A Novel Wavelet-Based Approach for Predicting Nucleosome Positions Using DNA Structural Information

Yanglan Gan, Guobing Zou, Jihong Guan, and Guangwei Xu

Abstract—Nucleosomes are basic elements of chromatin structure. The positioning of nucleosomes along a genome is very important to dictate eukaryotic DNA compaction and access. Current computational methods have focused on the analysis of nucleosome occupancy and the positioning of well-positioned nucleosomes. However, fuzzy nucleosomes require more complex configurations and are more difficult to predict their positions. We analyzed the positioning of well-positioned and fuzzy nucleosomes from a novel structural perspective, and proposed WaveNuc, a computational approach for inferring their positions based on continuous wavelet transformation. The comparative analysis demonstrates that these two kinds of nucleosomes exhibit different propeller twist structural characteristics. Well-positioned nucleosomes tend to locate at sharp peaks of the propeller twist profile, whereas fuzzy nucleosomes correspond to broader peaks. The sharpness of these peaks shows that the propeller twist profile may contain nucleosome positioning information. Exploiting this knowledge, we applied WaveNuc to detect the two different kinds of peaks of the propeller twist profile along the genome. We compared the performance of our method with existing methods on real data sets. The results show that the proposed method can accurately resolve complex configurations of fuzzy nucleosomes, which leads to better performance of nucleosome positioning prediction on the whole genome.

Index Terms—Nucleosome positioning, structural feature, continuous wavelet transformation, genome analysis

1 Introduction

Eukaryotic genomes are packaged into condensed chromatin structures, whose fundamental building units are nucleosomes. Each nucleosome is composed of a histone octamer wrapped by around 147 base pairs (bp) of DNA. The organization of nucleosomes can be described as an array of nucleosome units across the genome. On one hand, it provides the background of various histone modifications and variant turnovers [1]. On the other hand, by modulating the accessibility of underlying DNA sequences [2], nucleosome positioning directly or indirectly affects a variety of cellular processes, such as transcription factor binding kinetics, DNA replication and gene splicing [3]. To elucidate the complex interactions between chromatin and transcription factors, it is critical to understand how nucleosome positioning is established along the genome.

In different cells, the exact positions of nucleosomes are difficult to be determined. A part of nucleosomes are centering around the most preferred positions, while others may deviate more or less from their centers [4]. This deviation of nucleosome positions in a cell population is referred to as fuzziness. The fuzziness of nucleosomes can partly reflect the dynamics of nucleosome positioning, which provides the flexibility of responding to different environmental or physiological changes [5]. Studying fuzzy nucleosomes will facilitate the understanding of transcription regulation process through the complex interactions between chromatin and various transcription factors. As a result, the current challenge is to develop analytic tools to systematically explore the distribution characteristics of well-positioned and fuzzy nucleosomes, and further decipher the underlying mechanisms of fuzzy nucleosomes positioning.

The development of high-throughput sequencing techniques has substantially aided in the investigation of nucleosome positioning across the genomes of various model organisms [6], [7], [8]. The procedure for high-resolution mapping of nucleosomes involves an initial step to cross-link histones to nucleosomal DNA by formaldehyde treatment of living cells. Next, linker DNA is removed from isolated chromatin by MNase digestion [3]. Then, microarray or massive parallel sequencing techniques are adopted to determine nucleosome mappings. Based on the experimental nucleosome mappings, analyzing the involvement of DNA sequence on nucleosome positioning and occupancy is of great biological interest. Recently, the underlying mechanisms of chromatin organization are still under debate [9], nucleosome positioning is thought to be determined by a multitude of factors [10], including chromatin remodelers [11],

- *Y. Gan is with the School of Computer Science and Technology, Donghua University, Shanghai, China. E-mail: ylgan@dhu.edu.cn.*
- *G. Zou is with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. E-mail: gbzou@shu.edu.cn.*
- *J. Guan is with the Department of Computer Science and Technology, Tongji University, Shanghai, China. E-mail: jhguan@tongji.edu.cn.*
- *G. Xu is with the School of Computer Science and Technology, Donghua University, Shanghai, China, and with the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China. E-mail: dh.xuguangwei@gmail.com.*

Manuscript received 29 Dec. 2013; accepted 24 Jan. 2014. Date of publication 19 Feb. 2014; date of current version 4 Aug. 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2306837

specific DNA-binding proteins [12], and intrinsic DNA sequence preferences [13]. Previous experimental and bioinformatic studies have concerned about how and to what extent sequence compositional features can contribute to nucleosome organization [14]. In particular, AT and GC-rich dimeric and trimeric motifs were identified affecting nucleosome occupancy [15], [16]. Subsequently, several studies delineated that periodicity of dinucleotides and particular sequence patterns are associated with nucleosome-enriched sequences [7], [17]. Computational methods based on such sequence compositional features have been proposed to predict nucleosome occupancy and positioning [18], [19], [20]. Recently, another strategy to predict the propensity of nucleosome formation has gained much attention, which is based on the sequence-specific energy properties [21], [22]. These approaches are quite accurate for predicting the nucleosome occupancy level and are effective in inferring the positions of well-positioned nucleosomes. However, they still have great difficulty in properly handling more complex nucleosome positioning such as fuzzy nucleosomes.

Among various transcription factor binding sites (TFBSs) and other features thought to influence nucleosome positioning, the propeller twist structural property is regarded as the most correlated with nucleosome occupancy level [23]. It describes the angle of two aromatic bases in a base pair [24]. In general, dinucleotides with higher negative propeller twist angles are more rigid than those with smaller propeller twist values. Several previous studies have adopted this structural property to characterize nucleosome-depleted promoter regions [25]. The structural profile of propeller twist exhibits two distinct troughs in promoter sequences, where RNA polymerase II binds and nucleosome lacks [26], [27]. This observation is consistent with the finding that the rigidity of genomic sequences plays an important role in nucleosome exclusion from promoter regions [23]. In addition to promoter prediction, these studies also suggest that the propeller twist structural feature provides a powerful key to understanding the flexibility of DNA in forming nucleosomes.

In this article, we investigated the positioning of well-positioned and fuzzy nucleosomes from a novel structural perspective. In terms of the propeller twist structural feature, the systematic comparative analysis demonstrated that the two kinds of nucleosomes exhibit different structural characteristics. Well-positioned nucleosomes tend to locate at sharp peaks of propeller twist profile, whereas the positions of fuzzy nucleosomes correspond to broader peaks. The different signature of these peaks shows that the propeller twist structural profile may contain nucleosome positioning information. Exploiting this knowledge, we proposed a new computational approach based on continuous wavelet transformation (CWT) to predict nucleosome positioning, called WaveNuc. The application of WaveNuc to comprehensively search for the two kinds of peaks of the propeller twist structural profile along the genome. We compared the performance of our method on real data set against existing methods. The results demonstrate that our method can predict nucleosome positioning along the genome more accurately, especially can resolve complex configurations of fuzzy nucleosomes.

