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/© 2015 Elsevier Inc. All rights reserved.

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 $E_\mathrm{d}(n,k)$ 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

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

where the two characters right before the slash line (/) denote the 2-mer along one of its two single strands (ssDNA), while the two characters right after the slash line denote the 2-mer along the other strand; the two strands run opposite direction to each other as shown by the arrow above the nucleotide codes. The values of $E_{\rm d}(n,k)$ in Eq. (1) can be estimated by the following function based on the fluctuations of step parameters from their equilibrium values [42]:

$$E_d(n,k)=\frac{1}{2}\sum\nolimits_{i=1}^{6}\sum\nolimits_{j=1}^{6}f_{ij}(n,k)\Delta\theta_i\Delta\theta_j \tag{3}$$

where $i$ (or $j$) corresponds to the six base-pair parameters (twist, roll, tilt, shift, slide, and rise), $\Delta\theta_i=\theta_i-\theta_i^0$ is the deviation of the $i$-th step parameter $\theta_i$ from its equilibrium state $\theta_i^0$ caused by imposing the nucleosomal template characteristic of the corresponding 2-tuple base-pair, and $f_{ij}(n,k)$ is the element of the stiffness matrix $\mathbb{F}(m)$ associated with the $m$-th base-pair ($m=1,2,\cdots,10$) of Eq. (2) that can be formulated by a $6\times6$ matrix given below

$$\mathbb{F}(m)=\frac{1}{\rm k_BT}\,\mathbb{C}(m)$$
$$=\begin{bmatrix} f_{twist-twist} & f_{twist-tilt} & f_{twist-roll} & f_{twist-shift} & f_{twist-slide} & f_{twist-rise}\\ f_{twist-tilt} & f_{tilt-tilt} & f_{tilt-roll} & f_{tilt-shift} & f_{tilt-slide} & f_{tilt-rise}\\ f_{twist-roll} & f_{tilt-roll} & f_{roll-roll} & f_{roll-shift} & f_{roll-slide} & f_{roll-rise}\\ f_{twist-shift} & f_{tilt-shift} & f_{roll-shift} & f_{shift-shift} & f_{shift-slide} & f_{shift-rise}\\ f_{twist-slide} & f_{tilt-slide} & f_{roll-slide} & f_{shift-slide} & f_{slide-slide} & f_{slide-rise}\\ f_{twist-rise} & f_{tilt-rise} & f_{roll-rise} & f_{shift-rise} & f_{slide-rise} & f_{rise-rise} \end{bmatrix} \tag{4}$$

where $\rm k_B$ is the Boltzmann constant, T is the absolute temperature, and $\mathbb{C}(m)$ is the covariance matrix of the $m$-th base-pair step. More description about the covariance matrix and deformation energy can be found in [43,44] and [41,42], respectively.

Based on the 35 crystal structures of nucleosomes deposited in the Protein Data Bank (PDB), Yang and Yan [45] have deduced the deformation energy for each of the ten 2-tuple base-pairs in Eq. (2). Since the relative deformability of steps is independent of the value of $\rm k_BT$, similar to the treatment of Yang and Yan [45], we also set $\rm k_B$ and T to 1. Therefore, instead of being the real energy with the unit of joule, the deformation energy defined in Eq. (1) is actually a kind of energy score.

### 2.4. Operate prediction with discrimination function approach

Similar to the approach in identifying HIV protease cleavage sites [46] and predicting the enzyme's specificity [47], based on the deformation energy of a DNA segment as defined in Eq. (1), we can define a discrimination function given by [48]

$$\Delta(n)=E_{\rm D}(n)-\Re\ (1\le n\le L-150) \tag{5}$$

where $\Re$ is a modified factor [49] or cutoff threshold; its value is determined by optimizing outcome as will be mentioned later. Thus, we have

$$\begin{cases}\text{Nucleosomal sequence,} & \text{if } \Delta(n)>0\\ \text{Linker DNA,} & \text{otherwise}\end{cases} \tag{6}$$

the discrimination function approach has been successfully used to predict HIV protease cleavage sites (see, e.g., [46,48–52] and a review paper [53]).

