*Published online 9 June 2009*

*Nucleic Acids Research, 2009, Vol. 37, No. 14* ***4707–4722***
*doi:10.1093/nar/gkp475*

# Using DNA mechanics to predict *in vitro* nucleosome positions and formation energies

**Alexandre V. Morozov[1,*], Karissa Fortney[2], Daria A. Gaykalova[3], Vasily M. Studitsky[3], Jonathan Widom[2] and Eric D. Siggia[4]**

[1]Department of Physics & Astronomy and BioMaPS Institute for Quantitative Biology, Rutgers University, 136 Frelinghuysen Road, Piscataway, NJ 08854, [2]Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, IL 60208, [3]Department of Pharmacology, UMDNJ, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854 and [4]Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA

Received February 6, 2009; Revised May 6, 2009; Accepted May 18, 2009

## ABSTRACT

**In eukaryotic genomes, nucleosomes function to compact DNA and to regulate access to it both by simple physical occlusion and by providing the substrate for numerous covalent epigenetic tags. While competition with other DNA-binding factors and action of chromatin remodeling enzymes significantly affect nucleosome formation *in vivo*, nucleosome positions *in vitro* are determined by steric exclusion and sequence alone. We have developed a biophysical model, DNABEND, for the sequence dependence of DNA bending energies, and validated it against a collection of *in vitro* free energies of nucleosome formation and a set of *in vitro* nucleosome positions mapped at high resolution. We have also made a first *ab initio* prediction of nucleosomal DNA geometries, and checked its accuracy against the nucleosome crystal structure. We have used DNABEND to design both strong and weak histone-binding sequences, and measured the corresponding free energies of nucleosome formation. We find that DNABEND can successfully predict *in vitro* nucleosome positions and free energies, providing a physical explanation for the intrinsic sequence dependence of histone–DNA interactions.**

## INTRODUCTION

Genomic DNA is packaged into chromatin in eukaryotic cells. The building block of chromatin is the nucleosome, (1), a 147 bp DNA segment wrapped in ~1.8 superhelical coils around the surface of a histone octamer (2). The unstructured histone tails are targets of numerous covalent modifications (1) and may influence folding of nucleosome arrays into higher order chromatin structures. Chromatin can both block access to DNA (3) and juxtapose sites far apart on the linear sequence (4).

While nucleosome positions *in vitro* are determined only by intrinsic sequence preferences and steric exclusion, *in vivo* chromatin remodeling enzymes play a role that needs to be clarified. In one scenario, the role of such enzymes is purely catalytic, modifying the rate of assembly but not the final disposition of nucleosomes on DNA. In the other, chromatin remodeling enzymes actively reposition nucleosomes to control access to DNA, in analogy with motor proteins. It has not been possible to determine by genetics where living cells fall between these extremes. Therefore, to quantify the contribution of chromatin remodeling enzymes to *in vivo* chromatin structure a model is required that can accurately position nucleosomes *in vitro*.

Recent computational approaches used collections of nucleosomal sequences isolated *in vivo* (5–8) and *in vitro* (9) to train pattern matching tools that were then applied genome wide. However, the training data may not be representative of direct histone–DNA binding because other factors may reposition nucleosomes *in vivo*, while *in vitro* genomic data are affected by steric exclusion between neighboring nucleosomes and by the chromatin fiber formation which results in long-range contacts between distant nucleosomes. Furthermore, models based on alignments of nucleosome positioning sequences (5,6) require a choice of background or reference sequence and it is known that nucleotide composition varies among functional categories of DNA and among organisms.

Here, we focus on developing a biophysical model for the intrinsic sequence dependence of nucleosome formation—a first step towards quantitative description of *in vivo* chromatin. Our model resolves the nucleosome formation energy into the sum of two terms: histone–DNA interactions and DNA bending energy. The histone–DNA

*To whom correspondence should be addressed. Tel: +1 732 445 1387; Fax: +1 732 445 5958; Email: morozov@physics.rutgers.edu

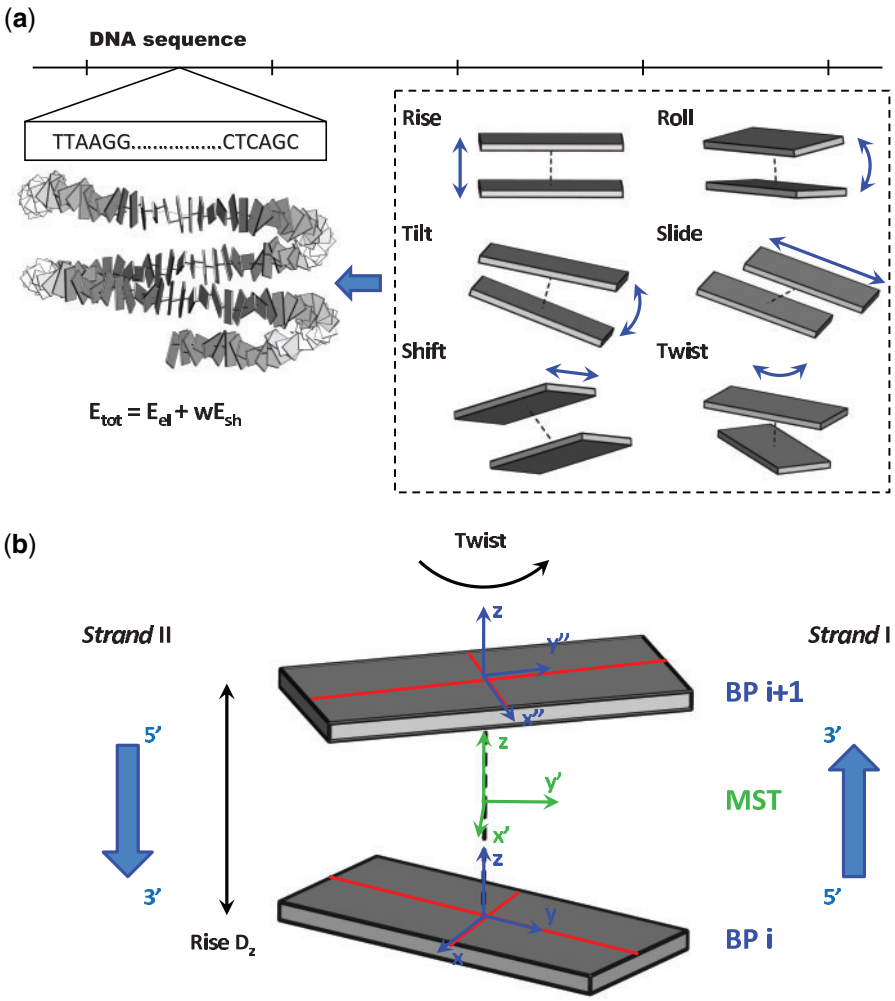


**Figure 1.** (**a**) DNA mechanics model of histone–DNA interactions. Conformation of a single DNA basestep (defined as two consecutive DNA base pairs in the $5' \rightarrow 3'$ direction) is described by six geometric degrees of freedom: rise, shift, slide, twist, roll and tilt. (11) DNA base pairs are shown as rectangular blocks. The minimized nucleosome energy (a weighted sum of the elastic energy $E_{el}$ and the restraint energy $E_{sh}$ which penalizes deviations of the DNA conformation from the ideal superhelix, see Methods section) is computed for each position along the DNA sequence. (**b**) Schematic illustration of a single dinucleotide (basestep) geometry. Coordinate frames attached to base pairs $i$ and $i + 1$ are shown in blue, and the MST coordinate frame is shown in green. For illustrative purposes, only rise $D_z$ and twist $\Omega$ are set to nonzero values. The origin of the MST frame is at the midpoint of the line connecting the origins of two base pair frames (which are separated by $D_z$ Å along the $z$-axis); the MST frame is rotated through $\Omega/2$ with respect to the frame $i$.

potential is assumed to be sequence independent because there are few direct contacts between histone side chains and DNA bases (10). For the DNA bending, we construct an empirical sequence-specific quadratic potential (11,12) using a database of 101 nonhomologous, nonhistone protein–DNA crystal structures to infer the elastic force constants.

In particular, we model DNA base stacking energies by defining three displacements (rise, shift and slide) and three rotation angles (twist, roll and tilt) for each dinucleotide [two adjacent base pairs, Figure 1a; (11)]. Together the six degrees of freedom completely specify the spatial position of base pair $i + 1$ in the local coordinate frame of base pair $i$ (Figure 1b), and can be used to reconstruct an arbitrary DNA conformation in global Cartesian coordinates (see Methods section). We assume that the histone–DNA potential is at a minimum along an ideal superhelix whose pitch and radius are inferred from the nucleosome crystal structure (2), and varies quadratically when the DNA deviates from the ideal superhelix. This sequence-independent term represents average attractive interactions between the histones and the DNA phosphate backbone (13) and steric exclusion between the histone octamer and the DNA.

The sum of the DNA bending and the histone–DNA potentials is minimized to yield the elastic energy and the DNA conformation for each nucleosomal sequence (Figure 1a). This is in contrast with the currently available DNA mechanics methods that impose DNA conformation from the nucleosome crystal structure (14) or from the ideal superhelix (15) regardless of the DNA sequence. Because the total energy is quadratic, energy minimization is equivalent to solving a system of linear equations for which efficient algorithms are available. There is no

genomic background in this model, but the results may depend on the quality of the protein–DNA structural data set. Since our bending energy is empirical and inferred from co-crystal structures, it lacks a physical energy scale. However, by comparison with the worm-like chain model our units can be converted to kilocalories per mole through multiplication by 0.26 (see Methods section). Our program DNABEND is not limited to nucleosomal superhelices—we can compute sequence-specific bending energies for DNA molecules of arbitrary length, restrained to follow arbitrary spatial curves imposed by DNA-bound proteins, experimental constraints, etc.

To model formation of multiple nucleosomes on longer DNA tracts, we have adapted a standard dynamic programming algorithm to positioning multiple nucleosomes and other factors on DNA (5,16). The algorithm uses standard thermodynamics and enforces steric exclusion between bound factors in any given configuration (see Methods section). The binding energy landscapes for each factor are used to infer their binding probabilities and base pair occupancies (defined as the probability for a base pair to be covered by any factor of a given type). DNABEND software and additional supporting data are available on the Nucleosome Explorer web site: http://nucleosome.rockefeller.edu.

We check the performance of DNABEND using sets of short DNA sequences for which free energies of nucleosome formation and nucleosome positions are not affected by steric exclusion and chromatin fiber formation, and for which we had high-resolution experimental measurements, either our own or from the literature. We also design and experimentally test nucleosomal sequences with both high and low free energies of nucleosome formation, and rank sets of sequences selected for their ability to facilitate or hinder nucleosome formation. We find that using DNABEND is preferable to keeping DNA geometry fixed for predicting free energies of nucleosome formation. Relaxing DNA geometries is less critical for predicting nucleosome positions, where all versions of the DNA mechanics model and the most recent bioinformatics algorithm (9) yield comparable results.

