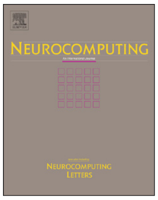
Contents lists available at ScienceDirect

Neurocomputing

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



An integrative analysis of nucleosome occupancy and positioning using diverse sequence dependent properties

Yun Lu [a,1], Yanglan Gan [b,1], Jihong Guan [a,2], Shuigeng Zhou [c,2]

[a] *Department of Computer Science and Technology, Tongji University, Shanghai, China*
[b] *School of Computer Science and Technology, Donghua University, Shanghai, China*
[c] *Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai, China*

ARTICLE INFO

Article history:
Received 9 August 2015
Received in revised form
17 October 2015
Accepted 22 November 2015
Available online 31 May 2016

Keywords:
Nucleosome positioning
Sequence-dependent properties
Structural properties

ABSTRACT

Nucleosome organization dictates eukaryotic DNA compaction and access, and plays a critical regulatory role in various eukaryotic genomic functions. It is well known that DNA sequence dependent histone octamer binding is important for nucleosome formation. However, many aspects of the phenomenon remain elusive. Here, we systematically analyze diverse sequence-dependent properties, consisting of multi-scale sequence compositional properties and twelve structural properties on the entire *Saccharomyces cerevisiae* genome. Based on comparative correlation analysis, the features that are highly correlated with nucleosome formation are identified and are integrated into a multilayer perceptron model for genome-wide nucleosome occupancy prediction. Furthermore, we predict nucleosome positions along genome based on the integrative profile of these properties. The results indicate the substantial effect on nucleosome formation and positioning along the entire genome.

1. Introduction

In eukaryotic organisms, 75–90% of genomes are packaged into nucleosomes, which are fundamental building units of condensed chromatin structure. Each nucleosome is composed of a histone octamer wrapped by around 147 bp DNA [1]. The formation of nucleosomes facilitates the storage and organization of long eukaryotic chromatin. Meanwhile, by modulating the accessibility of underlying DNA sequences, nucleosome formation directly or indirectly affects diverse cellular processes, such as transcription factor binding kinetics, gene splicing, DNA replication and DNA repair.

The recent availability of genome-wide nucleosome maps has provided researchers a great opportunity to understand the underlying mechanisms of nucleosome organization. Based on extensive research efforts, it is established that nucleosome formation is influenced by combined effects of multiple factors, including DNA sequence preference [2,4–7,14,29,30], DNA-binding proteins [8,9], nucleosome remodelers [10,11], the RNA polymerase II transcription machinery [12], DNA methylation [13], histone variant and histone post-translational modification [13,14]. Particularly, researchers have tried to describe nucleosome formation models based on DNA sequence signals.

Concerning how sequence dependent properties contribute to nucleosome organization, researchers have conducted extensive experimental and bioinformatic studies, which can be classified into two major classes. The first class has focused on local sequence compositional features [2,4–7,14,29,30]. A commonly cited nucleosome affinity feature is a ~10 bp periodicity of certain dinucleotides, which was first identified by the pioneer work of Trifonov [26], and has been confirmed in a variety of organisms including chicken [15], mouse [16], *Caenorhabditis elegans* [17] and *Homo sapiens* [18]. Furthermore, from all of tetranucleotides, the frequency of AAAA was identified as the most important sequence factor affecting nucleosome positioning [19]. Several studies further identified other sequence patterns associated with nucleosomal sequences [20,28]. G+C content was also exploited to explain the variation of nucleosome occupancy *in vitro* [5]. Subsequently, Computational methods based on such sequence compositional features have been proposed to predict nucleosome occupancy [21–23,31]. The other class is based on structural and mechanical features of DNA sequences. On the basis of lasso-based models, DNA structural features (*e.g.* tilts and propeller twists) were assigned the greatest significance to predict nucleosome occupancy [24]. Also, the DNA bending energy, DNA flexibility and intrinsic curvature were respectively adopted to characterize the histone–DNA interactions [25–27]. These structural and thermodynamic studies have shown the importance of structural features

[1] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
[2] To whom correspondence should be addressed.

http://dx.doi.org/10.1016/j.neucom.2015.11.107
0925-2312/© 2016 Elsevier B.V. All rights reserved.

of DNA sequences in the formation mechanism of nucleosomes. Nevertheless, among these diverse sequence-dependent features, which features are most correlated with nucleosome formation and have the highest predictive ability for nucleosome positioning need to be investigated.

In this paper, we systematically analyze two kinds of sequence-dependent properties, including multi-scale sequence compositional properties and structural properties on the entire *Saccharomyces cerevisiae* genome. Sequence compositional properties include all k-mer ($1 < k < 6$) features and G+C content. Structural properties consist of twelve structural features related to intrinsic flexibility, curvature and energy of DNA sequence. We calculate the correlation between each property and nucleosome occupancy. Based on comparative correlation analysis, the features that are highly significant in nucleosome formation are identified and are further integrated into a multilayer perceptron model for genome-wide nucleosome occupancy prediction. Furthermore, we predict nucleosome positions along genome based on the integrative profile of these properties. The results show that these selected properties perform well in predicting nucleosome organization and indicate the substantial effect of DNA sequence on binding affinity over the entire histone octamer.