### 2.5. A set of metrics for quantitative analysis

To facilitate the quantitative analysis, in this study we used a set of more intuitive and easier-to-understand metrics formulated with the symbols introduced by Chou [54] in studying signal peptide prediction. According to Chou's formulation, the sensitivity Sn, specificity Sp, overall accuracy Acc, and Matthews correlation coefficient MCC can be expressed as [55,56]

$$\begin{cases}\mathrm{Sn}=1-\dfrac{N_-^+}{N^+} & 0\le \mathrm{Sn}\le 1\\ \mathrm{Sp}=1-\dfrac{N_+^-}{N^-} & 0\le \mathrm{Sp}\le 1\\ \mathrm{Acc}=\Lambda=1-\dfrac{N_-^++N_+^-}{N^++N^-} & 0\le \mathrm{Acc}\le 1\\ \mathrm{MCC}=\dfrac{1-\left(\dfrac{N_-^++N_+^-}{N^++N^-}\right)}{\sqrt{\left(1+\dfrac{N_+^--N_-^+}{N^+}\right)\left(1+\dfrac{N_-^+-N_+^-}{N^-}\right)}} & -1\le \mathrm{MCC}\le 1\end{cases} \tag{7}$$

where $N^+$ is the total number of the positive samples or nucleosomal sequences investigated, while $N_-^+$ the number of nucleosomal sequences incorrectly predicted to be of linker sequences; $N^-$ the total number of the negative samples or linker sequences investigated, while $N_+^-$ the number of the linker sequences incorrectly predicted to be of nucleosomal sequences.

According to Eq. (7), the following are obvious. When $N_-^+=0$ meaning none of the nucleosomal sequences was incorrectly predicted belonging to linker sequences, we have the sensitivity $\mathrm{Sn}=1$. When $N_-^+=N^+$ meaning that all the nucleosomal sequences were incorrectly predicted belonging to linker sequences, we have the sensitivity $\mathrm{Sn}=0$. Likewise, when $N_+^-=0$ meaning none of the linker sequences was mispredicted, we have the specificity $\mathrm{Sp}=1$; whereas $N_+^-=N^-$ meaning that all the linker sequences were incorrectly predicted as nucleosomal sequences, we have the specificity $\mathrm{Sp}=0$. When $N_-^+=N_+^-=0$ meaning that none of nucleosomal sequences in the positive dataset and none of the linker sequences in the negative dataset were incorrectly predicted, we have the overall accuracy $\mathrm{Acc}=1$ and $\mathrm{MCC}=1$; when $N_-^+=N^+$ and $N_+^-=N^-$ meaning that all the nucleosomal sequences in the positive dataset and all the linker sequences in the negative dataset were incorrectly predicted, we have the overall accuracy $\mathrm{Acc}=0$ and $\mathrm{MCC}=-1$; whereas when $N_-^+=N^+/2$ and $N_+^-=N^-/2$ we have $\mathrm{Acc}=0.5$ and $\mathrm{MCC}=0$ meaning no better than random guess. As we can see from the above discussion, it would make the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient crystal clear by using the metrics formulated in Eq. (7) rather than the conventional formulation, particularly for the meaning of MCC, as concurred by a series of recent publications [6,22–31,57–65].

It should be pointed out, however, the set of equations defined in Eq. (7) is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in system biology [66–68] and system medicine [69], a completely different set of metrics is needed as elucidated in [70].

## 3. Results and discussion

### 3.1. Discrimination of nucleosomal and linker sequences

To analyze the correlation of the nucleosomal sequences with their deformation energies, a comparison was made between the nucleosomal and linker sequences based on their deformation energies calculated with Eqs. (1)–(4).

Quite interestingly, it was observed that the deformation energy of nucleosomal sequences was remarkably higher than that of the linker sequences as shown in Fig. 1, implying that DNA sequences need more deformation energy to wrap around the histone octamers to form nucleosomes.

