doi: 10.4149/gpb_2020002

Nucleosome sliding mechanism based on the potential energy of sequence

Hu Meng[1,2], Hong Li[3], Zhenhua Yang[1,3] and Yangming Si[3]

[1] *School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou China*
[2] *Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, China*
[3] *The Inner Mongolia Key Laboratory of Functional Genome Bioinformatics, Inner Mongolia University of Science and Technology, Baotou, China*

Abstract. Nucleosome sliding and nucleosome digestion are two main ways for regulating gene transcription. We constructed three characteristic parameters (CP) based on the information of CG0, CG1 and CG2 motifs, and used these parameters to analyze the sliding trend of −1 and +1 nucleosomes around TSS of genes with NFR in yeast. The CP distribution was used to describe the features of nucleosome sequences, and the slope of fit line of CP distribution curve was used to represent the potential energy of nucleosome sequences. Results show that nucleosome sliding trend could be reflected by CG0 and CG2 CP distributions, and CG0 CP distribution has a good correlation with nucleosome sliding trend. In addition, the sliding trend of nucleosomes is different in various expression level genes. For high expression gene, sliding trend of −1 nucleosome is weaker and that of +1 nucleosome is stronger.

Key words: Sliding trend — Transcription regulation — Sequence characteristic — *k*-mer — CG dinucleotide

Abbreviations: CG0, subset of 8-mer motif that contains no CG dinucleotide; CG1, subset of 8-mer motif that contains only one CG dinucleotide; CG2, subset of 8-mer motif that contains two or more CG dinucleotide; CP, characteristic parameter of a given sequence; NFR, nucleosome-free region; RF, relative frequency of m-mer (m < 8) in a given 8-mer set; RMN, relative motif number; TSS, transcription start site.

Introduction

k-mer, as an important frequency-based algorithm, is widely used in computational genomics. Because of its fast and efficient advantages, many *k*-mer-based approaches have been developed and used in similarity analyses (Blaisdell 1986; Sims et al. 2009; Huang et al. 2011; Yu 2013; Wen et al. 2014). For some special *k*-mer motifs, they could be used as signature signals to predict CpG islands (Yang et al. 2012; Mohamed Hashim et al. 2015), promoters (Li et al. 2006; Lin et al. 2011, 2014), noncoding RNA (Feng et al. 2016), DNA binding sites (Badis et al. 2009; Liu et al. 2012; Liu et al. 2015), and nucleosome positioning (Segal et al. 2009; Guo et al. 2014). *k-mer* has also been involved in probe design (Fofanov et al. 2004), repeat sequence annotation (Kurtz et al. 2008), genome assembly (Compeau et al. 2011), epigenetic analysis (Quante et al. 2016), and drug design (Chou 2015). In addition, based on *k-mer*, information entropy can be further calculated, and which has been used in biological computational analysis (Meng et al. 2017). *k-mer* spectra of whole genome sequences are a visualization approach to reveal the genomic features. It has been found that frequency distributions of *k*-mers $8 \le k \le 10$ is basically in agreement with that of *k*-mer $5 \le k \le 15$ (Das et al. 2007), and 8-mer spectra are multimodal in human and mouse, but unimodal in bacteria (Stacey et al. 2003). G+C content and CpG suppression are thought to contribute to the multimodal (Chor et al. 2009). Our previous study has also found that multimodal of 8-mer spectrum is only closely related to

Correspondence to: Hong Li, Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, 010021, China
E-mail: ndlihong@imu.edu.cn

the amount of CG dinucleotide in 8-mer motif, in analyzed model organism genomes, the spectra of only CG2, CG1, and CG0 subsets form independent unimodal distributions. The functions of 8-mer sets containing different amounts of CG dinucleotide are different, CG1 motifs are related to the nucleosome-binding and CG2 motifs are related to the modular units of CpG islands (Zheng et al. 2017).

In eukaryotic cells, DNA is highly packaged into nucleosome arrays, and a nucleosome core particle comprises 146/147 base pairs of DNA wrapped around an octamer of histone proteins (Luger et al. 1997; Zhu et al. 2016). Nucleosome positioning is very important for transcription regulation (Jiang et al. 2009; Bai et al. 2010; Radman-Livaja et al. 2010). It is generally recognized that there are two well positioned nucleosomes around TSS (transcription start site), and a nucleosome-free region (NFR) on the upstream of the TSS (Struhl et al. 2013; Lieleg et al. 2015). However, not all genes have the typical nucleosome organization (Kornberg et al. 1988; Becker 2014). Kubik et al. (2015) have pointed out that, in yeast, about 40% promoters have narrower NFR, and the formation of wider NFR in these promoters depends on the digestion of the "fragile" −1 nucleosome. For the "stable" −1 nucleosome promoters, the transcription regulation is mainly depended on the sliding of nucleosome. It has been proved that chromatin-remodeling complex is related to the sliding of nucleosome (Gangaraju et al. 2007; Zhou et al. 2016; Sinha et al. 2017), and sequence also has an effect on nucleosome sliding (McKnight et al. 2016; Niina et al. 2017; Brandani et al. 2018; Guoqing Liu et al. 2018). Here, we constructed three characteristic parameters based on the information of 8-mer sets containing different amounts of CG dinucleotide, and used these parameters to analyze the relations between sequence potential energy and nucleosome sliding trend.

