Nucleosome positioning signals in genomic DNA

Heather E. Peckham, Robert E. Thurman, Yutao Fu, John A. Stamatoyannopoulos, William Stafford Noble, Kevin Struhl and Zhiping Weng

Genome Res. published online Jul 9, 2007;
Access the most recent version at doi:10.1101/gr.6101007

Letter

Nucleosome positioning signals in genomic DNA

Heather E. Peckham,[1,2] Robert E. Thurman,[3] Yutao Fu,[1] John A. Stamatoyannopoulos,[4] William Stafford Noble,[4,5] Kevin Struhl,[6] and Zhiping Weng[1,2,7]

[1]Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA; [2]Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA; [3]Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA; [4]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [5]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA; [6]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA

Although histones can form nucleosomes on virtually any genomic sequence, DNA sequences show considerable variability in their binding affinity. We have used DNA sequences of *Saccharomyces cerevisiae* whose nucleosome binding affinities have been experimentally determined (Yuan et al. 2005) to train a support vector machine to identify the nucleosome formation potential of any given sequence of DNA. The DNA sequences whose nucleosome formation potential are most accurately predicted are those that contain strong nucleosome forming or inhibiting signals and are found within nucleosome length stretches of genomic DNA with continuous nucleosome formation or inhibition signals. We have accurately predicted the experimentally determined nucleosome positions across a well-characterized promoter region of *S. cerevisiae* and identified strong periodicity within 199 center-aligned mononucleosomes studied recently (Segal et al. 2006) despite there being no periodicity information used to train the support vector machine. Our analysis suggests that only a subset of nucleosomes are likely to be positioned by intrinsic sequence signals. This observation is consistent with the available experimental data and is inconsistent with the proposal of a nucleosome positioning code. Finally, we show that intrinsic nucleosome positioning signals are both more inhibitory and more variable in promoter regions than in open reading frames in *S. cerevisiae*.

[Supplemental material is available online at www.genome.org and http://zlab.bu.edu/NPS/.]

Nucleosomal DNA in *Saccharomyces cerevisiae* is 165 bp long, of which 146 bp wrap around the histone octamer in 1.65 turns. The histone octamer, composed of two copies of each histone protein—H2A, H2B, H3, and H4—has been highly conserved throughout evolution (Luger et al. 1997). Genomic DNA sequences show considerable variability in their binding affinity to the histone octamer, and this variability contributes to determining the location and distribution of nucleosomes (Drew and Travers 1985; Satchwell et al. 1986; Travers and Klug 1987; Baldi et al. 1996; Ioshikhes et al. 1996; Lowary and Widom 1998; Stein and Bina 1999; Widom 2001; Thastrom et al. 2004a,b,c; Gencheva et al. 2006). The strongest natural nucleosome positioning sequences have been shown to have affinities for histone binding that are less than that of some synthetic random DNA sequences, indicating that eukaryotic genomes have not evolved to maximize nucleosome positioning power with sequence alone (Thastrom et al. 1999). A program developed to recognize nucleosome sites found that nucleosome positioning in the promoter region may influence the regulation of gene expression (Levitsky et al. 2001). Nucleosomes are depleted from active regulatory elements throughout the *S. cerevisiae* genome in vivo (Lee et al. 2004), and yeast promoter regions are bound poorly by histones both in vivo and in vitro (Sekinger et al. 2005).

The sequence-dependent structure of DNA appears to determine the rotational positioning of DNA about the nucleosome (Drew and Travers 1985; Satchwell et al. 1986). Several other studies have also provided extensive evidence indicating a sequence-dependent positioning of nucleosomes along DNA (Simpson 1991; Thoma 1992; Lu et al. 1994; Wolffe 1994; Trifonov 1997; Levitsky et al. 1999; Kiyama and Trifonov 2002), and much work has been done to elucidate the nucleosome positioning signals that determine the preference of a particular region to bind to histones and form a nucleosome. The CA dinucleotide has been shown to be important for nucleosome positioning, and the decamer TATAAACGCC has a high affinity for histones (Widlund et al. 1997, 1999). TGGA repeats impair nucleosome formation (Cao et al. 1998), and poly (dA:dT) has been shown to increase the accessibility of transcription factors bound to nearby sequences (Struhl 1985; Chen et al. 1987; Iyer and Struhl 1995). It is well known that DNA containing short AT-rich sequences spaced by an integral number of DNA turns is easiest to bend around the nucleosome (Alberts 2002). There is evidence of a periodic repeat every 10.4 bases of the dinucleotides AA and TT in nucleosome forming sequences (Cohanim et al. 2005), and a ~10-bp periodicity of AA/TT/TA dinucleotides that oscillate in phase with each other and out of phase with ~10-bp periodic GC dinucleotides has been demonstrated (Segal et al. 2006). However, there appear to be no specific motifs responsible for nucleosome formation, and it is likely that an overall signal is produced by the composition of the DNA.

A recent study has reported that ~50% of in vivo nucleosome positioning is governed by an intrinsic organization encoded in genomic DNA (Segal et al. 2006). This was determined by using 199 stably wrapped and center-aligned mononucleosome DNA sequences to construct a probabilistic model representative of the DNA sequence preferences of nucleosomes in *S. cerevisiae*. The model was derived from dinucleotide probability distributions and included the reverse complement of each nucleosome sequence to represent the twofold symmetry axis of

[7]Corresponding author.
E-mail zhiping@bu.edu; fax (617) 353-6766.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6101007.

the nucleosome structure (Richmond and Davey 2003). Another recent study used comparative genomics to align six *Saccharomyces* genomes and derive nucleosome positioning sequence patterns based on the frequency of AA and TT dinucleotides (Ioshikhes et al. 2006). This study found agreement between predicted nucleosome positions and the experimentally mapped nucleosomes used in our study (Yuan et al. 2005) and concluded that nucleosome positioning is at least partially determined by the underlying DNA sequence throughout the *Saccharomyces* genomes. The investigators identified conserved nucleosome positioning sequence patterns across the various *Saccharomyces* species and concluded that the basic features they identified should be evident in higher eukaryotes given the evolutionary conservation of chromatin structure.

