Journal of Theoretical Biology 284 (2011) 92–98

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Correlation between the flexibility and periodic dinucleotide patterns in yeast nucleosomal DNA sequences

Qinqin Wu [a,b,*], Weiqiang Zhou [a], Jiajun Wang [b], Hong Yan [a,c]

[a] Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong
[b] School of Electronics and Information Engineering, Soochow University, Suzhou, China
[c] School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 9 June 2010
Received in revised form
20 June 2011
Accepted 21 June 2011
Available online 25 June 2011

Keywords:
Nucleosomes
DNA sequence flexibility
Periodic dinucleotides
Nucleosome positioning and formation

ABSTRACT

Nucleosome formation and positioning, which play important roles in a number of biological processes, are thought to be related to the distinctive periodic dinucleotide patterns observed in the DNA sequence wrapped around the protein octamer. Previous research shows that flexibility is a key structural property of a nucleosomal DNA sequence. However, the relationship between the flexibility and the periodic dinucleotide patterns has received little attention in research in the past. In this study, we propose the use of three different models to measure the flexibility of yeast DNA sequences. Although the three models involve different parameters, they deliver consistent results showing that yeast nucleosomal DNA sequences are more flexible than non-nucleosomal ones. In contrast to random flexibility values along non-nucleosomal DNA sequences, the flexibility of nucleosomal DNA sequences shows a clear periodicity of 10.14 base pairs, which is consistent with the periodicity of dinucleotide distributions. We also demonstrate that there is a strong relationship between the peak positions of the flexibility and the dinucleotide frequencies. Correlation between the flexibility and the dinucleotide patterns of CA/TG, CG, GC, GG/CC, AG/CT, AC/GT and GA/TC are positive with an average value of 0.5946. The highest correlation is shown by CA/TG with a value of 0.7438 and the lowest correlation is shown by AA/TT with a value of −0.7424. The source codes and data sets are available for downloading on http://www.hy8.com/bioinformatics.htm.

1. Introduction

Nucleosomes are the fundamental repetitive units of the eukaryotic chromatin, which play important roles in a number of biological processes (Grunstein, 1997; Kornberg and Lorch, 1999; Ashburner et al., 2000). A nucleosome is formed by ~147 base pairs (bp) of DNA wrapped around a disk-like histone octamer. Fig. 1 shows the three-dimensional (3D) structure of a nucleosome viewed from two different angles. The diagrams are created using UCSF Chimera (Pettersen et al., 2004). Previous studies suggest that the rotational positioning of DNA in the nucleosomes appears to be dominated by certain sequence-dependent modulations in the structure (Drew and Travers, 1985). Besides, there is an attractive hypothesis proposed in earlier research that the distinctive periodic variations in the DNA sequence may facilitate the sharp bending of the DNA around the nucleosome (Satchwell et al., 1986). However, this hypothesis simply comes from a comparison between the periodicity of the sequence modulation and that of the rotation of the double helix along the surface of the nucleosome (Drew and Travers, 1985; Satchwell et al., 1986). Although large-scale studies have been conducted on nucleosomes (Ababneh, 2009; Albert et al., 2007; Bina, 1994; Chela-Flores, 1994; Chen et al., 2008; Drew and Travers, 1985; Miele et al., 2008; Mobius et al., 2006; Peckham et al., 2007; Richmond and Davey, 2003; Satchwell et al., 1986; Sevinc et al., 2004; Widom, 2001; Wu et al., 2009; Yassour et al., 2008; Zhao and Yan, 2009), no comprehensive analysis of the structural properties of a large number of experimentally determined nucleosomal DNA sequences has been carried out.

The structural properties of DNA play important roles in gene expression regulation (Alberts et al., 2002; Brukner et al., 1995; Fukue et al., 2005; Gowers and Halford, 2003; Matthews, 1992; Packer et al., 2000b; Pedersen et al., 1998; Starr et al., 1995; Travers and Klug, 1990). A key structural property, sequence-dependent flexibility, has been found to be able to guide DNA-binding proteins efficiently to the target sites (Gowers and Halford, 2003). Earlier research on this structural property also suggests that the flexibility may influence DNA looping (Matthews, 1992), promoter activities (Fukue et al., 2005), nucleosome positioning (Pedersen et al., 1998) and transcription factor binding (Fukue et al., 2005; Starr et al., 1995).

* Corresponding author at: Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong.
E-mail address: suzhouqinqin@gmail.com (Q. Wu).

0022-5193/$ - see front matter © 2011 Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2011.06.026

The objective of this work is to study the structural properties of the experimentally determined nucleosomal DNA sequences as well as the correlation between the dinucleotide periodicity and flexibility. Based on different experimental or statistical flexibility models, we analyze the average flexibility of 34,876 nucleosomal DNA sequences in yeast. Consistent with the general concept that flexible DNA sequence segments wrap around a histone core more easily than rigid ones, which usually results in transcription repression (Mobius et al., 2006), we report that nucleosomal DNA sequences are highly flexible compared with other genomic regions. Furthermore, a comparative study demonstrates that there is a strong correlation between the flexibility and periodic dinucleotide patterns in nucleosomal DNA sequences.