Materials and Methods

Data sources

Nucleosome positioning data of *Saccharomyces cerevisiae* (unique map) were accessed from Brogaard et al. (2012). The reference genome sequence and gene annotation information of *Saccharomyces cerevisiae* were obtained from UCSC (SAC2 version) (http://genome.ucsc.edu/). The experimental data of gene expression levels were obtained from Holstege et al. (1998).

k-mer of genomic sequence

k-mer could be described as follows: supposing there is a genomic sequence *S* with length *L*, 'N_1, N_2,…N_L', where $N_i \in$ {A,T,C,G}. A string of consecutive *k* nucleotides within genetic sequence *S* is called a *k*-mer. The *k*-mers appearing in a sequence can be enumerated by using a sliding window of length *k*, shifting one base each time from position 1 to $L - k + 1$, until the entire sequence has been scanned. Given any *k*, there will be 4^k different possible permutations.

Relative motif number of 8-mers

According to the definition of *k*-mer, for a given DNA sequence with length L, all of the 8-mer frequencies could be counted. The relative motif number (RMN) of the 8-mers with frequency *i* could be calculated by the following equation:

$$RMN = \frac{N_i}{4^8} \tag{1}$$

N_i is the number of the 8-mers with frequency *i*. The distribution of RMN also called as the 8-mer spectrum.

Relative frequency of m-mer in 8-mers

The relative frequency (*RF*) of *m*-mer ($m < 8$) in a given 8-mer set is defined as follows:

$$RF = \frac{4^m}{(8-m+1)} \frac{\sum_{j=1}^{L_k} N_{ji} H_j}{\sum_{j=1}^{L_k} H_j} \tag{2}$$

where L_k is the number of the *k*-th 8-mer set, H_j is the frequency of the *j*-th 8-mer in the *k*-th 8-mer set, N_{ji} represents the occurrence number of the *i*-th *m*-mer in the *j*-th 8-mer, and $RF = 1$ denotes that the *i*-th *m*-mer appears randomly.

Characteristic parameter

CP value is used to represent the characteristic parameter of a given sequence. CP is defined by the following equation:

$$CP = \frac{1}{N} \sum_{i=1}^{N} RF_i \tag{3}$$

where *i* is the *i*-th *m*-mer and *N* is the total number of *m*-mers in a given sequence. RF_i is the relative frequency after the *i*-th *m*-mer in the given sequence is assigned by *m*-mer *RF* in a given 8-mer subset. The CP value represents the average abundance of *m*-mers in the given sequence. In our analysis, m = 3.

Results

Two types of nucleosome distribution modes around TSS

Based on the single-base pair resolution map of nucleosome positions given by Brogaard et al. (2012) we analyzed the

nucleosome distribution around TSS of yeast genes. The nucleosome located at TSS or nearest to the TSS is defined as +1 nucleosome, and the first nucleosome on the upstream of +1 nucleosome is defined as −1 nucleosome (Teves et al. 2014), and the distances between nucleosome dyad and TSS were calculated. Results show that not all genes have an obvious nucleosome free region on the upstream of TSS (Fig. 1). About 2/3 genes have a traditional nucleosome distribution pattern, and a longer interval appears between −1 and +1 nucleosome. These genes are called NFR genes. Meanwhile, about 1/3 genes have a different nucleosome distribution pattern. There is no NFR on the upstream of TSS, and the distance between −1 and +1 nucleosome dyad is much shorter than the average. These genes are called non-NFR genes. In Kubik's research, they have given a reasonable explanation through analyzing the stability of −1 nucleosome, and thought that the transcription regulation in non-NFR genes depends on the digestion of −1 nucleosome, and in NFR genes depends on the sliding of −1 and +1 nucleosome (Kubik et al. 2015). Sequence preference is an important factor in nucleosome positioning, and the difference of nucleosome stability could also be reflected in the interaction between histone and DNA. In our study, we made further analysis about these two transcription regulation mechanisms by the information content of −1 and +1 nucleosome sequences.