2 Methods

2.1 The Flowchart of WaveNuc Approach

Based on the propeller twist structural property of genomic sequences, a novel approach, named WaveNuc, is proposed to predict the most likely positions of nucleosomes, including well-positioned and fuzzy nucleosomes. Given a set of genomic sequences, the prediction of nucleosome positions is composed of three major steps. First, the genomic sequences are converted into the propeller twist structural profiles. Second, a peak detection method is utilized to comprehensively search for the significant peaks of the propeller twist profiles. Third, the positions of well-positioned and fuzzy nucleosomes are predicted according to the distinct signatures of peaks. Specifically, taking the advantage of multi-resolution analysis, the continuous wavelet transformation is introduced into the peak detection process. Thus, our approach can detect both sharp and broad peaks with a high degree of accuracy. The flowchart of the proposed approach is illustrated in Fig. 1.

2.2 Calculating the Propeller Twist Structural Profiles of Genomic Sequences

For a given genomic sequence se, the propeller twist structural profile is calculated in two steps. First, we scan the genomic sequence once and convert the DNA sequence into a numerical vector by replacing each dinucleotide with a numerical structural value. This conversion is based on experimentally determined structural model [25], in which 16 kinds of dinucleotide correspond to different propeller twist values. In the second step, we use a moving average window to smooth the raw structural profile, with a window size of 100 bp and a step of 10 bp. This process is aimed to initially remove noise signals in the structural profile of a single sequence. In this way, the final structural profile $s(t)$ of a genomic sequence is a vector of numerical values of the propeller twist property, which can be passed directly to the next peak detection procedure. By adjusting the moving step length, we can obtain a structural profile at a different resolution as needed.

2.3 Continuous Wavelet Transformation

Having the calculated and smoothed propeller twist profiles of genomic sequences, systematical comparative analysis is further conducted between the structural profiles of nucleosome-enriched and linker sequences (see the Section 3.1). It is noticed that nucleosome-enriched sequences and linker sequences exhibit different propeller twist signatures. Linker sequences usually do not contain any high signal in its structural profile, whereas nucleosomes locate at the significant peaks of the propeller twist profile. Specifically, the structural profiles of well-positioned nucleosomes exhibit narrow and sharp peaks. On the contrary, fuzzy nucleosomes exhibit relatively broad and blurring peaks. Based on these observations, the possible nucleosome positions are predicted through identifying the characteristic peak patterns in the propeller twist profile along the genomic sequence.

Due to the ability of multi-resolution analysis, the continuous wavelet transformation is applied to detect the significant peaks of structural profiles. CWT is one class of wavelet transformation methods. Different from the discrete

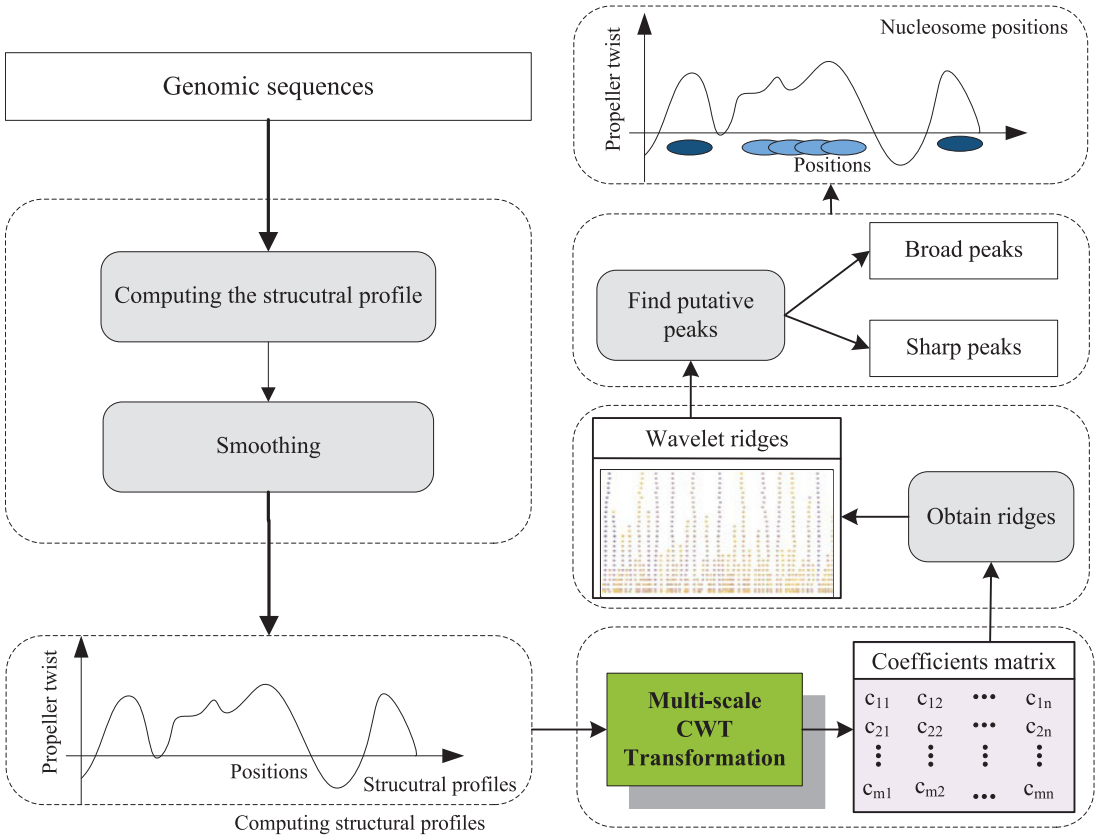


Fig. 1. Flowchart of the proposed WaveNuc approach. WaveNuc takes a set of genomic sequences as input, and outputs the possible nucleosome positions. It is composed of three major steps. First, the genomic sequences are converted into the propeller twist structural profiles, which are represented by the corresponding numerical vectors. Second, a peak detection method is utilized to search for the significant peaks of the propeller twist profiles. Third, the positions of well-positioned and fuzzy nucleosomes are predicted according to the distinct signatures of peaks.

wavelet transformation (DWT) operating over scales and positions based on the power of two, the CWT allows wavelet transformations at every scale with continuous translation [28]. As a result, the CWT can provide some redundancy, which makes the information available in peak shape and strength of the structural profiles much easier to be interpreted.

