 309(5734):626–630.
12. Kaplan N, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458(7236):362–366.
13. Tillo D, Hughes TR (2009) G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* 10:442–457.
14. Trifonov EN, Sussman JL (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci USA* 77(7):3816–3820.
15. Satchwell SC, Drew HR, Travers AA (1986) Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* 191(4):659–675.
16. van der Heijden T, van Vugt JJFA, Logie C, van Noort J (2012) Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proc Natl Acad Sci USA* 109(38):E2514–E2522.
17. Cui F, Zhurkin VB (2010) Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. *J Biomol Struct Dyn* 27(6):821–841.
18. Cornell W, et al. (1995) A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J Am Chem Soc* 117(19):5179–5197.
19. MacKerell A, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616.
20. Xu F, Olson WK (2010) DNA architecture, deformability, and nucleosome positioning. *J Biomol Struct Dyn* 27(6):725–739.
21. Zhou T, et al. (2013) DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 41(Web Server issue): W56–W62.

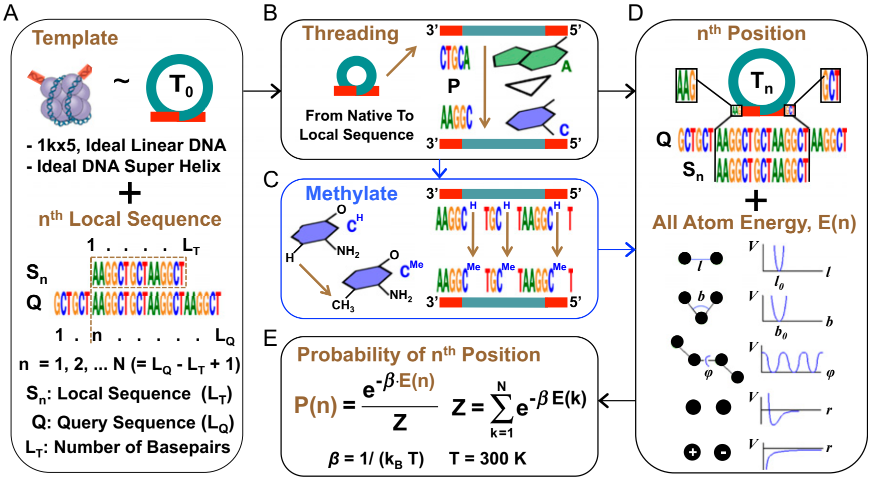


Fig. S1. Schematic illustration of the steps used in our ab initio approach to generate nucleosome occupancy profiles from first principles. (*A*) The use of an initial DNA template (T_0) of L_T base pairs and the query (genomic) sequence Q of length L_Q base pairs is shown. Local sequences S_n ($n = 1, \ldots, L_Q - L_T + 1$) are obtained by sliding the target sequence window (L_T base pairs) along the query sequence Q. (*B*) The two-stage process in which T_n, the nucleosome accommodating local sequence S_n, is built is shown. In the first stage each nucleotide base of T_0 is replaced with three pseudoatoms defining its plane (P). In the second stage, new nucleotide bases are built onto each plane. For example, how the fifth base (adenine) of the 3′–5′ strand of the template T_0 is replaced with the fifth base (cytosine) of the local sequence S_n is demonstrated. The complementary nucleotides on the 5′–3′ strand are also built. (*C*) If methylated DNA is modeled, then all cytosines are converted to 5Me-C on all template structures on both the nucleosomal DNA (PDB ID code 1kx5) and linear B-DNA forms. (*D*) The position of T_n along the genomic sequence and the terms of the energy function that are used to evaluate the score of this position are shown. (*E*) The all-atom energy scores are converted to probabilities.

