Genomics 110 (2018) 154–161

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Evolutionary analysis of nucleosome positioning sequences based on New Symmetric Relative Entropy

Hu Meng, Hong Li*, Yan Zheng, Zhenhua Yang, Yun Jia, Suling Bo

Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

ARTICLE INFO

Keywords:
Nucleosome sequence
Concerted evolution
Transcription regulation
Sequence preference
Constitutive property

ABSTRACT

New Symmetric Relative Entropy (*NSRE*) was applied innovatively to analyze the nucleosome sequences in *S. cerevisiae*, *S. pombe* and *Drosophila*. *NSRE* distributions could well reflect the characteristic differences of nucleosome sequences among three organisms, and the differences indicate a concerted evolution in the sequence usage of nucleosome. Further analysis about the nucleosomes around TSS shows that the constitutive property of $+1/-1$ nucleosomes in *S. cerevisiae* is different from that in *S. pombe* and *Drosophila*, which indicates that *S. cerevisiae* has a different transcription regulation mechanism based on nucleosome. However, in either case, the nucleosome dyad region is conserved and always has a higher *NSRE*. Base composition analysis shows that this conservative property in nucleosome dyad region is mainly determined by base A and T, and the dependence degrees on base A and T are consistent in three organisms.

1. Introduction

In eukaryotic cells, DNA is highly packaged into nucleosome arrays. The nucleosome core particle comprises 146/147 base pairs of DNA wrapped in 1.7 superhelical turns around an octamer of histone proteins [1,2]. Nucleosome has played a crucial role in gene transcription regulation [3,4], and lots of researches focus on the nucleosomes around TSS [5–9]. There is a conserved nucleosome organization around TSS with a nucleosome free region upstream of the TSS and a TSS-aligned regular array of evenly spaced nucleosomes downstream over the gene body [10,11]. Knowing the precise locations of nucleosomes in a genome is key to understanding how genes are regulated [12]. With the development of nucleosome positioning study, nucleosome occupancy data have been published in many organisms [13–18], and Mavrich et al. have given the more detailed data including nucleosome positioning centers. Lately, Brogaard et al. have published a single-base pair resolution map of nucleosome positions in yeast [19], and Moyle-Heyrman et al. have used the same method to give nucleosome position annotation in fusion yeast [20]. It is generally recognized that nucleosome positioning is determined by the combination of DNA sequence, nucleosome remodelers, and transcription factors [11], and DNA sequence preferences of nucleosomes have made a significant contribution to the nucleosome organization [11,21–23]. Many researchers could predict nucleosome positioning along genomes by their prediction models based on sequence information [24–32]. Some webservers, such as iNuc-PseKNC [33] and iNuc-PhysChem [34] have also been established for nucleosome prediction, which makes the prediction work more convenient. However, it is unclear that whether the sequence preferences of nucleosomes are the same in different organisms, and whether there is concerted evolution in nucleosome sequences.

Lots of methods have been applied in sequence analysis. Sequence alignment is a basic and important method in bioinformatics research. BLAST [35] and Smith-Waterman [36] are the most widely used algorithms for two sequence alignment, and CLUSTAL W [37] is the most widely used algorithm for multi-sequence alignment. However, sequence alignment algorithms become powerless to large biological sequence datasets. For these datasets, alignment-free sequence comparison is more efficient. Many algorithms of alignment-free sequence comparison are based on the probability distribution of *k*-mer [38–41]. Entropy used in the alignment-free sequence comparison is also based on the probability distribution of *k*-mer. Kullback and Leible proposed a Relative Entropy ($H(p\|q) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i} = -\sum_{i=1}^{m} p_i \ln \frac{q_i}{p_i}$) as early as 1951 to measure the similarity between two discrete probability distributions [42]. Relative Entropy has no symmetry, so it could not be directly used to describe the difference between two probability distributions. Fu revised Relative Entropy, and proposed a Symmetric Relative Entropy ($SRE(p\|q) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i} + \sum_{i=1}^{m} q_i \ln \frac{q_i}{p_i}$) [43]. *SRE* has symmetry, but it is sensitive to extreme values. For instance, if $p_i = 0$ or $q_i = 0$, there will be $SRE(p \| q) = \infty$. Shen improved *SRE*, and proposed a

Abbreviations: TSS, Transcription Start Site; *SRE*, Symmetric Relative Entropy; *NSRE*, New Symmetric Relative Entropy; Ac-*NSRE*, Accumulated New Symmetric Relative Entropy

* Corresponding author at: No.235, West Daxue Road, Hohhot, Inner Mongolia, China.
E-mail address: ndlihong@imu.edu.cn (H. Li).

http://dx.doi.org/10.1016/j.ygeno.2017.09.007
Received 4 July 2017; Received in revised form 6 September 2017; Accepted 12 September 2017
Available online 14 September 2017
0888-7543/

New Symmetric Relative Entropy in her doctoral thesis ($NSRE(p||q) = \sum_i p_i \log \frac{2p_i}{q_i+p_i} + \sum_i q_i \log \frac{2q_i}{q_i+p_i}$) [44], and *NSRE* worked well in sequence similarity analysis. *NSRE* was applied in preliminary analysis of nucleosome sequences in our recent study [45]. In this paper, we used *NSRE* for analyzing nucleosome sequences systematically, and except to get sequence preference characteristics of nucleosomes in different organisms.

2. Data and methods

2.1. Data sources

Nucleosome positioning data of *Saccharomyces cerevisiae* (unique map) were gotten from Brogaard [19]. The reference genome sequence and gene annotation information of *Saccharomyces cerevisiae* were obtained from UCSC (SAC2 version) (http://genome.ucsc.edu/). Nucleosome positioning data and gene annotation information of *Drosophila* were gotten from Mavrich [15]. The *Drosophila* reference genome was obtained from Flybase (ftp://ftp.flybase.net/releases/FB2007_01/dmel_r5.2/). Nucleosome positioning data of *Schizosaccharomyces pombe* (unique map) were gotten from Moyle-Heyrman [20]. The reference genome sequence and gene annotation information of *Schizosaccharomyces pombe* were obtained from Ensemble Genomes (ftp://ftp.ensemblgenomes.org/pub/fungi/release-15/).

2.2. k-mer of genomic sequence

k-mer could be described as follows: supposing there is a genomic sequence *S* with length *L*, '$N_1, N_2, \ldots N_L$', 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.

2.3. Calculation of New Symmetric Relative Entropy

NSRE could measure the difference between two probability distributions. *NSRE* is defined as follows:

$$NSRE(p||q) = \sum_i p_i \log \frac{2p_i}{q_i + p_i} + \sum_i q_i \log \frac{2q_i}{q_i + p_i} \quad (1)$$

NSRE has the following properties:

Nonnegativity: $\sum_i p_i \log \frac{2p_i}{q_i + p_i} + \sum_i q_i \log \frac{2q_i}{q_i + p_i} \geq 0$;

Minimality: $\sum_i p_i \log \frac{2p_i}{q_i + p_i} + \sum_i q_i \log \frac{2q_i}{q_i + p_i} = 0$, while and only while $p_i = q_i$.

