Characteristics of Nucleosome Core DNA and Their Applications in Predicting Nucleosome Positions

Hongde Liu, Jiansheng Wu, Jianming Xie, Xi'nan Yang, Zuhong Lu, and Xiao Sun
State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China

ABSTRACT By analyzing dinucleotide position-frequency data of yeast nucleosome-bound DNA sequences, dinucleotide periodicities of core DNA sequences were investigated. Within frequency domains, weakly bound dinucleotides (AA, AT, and the combinations AA-TT-TA and AA-TT-TA-AT) present doublet peaks in a periodicity range of 10–11 bp, and strongly bound dinucleotides present a single peak. A time-frequency analysis, based on wavelet transformation, indicated that weakly bound dinucleotides of core DNA sequences were spaced smaller (~10.3 bp) at the two ends, with larger (~11.1 bp) spacing in the middle section. The finding was supported by DNA curvature and was prevalent in all core DNA sequences. Therefore, three approaches were developed to predict nucleosome positions. After analyzing a 2200-bp DNA sequence, results indicated that the predictions were feasible; areas near protein-DNA binding sites resulted in periodicity profiles with irregular signals. The effects of five dinucleotide patterns were evaluated, indicating that the AA-TT pattern exhibited better performance. A chromosome-scale prediction demonstrated that periodicity profiles perform better than previously described, with up to 59% accuracy. Based on predictions, nucleosome distributions near the beginning and end of open reading frames were analyzed. Results indicated that the majority of open reading frames' start and end sites were occupied by nucleosomes.

INTRODUCTION

Seventy-five to ninety percent eukaryotic genomic DNA is packaged in nucleosomes (1). Nucleosome DNA can be divided into the core and the linker DNA. Core DNA, a ~147-bp DNA sequence, is sharply bent and tightly wrapped around a histone protein octamer (2). Nucleosome positioning refers to the position the DNA helix adopts with respect to the histone core; positioning has been implicated in the regulation of gene expression in eukaryotic cells, since packaging DNA into nucleosomes affects sequence accessibility (3–6).

Because DNA contains specific dinucleotides, in particular weakly bound dinucleotides (e.g., AA, TT, TA) (3,4), which are periodically spaced by 10–11 bp (7–10), DNA sequences are intimately associated to nucleosome positioning. The periodic spacing of dinucleotides facilitates a bend in the DNA helix, which allows wrapping of the histone protein octamer. Recently, computational models and experimental determinations have made advancements in the field of nucleosome positioning (3,4,11–13). Segal and co-workers (3) identified nucleosome-bound DNA sequences and employed those sequences to construct a dinucleotide-based model for nucleosome positioning (3). Similarly, Ioshikhes and co-workers utilized the occurrence of periodically distributed AA and TT dinucleotides to define a ''nucleosome positioning sequence'' (11). Both groups subsequently applied their models to predict nucleosome positioning on the entire *Saccharomyces cerevisiae* genome and compared their predictions to experimentally determined nucleosome locations. Results suggest that the genome DNA sequence partly determines the locations of nucleosomes. Yuan and co-workers developed a tiled microarray approach to identify nucleosomes on *S. cerevisiae* DNA (4). Using a nucleosome DNA sequence probe based on a specific dinucleotide periodical pattern, Salih and co-workers introduced a straightforward method for nucleosome mapping (12). More recently, Albert and co-workers sequenced DNA from 322,000 individual *S. cerevisiae* nucleosomes and analyzed the functions of nucleosome positioning in gene regulation (13). Support vector machines were also trained to predict nucleosome positions (14).

In this study, a new dinucleotide periodicity characteristic was discovered in core DNA sequences by analyzing the dinucleotide position frequency of yeast nucleosome-bound DNA sequences with wavelet-based techniques. Subsequently, based on the characteristics determined, three approaches were applied for predicting nucleosome positions. The results were compared with previously reported experimental data and predictions (3,13). Finally, the distance from the nucleosome dyad to the start and the stop codons was estimated.

MATERIALS AND METHODS

Data sets

The dinucleotide position-frequency data from yeast and chicken nucleosome-bound DNA sequences were retrieved (http://genie.weizmann.ac.il/pubs/nucleosomes06) (3). Each data set consisted of a matrix with 142 rows and 20 columns ($F_{142\times20}$), indicating the occurrence frequency of 20 dinucleotides (16 dinucleotides and 4 combination dinucleotides) at 142 positions within the core DNA sequence. The data sets served two purposes. First, periodicity analysis was used to explain the dinucleotide periodicity characteristics along the core DNA. Second, the score computation was used to predict nucleosome positions.

Submitted July 9, 2007, and accepted for publication January 18, 2008.

Address reprint requests to Xiao Sun, E-mail: xsun@seu.edu.cn.

Editor: Ruth Nussinov.

0006-3495/08/06/4597/08 $2.00

doi: 10.1529/biophysj.107.117028

To explore the relationship between DNA curvature and dinucleotide spacing, the crystal structure data of the DNA complex and its histone proteins were employed. Two crystal structure data sets (ID: 1ID3 and 1U35) were retrieved from the Protein Data Bank (PDB).

The nucleosome position predictions were tested on a 2200-bp DNA sequence (*S. cerevisiae* chr. II: 277,412-279,612 bp) as well as an entire chromosome (*S. cerevisiae* chr. I). Sequence data were downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org).