While progress has been made in identifying sequences of DNA that either favor or inhibit nucleosome formation, advances have been limited due to the lack of large-scale experimental data. The identification of nucleosome positions throughout the genome of *S. cerevisiae* (Yuan et al. 2005) has provided an unprecedented opportunity to study nucleosome positioning signals. Our study uses a support vector machine (SVM) classifier, trained on a data set of sequence features from the strongest and weakest nucleosome forming 50-bp fragments (Yuan et al. 2005). Our SVM-based approach to sequence-based prediction of preferences for nucleosome positioning is complementary to, and equally predictive as, the model proposed by Segal et al. (2006).

Table 1. The features responsible for distinguishing the nucleosome forming from the nucleosome inhibiting sequences in the training set

k-mer	ROC score	Nucleosome forming (+) or inhibiting (−)
A/T	0.91	−
C/G	0.91	+
TA	0.83	−
ATA/TAT	0.81	−
TAA/TTA	0.81	−
AAA/TTT	0.80	−
AT	0.78	−
AAT/ATT	0.77	−
AATA/TATT	0.76	−
ATAA/TTAT	0.76	−
AAAA/TTTT	0.76	−
GC	0.75	+
CC/GG	0.74	+
CCA/TGG	0.74	+
TAAA/TTTA	0.73	−
AAAT/ATTT	0.72	−
CAG/CTG	0.71	+

The features collectively impart nucleosome formation/inhibition potential on a sequence. A plus symbol indicates nucleosome formation potential, and a minus symbol indicates nucleosome inhibition potential. The higher the ROC score, the more significant the feature is in distinguishing the two groups. The GC/AT richness of a sequence is the strongest single factor among *k*-mer frequencies in determining its nucleosome formation potential. The entire table containing all 2772 *k*-mers is available in the supplemental material at http://zlab.bu.edu/NPS/.

Results

Training set

We used a SVM to distinguish between nucleosome forming and nucleosome inhibiting sequences and created a training set consisting of the 1000 highest (nucleosome forming) and 1000 lowest (nucleosome inhibiting) scoring fragments from chromosome III of the data set (Yuan et al. 2005) as shown in Figure 1. Using the frequencies of the *k*-mers for $k = 1$ to 6, the SVM can distinguish between the fragments within the training set with high accuracy. We measure the quality of a given classifier by measuring the area under the receiver operating characteristic (ROC) curve. By this metric, a random classifier achieves a ROC score of 0.5, and a perfect classifier receives a ROC score of 1.0. The ROC scores from 10-fold cross-validation have a mean of 0.951 (SD = 0.02), demonstrating that the differences between the frequencies of these *k*-mers can largely differentiate the two groups. Sequence elements with length greater than six did not significantly change the discrimination power of the SVM. The mean ROC scores from 10-fold cross-validation of shorter sequence elements are shown in Supplemental Figure 1.

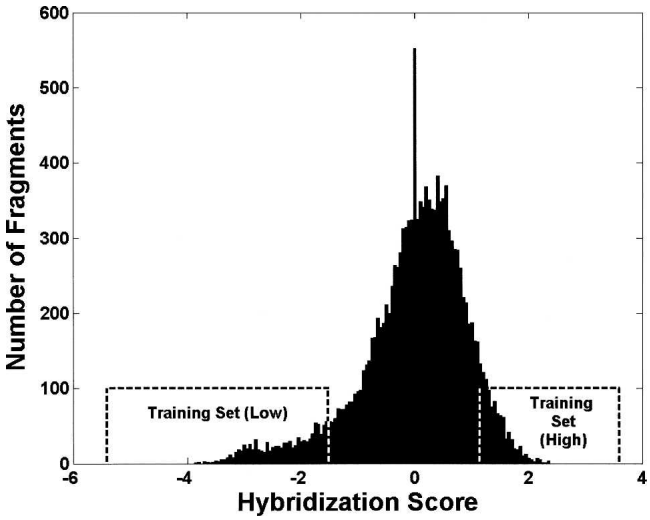

Figure 1. The training set is composed of the fragments with the 1000 highest and the 1000 lowest hybridization scores from chromosome III. High and low hybridization binding affinities indicate that a sequence is nucleosome forming or nucleosome inhibiting, respectively.

We also ranked the training set without the SVM to determine how well any given *k*-mer can separate the high-scoring fragments from the low-scoring fragments simply by counting the number of that *k*-mer in each of the sequences. For each of the 2772 *k*-mers, we ranked the training set by the *k*-mer composition and computed a ROC score from the resulting ranked list (see Table 1). Here we found that the single-nucleotide frequencies G+C (nucleosome forming) and A+T (nucleosome inhibiting) are the features most responsible for distinguishing the sequences, each with ROC scores of 0.91. This is consistent with findings that the AT-rich intergenic regions in *S. cerevisiae* are nucleosome-free (Lee et al. 2004; Sekinger et al. 2005). While the GC/AT-richness of a sequence is the strongest single factor among *k*-mer frequencies in determining its nucleosome formation potential, no individual *k*-mer can achieve the ROC score of the SVM using all *k*-mers; hence, it is the collection of features rather than any individual feature that leads to the best distinction of the two groups of sequences. Table 1 contains a ranked list of the abilities of the features to distinguish the two groups of sequences.

Application of the SVM to the test set

We used the trained SVM to classify the sequence fragments in the test set (those sequences not on chromosome III) and obtained the following ROC scores: entire test set, ROC = 0.71; extreme 2000 fragments, ROC = 0.90; extreme 1000 fragments, ROC = 0.93; and extreme 500 fragments, ROC = 0.97.

The extreme fragments are the fragments in the test set with the highest and lowest hybridization binding affinities. As shown in Figure 2, the fragments with the most extreme hybridization values score well, but the ROC scores decrease as the test set includes fragments with hybridization scores approaching zero. In other words, the fragments with extreme binding affinities are accurately distinguished by the SVM, while the fragments with hybridization values at or near zero are distinguished only as well as would be expected by chance. It is difficult for the SVM to correctly label the sequences with moderate hybridization values because the hybridization scores of these sequences are far from the hybridization scores of the sequences in the training set and indicate that they do not have a strong preference for either nucleosome formation or nucleosome inhibition.