We used *NSRE* distribution to describe the constitutive property of nucleosome core sequences, and the value of *NSRE* was calculated by the probability of *k*-mer ($k = 1, 2, \cdots 8$). Supposing there are *N* nucleosome sequences, and the length of each sequence is *L* bp. The sequences are aligned by the dyad to form a $N \times L$ matrix *Bp* ($Bp_{ij} \in \{A, T, C, G\}$):

$$\mathbf{Bp} = \begin{matrix} \\ P_{(i,l)} \\ \downarrow \end{matrix} \begin{matrix} \\ q_i \\ \downarrow \end{matrix} \begin{Bmatrix} \mathbf{B_{11}} & \mathbf{B_{12}} & \cdots & \mathbf{B_{1L}} \\ \mathbf{B_{21}} & \mathbf{B_{12}} & \cdots & \mathbf{B_{2L}} \\ \cdots & \cdots & & \cdots \\ \mathbf{B_{N1}} & \mathbf{B_{N2}} & \cdots & \mathbf{B_{NL}} \end{Bmatrix}$$

Let $p_{(i,l)}$ be the probability of *k*-mer in *l*-th column of *Bp* ($l \in \{1, L - k + 1\}$). *i* represents the elements of *k*-mer (i.e. while $k = 1$, i = A, T, C, G; while $k = 2$, i = AA, AT, AC, …, GT, GC, GG; that is to say, the amount of *i* is 4^k). Let q_i be the probability of *k*-mer of all the columns of *Bp*, so q_i represents the *k*-mer probability of all nucleosome sequences. Bringing $p_{(i,l)}$ and q_i into formula (1), and *NSRE* of the *l*-th site (totally $L - k + 1$ sites) could be worked out. Then we could get a *NSRE* distribution comprised of consecutive $L - k + 1$ values. By this way, we could measure the differences of sequence preferences between each site and all sites of nucleosome sequences. Thus, the *NSRE* distribution could describe the constitutive property of a group of nucleosome sequences. Higher the *NSRE* is, more specific the constitutive property will be.

3. Results

3.1. NSRE distributions of nucleosome sequences in different organisms

We calculated the *NSRE* of nucleosome core sequences based on the probability from 1-mer to 8-mer respectively in *Drosophila*, *S. pombe* and *S. cerevisiae*, and normalized the results (the normalization method was in S1). Then the differences of *NSRE* distributions were compared among different organisms and different *k*-mers (Fig. 1.). Results show that the dyad region always has a higher *NSRE*, in either case. However, there are some differences of *NSRE* distributions among different organisms: (1) The peak value of *NSRE* distribution in *Drosophila* is the highest, and that in *S. cerevisiae* is the lowest. (2) *NSRE* distributions of *Drosophila* have three peaks, and that in *S. pombe* and *S. cerevisiae* both have two peaks. (3) The peaks of *NSRE* distributions in *Drosophila* all appear in the downstream of the dyad, locating at + 3, + 13 and + 26 bp. The peaks in *S. pombe* are symmetrically located at + 3 bp and − 3 bp in the both sides of the dyad as well as *S. cerevisiae*. Furthermore, there are also differences among different *k*-mers: it is a general tendency that, smaller the *k* is, higher the peak value of *NSRE* distribution will be, at any peak location except for + 13 bp in *Drosophila*. The characteristics of *NSRE* distributions are obvious while $k \leq 3$. And while $k = 8$, *NSRE* distributions of nucleosome sequences hardly have any features, which is similar to the distributions of random sequences (Fig. S2). So we could indicate that: (1) The sequence usage of nucleosome has differences among different organisms, and such

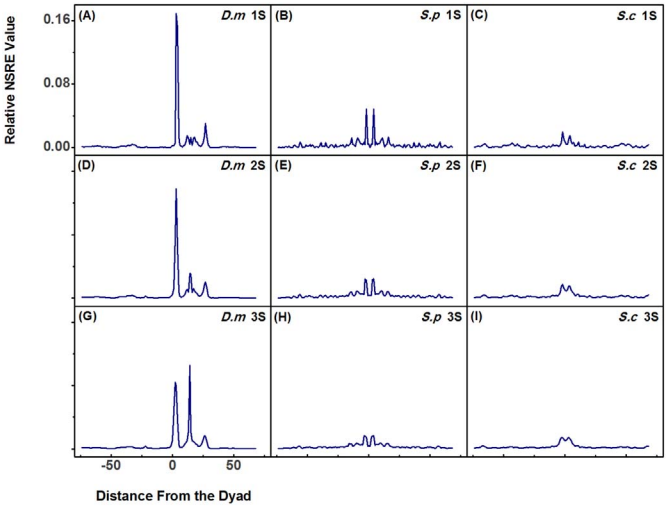

Fig. 1. *NSRE* distributions of nucleosome core sequences based on the probability of *k*-mer ($k = 1, 2, 3$) in 3 organisms. Nucleosome core sequence contains 75 bp on each side of dyad, totally 151 bp. *D.m* represents *Drosophila*. *S.p* represents *Schizosaccharomyces pombe*. *S.c* represents *Saccharomyces cerevisiae*. (A)–(C) *NSRE* distributions based on the information of 1-mer in 3 organisms. (D)–(F) *NSRE* distributions based on the information of 2-mer in 3 organisms. (G)–(I) *NSRE* distributions based on the information of 3-mer in 3 organisms.