Methods

Analysis of dinucleotide position-frequency data and periodicity patterns

Dinucleotides of DNA sequences display 10–11-bp periodicity (9), and it is thought that the periodicity could vary along the core DNA sequence. To explore this, two wavelet-based methods, time frequency spectrum (TFS) and wavelet frequency spectrum (WFS) (15), were used to analyze the dinucleotide position-frequency data ($F_{142\times20}$). TFS has the capacity to provide both time and frequency information synchronously, resembling the short-time Fourier transform. WFS is similar to the Fourier power spectrum, but is more accurate and flexible.

First, each dinucleotide occurrence frequency data set (each row of $F_{142\times20}$) was analyzed with WFS to determine what frequency (periodicity) components were contained. Second, TFS was employed to detect how the periodicity components evolve along the positions. The row data were separated into three sections (two ends, each 50 bp, and one middle section) before performing TFS. In TFS, the periodicities of weakly bound dinucleotides (AA, AT, TA, TT, and its combinations) present a regular variation along the positions. The regular variation is termed the ''periodicity pattern'' (see Results and Discussion).

Curvature pattern of nucleosome-bound DNA sequence

The periodicity pattern reflects the regular spacing of weakly bound dinucleotides along the core DNA sequence. Because DNA bending is sequence dependent, the DNA curvature was subsequently investigated with a theoretical model and an estimation of the crystal structure data for confirmation (16–19). The theoretical DNA curvature was estimated with curvature vector C(Eq. 1), which is calculated with a matrix of roll ρ and tilt τ angles obtained for the 16 dinucleotide steps (16),

$$C = \nu^0 (n_2 - n_1)^{-1} \sum_{j=n_1}^{n_2} (\rho_j - i\tau_j) \exp\left(\frac{2\pi ij}{\nu^0}\right), \qquad (1)$$

where ν^0 is the double-helix average periodicity (10.4 bp) and number $(n_2 - n_1)$ represents the integration step. The modulus of the vector represents deviation from B-DNA.

On the other hand, the crystal structure data depict the central tract of the DNA double helix. Each section of the tract is an arc. The true curvature was estimated from the length ratio of the arc to its chord (Eq. 2),

$$C = l/d, \qquad (2)$$

where l and d are the lengths of the arc and its chord, respectively. The ratio is a number between 1 and $\pi/2$.

The results demonstrate that the curvature of nucleosome-bound DNA sequences also presented a typical variation, which is called the ''curvature pattern''.

Three approaches for predicting nucleosome position

By virtue of the periodicity and the curvature patterns, two prediction approaches were constructed. First, the patterns were represented as two pattern signals. Second, the periodicity and curvature profiles were calculated. The periodicity profile was the signal that described weakly bound dinucleotides' periodicities along a DNA sequence and was calculated as follows: the DNA sequence was scanned at 1-bp increments with a 100-bp sliding window. In each sliding window, the correlation function of weakly bound dinucleotides (C_{WW-WW}) was calculated (9). The correlation function was an oscillating signal that contained dinucleotide periodicity information. Here, WFS was performed on the correlation function to extract the preponderant periodicity values in the range of 9.5–12 bp. The periodicity profile recorded the periodicity value and the window central position.

The curvature profile provided a theoretical curvature value for the DNA sequence at each position. The scanning was performed in 1-bp steps with a 10-bp sliding window. In each window, the curvature was estimated with Eq. 1.

The final step was to predict the nucleosome positions. This task was achieved by recognizing the pattern signal from its corresponding profile. However, noise often hinders recognitions. Two strategies were used to overcome this problem: one was a wavelet-based denoise for profiles; the second was a convolution operation, which was performed on the pattern signal and the corresponding profile. If a profile's segment resembled the pattern signal, the convolution would peak at the corresponding position, indicating a nucleosome.

The third prediction approach estimated nucleosome positions by appointing a score that correlated to the extent of the dinucleotide position frequency of a DNA sequence compared to the core DNA sequence. The score was the sum of all product matrix elements from the dinucleotide position-frequency matrix ($F_{142\times16}$) of core DNA sequences and the binary indicator matrix ($D_{16\times142}$) of the DNA sequence (Eq. 3),

$$score = sum(F_{142\times16} \cdot D_{16\times142}), \qquad (3)$$

where *score* represents the score, matrix $F_{142\times16}$ denotes the dinucleotide position-frequency matrix, and matrix $D_{16\times142}$ was the binary indictor matrix

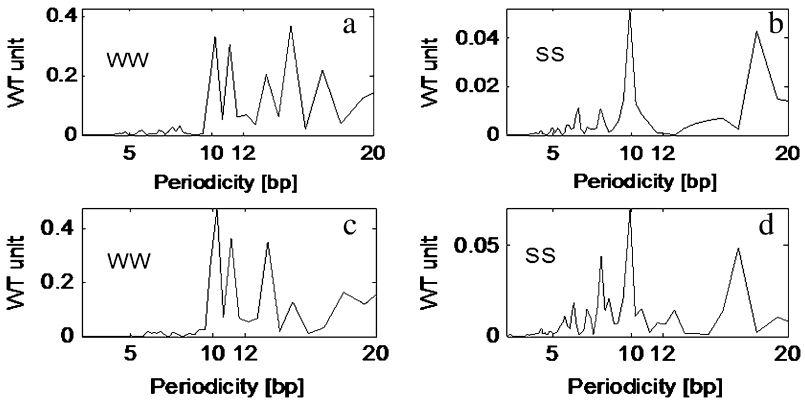

FIGURE 1 WFS of the dinucleotide position-frequency data in yeast and chicken nucleosome-bound DNA sequences. *a* and *b* represent yeast data; *c* and *d* represent chicken data. WW and SS indicate the weakly bound dinucleotides (AA-AT-TA-TT) and the strongly bound dinucleotides (GG-CC-GC-CG), respectively.