Highly accurate predictions on subsets of the data

The SVM produces more accurate predictions for test set sequences with extreme prediction scores. Figure 3 shows a contour plot of the ROC scores of the fragments in the test set with predicted values at various cutoffs. When all of the fragments are included [corresponding to the position (0, 0) on the plot], a ROC score of 0.71 is achieved. However, as the fragments with predicted values at or close to zero are removed from the test set, the ROC scores improve significantly. A ROC score of 0.8 is achieved with predicted values ≥ 0.70 and ≤ -0.55, and a ROC score of 0.9 is achieved with predicted values ≥ 1.55 and ≤ -1.25. As the magnitudes of the thresholds are increased, the number of fragments that are included in the test set decreases but the accuracy of the predictions increases. We are able to make highly accurate predictions on a subset of the data—the fragments with the strongest nucleosome forming and inhibiting signals.

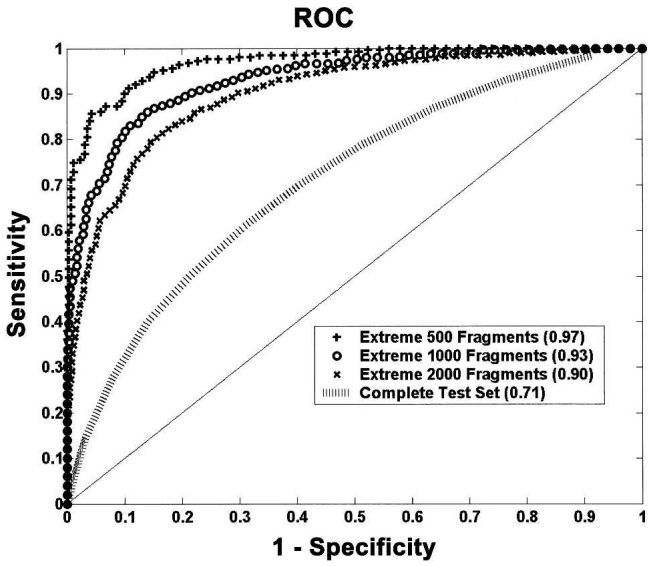

Figure 2. The ROC scores of the trained support vector machine applied to the test set composed of the fragments not on chromosome III. The support vector machine is trained on the fragments of chromosome III with the most extreme scores and most accurately separates the fragments in the test set with the most extreme scores.

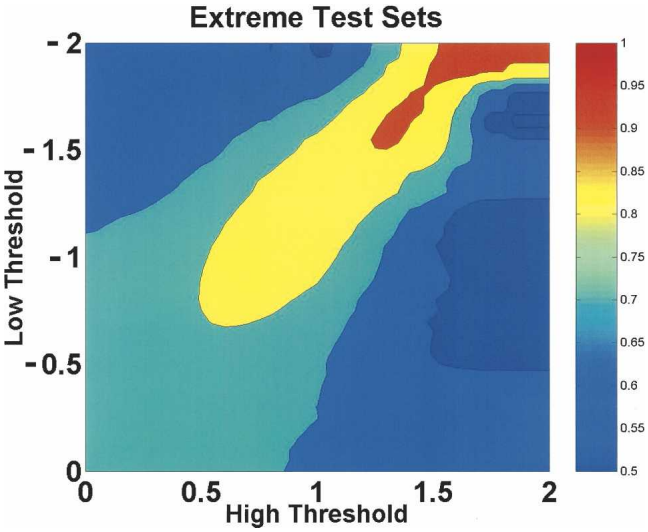

Figure 3. The fragments with the strongest nucleosome forming and inhibiting signals are the most accurately predicted within the test set. The *X*- and *Y*-axes are the high and low threshold values of the predicted scores of the fragments, respectively. The colors on the contour plot represent the ROC score. For example, at (0, 0) all the fragments in the test set are included and have a ROC score of 0.71, while at (1, −1) only the fragments in the test set with predicted scores ≥ 1 or ≤ -1 are included and have a ROC score of 0.83.

Taking the concept of thresholding a step further, we can consider what we call extreme neighbors, in which a fragment and the fragments adjacent to it must be above certain thresholds. In extreme neighbors 1, each fragment must be above the high threshold, T_h, or below the low threshold, T_l, and the fragment on each side of it must be $\geq [T_h - 0.1]$ or $\leq [T_l + 0.1]$, respectively. Likewise, in extreme neighbors 2 and extreme neighbors 3, each fragment must be above or below T_h and T_l and the two or three fragments, respectively, on each side of it must be $\geq [T_h - (0.2 \text{ or } 0.3)]$ or $\leq [T_l + (0.2 \text{ or } 0.3)]$. It is logical to require neighboring fragments to be above or below a threshold because a nucleosome corresponds to six to eight overlapping probes. Figure 4 illustrates the number of fragments in the test set and the corresponding ROC score achieved at various thresholds under the extreme and extreme neighbor conditions. The extreme neighbor test sets achieve higher ROC scores than the extreme test set but at the expense of fewer fragments being included at each set of thresholds. These fragments with strong nucleosome forming or inhibiting signals that are neighbored by fragments with the same characteristics and are thus part of nucleosome length stretches of genomic DNA with continuous nucleosome formation or inhibition signals are predicted with extraordinary accuracy.

