Genomics xxx (xxxx) xxx–xxx

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

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

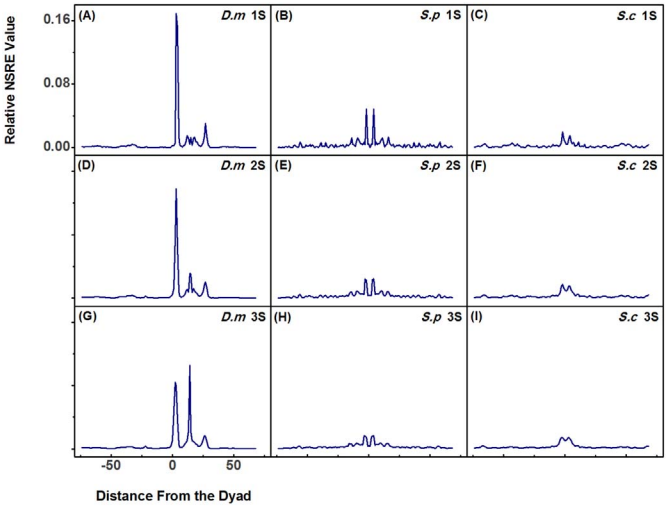

Fig. 1. *NSRE* distributions of nucleosome core sequences based on the probability of *k*-mer ($k = 1, 2, 3$) in 3 organisms. Nucleosome core sequence contains 75 bp on each side of dyad, totally 151 bp. *D.m* represents *Drosophila*. *S.p* represents *Schizosaccharomyces pombe*. *S.c* represents *Saccharomyces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