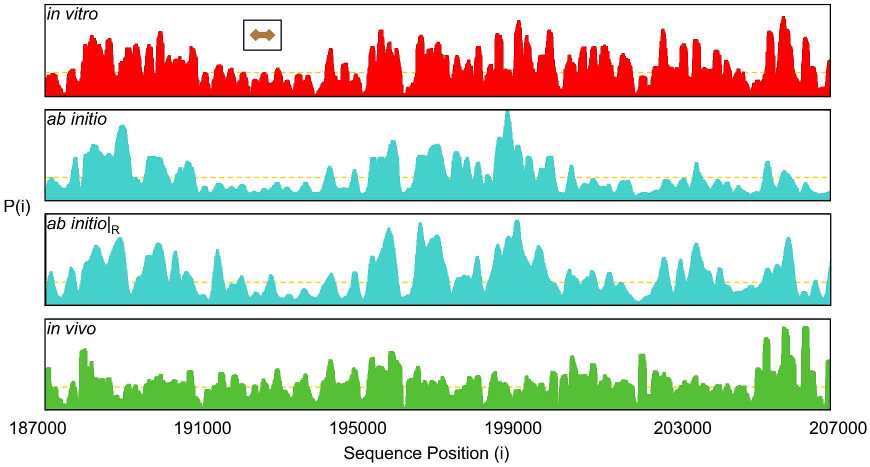


Fig. S2. Ab initio nucleosome occupancy profile along a 20,000-bp sequence of genomic DNA from yeast; sequence positions 187,000–207,000 in chromosome 14. The DNA template was taken from the crystal structure of PDB ID code 1kx5 and the molecular mechanics AMBER99-bsc0 force field with an implicit description of solvent was used for all calculations. The experimental profiles (red and green) have been placed on the same scale to facilitate comparison with the predicted data (cyan) using two separate approaches distinguished by the final energies [E(n) in Fig. S1*D*] used to generate the occupancy profile. In the first approach (ab initio), E(n) is defined as the energy obtained for the n^{th} local sequence minus the average energy obtained for all sequences considered. In the second approach (ab initio$|_R$). we define E(n) as the difference between the energy obtained for the nucleosome with the n^{th} sequence minus the energy obtained for an ideal linear B-DNA holding the same sequence. The double-headed arrow (*Inset*) shows the length of the sequence region accommodating a nucleosome. A similar length of 147 is also used in the filter that smoothed the ab initio data. The dashed yellow lines mark the average nucleosome occupancies over the 20,000-bp region.

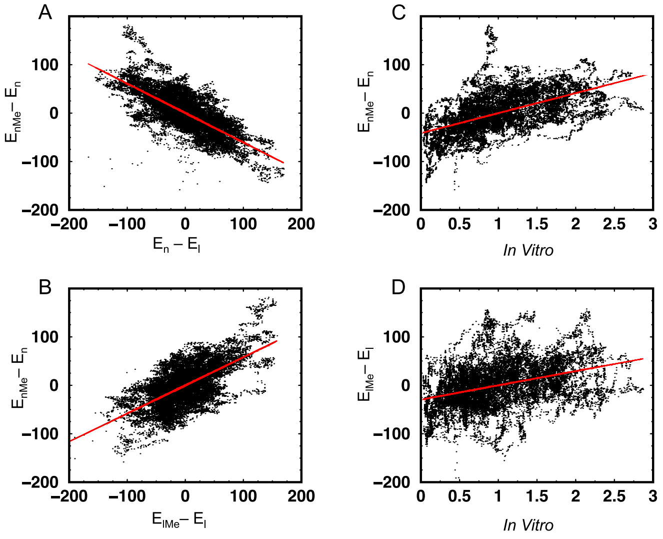


Fig. S3. Additional correlations are presented for Fig. 3. (*A*) The methylation energy on the nucleosome $E_{nMe} - E_n$ is plotted against nucleosome formation energy $E_n - E_l$. The strong anticorrelation is captured by the correlation coefficient CC = −0.739. (*B*) Plotting the methylation energy on the nucleosome $E_{nMe} - E_n$ as a function of the methylation energy on the linear form $E_{lMe} - E_l$ reveals the strong correlation (CC = 0.637) between the two quantities. (*C* and *D*) The methylation energies on the nucleosomal and linear forms are plotted against the in vitro occupancy. The correlation coefficients for the nucleosomal and linear forms are CC = 0.556 and CC = 0.360, respectively.

