Supplementary Material for:

High-throughput sequencing reveals a simple model of nucleosome energetics

George Locke[†], Denis Tolkunov[†], Zarmik Moqtaderi[‡], Kevin Struhl[‡] and Alexandre V. Morozov[†*]

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

[‡]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School Boston, MA 02115

[*]Corresponding author; email: morozov@physics.rutgers.edu; phone: 732-445-1387; fax: 732-445-5958

1 Supplementary Methods

1.1 Preprocessing high-throughput sequencing data

Mapping sequence read profiles.

We start from a collection of 25 bp-long Solexa sequence reads uniquely mapped onto the yeast genome with no more than two mismatches.[1] Each read is mapped onto either the forward (5') or the reverse (3') strand. For sequence reads mapped onto the forward (5') strand, we interpret the first base of a read as the start position of a nucleosome with the canonical length of 147 bp. For sequence reads mapped onto the reverse (3') strand, we interpret the last base of the read as the end position of a 147 bp nucleosome. Thus we create a "sequence read profile", a table which shows the number of nucleosomes starting at each genomic bp. This table is used to create a "read coverage profile" which shows how many nucleosomes cover each genomic bp.

Filtering sequence read profiles.

We observe that there are large gaps in our read profiles, possibly due to repetitive regions in the genome to which reads cannot be mapped uniquely, or to sequencing artifacts. We considered any stretch of ≥ 1000 bp without mapped reads to be anomalous and excluded such regions from further analysis. We also find regions where the read coverage was uncharacteristically high. For instance, our *in vitro* nucleosome measurement for chromosome 12 has an average nucleosome coverage of ~ 80 reads, but there is a small region near bp 460000 covered with 5000 reads. We exclude such regions according to the following procedure: For each chromosome, we find the average number of reads per bp. Next, for each bp we calculate the running

average number of reads in a window extending 75 bp in each direction. If this running average is more than three times the mean, we flag the region which extends out from the identified point in both directions until the running average equals the mean, and we remove this region from consideration. We then create a filter which marks the union of all excluded regions. Finally, each excluded region is extended 146 bp upstream so that there is no contribution to the nucleosome energy from filtered regions.

Normalizing sequence read profiles.

Next we use the sequence read profile to create nucleosome probability and occupancy profiles. First, we set sequence read counts to zero inside all filtered regions. Second, we use a Gaussian smoothing algorithm that replaces the number of sequence reads at a given bp with a normal distribution centered at that bp. The Gaussian is chosen to have $\sigma = 2$ or 20 depending on subsequent modeling, and the area under the curve is equal to the number of sequence reads at that bp. The smoothed sequence read profile is then constructed as a superimposition of all such Gaussians.

The smoothing procedure reflects a lack of bp precision in MNase digestion assays, which results in the uncertainty of the interpretation of sequence read coordinates as nucleosome start or end positions. In addition, because neighboring nucleosomes are expected to have similar binding affinities, collecting more sequence read data is assumed to result in a read profile that we approximate with the superposition of normal distributions centered on available reads.

We extend the smoothed read profile into a smoothed read coverage profile as described above, find the highest point N_{max} in the smoothed coverage profile and multiply the height of each point in the smoothed coverage profile and the smoothed read profile by $1/N_{max}$ so that the maximum coverage is one. Each point in the smoothed sequence read profile may now be interpreted as the probability for a nucleosome to start at a given position, and the coverage may be interpreted as the probability for any nucleosome to occupy a given position. We refer to the scaled results as nucleosome probability and occupancy profiles, respectively.

1.2 Energetics of DNA-binding one-dimensional particles of finite size

Consider particles of size a bp distributed along a DNA segment of length L bp. The particles can interact with DNA in a position-dependent manner and are also subject to steric exclusion (adjacent particles cannot overlap). A grand-canonical partition function for this system of DNA-bound particles is given by:

$$Z = \sum_{conf} e^{-[E(conf)-\mu N(conf)]}, \tag{1}$$

where $conf$ denotes an arbitrary configuration of DNA-bound non-overlapping particles, μ is the chemical potential, and $E(conf)$ and $N(conf)$ are the total DNA-binding energy and the number of particles in the current configuration (for simplicity we assume $k_BT = 1$, where k_B is the Boltzmann constant and T is the room temperature).

One can compute Z efficiently using a recursive relation:[2]

$$
\begin{aligned}
&Z_i^f = Z_{i-1}^f + Z_{i-a}^f e^{-(E_{i-a+1}-\mu)}, \quad i = a,\ldots,L \\
&Z_{a-1}^f = \cdots = Z_0^f = 1
\end{aligned}
\tag{2}
$$

which computes a set of partial partition functions in the forward direction. Likewise, partial partition functions can be computed in the reverse direction:

$$
\begin{aligned}
&Z_i^r = Z_{i+1}^r + Z_{i+a}^r e^{-(E_i-\mu)}, \quad i = L-a+1,\ldots,1 \\
&Z_{L-a+2}^r = \cdots = Z_{L+1}^r = 1
\end{aligned}
\tag{3}
$$

Note that $Z_L^f = Z_1^r = Z$ by construction. Furthermore, the probability of starting a particle at position i is given by:

$$
P_i = \frac{Z_{i-1}^f e^{-(E_i-\mu)} Z_{i+a}^r}{Z}, \quad i = 1,\ldots,L-a+1 \tag{4}
$$

Intuitively, Eq. (4) is a partition function for all configurations in which a particle is bound at position i (occupying positions i through $i+a-1$), divided by the partition function for all possible configurations. Using Eqs. (2), (3) and (4) we obtain:

$$
\begin{aligned}
&Z_i^f - Z_{i-1}^f = P_{i-a+1} Z / Z_{i+1}^r, \qquad i = a,\ldots,L \\
&Z_{i+1}^r - Z_i^r = -P_i Z / Z_{i-1}^f, \qquad i = L-a+1,\ldots,1
\end{aligned}
\tag{5}
$$

