Mathematical and Computer Modelling 52 (2010) 1932–1939

Contents lists available at ScienceDirect

Mathematical and Computer Modelling

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

A simulation model for nucleosome distribution in the yeast genome based on integrated cross-platform positioning datasets

Jihua Feng *, Xianhua Dai, Zhiming Dai, Qian Xiang, Jiang Wang, Yangyang Deng, Caisheng He

School of Information Science and Technology, Sun Yat-Sen University, 135 West Xin'gang Road, Guangzhou, PR China

ARTICLE INFO

Article history:
Received 23 September 2009
Received in revised form 25 December 2009
Accepted 31 January 2010

Keywords:
Nucleosome
Yeast
Histone
Chromatin

ABSTRACT

In-depth analysis of six recent experimental nucleosome datasets in the yeast reveals broad disagreements between the final nucleosome positions detected by previous studies. Our results, which combine different types of data, suggest that cross-platform information, including discrepancy and consistency, reflects the mechanisms of nucleosome packaging in vivo more faithfully than individual studies. Furthermore, nucleosomes can be divided into two classes according to their stable and dynamic characteristics. By combining the nucleosome distribution information with a statistical positioning idea, we constructed a novel simulation model for nucleosome distributions in promoter regions.

1. Introduction

In eukaryotic genomes, DNA is compacted through multiple steps into a protein–DNA complex known as chromatin. The first level of compaction involves wrapping the long genomic DNA strands into arrays of particles called nucleosomes, each containing a 146 bp long stretch of DNA that is sharply bent and tightly wrapped in nearly two superhelical turns around an octameric core of histone [1]. Any DNA sequence can be packaged into a nucleosome; however, homeostatic histone concentrations ensure that only 70–90% of the DNA is wrapped in nucleosomes, with consecutive nucleosomes typically separated by 15–50 bp of unwrapped linker DNA [2]. In particular, the histone components, as well as additional chromatin proteins, can interact to form higher-order chromosomal structures. Thus, nucleosomes are critical to the organization and maintenance of chromatin, and their position and modification state can significantly influence genetic activities, such as the plasticity or control of gene expression. As a result, studies of nucleosome positions, determined by either experimental measurements or computational methods, continue to be an active field of research.

Six high-resolution genome-scale nucleosome positioning studies have recently been completed in *Saccharomyces cerevisiae* [3–8]. In these assays, either tiling arrays or direct sequencing technologies were used to map the positions of nucleosomes. However, it is clear from previous work that nucleosome positions are subtle and diffuse, which makes it difficult to distinguish their true position data from random noise in a single experiment. The dynamic changes and experimental errors that may be responsible for inconsistencies among these studies led us to develop a criterion to assess these studies effectively. In addition, inconsistent assignment of nucleosome positions, derived from different detection methods, highlights the need for careful and comprehensive comparison of these experimental datasets.

Here, we overcame the limitations of single-study analyses by pooling the nucleosome distribution information from six independent datasets so that valid relationships were reinforced and experimental noise was suppressed. Our motivation for comparison of these experimental data was twofold. First, none of these independent studies were exactly consistent with

* Corresponding author.
E-mail addresses: fengjihua@21cn.com (J. Feng), issdxh@mail.sysu.edu.cn (X. Dai), chinawinners@163.com (Z. Dai), xiangq@mail.sysu.edu.cn (Q. Xiang), Wrdzsu2003@163.com (J. Wang), vijor@163.com (Y. Deng), chason@tom.com (C. He).

0895-7177/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mcm.2010.03.043

Table 1
The genome-wide correlations between datasets.

Authors	Lee	Albert	Shivaswamy	Whitehouse	Mavrich	Field
Lee	1.00	0.18	0.22	0.15	0.26	0.14
Albert	0.18	1.00	0.17	0.10	0.31	0.16
Shivaswamy	0.22	0.17	1.00	0.11	0.22	0.27
Whitehouse	0.15	0.11	0.11	1.00	0.14	0.09
Mavrich	0.26	0.31	0.22	0.14	1.00	0.19
Field	0.14	0.16	0.27	0.09	0.19	1.00

each other in genome-wide nucleosome positions according to the correlation coefficients between nucleosome position datasets published so far, and we thus sought to utilize these high-resolution maps to construct a set of new, high-confidence reference maps. Our second motivation was to utilize these new reference maps to construct a mathematical model that can capture the characteristics of nucleosome distributions surrounding transcription start sites (TSSs).