A wavelet is formulated as a function $\psi(t)$ in $L_2(R)$, such that $\psi(t)$ satisfies:

$$\int_{-\infty}^{+\infty} \psi(t)d(t) = 0. \tag{1}$$

For the detection of peak patterns, the selection of an appropriate mother wavelet is crucial for the CWT. In order to capture peak changes in the width and height, the mother wavelet should have the basic features of a peak, including one major positive peak and approximate symmetry. Based on the preliminary analysis, the shape of the Mexican Hat wavelet provides the best match to the peaks of the propeller twist structural profile. Therefore, the Mexican Hat wavelet is selected as the mother wavelet, which is proportional to the second derivative of the Gaussian probability density function [29]:

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}}(1-t^2)e^{-t^2/2}. \tag{2}$$

As a result, the CWT can be represented as :

$$C(a,b) = \int_R s(t)\psi_{a,b}(t)dt, \tag{3}$$

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \quad (a,b \in R, a>0), \tag{4}$$

where t is the genomic location, $s(t)$ is the propeller twist structural profile, a is the scale, b is the translation, $\psi(t)$ is the mother wavelet, $\psi_{a,b}(t)$ is the scaled and translated wavelet, and C is the 2-D matrix of wavelet coefficients.

Intuitively, the resulted CWT coefficients reflect the extent of pattern matching between the structural profile $s(t)$ and the wavelets $\psi_{a,b}(t)$. Higher coefficients indicate better matching. By varying a and b, the wavelets $\psi_{a,b}(t)$ can match different-scale peak patterns at different positions without complicated nonlinear curve fitting.

2.4 Detecting Peaks of the Structural Profiles

Based on the CWT, here we elaborate how to detect significant peaks of the propeller twist structural profile. The right part of Fig. 1 shows the workflow of the peak detection procedure. The procedure further break down into three specific substeps. (i) The structural profile $s(t)$ is transformed using continuous wavelet transformation at different scales. (ii) The local maxima coefficients of each scale are

determined, and they are further connected with nearby local maxima points at neighboring scales to form sequences of local maxima along scales. (iii) Remembering that a peak in the propeller twist profile is defined in the CWT coefficients as a local maximum sequence, the sequences of local maxima are subsequently selected according to given criteria. The maximum value of the local maximum sequence is identified as the peak. In the following, each substep is described in detail.

Computing the CWT coefficients. As any signal can be described as a function of individual periodic wavelets with different scales, a multi-scale CWT is performed on the propeller twist profile $s(t)$ at 37 scale levels:

$$A = \{1, 2, \ldots, 10; 12, 14, 16, \ldots, 62, 64\}.$$

The CWT coefficients are represented as matrix $C_{m\times n}$, where $m = 37$ is the number of scales used, and n is the length of the structural profile. Varying the scale can yield different-width wavelets $\psi_{a,b}(t)$. As the CWT coefficients contain the patterns of peaks, the change in the coefficients over different scales provides additional information for pattern matching. It can be exploited to determine the positions and significance of peaks, whose patterns are similar to $\psi_{a,b}(t)$ in the structural profile $s(t)$.

Creating ridges by linking local maxima. The local maxima of the CWT coefficients at each scale i are individually identified in the coefficients matrix. Each local maxima and its position are formed as a two-tuple $< lm_{i,j}, p_{i,j} >$. In the process, a window size w is defined, which is proportional to the wavelet support region at this scale. First, the global maximum of the coefficients is searched and all the positions within width w are marked invalid. Then, the next global maximum in the remaining valid regions is identified and its neighboring region is marked. This process iterates until no valid regions are left. Consequently, for the scale i, all identified maxima form the set of local maxima:

$$(< lm_{i,1}, p_{i,1} >, < lm_{i,2}, p_{i,2} >, \ldots, < lm_{i,k_i}, p_{i,k_i} >),$$

where k_i is the number of local maxima at scale i.

Based on the detected local maxima set of all 37 scales, the ridge lines can be created by linking the local maxima of CWT coefficients at each scale level. Here, the local maxima of the largest scale m are used to initialize the ridge lines. Meanwhile, since there might be lack of local maxima in adjacent scales, it is necessary to record the number of consecutive scales skipped by ridge lines. As a result, a gap parameter g is recorded for each ridge line.

The initial ridge set is represented as:

$$R = \left\{ \begin{array}{l} r_1 = (< lm_{m,1}, p_{m,1} >, g_1 = 0) \\ r_2 = (< lm_{m,2}, p_{m,2} >, g_2 = 0) \\ \ldots \\ r_{k_m} = (< lm_{m,k_m}, p_{m,k_m} >, g_{k_m} = 0) \end{array} \right\}.$$

For each ridge line r_j, the nearest local maxima at the next subsequent scale is identified within the window size w. If there are nearest points found, the two closest local maxima are linked as lines. Otherwise, the gap number g_j of the ridge line is increased by one. When the gap number of a ridge line exceeds six, this ridge line is marked as static and is no longer extended in the next round. At the current scale, the local maxima that are not linked to any point at the previous scale, they will be initiated as new ridge lines. These procedures repeat until the lowest scale *1*. Subsequently, these local maxima are linked as ridge lines.

To identify the final ridge line set, the next step is to filter those inappropriate ridge lines which may not converge to significant peaks. We calculate the ratio of the estimated signal strength and the local noise level. The signal/noise ratio (SNR) of each ridge line r_j is computed as:

$$SNR(r_j) = \frac{maxR_j}{noiseR_j}, \tag{5}$$

$maxR_j =$ the maximal coefficients of r_j,

$noiseR_j = 0.95$ quantile of the coefficients at scale 1.

More specifically, the noise value $noiseR_j$ is computed as 0.95 quantile of the coefficients at scale *1*, in the 150 bp region centering around the ridge line r_j. In order to determine an effective $SNR(r_j)$ threshold, different values in the range [0.1, 1] are tried. The peak detection method performs best when $SNR(r_j)$ is around 0.3, which is selected as the threshold. If the $SNR(r_j)$ is below 0.3, the ridge line r_j is deleted. Using this function, those small and noisy fluctuations are easily removed. Thus, the resulted set only include the ridge lines that likely converge toward peaks in the structural profile, as shown in Fig. 2a.

Identifying the peaks based on the ridge lines. Further, the peak positions are determined based on the maximum CWT coefficients of the ridge lines within certain scale range. In order to identify all significant peaks, two rules are applied:

R1. The scale determines the support region of wavelets. To identify significant peaks with certain width and strength, the scale corresponding to the maximum value on the ridge line should be within a certain range;

R2. The length of ridge lines should be larger than a certain threshold, which provides a good indication of peaks with different heights and widths. The major peaks correspond to long and high ridges, while the small peaks correspond to short and low ridges.

In some part of the structural profiles, there are small peaks existing around the major peaks. By reducing the length threshold of ridge lines, the small peaks in the surrounding region of major peaks can be easily identified. In Fig. 2b, the identified peaks of the structural profile are marked by red plus signs at the peak maxima positions, including both the major peaks and the nearby small peaks.