Note that both of these formulas can be extended to the $i = 1,\ldots,L$ range if we assume that $P_k = 0, \;\; k \notin [1, L-a+1]$. It is easy to show that $Z_i^f Z_{i+1}^r - Z_{i-1}^f Z_i^r = Z(P_{i-a+1} - P_i)$. This expression has the form of a complete differential and thus can be iterated as follows:

$$
Z_L^f Z_{L+1}^r - Z_{i-1}^f Z_i^r = Z \sum_{j=i}^{L} (P_{j-a+1} - P_j), \tag{6}
$$

yielding

$$
Z_{i-1}^f Z_i^r = Z\left(1 - \sum_{j=i-a+1}^{i-1} P_j\right), \quad i = 1,\ldots,L \tag{7}
$$

Using Eqs. (3), (4) and (7) we get:

$$
Z_{i+1}^r = Z_i^r \left(1 - \frac{P_i}{1 - \sum_{j=i-a+1}^{i-1} P_j}\right). \tag{8}
$$

Introducing $O_i = \sum_{j=i-a+1}^{i} P_j$ - the probability that position i is covered by any particle regardless of its starting position (also called the particle occupancy), we see that:

$$
Z_{i+1}^r = Z_i^r \left(1 - \frac{P_i}{1 - O_i + P_i}\right). \tag{9}
$$

Using Eq. (9) recursively (until $Z^r_{L+1} = 1$ is reached on the left-hand side), we obtain an explicit expression for Z^r_i:

$$Z^r_i = \prod_{j=i}^{L} (1 - \frac{P_j}{1 - O_j + P_j})^{-1}, \quad i = 1, \ldots, L \tag{10}$$

Likewise, using Eqs. (2), (4) and (7) together with $Z^f_0 = 1$ we get:

$$Z^f_i = \prod_{j=1}^{i} (1 - \frac{P_{j-a+1}}{1 - O_j + P_{j-a+1}})^{-1}, \quad i = 1, \ldots, L \tag{11}$$

Eqs. (10) and (11) are explicit expressions for forward and reverse partial partition functions in terms of particle probabilities and occupancies. Note that $Z^r_1 = Z^f_L = Z$ still holds, with Eqs. (10) and (11) providing alternative expressions for the partition function in this limit. Inserting Eqs. (10) and (11) into Eq. (4) and using Eq. (7) to express Z^f_{i-1} in terms of Z^r_i leads to the desired expression for the DNA-binding energy of the particle at position i:

$$E_i - \mu = \log \frac{1 - O_i + P_i}{P_i} + \sum_{j=i}^{i+a-1} \log \frac{1 - O_j}{1 - O_j + P_j}, \quad i = 1, \ldots, L - a + 1 \tag{12}$$

Alternatively, we can use Eq. (7) to express Z^r_{i+a} in terms of Z^f_{i+a-1}, leading to an equivalent expression for the DNA-binding energy:

$$E_i - \mu = \log \frac{1 - O_{i+a-1} + P_i}{P_i} + \sum_{j=i-a+1}^{i} \log \frac{1 - O_{j+a-1}}{1 - O_{j+a-1} + P_j}, \quad i = 1, \ldots, L - a + 1 \tag{13}$$

1.3 Hierarchical models of nucleosome energetics

We have created hierarchical models of nucleosome energetics which assign non-zero energies to nucleotide motifs of length N only if the nucleosome energies cannot be explained using nucleotide motifs of lengths $1 \ldots N-1$. This is implemented using constraints on word energies:

$$\sum_{\alpha_i} \varepsilon_{\alpha_1 \ldots \alpha_N} = 0, \ \ \forall i = 1 \ldots N \tag{14}$$

Here $\varepsilon_{\alpha_1 \ldots \alpha_N}$ is the energy of the word of length N with nucleotides $\alpha_1 \ldots \alpha_N$ at positions $1 \ldots N$.

With these constraints and the $\{A, C, G, T\}$ alphabet there are 3^N independent parameters describing energetics of words of length N. For example, for $N = 1$ we can choose $\{\varepsilon_A, \varepsilon_G, \varepsilon_T\}$ to be independent, while ε_C is fixed by the constraint: $\varepsilon_C = -(\varepsilon_A + \varepsilon_G + \varepsilon_T)$. For $N = 2$ there are 9 independent parameters: $\{\varepsilon_{AA}, \varepsilon_{AG}, \varepsilon_{AT}, \varepsilon_{GA}, \varepsilon_{GG}, \varepsilon_{GT}, \varepsilon_{TA}, \varepsilon_{TG}, \varepsilon_{TT}\}$, while the other 7 dinucleotide energies can be expressed through these using Eq. (14). The remaining 7 degrees of freedom are described by the lower order terms: 6 ε_α's (3 for each position in the dinucleotide) and the total offset ε^0.

In general, D^N degrees of freedom associated with words of length N drawn from an alphabet of size D can be described using constrained energies:

$$D^N = (D-1)^N + \binom{N}{1}(D-1)^{N-1} + \cdots + \binom{N}{N}(D-1)^0, \tag{15}$$

where each term describes the total number of constrained energies of order $(N,\ldots,0)$, computed as a product of the number of constrained energies at each possible position within the longer word, and the number of such positions. Note that the zeroth order term is simply the total offset ε^0. Furthermore, shorter words comprised of non-consecutive nucleotides are included in the expansion. If we set the energies of all non-consecutive words to zero, the total energy of a word of length N can be written as:

$$\varepsilon'_{\alpha_1\ldots\alpha_N} = \sum_{n=1}^{N}\sum_{j=1}^{N-n+1} \varepsilon_{\alpha_j\ldots\alpha_{j+n-1}} + \varepsilon^0 \tag{16}$$

Note that here and in Section 1.4 below we set $\mu = 0$ for simplicity. Although a set of constrained energies of order $0,\ldots,N$ on the right-hand side of Eq. (16) has fewer degrees of freedom than a set of unconstrained energies of order N, it provides the most complete description involving consecutive nucleotide motifs, and forms a basis of nucleosome models that have been further simplified by equating energies of motifs that occur at different positions within the nucleosomal site. Furthermore, since dinucleotides are too short to contain partial non-consecutive motifs, Eq. (16) entails no loss of degrees of freedom for $N = 2$.

