A comparative evaluation on prediction methods of nucleosome positioning

Hui Liu, Ruichang Zhang, Wei Xiong, Jihong Guan, Ziheng Zhuang and Shuigeng Zhou

Submitted: 7th June 2013; Received (in revised form): 4th August 2013

Abstract
Nucleosome positioning plays an essential role in cellular processes by modulating accessibility of DNA to proteins. Many computational models have been developed to predict genome-wide nucleosome positions from DNA sequences. Comparative analysis of predicted and experimental nucleosome positioning maps facilitates understanding the regulatory mechanisms of transcription and DNA replication. Therefore, a comprehensive evaluation of existing computational methods is important and useful for biologists to choose appropriate ones in their research. In this article, we carried out a performance comparison among eight widely used computational methods on four species including yeast, fruitfly, mouse and human. In particular, we compared these methods on different regions of each species such as gene sequences, promoters and 5’ UTR exons. The experimental results show that the performances of the two latest versions of the thermodynamic model are relatively steadier than the other four methods. Moreover, these methods are workable on four species, but their performances decrease gradually from yeast to human, indicating that the fundamental mechanism of nucleosome positioning is conserved through the evolution process, but more and more factors participate in the determination of nucleosome positions, which leads to sophisticated regulation mechanisms.

Keywords: nucleosome positioning; performance comparison; nucleosomal sequence; G:C content; prediction accuracy

INTRODUCTION
The nucleosome is the fundamental building unit of eukaryotic chromatin, consisting of ~147-bp double-helical DNA wrapped around a histone octamer in the left superhelix [1, 2]. In all, 75–90% of genomic DNA is packaged into nucleosomes, with adjacent nucleosomes being separated by stretches of DNA, which are referred as linker sequences. Further, nucleosomes are folded through a series of successively higher-order structures to eventually form a chromosome. In this manner, a large eukaryotic genome is packed into the tiny nucleus [3].

Besides the functionality of compressing chromosome, nucleosomes have been shown to play important roles in diverse biological processes, including transcription control, co-transcriptional splicing, DNA replication and repair [4–7]. In vivo eukaryotic cells, the chromosome structure is determined by multiple factors, including histone octamers, sequence-specific DNA-binding proteins and chromatin remodelers. Access to DNA wrapped on a histone octamer is occluded for polymerase, regulatory proteins, DNA repair and recombination complexes. Thus, the detailed locations of nucleosomes along
the DNA sequence may have important inhibitory or facilitatory roles in regulating gene expression [3, 5, 7]. Moreover, nucleosomes have been shown to carry epigenetically inherited marks in the form of covalent modifications of their histone tails [8]. Nucleosome positioning together with high-order chromatin structure creates a layer of regulatory control for gene regulation [9–11].

In recent years, the genome-wide nucleosome positioning maps of some model organisms were derived by virtue of the rapid progress of high-throughput array and sequencing techniques such as ChIP–ChIP and ChIP-seq [12–16], which greatly facilitate us to understand the mechanism of nucleosome positioning both in vivo and in vitro [9]. Some features of nucleosome positioning on yeast in vivo have been highlighted by many previous works as below: (1) Nucleosomes are significantly sparse in the regions of promoter, termination and transcription factor binding sites, but dense in the coding sequence regions, especially on constitute exons [17, 18]. (2) There exists a nucleosome free region (NFR) near the transcriptional start site (TSS) and transcriptional termination site, respectively [19]. Two nucleosomes flanking the 5 NFR (often called +1 and −1 nucleosomes) and the upstream nucleosome located closest to 3 NFR [20, 21] are strongly positioned. (3) The degree of uncertainty of the nucleosomes positioned in the coding sequence regions becomes larger with increase of their distances to either 5 NFR or 3 NFR [22]. However, in vitro reconstitutions of genomic DNA into nucleosomes by salt dialysis do not maintain all of these features. The relative importance of DNA sequence and protein factors in determining nucleosome positioning has been subject to considerable debate. A point of view is that intrinsic sequence preference has a dominant role in determining nucleosome positions in vivo, based on the observation that the genome-wide nucleosome mapping in vivo is strong correlated to that in vitro in yeast [19, 21]. Although another point of view, often referred as statistical positioning, has also been presented that intrinsic sequence preference is not the major determinants of nucleosome positioning in vivo [23–25]. In this view, the DNA-bound factors and DNA sequences unfavoring nucleosome formation act as barriers, and around these barriers the nucleosomes are organized into arrays by steric hindrance [22]. Nucleosome positioning in vivo is systematically controlled by many factors including sequence preference, ATP-dependent chromatin remodeling complex [26], protein factors [27, 28] and Pol II elongation [29]. Hughes et al. proposed a three-step model that integrates all the factors mentioned earlier in the text to establish the genome-wide pattern for identifying the determinants of nucleosome positioning [30]. Struhl and Segal reviewed the progress of nucleosome positioning and depicted the contribution of each type of factor to the genome-wide nucleosome positioning [31].