Accordingly, it is rational to use Eqs. (5)–(6) to discriminate the nucleosomal sequences from the linker sequences along a dsDNA. It was found by the 10-fold cross-validation on the benchmark

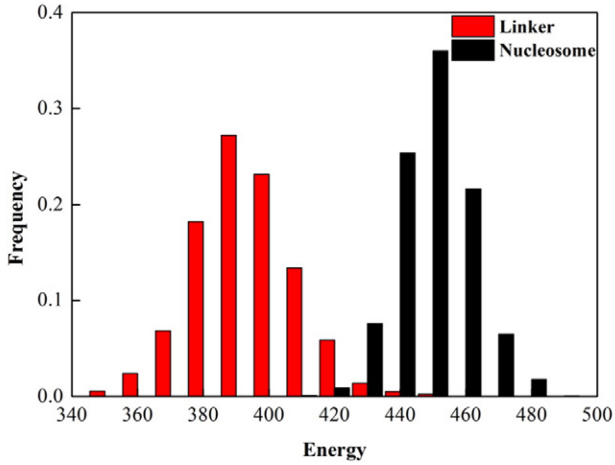

**Fig. 1.** Illustration to show the frequency spectrums of the deformation energy (cf. Eq. (1)) for the nucleosomal segments (black) and linker segments (red), respectively. The deformation energies of the former are remarkably higher than those of the latter. See the main text for further explanation.

dataset (Online Supporting Information S1) that, when the threshold $\mathfrak{R}$ of Eq. (5) was equal to 426.35, the success rates (Eq. (7)) in identifying nucleosomal sequences reached their peaks; i.e., when $\mathfrak{R} = 4$ 26.35, we have

$$\begin{cases} Sn = 0.982 \\ \mathrm{Sp} = 0.980 \\ \mathrm{Acc} = 0.981 \\ MCC = 0.963 \end{cases} \tag{8}$$

The above results indicate that the accuracy is very high regardless which one of the four metrics in Eq. (7) is used for the performance measurement.

To further show the performance of our model, we also compared the performance of our model with that of **iNuc-PhysChem** [15]. The predictive results obtained by **iNuc-PhysChem** with the 10-fold cross-validation on the same benchmark dataset (Online Supporting Information S1), was given below

$$\begin{cases} Sn = 0.972 \\ \mathrm{Sp} = 0.943 \\ \mathrm{Acc} = 0.967 \\ MCC = 0.936 \end{cases} \tag{9}$$

As we can see from Eqs. (8)–(9), the proposed biophysical model outperformed **iNuc-PhysChem** [15] in all the four metrics, indicating that the new model based on the deformation energy may become a useful tool in identifying nucleosomal sequences, or at the very least play a complementary role to the existing methods in this area.

Furthermore, the proposed model was also validated on the *S. cerevisiae* genome through a comparison between the nucleosome positions predicted by Eq. (6) and those determined by experiments [71]. The comparison was carried out in a 20 k-bp genomic region on chromosome 14 of the *S. cerevisiae* genome. As can be seen from Fig. 2, the profile of nucleosome occupancy predicted by the deformation energy approach is notably similar with the experimental maps of nucleosome organization, demonstrating once again that the deformation energy is indeed a very important factor for nucleosome positioning prediction.

Particularly, it is instructive to point out that, unlike most machine-learning predictors [12–16], our model is more like *ab initio* one since it basically needn't go through the tedious training process, as will be further manifested by using it to study the distribution of nucleosomes around the following DNA functional sites.

### *3.2. Nucleosome positioning around double strand break (DSB) hotspots*

Meiotic recombination is an important biological process. As a main driving force for evolution, recombination provides natural new combinations of genetic variations [55,72]. Recombination in meiosis occurs via a developmentally programmed pathway that forms numerous DNA double-strand breaks (DSBs) [73]. The regions where DSBs