TABLE 1 Dinucleotide periodicities of yeast nucleosome-bound DNA sequences

Single peak	Dinucleotide	CA	CC	CG	TA	TT	TG	GC	GT	SS	CC-GG	GA-TC
	Periodicity (bp)	10.69	9.88	9.88	10.27	11.14	9.88	11.14	10.27	9.88	9.88	9.88
Doublet peaks	Dinucleotide	AA	AT	CT	GA	AA-TT-TA	WW	CA-TG				
	Periodicity (bp)	10.27	10.27	9.88	10.27	10.27	10.27	9.88				
		11.14	11.14	10.69	11.14	11.14	11.14	10.69				
No peak	Dinucleotide	AC	AG	GG	TC	AC-GT	AG-CT					

of the DNA sequence, where 1 indicated the presence of a certain dinucleotide and 0 indicated its absence, i.e., a 142-bp sliding window was used to scan the DNA sequence from start to end in 1-bp steps. In each window, the score was calculated with Eq. 3. The window's position and the score were recorded in a score profile. A high peak in the score profile indicated a nucleosome. Some dinucleotides might be more important for nucleosome positioning than others, as reflected in nucleosome sequence patterns. In previously reported studies, the dinucleotide pattern AA-TT (11), AA-TA-TT-GC (3), and WW-SS (13) have been used. The pattern RR-YY was also suggested (20). In this study, these patterns were tested. Before calculating the score profile of specific dinucleotide patterns using Eq. 3, the position-frequency matrix ($F_{142\times16}$) was reconstructed by conserving the specific dinucleotide elements and setting the other elements to 0, which eliminated negative effects (noise) of other dinucleotides in nucleosome positioning. The final score profile was smoothed by a multipoint moving average.

The three methods all employed a sliding window to compute the profiles. The periodicity profile consumed the majority of computation time, and the score profile required the least amount of time. The results predicted here were compared with experimental data from previously published studies (13) as well as algorithmic results from Segal and co-workers (3). Following comparison to previously reported studies (13), a true positive (TP) was defined if the shift between the prediction and the experimental data was <30 bp. However, if the shift was more than 30 bp, the prediction was a false positive (FP). Finally, if the prediction resulted in a 30-bp miss from the experimental position, it was defined as a false negative (FN). Positive accuracy and sensitivity were defined by Eq. 4 and Eq. 5, respectively.

$$\text{Positive accuracy} = \frac{TP}{TP + FP} \times 100 \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \tag{5}$$

When comparing with Segal and co-workers' algorithm (3), their predictions were used as the standard for defining *TP*, *FP*, and *FN*.

RESULTS AND DISCUSSION

Periodicity analysis for the dinucleotide position-frequency data

The WFS of dinucleotide position-frequency data from yeast nucleosome-bound DNA indicated that 16 dinucleotides and their combinations have different periodicity components in the frequency domain (Fig. 1, Table 1). Strongly bound dinucleotides, including CA, CC, CG, and their combinations, resulted in a single peak around 9.88 bp. Two weakly bound dinucleotides, TT and TA, also demonstrate a single peak. Interestingly, their positions were distinctly different: TT presented at 11.14 bp, whereas TA presented at 10.27 bp. An additional significant characteristic was the doublet peaks that were observed between 10 and 11 bp for the weakly bound dinucleotides AA, AT, and the combination AA-TT-TA. Some dinucleotides exhibited no distinct peaks in the range. Moreover, in some chicken nucleosome-bound sequences, doublet peaks were also found for dinucleotides AA, TA, TT, AA-TT-TA, and SS-WW.

The doublet peaks indicated that there were two frequency (periodicity) components in the original signal. Therefore, it was important to locate where the two components occurred in the time (position) plane. To this end, the weakly bound dinucleotide position-frequency data were processed with TFS; the result is shown in Fig. 2. Clearly, the two data ends indicated a relatively low periodicity, whereas the middle region resulted in a high periodicity. These results indicate that the two ends of core DNA sequences adopt a denser arrangement of weakly bound dinucleotides compared to the middle region. For this study, this was termed the ''periodicity pattern'' of core DNA. Considering that DNA bending corresponds to weakly bound dinucleotides, the significance of the characteristic was quite clear. If weakly bound dinucleotides had greater periodicity, the DNA sequence would have a smaller curvature, because widely spaced and weakly bound dinucleotides hinder the bend of a large curvature.