1.4 Sequence-specific models of nucleosome energetics

Eq. (12) can be used to convert nucleosome probabilities and occupancies obtained from high-throughput sequencing data into histone-DNA interaction energies for each position i along the DNA, under the assumption that steric exclusion and specific interactions with DNA are the only factors that affect nucleosome positions *in vitro*. In order to understand which DNA sequence features explain the observed energy profile, we carried out linear fits of genome-wide Percus energies (Eq. (12)) to four sequence-specific models. Some models were designed to focus on the $\sim 10-11$ bp periodic distributions of sequence motifs, while others capture nucleosome-wide sequence signals such as motif enrichment and depletion in nucleosome-covered sequences.

Spatially resolved model.

In terms of unconstrained energies, the spatially resolved model is defined as:

$$E(S) = \sum_{i=I_1}^{I_2-1} \varepsilon'_{\alpha_i\alpha_{i+1}}, \tag{17}$$

where $E(S)$ is the nucleosome formation energy of a 147 bp-long sequence S, $\varepsilon_{\alpha_i\alpha_{i+1}}$ is the energy of the dinucleotide with bases α_i and α_{i+1} at positions i and $i+1$ respectively, and the

sum runs from $I_1 \geq 1$ to $I_2 \leq 147$ in the nucleosomal site. To minimize edge effects, we typically exclude 3 bps from each end of the nucleosome, setting $I_1 = 4$ and $I_2 = 144$.

Eq. (17) can be rewritten as:

$$E(S) = \sum_{i=I_1}^{I_2-1} \left(\varepsilon_{\alpha_i \alpha_{i+1}} + \bar{b}_{\alpha_i} + b_{\alpha_{i+1}} \right) + \varepsilon^0, \tag{18}$$

where

$$\begin{aligned} \varepsilon^0 &= \frac{1}{D^2} \sum_{i=I_1}^{I_2-1} \sum_{\alpha,\beta=1}^{D} \varepsilon'_{\alpha\beta} \equiv \sum_{i=I_1}^{I_2-1} \varepsilon^0_{i,i+1}, \\ \bar{b}_\alpha &= \frac{1}{D} \sum_{\beta=1}^{D} \left(\varepsilon'_{\alpha\beta} - \varepsilon^0_{i,i+1} \right), \\ b_\beta &= \frac{1}{D} \sum_{\alpha=1}^{D} \left(\varepsilon'_{\alpha\beta} - \varepsilon^0_{i,i+1} \right). \end{aligned}$$

Note that $\sum_{\alpha=1}^{D} \varepsilon_{\alpha\beta} = \sum_{\beta=1}^{D} \varepsilon_{\alpha\beta} = 0$ by construction. Eq. (18) is equivalent to the expansion in terms of constrained energies which is consistent with Eq. (16):

$$E(S) = \sum_{i=I_1}^{I_2-1} \varepsilon_{\alpha_i \alpha_{i+1}} + \sum_{i=I_1}^{I_2} \varepsilon_{\alpha_i} + \varepsilon^0, \tag{19}$$

where $\varepsilon_{\alpha_{I_1}} = \bar{b}_{\alpha_{I_1}}$, $\varepsilon_{\alpha_{I_1+1}} = \bar{b}_{\alpha_{I_1+1}} + b_{\alpha_{I_1+1}}, \ldots, \varepsilon_{\alpha_{I_2}} = b_{\alpha_{I_2}}$. Thus an unconstrained description of nucleosome energetics can be uniquely decomposed into a constrained description. However, the opposite is not true: for any p and q such that $p + q = 1$

$$\begin{aligned} \varepsilon'_{\alpha_{I_1}\alpha_{I_1+1}} &= \varepsilon_{\alpha_{I_1}\alpha_{I_1+1}} + \varepsilon_{\alpha_{I_1}} + q\varepsilon_{\alpha_{I_1+1}}, \\ \varepsilon'_{\alpha_i\alpha_{i+1}} &= \varepsilon_{\alpha_i\alpha_{i+1}} + p\varepsilon_{\alpha_i} + q\varepsilon_{\alpha_{i+1}}, \;\; I_1 < i < I_2 - 1 \\ \varepsilon'_{\alpha_{I_2-1}\alpha_{I_2}} &= \varepsilon_{\alpha_{I_2-1}\alpha_{I_2}} + p\varepsilon_{\alpha_{I_2-1}} + \varepsilon_{\alpha_{I_2}} \end{aligned}$$

are equally valid reconstructions that leave $E(S)$ unchanged. In this paper we use $p = 1, q = 0$ to compute unconstrained dinucleotide energies from constrained ones.

Position-independent model.

This model assigns the same energy to a given word within the nucleosome, regardless of its position in the site. Thus the position-independent model of order N is given by:

$$E(S) = \sum_{n=1}^{N} \sum_{\{\alpha_1 \ldots \alpha_n\}}^{3^n} n_{\alpha_1 \ldots \alpha_n} \varepsilon_{\alpha_1 \ldots \alpha_n} + \varepsilon^0, \tag{20}$$

where the outer sum is over word lengths, the inner sum is over all words of length n corresponding to constrained energies, $n_{\alpha_1 \ldots \alpha_n}$ is the number of words with the nucleotides $\alpha_1 \ldots \alpha_n$ at positions $1 \ldots n$, and $\varepsilon_{\alpha_1 \ldots \alpha_n}$ are word energies constrained by Eq. (14). As in the spatially

resolved model, the words are counted from bp $I_1 = 4$ to bp $I_2 = 144$, excluding 3 bp from each end of the site. The words are not allowed to extend outside this region. Note that both in this model and in the two partially position-dependent models described below there is no one-to-one correspondence between constrained models utilizing words of order $1 \ldots N$ and their unconstrained counterparts utilizing words of order N - the former require fewer fitting parameters.

Three-region model.

This model refines the position-independent model by dividing the 141 bp nucleosome site into 3 regions of equal length. Word energies are fitted separately inside each region. The total energy of sequence S is then given by:

$$E(S) = \sum_{r=1}^{3} \sum_{n=1}^{N} \sum_{\{\alpha_1 \ldots \alpha_n\}}^{3^n} n^r_{\alpha_1 \ldots \alpha_n} \varepsilon^r_{\alpha_1 \ldots \alpha_n} + \varepsilon^0, \tag{21}$$

where r refers to a particular 47 bp region.