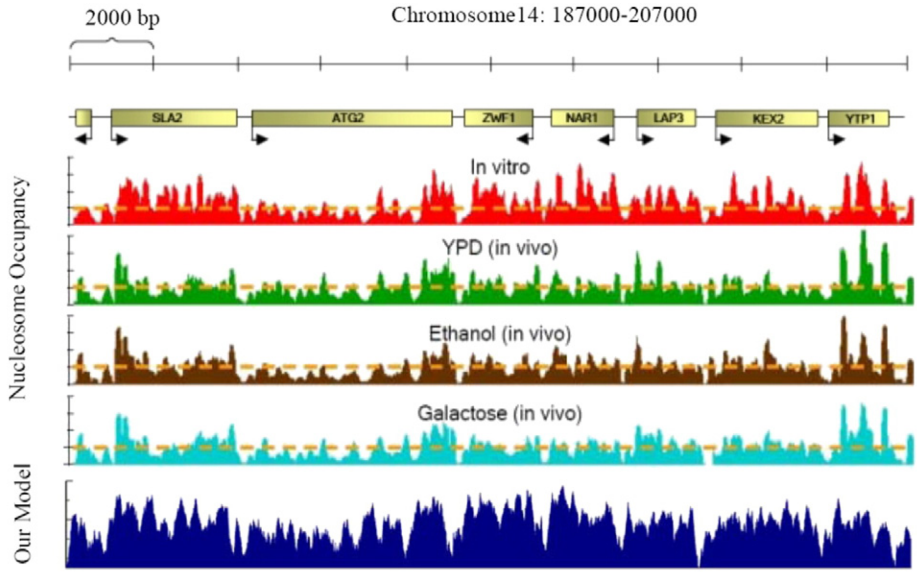

**Fig. 2.** Illustration to show the nucleosome occupancy profile predicted by the energy deformation model of Eq. (3). For facilitating comparison, the corresponding profiles by experiments are also shown. According to the top down order, they are the genomic position on the chromosome, the experimental map in vitro, as well as the experimental maps in vivo for YPD, ethanol, and galactose [71], respectively. Depicted in the lowest panel is the nucleosome occupancy profile by the proposed model. See the main text for more explanation.

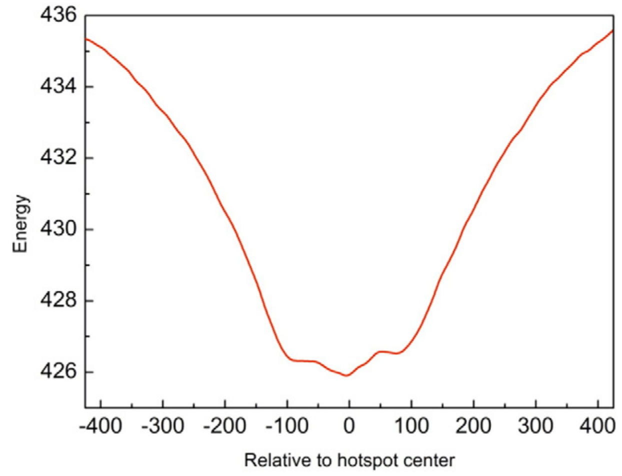

**Fig. 3.** The statistical profile of deformation energy around the DSB hotspot sites. The horizontal axis represents the genomic position ranging from −500 bp to +500 bp with the DSB hotspot site at the center or 0. The vertical axis represents DNA deformation energy. See the main text for more explanation.

form preferentially [36] are called DSB hotspots, which are usually with the presence of open chromatin structure, certain histone modifications, and sequence-specific transcription factors being bound at some loci [74]. But the detailed mechanism of how these factors affect the formation of the DSB hotspots is not quite clear yet.

To address this problem, we analyzed the nucleosome occupancy for *S. cerevisiae* genome via the following procedure. By sliding a 150-bp window (with 1-bp step at a time) along each of the 3600 samples in Online Supporting Information S2, we extracted $1000–150 + 1 = 851$ segments, followed by using Eq. (5) to calculate their deformation energies, respectively. The results thus obtained are given in Fig. 3, from which we can see that that DSB hotsopts occur nearly exclusively in nucleosome-depleted regions, fully consistent with the previous findings that nucleosome-depleted regions provide the opportunities for DSB formation [36]. The loose chromatin structure and the openness of the promixal regions surrounding DSB hotspots as shown in Fig. 3 may facilitate the binding of the topoisomerase-related Spo11 protein, which plays a predominant role in initiating meiotic recombinations [36].