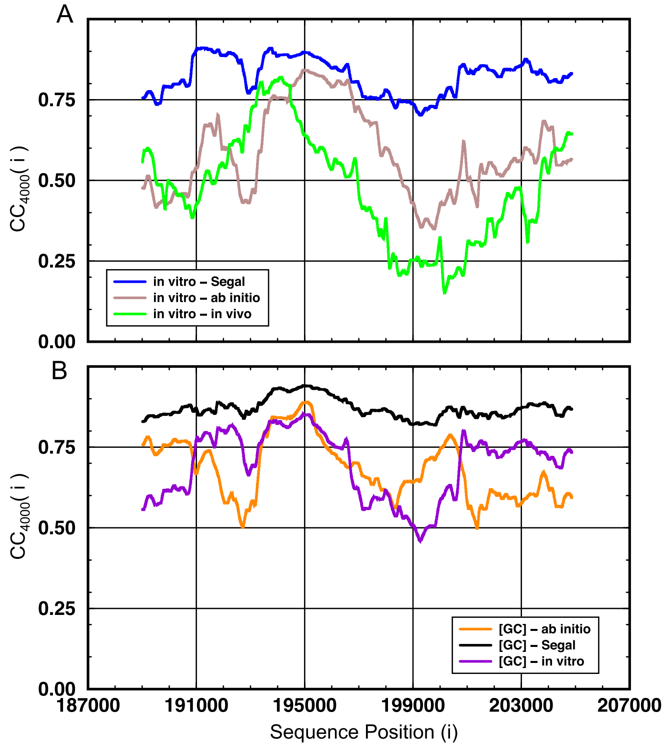
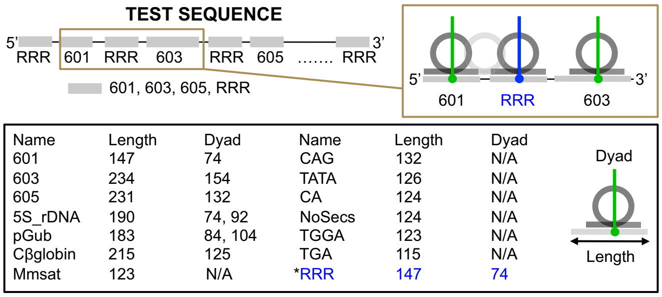
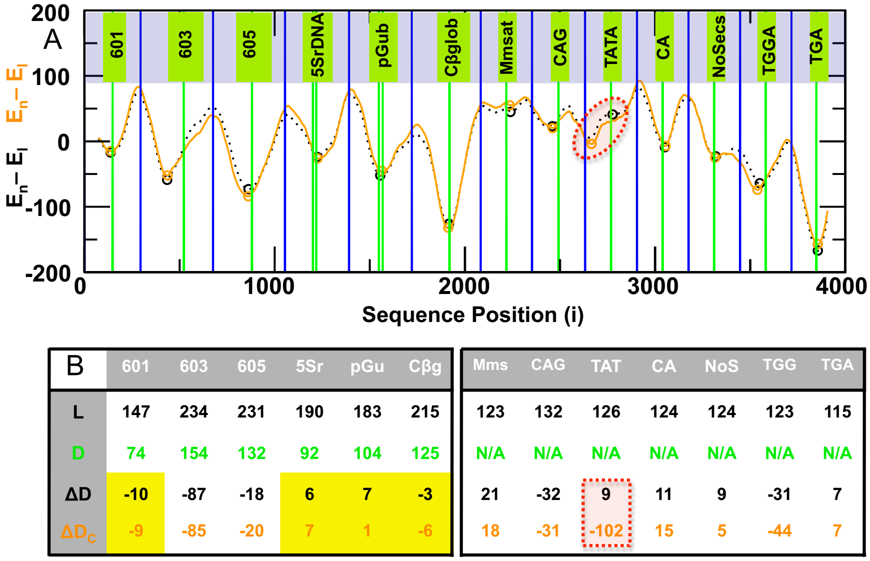


Fig. S4. Correlations among GC percentage, in vitro, in vitro model reconstruction (Segal), in vivo, and nucleosome formation energy ($E_n - E_l$)-based ab initio profiles. Position-dependent correlations are calculated as the correlation coefficient in a 4,000-bp sequence window that slides along a 20,000-bp sequence of genomic DNA from yeast; sequence positions 187,000–207,000 in chromosome 14. (*A*) Position-dependent correlations between in vitro and all other profiles. The overall correlations are 0.8185, 0.6155, and 0.5065 for in vitro and Segal (blue), in vitro and ab initio (brown), and in vitro and in vivo (green), respectively. (*B*) Position-dependent correlations with GC profile. The overall correlations are 0.8779, 0.7431, and 0.6843 for GC and Segal (black), GC and ab initio (orange), and GC and in vitro (magenta), respectively.