Although nucleosome positioning in vivo is a biological process determined by various DNA-binding protein factors, it is nontrivial to evaluate the contribution of DNA sequence on the nucleosome positioning. First, genome-wide comparative analysis of predicted and experimental nucleosome positioning maps facilitates to understand the mechanisms of nucleosome positioning. Particularly, comparative study on the genetic regions that are under stringent control, such as the promoter, enhancer and vicinity of TSS, reflects the competitive binding to DNA of histones, transcriptional factors (TFs) and RNA polymerases, and thus it helps downstream functional analysis, such as identifying the regulatory functions of TFs [32] and improving the prediction accuracy of transcription factor binding sites [33]. Second, the predicted nucleosome positioning based on underlying sequence may reveal the effect of chromatin secondary structure, as well as the roles of histone modification [8], DNA methylation [34] and chromosome remodeler. Finally, affinity of histone to sequence also affects the variation rate of nucleotide [35]. As a result, some methods based on sequence were developed to predict nucleosome positioning. These methods can be roughly divided into bioinformatic methods and biophysical methods. Most bioinformatic methods extract features from nucleosome sequences obtained from high-throughput experiments, including 10–11-bp dinucleotide periodicity [19, 20, 21, 36], nucleosomal sequence features [37–39], position-specific scoring matrix [40, 41] and structural constraint [42], to calculate the probability of forming a nucleosome for each ~147-bp sequence. In addition, information of linker DNA is incorporated to improve the performance [20, 43]. The biophysical methods are mainly based on sequence elasticity, histone–DNA binding energy [44–47]. To predict nucleosome locations along a long sequence, both types of methods account for the steric exclusion and thermodynamic equilibrium (i.e. at least some base spacing between two adjacent nucleosome is
required) and then adopt the dynamic programming algorithm to evaluate each legal configuration [19–21, 44].

This article focuses on a comprehensive evaluation on major nucleosome positioning prediction methods. We first describe briefly eight widely used methods, including Segal et al.’s method [19], Field et al.’s method [20], Kaplan et al.’s method [21], N-score [41], NuPoP [43], NucEnerGen [46], FineStr [40] and Heijden et al.’s method [48]. We then carry out a performance comparison on the four species, including the whole genome and promoter region of yeast, the largest chromosome, promoters and 5’UTR exons of fruitfly, mouse and human. Our comparative experiments show that most of these methods are workable on the four species, but their performances decrease gradually from yeast to human. This observation indicates that the fundamental mechanism of nucleosome positioning is conserved through the evolution process, but more and more factors participate in the determination of nucleosome positions, which leads to sophisticated regulation mechanisms.

OVERVIEW OF NUCLEOSOME POSITIONING PREDICTION METHODS

Segal et al.’s method
Segal et al. proposed a probabilistic model [19] to predict genome-wide nucleosome positioning in yeast. This model first estimates the dinucleotide distribution of each position i from millions of alignments of nucleosomal sequences obtained from high-throughout sequencing of yeast, and then calculates the probability \( P_N(S) \) for each 147-bp sequence \( S \) in favor of histone-DNA interaction based on the dinucleotide probability distribution as below

\[
P_N(S) = P_i(S_1) \prod_{i=2}^{147} P_i(S_i|S_{i-1}),
\]