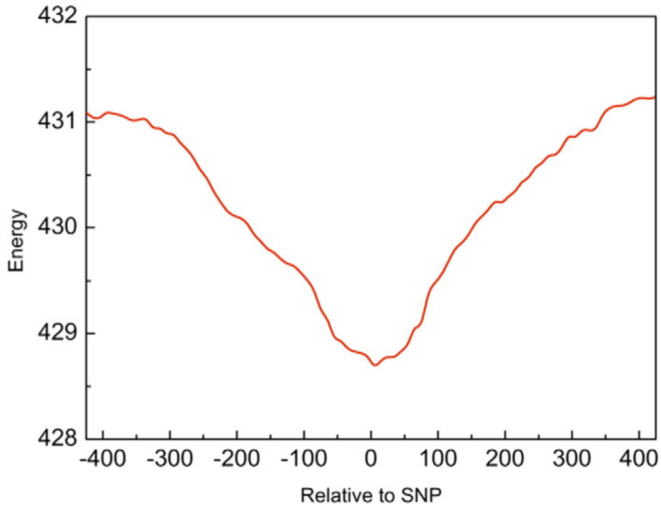

**Fig. 4.** DNA deformation energy profile around single nucleotide polymorphism (SNP) site. DNA deformation energy was smoothed using a 150 bp sliding window with 1 bp step. The horizontal axis represents the genomic position, which ranges from −500 bp to +500 bp relative to SNP site (denoted as 0). The vertical axis represents DNA deformation energy.

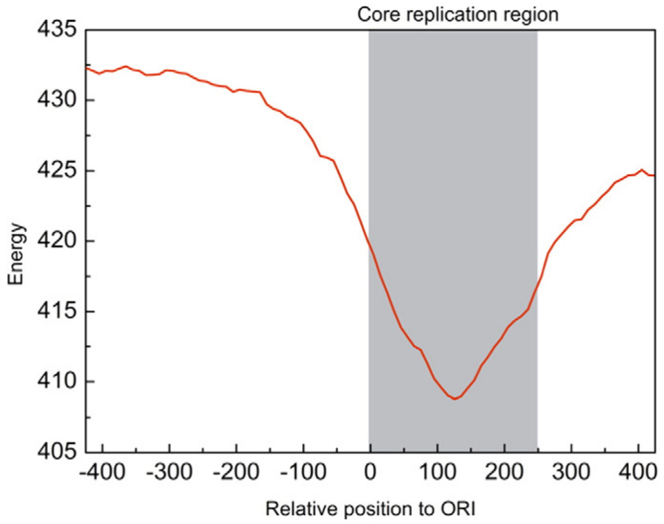

**Fig. 5.** DNA deformation energy profile around origin of replication (ORI). DNA deformation energy was smoothed using a 150 bp sliding window with 1 bp step. The horizontal axis represents the genomic position, which ranges from −500 bp to +500 bp relative to ORI (denoted as 0). The vertical axis represents DNA deformation energy.

### *3.3. Nucleosome positioning around single nucleotide polymorphism (SNP)*

It has been demonstrated that SNP sites are generally located at nucleosome-depleted regions in the human genome [75]. Yet the sequence patterns of nucleosomes near to the SNP sites in the *S. cerevisiae* genome remain to be clarified. Using the same approach as described in the last section to analyze the 6637 samples given in Online Supporting Information S3, we obtained the corresponding statistical (or average deformation energy) profile for the SNP sites, as shown in Fig. 4. It can be seen from the figure that nucleosomes are also depleted in the region near to the SNP site, suggesting a negative correlation between nucleosome occupancy and genetic variation. This is because the nucleosomal sequences are evolutionarily more conserved than the linker sequences [76], so as to protect them from mutations. That is why the SNP tends to locate at the nucleosome depleted regions.

### *3.4. Nucleosome positioning around origin of replication (ORI)*

DNA replication is thought of a most highly regulated process as far as the interactions between regulatory proteins and DNA sequences are concerned. The initiation of DNA replication is also regulated by chromatin structure. To in-depth study this problem, we used the same approach as described in Section 3.2 once again to calculate the deformation energy for the 322 experiment-confirmed ORIs in Online Supporting Information S4. The results thus obtained were used to depict the average deformation energy profile for the nucleosome positioning pattern in the vicinity of ORI, as shown in Fig. 5. It can be observed from the figure that the core replication region (0–+250 bp) is flanked by two well-positioned nucleosomes at its both sides, which is quite consistent with the previous experimental reports by the previous investigators [5,77] that nucleosomes are depleted in the core replication region but are well positioned in the flanking regions of ORI. The low nucleosome density in the core replication region suggests an open chromatin structure that may help the binding of recognition complex and facilitates origin firing [78].