## MATERIALS AND METHODS

### DNA geometry

We model each DNA base pair as a rigid body to which a local coordinate frame is attached (Figure 1b). The position of base pair $i$ in the local frame of base pair $i-1$ is uniquely specified by six geometric parameters: three Euler angles that define the unit vectors of frame $i$ and the displacement vector which gives its origin: $\alpha_i = (\vec{\Omega}_i, \vec{d}_i)$, $i = 1, \ldots, N$. Here $\vec{\Omega}_i \equiv \{\Omega_i, \rho_i, t_i\}$ are the helical twist, roll and tilt angles, and $\vec{d}_i$ is the displacement vector with the $x, y, z$ components called slide, shift and rise, respectively (Figure 1a) (17–20). These geometric parameters can be used to construct the global rotation matrix of the base pair $i$ recursively:

$$R_i = R_{i-1} T_i \ (i = 1, \ldots, N), \qquad \mathbf{1}$$

where each $T_i$ matrix is a product of three rotations:

$$T_i = \mathbb{R}_z(-\frac{\Omega_i}{2} + \phi_i)\mathbb{R}_y(\Gamma_i)\mathbb{R}_z(-\frac{\Omega_i}{2} - \phi_i). \qquad \mathbf{2}$$

Here, $\mathbb{R}_y(\theta)$ and $\mathbb{R}_z(\theta)$ are the rotation matrices around the $y$ and $z$ axes, and the middle term on the right-hand side of Equation (2) introduces both roll and tilt with a single rotation through $\Gamma_i$:

$$\mathbb{R}_y(\theta) = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix},$$

$$\mathbb{R}_z(\theta) = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\begin{cases} \Gamma_i = (\rho_i^2 + t_i^2)^{1/2}, \\ \cos\phi_i = \rho_i/\Gamma_i, \\ \sin\phi_i = t_i/\Gamma_i. \end{cases}$$

It is conventional to use the mid-step triads (MSTs) rather than the base pair triads to transform the displacement vector $\vec{d}_i$ into the global frame (18–20): $\vec{d}_i^{(g)} = R_i^{\mathrm{MST}} \vec{d}_i$, where MSTs are defined by:

$$R_i^{\mathrm{MST}} = R_{i-1} T_i^{\mathrm{MST}}, \qquad \mathbf{3}$$

$$T_i^{\mathrm{MST}} = \mathbb{R}_z\left(-\frac{\Omega_i}{2} + \phi_i\right)\mathbb{R}_y\left(\frac{\Gamma_i}{2}\right)\mathbb{R}_z(-\phi_i). \qquad \mathbf{4}$$

Thus, a complete set of local geometric parameters $\alpha_i$ is equivalent to knowing the global conformation of the DNA molecule with $(N+1)$ bp: first, the recursive relation (1) is employed to determine the orientations of all base pair coordinate frames (except for $R_0$ which is fixed and determines the overall orientation and position of the molecule). Second, Equation (3) is used to construct the MST frames, which are then employed to transform all local displacements $\vec{d}_i$ into the global frame. Finally, the displacement vectors are added up to determine the origins of the base pair coordinate frames. The inverse problem is also well-defined: a full set of base pair and MST rotation matrices in the global frame is sufficient for reconstructing all local degrees of freedom $\{\alpha_i\}$ (19).

Introducing nonzero roll and tilt angles imposes curvature onto the DNA conformation: consider the curvature vector $\vec{\kappa}_i = \vec{t}_{i+1} - \vec{t}_i$, where $\vec{t}_i$ is the tangent unit vector of base pair $i$. It can be shown that to the lowest order in roll and tilt $|\vec{\kappa}_i| = (\rho_i^2 + t_i^2)^{1/2} = \Gamma_i$. Furthermore, the roll and tilt contributions to curvature are shifted by 90° with respect to one another.

### Ideal superhelix

DNA in the nucleosome core particle approximately follows an ideal left-handed superhelix with pitch

$P$ and radius $R$ (2):

$$\vec{r}(s) = \begin{cases} R\cos(s/R_{\text{eff}}), \\ R\sin(s/R_{\text{eff}}), \\ -(P/2\pi R_{\text{eff}})s \end{cases} \quad \mathbf{5}$$

where $s$ is the arc length, and $R_{\text{eff}} = \sqrt{R^2 + (P/2\pi)^2}$.

A local frame at position $s$ is given by a set of three orthonormal Frenet vectors (tangent, normal and binormal) (21):

$$\begin{cases} \vec{t}(s) = \mathrm{d}\vec{r}/\mathrm{d}s, \\ \vec{n}(s) = \mathrm{d}\vec{t}/\mathrm{d}s/|\mathrm{d}\vec{t}/\mathrm{d}s|, \\ \vec{b}(s) = \vec{t} \times \vec{n}. \end{cases} \quad \mathbf{6}$$

We position $N + 1$ bp equidistantly on the ideal superhelix by distributing $N + 1$ sets of Frenet basis vectors along the superhelical curve: $s_i = 2\pi R_{\text{eff}}\alpha(i/N)$ $(i = 0, \ldots, N)$, where $\alpha$ is the number of nucleosomal superhelical turns. Fitting an ideal superhelix to the high-resolution crystal structure of the nucleosome core particle (2) gives $\alpha = 1.84$, $P = 25.9$ Å and $R = 41.9$ Å. To model helical twist, we impose a rotation around the tangent vector $\vec{t}$, such that the neighboring frames differ by $\Omega_0 = 34.696°$—the average helical twist from the structure. These rotations are applied to both base pair and MST frames [the MST frames are located at $s_i = 2\pi R_{\text{eff}}\alpha(i-1/2)/N$ $(i = 1, \ldots, N)$].

A superhelix described by Equation (5) has constant curvature: $|\vec{t}(s_{i+1}) - \vec{t}(s_i)| = 2\sin(\pi\alpha/N)$, corresponding to $\Gamma_i = 4.53°$, $\forall i$. In the ideal superhelix, roll and tilt make equal contributions to the curvature: $\vec{\Omega}_i = (\Omega_0, \Gamma_i\cos(\Omega^i_{tot} + \phi_0), \Gamma_i\sin(\Omega^i_{tot} + \phi_0))$ and $\vec{d}_i = (0, 0, d)$, where $d = 3.333$ Å, $\Omega^i_{tot} = i\Omega_0$, and $\phi_0$ is the initial phase determined by the first base pair. Thus, twist and rise are constant for every base pair in the ideal superhelix, slide and shift are zero, whereas roll and tilt exhibit oscillations resulting from the superhelical curvature and shifted by 90° with respect to one another (cf. green curves in Figure 2).

### DNA conformational energy

The total energy of a nucleosomal DNA is given by a weighted sum of two quadratic potentials:

$$E = E_{el} + wE_{sh}, \quad \mathbf{7}$$

where $E_{el}$ is sequence-specific DNA elastic energy (11,12) and $E_{sh}$ is nonspecific histone–DNA interaction energy. Below we elaborate on each energy in turn.

*DNA elastic energy*. We represent DNA elastic energy as (11):

$$E_{el} = \frac{1}{2}\sum_{s=1}^{N}[\alpha^s - \langle\alpha^{n(s)}\rangle]^T F^{n(s)}[\alpha^s - \langle\alpha^{n(s)}\rangle], \quad \mathbf{8}$$

where $\alpha^s$ is the six component vector of angles and displacements, the sum runs over all consecutive dinucleotides (basesteps) $s$ and $\langle\alpha^n\rangle$ are the average values of the local degrees of freedom computed for all basestep types $(n = AA, AC, AG, \ldots, TT)$ using a collection of oligonucleotides extracted from a set of 101 nonhomologous protein–DNA structures (12). The matrix of force constants $F^n$ is evaluated by inverting the covariance matrix $C^n$ of deviations of local geometric parameters from their average values $(\alpha^n - \langle\alpha^n\rangle)$ (11):

$$(F^n)^{-1} = C^n, \text{ where } C^n_{ij} = \langle(\alpha^n_i - \langle\alpha^n_i\rangle)(\alpha^n_j - \langle\alpha^n_j\rangle)\rangle. \quad \mathbf{9}$$

Note that our elastic energy model utilizes only the first and second moments of the empirical distributions of dinucleotide geometries. Strongly bent dinucleotides (with one or more geometric parameters further than 3 SD from the mean) are iteratively excluded from the data set (11). Our model does not use any higher order moments of empirical geometry distributions, which would lead to a nonquadratic elastic potential; nor are there sufficient data to model more than successive base pairs. The empirical parameters of the elastic model consist of 10 basestep-type dependent averages for each of the six local degrees of freedom (out of 16 basesteps only 10 are unique, i.e. not related by reverse complementarity) and 15 $(6 \times 5/2)$ independent force constants in each of the 10 symmetric $6 \times 6$ matrices $F^n$. All elastic model parameters are listed in the Supplementary Tables 1–3.

To transform the local displacements (shift, slide and rise) into the global frame, we apply the following coordinate transformation to Equation (8):

$$\mathbf{R}_s = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & R^{\text{MST}}_s \end{pmatrix} \quad \mathbf{10}$$

($\mathbf{1}$ and $\mathbf{0}$ denote $3 \times 3$ unit and zero matrices, respectively). The transformed elastic force constants are $\mathbf{F}^n = \mathbf{R}_s F^n \mathbf{R}_s^{-1}$, leading to:

$$E_{el} = \frac{1}{2}\sum_{s=1}^{N}[\mathbf{R}_s(\alpha^s - \langle\alpha^{n(s)}\rangle)]^T \mathbf{F}^{n(s)}[\mathbf{R}_s(\alpha^s - \langle\alpha^{n(s)}\rangle)]. \quad \mathbf{11}$$

Finally, we express all degrees of freedom in terms of their deviations from the ideal superhelix:

$$\begin{cases} \vec{d}^{(g)}_s = \vec{d}^{(g),0}_s + \delta\vec{d}_s, \\ \vec{\Omega}_s = \vec{\Omega}^0_s + \delta\vec{\Omega}_s. \end{cases} \quad \mathbf{12}$$

Note that all the transformations described in this section involve no additional approximations to the original DNA elastic energy [Equation (8)]. Thus, we are free to choose the most convenient rotation matrix in Equation (10), and use $R^{\text{MST}}_s|_0$ from the ideal superhelix in Equation (11).