We separately analyzed −1 and +1 nucleosome sequences of NFR genes and non-NFR genes by 3 characteristic parameters constructed by the information of 8-mer set of CG0, CG1 and CG2, and the CP distribution was used to describe the characteristic of nucleosome sequences. Results show that (Fig. 2), no matter which parameter was used, the CP distributions of +1 nucleosome sequences have no obvious differences between NFR genes and non-NFR genes, but CP distributions of −1 nucleosome sequences are different. For the CP distributions obtained by the information of CG1 set, only the CP distribution of −1 nucleosome sequence in non-NFR genes deviates obviously from that of all nucleosome sequences. Our previous research showed that the CG1 motif was the important positioning signal for nucleosome (Zheng et al. 2017), so this deviation could reflect the instability of −1 nucleosome in non-NFR genes to some extent. For the CP distributions obtained by the information of CG0 set, in NFR genes, the downstream of CP distribution of −1 nucleosome shows an increasing trend, so the CP distribution line of −1 nucleosome is tilted. In non-NFR genes, either the downstream or the upstream of CP distribution of −1 nucleosome shows an increasing trend, so the CP distribution line of −1 nucleosome is aclinic. For the CP distributions obtained by the information of CG2 set, the CP distribution of −1 nucleosome shows a decreasing trend in NFR genes, but not in non-NFR genes. In NFR genes, we thought that transcription regulation should depend on the slide of −1 and +1 nucleosome, and the slide of nucleosomes must be closely related to the sequence construction. In our

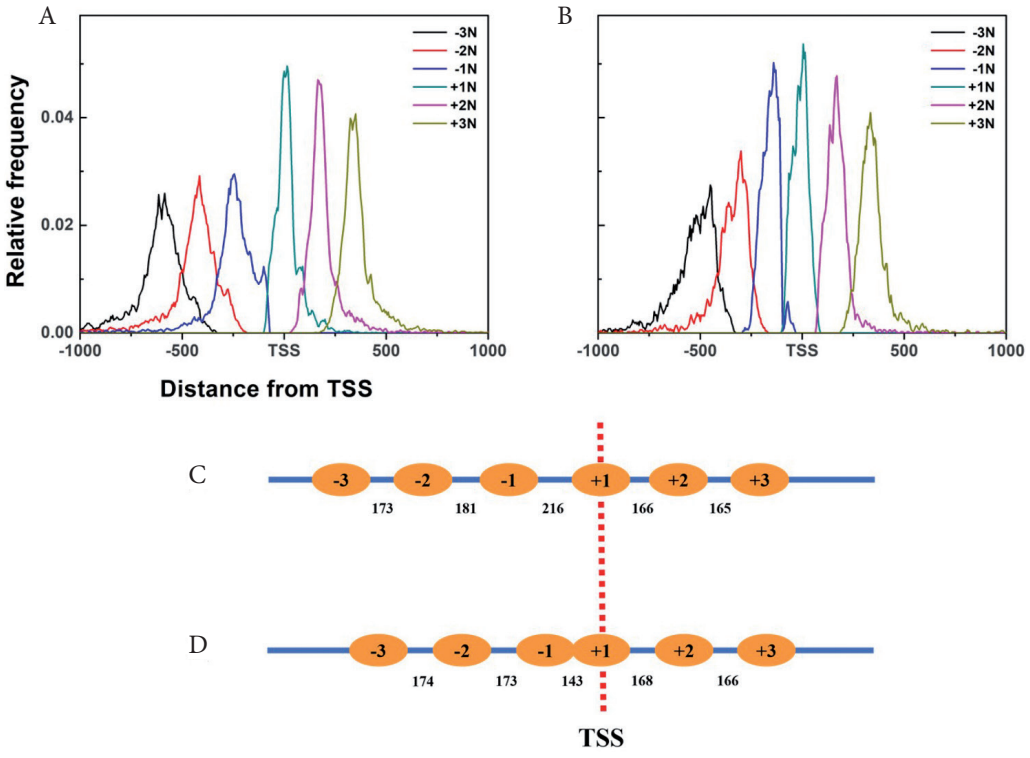

Figure 1. Two types of nucleosome distributions around TSS. **A.** Nucleosome distributions around TSS in NFR genes. **B.** Nucleosome distributions around TSS in non-NFR genes. **C.** Schematic diagram of the most probable position of nucleosomes in NFR genes. **D.** Schematic diagram of the most probable position of nucleosomes in non-NFR genes. NFR, nucleosome-free region; TSS, transcription start site.