## 4. Conclusions

Our model has demonstrated that the deformation energy plays a key role in discriminating nucleosomal sequences from DNA linker sequences.

Unlike most machine-learning predictors, our model is more like *ab initio* one as reflected by the fact that, although it has basically not undergone the tedious training process, it can be directly and successfully used to predict the distribution of nucleosomes around some DNA functional sites, such as DSB, SNP, and ORI.

It is anticipated that the deformation energy model as presented in this paper will stimulate a series of new and more powerful methods for predicting nucleosome positioning.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2015.12.005.

## Acknowledgments

The authors would like to thank the two anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper. This work was supported by National Nature Scientific Foundation of China (No. 61202256), Nature Scientific Foundation of Hebei Province (No.C2013209105), Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

## References

[1] K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, T.J. Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution, Nature 389 (1997) 251–260.

[2] T. Yasuda, K. Sugasawa, Y. Shimizu, S. Iwai, T. Shiomi, F. Hanaoka, Nucleosomal structure of undamaged DNA regions suppresses the non-specific DNA binding of the XPC complex, DNA Repair (Amst) 4 (2005) 389–395.

[3] S. Schwartz, E. Meshorer, G. Ast, Chromatin organization marks exon–intron structure, Nat. Struct. Mol. Biol. 16 (2009) 990–995.

[4] W. Chen, L. Luo, L. Zhang, The organization of nucleosomes around splice sites, Nucleic Acids Res. 38 (2010) 2788–2798.

[5] N.M. Berbenetz, C. Nislow, G.W. Brown, Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure, PLoS Genet. 6 (2010).

[6] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biomed Res. Int. (BMRI) 2014 (2014) 623149.

[7] W. Lee, D. Tillo, N. Bray, R.H. Morse, R.W. Davis, T.R. Hughes, C. Nislow, A high-resolution atlas of nucleosome occupancy in yeast, Nat. Genet. 39 (2007) 1235–1244.

[8] D.E. Schones, K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, K. Zhao, Dynamic regulation of nucleosome positioning in the human genome, Cell 132 (2008) 887–898.

[9] G.C. Yuan, J.S. Liu, Genomic sequence is highly predictive of local nucleosome depletion, PLoS Comput. Biol. 4 (2008), e13.

[10] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J.A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, S.M. Johnson, A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning, Genome Res. 18 (2008) 1051–1063.

[11] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang, J. Widom, A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.

[12] I.P. Ioshikhes, I. Albert, S.J. Zanton, B.F. Pugh, Nucleosome positions predicted through comparative genomics, Nat. Genet. 38 (2006) 1210–1215.

[13] S. Gupta, J. Dennis, R.E. Thurman, R. Kingston, J.A. Stamatoyannopoulos, W.S. Noble, Predicting human nucleosome occupancy from primary sequence, PLoS Comput. Biol. 4 (2008), e1000134.

[14] H.E. Peckham, R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W.S. Noble, K. Struhl, Z. Weng, Nucleosome positioning signals in genomic DNA, Genome Res. 17 (2007) 1170–1177.

[15] W. Chen, H. Lin, P.M. Feng, C. Ding, iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties, PLoS One 7 (2012), e47843.

[16] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics (2014) 1522–1529.

[17] V. Miele, C. Vaillant, Y. d'Aubenton-Carafa, C. Thermes, T. Grange, DNA physical properties determine nucleosome occupancy from yeast to fly, Nucleic Acids Res. 36 (2008) 3746–3756.

[18] A.V. Morozov, K. Fortney, D.A. Gaykalova, V.M. Studitsky, J. Widom, E.D. Siggia, Using DNA mechanics to predict in vitro nucleosome positions and formation energies, Nucleic Acids Res. 37 (2009) 4707–4722.