To confirm this result, the curvature of a nucleosome-bound DNA was estimated on the basis of the theoretical model as well as from the crystal structure data. Fig. 3 depicts the center tract of nucleosome-bound DNA (PDB ID: 1ID3) and the positions of nucleotide A and T on the helix. Clearly, the DNA backbone bent sharply into a helix to fit the histone protein shape. Furthermore, the weakly bound nucleotides

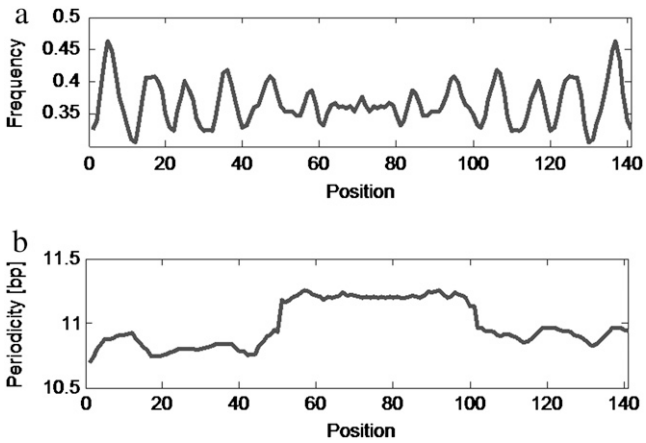

FIGURE 2 Position-frequency data of dinucleotide WW of yeast nucleosome-bound DNA sequences (*a*) and its TFS (*b*).

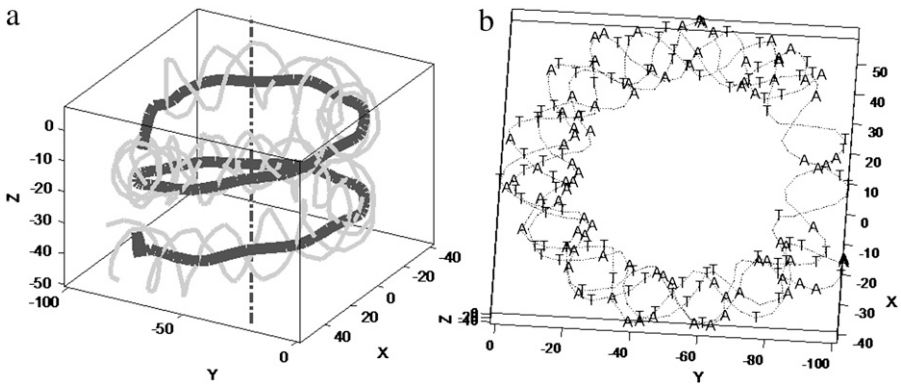

FIGURE 3 Center tract of a nucleosome-bound DNA (PDB ID: 1ID3) (*a*) and the distribution of nucleotide A and T on the backbone (*b*). In *a*, the gray line represents the backbone of the DNA double helix, the bold black line is the center tract of DNA, and the dotted line is the center axis of the complex of DNA and its histone. In *b*, the dotted line represents the DNA helix.

occurred frequently at the edge of the DNA helix. Fig. 4, *a* and *c*, demonstrates the curvatures of the yeast nucleosome-bound DNA sequence (PDB ID: 1ID3). Although the curves from the theoretical model and the estimation from the crystal structure data varied in some of the local positions, it was obvious that both curves had high values at the two ends and low values in the middle region. These results demonstrate that the two ends had a large curvature and the middle region had a small curvature, which provided powerful evidence for a periodicity characteristic of core DNA. The curvature characteristic was termed the ''curvature pattern''. Moreover, the curvature pattern not only was exhibited in an individual case but also was prevalent in many nucleosome-bound DNA sequences (Fig. 4, *b* and *d*, for human nucleosome-bound DNA (PDB ID: 1U35)). It was determined that the random sequence showed no such tendency (Fig. 4 *e*), which indicated that a bend path was unique for core DNA. Therefore, the curvature pattern can also be applied for predicting nucleosome positions.

To predict nucleosome positions, the periodicity pattern and the curvature pattern were represented with corresponding pattern signals (Fig. 5).

The predictions of nucleosomes positions

Nucleosome positions from the test sequence (*S. cerevisiae* chr. II: 277,412-279,612 bp) were predicted with three approaches. The DNA sequence consisted of the downstream portion of gene GAL10, an intergenic sequence with protein-binding sites, and the upstream portion of gene GAl1 (5′ to 3′). In Fig. 6 *a*, the previously predicted nucleosomes are represented by the white (3) and gray (13) ovals. The small ovals indicated the conserved and bound DNA-binding sites (21).

Fig. 6 *b* represents the denoised periodicity and Fig. 6 *c* the convolution periodicity profile. The arc form peaks in the convolution profile (Fig. 6 *c*) indicate stable nucleosomes; the large and sharp peaks indicate unstable nucleosomes or no nucleosomes. The nucleosome positions were predicted by combining the proofs of Fig. 6, *b* and *c*. The black ovals between the two subplots represented the predicted stable nucleosomes. Table 2 listed the prediction deviations from the experimental data (13) and Segal and co-workers' report (3). Compared with Segal and co-workers' predictions (3), five nucleosomes were correctly detected, with deviations <30 bp and positive accuracy of 55.56%. These results, however, do