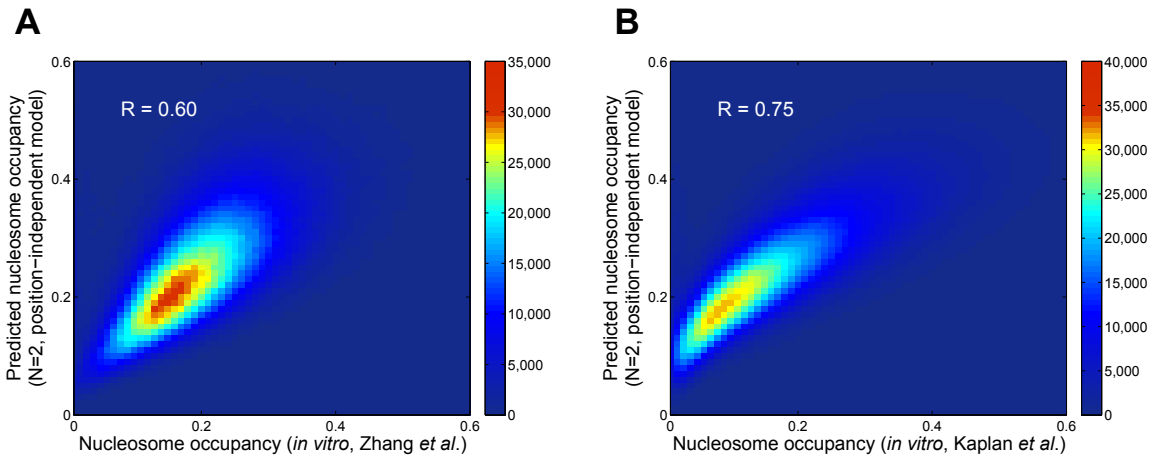
Periodic model.

This model enforces DNA helical twist periodicity by equating the energies of words separated by a multiple of 10 bp. To reduce the number of fitting parameters, we also grouped energies of words at positions $1 \ldots 10$ into 5 distinct bins. Thus an AGT motif starting at position 1 within the nucleosome site would have the same energy as the AGT motif starting at positions $11, 21, 31 \ldots$ as well as positions $2, 12, 22 \ldots$, whereas the energy of the same motif starting at positions 3 and 4 is grouped into a different bin. The total energy is then computed as:

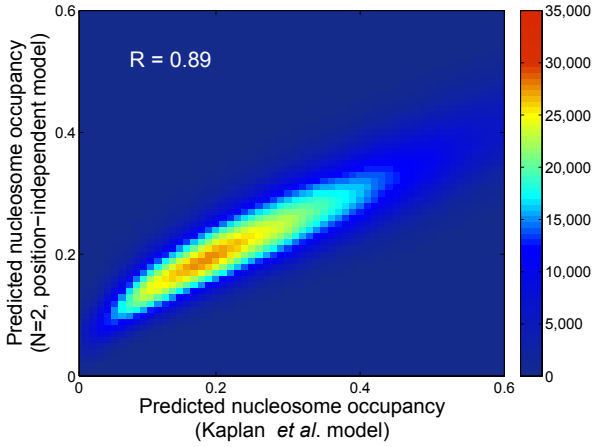
$$E(S) = \sum_{b=1}^{5} \sum_{n=1}^{N} \sum_{\{\alpha_1 \ldots \alpha_n\}}^{3^n} n^b_{\alpha_1 \ldots \alpha_n} \varepsilon^r_{\alpha_1 \ldots \alpha_n} + \varepsilon^0, \tag{22}$$

where b is the bin index used to group motifs separated by the helical twist as described above. As before, all words overlapping with the 3 bp edge regions are excluded from the counts.