[19] T. Nozaki, N. Yachie, R. Ogawa, R. Saito, M. Tomita, Computational analysis suggests a highly bendable, fragile structure for nucleosomal DNA, Gene 476 (2011) 10–14.

[20] Q. Wu, W. Zhou, J. Wang, H. Yan, Correlation between the flexibility and periodic dinucleotide patterns in yeast nucleosomal DNA sequences, J. Theor. Biol. 284 (2011) 92–98.

[21] W. Chen, H. Lin, P.M. Feng, DNA physical parameters modulate nucleosome positioning in the *Saccharomyces cerevisiae* genome, Curr. Bioinforma. 9 (2014) 188–193.

[22] W. Chen, P. Feng, H. Ding, iRNA-Methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33 (also, Data in Brief, 2015, 5: 376-378).

[23] J. Jia, Z. Liu, X. Xiao, iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, J. Theor. Biol. 377 (2015) 47–56.

[24] B. Liu, L. Fang, F. Liu, X. Wang, Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS One 10 (2015), e0121501.

[25] B. Liu, L. Fang, F. Liu, iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach, J. Biomol. Struct. Dyn. (2015), http://dx.doi.org/10.1080/07391102.2015.1014422.

[26] Z. Liu, X. Xiao, W.R. Qiu, iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77 (also, Data in Brief, 2015, 4: 87-89).

[27] W.R. Qiu, X. Xiao, W.Z. Lin, iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model, J. Biomol. Struct. Dyn. 33 (2015) 1731–1742.

[28] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. 33 (2015) 2221–2233.

[29] R. Xu, J. Zhou, B. Liu, Y.A. He, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, J. Biomol. Struct. Dyn. 33 (2015) 1720–1730.

[30] J. Jia, Z. Liu, X. Xiao, B. Liu, Identification of protein–protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition, J. Biomol. Struct. Dyn. (2015), http://dx.doi.org/10.1080/07391102.2015.1095116.

[31] B. Liu, L. Fang, S. Wang, X. Wang, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, J. Theor. Biol. 385 (2015) 153–159.

[32] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2 L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics (2015)http://dx.doi.org/10.1093/bioinformatics/btv604.

[33] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review), J. Theor. Biol. 273 (2011) 236–247.

[34] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.

[35] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[36] J. Pan, M. Sasaki, R. Kniewel, H. Murakami, H.G. Blitzblau, S.E. Tischfield, X. Zhu, M.J. Neale, M. Jasin, N.D. Socci, A. Hochwagen, S. Keeney, A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation, Cell 144 (2011) 719–731.

[37] J. Schacherer, D.M. Ruderfer, D. Gresham, K. Dolinski, D. Botstein, L. Kruglyak, Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains, PLoS One 2 (2007).

[38] C.A. Nieduszynski, S. Hiraga, P. Ak, C.J. Benham, A.D. Donaldson, OriDB: a DNA replication origin database, Nucleic Acids Res. 35 (2007) D40–D46.

[39] T. Abeel, Y. Saeys, E. Bonnet, P. Rouze, Y. Van de Peer, Generic eukaryotic core promoter prediction using structural features of DNA, Genome Res. 18 (2008) 310–323.

[40] J.R. Goni, C. Fenollosa, A. Perez, D. Torrents, M. Orozco, DNAlive: a tool for the physical analysis of DNA at the genomic scale, Bioinformatics 24 (2008) 1731–1732.

[41] M.Y. Tolstorukov, A.C. Colasanti, D. McCandlish, W.K. Olson, V.B. Zhurkin, A novel 'roll-and-slide' mechanism of DNA folding in chromatin. Implications for nucleosome positioning, J. Mol. Biol. 371 (2007) 725–738.

[42] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 11163–11168.

[43] K.C. Chou, A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins: structure, Funct. Genet. 21 (1995) 319–344.

[44] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[45] X. Yang, H. Yan, Statistical analysis of conformational properties of periodic dinucleotide steps in nucleosomes, J. Biomed. Sci. Eng. 3 (2010) 9.

[46] C.T. Zhang, F.J. Kezdy, A vector approach to predicting HIV protease cleavage sites in proteins, Proteins Struct. Funct. Genet. 16 (1993) 195–204.