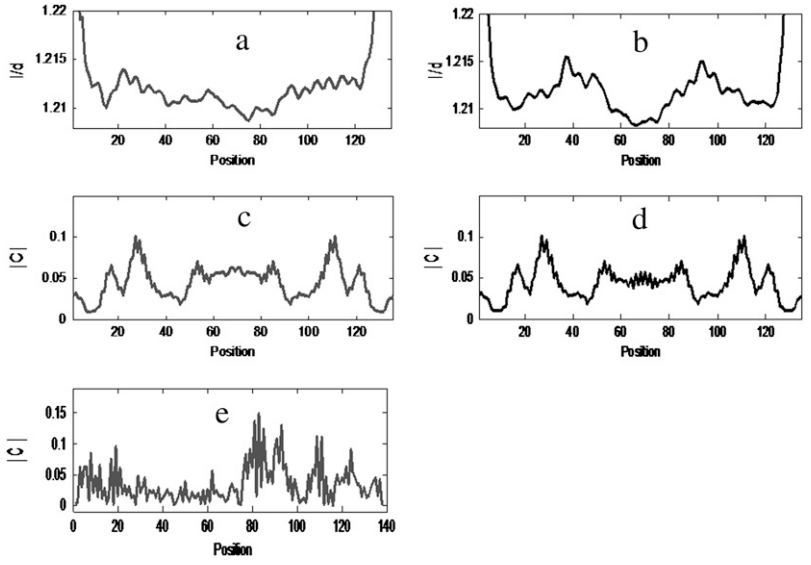

FIGURE 4 *a* and *b* are curvatures calculated from the crystal structure data, *a* for 1ID3 and *b* for 1U35. *c* and *d* are theoretical curvatures for DNA of 1ID3 and 1U35, respectively. *e* represents the theoretical curvature for a random DNA sequence.

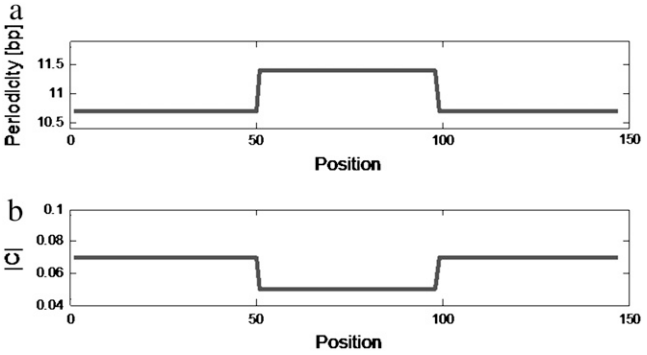

FIGURE 5 The periodicity pattern signal (*a*) and the curvature pattern signal (*b*).

not indicate that the proposed methods were less accurate than Segal and co-workers' method; the difference between them should be attributed to inaccuracies in both methods.

When compared to experimental data (13), the predictions appeared to contain more FPs, which most likely indicates that nucleosome positioning was determined by multiple factors, not a single sequence-dependent one. Interestingly, within the area surrounding the protein-DNA binding sites (between gene GAL1 and GAL10), both the denoised and the convolution profile displayed an irregular signal, indicating unstable nucleosomes. If the dinucleotide arrangement of a DNA sequence was not in accordance with core DNA sequences, the DNA sequence bends would not appropriately match the shape of the histone and would result in an unstable DNA complex and nucleosome. If a DNA sequence segment served as a particular protein's target site, the sequence was conserved in most cases. The conserved sequence might disrupt the dinucleotide position-frequency pattern and result in an unstable nucleosome. Therefore, the position of an unstable nucleosome could be a potential protein-binding site. Specifically, the irregular signals in a periodicity profile might be employed to predict potential protein-binding sites.

The curvature profiles are depicted in Fig. 6, *d* and *e*. Although they provide some correct predictions, the overlapping peaks hindered proper recognition. Taken together, this study concluded that the periodicity and curvature patterns are intrinsic properties of core DNA sequences. The prediction approaches that were based on these patterns are feasible; however, the results can still be improved.

The third approach predictions (score profiles) are depicted in Fig. 7. The performance of various dinucleotide patterns is summarized in Table 2. The score profiles of the YY-RR