which captures the periodic signal of dinucleotides along the nucleosome sequences. These probabilities are converted to an apparent free energy score by taking their log ratio to a predefined background model \( P_B \), i.e. \( \text{Score}(S) = \log \frac{P_N(S)}{P_B(S)} \). The background model \( P_B \) can be set to uniform distribution \( \{\text{A}=P(\text{C})=P(\text{G})=P(\text{T})=0.25\} \) or the probability distribution computed from the nucleotide distribution in the genome in which the computations are performed. Next, a thermodynamic model that accounts for the steric hindrance effect between nucleosomes on a sequence is used to predict genome-wide nucleosome positions. Specifically, for each sequence \( S \) and legal configuration \( \epsilon \) with \( k \) nucleosomes positioned at \( \epsilon[1], \ldots, \epsilon[k] \), a statistical weight \( W_\epsilon[S] \) is defined as

\[
W_\epsilon[S] = \left( \prod_{i=1}^k \tau \exp(\beta \cdot \text{Score}(S_{\epsilon[i]}, S_{\epsilon[i]+146})) \right),
\]

where \( \tau \) represents an apparent nucleosome concentration, and \( \beta \) is an apparent inverse temperature parameter. In accord with the Boltzmann distribution and under the assumption of thermodynamic equilibrium, it follows that the probability of every configuration is then given by

\[
P(W_\epsilon[S]) = W_\epsilon[S] / \sum_{\epsilon \in C} W_\epsilon[S].
\]

The partition function is the sum of all legal configurations of nucleosomes on a sequence \( S \), where a legal configuration specifies a set of 147-bp nucleosomes and a start position for each of these nucleosomes on \( S \), such that no two nucleosomes overlap and the minimum distance between the end of one nucleosome and the beginning of the next is 10 bp. Although the space of all legal configurations \( C \) for a sequence is huge, a dynamic programming method similar to Hidden Markov Model is used to efficiently compute the probability.

Field et al.’s method
Field et al. improved the Segal et al.’s model by incorporating the information of linker sequences to calculate the probability of forming a nucleosome for each 147-bp sequence [20]. Besides \( P_N(S) \) introduced in the Segal et al.’s model, another component, \( P_L(S) \), was introduced to represent the position-independent distribution over 5-mers in linker regions compared with nucleosomal DNA, which thus captures sequences that are generally favored or disfavored by nucleosomes regardless of their detailed positions within the nucleosome. In particular, each 147-bp sequence \( S \) is assigned a score that is defined as the log ratio of the two components

\[
\text{Score}(S) = \log \frac{P_N(S)}{P_L(S)} = \log \frac{P_N(S_1) \prod_{i=2}^{147} P_N(S_i|S_{i-1})}{P_L(S_1) \prod_{i=2}^{147} P(L_{\max(1,i-4)}, \ldots, S_{i-1})},
\]

where \( P_L \) is the product of a position-independent Markov model of order 4, i.e. the probability.
distribution over each one of the 1024 possible 5-mers. \( P_i \) is introduced to capture the contributions from both sequences that are disfavored by nucleosomes and sequences that are favored by nucleosomes, because it models the distribution over all sequences of length 5, with the disfavored sequences having a relatively high probability and the favored sequences having a relatively low probability. Finally, the scores were incorporated into thermodynamic model mentioned in the Segal et al.’s method to predict the genome-wide nucleosome positioning.

Kaplan et al’s method

As in vivo nucleosome maps reflect the combined action of all influential factors, it is difficult to estimate the relative contribution of sequence preference on nucleosome positioning. Kaplan et al. constructed an in vitro nucleosome map [21] by measuring the genome-wide occupancy of nucleosomes assembled on purified yeast genomic DNA, where the nucleosome locations are affected solely by intrinsic histone-DNA interactions and the formation of higher-order chromatin structures. The resulting map is similar to in vivo nucleosome maps of yeast generated in three different growth conditions. The in vitro map was subsequently used to learn the probability model that is identical to the Field et al’s model [20], and the learned model was used to predict the nucleosome occupancies that are significantly correlated with in vivo nucleosome occupancies in Caenorhabditis elegans. Therefore, the intrinsic DNA sequence preferences of nucleosomes were supposed to play a central role in determining the organization of nucleosomes in vivo.