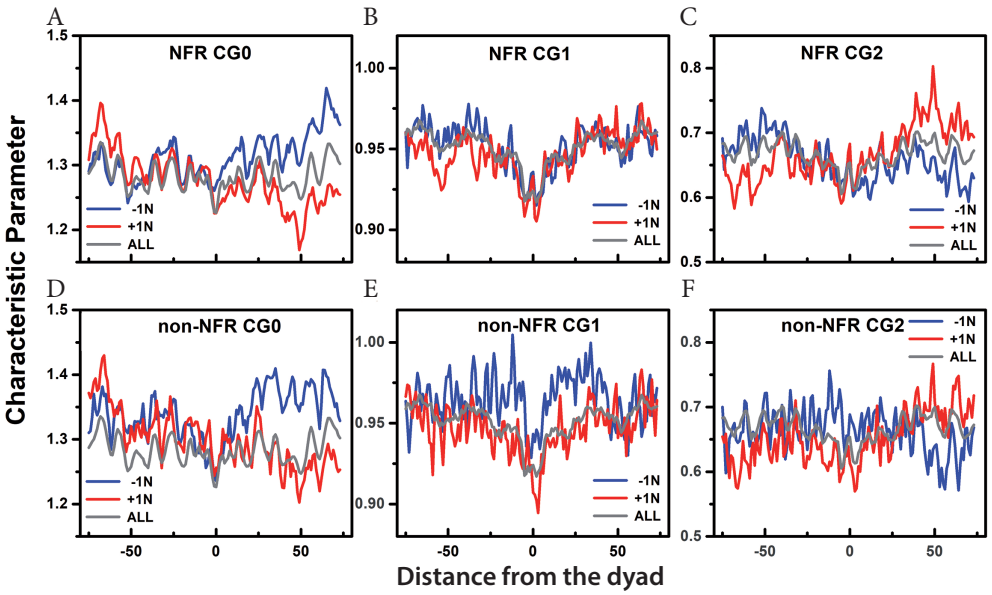

Figure 2. The CP distributions of nucleosome sequences in NFR genes and non-NFR genes. The blue line represents −1 nucleosome. The red line represents +1 nucleosome. The gray line represents all nucleosomes. **A.–C.** The CP distributions obtained by the information of CG0, CG1 and CG2 sets separately in NFR genes. **D.–F.** The CP distributions obtained by the information of CG0, CG1 and CG2 sets separately in non-NFR genes. CP, characteristic parameter of a given sequence; CG0, subset of 8-mer motif that contains no CG dinucleotide; CG1, subset of 8-mer motif that contains only one CG dinucleotide; CG2, subset of 8-mer motif that contains two or more CG dinucleotide; NFR, nucleosome-free region. (See online version for color figure.)

previous study, it has been shown that CG1 motifs are associated with nucleosome positioning. The distribution of CG1 motifs in nucleosome sequences has the balanced and symmetrical characteristic, and this characteristic could be used to predict the nucleosome positioning. However, the distribution characteristics of CG1 motifs have no difference between special nucleosomes (−1 and +1 nucleosome) and normal nucleosomes. Oppositely, the distribution characteristics of CG0 and CG2 motifs have difference between special nucleosomes and normal nucleosomes, so the potential energy of sequence reflected by the CP distribution obtained by the information of CG0 and CG2 sets could be an appropriate parameter to describe the sliding trend of nucleosome.

Potential energy of nucleosome sequences with different length of NFR

For discussing the correlation between the potential energy of sequence and the sliding trend of nucleosome, we analyzed −1 and +1 nucleosome sequences with different length of intervals in NFR genes by the CP distribution obtained by the information of CG0 and CG2 sets. NFR genes were divided into 3 groups: (1) when the distance between −1 and +1 nucleosome dyad is shorter than 217 bp (or the length of NFR is shorter than 70 bp), these genes were called short NFR genes (S-NFR); (2) when the distance between −1 and +1 nucleosome dyad is between 217 bp and 287 bp (or the length of NFR is between 70 bp and 140 bp), these genes were called middle NFR genes (M-NFR); (3) when the distance between −1 and +1 nucleosome dyad is longer than 287 bp (or the length of NFR is longer than 140 bp), these genes were called long NFR genes (L-NFR). The CP distributions of −1 and +1 nucleosomes in each group of NFR genes obtained by the information of CG0 and CG2 sets are shown in Fig. 3A–D. The main characteristics of −1 and +1 nucleosome sequences in 3 groups of genes have not changed. Though it shows some differences, the correlation between the potential energy of sequence and the sliding trend of nucleosome could not be described directly. In physics, potential energy curve is often used to describe the change of potential energy with relative position. In this study, we introduced the idea of potential energy curve and expected it to describe the relation between nucleosome sliding tendency and position of sequence. To visualize this relation, the linear fitting analysis was applied in these distribution curves, and the slope of the fit line was used to represent the difference of the potential energy of sequences (Fig. 3E–H). In CG0 linear fitting analysis, −1 and +1 nucleosomes both show a reasonable relation between potential energy of sequence and sliding trend of nucleosome. With the decrease of the interval between −1 and +1 nucleosomes, the absolute values of the slope of the fit line of −1 and +1 nucleosome increase. It is appropriate that the −1 and +1 nucleosomes with shorter interval need more sliding trend to facilitate transcription initiation. So CG0 CP distribution could well reflect the sliding trend of nucleosome. The slope of fit line obtained by CG2 could not show a reasonable relation between potential energy of sequence and sliding trend of nucleosome, because the