2. Comparisons among cross-platform nucleosomal datasets in different genome scales

To objectively compare published nucleosome position data, we first collected all available basic information from these studies. All six experiments measured genome-wide nucleosome positions, but differed in their focus, emphasis and platforms. We found that these experiments can be divided into two groups according to the strains used and the experimental conditions. The normal group was defined as those studies that primarily made nucleosome preparations from BY4741 strains under normal conditions: Lee et al. [7], Albert et al. [8], Mavrich et al. [4] and Field et al. [5]. The conditional group consisted of the studies that used S288C strains and conducted experiments in the context of a physiological or genetic perturbation: Whitehouse et al. [6] and Shivaswamy et al. [3].

In order to evaluate the disagreement between datasets, we used binary sequences to reconstruct the nucleosome position datasets. Under this method, packaging DNA is represented by the logical symbol 1 and linker DNA is represented by 0 along the chromosome coordinates. Then, we calculated the Pearson correlation coefficients between the nucleosome positioning maps in different genomic regions (Table 1). According to positioning data, the average correlation coefficient is 0.21 across the entire genome and 0.29 in promoter regions. In the analysis of different gene segments, the most consistent regions for nucleosome positions are the TSSs (Fig. 1).

It is important to point out that to date there exist only two high-resolution genome-wide occupancy ratio datasets derived from microarrays [6,7]. Therefore, in a statistical sense, the analysis of nucleosomal occupancy data (i.e., raw data) is more uncertain than that of positional data (i.e., analyzed data). Based on the previously discussed factors, our study was restricted to the six recent nucleosomal positioning datasets.

3. Two distinct nucleosome positioning patterns

To decipher nucleosome positioning patterns from the cross-platform datasets, it is vital to determine the agreements and disagreements between these datasets. In a computationally intensive analysis, we identified these relationships by using the conjunction operation (logic AND) and the exclusive or operation (logic XOR). We chose these two methods for the following reasons: (1) the logical method is simple, rapid and accurate, which is very suitable for our restructuring binary data, and (2) most importantly, the logical algorithm has a good performance in large binary datasets; its computational cost is smaller than that of real number.

First, we performed a logical AND to extract common information from the six binary datasets. In principle, the intersection of the six binary datasets represents the stable nucleosome distribution among these datasets. Indeed, when we aligned and averaged the intersection signals with the TSS, an orderly organizational profile emerged from the promoter regions. Unlike the results of previous studies [4,5,7], however, we observed a natural order decay of stable nucleosome positioning peaks in the flanks of the nucleosome-free regions (NFRs) that depends on the distance from the TSSs. Despite the fact that our data were gathered from different platforms and under various conditions, all of the common combinatorial signals exhibit uniform phases and are distributed symmetrically around the TSS, both findings that have not been previously reported (Fig. 2). We refer to these in-phase signals as the stable nucleosome profile. This finding suggests that nucleosome organization in the genome may be more conserved than previously thought.

The observed discrepancies suggested that dynamic characteristics are inherent to the nucleosomes. In this study, the discrepant data is defined as follows. If and only if a binary value in a binary positioning dataset differs from that in any other binary datasets at the same loci do we consider it as a different datum and refer to the corresponding loci as the dynamic nucleosome-occupied domain (the dynamic nucleosome for short). The logic XOR was used to evaluate the differences between these datasets. We finally integrated those XOR results by using the arithmetic mean. Correspondingly, we refer to this integrated data as the dynamic nucleosome profile. Interestingly, similar to the observation regarding the stable nucleosome profile, we found that the integrated map of the dynamic nucleosome profiles is well organized around the TSS (Fig. 2).