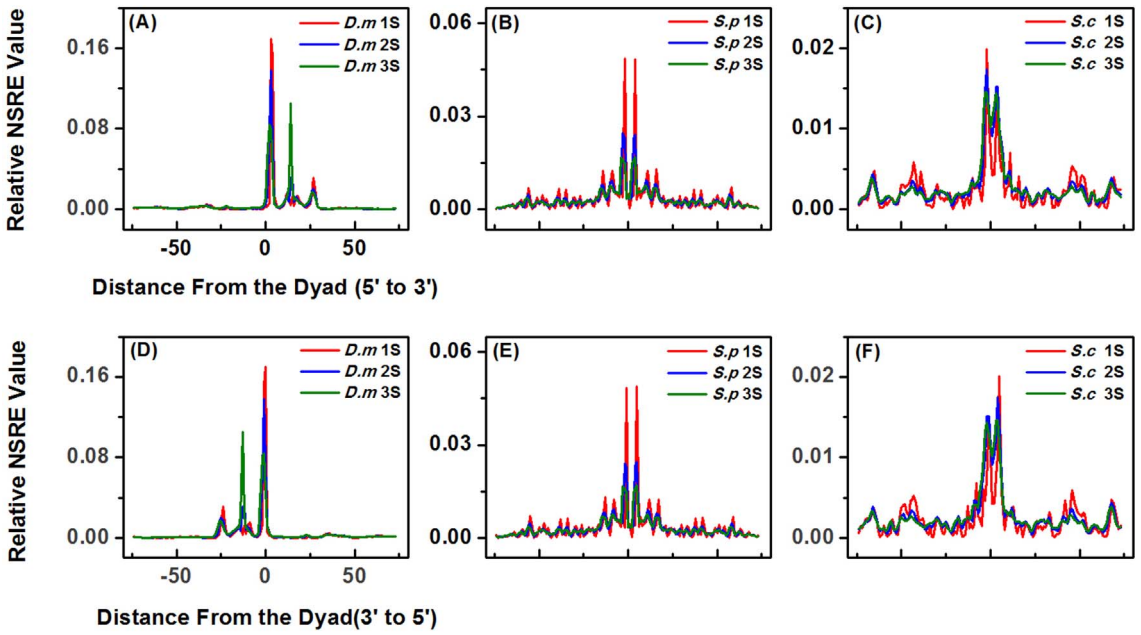

Fig. 2. *NSRE* distributions of normal sequences and their reverse complements of nucleosome in three organisms. (A)–(C) *NSRE* calculated by the normal sequences of nucleosome. (D)–(F) *NSRE* calculated by the reverse complements of nucleosome. (A) (D) *NSRE* distributions obtained by the probabilities of 1-mer, 2-mer and 3-mer in *Drosophila*. (B) (E) *NSRE* distributions obtained by the probabilities of 1-mer, 2-mer and 3-mer in *S. pombe*. (C) (F) *NSRE* distributions obtained by the probabilities of 1-mer, 2-mer and 3-mer in *S. cerevisiae*.

differences are mainly reflected in some certain sites on nucleosome sequence. (2) Single nucleotide ($k = 1$) makes major contribution to the constitutive property of nucleosome sequence, but it seems that 3-mer is more important at + 13 bp in *Drosophila*'s nucleosome. (3) The degree of sequence dependence to nucleosome is *Drosophila* > *S. pombe* > *S. cerevisiae*. *Drosophila* is a typical metazoan, and *S. cerevisiae* is a typical monad. Though *S. pombe* is also monad, it is distinguished from *S. cerevisiae* by sharing important characteristics of chromosome structure with metazoan [20,46–49]. So *NSRE* distributions could well reflect the evolutionary relationship of nucleosome sequence among these three organisms.

Above *NSRE* calculations are all based on the sequences in normal chain. For discussing whether the Watson strand or the Crick strand affects the *NSRE* distribution characteristics of nucleosome sequences, we also analyzed the reverse complements as a contrast. *NSRE* was calculated by the probabilities of 1-mer, 2-mer and 3-mer respectively in three organisms based on nucleosome sequences both in normal sequences and their reverse complements. Results have shown in Fig. 2. *NSRE* distributions of the normal sequences and their reverse complements are symmetrical. When in the same orientation, normal sequences and their reverse complements have a consistent *NSRE* distribution feature. It is indicated that the sequence information which determines nucleosome positioning is strand-insensitive. No matter it is Watson strand or Crick strand, it contains the characteristic information, and the information is highly consistent. Therefore, for nucleosome sequence research, it just needs the information of either strand. In the following *NSRE* calculation, the sequences in normal chain were selected.

3.2. NSRE distributions of nucleosome sequences around TSS

Above results show that there are differences of constitutive property of nucleosome sequences among three organisms. Nucleosome organization around TSS is important to gene transcription regulation. So the nucleosomes around TSS were studied in detail. *NSRE* distribution of 1-mer has more significant feature, but it lacks of the information about bases correlation. In order to reflect more comprehensive information, we added up the values of *NSRE* obtained by the probability of 1-mer, 2-mer and 3-mer, and got an Accumulated *NSRE* (Ac-*NSRE*). Then the Ac-*NSRE* was used to analyze − 3, − 2, − 1, + 1, + 2 and + 3 nucleosome on TSS flanking region (the definition of + 1/− 1 nucleosomes was shown in S3).

First, above nucleosomes were divided into two groups by the location relative to TSS: Upstream(Up)-nucleosomes (− 3, − 2, − 1 nucleosome) and Downstream(Dn)-nucleosomes (+ 1, + 2, + 3 nucleosome). Then the Ac-*NSRE* of nucleosome sequences of these two groups were calculated respectively, and were compared with the Ac-*NSRE* of all nucleosome sequences in each organism. Results have shown in Fig. 3. In *Drosophila*, Ac-*NSRE* distribution characteristics of Up-nucleosomes and Dn-nucleosomes have no obvious difference from all nucleosomes', and just the Up-nucleosome sequences have a stronger constitutive property. In *S. pombe*, Ac-*NSRE* is significantly decreased in the dyad region both in Up-nucleosomes and Dn-nucleosomes. In *S. cerevisiae*, Ac-*NSRE* distribution characteristics of Up-nucleosomes have a slight difference from all nucleosomes', and it changes more obviously in Dn-nucleosomes. Ac-*NSRE* increases in the region close to TSS, which exists in both two groups.

We also calculated the Ac-*NSRE* of sequences of each nucleosome (− 3 to + 3) around TSS. Results have shown in Figs. 4 and S4. Totally, − 1, − 2 and − 3 nucleosomes upstream of TSS have consistent distribution characteristics. + 1, + 2 and + 3 nucleosomes downstream of TSS also have consistent distribution characteristics, but different from nucleosomes' upstream. In *Drosophila* and *S. cerevisiae*, Ac-*NSRE* distribution characteristic of individual nucleosome has no obvious differences from that of grouping nucleosome. But in *S. pombe*, the dyad region of individual nucleosome still has a high Ac-*NSRE*, which is different from the case after grouping. It is indicated that, in *S. pombe*, the constitutive property of the individual nucleosome around TSS is consistent, but the nucleotide element which determines the constitutive property of nucleosome sequence is not the same.

By comparing Ac-*NSRE* distributions of + 1/− 1 nucleosomes among three organisms (Fig. 4), we could see that, the altered degree of Ac-*NSRE* distribution characteristic in + 1/− 1 nucleosomes compared to all nucleosomes is significantly different from *Drosophila* to *S. cerevisiae*. In *Drosophila*, + 1/− 1 nucleosomes have maintained a fundamental Ac-*NSRE* distribution characteristic compared to all nucleosomes, and just − 1 nucleosome seems to have a stronger constitutive property. In *S. cerevisiae*, Ac-*NSRE* distribution characteristic of + 1/