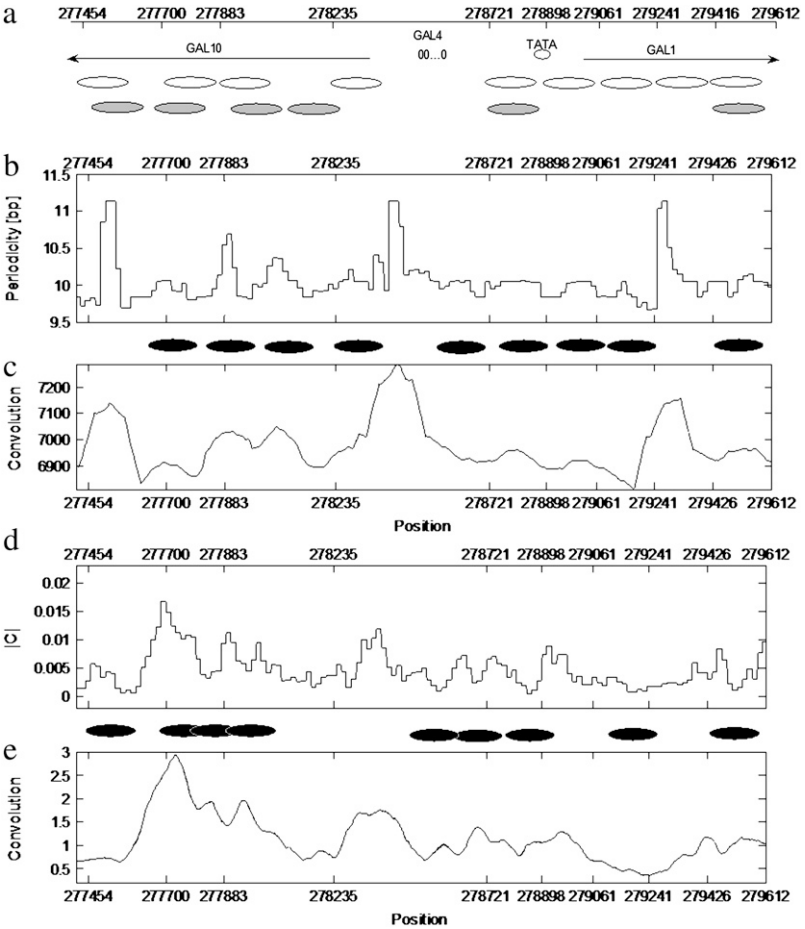

FIGURE 6 Predictions of nucleosome positions. (*a*) the test DNA sequence's structure and its nucleosome positions reported in previously published studies. White and gray ovals indicate the nucleosomes predicted in previously published studies (3,13), respectively. Small ovals indicate conserved and DNA-binding sites (14). *b* and *c* are the denoised periodicity profile and the convolution periodicity profile, respectively. Black ovals between *b* and *c* are the predicted nucleosomes. *d* and *e* are the predictions based on the curvature pattern.

TABLE 2 Prediction comparisons of the experimental data and reports in the literature

	Comparison with Segal and co-workers' (3) predictions ($p > 0.2$)							Comparison with the experimental data (13) (threshold >0.3)							
			Score profile							Predictions by Segal and	Score profile				
Method	Periodicity profile	Curvature profile	YY-RR	SS-WW	AA-TT	AA-TT-TA-GC	All	Periodicity profile	Curvature profile	co-workers (3) ($p > 0.2$)	YY-RR	SS-WW	AA-TT	AA-TT-TA-GC	All
TP (shift ≤30 bp)	5	3	7	5	5	4	4	2	3	4	2	3	4	2	2
FP (shift >30 bp)	4	6	3	4	4	5	5	7	6	5	8	6	5	7	7
FN (miss >30 bp)	2	2	1	1	0	0	2	1	0	1	0	1	0	1	1
Sensitivity (%)	71.43	60.00	87.50	83.33	100.00	100.00	66.67	66.67	100.00	80.00	100.00	75.00	100.00	66.67	66.67
Positive accuracy (%)	55.56	33.33	70.00	55.56	55.56	44.44	44.44	22.22	33.33	44.44	20.00	33.33	44.44	22.22	22.22

pattern and the 16-dinucleotide pattern exhibited more noise because they required a larger smooth window (41 point). Compared with previously reported data (13), the AA-TT pattern displayed the highest positive accuracy (44.44%) and sensitivity (100%), followed by the AA-TT-TA-GC pattern and SS-WW pattern. These three patterns generated similar score profiles. The YY-RR and 16-dinucleotide patterns resulted in more overlapping peaks, with an unclear baseline. In addition, the score profiles were different from the other three patterns. These results indicate that dinucleotide AA-TT played an important role in nucleosome positioning. After the addition of dinucleotide TA-GC to AA-TT or the use of SS-WW, predictions were not improved. In addition, the use of the YY-RR pattern introduced even more noise.

A large-scale prediction was completed on an entire chromosome (*S. cerevisiae* chr. I). Table 3 lists the performance. The periodicity profile was slightly better than the curvature profile, and both proposed methods resulted in better perfor-

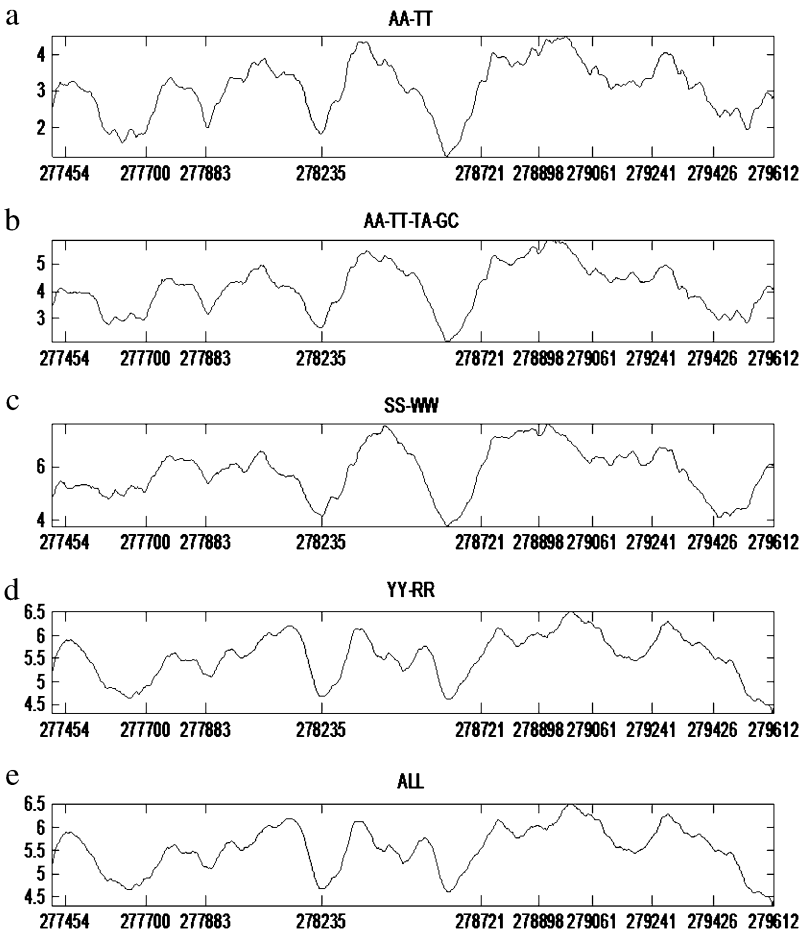

FIGURE 7 Predictions of nucleosome positions by the score profile. *a*, *b*, *c*, and *d* are the score profiles of specific dinucleotide patterns AA-TT, AA-TT-TA-GA, SS-WW, and YY-RR, respectively. *e* is the score profile based on 16 dinucleotides. A 47-point smoothing was used for *d* (YY-RR) and *e* (16 dinucleotides); other score profiles were smoothed with 21 points.