2.5 Predicting Nucleosome Positioning Based on the Detected Significant Peaks

Due to the difference between the peak signatures of well-positioned and fuzzy nucleosomes, the peak centers can not directly represent the nucleosome positions. Well-positioned nucleosomes tend to locate at narrow and sharp peaks of the propeller twist profile. Conversely,

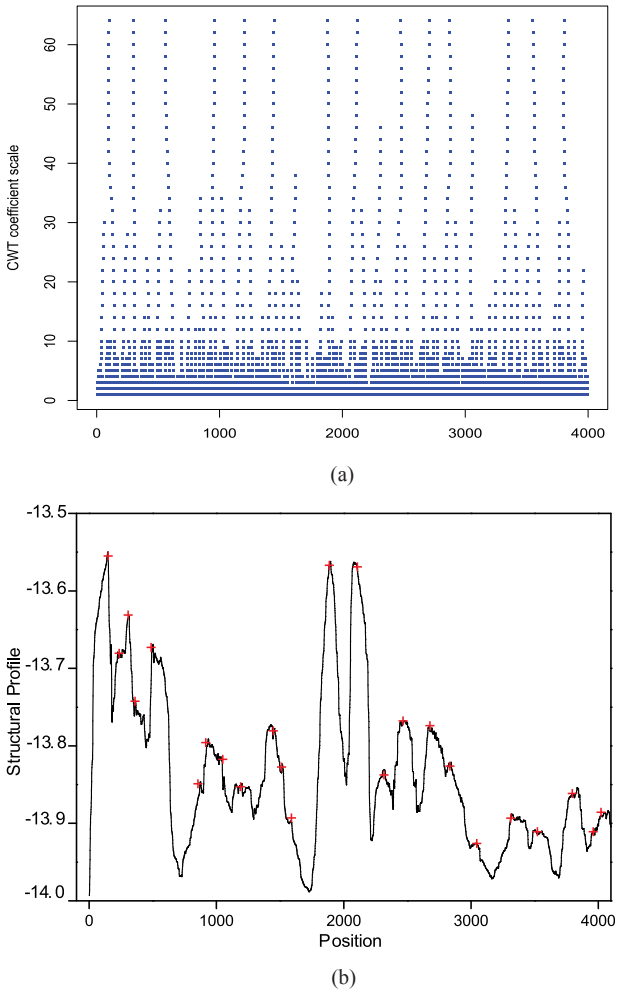


Fig. 2. Example of identifying peaks of the propeller twist profile. (a) Plots of the ridge lines likely converging toward peaks. (b) The identified peaks marked by red plus signs.

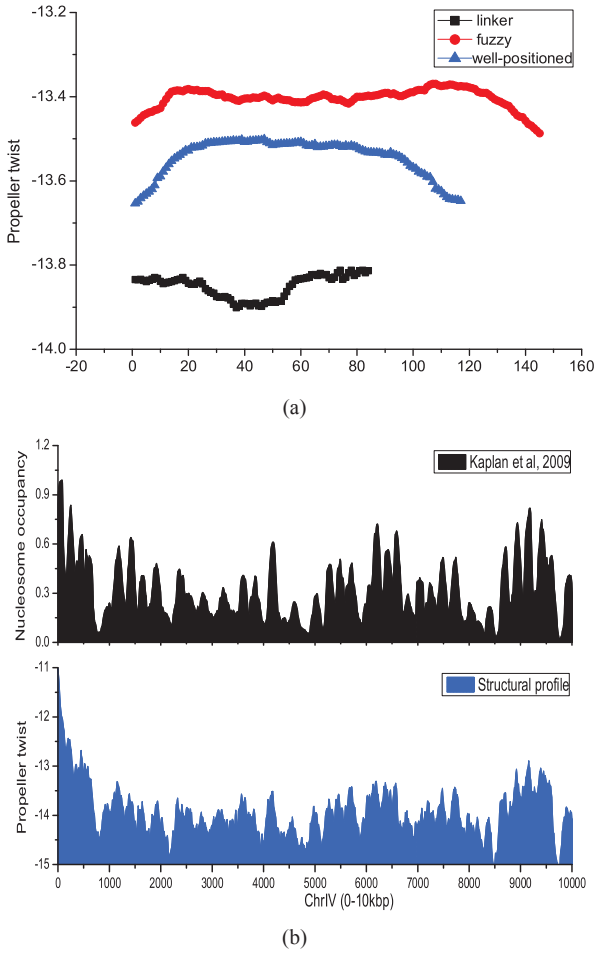


Fig. 3. Propeller twist structural profile analysis. (a) Structural profiles of linker sequences, well-positioned and fuzzy nucleosomes. (b) The comparison between the structural profile and experimental nucleosome occupancy.

fuzzy nucleosomes locate at broad and blurring peaks. The broad peaks are usually comprised of major peaks surrounded by several small peaks, which are commonly considered as a preferred position and several shifted positions. This is in accordance with previous studies [5]. Specifically, molecular biologists distinguish fuzzy nucleosomes from overlapping nucleosomes. In the overlapping case, the overlap between neighboring nucleosomes is a small fraction of their size. For the fuzzy nucleosomes, nucleosomes are mutually overlapping for a significant fraction [30]. By introducing a threshold parameter on the allowed overlapping, we can define precisely fuzzy nucleosomes.

Here, to further account for broad peaks seen in the structural profile of fuzzy nucleosome sequences, our algorithm set the overlapping threshold as 30 percent. If the centers of two peaks make two possible adjacent nucleosomes overlap more than 30 percent [4], the two peaks are merged to form a broader peak. After iterating the merging process, all possible peaks are finally determined. The locations of those peaks are then determined as the nucleosome positions. Meanwhile, the nucleosomes corresponding to the merged broad peaks are labeled as fuzzy, while other nucleosomes are considered as well-positioned.

2.6 Data

The analysis is performed using publicly available genome-scale data sets. Nucleosome occupancy and positioning data are obtained from previous experimental studies. In these studies, MNase assay is used for the digestion of genomic sequences, and then high throughput sequencing techniques are adopted to determine nucleosome occupancy and positioning. We retrieve in vivo and in vitro nucleosome occupancy data of S. cerevisiae [13], [31]. Two experimental maps of the nucleosome locations of the S. cerevisiae are respectively obtained from the author's website [23], [32]. The whole S. cerevisiae genome sequences are retrieved from the 2006 assembly of the Saccharomyces Genome Database [33].

3 Results

3.1 The Propeller Twist Structural Profile is Highly Predictive of Nucleosome Occupancy

In yeast, we first compared the average structural profiles of propeller twist on nucleosome-depleted and nucleosome-enriched sequences, including fuzzy and well-positioned nucleosomes. We observed that their structural profiles exhibit different signatures, as presented in Fig. 3a. The propeller twist structural profile of nucleosome-depleted

sequences exhibits a trough. In contrast, the nucleosome-enriched sequences correspond to significant peaks. Meanwhile, there is obvious difference between the peak patterns of well-positioned and fuzzy nucleosome sequences. Well-positioned nucleosome sequences tend to locate at significant narrow and sharp peaks of the propeller twist profile, whereas the peak patterns of fuzzy nucleosomes are much broader.