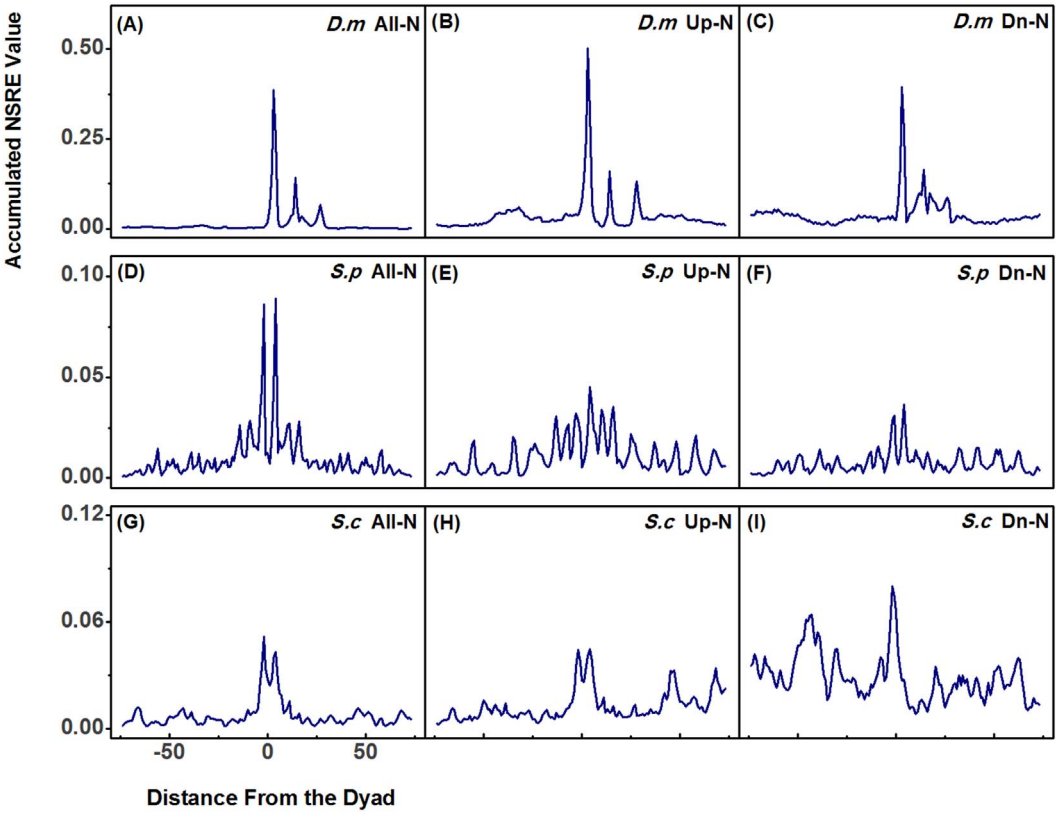

Fig. 3. Ac-*NSRE* distributions of nucleosome core sequences around TSS in 3 organisms. Ac-*NSRE* was calculated in All-N, Up-N and Dn-N respectively. All-N represents Ac-*NSRE* obtained by all nucleosome sequences of an organism. Up-N represents Ac-*NSRE* obtained by nucleosome sequences upstream of TSS. Dn-N represents Ac-*NSRE* obtained by nucleosome sequences downstream of TSS. (A)–(C) Ac-*NSRE* distributions of different groups of nucleosomes in *Drosophila*. (D)–(F) Ac-*NSRE* distributions of different groups of nucleosomes in *S. pombe*. (G)–(I) Ac-*NSRE* distributions of different groups of nucleosomes in *S. cerevisiae*.

-1 nucleosomes alter obviously compared to all nucleosomes, and especially in the region close to TSS, Ac-*NSRE* increases significantly. Moreover, the Ac-*NSRE* distribution characteristic in the region of downstream of -1 nucleosome and upstream of $+1$ nucleosome is similar to that in dyad region. Overall, in these three organisms, from monad to metazoan (*S. cerevisiae* → *S. pombe* → *Drosophila*), the constitutive property of -1 nucleosome becomes stronger and stronger, and the constitutive property of $+1$ nucleosome becomes more and more conserved. In other words, -1 nucleosome trends to be more functional and $+1$ nucleosome trends to be more stable.

3.3. The determined elements of NSRE distribution peaks

NSRE distributions have shown the nucleosome sequence characteristics in three organisms, and the most important feature is the distinctiveness of nucleosome dyad region. *NSRE* distribution peaks always locate near the dyad. It is indicated that nucleosome stability is mainly determined by the sequence of dyad region. However, it is unclear that which element contributes to *NSRE* distribution peaks, and whether it is the same in different organisms. So we analyzed the base composition of nucleosome core sequences in *Drosophila*, *S. pombe* and *S. cerevisiae*. Results show that, no matter in which organism, there are always two special sites in nucleosome core sequences: one is A-rich and T-poor, and the other is just opposite (Fig. 5). In *Drosophila*, the two special sites locate on $+2$ (A-poor and T-rich) and $+3$ bp (A-rich and T-poor). In *S. pombe* and *S. cerevisiae*, the two special sites locate on -3 (A-rich and T-poor) and $+3$ bp (A-poor and T-rich). These sites are just the positions of *NSRE* distribution peaks. Brogaard et al. have pointed out that these two special sites exist in yeast nucleosome sequences, and it could not exclude the possibility that chemical map might present a bias [19]. The *NSRE* distribution features in the dyad region also may be caused by the bias of chemical map.

Base composition analysis reflects the difference in base usage frequency. Especially in some special sites, base A & T have a variable usage frequency, and base C & G have a steady usage frequency. It indicates that the *NSRE* based on the probabilities of A & T should be different from that based on the probabilities of C & G. Then *NSRE* was calculated separately by the probabilities of base A & T and C & G. Results have shown in Fig. 6. We could see that, in all three organisms, *NSRE* obtained by A & T is significantly higher than the *NSRE* obtained by C & G, and *NSRE* distribution of base A & T is similar as the *NSRE* distribution of all bases. That is to say, the *NSRE* distribution characteristics of nucleosome core sequences are mainly depended on the usage frequency of base A & T, and it is almost irrelevant to base C & G.

Base A & T have made a majority of contribution to *NSRE* distribution characteristics of all nucleosome core sequences in three organisms. Further, *NSRE* was calculated separately by base A & T and base C & G only in $+1/-1$ nucleosomes. Results have shown in Fig. 7. In *Drosophila* and *S. pombe*, $+1/-1$ nucleosomes have the same way in base frequency usage for *NSRE*: base A & T are major determinants for *NSRE* distribution characteristics, and base C & G are nearly useless. But in *S. cerevisiae*, there are some differences: base C & G have also taken part in the construction of *NSRE* distribution characteristics, though it only exists in the flanking region of nucleosome sequence but not in the dyad region. Results indicate that the functional mechanism based on sequence of $+1/-1$ nucleosomes in *S. cerevisiae* is different from that in *Drosophila* and *S. pombe*. That is to say the transcription regulation mechanism based on nucleosomes is different from *S. cerevisiae* to the other two organisms. In addition, in either case, base C & G are not related to nucleosome dyad region, and only base A & T determine the dyad region features.