N-score

Yuan et al. developed a statistical multiresolution approach, called N-score [41], to distinguish nucleosomal sequences from non-nucleosome sequences. This approach differs from the traditional methods in two main aspects. First, instead of characterizing each aligned nucleotide position independently using a position-specific scoring matrix, it applies a wavelet analysis to extracting spatially periodic signals. Second, instead of extracting information from only nucleosomal DNA sequences, it uses a logistic regression model to identify signals that help differentiate nucleosome and linker sequences. Specifically, the central 131 bp were extracted from each nucleosome and linker sequence for training. For each of the 16 dinucleotides, each extracted sequence was represented by 130-dimensional numerical signals that correspond to the position-specific dinucleotide frequencies. Taking average over the signals of three consecutive positions results in a 128-dimensional vector; each sequence is thus represented by 16 numerical vectors of 128 dimensions.

The probability of a sequence being nucleosomal is then modeled as the logit of a linear combination of the wavelet energies (covariates), which are defined as the total variations at each length scale and are supposed to characterize the periodic patterns embedded in a dinucleotide frequency signal, based on the rationale that if a particular frequency is significantly associated with nucleosome positioning, it is supposed to be detected by comparing the corresponding wavelet energies for nucleosome vs linker sequences. For prediction of genome-wide nucleosome positions, N-score also takes into account the effect of steric hindrance imposed by competing nucleosomes, and adopts the Hidden Markov Model to derive the nucleosome locations along a long stretch of DNA sequence.

NuPoP

Xi et al. proposed NuPoP [43] to improve the prediction performance by incorporating the linker DNA information. Also, a rescaling method was developed to take the base composition variation into account when using yeast models to make predictions for other species. Concretely, it first models the DNA sequence with a duration hidden Markov model of two alternative states: nucleosome (N) and linker (L), where the nucleosome state has a fixed length of 147 bp and the linker state has a variable length. At the end of each state, the chain must transit to the other state. A complete chromatin sequence must start with and end in a linker state. A fourth-order time-dependent Markov chain is trained for the N state, and a homogeneous fourth-order Markov chain for the L state to distinguish the \( k \)-mer usage preferences for \( k \) up to 5 between the nucleosome and linker states. For other species, the transition probability trained on yeast is scaled by a factor determined by the difference of base composition between the species of interest and yeast.

NucEnerGen

Locke et al. proposed a biophysical model, NucEnerGen [46], to predict free energies of nucleosome formation directly from high-throughput
maps of nucleosome positions. This model uses an exact relation between measured nucleosome occupancies and free energies, treating steric exclusion rigorously in the presence of histone-DNA interactions of arbitrary strength and sequence specificity. In particular, the free energies of nucleosome formation, referred as Percus energies by the authors, were obtained from exact decomposition from experimentally available nucleosome probabilities and occupancies.

In summary, NucEnerGen consists of three steps: (1) the nucleosome formation energy for each base pair is derived directly from the high-throughput nucleosome map, under the assumption that the observed nucleosome positions are affected solely by intrinsic histone-DNA interactions and steric exclusion, (2) a linear fitting model is used to establish the correlation between Percus energies and sequence features found in nucleosomal and linker DNA and (3) the nucleosome probabilities and occupancies are predicted from fitted energies using a standard recursive algorithm. As a result, the model demonstrates that the nucleotide component (A:T/G:C content) is the primary determinant of nucleosome sequence preferences in *Saccharomyces cerevisiae*, although periodic dinucleotide distributions and longer sequence motifs play only a minor role.