Nucleosome formation potential

We can use the predicted values from the trained SVM to assess the nucleosome formation potential along any given DNA sequence. The *MRM1-HIS3* promoter region in *S. cerevisiae* has preferential accessibility that is determined by a general property of the DNA sequence rather than by defined sequence elements (Sekinger et al. 2005). Figure 5 shows the binding affinities (Yuan

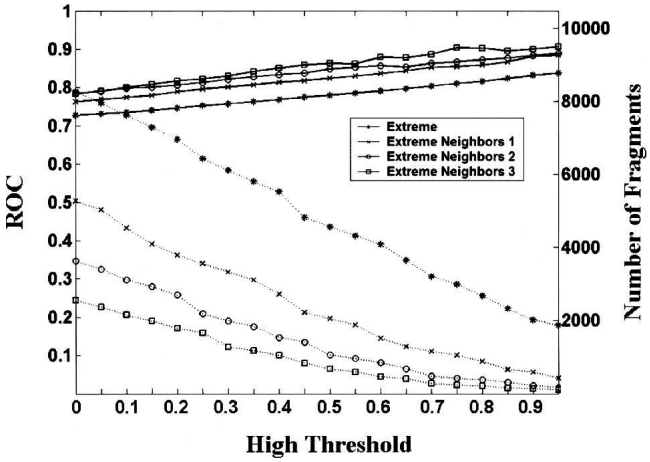

Figure 4. The number of fragments in the test set and the corresponding ROC score achieved under the extreme and extreme neighbor conditions. The *X*-axis represents 0.05 increments in the high threshold, T_h. Within each increment of T_h, increments of 0.05 from 0.00 to −2.00 of the low threshold, T_l, are computed, and the highest ROC score from these increments is selected. For example, the highest ROC score achieved between a T_h of 0 and a T_l between 0.00 and −2.00 is 0.73 from 8264 fragments. The *left* and *right Y*-axes represent the ROC score (solid lines) and the number of fragments in the test sets (dashed lines), respectively.

et al. 2005) and the predicted values of nucleosome formation potential determined by our SVM along the *MRM1-HIS3* promoter region. The predictions reproduce the pattern of nucleosome formation potential identified by the experimentally determined hybridization values ($r = 0.75$), thus confirming that the general properties of the DNA sequence can determine the positions of the nucleosomes in this promoter region.

To further test our ability to predict nucleosome formation potential, we used our trained SVM to predict the nucleosome formation potential along 199 center-aligned mononucleosomes studied recently (Segal et al. 2006). Figure 6 shows the mean predicted nucleosome formation potential over the forward and reverse strands of the 199 nucleosome sequences, in which the positions represent the center of the 50-mers on which the SVM made a prediction. The nucleosome formation potentials form an arc along the nucleosome and peak every 10 bp, except in the center of the nucleosome where the peaks are every 20 bp with a slight bulge 10 bp in between them. In contrast, the GC-content of the mononucleosomes shows much weaker periodicity (to facilitate comparison, the ranges of the SVM scores and GC-content in Figure 6 are set to 5% of their respective total ranges). The periodicity of the nucleosome formation potential is identified by the SVM even though there is no periodicity information used in its training, and the pattern reproduces the ~10-bp periodicity found using the fraction of AA/TT/TA dinucleotides at each position of the center-aligned nucleosome-bound DNA sequences (Segal et al. 2006). A strong periodicity of AA and TT dinucleotides along *Caenorhabditis elegans* nucleosomal DNA at intervals of ~10 bp that becomes less pronounced in the sequence surrounding the putative dyad of the nucleosome (base pair 73) has also been observed (Johnson et al. 2006). A ~10-bp periodicity around the dyad has also been seen in nucleosomes of the ovine β-lactoglobulin gene in which the investigators suggested that in vivo positioning sites are not necessarily aligned with the strongest available in vitro positioning site but rather tend to be located at a fixed distance (Gencheva et al. 2006). The SVM implicitly reveals the AA/TT periodicity in the mononucleosomes and clearly demonstrates its advantage over using the GC-content alone to identify nucleosome formation potential.

To determine the influence of nucleosome formation potential on in vivo nucleosome positions, we used a hidden Markov model to derive the boundaries of the predicted nucleosomes in our test set and compared them to the nucleosome boundaries determined by the experimental data. The hidden Markov model was developed to use the hybridization values of the tiled probes as input to assign nucleosome/linker boundaries (Yuan et al. 2005). Here, we use it for the same purpose, but with scores predicted by the SVM as input. Overall, 41.3% and 49.8% of our predicted well-positioned nucleosomes are within 30 and 40 bp, respectively, of those determined experimentally (Yuan et al. 2005), compared with 24.05 ± 0.02% expected by chance ($P < 10^{-8}$). A previous study reported that 54% of predicted stable nucleosomes were within 35 bp of literature positions compared with 39 ± 1% expected by chance ($P < 10^{-16}$) (Segal et al. 2006). Despite the differences between the methods for predicting nucleosome formation, they both position 15%–17% more nucleosomes than is expected by chance. Since the chance occurrences of well-positioned nucleosomes provide no information on intrinsic positioning, the percentage of nucleosomes that are intrinsically positioned is determined as the proportion of non-chance nucleosomes that are aligned. Since both approaches reveal 15%–17% more well-positioned nucleosomes than is expected by chance, this corresponds to 22%–25% of nucleosome positioning due to intrinsic sequence signal. A comparison of the experimental and predicted nucleosome mapping is available in the Supplemental Material.

For comparison, we trained another SVM with the regions assigned as well-positioned nucleosomes and linkers in chromosome III (Yuan et al. 2005). We then used this trained SVM to make new predictions on the test set (nonchromosome III regions) and used these predicted values as input to the aforemen-

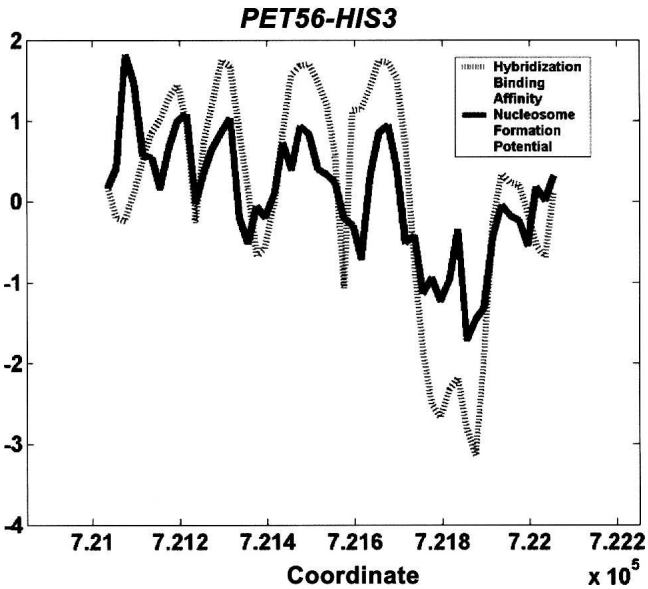

Figure 5. The predicted nucleosome formation potentials reproduce the pattern of positioned nucleosomes found with the experimentally determined hybridization binding affinities in the *MRM1-HIS3* promoter region.