*Histone-DNA interaction energy*. Nonspecific histone–DNA interactions are modeled with a quadratic potential that penalizes deviations of nucleosomal DNA from the ideal superhelix:

$$E_{sh} = \sum_{s=1}^{N}(\vec{r}_s - \vec{r}^{\,0}_s)^2, \quad \mathbf{13}$$

where $\vec{r}_s$ and $\vec{r}_s^0$ are the nucleosomal DNA and the ideal superhelix radius vectors in the global frame ($s = 1, \ldots, N$):

$$\begin{cases} \vec{r}_s = \vec{r}_0 + \sum_{j=1}^{s} \vec{d}_j^{(g)}, \\ \vec{r}_s^0 = \vec{r}_0^0 + \sum_{j=1}^{s} \vec{d}_j^{(g),0}. \end{cases} \quad \mathbf{14}$$

Then to the lowest order the difference between the radius vectors is given by:

$$r_s^\beta - r_s^{0,\beta} = \sum_{j=1}^{s} \left[ \delta d_j^\beta + \sum_{\alpha=1}^{3} \sum_{j'=1}^{j} b_{jj'}^{\alpha\beta} \delta\Omega_{j'}^\alpha \right], \quad \mathbf{15}$$

where $\alpha, \beta = 1, \ldots, 3$ label the vector components, and

$$b_{jj'}^{\alpha\beta} = \frac{\partial R_j^{MST}}{\partial \Omega_{j'}^\alpha} |_0 \left( \vec{d}_j^0 \right)^\beta$$

are the connectivity coefficients. The first term in Equation (15) represents the net change in the global radius vector $r_s^\beta$ caused by the changes in the preceding displacements $d_j$, while the second term reflects the change in the global radius vector resulting from modifying one of the rotation angles $\Omega_j^\alpha, j \leq s$. Note that changing a rotation angle at position $j$ affects downstream base pair positions linearly by introducing a bend into the DNA chain, whereas changing a displacement at position $j$ results in a parallel shift of all downstream coordinates. The first derivative of the rotation matrix is evaluated as:

$$\frac{\partial R_j^{MST}}{\partial \Omega_{j'}^\alpha} |_0 = \begin{cases} T_0 \ldots T_{j-1} \frac{\partial T_j^{MST}}{\partial \Omega_j^\alpha} & j' = j \\ T_0 \ldots T_{j'-1} \frac{\partial T_{j'}}{\partial \Omega_{j'}^\alpha} T_{j'+1} \ldots T_{j-1} T_j^{\mathrm{MST}} & 1 \leq j' < j. \end{cases}$$

Upon substitution of the expansion (15) into the restraint energy, we obtain an effective quadratic potential:

$$E_{sh} = \sum_{i,j=1}^{N} \delta\alpha^{i\,T} \mathbf{G}^{ij} \delta\alpha^j, \quad \mathbf{16}$$

where $\delta\alpha^i = (\delta\vec{\Omega}_i, \delta\vec{d}_i)$, and the $6 \times 6$ matrix of force constants is given by three distinct $3 \times 3$ submatrices:

$$\mathbf{G}^{ij} = \begin{pmatrix} H_{ij} & F_{ij} \\ F_{ij} & G_{ij} \end{pmatrix}.$$

Here,

$$G_{ij}^{\alpha\beta} = \delta_{\alpha\beta} M(i,j),$$

$$F_{ij}^{\alpha\beta} = \sum_{l=j}^{N} M(i,l) b_{lj}^{\alpha\beta},$$

$$H_{ij}^{\alpha\beta} = \sum_{k=i}^{N} \sum_{l=j}^{N} M(k,l) \sum_{\gamma=1}^{3} b_{ki}^{\alpha\gamma} b_{lj}^{\beta\gamma},$$

where $M(i,j) = N + 1 - \max(i,j)$, and $\delta_{\alpha\beta}$ is the Kronecker delta ($\alpha, \beta = 1, \ldots, 3$). $G_{ij}^{\alpha\beta}$ couples displacements with displacements, $F_{ij}^{\alpha\beta}$ couples displacements with angles (and thus has one connectivity coefficient), and $H_{ij}^{\alpha\beta}$ couples angles with angles through two connectivity coefficients.

*Total energy and DNA conformation minimization*. The total energy of nucleosomal DNA is a function of the fitting weight $w$, introduced to capture the balance between favorable histone–DNA interactions and the unfavorable energy of bending DNA into the nucleosomal superhelix. We fit $w$ to maximize the average correlation coefficient between the distributions of geometric parameters observed in the high-resolution crystal structure of the nucleosome core particle [PDB code 1kx5; (2)] and the corresponding DNABEND predictions (Figure 2). This procedure yields $w = 0.1$ for the 147 bp superhelix and $w = 0.5$ for the 71 bp superhelix bound by the $H3_2H4_2$ tetramer (base pairs 39 through 109 in the 147 bp superhelix). DNABEND is not very sensitive to the exact value of $w$: we found a correlation of 0.99 between the free energies computed using $w = 0.1$ and $w = 0.5$ for the 71 bp superhelix (data not shown).

The final conformation of the DNA molecule is the one that minimizes its total energy $E$ [Equation (7)]:

$$\frac{\partial E}{\partial \delta\alpha_i^s} = 0 \quad s = 1, \ldots, N, \quad i = 1, \ldots, 6, \quad \mathbf{17}$$

where $\delta\alpha^s = (\delta\vec{\Omega}_s, \delta\vec{d}_s)$ and $N$ is the number of dinucleotides. The numerical solution of the system of linear equations (17) yields a set of geometric parameters $(\delta\vec{\Omega}_s, \delta\vec{d}_s)$ and the corresponding elastic energy $E_{el}$ for a given nucleosome position. The geometric parameters in the local frame can also be reconstructed: $(\Omega_s^0 + \delta\vec{\Omega}_s, \vec{d}_s^0 + R_s^{\mathrm{MST}-1} \delta\vec{d}_s)$.

*Worm-like chain*. We use estimates based on the worm-like chain model of DNA to convert elastic energies into kilocalories per mole. According to the worm-like chain model, the energy required to bend a DNA molecule is given by (22):

$$E_{wlc} = \frac{k_B T L_p}{2} \int_0^{L_0} \mathrm{d}s |\frac{\mathrm{d}\vec{t}}{\mathrm{d}s}|^2, \quad \mathbf{18}$$

where $L_0$ is the Contour length of the molecule, $L_p$ is the persistence length (estimated to be $\simeq$400 Å) (22), $k_B T \simeq 0.6$ kcal/mol and $\vec{t}$ is the tangent unit vector. The contour length of the ideal 147 bp superhelix is given by $146 \times 3.333$ Å. From Equations (5) and (6), we obtain $|\mathrm{d}\vec{t}/\mathrm{d}s|^2 = R^2/R_{\mathrm{eff}}^4$, and thus

$$E_{wlc} = \frac{k_B T L_p L_0}{2} \frac{R^2}{R_{\mathrm{eff}}^4} \simeq 32.6 \,\mathrm{kcal/mol}.$$

In Figure 4c, the mean and the standard deviation $\sigma$ of DNABEND energies computed for chromosome III are $127.0 \pm 6.1$. Equating the worm-like chain model estimate with the mean DNABEND energy, we obtain a scaling coefficient of 0.26. This yields a difference of 15.2 kcal/mol between the best and the worst chromosome III sequences, and $\sigma = 1.6$ kcal/mol. Most sequences differ by $2\sigma$ or less in Figure 4c and are thus separated by

$\leq$6.4 kcal/mol. A similar value of the scaling coefficient (0.21) arises from a linear model fit between experimental and DNABEND-predicted free energies in Figure 4a (red circles).

### Predicting genome-wide occupancies of DNA-binding proteins

Nucleosome formation energies (given by $E_{el}$) and DNA-binding energies of other factors at each genomic position can be used as input to a dynamic programming algorithm (16) that outputs factor binding probabilities and base pair occupancies for each DNA element.

Here, we develop this approach for a general case of $M$ objects of length $L_j$ ($j = 1, \ldots, M$) placed on genomic DNA. The objects could represent nucleosomes, transcription factors (TFs) or any other DNA-binding proteins. The binding energy of object $j$ at each position $i$ is assumed to be known: $E_i^j (i = 1, \ldots, N - L_j + 1)$, where $N$ is the number of base pairs. We assign index 0 to the background which can be formally considered to be an object of length 1: $L_0 = 1$. In the simplest case which we consider here the background energy is zero everywhere, but more sophisticated models could incorporate a global bias by making positions near DNA ends less favorable, etc.

We wish to compute a statistical sum over all possible configurations in which object overlap is not allowed (including the background 'object'):

$$Z = \sum_{conf} e^{-E(conf)}, \quad \mathbf{19}$$

where $E(conf) = \sum_{j=0}^{M} \sum_{i=1}^{N_{obj}^j} E_{c(i)}^j$ is the total dimensionless energy of an arbitrary configuration of nonoverlapping objects, $N_{obj}^j$ is the number of objects of type $j$ and $E_{c(i)}^j$ is the precomputed energy of the object of type $j$ which occupies positions $c(i)$ through $c(i) + L_j - 1$.

It is possible to evaluate $Z$ [or the free energy $F = \log(Z)$] efficiently by recursively computing the partial statistical sums:

$$Z_i^f = \sum_{j=0}^{M} Z_{i-L_j}^f e^{-E_{i-(L_j-1)}^j} \theta_{i-(L_j-1)}, \; i = 1, \ldots, N, \quad \mathbf{20}$$

with the initial condition $Z_0^f = 1$. The theta function is defined as:

$$\theta_i = \begin{cases} 1, i > 0 \\ 0, i \leq 0 \end{cases} \quad \mathbf{21}$$

The partial free energies $F_i = \log Z_i^f$ can be calculated in a similar way:

$$F_i = F_{i-1} + \log\left( \sum_{j=0}^{M} e^{F_{i-L_j} - F_{i-1} - E_{i-L_j+1}^j} \theta_{i-L_j+1} \right), \; i = 1, \ldots, N, \quad \mathbf{22}$$

with the initial condition $F_0 = 0$. Since the algorithm proceeds by computing partial sums from 1 to $N$ it is often called the forward pass. Similar equations can be constructed for the backward pass which proceeds from $N$ to 1:

$$Z_i^r = \sum_{j=0}^{M} Z_{i+L_j}^r e^{-E_i^j} \theta_{N-i-L_j+2}, \; i = N, \ldots, 1, \quad \mathbf{23}$$

with the initial condition $Z_{N+1}^r = 1$. In terms of the backward partial free energies $R_i = \log Z_i^r$:

$$R_i = R_{i+1} + \log\left( \sum_{j=0}^{M} e^{R_{i+L_j} - R_{i+1} - E_i^j} \theta_{N-i-L_j+2} \right), i = N, \ldots, 1. \quad \mathbf{24}$$

with the initial condition $R_{N+1} = 0$. Note that $R_1 = F_N = F = \log(Z)$.

With the full set of forward and backward partial free energies, we can evaluate any statistical quantity of interest. For example, the probability of finding an object of type $j$ at positions $(i, \ldots, i + L_j - 1)$ is given by:

$$P_i^j = \frac{Z_{i-1}^f e^{-E_i^j} Z_{i+L_j}^r}{Z} = e^{F_{i-1} - E_i^j + R_{i+L_j} - F}, i = 1, \ldots, N - L_j + 1. \quad \mathbf{25}$$

Another quantity of interest is the occupancy of the base pair $i$ by object $j$, defined as the probability that base pair $i$ is covered by any object of type $j$ (5):

$$O_i^j = \sum_{k=i-(L_j-1)}^{i} P_k^j = O_{i-1}^j + P_i^j - P_{i-L_j-1}^j, i = 1, \ldots, N. \quad \mathbf{26}$$

(note that $P_k^j = 0$ for $k < 1$ and $k > N - L_j + 1$, and $O_0^j = 0$).