To explore the relationship between nucleosome occupancy and the propeller twist property of genomic sequences, we compared the structural profile and nucleosome occupancy along the whole genome [31]. The genome-wide nucleosome occupancy was determined by using parallel sequencing techniques. The nucleosome intensity signals were represented as log ratio between nucleosomal DNA and genomic DNA, showing nucleosomes as peaks of about 150 bp long, surrounded by lower values corresponding to linker regions. We observed a significant consistency between the structural profile and experimental nucleosome occupancy data. Here, we showed the results of a representative segment on chromosome IV. In Fig. 3b, the values for the experimental nucleosome occupancy data represent the nucleosome coverage along the genomic sequence. The oscillations of the calculated propeller twist structural profile coincide with those of nucleosome occupancy in different genomic regions, including the coding regions and intergenic regions.

Quantitatively, we calculated the correlation between the propeller twist profile and experimental in vitro and in vivo nucleosome occupancy on the genome-wide scale. The Pearson correlation coefficients are respectively about 0.82 and 0.67. The results indicate high correlations, which imply that the propeller twist property is an important factor of in vitro and in vivo nucleosome organization. Meanwhile, unlike in vitro situation, in vivo nucleosome occupancy data is less correlated with the structural profile, suggesting that in vivo nucleosome organization may also be influenced by the actions of some additional external factors like DNA binding proteins and chromatin remodelers [34], [35].

For the propeller twist structural profile, significant peaks are associated with nucleosome-enriched regions, while trough-like regions correspond to nucleosome-depleted sequences. In most cases, the peak centers coincide with the centers of nucleosomes. In different cells, the exact positions of nucleosomes may deviate more or less from the most preferred positions. This deviation of nucleosome positions varies in different genomic regions. For example, the proximal promoter region is largely composed of well-positioned nucleosomes. In yeast genome, more than 10 well-positioned nucleosomes are observed flanking transcription start sites. On the contrary, distal regulatory regions are more fluid, where the distribution of nucleosomes becomes much fuzzier. Up to now, the exact positioning of nucleosomes is still poorly defined, especially in the fluid regions. Based on the comparative analysis, we observed that the changing pattern of the propeller twist structural profile is quite predictive of the fuzziness of nucleosomes, as shown in Fig. 4. For the well-positioned nucleosomes, the propeller twist structural profile exhibits a narrow and sharp peak. In contrast, the peaks of the structural profile related to fuzzy nucleosomes are relatively

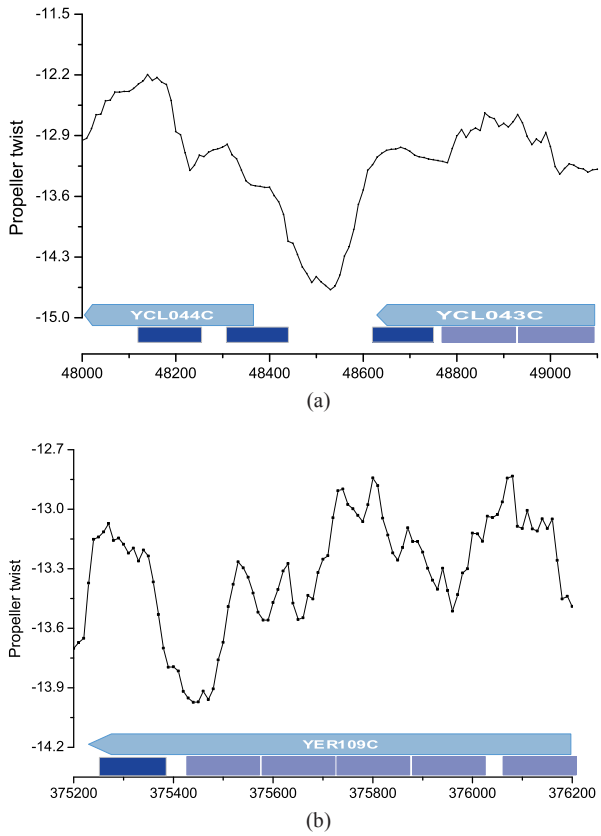


Fig. 4. The propeller twist structural profile of well-positioned and fuzzy nucleosomes. The dark blue rectangle represents well-positioned nucleosome, whereas fuzzy nucleosome is represented by light blue rectangle. (a) Local views of chromosome III regions. (b) Local views of chromosome V regions.

broad. In other words, there usually exist blurring peaks in the fluid regions. This finding is consistent with the previous study that, in fuzzy case, some nucleosomes deviate from the most preferred nucleosome positions [5].

3.2 Prediction of Nucleosome Positioning Using WaveNuc

The above analyses indicate that the propeller twist structural profile is informative of nucleosome positioning. Most nucleosome sequences have an obvious peak in the profile, while there is virtually no high signal in nucleosome-depleted regions. This implies that the propeller twist property is sufficiently distinctive to allow effective prediction of nucleosome positions. We thus used the structural profile to determine the nucleosome positions on the whole genome. In detail, we developed a computational method WaveNuc to predict nucleosome positions by detecting the significant peaks of the structural profile based on the continuous wavelet transformation. A direct test of how accurately WaveNuc positions nucleosomes on genomic sequences can be provided by a collection of nucleosome maps, where in vitro nucleosome positions are experimentally determined.

To assess the performance of our proposed method WaveNuc, we carried out comparisons with four widely used existing methods, including DNABEND [36], NuScore [37], NuPoP [38] and DLaNe [39]. DNABEND and NuScore

are based on energy-related features of genomic sequences. DLaNe applies an ensemble technique to combine several structural features for predicting nucleosome positions. Differently, NuPoP uses sequence compositional features and adopts a duration HMM to infer nucleosome positioning. For these four methods, we assumed that authors provided the optimal parameters for their own methods. Accordingly, we ran these programs as their suggested settings to obtain the predictions of nucleosome positions. All nucleosome positions predicted by different methods were equally validated by the genome-wide reference nucleosome positioning map [32].

These automated methods allow for an objective comparison of their predictions with experimentally determined nucleosome positions, by measuring the distance between predicted and experimental centers. If a predicted nucleosome center is within a certain distance L of a true position, we considered it as a correct prediction, where L is a parameter of distance cutoff. To obtain a fair comparison, we evaluated these predicted positions by different distance cutoffs. We used six cutoff values, ranging from 10 to 60 bp with an increment of 10 bp. As previous studies evaluated their prediction accuracy in terms of sensitivity (Se) and positive predictive value (PPV) [38], [39], here we also adopted these two criteria. Specifically, sensitivity represents the fraction of experimentally verified nucleosome positions that are correctly predicted, and PPV is the fraction of correctly predicted positions out of all predictions. Fig. 5 presents the performance comparisons under different distance cutoffs. The results demonstrate that DNABEND and DLaNe achieve good sensitivity. However, there is no superiority in their PPV. We observed that WaveNuc achieves more balanced sensitivity and PPV, which leads to better performance than previous methods.