**Heijden et al.’s method**

Heijden *et al.* presented an empirical statistical mechanics model [48] based solely on the periodic occurrences of the dinucleotides TA, TT, AA and GC. This model first assigns a probability score to each dinucleotide that depends only on the phasing of the dinucleotide, converts the resulting nucleosome-positioning likelihood map to a sequence-specific energy landscape for nucleosome binding and predicts relative nucleosome affinities with high accuracy. Specifically, the first three dinucleotides TA, TT and AA are spaced with a periodicity p of \( \sim 10 \) bp, where the dinucleotide GC is 5 bp out of phase. Their periodic distributions within the nucleosome were modeled by using the sine function as below

\[
P(Y_s|Y_{s-1}) = P_0 + B \sin\left(2\pi \left(\frac{s}{p} + \frac{1}{2} \delta_{GC}\right)\right)
\]

where \( Y \) denotes the nucleotide A, C, G or T. The subscript \( s \) indicates the position of the nucleotide in the nucleosome relative to the dyad. \( P_0 \) is the average probability to find one of the four nucleotides at position \( s \) and equals 0.25. \( B \) is the amplitude with which the dinucleotides are distributed in nucleosomes. \( \delta_{GC} \) is the Kronecker delta function, being 1 for a GC dinucleotide and 0 otherwise. For the remaining dinucleotides, the conditional probability is normalized to \( \sum_{Y_s} P(Y_s|Y_{s-1}) = 1 \).

Therefore, the likelihood to find the dyad of a nucleosome at position \( i \) in a specific DNA sequence is proportional to the product of the conditional probabilities of all of the dinucleotides in the nucleosome, \( P(i,N) = \prod_{i=-N/2}^{i+N/2} P(Y_i|Y_{i-1}) \) with \( N \) equal to the DNA footprint of the histone octamer or tetramer. After converting the likelihood ratio to the free energy landscape, Percus’s equation [49] is used to calculate the thermodynamic equilibrium density of nucleosome positions, by introducing another parameter called *chemical potential* to account for the average sequence-independent nucleosome affinity. This model has been shown to relate well-documented high-resolution nucleosome positions obtained from *in vitro* reconstitution reactions on strong positioning sequences to genome-wide nucleosome occupancy maps. The method was shown to isolate sequence effects on genome-wide nucleosome occupancy from other factors that may influence nucleosome positioning.

**FineStr**

Based on the observations that a nucleosome DNA sequence has 12 close contacts of oriented inward minor groove with arginines of the histones [50], and at least 11 periods of DNA helix turns tightly wrap around the octamer, Gabdank *et al.* developed a method called FineStr [40] to calculate the single-base-resolution nucleosome positioning according to a nucleosome DNA bendability matrix of *C. elegans*. The DNA bendability matrix is derived from the positional preferences of various dinucleotides within the 10.4 base nucleosome DNA repeat. Subsequently, the suitability of any DNA sequence to bend around the histone octamer can be evaluated using the bendability matrix, simply by aligning the matrix with the sequence and summing the scores at all 116 matrix positions.

**PERFORMANCE EVALUATION**

**Data sets**

For a comprehensive and impartial performance comparison of these methods, we conducted experiments not only on different species but also on different sequence regions of each species. Four species, including yeast, fruitfly, mouse and human, were
chosen to take both small and big genomes into account. For each species, three different sequence regions, including genetic regions, promoters (−2000 bp upstream to TSS) and 5′UTR exons, were taken into account to evaluate these algorithms. We used experimental nucleosome positioning maps from ChIP–ChIP or ChIP-seq as gold standard to evaluate these methods. Furthermore, we chose the experimental data sets that have not been used to learn any algorithm included in our performance comparison experiments.

For yeast, we carried out experiments on only two regions, the whole genome and promoter region, due to few sequences included in the yeast 5′UTR exons. The ChIP–ChIP data published by Lee et al. [51] were used to extract nucleosomal and linker sequences, and the sequences compiled in SGD/sacCer2 release were downloaded from UCSC Table Browser [52]. For the other three species, we used the largest chromosome instead of the whole genome to evaluate these algorithms, as the largest chromosome of each species is enough to evaluate impartially these algorithms and