Name	Length	Dyad	Name	Length	Dyad
601	147	74	CAG	132	N/A
603	234	154	TATA	126	N/A
605	231	132	CA	124	N/A
5S_rDNA	190	74, 92	NoSecs	124	N/A
pGub	183	84, 104	TGGA	123	N/A
Cβglobin	215	125	TGA	115	N/A
Mmsat	123	N/A	*RRR	147	74

Fig. S5. Illustrating the composition of the test sequence that is constructed by concatenating 13 established nucleosome-positioning target sequences (used in ref. 13), 601, 603, 605, 5S_rDNA, pGub, chicken β-globin, mouse minor satellite (Mmsat), CAG, TATA, CA, NoSecs, TGGA and TGA with RRR, a 147-nt long random DNA sequence generated by an online engine (www.faculty.ucr.edu/~mmaduro/random.htm) using a C+G/A+T ratio of 0.40. The table presents the length and known dyad locations of nucleosome-positioning sequences. Note, that some target sequences (601) have the length equal to that of the nucleosome DNA (147 bp) while others may be longer (603, 605) or shorter (Mmsat). For some nucleosome-positioning target sequences such as 601, 603, 605, 5S rDNA, pGub, and chicken β-globin, information for the exact dyad location is also provided.



B	601	603	605	5Sr	pGu	Cβg	Mms	CAG	TAT	CA	NoS	TGG	TGA
L	147	234	231	190	183	215	123	132	126	124	124	123	115
D	74	154	132	92	104	125	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ΔD	-10	-87	-18	6	7	-3	21	-32	9	11	9	-31	7
ΔD_c	-9	-85	-20	7	1	-6	18	-31	-102	15	5	-44	7

Fig. S6. (*A*) Nucleosome formation energies for normal DNA modeled by using implicit (black dotted line) and explicit (orange) counterions as a function of the dyad position along the test sequence that is constructed by concatenating nucleosome-positioning sequences separated by a random DNA sequence of 147 nt (Fig. S5). The vertical green lines indicate dyad locations where the nucleosome is expected to be centered. If the dyad location is not known, the green lines refer to the center nucleotide of the sequence. The blue lines indicate the center of the random sequence on our nucleosome template. Circles mark minima or saddle points of the computed energy. (*B*) Showing four properties of 13 established positioning sequences (Fig. S5), 601, 603, 605, 5Sr DNA, pGub, chicken β-globulin, Mmsat, CAG, TATA, CA, NoSecs, TGGA and TGA. (Row 1) L is the length or the number of nucleotides in the sequence. (Row 2) D is an experimentally verified dyad location (if available). (Row 3) ΔD is the difference between the dyad location and the nearest energy minimum (or saddle point) for normal DNA modeled with implicit counterions. Yellow shading highlights the accurate prediction of nucleosome positions (within 10 nt) for 4 of the 6 sequences with verified dyad locations. If dyad locations are not known, ΔD represents the difference between the location of the center nucleotide and the nearest energy minimum or saddle point. (Row 4) ΔD_C is the same as ΔD for DNA modeled with explicit counterions. The red box highlights one example where the explicit treatment of counterions leads to a different prediction.