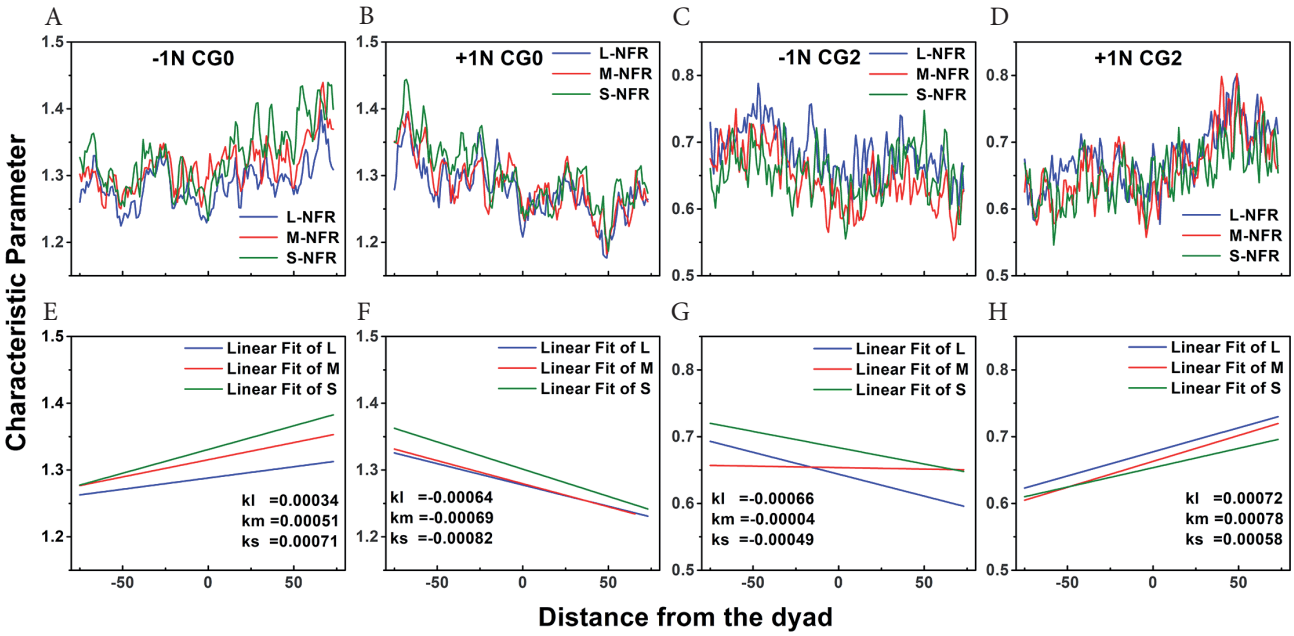

Figure 3. The CG0/CG2 CP distributions and fit lines of −1 and +1 nucleosome sequences in NFR genes with different length of intervals between −1 and +1 nucleosomes. The blue line represents nucleosomes in long NFR genes (L-NFR). The red line represents nucleosomes in middle NFR genes (M-NFR). The green line represents nucleosomes in short NFR genes (S-NFR). *k* value represents the slope of the fit line. **A.–D.** The CG0/CG2 CP distributions of −1 and +1 nucleosome sequences. **E.–H.** The fit lines of CP distribution curves. For abbreviations, see Fig. 2. (See online version for color figure.)

amount of CG2 motif is so small that it has no universality. However, the CG2 motif also affects nucleosome sliding in some case, and this affection is just the opposite to that of CG0 motif. Totally, the potential energy of sequences reflected by CG0 CP distribution is the most appropriate to describe the nucleosome sliding trend.