2. Materials and methods

2.1. The overview of the proposed method

It is known that nucleosome distribution plays a very important role in various cellular phenomenon and is influenced by a multitude of factors. Here, to identify which features are most correlated with nucleosome formation, we developed a pipeline to evaluate the effect of diverse sequence-dependent properties and twelve kinds of structural properties on nucleosome positioning. As shown in Fig. 1, our pipeline starts from the given genomic sequence. Then it performs the following steps, including calculating the sequence-dependent properties, estimating the correlation between each property and nucleosome occupancy, and selecting the most correlated properties. Further, to improve the prediction accuracy of nucleosome occupancy and positioning, we integrated these selected properties into a multilayer perceptron to predict nucleosome occupancy, and to predict nucleosome positions by detecting peaks. In what follows, we will describe the details of the major steps in our pipeline.

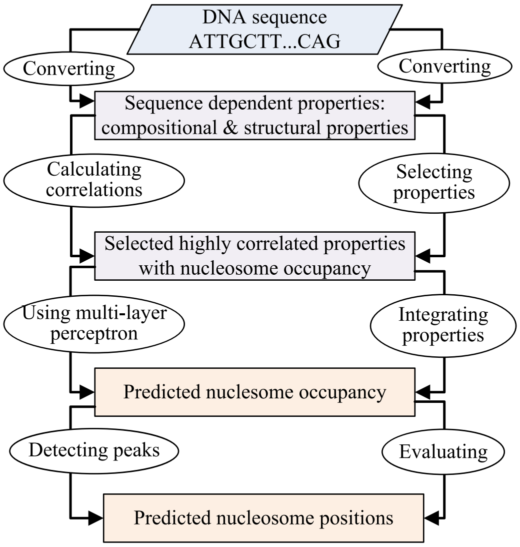

DNA sequence
ATTGCTT...CAG
Converting
Converting
Sequence dependent properties:
compositional & structural properties
Calculating correlations
Selecting properties
Selected highly correlated properties with nucleosome occupancy
Using multi-layer perceptron
Integrating properties
Predicted nuclesome occupancy
Detecting peaks
Evaluating
Predicted nucleosome positions

Fig. 1. The overview of the proposed method.

2.2. Calculating sequence compositional properties

To fully investigate the effect of the local sequence in nucleosome formation, we used multi-scale sequence patterns. We examined all k-mer as the sequence compositional properties, k ranging from 2 to 5. For a given k, there are at most 4^k k-mers in a DNA sequence. As each DNA fragment can be obtained from either strand of the DNA genome, a k-mer and its opposite complement k-mer can be regarded as a feature, and this process will reduce the size of vector by half, so $N(k) = 4^k/2$. Take $k=2$ as an example, $N(2) = 4^2/2 = 8$. That is, the number of 2-mers is 8. Similarly, there are 32 3-mers, 128 4-mers and 512 5-mers. For each individual k-mer, we counted its frequency appearing in a sliding window. Along the sequence, we constructed the frequency vector of the k-mer. As G+C content has been indicated predictive for nucleosome occupancy, this property is also added in the property set. We additionally calculated G+C content along the genome as in [5].

2.3. Calculating structural properties

As previous studies [27], we selected 12 structural properties to evaluate different structures and mechanics of DNA sequences. These properties are as follows: DNA denaturation (the ability of DNA to denature), Propeller twist (the angle of the two aromatic bases in a base pair), Protein-DNA twist (the twist angle of DNA deformed by protein), B-DNA twist (the mean twist angle in B-DNA), Duplex free energy (the thermodynamic energy content), Duplex disrupt energy (DNA duplex energy), Stacking energy (energy scale of dinucleotide base-stacking energy scale), Z-DNA (the ability to be covered from B- to Z-DNA), DNA-bending stiffness (the anisotropic flexibility of DNA), Bendability (the trinucleotide bendability), Protein-induced deformation (the ability to be changed by proteins) and Aphilicity (the free energy value for a transition from B- to A-DNA form).

The structural properties were calculated in two steps. For each property, we first converted each DNA sequence into a numerical sequence by replacing each dinucleotide or trinucleotide with the corresponding structural value. This transformation was based on experimentally determined structural models [27]. Then, we used a moving average to smooth the raw structural profile, with a step of 4 bp and a window size of 100 bp.

2.4. Predicting nucleosome occupancy based on multilayer perceptron model

By computing the correlation between the profile of each property and nucleosome occupancy, we selected the most correlated properties to predict nucleosome occupancy. Considering the good nonlinear mapping ability, we chose multilayer perceptron as the prediction model. Specifically, perceptron is a sort of multilayer feed-forward single direct propagation network model, which has been widely applied in function approximation. Multilayer perceptron has input layer, output layer and one or more layers of units between the input and output layers. We used the property matrix as the multilayer perceptron input and then built the model. The result of the output layer is the predicted nucleosome occupancy.

2.5. Predicting nucleosome positioning based on peak detection

Based on the multilayer perceptron model, we obtained the predicted nucleosome occupancy by our selected property set. As peak regions of the nucleosome occupancy often correspond to locations of nucleosomes, we further located the positions of nucleosomes by detecting the peaks of the predicted nucleosome occupancy. Here, we defined three variables in the peak detection