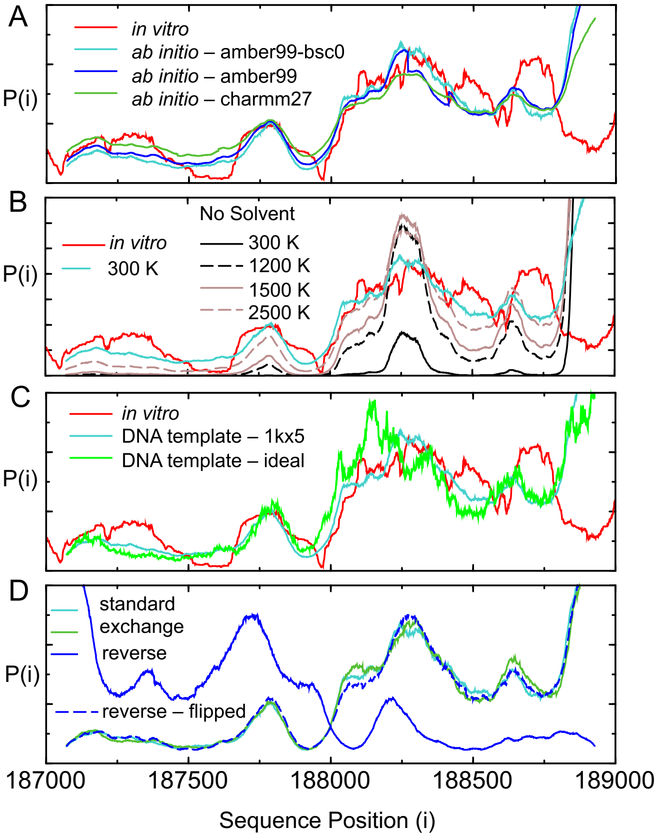


Fig. S7. Sensitivity of computed nucleosome occupancy profile to different force fields, solvent conditions, different initial DNA templates, and different ways of threading the same (reference) sequence onto the template. (*A*) Ab initio nucleosome occupancy generated along sequence region 187,000–189,000 in yeast chromosome 14, using the DNA template from PDB crystal structure 1kx5 and three different force fields—AMBER99-bsc0 (cyan), AMBER99 (blue), and CHARMM27 (green)—each used with the same implicit description of solvent. The renormalized experimental profile (red) is shown for comparison. (*B*) Ab initio occupancy profiles based on template 1kx5 and the AMBER99-bsc0 force field were generated both with and without the distance dependent dielectric treatment that we use to account for solvent effects. Solvent-free profiles at various temperatures: T = 300 K (solid black), 1,200 K (dashed black), 1,500 K (solid brown), and 2,500 K (dashed brown) are compared with the profile with solvent at T = 300 K (cyan) as well as that from in vitro (red) experiment. (*C*) Ab initio nucleosome occupancy profiles based on AMBER99-bsc0 force field with solvent were generated using two different templates: 1kx5 (cyan) and ideal DNA superhelix (green). The correlation coefficients to the experiment are 0.7714, 0.7800, and 0.8210 for force fields AMBER99-bsc0, AMBER99, and CHARMM27, respectively, and 0.7458 and 0.7187 for template DNA from 1kx5 and ideal superhelix, respectively. (*D*) The three different threading scenarios are tests. In the original case (cyan), the reference sequence is threaded onto strand A starting at the 3′ end while strand B accommodates the complementary sequence. In the exchange case (green), the reference and its complementary sequences are exchanged between strands. In the reverse case (blue), the reference sequence is threaded onto strand A starting at the 5′ end; it is also shown in a flipped version (around a vertical axis) to aid comparison (dashed blue). In the main results presented here, AMBER99-bsc0 force field and the DNA template from crystal structure 1kx5 were used.

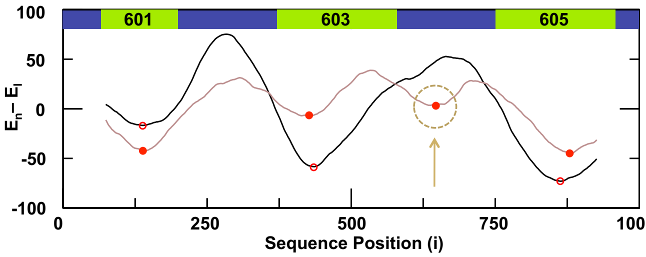


Fig. S8. Nucleosome formation energies for normal DNA modeled by using our method with a standard cutoff of 12 Å (black line) and an artificially small cutoff of 7 Å (brown line) for the first 1,000 positions of Fig. 5*A*. The energy minima are denoted by red circles and dots. For both cutoffs, the energy minima identify nucleosome positions located on each of the 601, 603, and 605 nucleosome-positioning sequences. In addition, the artificially small cutoff protocol predicts a false positive minimum (brown dashed circle), and identifies a nucleosome-binding position where no nucleosome-positioning sequence is present.