TABLE 3 Prediction performance of the periodicity profile and the curvature profile for *S. cerevisiae* chr. I

	Positive accuracy	Sensitivity	Averaged error of TP	Resolution
Periodicity profile	59.76%	70.34%	15 bp	212 bp
Curvature profile	59.51%	69.53%	20 bp	235 bp
Predictions by Segal and co-workers (3) ($p > 0.2$)	42.32%	68.04%	29 bp	172 bp

Prediction resolution refers to the most frequent distance of neighboring nucleosomes.

mance than previously published studies (3). In addition, nucleosome distributions near the start and stop codons were analyzed for 71 verified ORFs of chr. I; the results are shown in Fig. 8. The number of nucleosomes that were present at varying distances, from either an ORF start (Fig. 8, *a–c*) or an ORF end (Fig. 8, *d–f*), were binned in 10-bp intervals then plotted as a seven-point smoothed frequency distribution. Fig. 8, *a* and *d*, depicts the experimental data (13); Fig. 8, *b* and *e*, shows the predictions using the periodicity profile; and Fig. 8, *c* and *f*, demonstrates the predictions from Segal and co-workers' study (3).

Fig. 8, *a* and *b*, indicates that some minor peaks still remain separate from the main peaks, which coincides with other published studies (4,11,13). Fig. 8 *c*, however, demonstrates a lack of minor peaks. Although Fig. 8 *b* differs from Fig. 8 *a* at the local region (~ -250 to ~ -100 bp), our impression was that Fig. 8 *b* was more similar to Fig. 8 *a* than to Fig. 8 *c*. With respect to the nucleosome-free region, Fig. 8 *b* exhibits differing results from the experimental data. In previous studies (11), it was concluded that the nucleosome distributions presented substantial differences in the region spanning -300 to -80 bp, in terms of TATA-less and TATA-containing genes. Even for TATA-containing genes, the nucleosome distributions of three gene clusters differed. Moreover, Fig. 8 *b* demonstrates an aggregate nucleosome distribution of chr. I. In this respect, the characteristics of the nucleosome-free region shown in Fig. 8 *b* are reasonable.

In Fig. 8 *b* (and Fig. 8 *a*), there is a peak at ~ -27 bp, which indicates that the ORF start sites were occupied by nucleosomes; these results have also been previously reported (4,11,13). Similarly, most ORF end sites were also occupied by nucleosomes at ~ -13 bp upstream (Fig. 8, *d* and *e*). This finding, however, has not been previously reported. Nucleosome density in various gene regions was also investigated. As previously shown (2), nucleosomes possess different densities at ORFs and intergenic regions. Intergenic regions had a low occupying ratio ($\sim$10.4%), whereas ORFs presented a higher occupying ratio ($>$32.7%).