Finally, we need to take into account the fact that the objects can bind DNA in both directions, and thus there are two binding energies for each position: $E_i^j$ (object $j$ starts at $i$ and extends in the 5′ to 3′ direction) and $E_i^{j(rc)}$ (object $j$ starts at $i + L_j - 1$ and extends in the 3′ to 5′ direction). It is easy to show that the formalism developed above applies without change if the binding energies $E_i^j$ are replaced by the free energies $\bar{E}_i^j$ which take both the binding orientations into account:

$$\bar{E}_i^j = -\log(e^{-E_i^j} + e^{-E_i^{j(rc)}}). \quad \mathbf{27}$$

In the case of a single type of DNA-binding object (such as nucleosomes described by DNABEND), there are two free parameters: the mean nucleosome energy $\langle E^{\text{nuc}} \rangle$ over a chromosome or a given DNA region which plays the role of the chemical potential (note that $E^{\text{nuc}}$ stands for $E^{\text{nuc}} - \mu$ in the grand canonical ensemble formulation used above), and the SD $\sigma(E^{\text{nuc}})$ which plays the role of the inverse temperature. The temperature determines the fraction of stable nucleosomes (defined as $P_i > 0.5$, where $P_i$ is the probability to start a nucleosome at base pair $i$), while the chemical potential determines the average nucleosome occupancy. We have set $\langle E^{\text{nuc}} \rangle = 0.0$ to produce the genome-wide nucleosome occupancy of 0.797 [which is close to previously published models (5,23)]. We also set $\sigma^2(E^{\text{nuc}}) = 45.0$, which results in stable, nonoverlapping nucleosomes covering 16.3% of

the yeast genome. Thus, the DNABEND energy landscape was rescaled to fit the bulk nucleosome occupancy in yeast.

### Histone–DNA binding affinity measurements

We used a standard competitive reconstitution procedure to measure the relative affinity of different DNA sequences for binding to histones in nucleosomes (24). In this method, differing tracer DNA molecules compete with an excess of unlabeled competitor DNA for binding to a limiting pool of histone octamer. The competition is established in elevated [NaCl], such that histone–DNA interactions are suppressed and the system equilibrates freely. The [NaCl] is then slowly reduced by dialysis, allowing nucleosomes to form; further reduction in [NaCl] to physiological concentrations or below 'freezes-in' the resulting equilibrium, allowing subsequent analysis, by native gel electrophoresis, of the partitioning of each tracer between free DNA and nucleosomes. The distribution of a given tracer between free DNA and nucleosomes defines an equilibrium constant and a corresponding free energy, valid for that competitive environment. Comparison of the results for a given pair of tracer DNAs in the identical competitive environment eliminates the dependence on the details of the environment, yielding the free energy difference ($\Delta\Delta G$) of histone interaction between the two tracer DNAs. To allow for comparison with other work, we include additional tracer DNAs as reference molecules: a derivative of the 5S rDNA natural nucleosome positioning sequence 24 and the 147 bp nucleosome-wrapped region of the selected high affinity nonnatural DNA sequence 601 (25).

The 5S and 601 reference sequences were prepared by PCR using plasmid clones as template. The 146- and 147-bp long DNAs analyzed in X-ray crystallographic studies of nucleosomes (PDB codes 1aoi and 1kx5, respectively) were prepared as described (26) using clones supplied by Professors K. Luger and T.J. Richmond, respectively. New 147-bp long DNA sequences designed in the present study were prepared in a two-step PCR-based procedure using chemically synthesized oligonucleotide primers. All synthetic oligonucleotides were gel-purified prior to use. The central 71 bp were prepared by annealing the two strands. The resulting duplex was gel purified and used as template in a second stage PCR reaction to extend the length on each end creating the final desired 147-bp long DNA. The resulting DNA was again purified by gel electrophoresis.

DNA sequences to be analyzed were 5′-end labeled with $^{32}$P, and added in tracer quantities to competitive nucleosome reconstitution reactions. Reconstitution reactions were carried out as described (24) except that each reaction included 10 μg purified histone octamer and 30 μg unlabeled competitor DNA (from chicken erythrocyte nucleosome core particles) in the 50 μl microdialysis button.

### Hydroxyl radical footprinting of nucleosomal templates

*DNA templates*. Plasmids pGEM-3Z/601, pGEM-3Z/603 and pGEM-3Z/605 containing nucleosome positioning sequences 601, 603 and 605, respectively, were described previously (27). To obtain templates for hydroxyl radical footprinting experiments the desired ~200-bp DNA fragments were PCR-amplified using various pairs of primers and Taq DNA polymerase (New England BioLabs). The sequences of the primers can be provided on request. To selectively label either upper or lower DNA strands one of the primers in each PCR reaction was 5′-end radioactively labeled with polynucleotide kinase and $\gamma^{32}$P-ATP (28). The single-end-labeled DNA templates were gel-purified and single nucleosomes were assembled on the templates by dialysis from 2 M NaCl (28). Nucleosome positioning was unique on at least 95% of the templates.

*Hydroxyl radical footprinting*. Hydroxyl radicals introduce nonsequence-specific single nucleotide gaps in DNA, unless DNA is protected by DNA-bound proteins (29). Hydroxyl radical footprinting was conducted using single-end-labeled histone-free DNA or nucleosomal templates as previously described (29). In short, 20–100 ng of single-end-labeled DNA or nucleosomal templates were incubated in 10 mM HEPES buffer (pH 8.0) in the presence of hydroxyl radical-generating reagents present at the following final concentrations [2 mM Fe(II)-EDTA, 0.6% H2O2, 20 mM Na-ascorbate] for 2 min at 20°C. Reaction was stopped by adding thiourea to 10 mM final concentration. DNA was extracted with Phe:Chl (1:1), precipitated with ethanol, dissolved in a loading buffer and analyzed by 8% denaturing PAGE.

*Data analysis and sequence alignment*. The denaturing gels were dried on Whatman 3MM paper, exposed to a Cyclone screen, scanned using a Cyclone and quantified using OptiQuant software (Perklin Elmer). Positions of nucleotides that are sensitive to or protected from hydroxyl radicals were identified by comparison with the sequence-specific DNA markers (Supplementary Figures 2 and 3). The dyad was localized by comparison of the obtained footprints with the footprints of the nucleosome assembled on human α-satellite DNA. The latter footprints were modeled based on the available 2.8 Å resolution X-ray nucleosome structure (10).

## RESULTS

### Prediction of DNA geometries from the nucleosome crystal structure

Unlike the previous approaches that keep DNA conformation frozen regardless of the sequence (14,15), DNABEND finds both sequence-specific nucleosome formation energies and the corresponding DNA geometries. Therefore, one way to validate DNABEND is to predict the DNA conformation in the high-resolution (1.9 Å) nucleosome crystal structure [(2); PDB code 1kx5], using only DNA sequence as input. As can be seen in Figure 2, DNABEND predictions are significantly correlated with the experimental geometries for twist, roll, tilt and slide ($r \geq 0.49$), but are less successful with shift and rise, although the peak positions are generally correct. The correlation between our model and 1kx5 is significantly

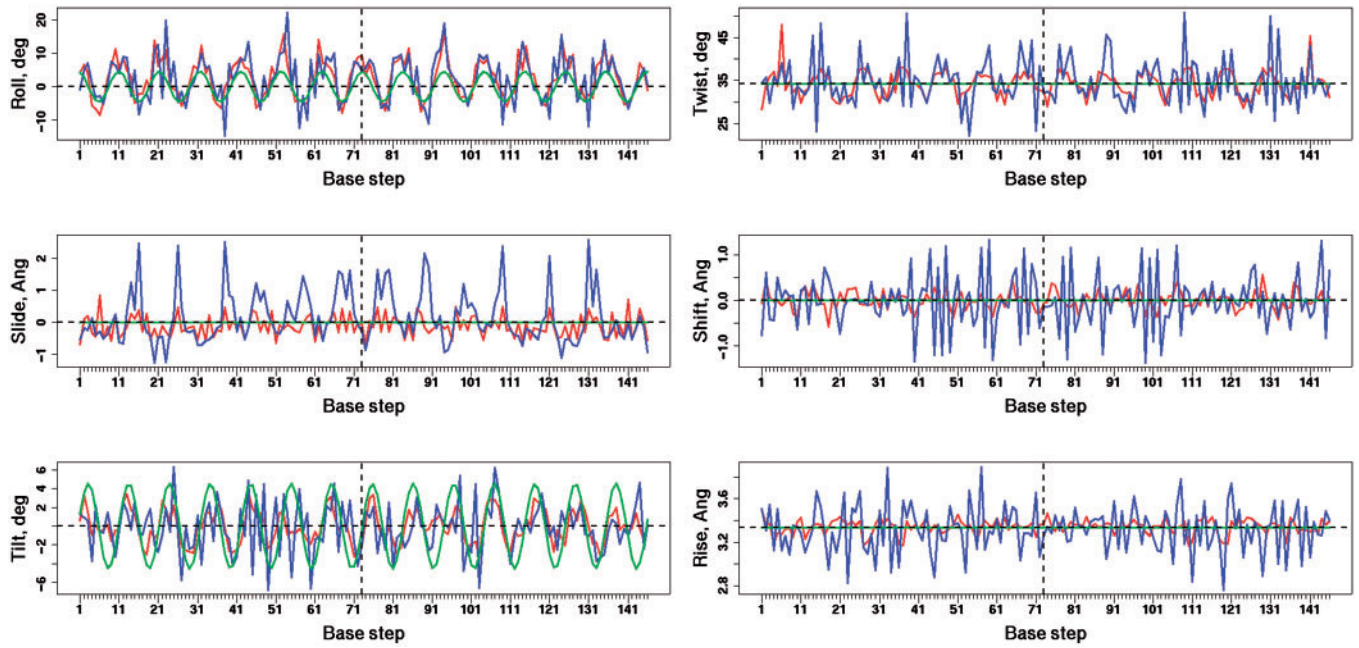


**Figure 2.** DNABEND-predicted and experimentally observed DNA geometries. Six dinucleotide degrees of freedom in the crystal structure of the nucleosome core particle [(2); PDB code: 1kx5] (blue), in the minimum energy structure obtained using 1kx5 DNA sequence as input to DNABEND (red) and in the ideal superhelix with no energy relaxation (green). The 2-fold nucleosome symmetry axis is shown as a dashed vertical line. Mean values of the geometric degrees of freedom in the ideal superhelix are shown as dashed horizontal lines. Correlation coefficients between the degrees of freedom from the native and the minimized structures are: ($r_{twist}$, $r_{roll}$, $r_{tilt}$, $r_{slide}$, $r_{shift}$, $r_{rise}$) = (0.489, 0.709, 0.539, 0.536, 0.247, 0.238) ($\langle r \rangle$ = 0.460). Correlation coefficients between the degrees of freedom from the native structure and the ideal superhelix are: ($r_{twist}$, $r_{roll}$, $r_{tilt}$, $r_{slide}$, $r_{shift}$, $r_{rise}$) = (−0.066, 0.669, 0.322, −0.550, 0.027, 0.021) ($\langle r \rangle$ = 0.071).

higher than the correlation between 1kx5 and the ideal superhelix (see Figure 2, caption), although part of the predicted correlation in roll and tilt is simply due to the ideal helical oscillations.

There are several inaccuracies in DNABEND predictions: e.g. rapid shift oscillations in the region between bps 35 and 105 are not reproduced, and in general the magnitude of observed oscillations in rise and shift is underestimated. Furthermore, most slide peaks are under-predicted, which is especially important because positive slide makes a significant contribution to the superhelical trajectory in the nucleosome crystal structure (14). Slide is under-predicted by DNABEND because for certain key base pairs the protein–DNA structures in our training set imply a much smaller mean value of slide (e.g. 0.18 Å for CA basesteps) than that observed in a currently available, limited set of nucleosome structures (0.91 Å for CA basesteps). Changing just this one mean value produces slide peaks of a more reasonable magnitude, but does not improve the correlation coefficient (data not shown). Because nucleosomal DNA is highly bent, different degrees of freedom are strongly coupled (Supplementary Figure 1a): for example, base pairs tend to tilt and shift simultaneously to avoid a steric clash. These couplings are much less pronounced in the nonhistone protein–DNA complexes used to derive the elastic energy model (Supplementary Figure 1b), but nonetheless appear prominently when the 1kx5 DNA geometry is predicted by DNABEND (Supplementary Figure 1c). Thus, DNABEND is reasonably successful in reproducing nucleosomal DNA geometries *ab initio*; remaining discrepancies can be attributed to the deficiencies of the elastic energy model (which by necessity is based on the currently available set of protein–DNA complexes), and to the approximation inherent in expanding DNA geometries around the ideal superhelix.