Dinucleotide (2-mer) and trinucleotide (3-mer) contain the

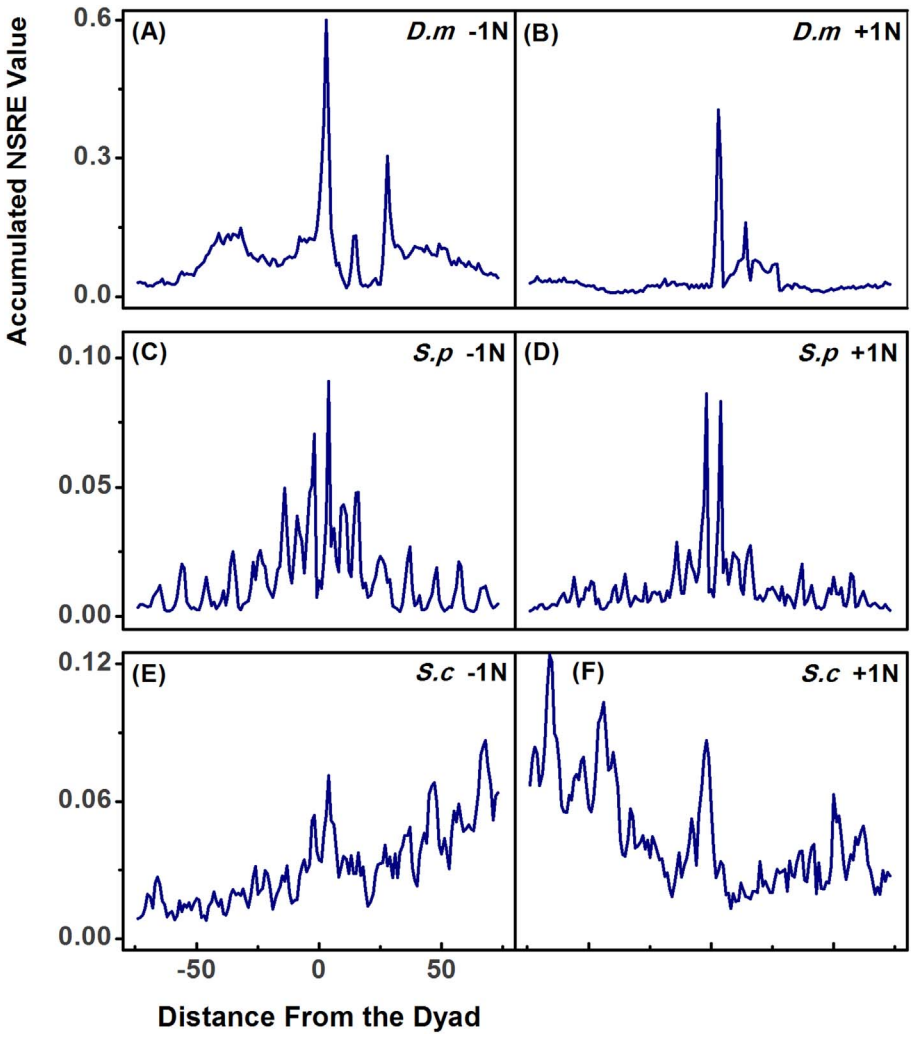

Fig. 4. Ac-*NSRE* distributions of +1/−1 nucleosome sequences in 3 organisms. (A)–(B) Ac-*NSRE* distributions of −1/+1 nucleosome sequences in *Drosophila*. (C)–(D) Ac-*NSRE* distributions of −1/+1 nucleosome sequences in *S. pombe*. (*E*)–(F) Ac-*NSRE* distributions of −1/+1 nucleosome sequences in *S. cerevisiae*.

information of bases correlation. *NSRE* distributions obtained by the probabilities of dinucleotide and trinucleotide also have shown a conserved feature in nucleosome dyad region. In single base *NSRE* distributions, base A & T determine the dyad region features. For discussing the effect of base A & T for dinucleotide *NSRE* and trinucleotide *NSRE*, we calculated *NSRE* separately by the probabilities of dinucleotide and trinucleotide with different numbers of A & T. Results have shown in Fig. 8. In dinucleotide *NSRE* distributions, it shows a tendency that the peak value of *NSRE* distributions increases as the number of A & T in dinucleotide increases, and this tendency is also reflected in trinucleotide *NSRE* distributions. The higher amount of base A & T in dinucleotide or trinucleotide makes the higher peak value of *NSRE* distribution, and this is the same in different organisms. However, there is a special case. In *Drosophila*, base C & G are not useless for *NSRE*. The analysis of *NSRE* distribution based on various *k*-mer has shown that the peak value of *NSRE* distribution decreases as *k* increases generally, but the situation at the location of +13 bp in *Drosophila* nucleosome sequence is just the opposite (Fig. 1). Especially in trinucleotide *NSRE* distribution (Fig. 1G), the location of +13 bp has an extremely high peak value, and we can describe it as a Special Peak. Through calculating *NSRE* separately by base A & T and base C & G, we found this Special Peak is just caused by the trinucleotide only composed with C & G.

The peaks in the dyad region are conserved in three organisms, and

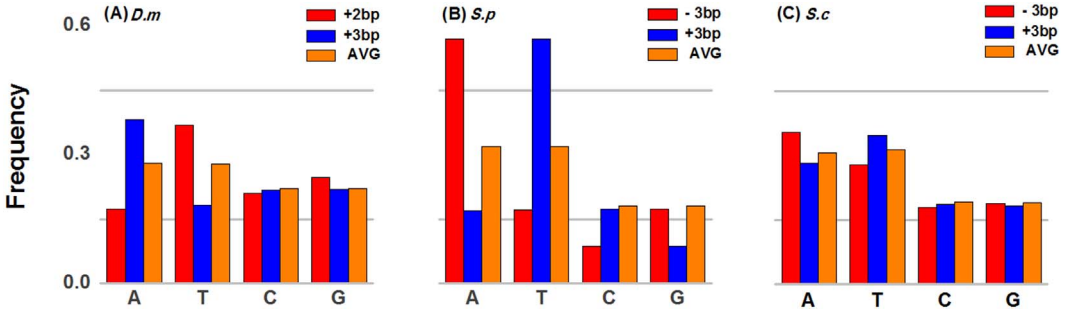

Fig. 5. The base frequency on two special sites in nucleosome core sequence in 3 organisms. (A) The base frequency on +2 bp & +3 bp and average frequency in *Drosophila*. (B) The base frequency on −3 bp & +3 bp and average frequency in *S. pombe*. (C) The base frequency on −3 bp & +3 bp and average frequency in *S. cerevisiae*.

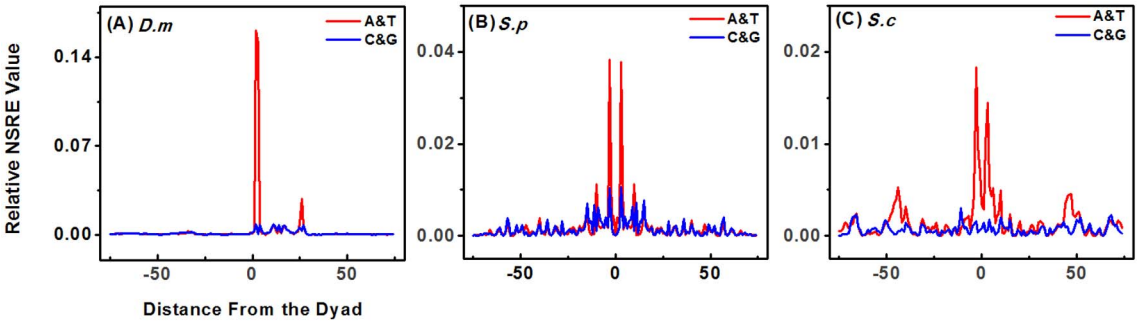

Fig. 6. Two *NSRE* distributions obtained separately by base A & T and C & G in 3 organisms. *NSRE* in red line is calculated by the probabilities of base A & T. *NSRE* in blue line is calculated by the probabilities of base C & G. The comparisons are shown respectively in *Drosophila* (A), *S. pombe* (B) and *S. cerevisiae* (C). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the formation of these peaks is strongly associated with base A & T. Further, we analyzed the correlation of the dependence degree on base A & T for dyad peaks among three organisms, and three organisms show a good correlation(correlation coefficient was shown in S5). The dependence degree on base A & T for dyad peaks in three organisms is nearly the same (Fig. 9). It is indicated that the core characteristic of nucleosome sequences is conserved, and the mechanism of sequence usage for constructing this characteristics is also conserved.