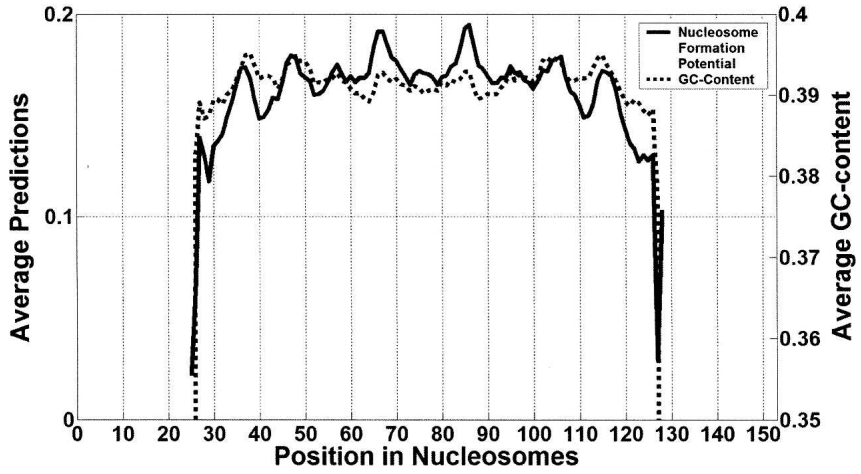

Figure 6. The trained support vector machine reveals the periodicity of the nucleosome formation potential along center-aligned mononucleosomes. The mean predicted nucleosome formation potential over the forward and reverse strands of 199 nucleosome sequences is shown with a solid line, while the GC-content of the sequences is shown with a dashed line. The positions represent the center of the 50-mers on which either the SVM made a prediction or the GC-content was calculated.

tioned hidden Markov model. We found 38% and 46% of the predicted well-positioned nucleosomes are within 30 and 40 base pairs, respectively, of those determined experimentally. This performance is slightly worse than that of the SVM trained on the 2000 probes with the most extreme hybridization values, supporting the validity of using individual probe sequences as training data.

Intrinsic variability of promoter regions

Previous evidence has shown that intrinsic depletion of nucleosomes is a mechanism commonly used by promoter regions in *S. cerevisiae* and that the intrinsic positioning of nucleosomes within coding regions in this species is more modest (Sekinger et al. 2005). We hypothesize that the organizational plan of *S. cerevisiae* may be that promoter regions are specifically designed to inhibit nucleosome formation. If this hypothesis is true, then we would expect the average nucleosome formation potential within each of the promoter regions to be lower than that within each of the coding regions, indicating that promoter regions are more inhibitory. Figure 7 shows the averages and standard deviations of the predicted nucleosome formation potentials within each of the open reading frames and corresponding promoter regions in *S. cerevisiae* obtained from the Saccharomyces Genome Database (http://www.yeastgenome.org). Indeed, the average nucleosome formation potential within each of the promoter regions is significantly lower than that within each of the coding regions ($P < 10^{-15}$). Meanwhile, the standard deviation of the nucleosome formation potentials within each of the promoter regions is significantly higher than that within each of the coding regions ($P < 10^{-15}$), perhaps because the complex regulatory functions of promoters require some regions to be nucleosome free and others to be tightly bound by nucleosomes, while the primary function of coding regions is not regulated by the variable binding of individual nucleosomes. Nucleosome positioning has been suggested to play a role in regulating gene expression in human promoter regions in which the nucleosome formation potential has been shown to be higher in tissue-specific genes than in housekeeping genes (Levitsky et al. 2001).

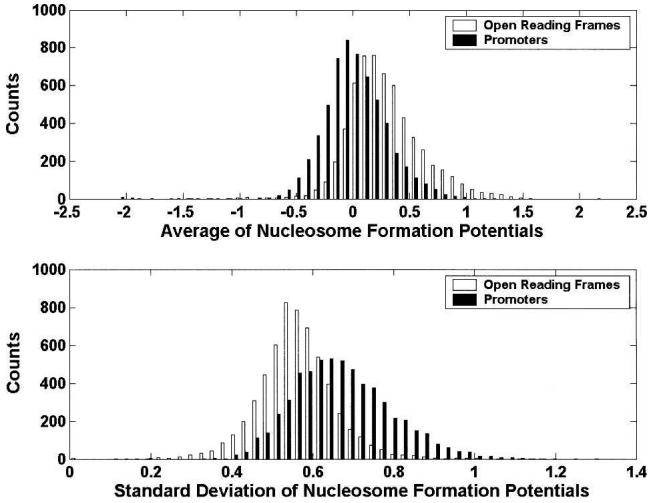

Figure 7. Intrinsic nucleosome positioning signals are more inhibitory and more variable in promoter regions than in open reading frames in *S. cerevisiae*. The distributions of the average and standard deviation of the nucleosome formation potentials within each of the promoter regions and open reading frames suggest that promoter regions in *S. cerevisiae* are specifically designed to inhibit nucleosome formation.