Potential energy of nucleosome sequences with different gene expression level

Further, the −1 and +1 nucleosome sequences in NFR genes with different expression levels were analyzed by CG0 CP distributions. NFR genes were divided into 3 groups according to gene expression level: (1) genes with expression values less than 1 mRNA/h were called low expression genes; (2) genes with expression values among 1–4 mRNA/h were called middle expression genes; (3) genes with expression values higher than 4 mRNA/h were called high expression genes. The CG0 CP distributions of −1 and +1 nucleosomes in each group of NFR genes are shown in Fig. 4A, B. and the fit lines of distribution curves are shown in Fig. 4C, D. For −1 nucleosome, the slope of fit line of middle expression genes is the highest, and that of high expression genes is the lowest. This phenomenon is reasonable, because the statistical distance between −1 and +1 nucleosome in high expression genes is obviously higher than those in other two groups of genes, and the distances in other two groups of genes are similar (Fig. 4). In high expression genes, the length of NFR is sufficient, so the −1 nucleosome has a weak sliding trend. The potential energy of −1 nucleosome sequences in middle expression level genes is different from that in low expression level genes, which indicates that the −1 nucleosome with stronger sliding trend would be more beneficial to gene transcription. For +1 nucleosome, the absolute value of the slope of fit line of high expression genes is obviously higher (about 50% higher) than those in other two groups of genes. For most of yeast genes, +1 nucleosome occupies the TSS. The sliding of +1 nucleosome is for exposing TSS, and not only for extending NFR. So the stronger sliding trend of +1 nucleosome in high expression genes is necessary.

Discussion

The position of nucleosomes in the genome is dynamic. The functions of nucleosomes in different positions in the genome are different. −1 and +1 nucleosomes, as two special nucleosomes, are closely related to transcription regulation, and the nucleosome sliding is an important way of regulation. The nucleosome sliding mechanism is complex, and it is regulated by many factors, such as chromatin-remodeling

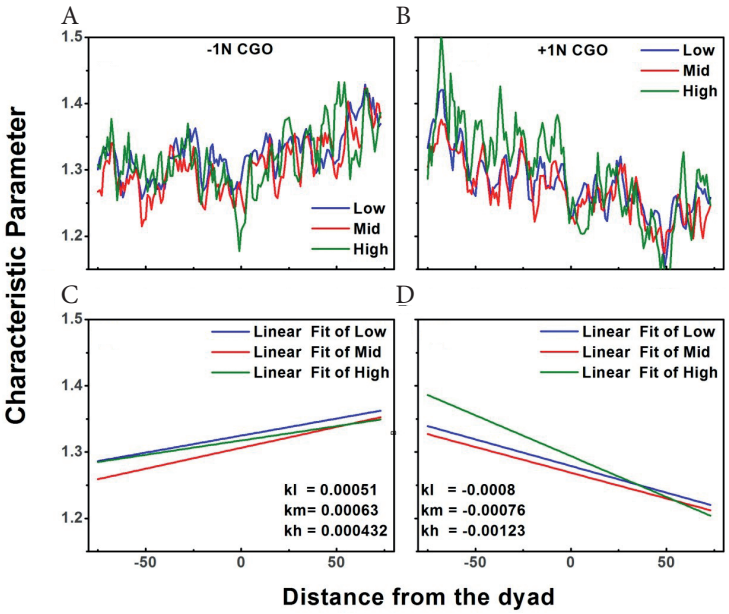

Figure 4. The CG0 CP distributions and fit lines of −1 and +1 nucleosome sequences in NFR genes with different expression levels. The blue line represents nucleosomes in low expression level genes. The red line represents nucleosomes in middle expression level genes. The green line represents nucleosomes in high expression level genes. *k* value represents the slope of the fit line. **A.** and **B.** The CG0 CP distributions of −1 and +1 nucleosome sequences. **C.** and **D.** The fit lines of CP distribution curves. For abbreviations, see Fig. 2. (See online version for color figure.)

complex and sequence matter. Sequence preference determines nucleosome positioning, and it also affects the stability of nucleosomes. However, not all sequence motifs are related to the nucleosome sliding. Our results show that the usage of CG1 motifs is always equalizing, and these motifs should determine the stability of nucleosomes but not sliding. CG0 and CG2 motifs both reflect the changes of potential energy of nucleosome sequence, but the correlation between potential energy and sliding trend obtained by CG2 motifs is poor. This may be attributed to the small amount of CG2 motifs. The CP constructed by the information of CG0 motifs is the most appropriate parameter to describe the sliding trend of nucleosomes. Because of our previous study has found that multimodal of 8-mer spectrum is only closely related to the amount of CG dinucleotide, we payed more attention on CG dinucleotide. Therefore, we can't yet conclude that CG determines nucleosome sliding. The potential energy of sequences reveals the nucleosome sliding mechanism. The sliding trend of −1 and +1 nucleosomes is mainly determined by the length of NFR. The longer the NFR is, the weaker the trend will be. For middle and low expression genes, the higher expression level genes have a stronger sliding trend of −1 nucleosome. For high expression level genes, it only needs the strong sliding trend of +1 nucleosome to expose TSS.