4. Discussion

New Symmetric Relative Entropy was innovatively used in analysis of the differences between partial and whole. The differences of *k*-mer usage between each site and all sites on nucleosome sequence are reflected by *NSRE*. Compared with normal information content, *NSRE* introduced a background probability, which involved different background compositions of different organisms. Meanwhile, with its good properties, *NSRE* could well describe the constitutive property and better reflect the key feature of nucleosome sequences (Fig. S6). Therefore, *NSRE* is an effective method for evolution analysis of nucleosome sequences. *S. cerevisiae* and *S. pombe*, as monad, have a similar *NSRE* distribution characteristic of nucleosome sequences. Two significant peaks appear at -3 and $+3$ bp on nucleosome sequence, which is different from that in metazoan. In *Drosophila*, there is only one significant peak appearing at $+3$ bp, but the value of the peak is obviously higher than that in *S. cerevisiae* and *S. pombe*. Though *S. pombe* is

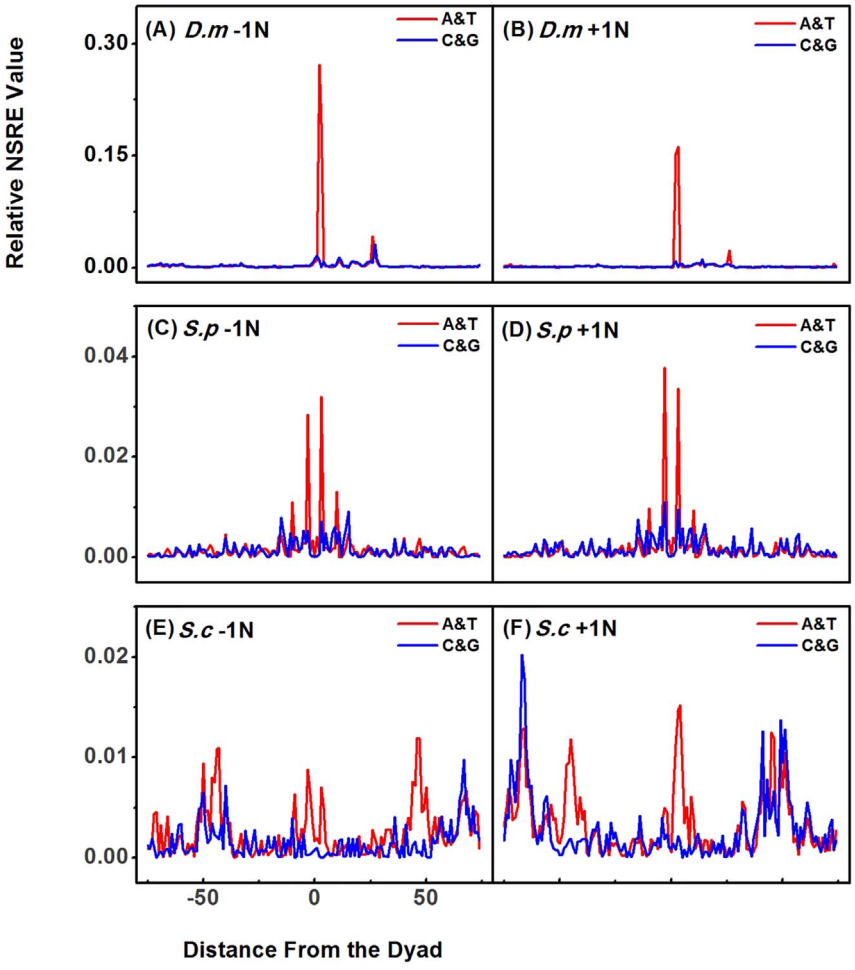

Fig. 7. Two *NSRE* distributions obtained separately by base A & T and base C & G in $+1/-1$ nucleosomes. (A)–(B) The comparisons in -1 and $+1$ nucleosome in *Drosophila*. (C)–(D) The comparisons in -1 and $+1$ nucleosome in *S. pombe*. (*E*)–(F) The comparisons in -1 and $+1$ nucleosome in *S. cerevisiae*.

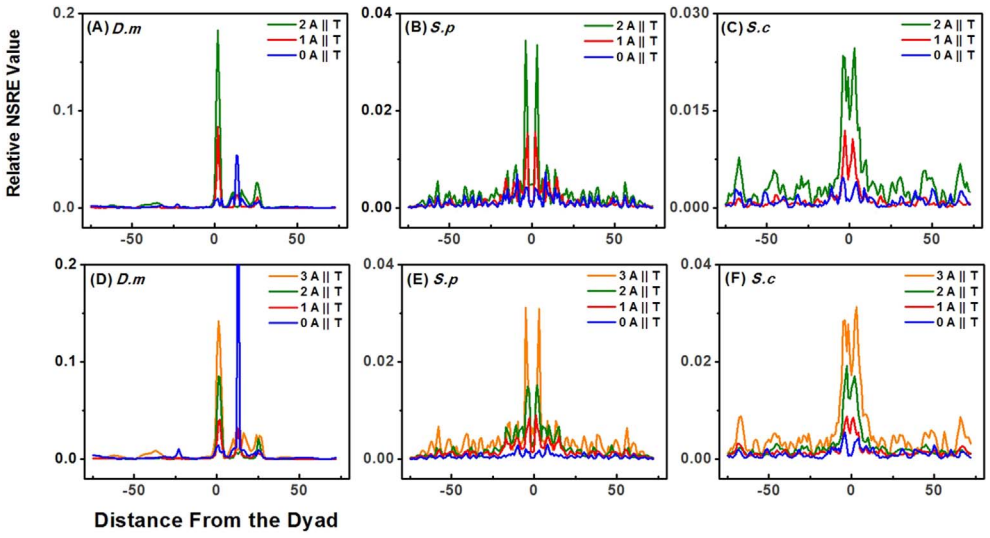

Fig. 8. *NSRE* distributions obtained separately by dinucleotide and trinucleotide with different numbers of A & T. The dinucleotide comparisons are shown respectively in *Drosophila* (A), *S. pombe* (B) and *S. cerevisiae* (C). *NSRE* in green line is calculated by the probabilities of dinucleotide containing only A/T. *NSRE* in red line is calculated by the probabilities of dinucleotide containing 1 A/T. *NSRE* in blue line is calculated by the probabilities of dinucleotide containing 0 A/T. The trinucleotide comparisons are shown respectively in *Drosophila* (D), *S. pombe* (E) and *S. cerevisiae* (F). *NSRE* in yellow line is calculated by the probabilities of trinucleotide containing only A/T. *NSRE* in green line is calculated by the probabilities of trinucleotide containing 2 A/T. *NSRE* in red line is calculated by the probabilities of trinucleotide containing 1 A/T. *NSRE* in blue line is calculated by the probabilities of trinucleotide containing 0 A/T. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

