Genomics 107 (2016) 69–75

Contents lists available at ScienceDirect

Genomics

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

Using deformation energy to analyze nucleosome positioning in genomes

Wei Chen [a,*], Pengmian Feng [b], Hui Ding [c], Hao Lin [c,d,**], Kuo-Chen Chou [d,e,***]

[a] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China
[b] School of Public Health, North China University of Science and Technology, Tangshan 063000, China
[c] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, and Center for Information in Biomedicine, School of Life Scienceand Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[d] Gordon Life Science Institute, Boston, MA 02478, USA
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form 6 December 2015
Accepted 22 December 2015
Available online 24 December 2015

Keywords:
Deformation energy
Nucleosome
SNP site
DSB site
ORI

ABSTRACT

By modulating the accessibility of genomic regions to regulatory proteins, nucleosome positioning plays important roles in cellular processes. Although intensive efforts have been made, the rules for determining nucleosome positioning are far from satisfaction yet. In this study, we developed a biophysical model to predict nucleosomal sequences based on the deformation energy of DNA sequences, and validated it against the experimentally determined nucleosome positions in the *Saccharomyces cerevisiae* genome, achieving very high success rates. Furthermore, using the deformation energy model, we analyzed the distribution of nucleosomes around the following three types of DNA functional sites: (1) double strand break (DSB), (2) single nucleotide polymorphism (SNP), and (3) origin of replication (ORI). We have found from the analyzed energy spectra that a remarkable "trough" or "valley" occurs around each of these functional sites, implying a depletion of nucleosome density, fully in accordance with experimental observations. These findings indicate that the deformation energy may play a key role for accurately predicting nucleosome positions, and that it can also provide a quantitative physical approach for in-depth understanding the mechanism of nucleosome positioning.

1. Introduction

In eukaryotes, 75%–95% of genomic DNAs are packaged into chromatins. The elementary structural unit of chromatin is nucleosome, formed by ~147 base pairs (bp) of DNA wrapped in superhelical turns around the surface of a histone octamer (composed of pairs of the four core histones H2A, H2B, H3 and H4) [1]. The packaging of DNA around the histone–octamer not only facilitates the storage of DNA in the limited cell space but also makes it possible to modulate the access of regulatory proteins to genomic regions. A growing body of evidence shows that nucleosomes play important roles in various biological processes, such as mRNA splicing, DNA replication and DNA repair [2–6]. Consequently, revealing the mechanism involved in controlling nucleosome positioning is fundamentally important for in-depth understanding the subsequent steps of gene expression.

High-resolution genome-wide nucleosome maps are now available for yeast, worms, flies and human genomes [7–10]. These high-resolution data provide unprecedented opportunities for further investigation of the mechanism of nucleosome positioning and its roles in gene regulation.

Since the nucleosome positioning code in yeast [11] was reported, various models have been proposed to elucidate nucleosome occupancy signals that determine the preference of a particular region to bind to histone and form a nucleosome [12–14], stimulating the recent breakthrough in developing computational predictors for identifying nucleosome positioning in genomes [15,16]. Although quite interesting and encouraging, the predictors based on the sequence information alone have been limited in their accuracy and resolution. Besides, the benchmark dataset used to train the sequence-based predictors may not be representative of direct histone–DNA binding. Therefore, it is highly desirable to develop a novel model that will have more direct and close correlation with nucleosome positioning.

* Correspondence to: W. Chen, Department of Physics, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China.
** Correspondence to: H. Lin, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China.
*** Correspondence to: K.-C. Chou, Gordon Life Science Institute, Boston, MA 02478, USA.
E-mail addresses: chenweiimu@gmail.com, wchen@gordonlifescience.org (W. Chen), fengpengmian@gmail.com (P. Feng), hding@uestc.edu.cn (H. Ding), hlin@uestc.edu.cn (H. Lin), kcchou@gordonlifescience.org (K.-C. Chou).

http://dx.doi.org/10.1016/j.ygeno.2015.12.005
0888-7543/

Recently, Miele et al. [17] reported that DNA physical properties were able to determine nucleosome occupancy from yeast to fly. Morozov et al. [18] proposed an *ab initio* model to predict nucleosomes by measuring the free energies of nucleosome formation. Nozaki et al. [19] and Wu et al. [20] suggested the existence of a highly bendable, fragile structure for nucleosomal DNA. By comparing the six DNA physical parameters (twist, roll, tilt, shift, slide, and rise) between nucleosomal and linker DNA sequences, we found that these DNA physical parameters are also quite useful for characterizing the description of nucleosomal DNA sequences [21]. All these facts indicate that there exists some structural code in DNA sequences that may be of use for determining the genome-wide nucleosome positioning.

The present study was devoted to investigate the deformation energy of DNA sequences and use it to develop a new model for predicting nucleosome positions. Since nucleosome positioning may affect all DNA-templated processes, it is important to analyze how those processes occur on nucleosome-structure DNA. But except for the transcriptional regulation, there are many unknowns yet for the molecular mechanisms of nucleosome positioning around the other functional sites. In order to dissect the roles of nucleosome positioning on them, we are to propose a biophysical model to analyze the distribution pattern of nucleosomes near some important functional sites, such as double strand break (DSB) site, single nucleotide polymorphism (SNP) site, and origin of replication (ORI). Using the proposed model, we not only have obtained the prediction results quite consistent with experimental observations, but also can reveal the distribution pattern of those nucleosomes that are near the aforementioned important functional sites.