## Relationship between DNA geometries and sequence specificity

Analysis of available nucleosome crystal structures shows that tight DNA wrapping is facilitated by sharp DNA kinks if flexible dinucleotides (e.g. 5′-CA/TG-3′ or 5′-TA-3′) are introduced into the region where the minor groove faces the histone surface. For other sequences and other structural regions the bending is distributed over several dinucleotides (2). We substituted all possible dinucleotides into the 1kx5 atomic structure (keeping DNA conformation fixed), and computed the elastic energy for each sequence variant. The most sequence-specific regions are those where the minor groove faces the histone octamer (Figure 3a). The specificity is especially dramatic if DNA is strongly kinked (e.g. at positions 109, 121 and 131; Figure 2) (2). Although these positions are occupied by the CA/TG dinucleotides in the crystal structure, the model assigns the lowest energy to the TA dinucleotide, consistent with the periodic TA signal previously observed in good nucleosome positioning sequences (30) (Figure 3b). The observed dinucleotide ranking is in agreement with the averages, standard deviations and force constants inferred from the structural database

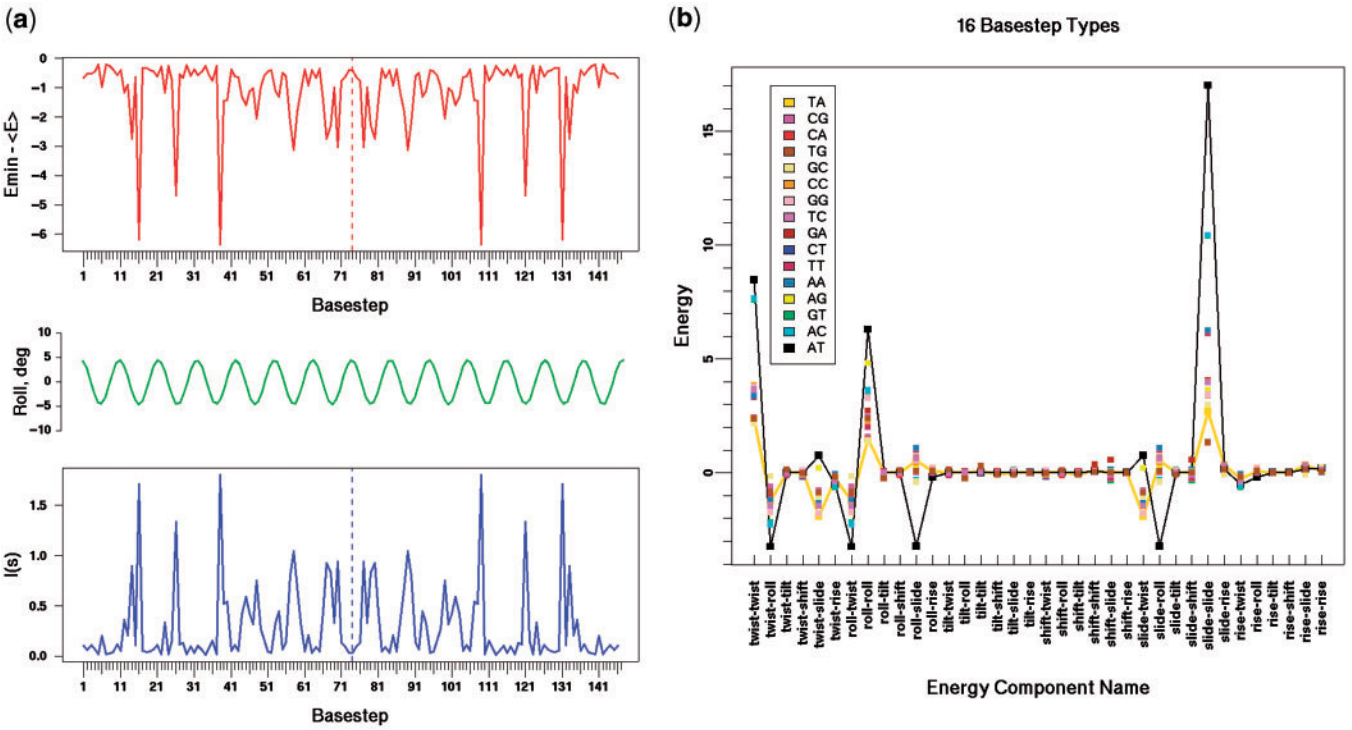


**Figure 3.** Elastic energy analysis of the nucleosome crystal structure. (**a**) Position-dependent sequence specificity in the nucleosomal DNA revealed by the energetic analysis of dinucleotides substituted into the crystal structure of the nucleosome core particle [PDB code: 1kx5; (2)] All possible dinucleotides were introduced at every position into the 147 bp nucleosomal site using DNA dihedral angles from the native dinucleotide, and DNA elastic energy was computed for every sequence variant. Upper panel: the difference between the energy of the most favorable dinucleotide and the average energy of all dinucleotides at this position. Lower panel: information entropy, defined as $I(s) = \log_2(16) + \sum_{i=1}^{16} p_i^s \log_2(p_i^s)$, where $p_i^s = \exp(-E_i^s)/\sum_{i=1}^{16}\exp(-E_i^s)$, and $E_i^s$ is the elastic energy change which results from introducing a dinucleotide of type $i = 1, \ldots, 16$ at position $s$: $E_i^s = E_i^{s(mut)} - E_i^{s(wt)}$. To enforce the 2-fold symmetry of the nucleosome core particle, all dinucleotide energies were symmetrized around the middle of the DNA site, shown as a dashed vertical line. Middle panel: roll angle of the ideal superhelix showing DNA geometry in relation to the histone octamer. Negative roll angles correspond to the minor groove facing the histone surface. (**b**) Elastic energy components for all possible dinucleotides substituted into the 1kx5 crystal structure at position 109 where the DNA conformation is kinked (Figure 2) (2). Dinucleotides are ranked by their total energy as shown in the legend (best to worst energy from top to bottom). TA is the lowest energy dinucleotide (thick golden line). The energy component analysis reveals that it is the degrees of freedom related to slide (slide–slide and slide–twist components) and roll (roll–roll component) that make the TA dinucleotide most favorable, although the slide–slide component is slightly better in the native CA/TG dinucleotide (red/brown dots). In contrast, the AT dinucleotide (black lines) has the highest energy due to its low flexibility with respect to roll, slide and twist (Supplementary Table 2).

of protein–DNA complexes: for example, standard deviations of the roll, slide and twist degrees of freedom are the highest for the TA dinucleotide (Supplementary Table 2).

### *In vitro* free energies of nucleosome formation and ranking of selected sequence sets

DNABEND accurately predicts experimental free energies of nucleosome formation (31–33) (Figure 4a and b). Geometry minimization is essential for these predictions—the same calculations are much less successful if DNA geometries are taken from the nucleosome crystal structure (1kx5) (14), or if DNA is threaded along an ideal superhelix [Supplementary Figure 4; (15)]. DNA base pairs threaded along the ideal superhelix do not form an accurate representation of nucleosomal DNA, which follows a 'zig-zag' path and exhibits numerous kinks and irregularities in the crystal structure (2). At the same time, DNA conformation taken from 1kx5 or any other crystal structure is necessarily sequence specific—for example, it is likely that major slide kinks at positions 109, 121 and 131 in 1kx5 are caused simply by the presence of the CA dinucleotides, and will occur elsewhere or disappear altogther if these dinucleotides are mutated or moved. Thus, it is not *a priori* obvious that any particular geometry which reflects a single DNA sequence provides a universal structural template.

DNABEND can also separate sequences selected *in vitro* for their ability to form stable nucleosomes (27) or occupied by nucleosomes *in vivo* (5) from genomic yeast sequences (Figure 4c). Note that the lowest energies are assigned to a set of sequences that were chemically synthesized and then subjected to multiple rounds of selection for binding affinity (Figure 4c, black histogram) (27). Thus, these sequences can be expected to form more stable nucleosomes than those typically found in the yeast genome. In contrast, a selection experiment on yeast genomic sequences, or an MNase digestion assay used to find stable *in vivo* nucleosomes (5) result in the energy distributions that overlap much more with the histogram of genomic energies. Finally, DNABEND correctly ranks mouse genome sequences selected *in vitro* on