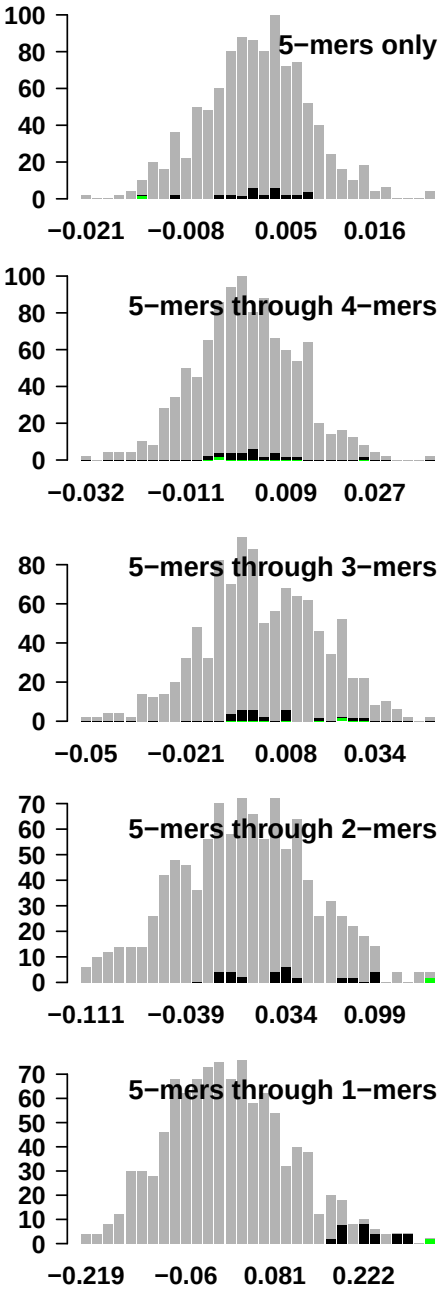
Supplementary Figures



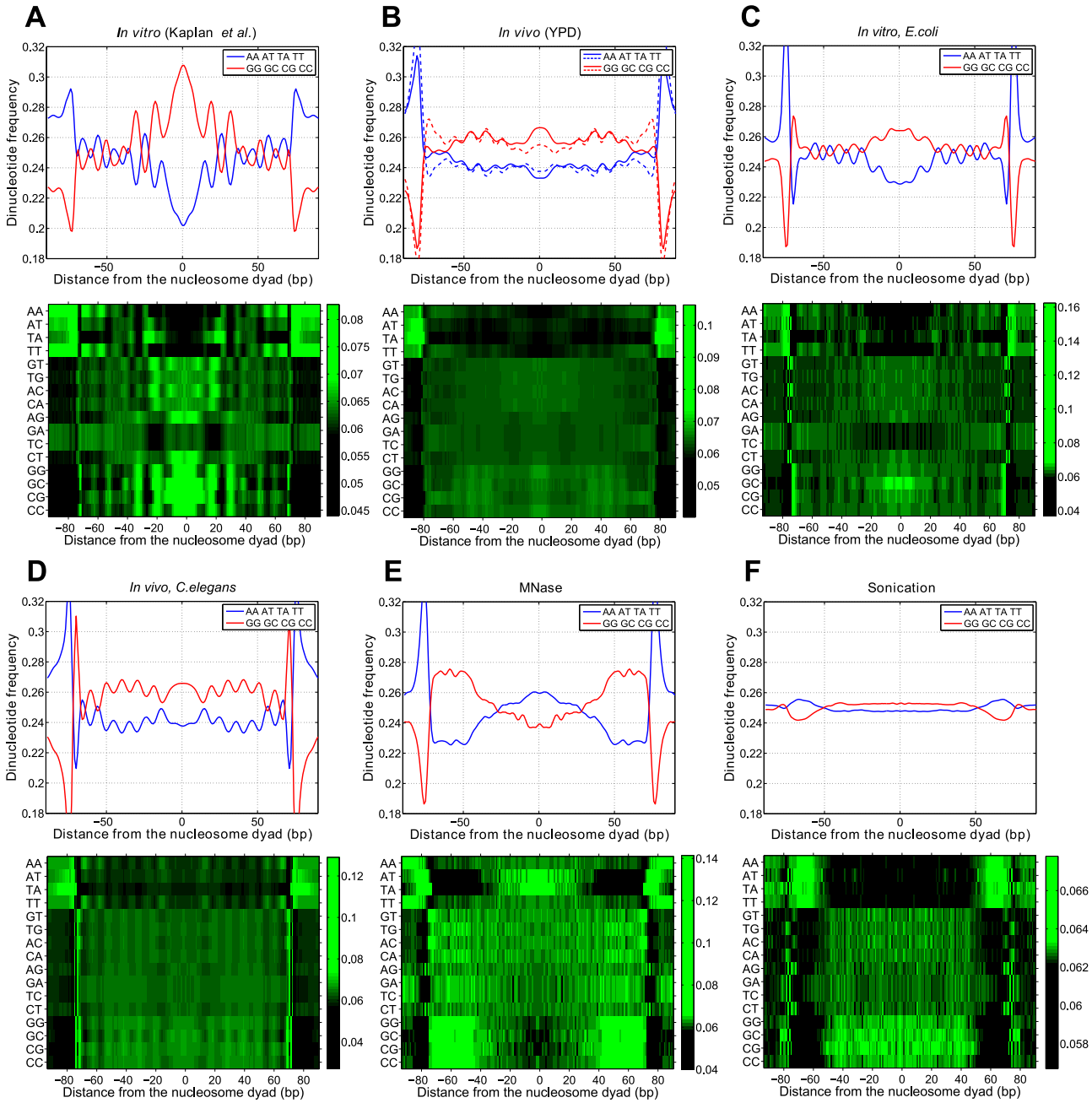
Supplementary Figure 1. $N = 2$ **position-independent model is sufficient to explain nucleosome occupancy in *S.cerevisiae*.** a) Density scatter plot for the nucleosome occupancy at each genomic base pair (predicted with the $N = 2$ position-independent model) vs. *in vitro* occupancy observed by Zhang *et al.*[1] b) Same as (a) except that *in vitro* occupancy is from Kaplan *et al.*[3]



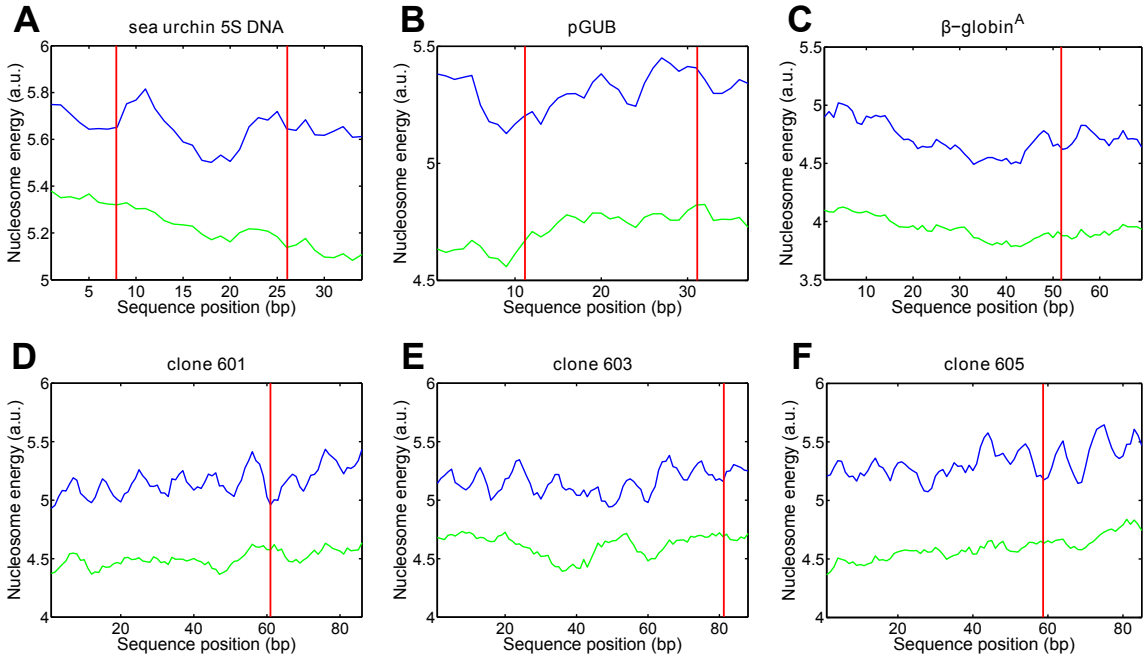
Supplementary Figure 2. Similar predictive power of the $N = 2$ position-independent model and a bioinformatics model based on periodic dinucleotide distributions and frequencies of 5 bp-long words.[3] Density scatter plot for the nucleosome occupancy at each genomic base pair (predicted with the $N = 2$ position-independent model) vs. nucleosome occupancy predicted by Kaplan *et al.*[3]