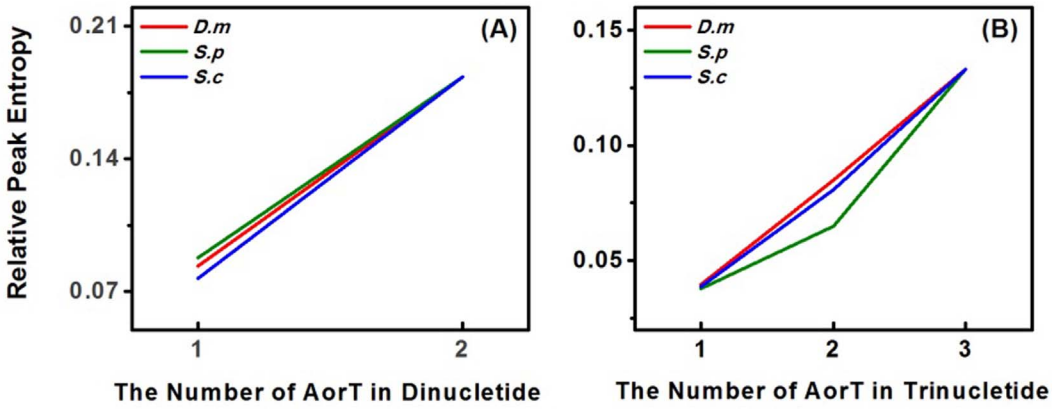

Fig. 9. Trends of the dependence degree on base A & T. Axis Y is the normalized peak value, and the value in *Drosophila* is the reference value. Axis X is the number of base A & T in dinucleotide or trinucleotide. About the peak value selection, in *Drosophila*, the average value of + 2 bp and + 3 bp is the peak value, and in *S. pombe* and *S. cerevisiae*, the average value of − 3 bp and + 3 bp is the peak value. (A) Trends of the dependence degree on base A & T in dinucleotide. (B) Trends of the dependence degree on base A & T in trinucleotide.

monad, it has some important characteristics of chromosome structure with metazoan, and its peak value is between the other two organisms'. So *NSRE* distributions reflect an evolutionary trend of nucleosome sequence from monad to metazoan: the constitutive property of nucleosome sequence becomes stronger and stronger. That is to say, DNA sequence preference is more important to nucleosome positioning in metazoan.

Ac-*NSRE* distribution characteristics of + 1 and − 1 nucleosomes have shown the differences of transcription regulatory mechanisms from monad to metazoan. In *S. cerevisiae*, Ac-*NSRE* distribution characteristics of + 1 and − 1 nucleosomes changed obviously compared with all nucleosomes'. The regions of downstream of − 1 nucleosome and upstream of + 1 nucleosome have shown a distribution characteristic which is similar to that in dyad region, and peak locations in these two regions would be potential nucleosome dyad. It is inferred that the position of + 1/− 1 nucleosomes is not fixed, and + 1/− 1 nucleosomes could slide. Gene transcription regulation could be done by the position modification of + 1/− 1 nucleosomes in *S. cerevisiae*. But in *Drosophila*, Ac-*NSRE* distribution characteristics of + 1/− 1 nucleosomes changed slightly compared with all nucleosomes'. It is indicated that the position of + 1/− 1 nucleosomes is stable and conserved, and genes in *Drosophila* may use another transcription regulatory mechanism, such as the dispelling of nucleosomes. Though the Ac-*NSRE* distribution characteristic of all nucleosomes in *S. pombe* is more similar to that in *S. cerevisiae*, the change of Ac-*NSRE* distribution characteristic of + 1/− 1 nucleosomes in *S. pombe* is more similar to that in *Drosophila*. *S. pombe* is in a transition stage from monad to metazoan. In addition, it could be inferred that the evolution of gene transcription regulation mechanism should precede the evolution of nucleosome sequence.

Consistently, the dyad region of nucleosome sequence is conserved in three organisms. The conservative property is reflected in not only the high *NSRE* but also the formation mechanism of the high *NSRE*. *NSRE* distribution characteristics of nucleosome sequences show that the stability of nucleosome mainly depends on the dyad region, and the feature of the dyad region mainly depends on the base A & T, and the dependence degree on base A & T is basically the same in three organisms. Overall, *NSRE* distributions have shown differences of nucleosome sequence characteristics among different organisms, and it has also reflected the conservative property in sequence usage of nucleosome. *NSRE* is an effective method in constitutive property analysis for a set of sequences. However, the basement of using *NSRE* to analyze nucleosome sequences is single base pair resolution map of nucleosome position, and related data are limited in a few of model organisms. Optimistically, with the development of experimental technology of nucleosome positioning, the base pair resolution data of nucleosome position will be obtained in more organisms. Then *NSRE* will be a powerful tool for nucleosome sequences evolution research.

Acknowledgments

This work was supported by grants from the National Natural

Science Foundation of China (No. 31260219), doctor subject Foundation of the Ministry of Education of China (No. 20121501110006).

Competing financial interests

The authors declare no competing financial interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2017.09.007.