Discussion

We have developed a method to predict the nucleosome formation potential of any given sequence of DNA and identified the sequence features that contribute to this potential. Our analysis gives additional insight into previously studied regulatory regions (Yuan et al. 2005), and we can now identify the segments of DNA in this data set that have nucleosome formation potential and those that lack it. It is likely that when nucleosomes are formed with DNA that lacks nucleosome formation potential, there are other factors involved in their formation. We can also expect that in regions showing high nucleosome formation potential but no nucleosomes, there must be an explanation for why the area is devoid of nucleosomes. For example, it may be an active regulatory region that requires the DNA to be free of chromatin. There are two well-known mechanisms for disrupting nucleosomes and causing lower nucleosome occupancy—activator-dependent recruitment of chromatin-modifying activities at enhancers (Deckert and Struhl 2001; Boeger et al. 2003, 2004; Reinke and Horz 2003; Korber et al. 2004) and chromatin alterations during the process of Pol II elongation (Kristjuhan and Svejstrup 2004; Lee et al. 2004; Schwabish and Struhl 2004). These mechanisms occur largely independently of the DNA sequences wrapped around nucleosomes and work in conjunction with the nucleosome formation potential of the sequences to produce the observed nucleosome positions.

The features responsible for distinguishing the two groups

of sequences in the training set show that the GC/AT-richness of a sequence is the strongest factor with regard to single *k*-mer frequencies in determining the nucleosome formation potential. GC- and AT-rich sequences favor and inhibit nucleosome formation, respectively. An enrichment of AT/TA dinucleotides in linker DNA and of GC/CG dinucleotides in nucleosome core DNA has also been found in *C. elegans* (Johnson et al. 2006). The differences between sequences that prefer to bind to histones and those that do not may lie in the flexibility of the DNA. If the sequence is flexible, it will more easily wrap around the histones than if it is rigid, and the *k*-mers identified as contributing to or resisting nucleosome formation may be a result of the deformability of the sequence.

The sequence dependence of nucleosomal positioning has been examined from many perspectives (Widom 2001) including modeling DNA structure and deformability (Sivolob and Khrapunov 1995; Anselmi et al. 1999; Kiyama and Trifonov 2002). It has recently been shown that histone proteins impose large shearing deformations of adjacent base pairs at sites of sharp local bending into the minor groove, and this appears to govern both the superhelical pathway and the positioning of the nucleosome (Tolstorukov et al. 2007). These shear deformations, called Slide, describe the displacement of two adjacent base pairs along the long axis of the dinucleotide step (Dickerson 1989). Tolstorukov and colleagues found that the computed cost of deforming DNA on the nucleosome increases substantially if the crystallized sequence is displaced relative to its observed positioning, indicating that the flexibility of each type of dinucleotide step with respect to Slide is related to the nucleosome formation potential of that step.

The flexibilities of the 10 unique dinucleotide steps with respect to the parameter Slide have been used to determine that although the pyrimidine-purine steps have traditionally been considered the most flexible, TA has context-dependent flexibility while CA/TG and CG are generally more flexible (Packer et al. 2000a). According to this study, CG, GC, and GG/CC steps are flexible; AT and AA/TT steps are rigid; and the TA step has context-dependent flexibility, providing a sound explanation of why G+C content and A+T content favor and inhibit nucleosome formation, respectively. CA/TG is a highly flexible dinucleotide, and while it is not as distinguishable as GC, CC/GG, and CG in terms of frequency between nucleosome forming and nucleosome inhibiting sequences, each incidence of CA/TG in a sequence likely has a large impact in imparting nucleosome formation potential. A complementary study on tetranucleotide structure has shown that the dinucleotide steps AT/TT, AT, and TA are context independent, while CC/GG, CG, and GC are strongly context dependent and the remaining steps are weakly context dependent (Packer et al. 2000b). Thus dinucleotides that inhibit nucleosome formation are generally rigid regardless of their context, while those that favor nucleosome formation are flexible with their structure depending on their tetranucleotide context.

A-tracts, straight and rigid sequences that cause a sharp bend in the dinucleotide step following them, are strong nucleosome breakers and appear to be used as part of a nucleosome prevention system (Iyer and Struhl 1995). A-tracts are present in significantly higher amounts in the low scoring (nucleosome inhibiting) compared with the high scoring (nucleosome forming) fragments of our training set. A comparison of A-tracts and C-tracts in the training set reveals striking differences. While C-tracts of size 4 to 6 cannot distinguish the two groups of fragments (ROC = 0.55), A-tracts of the same size do a fair job (ROC = 0.73). As shown in the supplemental data to Table 1, among the frequencies of single features, AAAAA and AAAAAA are the highest-scoring 5- and 6-mers, respectively, indicating that these homopolymers have the most influence among the longest *k*-mers in imparting or inhibiting nucleosome formation. These results are consistent with other studies concluding that A-tracts contribute to the prevention of nucleosome formation but are not sufficient on their own (Iyer and Struhl 1995; Suter et al. 2000). The role of A-tracts in increasing transcription and protein accessibility has been well demonstrated (Russell et al. 1983; Struhl 1985; Chen et al. 1987), and nucleosome-free regions have been found to be enriched for A-tracts (Yuan et al. 2005).

We compared the locations of our predicted well-positioned nucleosomes to experimentally determined locations using the same definition of accuracy used previously (Segal et al. 2006) so that a direct comparison could be made. Our demonstration that 41% and 50% of our predicted well-positioned nucleosomes are within 30 and 40 bp, respectively, of those determined experimentally (Yuan et al. 2005), compared with 24% and 33% expected by chance, indicates that 17% of the nucleosome positioning above what is expected by chance is determined by intrinsic signals in the genomic DNA. This corresponds to 22%–25% of non-chance nucleosome positioning. These results are in accord with a previous study claiming that ~50% of nucleosome positioning is determined by sequence (Segal et al. 2006). In this study, 54% of the predicted stable nucleosomes were within 35 bp of literature positions, while 39% were expected to be by chance, indicating an intrinsic sequence signal effect of 15% above random and 25% nonchance nucleosome positioning. This lower rate of intrinsic sequence signal that determines nucleosome positioning goes against the idea of a genomic code and rather reveals that the discrimination of different sequences by histones is relatively subtle except in some extreme cases. This is in agreement with a previous study suggesting that nucleosome location has a greater role than positioning strength in nucleosome remodeling (Ioshikhes et al. 2006). It is also in excellent accord with the only study to explicitly reveal intrinsic nucleosome positioning in which two out of seven nucleosomes are positioned similarly in vitro and in vivo (Sekinger et al. 2005).