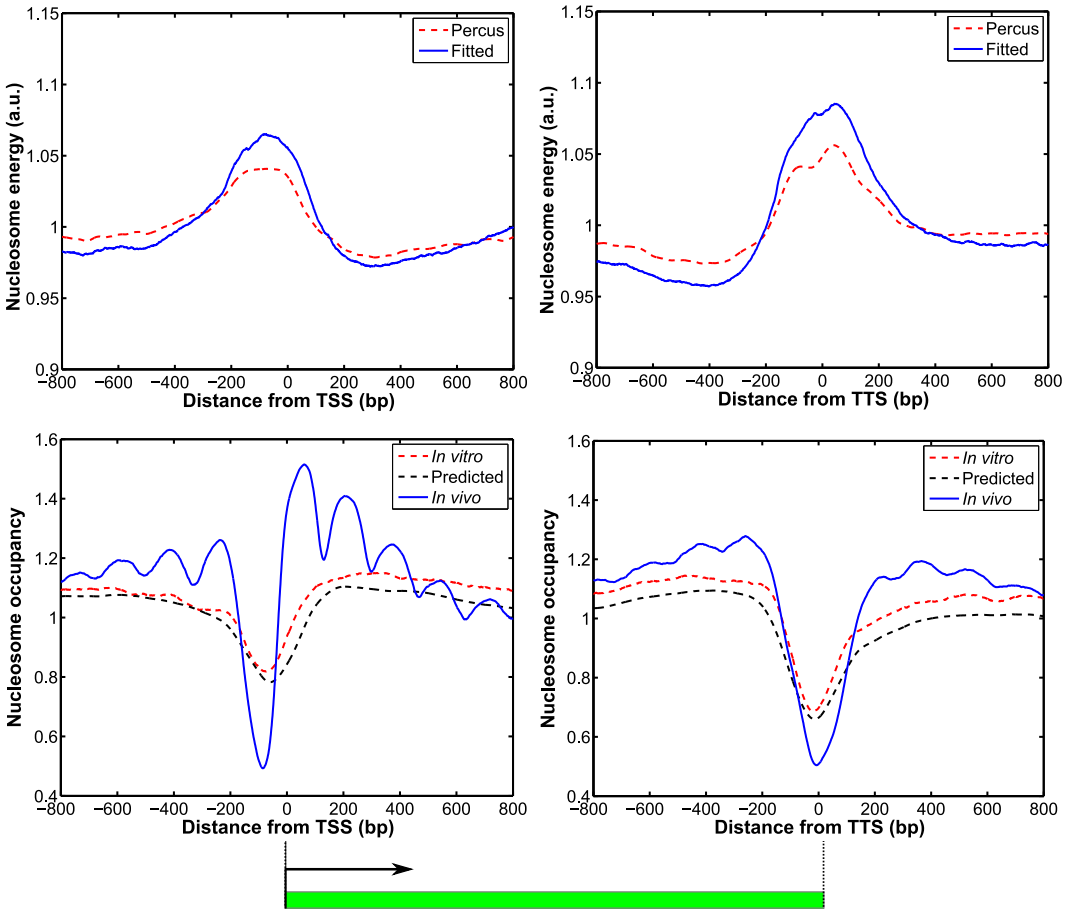
Supplementary Figure 3. Minor role of the higher-order contributions to the energies of 5 bp-long words. $N = 5$ position-independent model was trained on nucleosomes reconstituted *in vitro* on the yeast genome,[1] yielding energies of all motifs of 1 through 5 bp in length. Energies of 5 bp-long words were then computed by summing contributions from a subset of shorter motifs: $E(S) = \sum_{n=L}^{5} \sum_{\{\alpha_1 \ldots \alpha_n\}}^{4^n} n_{\alpha_1 \ldots \alpha_n} \varepsilon_{\alpha_1 \ldots \alpha_n}$, where $n_{\alpha_1 \ldots \alpha_n}$ is the number of times a given word was found in the 5 bp-long sequence S and $\varepsilon_{\alpha_1 \ldots \alpha_n}$ is the fitted energy of that word. $L = 5 \ldots 1$ is the length of the shortest motif included into $E(S)$. Grey: all 5 bp-long words, black: A:T-containing words, green: the poly(dA:dT) tract (AAAAA).


A
In vitro (Kaplan et al.)
B
In vivo (YPD)
C
In vitro, E.coli
D
In vivo, C.elegans
E
MNase
F
Sonication
AA AT TA TT
GG GC CG CC
Dinucleotide frequency
Distance from the nucleosome dyad (bp)

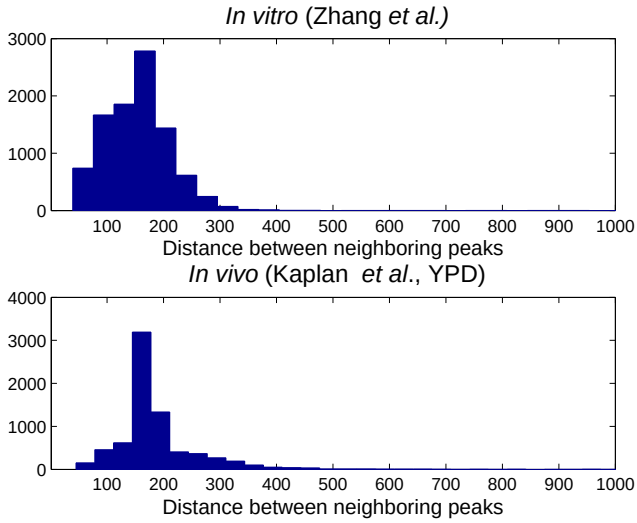
Supplementary Figure 4. Dinucleotide distributions in nucleosome and linker sequences. *Upper panel*: average relative frequencies of WW (AA, TT, AT and TA) and SS (CC, GG, CG and GC) dinucleotides at each position within the nucleosome are plotted with respect to the nucleosome dyad. The relative frequency of each dinucleotide is defined as its frequency at a given position divided by genome-wide frequency. All frequencies are smoothed using a 3 bp moving average. *Lower panel*: heat map of relative frequencies for each dinucleotide, plotted with respect to the nucleosome dyad. a) Nucleosomes assembled *in vitro* on the yeast genome (defined by more than five sequence reads), from Kaplan *et al.*[3] b) *In vivo* nucleosomes (defined by more than five sequence reads) from yeast cells grown in YPD medium.[3] Upper panel: dashed lines - cross-linked nucleosomes, solid lines - no cross-linking. Lower panel: dinucleotide counts based on a combination of all YPD replicates. c) Nucleosomes assembled *in vitro* on the *E.coli* genome (defined by more than one sequence read).[1] d) *In vivo* nucleosomes (defined by more than three sequence reads) from *C.elegans*.[4] e) Same as (a)-(d) except the dinucleotide frequencies are from mononucleosome-size DNA sequences (defined by more than five sequence reads) from yeast genomic DNA digested by MNase in the absence of nucleosomes. f) Same as (e) except mononucleosome-size DNA sequences (defined by more than one sequence read) were obtained by sonication.