<table>
<thead>
<tr>
<th>Species</th>
<th>Region</th>
<th>The numbers of nucleosomal sequence</th>
<th>The numbers of linker sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>Genome</td>
<td>39 661</td>
<td>4824</td>
</tr>
<tr>
<td></td>
<td>promoters</td>
<td>27 373</td>
<td>44 63</td>
</tr>
<tr>
<td>Fruity</td>
<td>chr3R</td>
<td>46 053</td>
<td>30 457</td>
</tr>
<tr>
<td></td>
<td>promoters</td>
<td>48 250</td>
<td>28 761</td>
</tr>
<tr>
<td></td>
<td>5′UTR exons</td>
<td>4666</td>
<td>2703</td>
</tr>
<tr>
<td>Mouse</td>
<td>chr1</td>
<td>59 776</td>
<td>56 762</td>
</tr>
<tr>
<td></td>
<td>promoters</td>
<td>36 978</td>
<td>47 787</td>
</tr>
<tr>
<td></td>
<td>5′UTR exons</td>
<td>3401</td>
<td>3412</td>
</tr>
<tr>
<td>Human</td>
<td>chr1</td>
<td>97 208</td>
<td>60 238</td>
</tr>
<tr>
<td></td>
<td>promoters</td>
<td>56 337</td>
<td>44 120</td>
</tr>
<tr>
<td></td>
<td>5′UTR exons</td>
<td>11 754</td>
<td>48 484</td>
</tr>
</tbody>
</table>

Performance evaluation flowchart
The flowchart of our experimental process is shown in Figure 1. We first extract the coordinates of the midpoints of each nucleosomal sequence and linker sequence from experimental nucleosome maps and then extend 75 bp from the midpoint on both ends to get 150-bp coordinate pairs. We then fetch the nucleosomal sequences and linker sequences from UCSC Table Browser. As the numbers of generated nucleosomal and linker sequences of each species are
large, we adopt the sampling with replacement method to evaluate each method. In particular, we sample respectively 100 sequences from the generated nucleosomal and linker sequences and take them as positive and negative samples to run these methods. Based on the prediction results, we draw the receiver operating characteristic curve (ROC) and compute the area under curve (AUC). We repeat the sampling process 100 times and average the performance measures over the 100 tests as the final result.

As the binding affinity of underlying sequences to histone octamers varies with nucleotide composition, each method outputs a score for each base pair that indicates the potential to participate in forming a nucleosome, or a score that represents the probability of the whole input sequence being a nucleosomal sequence. Based on the score, we compute the specificity and sensitivity measures for each method by setting a score threshold, i.e. a sequence is classified as a nucleosomal sequence if its score is greater than the predefined threshold or otherwise linker sequence. The ROC curve is subsequently obtained for each method by adjusting the threshold. Each method was run with the default parameter setting in all experiments. For NuPoP, the fourth-order time-dependent Markov chain model is used. For NucEnerGen, the two position-independent model is used in our performance evaluation.

**Performance comparison**

Figure 2 shows the ROC curves of eight methods on yeast: the left subgraph shows the result on the genome and the right shows the result on promoters. It is shown that Field et al.’s method performs best on both types of sequence regions, its AUC reaches 0.77372 and 0.78662, respectively. NuPoP also has relatively high performance on both types of regions (AUC = 0.74409 and 0.74446). It is interesting to note that N-score obtains remarkable performance on genetic region and its AUC (0.76903) is only slightly smaller than that of Field et al.’s method, but its performance deteriorates seriously on promoter (AUC = 0.7017). Segal et al.’s method has the worst performance and approximates to random guess (AUC = 0.49556 and 0.51755). NucEnerGen and Kaplan et al.’s method perform slightly worse than NuPoP; FineStr and Heijden et al.’s method are poorer than NucEnerGen, though they are much better than Segal et al.’s method.

For fruitfly, performance comparison was also conducted on 5'UTR exons in addition to the gene sequences of chr3R and promoters, and the ROC curves are shown in Supplementary Figure S1 included in the Supplementary File. Each method performs steady and the performance ranking keeps unchanged among the three fruitfly sequence regions. However, it was found that the performance of these methods drops slightly.

![Figure 2](http://bib.oxfordjournals.org)
comparing with that on yeast. The performance of N-score deteriorates mostly and is only slightly better than random guess (its AUC is equal to 0.58863, 0.61992 and 0.53792 on the three regions, respectively). Field et al.’s method and Kaplan et al.’s method have relatively high precision on the three regions and their performances are close. NucEnerGen performs well on fruitfly, and on 5’UTR its performance approaches that of Field et al.’s method. Segal et al.’s method still performs worst on the three fruitfly regions and its precision is even poorer than random guess (AUC = 0.3164, 0.3160 and 0.38404). FineStr and Heijden et al.’s method are just slightly better than random guess.