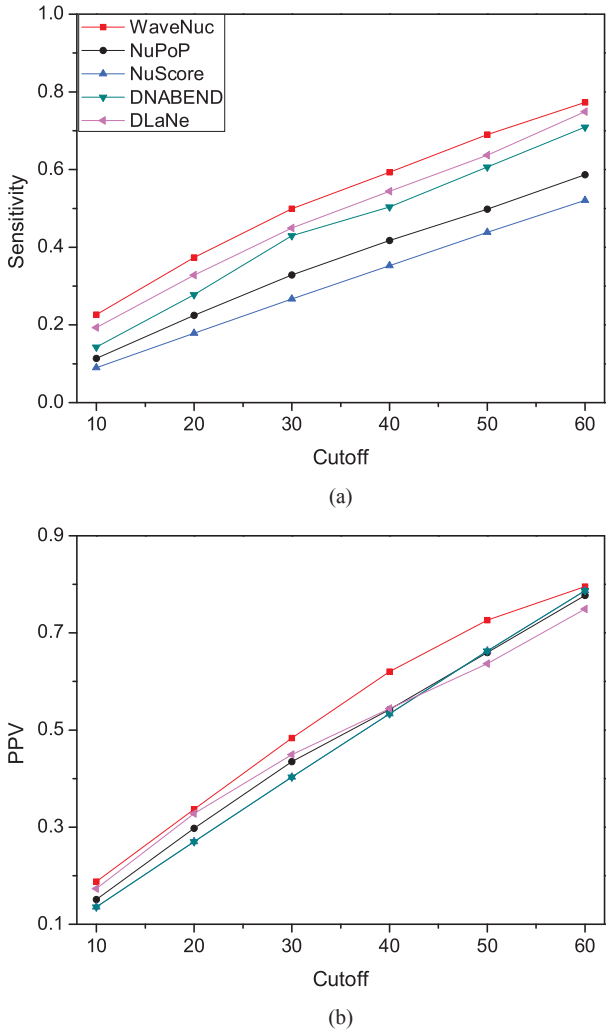


Fig. 5. Comparison of our method with four existing methods. Here we present the results of these methods at six different distance cutoff values from 10 to 60. (A) Sensitivity. (B) PPV.

3.3 Prediction of Well-Positioned and Fuzzy Nucleosomes

Experimental nucleosome maps of various model organisms have demonstrated that the nucleosomes downstream of the TSSs exhibit strong phasing. With distance from the TSSs, the phasing decays in a ripple-like manner and becomes quite fuzzy [40]. Previous study of Lee et al. determined that about 81 percent of the yeast genome is covered by nucleosomes: $40,096$ well-positioned and $30,777$ fuzzy nucleosomes [23]. Fuzzy nucleosomes represent a significant proportion of nucleosomes, having important implication in nucleosome dynamics. Most of previous studies can only predict the nucleosome occupancy level and the positioning of well-positioned nucleosomes. However, it is very beneficial to supply molecular biologists additional information about nucleosomes whether they are well-positioned or fuzzy on genomic sequences and predict the positions of fuzzy nucleosomes.

By analyzing the propeller twist profile, we noticed that strongly positioned nucleosomes appear as sharp peaks of structural profile, in contrast to fuzzy positioning where the peak locations are blurred as a result of nucleosome delocalization. Due to the multi-resolution property, we applied continuous wavelet transformation to search for all possible significant peaks in the structural profile. In general, the sharpness is considered as a measurement of the fuzziness of nucleosomes. If a peak is narrow and the surrounding regions are depleted of small peaks, this is an indicator of a well-positioned nucleosome. In contrast, wide peaks or peaks very close to each other probably implicate fuzzy nucleosomes. As mentioned previously (see Method section), this measurement can be improved by accounting the fuzziness of a nucleosome call (the sharpness of the peak). We used an overlapping argument accounting for longer range diffuse peaks. In this way, we can merge those overlapping peaks to a broader peak which is considered as a fuzzy region. In particular, fuzzy nucleosomes are placed at the broader peaks of the structural profile, whereas the peaks corresponding to well-positioned nucleosomes are much sharper. Then we evaluated our predicted nucleosomes against the reference nucleosome map. The results are shown in Fig. 6. WaveNuc is able to detect a total of 24,770 well-positioned nucleosome calls, about 61 percent of the centers that we predicted were within 40 bp of the centers predicted by HMM analysis of the experimental data. For fuzzy nucleosomes, a total of 15,062 nucleosome

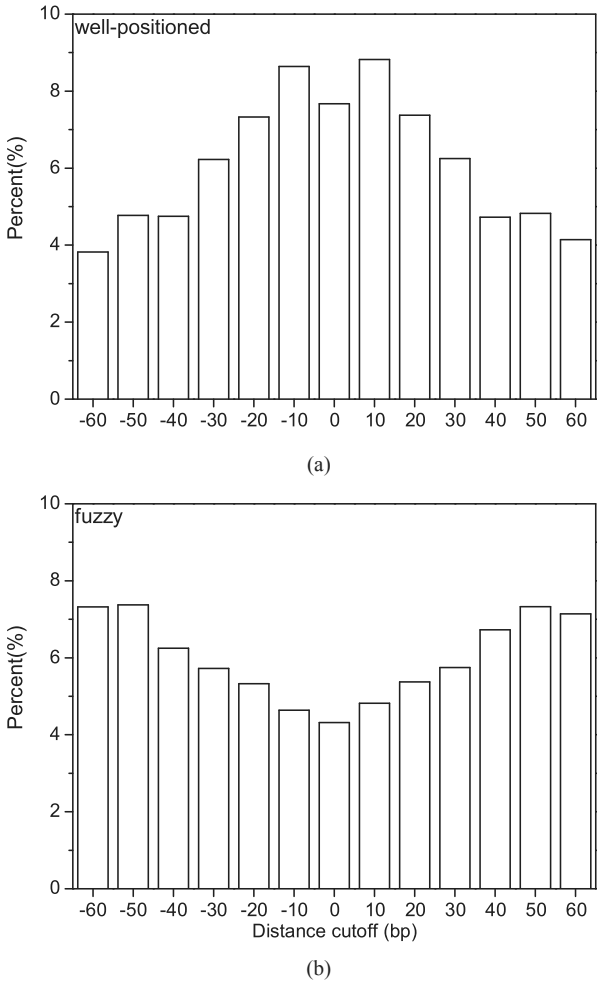


Fig. 6. Distribution of distances between predicted experimental nucleosomes and predicted nucleosomes based on the propeller twist profile. (a) Well-positioned nucleosomes. (b) Fuzzy nucleosomes.

calls, about 49 percent of the prediction were within 40 bp of the reference positions. With the strict evaluation criteria, WaveNuc predicts well-positioned nucleosomes with a higher accuracy. As the distance cutoff increases, the prediction accuracy of fuzzy nucleosomes increases faster than that of well-positioned nucleosomes. The result suggests that fuzzy nucleosome organization requires more complex mechanism and is harder to be predicted.