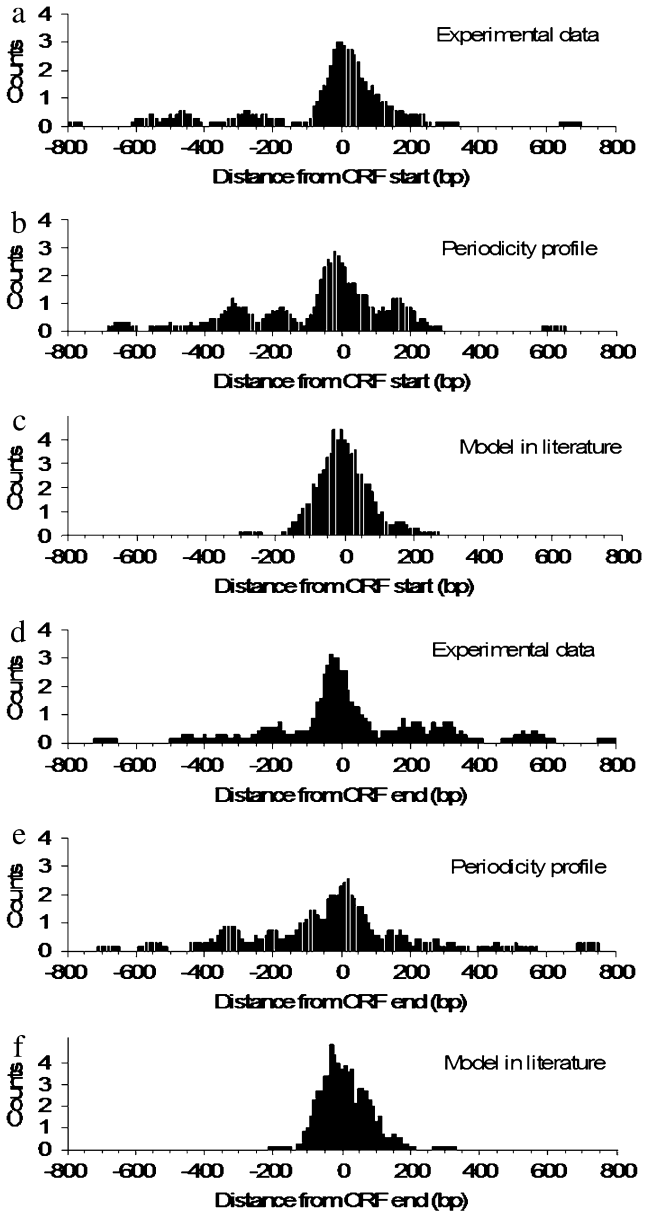

FIGURE 8 Nucleosome distribution near the ORF start and ORF end. The number of nucleosomes located at varying distances from either an ORF start (*a*, *b*, and *c*) or an ORF end (*d*, *e*, and *f*) were binned in 10-bp intervals then plotted as a smoothed frequency distribution. *a* and *d* are calculated with nucleosome positions determined from experiments (13); *b* and *e* are calculated with those predicted by the periodicity profile; *c* and *f* are calculated according to Segal and co-workers' (3) predictions.

CONCLUSIONS

This study represents a thorough investigation of the dinucleotide periodicity characteristics of nucleosome-bound DNA sequences. Weakly bound dinucleotides of core DNA sequences were spaced smaller at the two ends, with larger spacing in the middle section. The periodicity pattern was also reflected in the curvature of the DNA sequence. Based on these findings, three approaches were constructed to predict nucleosome positions, and the results from this study confirmed that these approaches were feasible. Using the prediction data, nucleosome distributions were examined

near the start and stop codons, as well as the nucleosome densities in ORFs and the intergenic regions.

The periodicity and curvature patterns were determined to be intrinsic properties of core DNA sequences; they were prevalent in almost all core DNA sequences. The periodicity profile is derived from dinucleotide frequencies. Therefore, if it is applied to new genomic sequences, new dinucleotide position-frequency data are still needed to refine the characteristic.

We thank the anonymous reviewers for their suggestions to improve the article's quality.

This work was supported by the Natural Science Foundation of China (60671018, 60121101), the National Postdoctoral Foundation (20070420959), and the Jiangsu Planned Projects for Postdoctoral Research Funds (1660631153).

REFERENCES

1. van Holde, K. E. 1989. Chromatin. Springer, New York.
2. Lewin, B. 2004. Gene VIII. Prentice Hall, Chap. 20.
3. Segal, E., Y. F. Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J. Z. Wang, and J. Widom. 2006. A genomic code for nucleosome positioning. *Nature*. 442:772–778.
4. Yuan, G. C., Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. 2005. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*. 39:626–630.
5. Audit, B., C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes. 2002. Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J. Mol. Biol.* 316:903–918.
6. Li, S., and M. J. Smerdon. 2002. Nucleosome structure and repair of *N*-methylpurines in the GAL1–10 genes of Saccharomyces cerevisiae. *J. Biol. Chem.* 277:44651–44659.
7. Ioshikhes, I., A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E. N. Trifonov. 1996. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* 262:129–139.
8. Thåström, A., P. T. Lowary, H. R. Widlund, H. Cao, M. Kubista, and J. Widom. 1999. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* 288:213–229.
9. Schieg, P., and H. Herzel. 2004. Periodicities of 10–11bp as indicators of the supercoiled state of genomic DNA. *J. Mol. Biol.* 234:891–901.
10. Herzel, H., O. Weiss, and E. N. Trifonov. 1999. 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*. 15:187–193.
11. Ioshikhes, I., I. Albert, S. J. Zanton, and B. F. Pugh. 2006. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* 38:1210–1215.
12. Salih, F., B. Salih, and E. N. Trifonov. 2007. Sequence-directed mapping of nucleosome positions. *J. Biomol. Struct. Dyn.* 24:429–514.
13. Albert, I., T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, and B. F. Pugh. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*. 446:572–576.
14. Peckham, H., R. Thurman, Y. Fu, J. Stamatoyannopoulos, W. Noble, K. Struhl, and Z. Weng. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res.* 17:1170–1177.
15. Lu, X. Q., H. D. Liu, J. W. Kang, and J. Chen. 2003. Wavelet frequency spectrum and its application in analyzing oscillating chemical system. *Anal. Chim. Acta*. 484:201–210.
16. Cacchione, S., P. D. Santis, D. Foti, A. Palleschi, and M. Savino. 1989. Periodical polydeoxynucleotides and DNA curvature. *Biochemistry*. 28:8706–8713.
17. Santis, P. D., A. Palleschi, M. Savino, and A. Scipioni. 1990. Validity of the nearest-neighbor approximation in the evaluation of the electrophoretic manifestations of DNA curvature? *Biochemistry*. 29:9269–9273.
18. Widlund, H. R., P. N. Kuduvalli, M. Bengtsson, H. Cao, T. D. Tulliusi, and M. Kubista. 1999. Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J. Biol. Chem.* 274:31847–31852.
19. Olson, W. K., A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*. 95:11163–11168.
20. Kogan, S., M. Kato, R. Kiyama, and E. N. Trifonov. 2006. Sequence structure of human nucleosome DNA. *J. Biomol. Struct. Dyn.* 24:43–48.
21. Harbison, C. T., D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, K. Manolis, P. A. Rolfe, and R. A. Young. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 431:99–104.