Acknowledgments. This work was supported by grants from the National Natural Science Foundation of China (No. 31860304), Inner Mongolia university scientific research project (NJZY19126), Inner Mongolia natural science foundation project (2019BS03024).

Conflict of interest. The authors declare no competing financial interests.

References

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. (2009): Diversity and complexity in DNA recognition by transcription factors. Science **324,** 1720-1723 https://doi.org/10.1126/science.1162327

Bai L, Morozov AV (2010): Gene regulation by nucleosome positioning. Trends Genet. **26,** 476-483 https://doi.org/10.1016/j.tig.2010.08.003

Becker PB (2014): Nucleosome sliding: facts and fiction. EMBO J. **21,** 4749-4753 https://doi.org/10.1093/emboj/cdf486

Blaisdell BE (1986): A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. U.S.A. **83,** 5155-5159 https://doi.org/10.1073/pnas.83.14.5155

Brandani GB, Niina T, Tan C, Takada S (2018): DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations. Nucleic Acids Res. **46,** 2788-2801 https://doi.org/10.1093/nar/gky158

Brogaard K, Xi L, Wang JP, Widom J (2012): A map of nucleosome positions in yeast at base-pair resolution. Nature **486,** 496-501 https://doi.org/10.1038/nature11142

Chor B, Horn D, Goldman N, Levy Y, Massingham T (2009): Genomic DNA k-mer spectra: models and modalities. Genome Biol. **10,** R108 https://doi.org/10.1186/gb-2009-10-10-r108

Chou KC (2015): Impacts of bioinformatics to medicinal chemistry. Med Chem. **11,** 218-234

https://doi.org/10.2174/1573406411666141229162834
Compeau PE, Pevzner PA, Tesler G (2011): How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. **29,** 987-991 https://doi.org/10.1038/nbt.2023
Das MK, Dai HK (2007): A survey of DNA motif finding algorithms. BMC Bioinformatics **8,** S21 https://doi.org/10.1186/1471-2105-8-S7-S21
Feng P, Zhang J, Tang H, Chen W, Lin H (2016): Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions. Interdiscip. Sci. **9,** 1-5 https://doi.org/10.1007/s12539-016-0193-4
Fofanov Y, Luo Y, Katili C, Wang J, Belosludtsev Y, Powdrill T, Belapurkar C, Fofanov V, Li TB, Chumakov S, Pettitt BM (2004): How independent are the appearances of n-mers in different genomes? Bioinformatics **20,** 2421-2428 https://doi.org/10.1093/bioinformatics/bth266
Gangaraju VK, Bartholomew B (2007): Mechanisms of ATP dependent chromatin remodeling. Mutat. Res. **618,** 3-17 https://doi.org/10.1016/j.mrfmmm.2006.08.015
Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014): iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics **30,** 1522-1529 https://doi.org/10.1093/bioinformatics/btu083
Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998): Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95,** 717-728 https://doi.org/10.1016/S0092-8674(00)81641-4
Huang G, Zhou H, Li Y, et al. (2011): Alignment-free comparison of genome sequences by a new numerical characterization. J. Theor. Biol. **281,** 107-112 https://doi.org/10.1016/j.jtbi.2011.04.003
Jiang C, Pugh BF (2009): Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. **10,** 161-172 https://doi.org/10.1038/nrg2522
Kornberg RD, Stryer L (1988): Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucleic Acids Res. **16,** 6677 https://doi.org/10.1093/nar/16.14.6677
Kubik S, Bruzzone MJ, Jacquet P, Falcone JL, Rougemont J, Shore D (2015): Nucleosome stability distinguishes two different promoter types at all protein-coding genes in yeast. Mol. Cell **60,** 422-434 https://doi.org/10.1016/j.molcel.2015.10.002
Kurtz S, Narechania A, Stein JC, Ware D (2008): A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics **9,** 517 https://doi.org/10.1186/1471-2164-9-517
Li QZ, Lin H (2006): The recognition and prediction of sigma70 promoters in Escherichia coli K-12. J. Theor. Biol. **242,** 135-141 https://doi.org/10.1016/j.jtbi.2006.02.007
Lieleg C, Krietenstein N, Walker M, Korber P (2015): Nucleosome positioning in yeasts: methods, maps, and mechanisms. Chromosoma **124,** 131-151 https://doi.org/10.1007/s00412-014-0501-x
Lin H, Li QZ (2011): Eukaryotic and prokaryotic promoter prediction using hybrid approach. Theory Biosci. **130,** 91-100 https://doi.org/10.1007/s12064-010-0114-8
Lin H, Deng EZ, Ding H, Chen W, Chou KC (2014): iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. **42,** 12961-12972 https://doi.org/10.1093/nar/gku1019
Liu B, Liu F, Fang L, Wang X, Chou KC (2015): repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics **31,** 1307-1309 https://doi.org/10.1093/bioinformatics/btu820
Liu G, Liu J, Cui X, Cai L (2012): Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. J. Theor. Biol. **293,** 49-54 https://doi.org/10.1016/j.jtbi.2011.10.004
Liu G, Xing Y, Zhao H, Cai L, Wang J (2018): The implication of DNA bending energy for nucleosome positioning and sliding. Sci. Rep. **8,** 8853 https://doi.org/10.1038/s41598-018-27247-x
Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997): Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature **389,** 251-260 https://doi.org/10.1038/38444
McKnight JN, Tsukiyama T, Bowman GD (2016): Sequence-targeted nucleosome sliding in vivo by a hybrid Chd1 chromatin remodeler. Genome Res. **26,** 693-704 https://doi.org/10.1101/gr.199919.115
Meng H, Li H, Zheng Y, Yang Z, Jia Y, Bo S (2018): Evolutionary analysis of nucleosome positioning sequences based on new symmetric relative entropy. Genomics **110,** 154-161 https://doi.org/10.1016/j.ygeno.2017.09.007
Mohamed Hashim EK, Abdullah R (2015): Rare k-mer DNA: Identification of sequence motifs and prediction of CpG island and promoter. J. Theor. Biol. **387,** 88-100 https://doi.org/10.1016/j.jtbi.2015.09.014
Niina T, Brandani GB, Tan C, Takada S (2017): Sequence-dependent nucleosome sliding in rotation-coupled and uncoupled modes revealed by molecular simulations. PLoS Comput. Biol. **13,** e1005880 https://doi.org/10.1371/journal.pcbi.1005880
Quante T, Bird A (2016): Do short, frequent DNA sequence motifs mould the epigenome? Nat. Rev. Mol. Cell Biol. **17,** 257-262 https://doi.org/10.1038/nrm.2015.31
Radman-Livaja M, Rando OJ (2010): Nucleosome positioning: how is it established, and why does it matter? Dev. Biol. **339,** 258-266 https://doi.org/10.1016/j.ydbio.2009.06.012
Segal E, Widom J (2009): Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr. Opin. Struct. Biol. **19,** 65-71 https://doi.org/10.1016/j.sbi.2009.01.004
Sims GE, Jun SR, Wu GA, Kim SH (2009): Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. U.S.A. **106,** 2677-2682 https://doi.org/10.1073/pnas.0813249106