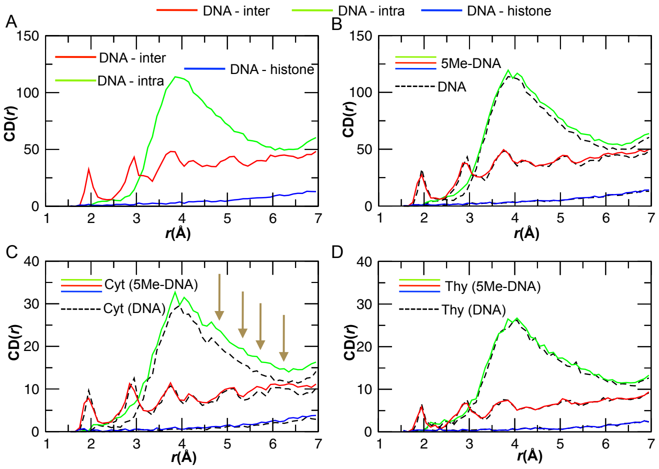


Fig. S9. Distance-dependent contact density (CD) in the reconstructed nucleosome core particle composed of our optimized nucleosomal DNA structure (representing the first minimum in Fig. 5*A*) and all histone proteins (obtained from PDB ID code 1kx5) with added hydrogens. The number of contacts made by DNA base atoms is calculated in bins 0.1-Å wide in the interval from 1.5 to 7 Å and the resulting numbers are normalized by the volume ($4/3\ \pi[(r + dr)^3 - r^3]$) of successive spherical shells. (*A*) Contact density within DNA strands (green) and between DNA strands (red) and between DNA and protein (histones) (blue). (*B*) The contact density presented in *A* is compared for pure (black dashed line) and methylated (colored solid line) DNA. (*C*) The contact density for the cytosine base in pure and 5Me-C DNA. Arrows indicate the noticeable deviation due to methylation in the contact densities within strands. (*D*) The same as in *C* for the thymine base. In all calculations contacts within the same nucleotide are excluded.

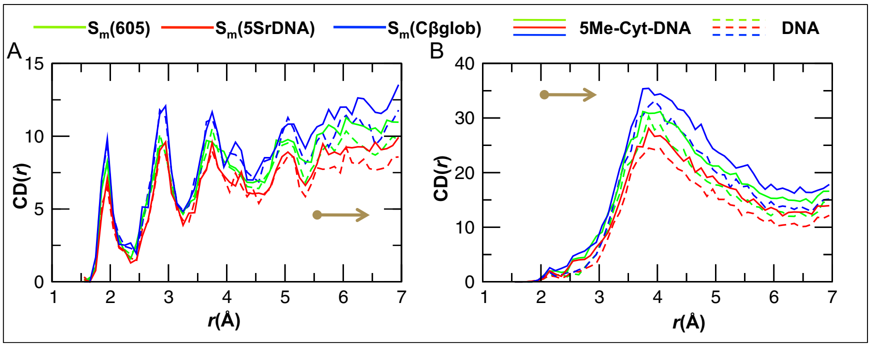


Fig. S10. The distance-dependent CD (Fig. S9) between DNA strands (*A*) and within DNA strands (*B*) is calculated for different DNA sequences—$S_m(605)$ (green), $S_m(5SrDNA)$ (red), and $S_m(C\beta glob)$ (blue)—that represent local minima on the nucleosome formation energy landscape presented in Fig. 5*A*. Results are shown for both methylated (solid line) and pure (dashed line) DNA. The brown arrows mark the minimum distance above which the methylation effect becomes comparable to the sequence effect.