References

[1] K. Luger, A.W. Mäder, R.K. Richmond, D.F. Sargent, T.J. Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution, Nature 389 (1997) 251–260.
[2] P. Zhu, G. Li, Structural insights of nucleosome and the 30-nm chromatin fiber, Curr. Opin. Struct. Biol. 36 (2016) 106–115.
[3] M. Radman-Livaja, O.J. Rando, Nucleosome positioning: how is it established, and why does it matter? Dev. Biol. 339 (2010) 258–266.
[4] L. Bai, A.V. Morozov, Gene regulation by nucleosome positioning, Trends Genet. 26 (2010) 476–483.
[5] S. Kubik, M.J. Bruzzone, P. Jacquet, J.L. Falcone, J. Rougemont, D. Shore, Nucleosome stability distinguishes two different promoter types at all protein-coding genes in yeast, Mol. Cell 60 (2015) 422–434.
[6] R.Z.V. Chereji, A.V. Morozov, Ubiquitous nucleosome crowding in the yeast genome, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 5236–5241.
[7] I. Tirosh, N. Barkai, Two strategies for gene regulation by promoter nucleosomes, Genome Res. 18 (2008) 1084–1091.
[8] J. Ocampo, R.V. Chereji, P.R. Eriksson, D.J. Clark, The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo, Nucleic Acids Res. 44 (2016) 4625–4635.
[9] A. Weiner, A. Hughes, M. Yassour, O.J. Rando, N. Friedman, High-resolution nucleosome mapping reveals transcription-dependent promoter packaging, Genome Res. 20 (2010) 90–100.
[10] C. Lieleg, N. Krietenstein, M. Walker, P. Korber, Nucleosome positioning in yeasts: methods, maps, and mechanisms, Chromosoma 124 (2015) 131–151.
[11] K. Struhl, E. Segal, Determinants of nucleosome positioning, Nat. Struct. Mol. Biol. 20 (2013) 267–273.
[12] C. Jiang, B.F. Pugh, Nucleosome positioning and gene regulation: advances through genomics, Nat. Rev. Genet. 10 (2009) 161–172.
[13] W. Lee, D. Tillo, N. Bray, R.H. Morse, R.W. Davis, T.R. Hughes, C. Nislow, A high-resolution atlas of nucleosome occupancy in yeast, Nat. Genet. 39 (2007) 1235–1244.
[14] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J.A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, S.M. Johnson, A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning, Genome Res. 18 (2008) 1051–1063.
[15] T.N. Mavrich, C. Jiang, I.P. Ioshikhes, X. Li, B.J. Venters, S.J. Zanton, L.P. Tomsho, J. Qi, R.L. Glaser, S.C. Schuster, D.S. Gilmour, I. Albert, B.F. Pugh, Nucleosome organization in the drosophila genome, Nature 453 (2008) 358–362.
[16] F. Ozsolak, J.S. Song, X.S. Liu, D.E. Fisher, High-throughput mapping of the chromatin structure of human promoters, Nat. Biotechnol. 25 (2007) 244–248.
[17] D.E. Schones, K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, K. Zhao, Dynamic regulation of nucleosome positioning in the human genome, Cell 132 (2008) 887–898.
[18] C.D. Schmid, P. Bucher, ChIP-Seq data reveal nucleosome architecture of human promoters, Cell 131 (2007) 831–832.
[19] K. Brogaard, L. Xi, J.P. Wang, J. Widom, A map of nucleosome positions in yeast at base-pair resolution, Nature 486 (2012) 496–501.
[20] G. Moyle-Heyrman, T. Zaichuk, L. Xi, Q. Zhang, O.C. Uhlenbeck, R. Holmgren, J. Widom, J.P. Wang, Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 20158–20163.
[21] B. Eslami-Mossallam, H. Schiessel, J. van Noort, Nucleosome dynamics: sequence matters, Adv. Colloid Interf. Sci. 232 (2016) 101–113.
[22] N. Kaplan, I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, E.M. LeProust, T.R. Hughes, J.D. Lieb, J. Widom, E. Segal, The DNA-encoded nucleosome organization of a eukaryotic genome, Nature 458 (2008) 362–366.
[23] D.J. Clark, Nucleosome positioning, nucleosome spacing and the nucleosome code, J. Biomol. Struct. Dyn. 27 (2010) 781–793.
[24] G. Liu, J. Liu, X. Cui, L. Cai, Sequence-dependent prediction of recombination hotspots in *Saccharomyces cerevisiae*, J. Theor. Biol. 293 (2012) 49–54.
[25] R. Ogawa, N. Kitagawa, H. Ashida, R. Saito, M. Tomita, Computational prediction of nucleosome positioning by calculating the relative fragment frequency index of nucleosomal sequences, FEBS Lett. 584 (2010) 1498–1502.
[26] H.E. Peckham, R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W.S. Noble, K. Struhl, Z. Weng, Nucleosome positioning signals in genomic DNA, Genome Res. 17 (2007) 1170–1177.
[27] A. Scipioni, P. De Santis, Predicting nucleosome positioning in genomes: physical and bioinformatic approaches, Biophys. Chem. 155 (2011) 53–64.
[28] E. Segal, Y. Fondufemittendorf, L. Chen, A.C. Thåström, Y. Field, I.K. Moore, J.P.Z. Wang, J. Widom, A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.
[29] S. Tesoro, I. Ali, A.N. Morozov, N. Sulaiman, D. Marenduzzo, A one-dimensional statistical mechanics model for nucleosome positioning on genomic DNA, Phys. Biol. 13 (2016) 016004.
[30] Y. Xing, X. Zhao, L. Cai, Prediction of nucleosome occupancy in *Saccharomyces cerevisiae* using position-correlation scoring function, Genomics 98 (2011) 359–366.
[31] G.-C. Yuan, J.S. Liu, Genomic sequence is highly predictive of local nucleosome depletion, PLoS Comput. Biol. (2005) e13 (preprint).
[32] G. Liu, Y. Xing, H. Zhao, J. Wang, Y. Shang, L. Cai, A deformation energy-based model for predicting nucleosome dyads and occupancy, Sci Rep 6 (2016) 24133.
[33] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, K.C. Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[34] C. Wei, L. Hao, P.M. Feng, D. Chen, Y.C. Zuo, K.C. Chou, iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties, PLoS One 7 (2012) e47843.
[35] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[36] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[37] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[38] B. Hao, H. Xie, Z. Yu, G. Chen, Avoided strings in bacterial complete genomes and a related combinatorial problem, Ann. Comb. 4 (2000) 247–255.
[39] C. Wu, The distributions of the frequency of occurrence of nucleotide subsequences, Methodol. Comput. Appl. Probab. 7 (2005) 325–334.
[40] T. Bao, L. Hong, X. Zhao, G. Liu, Predicting nucleosome binding motif set and analyzing their distributions around functional sites of human genes, Chromosom. Res. 20 (2012) 685–698.
[41] J. Wen, R.H. Chan, S.C. Yau, R.L. He, S.S. Yau, K-mer natural vector and its application to the phylogenetic analysis of genetic sequences, Gene 546 (2014) 25–34.
[42] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–86.
[43] Q. Fu, M.P. Qian, L.B. Chen, Y.X. Zhu, Features of coding and noncoding sequences based on 3-tuple distributions, Yi Chuan Xue Bao Acta Genet. Sin. 32 (2005) 1018–1026.
[44] J. Shen, New Symmetric Relative Eetropy and Similarity Analysis for DNA Sequences in, Northwest A & F University, (2010).
[45] Y. Zheng, H. Li, Y. Wang, H. Meng, Q. Zhang, X. Zhao, Evolutionary mechanism and biological functions of 8-mers containing CG dinucleotide in yeast, Chromosom. Res. (2017) 1–17.
[46] B.J. Monahan, J. Villen, S. Marguerat, J. Bahler, S.P. Gygi, F. Winston, Fission yeast SWI/SNF and RSC complexes show compositional and functional differences from budding yeast, Nat. Struct. Mol. Biol. 15 (2008) 873–880.
[47] F.E. Reyes-Turcu, S.I. Grewal, Different means, same end-heterochromatin formation by RNAi and RNAi-independent RNA processing factors in fission yeast, Curr. Opin. Genet. Dev. 22 (2012) 156–163.
[48] L. Clarke, Centromeres of budding and fission yeasts, Trends Genet. 6 (1990) 150–154.
[49] J. Xu, Y. Yanagisawa, A.M. Tsankov, C. Hart, K. Aoki, N. Kommajosyula, K.E. Steinmann, J. Bochicchio, C. Russ, A. Regev, O.J. Rando, C. Nusbaum, H. Niki, P. Milos, Z. Weng, N. Rhind, Genome-wide identification and characterization of replication origins by deep sequencing, Genome Biol. 13 (2012) R27.