As done in a series of recent publications [22–32] in proposing new analysis/prediction methods for biological systems, to make the presentation logically more clear and the results objectively more reliable, the following procedures [33] are followed: (1) construct or select a valid benchmark dataset to train and test the proposed model; (2) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be analyzed/predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the analysis/prediction; and (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the model. Below, we are to elaborate how to deal with these steps one by one.

2. Materials and methods

In this study, the benchmark dataset consists of two parts of DNA sequences. The first one is for analyzing nucleosome positioning, and the 2nd one for studying the genomic sequence patterns around some important functional sites.

2.1. Benchmark dataset for nucleosomal and linker sequences

In literature, the benchmark dataset usually consists of a training dataset and a testing dataset: the former is constructed for the purpose of training a proposed model, while the latter for the purpose of testing it. As pointed out in a comprehensive review [34], however, there is no need to separate a benchmark dataset into a training dataset and a testing dataset for validating a prediction method if it is tested by the jackknife or subsampling (K-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests. Therefore, the benchmark dataset for the current study may consist of a positive subset and a negative subset: the former contains only nucleosomal DNA sequences while the latter contains only the linker DNA sequences.

The reference genome sequence of *Saccharomyces cerevisiae* was obtained from the *Saccharomyces* Genome Database (SGD, http://www.yeastgenome.org/). The experiment-confirmed nucleosome positions of *S. cerevisiae* were taken from Lee et al. [7], where each of the 1,206,683 DNA fragments in the dataset constructed by these authors had been assigned a nucleosome formation score using a lasso model, with the high or low score to reflect its high or low propensity in forming nucleosome, respectively. The low score can also be interpreted as the propensity to inhibit the formation of nucleosome. Thus, the 5000 fragments of 150 bp with the highest scores were selected as the nucleosomal sequences and the 5000 fragments of 150 bp with the lowest scores were selected as the non-nucleosomal (or linker) sequences.

Also, as elaborated in [33], a benchmark dataset containing high similar samples would be lack of statistical representativeness. In the present study, to avoid the redundancy and reduce the homology bias, sequences with more than 80% sequence similarity were removed by using the CD-HIT program [35]. After such a screening procedure, the final benchmark dataset contains 3620 samples, of which 1880 are nucleosomal sequences belonging to the positive subset, and 1740 are linker sequences belonging to the negative subset. The detailed sequences thus obtained are given in Online Supporting Information S1.

2.2. Benchmark datasets for genomic sequences around functional sites

The experiment-confirmed 3600 DSB hotspots in endogenous chromosomal sequences were taken from Pan et al. [36]. The DNA sequence contexts from −500 bp to +500 bp flanking each of the DSB hotspot centers were extracted. The detailed sequences thus obtained are given in Online Supporting Information S2.

The 6637 SNP data for the *S. cerevisiae* were taken from Schacherer et al. [37]. The DNA sequence contexts from −500 bp to +500 bp flanking each of the SNP sites were extracted. The detailed sequences are given in Online Supporting Information S3.

The 322 experiment-confirmed ORIs were extracted from the OriDB database [38]. The DNA sequence contexts from −500 bp to +500 bp flanking each of the ORIs were extracted. The detailed sequences are given in Online Supporting Information S4.

2.3. Use deformation energy scores to represent DNA samples

Deformability of DNA is important for its superhelical folding in the nucleosome and can be reflected by the DNA step parameters, including three local angular parameters (twist, tilt, and roll) and three translational parameters (shift, slide, and rise). This suite of parameters has important roles in various biological processes, such as protein–DNA interactions, formation of chromosomes, and higher-order organization of the genetic material in a cell nucleus [21,39,40].

As demonstrated by Tolstorukov et al. [41], the deformation energy of the n-th segment generated by 150-bp sliding window along a DNA sequence of L in length can be defined by [41]

$$E_D(n) = \sum_{k=1}^{150} E_d(n,k),\ (1 \le n \le L-150) \tag{1}$$

where $E_d(n,k)$ is the deformation energy of the 2-tuple base pair at the k-th step. There are total ten possible 2-tuple base pairs in a DNA double stranded structure (dsDNA), as given by

$$\begin{cases} \overrightarrow{AA}/\overleftarrow{TT} & \overrightarrow{AC}/\overleftarrow{TG} & \overrightarrow{AG}/\overleftarrow{TC} & \overrightarrow{AT}/\overleftarrow{TA} & \overrightarrow{CA}/\overleftarrow{GT} \\ \overrightarrow{CC}/\overleftarrow{GG} & \overrightarrow{CG}/\overleftarrow{GC} & \overrightarrow{GA}/\overleftarrow{CT} & \overrightarrow{GC}/\overleftarrow{CG} & \overrightarrow{TA}/\overleftarrow{AT} \end{cases} \tag{2}$$

where the two characters right before the slash line (/) denote the 2-mer along one of its two single strands (ssDNA), while the two