4 Discussion and Conclusion

We proposed a computational method, WaveNuc, using the propeller twist structural profile of genomic sequences to accurately predict nucleosome positions. Based on public nucleosome positioning data, the results demonstrated that WaveNuc is very effective for detecting well-positioned nucleosomes as well as fuzzy nucleosomes, not only in promoter regions, but also in distal regulatory regions with more fuzzy nucleosomes.

Our study provided new structure-based perspectives in three major aspects: (1) The propeller twist structural feature is closely correlated with nucleosome occupancy. Specifically, two kinds of nucleosomes exhibit different propeller twist characteristics, which contain nucleosome positioning information. (2) Unlike most of previous approaches that predict nucleosomes from only well-positioned condition. WaveNuc is a novel approach explicitly designed for the analysis of both well-positioned and fuzzy nucleosomes. (3) WaveNuc is independent of any experimental nucleosome occupancy data. This allows us to predict nucleosome positioning directly from the propeller twist structural profile of genomic sequences. The comparison of different methods reveals considerable effectiveness in predicting nucleosome positioning.

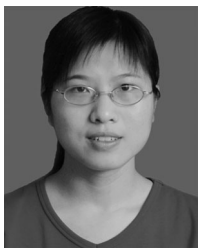
Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61300100, 61303096, 61272380, 61173118 and 31100954), Shanghai Natural Science Foundation (13ZR1451000, 13ZR1454600), Open Research Foundation of Shanghai Key Laboratory of Intelligent Information Processing (IIPL-2012-001), the Fundamental Research Funds for the Central Universities (13D111206), and the foundation of Donghua University for Young Faculty.

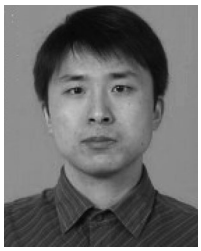
References

[1] B.D. Strahl and C.D. Allis, "The Language of Covalent Histone Modifications," *Nature*, vol. 403, no. 6765, pp. 41-45, 2000.
[2] O. Bell, V.K. Tiwari, N.H. Thomä, and D. Schübeler, "Determinants and Dynamics of Genome Accessibility," *Nature Rev. Genetics*, vol. 12, no. 8, pp. 554-564, 2011.
[3] C. Jiang and B.F. Pugh, "Nucleosome Positioning and Gene Regulation: Advances through Genomics," *Nature Rev. Genetics*, vol. 10, no. 3, pp. 161-172, 2009.
[4] A. Polishko, N. Ponts, K.G. Le Roch, and S. Lonardi, "Normal: Accurate Nucleosome Positioning Using a Modified Gaussian Mixture Model," *Bioinformatics*, vol. 28, no. 12, pp. i242-i249, 2012.
[5] K. Chen, Y. Xi, X. Pan, Z. Li, K. Kaestner, J. Tyler, S. Dent, X. He, and W. Li, "Danpos: Dynamic Analysis of Nucleosome Position and Occupancy By Sequencing," *Genome Research*, vol. 23, no. 2, pp. 341-351, 2013.
[6] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P.Z. Wang, and J. Widom, "A Genomic Code for Nucleosome Positioning," *Nature*, vol. 442, no. 7104, pp. 772-778, 2006.
[7] T.N. Mavrich et al., "Nucleosome Organization in the Drosophila Genome," *Nature*, vol. 453, no. 7193, pp. 358-362, 2008.
[8] D.E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, "Dynamic Regulation of Nucleosome Positioning in the Human Genome," *Cell*, vol. 132, no. 5, pp. 887-898, 2008.
[9] Y. Zhang, Z. Moqtaderi, B.P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X.S. Liu, and K. Struhl, "Intrinsic Histone-DNA Interactions Are Not the Major Determinant of Nucleosome Positions in Vivo," *Nature Structural & Molecular Biology*, vol. 16, no. 8, pp. 847-852, 2009.
[10] K. Struhl and E. Segal, "Determinants of Nucleosome Positioning," *Nature Structural & Molecular Biology*, vol. 20, no. 3, pp. 267-273, 2013.
[11] R. Sadeh and C.D. Allis, "Genome-Wide-Modeling of Nucleosome Positions," *Cell*, vol. 147, no. 2, pp. 263-266, 2011.
[12] T. Bartke, M. Vermeulen, B. Xhemalce, S.C. Robson, M. Mann, and T. Kouzarides, "Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation," *Cell*, vol. 143, no. 3, pp. 470-484, 2010.
[13] Y. Field, N. Kaplan, Y. Fondufe-Mittendorf, I.K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal, "Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals," *PLoS Computational Biology*, vol. 4, no. 11, p. e1000216, 2008.
[14] A. Stein, T.E. Takasuka, and C.K. Collings, "Are Nucleosome Positions in Vivo Primarily Determined by Histone–DNA Sequence Preferences?" *Nucleic Acids Research*, vol. 38, no. 3, pp. 709-719, 2010.