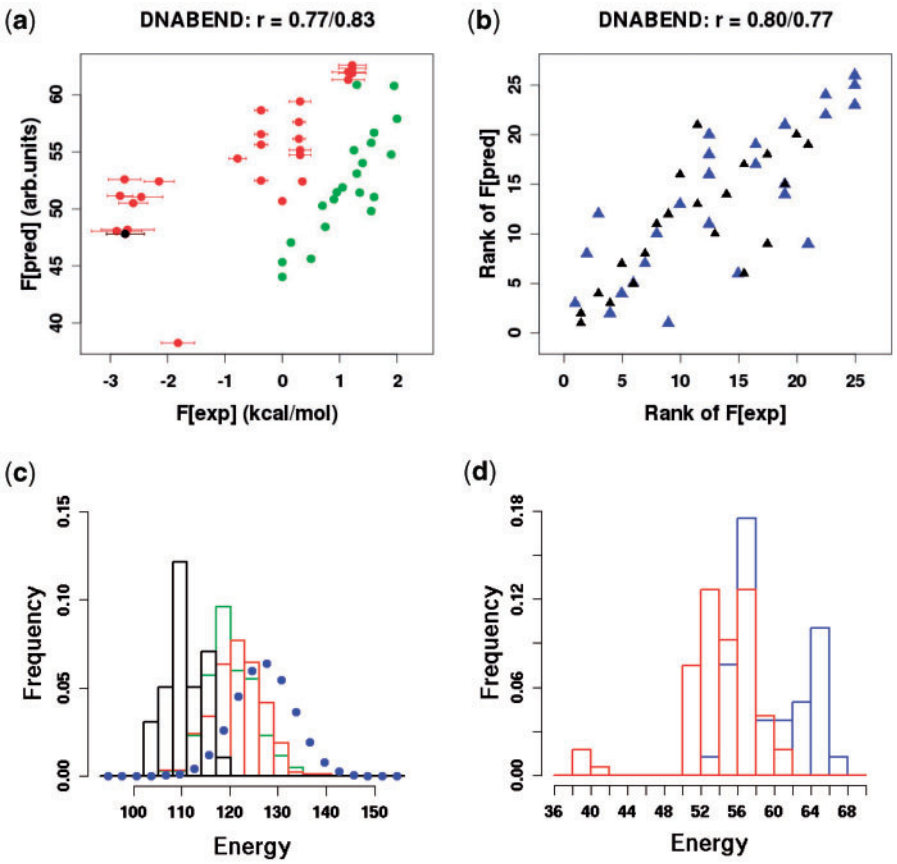


**Figure 4.** DNABEND accurately ranks free energies of nucleosome formation and sets of nucleosome sequences. (**a**) Prediction of *in vitro* free energies of nucleosome formation measured using nucleosome dialysis (red circles) (31) and nucleosome exchange (green circles) (32,33). High affinity sequence 601 (25,31) is shown in black. Free energies were computed using only the central 71 bp of the 147 bp nucleosomal site, because competitive nucleosome reconstitution on DNAs with any lengths between 71 and 147 bp gives identical apparent free energies, and quantitatively equivalent free energies are obtained using either the full histone octamer or just the core histone tetramer (5,25). (**b**) Ranking of the nucleosome free energies shown in (a). Blue triangles: nucleosome dialysis (31). Black triangles: nucleosome exchange (32,33). (**c**) Histograms of DNA elastic energies (in arbitrary units) computed using the 147 bp nucleosomal site, consistent with the sequence lengths found in the *in vitro* selection on the yeast genome. (5) Yeast genomic sequences are compared with three sets of sequences selected for their nucleosome positioning ability. Blue: energies of all 147-bp long sequences from *Saccharomyces cerevisiae* chromosome III, green: energies of sequences from a genome-wide *in vivo* mononucleosome extraction assay (5), red: energies of sequences from an *in vitro* selection assay on yeast genomic DNA (5), black: energies of sequences from a SELEX experiment on a large pool of chemically synthesized random DNA molecules (27). Sequences shorter than 147 bp were omitted from all selected sequence sets; in sequences longer than 147 bp the most favorable energy was reported, taking both forward and reverse strands into account. (**d**) Histograms of DNA elastic energies for the mouse genome sequences selected for their ability to position nucleosomes (red) (34), or to impair nucleosome formation (blue) (35). Because most of these sequences are shorter than 147 bp, it was assumed that selective pressure was exerted mainly on the central 71 bp stretch of the nucleosomal DNA which interacts with the $H3_2H4_2$ tetramer. In sequences longer than 71 bp the most favorable energy of binding with the tetramer was computed, taking both forward and reverse strands into account.

the basis of their high or low binding affinity (34,35) ($P = 1.42 \times 10^{-9}$), although there is still substantial overlap between the two sets (Figure 4d).

### Prediction of footprinted nucleosome positions

A direct test of how accurately DNABEND positions nucleosomes on DNA can be provided by a collection of sequences where *in vitro* nucleosome positions are known with 1–2 bp accuracy. We have determined nucleosome positions on synthetic high-affinity sequences 601, 603, and 605 (27) using hydroxyl radical footprinting (Supplementary Figures 2 and 3), and combined these data with three more footprinting experiments from the literature (36–38). We use DNABEND energies (computed using the full 147 bp superhelix) followed by the dynamic programming algorithm (16) to make a probabilistic prediction of nucleosome positions. Whereas with longer genomic sequences a typical configuration consists of many nonoverlapping nucleosomes, only one nucleosome can form on the shorter sequences considered here, its position typically determined by the global energy minimum (except for two sequences with two experimentally mapped alternative positions, cf. below). We compute the grand canonical partition function and the probability for the histone octamer to bind DNA starting at every possible position along the sequence. We also compute the nucleosome occupancy for each base pair, defined as its probability to be covered by any nucleosome, regardless of the nucleosome's starting position (see Methods section).

We find that DNABEND predicts footprinted nucleosome positions reasonably well: the measured position is

always within 1–2 bp of a local minimum in our energy profiles, and that energy minimum in 5 out of 6 cases is within 0.5–1.0 kcal/mol of the global energy minimum (Figure 5; note that the total range of sequence-dependent binding energies is ~5 kcal/mol). We expect nucleosome energies to be approximately equal for positions that are in phase with respect to the helical twist, and indeed consecutive energy minima and maxima in Figure 5 are separated by 10–11 bp. However, in several cases (e.g. for clone 601) where the experimental nucleosome position coincides with a local rather than a global minimum, the relatively small energy difference between these minima is sufficient to misplace the predicted nucleosome by tens of base pairs from its experimentally known location.

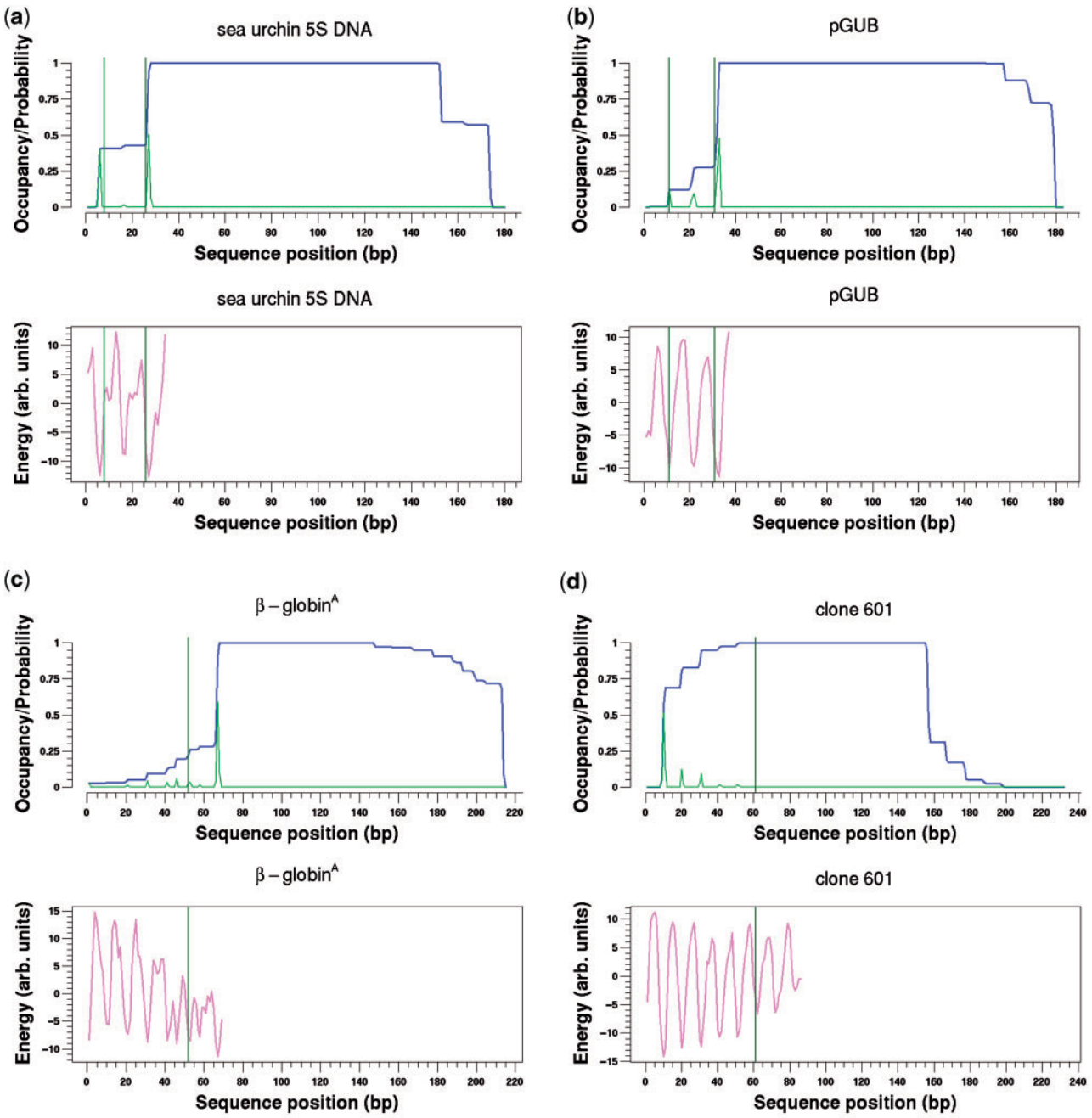


**Figure 5.** DNABEND predictions of *in vitro* nucleosome positions. Probability of a nucleosome to start at each base pair (green), nucleosome occupancy (blue) and nucleosome formation energy (violet). Vertical lines: experimentally known nucleosome starting positions, with base pair coordinates listed in parentheses below. (**a**) The 180 bp sequence from the sea urchin 5S rRNA gene (bps 8,26) (36). (**b**) The 183 bp sequence from the pGUB plasmid (bps 11,31) (37). (**c**) The 215 bp fragment from the sequence of the chicken β–globin$^A$ gene (bp 52) (38). (**d,e,f**) Synthetic high-affinity sequences (27) 601 (bp 61), 603 (bp 81) and 605 (bp 59). Nucleosomes on sequences 601, 603 and 605 were mapped by hydroxyl radical footprinting (Supplementary Figures 2 and 3). All DNA sequences used in this calculation are available on the Nucleosome Explorer web site: http://nucleosome.rockefeller.edu.

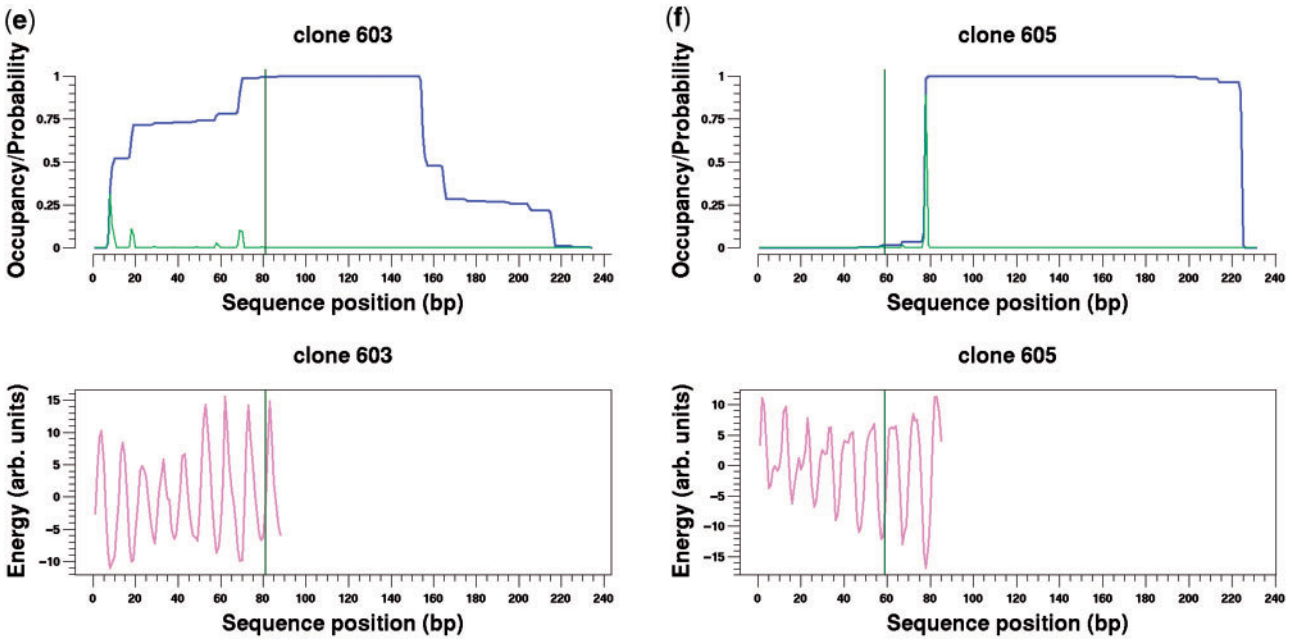


**Figure 5.** Continued.

Errors of similar magnitude are made if DNA geometry is taken from the ideal superhelix [(15); Figure 5], from the 1kx5 crystal structure [(14); Supplementary Figure 6], or if nucleosome positions are predicted using the latest bioinformatics model which was trained on sequences from nucleosomes reconstituted *in vitro* by salt dialysis on yeast genomic DNA [(9); Supplementary Figure 7]. We compare all four prediction methods in Figure 6, using the number and the height of predicted probability peaks near experimentally known positions as a metric. Our results underscore the exacting level of accuracy ($\leq$0.5 kcal/mol) required to position nucleosomes precisely on genomic or synthetic DNA. It is also possible that suboptimal nucleosome positions are in fact produced in experiments but not detected because such sub-populations would be relatively small.