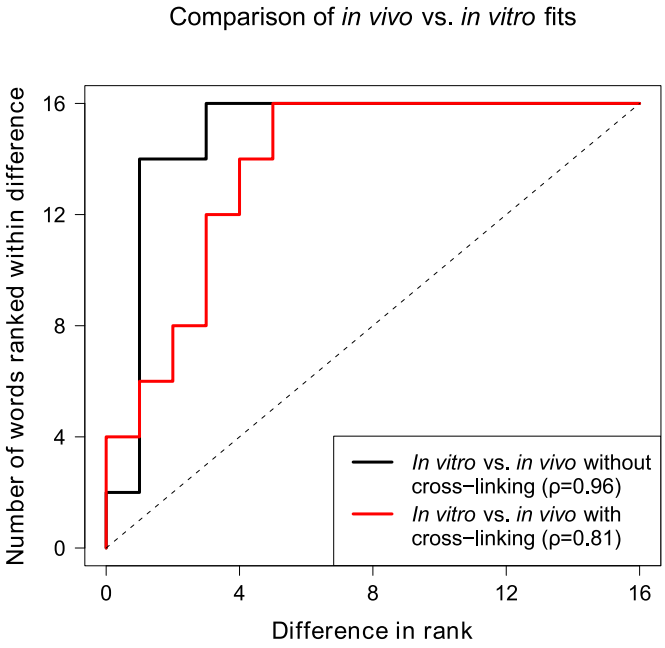
Supplementary Figure 5. Prediction of six nucleosome positions mapped *in vitro* at high resolution. Shown are nucleosome formation energies computed using the $N = 2$ position-independent model (green curves) and the spatially resolved model (blue curves). Vertical lines: known nucleosome starting positions, also listed in parentheses below. (a) The 180 bp sequence from the sea urchin 5S rRNA gene (bps 8,26).[5] (b) The 183 bp sequence from the pGUB plasmid (bps 11,31).[6] (c) The 215 bp fragment from the sequence of the chicken $\beta - \mathrm{globin}^{\mathrm{A}}$ gene (bp 52).[7] (d,e,f) Synthetic high-affinity sequences[8] 601 (bp 61), 603 (bp 81) and 605 (bp 59).[2]



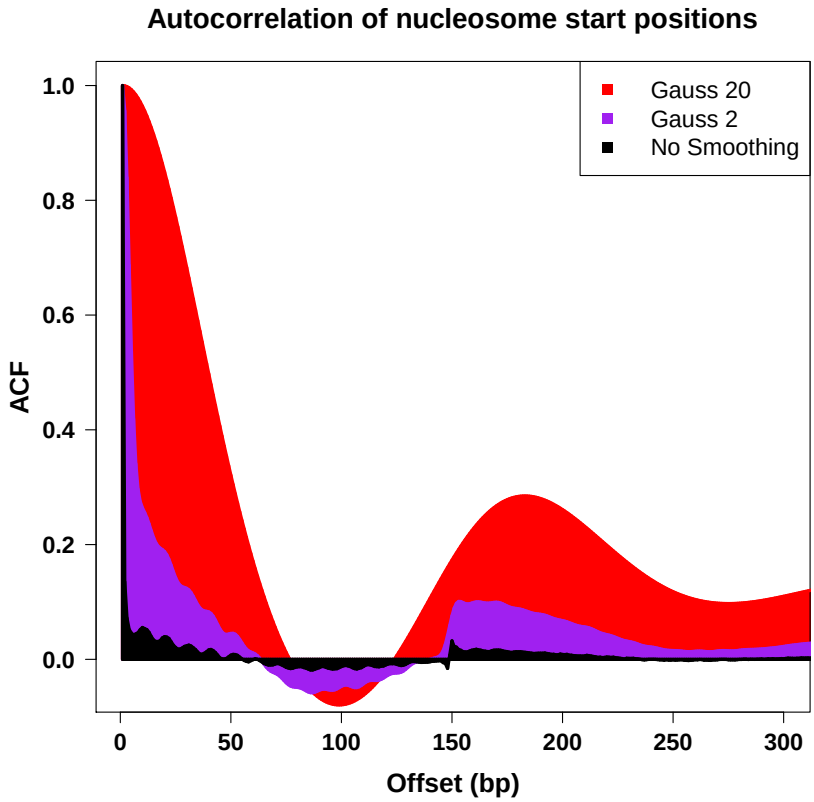
Supplementary Figure 6. Nucleosome energies and occupancies in the vicinity of transcription start and termination sites. a) Percus energy (red) and the sequence-specific energy predicted using the $N = 2$ position-independent model (blue). The energies were inferred from nucleosomes positioned *in vitro* on the yeast genome,[1] averaged over all genes for which transcript coordinates were available,[9] and plotted with respect to the transcription start and termination sites (TSS and TTS, respectively). All energies were divided by a genome-wide average. b) *In vitro* nucleosome occupancy (red),[1] *in vivo* nucleosome occupancy in YPD medium without cross-linking (blue),[3] and occupancy predicted using the $N = 2$ position-independent model (black). All occupancies were divided by the genome-wide average and plotted as described in (a).