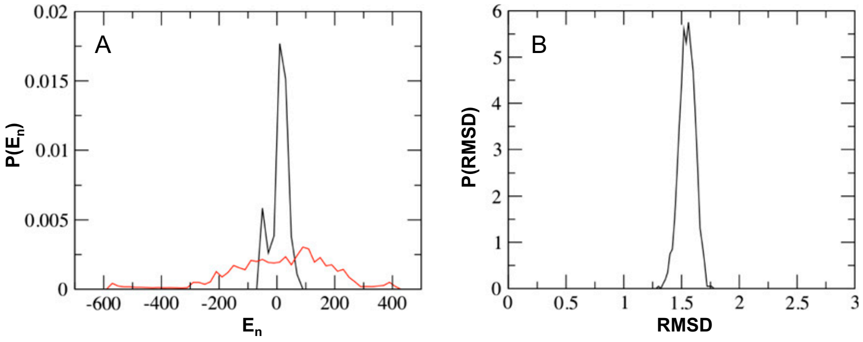


Fig. S11. (*A*) Distribution of minimized nucleosomal DNA energies for the native sequence (black) of 1kx5 accommodating all template structures initially minimized for each individual local sequence used in the calculation for Fig. 5*A*. The distribution of nucleosomal DNA energies optimized for each individual local sequence starting from the PDB template structure of 1kx5 is also shown (red). (*B*) Distribution of the rmsd of the reconstructed native structures of 1kx5 used in *A* from the original 1kx5 structure from the PDB.

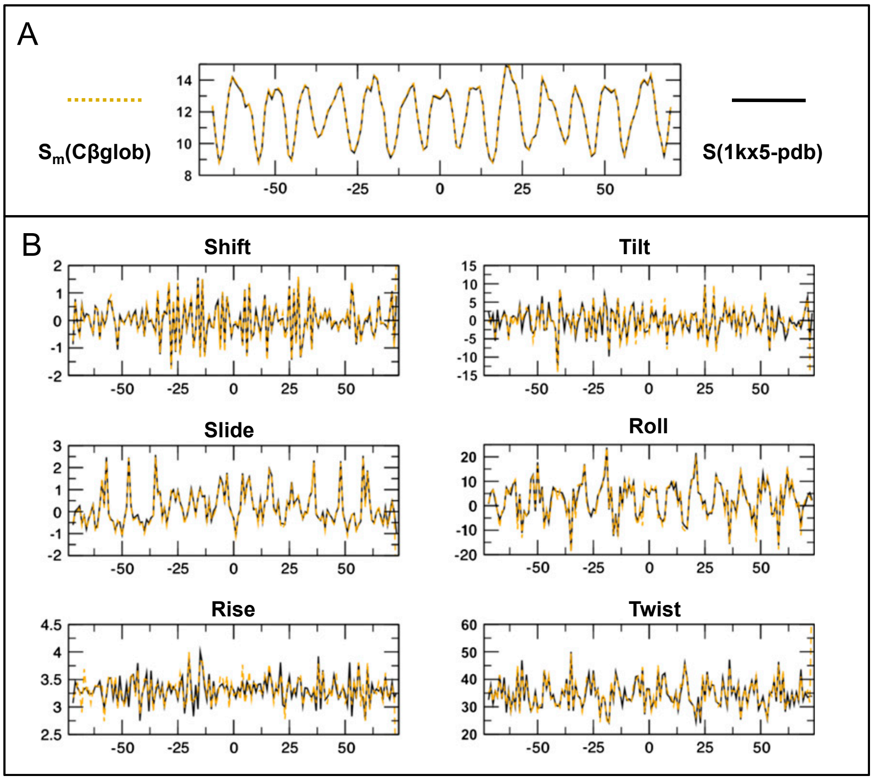


Fig. S12. The minor groove width and the helical parameters of the nucleosomal DNA accommodating the S_m(Cβglob) sequence (orange dashed line) used in Fig. S10 are compared with the same parameters for the crystal structure (black solid line) PDB ID code 1kx5. (*A*) Minor groove width. (*B*) Helical parameters: shift, slide, and rise in Å and tilt, roll, and twist in degrees. Our procedure preserved the detailed helical geometry of the X-ray structure.