Nucleosome positioning sequence patterns have been found to be conserved across related *Saccharomyces* species (Ioshikhes et al. 2006). We have shown evidence that promoter regions deliberately inhibit nucleosome formation to a greater extent than coding regions, suggesting an evolutionary selection that makes these regions distinct with respect to intrinsic histone patterns. We would expect nucleosome positioning in promoter regions to be conserved throughout *Saccharomyces* species even if the underlying sequence is not strongly conserved. If this proves to be the case, it would add a new level of understanding to sequence conservation as it would be a general sequence property rather than individual motifs that is conserved. Other non-yeast species can also be examined to reveal the extent to which the nucleosome positioning features are universal. It is reasonable that the high conservation of histone proteins between species would contribute to similar nucleosome positioning signals being found within the genomes of distant organisms.

Methods

We used experimental data consisting of nucleosome hybridization affinities of overlapping 50-mers within the *S. cerevisiae* genome (Yuan et al. 2005). The *S. cerevisiae* DNA was treated with

MNase, a DNA-digesting enzyme that removes the regions connecting one nucleosome bead with the next. The nucleosomal DNA itself survived because it was protected by histones. The nucleosomal DNA was isolated, labeled with a green fluorescent dye, and mixed with digested pieces of total genomic yeast DNA labeled with a red fluorescent dye. The mixture was then applied to a microarray chip covered with overlapping 50-base DNA sequences. The DNA pieces hybridize to the probes on the chip, and by plotting the green-to-red ratio for each spot on the chip, the positions of the nucleosomes could be determined. The overlapping 50 base pair fragments cover ~4% of the *S. cerevisiae* genome (most of chromosome III and 223 additional regulatory regions), and for each fragment, we have the genomic coordinates, sequence, and hybridization score. Scores represent hybridization strength and are inversely proportional to the amount of cleavage; i.e., high-scoring fragments are nucleosome forming and low-scoring fragments are non-nucleosome forming.

Support vector machine

The SVM is a supervised classification algorithm that separates two groups of data according to given characteristics (Vapnik 1998). The trained classifier can subsequently be used to assign new data points to one of the two given classes. A training set is mapped onto a feature space, and a plane separating the positive and negative examples is chosen that maintains a maximum margin from any point in the training set. The classification of an unlabeled example can then be predicted by mapping it into the feature space and determining on which side of the separating plane it lies.

We used the Gist software package for SVM classification (Pavlidis et al. 2004). In our application, the data are 50-mer sequences, and the two groups are "nucleosome forming" and "nucleosome inhibiting." We represent each sequence using the frequencies of each overlapping *k*-mer, where $k = 1$ to 6 (A, C, AA, AT, TA, etc.). Thus, each sequence is converted into a fixed-length vector of *k*-mer frequencies and is labeled indicating whether it is initially identified as nucleosome forming or nucleosome inhibiting. A ROC score indicates the accuracy with which the frequencies of the *k*-mers separate the two groups of sequences in a held-out test set (Hanley and McNeil 1982). The values of the ROC score range from zero to one, with a score of one being perfect (the test correctly predicts the classification of each object) and a score of 0.5 indicating that the predictions are random.

Determination of nucleosome states

Hidden Markov models (HMMs) allow us to look at time series data in which what we wish to predict is not what we observe; i.e., the underlying system is hidden. A HMM has been developed in which the observed states are the hybridization value of each 50-bp fragment of DNA and the hidden states are the designation of a well-positioned nucleosome (N), a delocalized nucleosome (D), or a nucleosome free/linker region (L) for each fragment (Yuan et al. 2005). The HMM uses both the hybridization value of the DNA fragment as well as the probable state of the DNA fragments preceding it to determine its state. We used the predicted values derived from the SVM in place of the hybridization values as input to the HMM in order to determine the predicted boundaries of the nucleosomes and linker regions within the continuous regions of our test set. The HMM uses overlapping 50-mers with a step size of 20 as input and results in six to eight probes indicating a well-positioned nucleosome and nine or more indicating a delocalized nucleosome. As a result, distances between centers of nucleosomes are in 10-bp increments. The number of well-positioned nucleosomes that are expected to align by chance with those determined experimentally were determined by repeated random shuffling of the placement of the predicted well-positioned nucleosomes in each continuous region.

Acknowledgments

We thank Dr. Guo-Cheng Yuan and Dr. Oliver J. Rando of the Bauer Center for Genomics Research at Harvard University for kindly sharing with us the source code of the hidden Markov model that was used in their manuscript (Yuan et al. 2005). We also thank Mary Ellen Fitzpatrick and Stephen Peckham for continuous system administration support. This work was supported by the National Institutes of Health grants ENCODE R01HG03110, U01HG003161 and R01GM071923.