Sinha KK, Gross JD, Narlikar GJ (2017): Distortion of histone octamer core promotes nucleosome mobilization by a chromatin remodeler. Science **355,** eaaa3761
https://doi.org/10.1126/science.aaa3761

Stacey KJ, Young GR, Clark F, Sester DP, Roberts TL, Naik S, Sweet MJ, Hume DA (2003): The molecular basis for the lack of immunostimulatory activity of vertebrate DNA. J. Immunol. **170,** 3614-3620
https://doi.org/10.4049/jimmunol.170.7.3614

Struhl K, Segal E (2013): Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. **20,** 267-273
https://doi.org/10.1038/nsmb.2506

Teves SS, Weber CM, Henikoff S (2014): Transcribing through the nucleosome. Trends Biochem. Sci. **39,** 577-586
https://doi.org/10.1016/j.tibs.2014.10.004

Wen J, Chan RH, Yau SC, He RL, Yau SS (2014): K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene **546,** 25-34
https://doi.org/10.1016/j.gene.2014.05.043

Yang Y, Nephew K, Kim S (2012): A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters. BMC Bioinformatics **13,** S15
https://doi.org/10.1186/1471-2105-13-S3-S15

Yu HJ (2013): Segmented k-mer and its application on similarity analysis of mitochondrial genome sequences. Gene **518,** 419-424
https://doi.org/10.1016/j.gene.2012.12.079

Zheng Y, Li H, Wang Y, Meng H1, Zhang Q4, Zhao X (2017): Evolutionary mechanism and biological functions of 8-mers containing CG dinucleotide in yeast. Chromosome Res. **25,** 173-189
https://doi.org/10.1007/s10577-017-9554-z

Zhou CY, Narlikar GJ (2016): Analysis of nucleosome sliding by ATP-dependent chromatin remodeling enzymes. Methods Enzymol. **573,** 119-135
https://doi.org/10.1016/bs.mie.2016.01.015

Zhu P, Li G (2016): Structural insights of nucleosome and the 30-nm chromatin fiber. Curr. Opin. Struct. Biol. **36,** 106-115
https://doi.org/10.1016/j.sbi.2016.01.013

Received: September 11, 2019
Final version accepted: January 21, 2020