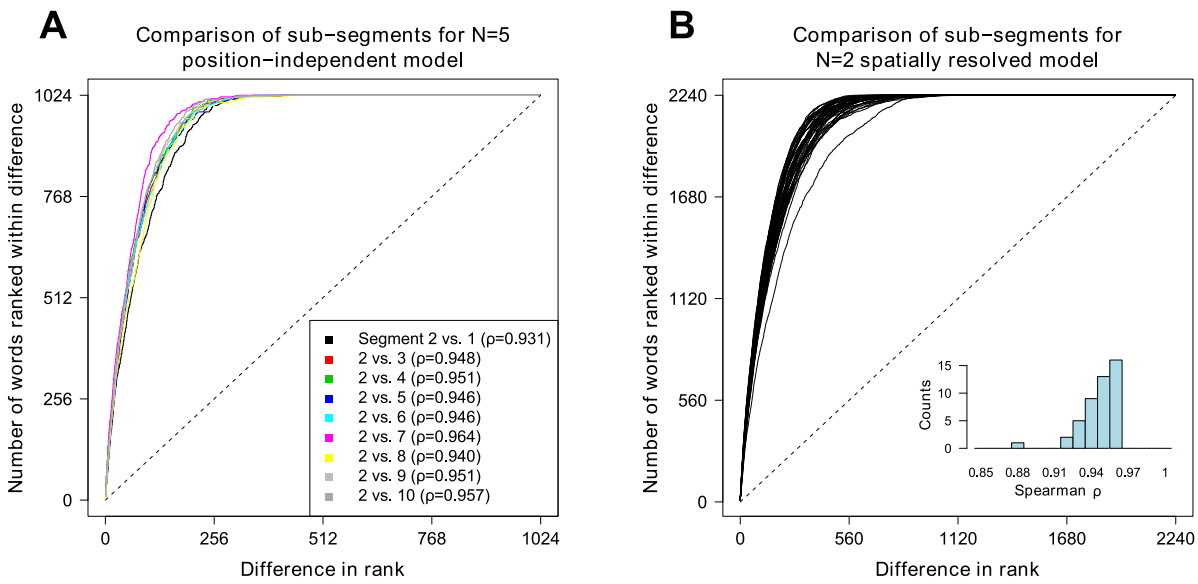
Supplementary Figure 7. Histogram of distances between neighboring peaks from *in vitro* and *in vivo* nucleosome sequence read profiles in *S.cerevisiae*. Mapped sequence reads were smoothed with a $\sigma = 20$ Gaussian. Neighboring peaks are defined by local maxima in the sequence read profile.



Supplementary Figure 8. Comparison of $N = 2$ position-independent models trained on *in vitro* and *in vivo S.cerevisiae* nucleosomes. Rank-order plots of energies of 2 bp words: the energy of each word is ranked using a position-independent model of order $N = 2$ trained on either *in vivo* (with and without cross-linking) or *in vitro* nucleosome positioning data. Each curve shows the number of words whose ranks are separated in the *in vivo* vs. *in vitro* fits by a given distance or less.



Supplementary Figure 9. Autocorrelation functions of nucleosome starting positions. Nucleosomes were assembled *in vitro* on the yeast genome.[1] Black: original starting positions, violet: starting positions smoothed with a $\sigma = 2$ Gaussian, red: starting positions smoothed with a $\sigma = 20$ Gaussian (see Supplementary Methods).



Supplementary Figure 10. Cross-validation of the $N = 5$ position-independent and $N = 2$ spatially resolved models in *S.cerevisiae*. a) Rank-order plots of energies of 5 bp words: yeast genome is divided into 4 segments of equal size and the energy of each word is ranked using $N = 5$ position-independent models independently trained on each segment. Each curve shows the number of words whose ranks are separated by a given distance or less. Energies of 5 bp-long words contain contributions from all shorter motifs: $E(S) = \sum_{n=1}^{5} \sum_{\{\alpha_1 \ldots \alpha_n\}}^{4^n} n_{\alpha_1 \ldots \alpha_n} \varepsilon_{\alpha_1 \ldots \alpha_n}$, where $n_{\alpha_1 \ldots \alpha_n}$ is the number of times a given word was found within the 5 bp-long sequence S and $\varepsilon_{\alpha_1 \ldots \alpha_n}$ is the fitted energy of that word. b) Rank-order plots of dinucleotide energies at each position predicted with $N = 2$ spatially resolved models independently trained on 47 segments of equal size. Dinucleotide energies at each position are computed using $E_{\alpha_i \alpha_{i+1}} = \varepsilon_{\alpha_i \alpha_{i+1}} + \varepsilon_{\alpha_i}$, $i = 4 \ldots 142$, $E_{\alpha_{143} \alpha_{144}} = \varepsilon_{\alpha_{143} \alpha_{144}} + \varepsilon_{\alpha_{143}} + \varepsilon_{\alpha_{144}}$ (Supplementary Methods) and ranked across all positions. The inset shows a histogram of rank-order correlation coefficients between dinucleotide energies trained on one of the segments, and all other segments.

Supplementary Tables

Supplementary Table 1. Table of correlation coefficients between predicted or observed occupancy profiles on the yeast genome. All observed profiles have been filtered for abnormally high- and low-density regions as described in the Supplementary Methods, with each correlation coefficient computed only for those basepairs that have not been removed from either dataset (predicted occupancies do not have filtered regions).

Supplementary Table 2. Table of dinucleotide energies predicted by training $N = 2$ position-independent models on several nucleosome positioning maps and nucleosome-free control experiments. Energies for each model have been rescaled to the variance of 1 a.u.

References

[1] Zhang Y, et al. (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*. *Nature Struct. Mol. Biol.* 16:847–852.

[2] Morozov AV, et al. (2009) Using DNA mechanics to predict *in vitro* nucleosome positions and formation energies. *Nucleic Acids Res.* 37:4707–4722.

[3] Kaplan N, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458:362–366.

[4] Valouev A, et al. (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18:1051–1063.

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

[6] Kassabov S, Henry N, Zofall M, Tsukiyama T, Bartholomew B (2002) High-resolution mapping of changes in histone-DNA contacts of nucleosomes remodeled by ISW2. *Mol. Cell. Biol.* 22:7524–7534.

[7] Davey C, Pennings S, Reilly C, Meehan R, Allan J (2004) A determining influence for CpG dinucleotides on nucleosome positioning *in vitro*. *Nucl. Acids Res.* 32:4322–4331.

[8] Lowary P, 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.

[9] Nagalakshmi U, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349.