[47] K.C. Chou, A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase, Protein Sci. 4 (1995) 1365–1383.

[48] K.C. Chou, A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins, J. Biol. Chem. 268 (1993) 16938–16948.

[49] A.L. Tomasselli, I.M. Reardon, R.L. Heinrikson, Predicting HIV protease cleavage sites in proteins by a discriminant function method, Proteins Struct. Funct. Genet. 24 (1996) 51–72.

[50] C.T. Zhang, An alternate-subsite-coupled model for predicting HIV protease cleavage sites in proteins, Protein Eng. 7 (1993) 65–73.

[51] J.J. Chou, Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach, J. Protein Chem. 12 (1993) 291–302.
[52] C.T. Zhang, Studies on the specificity of HIV protease: an application of Markov chain theory, J. Protein Chem. 12 (1993) 709–724.
[53] K.C. Chou, Review: prediction of human immunodeficiency virus protease cleavage sites in proteins, Anal. Biochem. 233 (1996) 1–14.
[54] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[55] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[56] Y. Xu, J. Ding, L.Y. Wu, iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, PLoS One 8 (2013) e55844.
[57] W. Chen, P.M. Feng, E.Z. Deng, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.
[58] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels, BioMed Res. Int. (BMRI) 2014 (2014) 286419.
[59] Y.N. Fan, X. Xiao, J.L. Min, iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking, Int. J. Mol. Sci. (IJMS) 15 (2014) 4915–4937.
[60] H. Lin, E.Z. Deng, H. Ding, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.
[61] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS One 9 (2014), e106691.
[62] W.R. Qiu, X. Xiao, W.Z. Lin, iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach, BioMed Res. Int. (BMRI) 2014 (2014) 947416.
[63] Y. Xu, X. Wen, X.J. Shao, N.Y. Deng, iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, Int. J. Mol. Sci. (IJMS) 15 (2014) 7594–7610.
[64] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS One 9 (2014), e105018.
[65] W. Chen, H. Lin, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. BioSyst. 11 (2015) 2620–2634.
[66] Z.C. Wu, X. Xiao, iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. BioSyst. 8 (2012) 629–641.
[67] W.Z. Lin, J.A. Fang, X. Xiao, iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, Mol. BioSyst. 9 (2013) 634–644.
[68] X. Xiao, Z.C. Wu, iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, J. Theor. Biol. 284 (2011) 42–51.
[69] X. Xiao, P. Wang, W.Z. Lin, iAMP-2 L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal. Biochem. 436 (2013) 168–177.
[70] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. BioSyst. 9 (2013) 1092–1100.
[71] N. Kaplan, I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, E.M. LeProust, T.R. Hughes, J.D. Lieb, J. Widom, E. Segal, Nature 458 (2009) 362–366.
[72] W.R. Qiu, X. Xiao, IRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. (IJMS) 15 (2014) 1746–1766.
[73] S. Keeney, Spo11 and the formation of DNA double-strand breaks in meiosis, Genome Dyn. Stab. 2 (2008) 81–123.
[74] T.D. Petes, Meiotic recombination hot spots and cold spots, Nat. Rev. Genet. 2 (2001) 360–369.
[75] H. Liu, X. Duan, S. Yu, X. Sun, Analysis of nucleosome positioning determined by DNA helix curvature in the human genome, BMC Genomics 12 (2011) 72.
[76] M.Y. Tolstorukov, N. Volfovsky, R.M. Stephens, P.J. Park, Impact of chromatin structure on sequence variability in the human genome, Nat. Struct. Mol. Biol. 18 (2011) 510–515.
[77] S. Yin, W. Deng, L. Hu, X. Kong, The impact of nucleosome positioning on the organization of replication origins in eukaryotes, Biochem. Biophys. Res. Commun. 385 (2009) 363–368.
[78] R. Lombrana, R. Almeida, I. Revuelta, S. Madeira, G. Herranz, N. Saiz, U. Bastolla, M. Gomez, High-resolution analysis of DNA synthesis start sites and nucleosome architecture at efficient mammalian replication origins, EMBO J. 32 (2013) 2631–2644.