2. Materials and methods

2.1. The genome sequences

All 34,876 experimentally determined non-overlapping nucleosomal DNA sequences (Albert et al., 2007) of 16 yeast chromosomes are extracted from the GenBank (http://www.ncbi.nlm.nih.gov/genbank/). Early research suggests that the nucleosomes are mainly 147 bp long (Widom, 2001) but some of them may be longer, say, 160 bp (Bryant et al., 2008) or even longer than that. Thus, to include as many nucleosomes as possible, the length of these sequences is set to 180 bp, that is, the sequences range from 90 bp upstream to 90 bp downstream relative to the nucleosomal centers determined by experiments (Albert et al., 2007).

To validate if the structural properties observed in the flexibility are unique in the nucleosomal DNA sequences, it is necessary to calculate the flexibility of non-nucleosomal DNA sequences and make a comparison between them. Thus, we retrieve DNA sequences which are determined as non-nucleosomes in the experiments of 16 yeast chromosomes (Albert et al., 2007) from the GenBank. These sequences also contain 180 bp in their lengths.

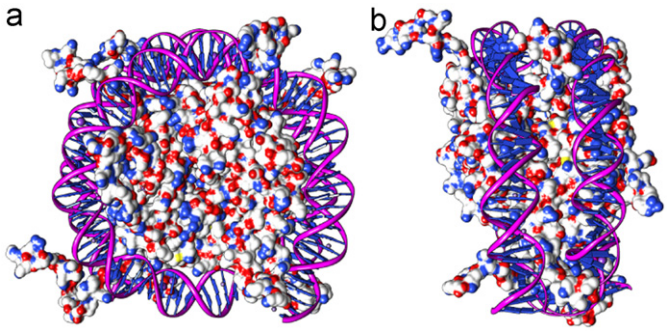

Fig. 1. The three-dimensional structure of a nucleosome viewed from different angles: (a) vertical view of the nucleosome structure and (b) horizontal view of the nucleosome structure.

Another group of sequences is also extracted from the GenBank, to make a general analysis of structural properties of the nucleosomal DNA sequences as well as their genomic surrounding regions based on the flexibility of 'non-nucleosome'–'nucleosome'–'non-nucleosome' sequences. However, a large number of nucleosomes are found to be closely positioned along the DNA sequence. Therefore, it is difficult to find a nucleosome which has a non-nucleosome sequence long enough on each side. Due to this constraint, this group includes only 235 sequences, ranging from 400 bp upstream of the nucleosome center position to 400 bp downstream of the nucleosome end site, i.e. [−400, nucleosomal region (147 bp long), +400]. After calculating the flexibility for each sequence according to the dinucleotide, trinucleotide and tetranucleotide models, we average the flexibility values of the nucleosome region for each model and obtain a single flexibility value of the nucleosome region against the center position. For a fair comparison, we also average the flexibility values of the non-nucleosome region for every overlapping 147 bp. Using this method, we obtain the final flexibility of [−400, +400] in length with the position 0 corresponding to the nucleosome center.

2.2. DNA flexibility models

Recently, various DNA flexibility models were proposed based on either experimental or statistical methods. These models make use of various structural properties including the angular parameters (twist, roll and tilt) as well as the translational parameters (shift, slide and rise). Using the flexibility parameters listed in the models, one can easily calculate the flexibility for DNA sequences. There are mainly three kinds of flexibility models, the dinucleotide model (Packer et al., 2000a), the trinucleotide model (Brukner et al., 1995) and the tetranucleotide model (Packer et al., 2000b). These three models are widely used to analyze the structural properties of, human promoters (Cao et al., 2008), mammalian and plant genomes (Florquin et al., 2005) as well as some other genomic regions. Here we intend to use these three flexibility models to find a common structural property of the sequences to be analyzed. Table 1 summarizes the characteristics of the flexibility models used in this paper. If all three models show consistent flexibility patterns, it can be confidently concluded that consensus physical signals do exist in the DNA sequences.

The three flexibility models used in this paper, as mentioned above, measure flexibility in different ways. The parameters for the dinucleotide model (Packer et al., 2000a) range from the smallest CA/TG=1.35 to the largest AA/TT=13.72 where larger values correspond to more rigid dinucleotides. For the trinucleotide model (Brukner et al., 1995), larger values represent more flexible trinucleotides. Each trinucleotide flexibility parameter ranges from the smallest AAT/ATT=−0.28 to the largest TCA/TGA=0.194. The tetranucleotide model (Packer et al., 2000b) is similar to the dinucleotide model, and its higher values correspond to more rigid tetranucleotides. The smallest flexibility parameter is TACA/TGTA=1.9 while the largest is AAAC/GTTT=27.2. Although different models exhibit different scales

Table 1
Characteristics of DNA flexibility models.

Model	The smallest value	Corresponding base pair steps	The largest value	Corresponding base pair steps	Characteristics
Dinucleotide model (Packer et al., 2000a)	1.35	CA/TG	13.72	AA/TT	The larger the more rigid
Trinucleotide model (Brukner et al., 1995)	−0.28	AAT/ATT	0.194	TCA/TGA	The larger the more flexible
Tetranucleotide model (Packer et al., 2000b)	1.9	TACA/TGTA	27.2	AAAC/GTTT	The larger the more rigid