[15] H.R. Widlund, H. Cao, S. Simonsson, E. Magnusson, T. Simonsson, P.E. Nielsen, J.D. Kahn, D.M. Crothers, and M. Kubista, "Identification and Characterization of Genomic Nucleosome-Positioning Sequences," *J. Molecular Biology*, vol. 267, no. 4, pp. 807-817, 1997.
[16] H.E. Peckham, R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng, "Nucleosome Positioning Signals in Genomic DNA," *Genome Research*, vol. 17, no. 8, pp. 1170-1177, 2007.
[17] S.C. Satchwell, H.R. Drew, and A.A. Travers, "Sequence Periodicities in Chicken Nucleosome Core DNA," *J. Molecular Biology*, vol. 191, no. 4, pp. 659-675, 1986.
[18] K. Chen, Q. Meng, L. Ma, Q. Liu, P. Tang, C. Chiu, S. Hu, and J. Yu, "A Novel DNA Sequence Periodicity Decodes Nucleosome Positioning," *Nucleic Acids Research*, vol. 36, no. 19, pp. 6228-6236, 2008.
[19] G.-C. Yuan and J.S. Liu, "Genomic Sequence is Highly Predictive of Local Nucleosome Depletion," *PLoS Computational Biology*, vol. 4, no. 1, p. e13, 2008.
[20] D. Tillo and T.R. Hughes, "G+C Content Dominates Intrinsic Nucleosome Occupancy," *BMC Bioinformatics*, vol. 10, no. 1, article 442, 2009.
[21] T. van der Heijden, J.J. van Vugt, C. Logie, and J. van Noort, "Sequence-Based Prediction of Single Nucleosome Positioning and Genome-Wide Nucleosome Occupancy," *Proc. Nat'l Academy of Sciences USA*, vol. 109, no. 38, pp. E2514-E2522, 2012.
[22] V. Miele, C. Vaillant, Y. d'Aubenton Carafa, C. Thermes, and T. Grange, "DNA Physical Properties Determine Nucleosome Occupancy from Yeast to Fly," *Nucleic Acids Research*, vol. 36, no. 11, pp. 3746-3756, 2008.
[23] W. Lee, D. Tillo, N. Bray, R.H. Morse, R.W. Davis, T.R. Hughes, and C. Nislow, "A High-Resolution Atlas of Nucleosome Occupancy in Yeast," *Nature Genetics*, vol. 39, no. 10, pp. 1235-1244, 2007.
[24] M. El Hassan and C. Calladine, "Propeller-Twisting of Base-Pairs and the Conformational Mobility of Dinucleotide Steps in DNA," *J. Molecular Biology*, vol. 259, no. 1, pp. 95-103, 1996.
[25] K. Florquin, Y. Saeys, S. Degroeve, P. Rouze, and Y. Van de Peer, "Large-Scale Structural Analysis of the Core Promoter in Mammalian and Plant Genomes," *Nucleic Acids Research*, vol. 33, no. 13, pp. 4255-4264, 2005.
[26] T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, and Y. Van de Peer, "Generic Eukaryotic Core Promoter Prediction Using Structural Features of DNA," *Genome Research*, vol. 18, no. 2, pp. 310-323, 2008.
[27] Y. Gan, J. Guan, and S. Zhou, "A Pattern-Based Nearest Neighbor Search Approach for Promoter Prediction Using DNA Structural Profiles," *Bioinformatics*, vol. 25, no. 16, pp. 2006-2012, 2009.
[28] C.E. Heil and D.F. Walnut, "Continuous and Discrete Wavelet Transforms," *SIAM Rev.*, vol. 31, no. 4, pp. 628-666, 1989.
[29] I. Daubechies et al., *Ten Lectures on Wavelets*, vol. 61, SIAM, 1992.
[30] Z. Zhang and B.F. Pugh, "High-Resolution Genome-Wide Mapping of the Primary Structure of Chromatin," *Cell*, vol. 144, no. 2, pp. 175-186, 2011.
[31] N. Kaplan et al., "The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome," *Nature*, vol. 458, no. 7236, pp. 362-366, 2008.
[32] K. Brogaard, L. Xi, J.-P. Wang, and J. Widom, "A Map of Nucleosome Positions in Yeast at Base-Pair Resolution," *Nature*, vol. 486, no. 7404, pp. 496-501, 2012.
[33] J.M. Cherry et al., "Saccharomyces Genome Database: The Genomics Resource of Budding Yeast," *Nucleic Acids Research*, vol. 40, no. D1, pp. D700-D705, 2012.
[34] E. Segal and J. Widom, "What Controls Nucleosome Positions?" *Trends in Genetics*, vol. 25, no. 8, pp. 335-343, 2009.
[35] K. Luger, M.L. Dechassa, and D.J. Tremethick, "New Insights Into Nucleosome and Chromatin Structure: An Ordered State or a Disordered Affair?" *Nature Rev. Molecular Cell Biology*, vol. 13, no. 7, pp. 436-447, 2012.
[36] A.V. Morozov, K. Fortney, D.A. Gaykalova, V.M. Studitsky, J. Widom, and E.D. Siggia, "Using DNA Mechanics to Predict in Vitro Nucleosome Positions and Formation Energies," *Nucleic Acids Research*, vol. 37, no. 14, pp. 4707-4722, 2009.
[37] M.Y. Tolstorukov, V. Choudhary, W.K. Olson, V.B. Zhurkin, and P.J. Park, "nuScore: A Web-Interface for Nucleosome Positioning Predictions," *Bioinformatics*, vol. 24, no. 12, pp. 1456-1458, 2008.
[38] L. Xi, Y. Fondufe-Mittendorf, L. Xia, J. Flatow, J. Widom, and J.-P. Wang, "Predicting Nucleosome Positioning Using a Duration Hidden Markov Model," *BMC Bioinformatics*, vol. 11, no. 1, article 346, 2010.
[39] Y. Gan, J. Guan, S. Zhou, and W. Zhang, "Structural Features Based Genome-Wide Characterization and Prediction of Nucleosome Organization," *BMC Bioinformatics*, vol. 13, no. 1, article 49, 2012.
[40] G. Arya, A. Maitra, and S.A. Grigoryev, "A Structural Perspective on the Where, How, Why, and What of Nucleosome Positioning," *J. Biomolecular Structure and Dynamics*, vol. 27, no. 6, pp. 803-820, 2010.



Yanglan Gan received the PhD degree in computer science from Tongji University, China, in 2012. Her research interests include Bioinformatics, data mining, and web services. She is an assistant professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. She has published more than 15 papers on international journals and conferences, including *Bioinformatics*, *BMC Bioinformatics*, *Knowledge-based Systems*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *Soft Computing*, and FSKD. She served as a program committee member on APBC-12 and ADMA-13. She worked as a reviewer for a variety of international journals and conferences, such as *BMC Bioinformatics*, *Journal of Molecular Biology Knowledge-based Systems*, KDD-10, APBC-12, and ADMA-13.



Guobing Zou received his PhD degree in computer science from Tongji University, China, in 2012. He is an assistant professor in the School of Computer Engineering and Science at Shanghai University. His research interests include web service composition, automated planning, Bioinformatics, and cloud computing. He has published more than 20 papers on international journals and conferences, including *IEEE Transactions on Services Computing*, AAAI-12, Soft Computing, *Journal of Tongji University*, and CCV-10. He served as a program committee member on CIT-12, UUMA-12, UUMA-13, and organization member on DISD-13. He worked as a reviewer for *Journal of Artificial Intelligence Research*, *IEEE Transactions on Services Computing*, KDD-09, AAAI-10, IJCAI-11, and ICDM-10.

Jihong Guan received the PhD degree from Wuhan University, China, in 2002. She is a professor and dean in the Department of Computer Science and Technology at Tongji University, Shanghai, China. She has currently published more than 200 papers on international journals and conferences, including *Bioinformatics* NAR, *Physical Review E*, ICDE, SIGIR, DASFAA, and PAKDD. She served as a reviewer for international journals and conferences, such as *BMC Bioinformatics, IEEE Transactions on Knowledge and Data Engineering,* ICDM, and FSKD. Her research interests include database, Bioinformatics, distributed computing, data mining, and complex network.

Guangwei Xu received the MS degree from Nanjing University, Nanjing, China, in 2000, and the PhD degree from Tongji University, Shanghai, China, in 2003. He is an associate professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. His research interests include data mining, the verification of data integrity, QoS and security of the wireless Ad Hoc networks, and sensor networks. He has published more than 20 papers on international journals and conferences, including *IEEJ Transactions on Electrical and Electronic Engineering*, *Journal of Tongji University*, and ECWAC 2011. He reviewed articles for several academic journals and international conferences.