## Design of nucleosomal sequences

We also asked if DNABEND could be used to design *de novo* DNA sequences with intrinsically high and low histone binding affinities. We used simulated annealing to search for 147 bp sequences whose free energy of nucleosome formation would be either minimum or maximum. The sequences with lowest free energies were created using a computationally optimized 71 bp histone tetramer binding site, because competitive nucleosome reconstitution on DNAs with any lengths between 71 and 147 bp gives identical free energies under our experimental conditions (25). The 71 bp designed site was annealed to two different fixed flanking sequences (sequences B71S1 and B71S2 in Table 1). The free energy of nucleosome formation is dominated by the computationally designed site: both predicted and experimental free energies depend on the flanking sequence very weakly (Table 1). For the worst histone tetramer binder, we have designed a 147 bp sequence

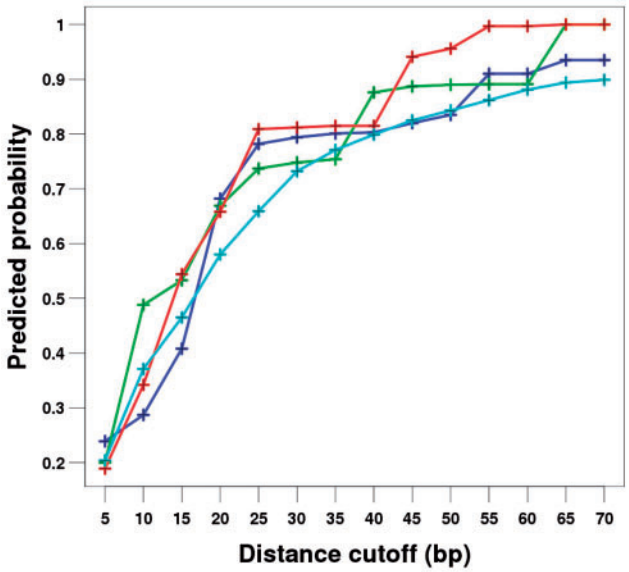


**Figure 6.** Comparison of four alternative methods for predicting nucleosome positions. For each experimentally mapped nucleosome position, we compute a sum over all predicted probabilities to start a nucleosome (green curves in Figure 5 and Supplementary Figures 5–7) that are separated by $\leq D$ bp from the experimental position. Shown are the averages over all experimental positions for a given method, as a function of $D$. Blue—DNABEND, green—DNABEND with geometries from the nucleosome crystal structure, red—DNABEND with geometries from the ideal superhelix, cyan—an alignment-based method from Kaplan *et al.* (9). Alignment-based nucleosome positioning software was downloaded from http://genie.weizmann.ac.il/pubs/nucleosomes08/ and run with default parameters. Note that the provided software does not output nucleosome energies or scores.

(W147S in Table 1) that has no favorable $H3_2H4_2$ binding sites anywhere along the sequence.

We measured *in vitro* free energies of nucleosome formation using salt gradient dialysis [(25,31); see Methods

**Table 1.** Predicted ($F_{pred}$) and measured ($F_{exp}$) free energies of computationally designed and control sequences

| | $F_{pred}$ (arb.units) | $F_{exp}$ (kcal/mol) |
|---|---|---|
| B71S1 | −20.09 | $-1.57 \pm 0.41$ |
| B71S2 | −20.10 | $-1.51 \pm 0.27$ |
| 601S1 | −0.49 | $-2.99 \pm 0.55$ |
| 601S2 | −1.74 | $-2.46 \pm 0.18$ |
| W147S | 15.26 | $0.09 \pm 0.23$ |
| X147 | 5.92 | $0.75 \pm 0.29$ |
| X146 | 6.86 | $0.45 \pm 0.91$ |

Experimental free energies are shown relative to the reference sequence from the *Lumbriculus variegatus* 5S rRNA gene. Histone binding is dominated by the contribution from the $H3_2H4_2$ tetramer with the 71 bp binding site (25). The best binder was created by using simulated annealing to introduce mutations and thus minimize the energy of a 71 bp DNA molecule. B71S1 and B71S2 have different sequences flanking the 71 bp designed site (whose contribution dominates the total free energy). 601S1 and 601S2 consist of the 71-bp site from the center of the 601 sequence (27) and flanking sequences from B71S1 and B71S2, respectively. W147S is a 147 bp sequence whose free energy (with contributions from multiple $H3_2H4_2$ binding sites) was maximized by simulated annealing. X146 and X147 are 146 and 147 bp DNA sequences from nucleosome crystal structures 1aoi (10) and 1kx5 (2).

section] for the computationally designed sequences, the 71 bp tetramer site from the 601 sequence flanked in the same way as B71S1 and B71S2, and two sequences from the nucleosome crystal structures. Although the free energy of the designed best sequence was lower than the free energy of the designed worst sequence, the free energy difference was only 1.6 kcal/mol, less than the experimentally known range of free energies (see e.g. the difference between 601S1 and X146 in Table 1). These results underscore both the ranking power and the limitations of our current DNA mechanics model.

### Periodic dinucleotide distributions in high and low energy sequences

DNABEND-selected nucleosome sequences exhibit periodic dinucleotide patterns that are consistent with those determined experimentally (27): for example, with lowest energy sequences, 5′-AA/TT-3′ and 5′-TA-3′ dinucleotide frequencies are highest in the negative roll regions (where the minor groove faces inward), while 5′-GC-3′ frequencies are shifted by ~5 bp (Supplementary Figure 8). Surprisingly, the distributions of AT and AA/TT, TA dinucleotides are in phase, despite a very low flexibility of the former (Figure 3). It is possible that rigid AT steps are used to flank and 'anchor' more flexible kinked dinucleotides. We estimate the energy difference between the best and the worst 147 bp nucleosome forming sequences to be 15.2 kcal/mol, with the energies of 95% of genomic sequences separated by <6.4 kcal/mol. This is larger than the experimentally accessible range (Figure 4) because nucleosomes cannot be forced in experiments to occupy the worst possible location on DNA, but instead find a local energy minimum with respect to the 10–11 bp helical twist.

### Nucleosome stability and gene expression levels

It is at present unclear whether nucleosome positions and stabilities are fine-tuned genome wide to achieve optimal transcriptional response. Here, we show that nucleosome stabilities may play a crucial role in gene activation and background gene expression levels of two yeast promoters: *MEL1* and *CYC1*. Specifically, DNABEND predicts that both TATA boxes in the promoter of the yeast *MEL1* (α-galactosidase) gene are occupied by a stable nucleosome, in agreement with the extremely low level of background gene expression observed in *MEL1* promoter-based reporter plasmids (39,40) (Figure 7). The nucleosome is not displaced in competition with TBP. In contrast, the TATA elements of the *CYC1* promoter were previously shown to be intrinsically accessible *in vivo* (41,42) resulting in high background expression levels. Consistent with these findings, we predict that one of the *CYC1* TATA boxes has intrinsically low nucleosome occupancy, and moreover that the nucleosome is easily displaced in competition with TBP (Figure 7).

## DISCUSSION AND CONCLUSIONS

We have developed a nucleosome model based on a combination of an empirical DNA elastic potential (11) with another potential designed to capture favorable histone–DNA interactions that bend nucleosomal DNA into a superhelix. Despite several approximations (such as neglecting direct interactions between DNA base pairs and amino acid side chains, and assuming that on average nucleosomal DNA forms an ideal superhelix), the model is reasonably successful in predicting *in vitro* free energies of nucleosome formation (Figure 4). Energy minimization is essential for this success: using static DNA geometries from the ideal superhelix or from the nucleosome crystal structure was found to be detrimental to the prediction accuracy (compare Figure 4 and Supplementary Figure 4). Presumably, this is because the DNA conformation (and in particular the pattern of kinks that facilitate superhelix formation) is strongly sequence dependent. Minimized DNA geometries for the 1kx5 DNA thus constitute an *ab initio* prediction that can be compared with the crystal structure; we find both overall correlation and significant discrepancies (Figure 2). In particular, we underestimate the magnitude of the slide peaks that are very prominent in the crystal structure. This could be fixed by constraining our geometries around the 'piecewise' ideal superhelix into which a ladder of slide steps is built in by hand. However, this approach can only be justified in our framework if the slide steps come from the direct contacts between base pairs and histone side chains and do not appear 'spontaneously' when DNA is bent. There is currently no direct evidence to support the former point of view; it is equally likely that the elastic potential parameters are inaccurate for some of the flexible dinucleotides and so improving DNA mechanics models would result in better correspondence with the crystal structure.

Despite the ability of DNABEND to rank sets of sequences selected for binding affinity (Figure 4c and d), our designs of extremely stable and unstable nucleosomes