On mouse genetic regions and 5’UTR exons, both the performance and performance ranking of the eight methods are similar to that on fruitfly, as shown in Figure 3. However, most methods perform significantly worse on mouse promoters than on the other two sequence regions. It can also be found in Figure 3 that on mouse promoter, N-score achieves the best performance (AUC = 0.65683), the next is FineStr (AUC = 0.64717), whereas Field et al.’s method, NucEnerGen and Kaplan et al.’s method deteriorate greatly and their performances are only slightly better than that of Segal et al.’s method, whose accuracy increases considerably in comparison with that on other species and regions.

From the experimental results on human, Field et al.’s method and Kaplan et al.’s method still have similar accuracy on the three sequence regions, but their performances show a significant degradation in comparison to that on yeast and fruitfly, as shown in Supplementary Figure S2. On human promoter, NuPoP outperforms all the other methods (AUC = 0.6668), whereas on 5’UTR exons NucEnerGen achieves the highest accuracy; Segal et al.’s method is still the worst.

In summary, we can draw the following preliminary conclusions: (1) the performance of each method varies on different species and sequence regions. (2) Roughly, the performances of Field et al.’s method and Kaplan et al.’s method are steadier compared with the other methods; we thus recommend them for predicting nucleosome positioning. (3) In the view of evolution from yeast to human, it can be found that the performance of each method decreases gradually. Supplementary Figure S3 is more illustrative for this observation, which shows the AUCs of eight methods in histogram for each species. In addition, we note an unexpected phenomenon that on mouse promoters most methods perform poorly except Segal et al.’s method, as shown in Figure 4, whereas on the other three species all methods have better or at least comparable performance on promoters than on the other sequence regions.

**DISCUSSION**

Although these methods, except FineStr, included in our evaluation experiments were trained on the experimental nucleosome positioning map of yeast, they are applicable to predicting the nucleosome positioning of other species [6, 41, 46]. Their prediction accuracies gradually decrease from yeast to human. This observation indicates that on the one hand the fundamental mechanism of nucleosome
positioning is highly conserved through the evolution process and on the other hand more and more endonuclear factors involving in transcription regulation also influence the locations of nucleosomes nearby, which leads to more sophisticated mechanism of nucleosome positioning.

The methods using solely sequence information are able to predict in vivo nucleosome positioning, suggesting that sequence preference to histone octamer is critical for the overall organization of nucleosomes in living cells. However, we note that predictions of genome-wide nucleosome positions by all the tested methods are only modestly better than random guess, which is consistent with the conclusion in Peckham et al. [37]. Although it is expectable that new methods may uncover new sequence information and greatly improve the prediction accuracy, we tend to think that exact nucleosome positions can be further fine-tuned through interactions with chromatin modifying complexes and TFs, which are environment and stage dependent, but cannot be accurately predicted from sequence information alone.

We went further to explore the distributions of G:C (GG, GC, CG, CC) enrichment in nucleosomal and linker sequences. Consistent with previous studies [15,19,22], the nucleosomal sequences are remarkably enriched in G:C dinucleotide while linker sequences are G:C-depleted, and the nucleosomal sequences are characterized by the G:C-A:T dinucleotide periodicity, as shown in Supplementary Figure S3 (For detail of calculation of GC-content distribution see Supplementary File). We have noticed the sharp decrease of G:C dinucleotide content across the nucleosome boundary that would be highly predictive of nucleosome positions. The absolute difference of G:C content between nucleosomal and linker sequences maybe also influence the prediction accuracy. For instance, on mouse promoters the G:C enrichment difference between nucleosomal and linker sequences is much less than that on mouse chr1 genes and 5'UTR exons, as shown in

![Figure 4: Histograms of AUC of eight methods in the view of sequence regions.](image)
Figure 5. Correspondingly, the accuracies of almost all methods on mouse promoter are significantly lower than those on the other two regions of mouse, as well as on the promoters of other species, as shown in Figure 4 and Supplementary Figure S3.