References

Alberts, B. 2002. *Molecular biology of the cell*. Garland Science, New York.
Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M., and Scipioni, A. 1999. Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability. *J. Mol. Biol.* **286:** 1293–1301.
Baldi, P., Brunak, S., Chauvin, Y., and Krogh, A. 1996. Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.* **263:** 503–510.
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.
Boeger, H., Griesenbeck, J., Strattan, J.S., and Kornberg, R.D. 2004. Removal of promoter nucleosomes by disassembly rather than sliding in vivo. *Mol. Cell* **14:** 667–673.
Cao, H., Widlund, H.R., Simonsson, T., and Kubista, M. 1998. TGGA repeats impair nucleosome formation. *J. Mol. Biol.* **281:** 253–260.
Chen, W., Tabor, S., and Struhl, K. 1987. Distinguishing between mechanisms of eukaryotic transcriptional activation with bacteriophage T7 RNA polymerase. *Cell* **50:** 1047–1055.
Cohanim, A.B., Kashi, Y., and Trifonov, E.N. 2005. Yeast nucleosome DNA pattern: deconvolution from genome sequences of *S. cerevisiae*. *J. Biomol. Struct. Dyn.* **22:** 687–694.
Deckert, J. and Struhl, K. 2001. Histone acetylation at promoters is differentially affected by specific activators and repressors. *Mol. Cell. Biol.* **21:** 2726–2735.
Dickerson, R.E. 1989. Definitions and nomenclature of nucleic acid structure parameters. *J. Biomol. Struct. Dyn.* **6:** 627–634.
Drew, H.R. and Travers, A.A. 1985. DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* **186:** 773–790.
Gencheva, M., Boa, S., Fraser, R., Simmen, M.W., Whitelaw, C.B.A., and Allan, J. 2006. In vitro and in vivo nucleosome positioning on the ovine beta-lactoglobulin gene are related. *J. Mol. Biol.* **361:** 216–230.
Hanley, J.A. and McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143:** 29–36.
Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., and Trifonov, E.N. 1996. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* **262:** 129–139.
Ioshikhes, I.P., Albert, I., Zanton, S.J., and Pugh, B.F. 2006. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* **38:** 1210–1215.
Iyer, V. and Struhl, K. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14:** 2570–2579.
Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P., and Fire, A.Z. 2006. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* **16:** 1505–1516.
Kiyama, R. and Trifonov, E.N. 2002. What positions nucleosomes?—A model. *FEBS Lett.* **523:** 7–11.
Korber, P., Luckenbach, T., Blaschke, D., and Horz, W. 2004. Evidence for histone eviction in trans upon induction of the yeast PHO5 promoter. *Mol. Cell. Biol.* **24:** 10965–10974.
Kristjuhan, A. and Svejstrup, J.Q. 2004. Evidence for distinct mechanisms facilitating transcript elongation through chromatin in vivo. *EMBO J.* **23:** 4243–4252.
Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D., and Lieb, J.D. 2004. Evidence

for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* **36:** 900–905.
Levitsky, V.G., Ponomarenko, M.P., Ponomarenko, J.V., Frolov, A.S., and Kolchanov, N.A. 1999. Nucleosomal DNA property database. *Bioinformatics* **15:** 582–592.
Levitsky, V.G., Podkolodnaya, O.A., Kolchanov, N.A., and Podkolodny, N.L. 2001. Nucleosome formation potential of eukaryotic DNA: Calculation and promoters analysis. *Bioinformatics* **17:** 998–1010.
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.
Lu, Q., Wallrath, L.L., and Elgin, S.C. 1994. Nucleosome positioning and gene regulation. *J. Cell. Biochem.* **55:** 83–92.
Luger, K., Mader, 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.
Packer, M.J., Dauncey, M.P., and Hunter, C.A. 2000a. Sequence-dependent DNA structure: Dinucleotide conformational maps. *J. Mol. Biol.* **295:** 71–83.
Packer, M.J., Dauncey, M.P., and Hunter, C.A. 2000b. Sequence-dependent DNA structure: Tetranucleotide conformational maps. *J. Mol. Biol.* **295:** 85–103.
Pavlidis, P., Wapinski, I., and Noble, W.S. 2004. Support vector machine classification on the web. *Bioinformatics* **20:** 586–587.
Reinke, H. and Horz, W. 2003. Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. *Mol. Cell* **11:** 1599–1607.
Richmond, T.J. and Davey, C.A. 2003. The structure of DNA in the nucleosome core. *Nature* **423:** 145–150.
Russell, D.W., Smith, M., Cox, D., Williamson, V.M., and Young, E.T. 1983. DNA sequences of two yeast promoter-up mutants. *Nature* **304:** 652–654.
Satchwell, S.C., Drew, H.R., and Travers, A.A. 1986. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191:** 659–675.
Schwabish, M.A. and Struhl, K. 2004. Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Mol. Cell. Biol.* **24:** 10111–10117.
Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772–778.
Sekinger, E.A., Moqtaderi, Z., and Struhl, K. 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* **18:** 735–748.
Simpson, R.T. 1991. Nucleosome positioning: Occurrence, mechanisms, and functional consequences. *Prog. Nucleic Acid Res. Mol. Biol.* **40:** 143–184.
Sivolob, A.V. and Khrapunov, S.N. 1995. Translational positioning of nucleosomes on DNA: The role of sequence-dependent isotropic DNA bending stiffness. *J. Mol. Biol.* **247:** 918–931.
Stein, A. and Bina, M. 1999. A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res.* **27:** 848–853.
Struhl, K. 1985. Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc. Natl. Acad. Sci.* **82:** 8419–8423.
Suter, B., Schnappauf, G., and Thoma, F. 2000. Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Res.* **28:** 4083–4089.
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.
Thastrom, A., Bingham, L.M., and Widom, J. 2004a. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.* **338:** 695–709.
Thastrom, A., Gottesfeld, J.M., Luger, K., and Widom, J. 2004b. Histone-DNA binding free energy cannot be measured in dilution-driven dissociation experiments. *Biochemistry* **43:** 736–741.
Thastrom, A., Lowary, P.T., and Widom, J. 2004c. Measurement of histone-DNA interaction free energy in nucleosomes. *Methods* **33:** 33–44.
Thoma, F. 1992. Nucleosome positioning. *Biochim. Biophys. Acta* **1130:** 1–19.
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.* doi: 10.1016/j.jmb.2007.05.048.
Trifonov, E.N. 1997. Genetic level of DNA sequences is determined by superposition of many codes. *Mol. Biol. (Mosk.)* **31:** 759–767.
Vapnik, V.N. 1998. *Statistical learning theory*. Wiley, New York.
Widlund, H.R., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P.E., Kahn, J.D., Crothers, D.M., and Kubista, M. 1997. Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.* **267:** 807–817.
Widlund, H.R., Kuduvalli, P.N., Bengtsson, M., Cao, H., Tullius, T.D., and Kubista, M. 1999. Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J. Biol. Chem.* **274:** 31847–31852.
Widom, J. 2001. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.* **34:** 269–324.
Wolffe, A.P. 1994. Nucleosome positioning and modification: Chromatin structures that potentiate transcription. *Trends Biochem. Sci.* **19:** 240–244.
Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309:** 626–630. http://www.yeastgenome.org/.

Received November 5, 2006; accepted in revised form June 5, 2007.