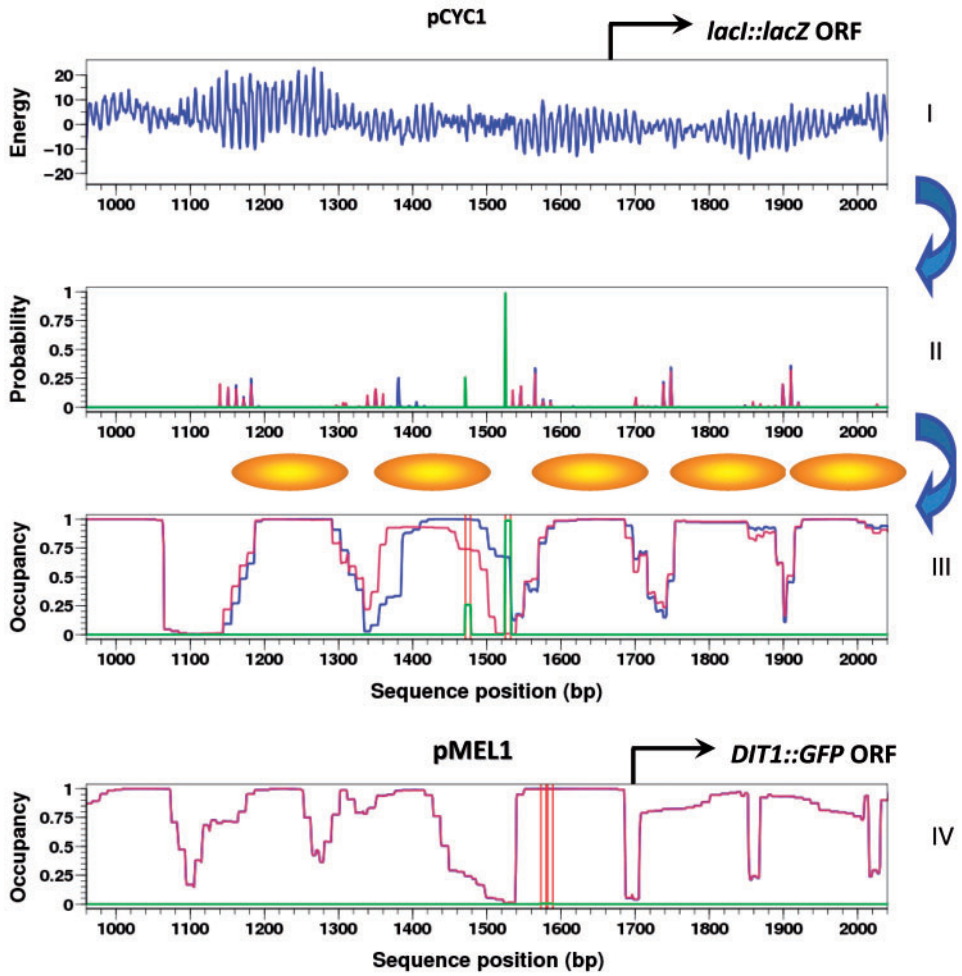


**Figure 7.** Nucleosome positioning explains background gene expression levels observed in reporter plasmids. Panel I (from top). Blue: nucleosome energies (in arbitrary units, au) in the *CYC1* promoter region from the lacI::lacZ reporter plasmid (42). Note the 10-11 bp periodicity due to DNA helical twist. Panel II. Probability of a nucleosome to start at each base pair, in the absence (blue) and presence (maroon) of TBP. Some of the latter nucleosomes are also shown as orange ovals (note that in general nucleosome positions with $P < 0.5$ may overlap). Green: probability of a TBP to bind a TATA box. Panel III. Nucleosome occupancy in the absence (blue) and presence (maroon) of TBP. Green: TBP occupancy, red vertical lines: known TATA box positions. Arrows on the right correspond to the order of calculations. Panel IV. Nucleosome occupancy of the *MEL1* promoter region from the DIT1::GFP reporter vector [(39); see *CYC1* legend above for the color scheme]. Red vertical lines: known TATA box positions. Note that blue and maroon occupancy profiles completely overlap. TBP DNA-binding energies were computed as weight matrix log scores. The weight matrix was constructed using the alignment of TATA box sites from Basehoar *et al.* (43). From left to right the TBP binding energies were set to −5.819 au (TATATATA site) and −5.327 au (TATATAAA site) for *CYC1*, and to −4.726 au for both *MEL1* sites (TATAAAAA).

were only partially successful: although free energies measurements confirmed our computational ranking of designed sequences (with the free energy difference of 1.6 kcal/mol), nucleosomes with free energies that are both lower and higher compared with our designed sequences are already available in the literature (Table 1). Evidently, the accuracy of our model is not sufficient to handle these extreme cases. Finally, our nucleosome footprinting experiments reveal the exacting level of accuracy required for predicting *in vitro* nucleosome positions: even a small discrepancy on the order of 0.5 kcal/mol may result in the positioning error of 10–20 bp in the absence of steric exclusion. This is true both for DNABEND, for the models that employ fixed DNA geometries (14,15), and for the alignment-based bioinformatics model [(9); Figures 5 and 6; Supplementary Figures 5–7]. This problem may be alleviated *in vivo* where formation of nucleosome arrays is guided as much by steric exclusion and other factors as by sequence specificity.

DNABEND presents a useful biophysical framework for the analysis of *in vivo* and *in vitro* nucleosome positions and TF-nucleosome competition. *In vivo* chromatin structure is affected by intrinsic sequence preferences, steric exclusion and extrinsic factors such as nonhistone DNA-binding proteins and chromatin remodeling enzymes. Our approach helps disentagle these contributions to *in vivo* nucleosome positioning, and should provide a useful foundation for future models of chromatin. In particular, our results linking nucleosome stability with gene expression in *MEL1* and *CYC1* promoters (Figure 7) open a pathway towards modulating gene expression levels in model systems through computational

design of nucleosome occupancy profiles. Finally, unlike previously published bioinformatics approaches, our model is not trained on genomic data and thus should be equally applicable to other eukaryotic organisms, including longer metazoan genomes.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

J.W. acknowledges the use of instruments in the Keck Biophysics Facility at Northwestern University. We thank Eran Segal for many useful discussions and for sharing his data and software prior to publication.

## FUNDING

Alfred P. Sloan Research Fellowship (to A.V.M.); National Science Foundation (DMR-0129848 to E.D.S., 0549593 to V.M.S.); National Institutes of Health (R01 GM054692 and R01 GM058617 to J.W., R01 GM58650 to V.M.S., R01 HG004708 to A.V.M.). Funding for open access charge: National Institutes of Health.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Khorasanizadeh,S. (2004) The nucleosome: from genomic organization to genomic regulation. *Cell*, **116**, 259–272.
2. Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
3. Boeger,H., Griesenbeck,J., Strattan,J.S. and Kornberg,R.D. (2003) Nucleosomes unfold completely at a transcriptionally active promoter. *Mol. Cell*, **11**, 1587–1598.
4. Wallrath,L.L., Lu,Q., Granok,H. and Elgin,S.C.R. (1994) Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. *Bioessays*, **16**, 165–170.
5. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thastrom,A., Field,Y., Moore,I.K., Wang,J.Z. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
6. Ioshikhes,I.P., Albert,I., Zanton,S.J. and Pugh,B.F. (2006) Nucleosome positions predicted through comparative genomics. *Nat. Genet.*, **38**, 1210–1215.
7. Peckham,H.E., Thurman,R.E., Fu,Y., Stamatoyannopoulos,J.A., Noble,W.S., Struhl,K. and Weng,Z. (2007) Nucleosome positioning signals in genomic DNA. *Genome Res.*, **17**, 1170–1177.
8. Yuan,G. and Liu,J.S. (2008) Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, **4**, 0164–0174.
9. Kaplan,N., Moore,I.K., Fondufe-Mittendorf,Y., Gossett,A.J., Tillo,D., Field,Y., LeProust,E.M., Hughes,T.R., Lieb,J.D., Widom,J. *et al*. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
10. Luger,K., Mäder,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
11. Olson,W.K., Gorin,A.A., Lu,X., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
12. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
13. Arents,G. and Moudrianakis,E.N. (1993) Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA. *Proc. Natl Acad. Sci. USA*, **90**, 10489–10493.
14. Tolstorukov,M.Y., Colasanti,A.V., McCandlish,D.M., Olson,W.K. and Zhurkin,V.B. (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**, 725–738.
15. Miele,V., Vaillant,C., d'Aubenton-Carafa,Y., Thermes,C. and Grange,T. (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.
16. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, MA.
17. Lu,X. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structure. *Nucleic Acids Res.*, **31**, 5108–5121.
18. El Hassan,M.A. and Calladine,C.R. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *J. Mol. Biol.*, **251**, 648–664.
19. Lu,X., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: analysis program (SCHNAaP). *J. Mol. Biol.*, **273**, 668–680.
20. Lu,X., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: rebuilding program (SCHNArP). *J. Mol. Biol.*, **273**, 681–691.
21. Goldstein,H. (1980) *Classical Mechanics* 2nd edn. Addison-Wesley Publishing Company, Reading, MA.
22. Bouchiat,C., Wang,M.D., Allemand,J.-F., Strick,T., Block,S.M. and Croquette,V. (1999) Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophys. J.*, **76**, 409–413.
23. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet.*, **39**, 1235–1244.
24. Thastrom,A., Lowary,P.T. and Widom,J. (2004) Measurement of histone-DNA interaction free energy in nucleosomes. *Methods*, **33**, 33–44.
25. Thastrom,A., Bingham,L.M. and Widom,J. (2004) Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.*, **338**, 695–709.
26. Dyer,P.N., Edayathumangalam,R.S., White,C.L., Bao,Y., Chakravarthy,S., Muthurajan,U.M. and Luger,K. (2004) Reconstitution of nucleosome core particles from recombinant histones and DNA. *Methods Enzymol.*, **375**, 23–44.
27. Lowary,P.T. and Widom,J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **276**, 19–42.
28. Walter,W. and Studitsky,V.M. (2004) Construction, analysis, and transcription of model nucleosomal templates. *Methods*, **33**, 18–24.
29. Tullius,T.D., Dombroski,B.A., Churchill,M.E. and Kam,L. (1987) Hydroxyl radical footprinting: a high-resolution method for mapping protein-DNA contacts. *Methods Enzymol.*, **155**, 537–558.
30. Widom,J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.*, **34**, 269–324.
31. Thastrom,A., Lowary,P.T., Widlund,H.R., Cao,H., Kubista,M. and Widom,J. (1999) Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.*, **288**, 213–229.
32. Shrader,T.E. and Crothers,D.M. (1989) Artificial nucleosome positioning sequences. *Proc. Natl Acad. Sci. USA*, **86**, 7418–7422.
33. Shrader,T.E. and Crothers,D.M. (1990) Effects of DNA sequence and histone-histone interactions on nucleosome placement. *J. Mol. Biol.*, **216**, 69–84.
34. Widlund,H.R., Cao,H., Simonsson,S., Magnusson,E., Simonsson,T., Nielsen,P.E., Kahn,J.D., Crothers,D.M. and Kubista,M. (1998) Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.*, **267**, 807–817.
35. Cao,H., Widlund,H.R., Simonsson,T. and Kubista,M. (1998) TGGA repeats impair nucleosome formation. *J. Mol. Biol.*, **281**, 253–260.
36. Flaus,A., Luger,K., Tan,S. and Richmond,T.J. (1996) Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals. *Proc. Natl Acad. Sci. USA*, **93**, 1370–1375.

37. Kassabov,S.R., Henry,N.M., Zofall,M., Tsukiyama,T. and Bartholomew,B. (2002) High-resolution mapping of changes in histone-DNA contacts of nucleosomes remodeled by ISW2. *Mol. Cell Biol.*, **22**, 7524–7534.
38. Davey,C.S., Pennings,S., Reilly,C., Meehan,R.R. and Allan,J. (2004) A determining influence for CpG dinucleotides on nucleosome positioning *in vitro*. *Nucleic Acids Res.*, **32**, 4322–4331.
39. Ligr,M., Siddharthan,R., Cross,F.R. and Siggia,E.D. (2006) Gene expression from random libraries of yeast promoters. *Genetics*, **172**, 2113–2122.
40. Melcher,K., Sharma,B., Ding,W.V. and Nolden,M. (2000) Zero background yeast reporter plasmids. *Gene*, **247**, 53–61.
41. Kuras,L. and Struhl,K. (1999) Binding of TBP to promoters *in vivo* is stimulated by activators and requires Pol II holoenzyme. *Nature*, **399**, 609–613.
42. Chen,J., Ding,M. and Pederson,D.S. (1994) Binding of TFIID to the CYC1 TATA boxes in yeast occurs independently of upstream activating sequences. *Proc. Natl Acad. Sci. USA*, **91**, 11909–11913.
43. Basehoar,A.D., Zanton,S.J. and Pugh,B.F. (2004) Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, **116**, 699–709.