In addition, we conducted experiments to investigate whether the information of linker sequence is helpful to improve prediction accuracy. We ran Field et al.’s method on 150-bp sequences generated from experimental data and 150-bp sequences plus 5000-bp flanking sequences of both sides for the four species; the length of extended sequence is as suggested by Segal’s Lab Web site. As shown in Figure 6, the performance on extended sequences does not exceed significantly that on nonextended sequences, implying that the information of linker sequences may not be so helpful for improvement on prediction accuracy as expected.

Moreover, we inspected whether the performances of these methods are affected by the rules of selecting nucleosomal and linker sequences from experimental data. By imposing the requirement of at least 120-bp length of consecutive nucleosome coverage signals, a higher signal threshold leads to well-positioned, but a fewer number of, nucleosomal sequences. On mouse genes of chr1, we generated six different data sets by setting the signal threshold to 15, 20, 30, 40, 45 and 50, respectively. We ran the eight methods on each data set and computed their AUC values, which are shown in Figure 7. We have assumed that all methods should have better accuracy on the data sets generated by higher signal thresholds. However, to our surprise, the performances of all methods except Segal et al.’s method and Heijden et al.’s method turn worse as the signal threshold increases. The unexpected results may indicate that the well-positioned nucleosomes such as +1 and -1 nucleosomes around TSS are mainly governed by RNA polymerases, chromatin remodelers and other factors involved in transcription, rather than sequence preference. On the contrary, the
data set generated from a lower signal threshold includes much more sequences of high preference to octamers, according to the statistical positioning models [22]. As a result, the methods based on sequence information perform worse on the sequences obtained from well-positioned nucleosomes.

Finally, it should be noted that the thermodynamic equilibrium of nucleosome positioning in living cell is systematically controlled by various factors [55], and the equilibrium state of nucleosome positioning in vivo transfers from one state to another in different cell types [56,57], external stimulus [58] or different development stages [59, 60]. The experimental nucleosome maps are the snapshots of the genome-wide nucleosome positioning in vivo, and the dyads of nucleosome sequences extracted from experimental maps may deviate some base pairs from those with most potential to bind the histone octamers, reducing the performance of the calculation methods to some extent.

**CONCLUSION**

In this article, we carried out extensive experiments to compare the performances of eight methods for nucleosome positioning prediction. To our knowledge,
this is the most comprehensive comparison of the methods for predicting nucleosome positioning. First, from the performance comparison on different species and sequence regions, it can be concluded that most methods trained on the experimental nucleosome positioning map of yeast are workable on other species, indicating that the fundamental mechanism of nucleosome positioning is highly conserved through the evolution process. Second, our experimental results also show that Field et al.’s method and Kaplan et al.’s method achieve steadier performance on different species and genetic regions in comparison to other methods. Third, we explored the influence of the G:C content difference between nucleosomal and linker sequences, and the information of flanking sequences of nucleosome on the prediction accuracy. Last, but not least, the effect of signal thresholds of extracting nucleosomal and linker sequences was also investigated, and the result implies that the sequences bearing well-positioned nucleosomes may not be high nucleosome preference.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points
- Many computational models have been developed to predict genome-wide nucleosome positions from DNA sequences. However, their performances have not been evaluated comprehensively.
- To help biologists choose appropriate methods for predicting nucleosome positioning, this article conducted a comprehensive comparison study on eight existing methods over four species and with different genetic regions.
- Our results show that Field et al’s method and Kaplan et al’s method are relatively steadier than the other methods, and the performance of each method decreases gradually from yeast to human, indicating that the fundamental mechanism of nucleosome positioning is conserved through the evolution process.
- The results also show that the difference of G:C content between nucleosomal and linker sequences may affect prediction accuracy, and the sequences bearing well-positioned nucleosomes may not be high nucleosome preference.

Acknowledgement
The authors would like to thank Idan Gabdank and John van Noort for providing the executables of FineStr and Heijden et al.’s method, respectively, and their helpful suggestions for our experiments.

FUNDING
The National Basic Research Program of China (grant 2010CB126604), the National Natural Science Foundation of China (grants 31100954, 61173118 and 61272380).

References


