FEBS Letters 586 (2012) 934–938

journal homepage: www.FEBSLetters.org

Prediction of replication origins by calculating DNA structural properties

Wei Chen [a,*], Pengmian Feng [b], Hao Lin [c,*]

[a] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China
[b] School of Public Health, Hebei United University, Tangshan 063000, China
[c] Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE INFO

Article history:
Received 28 November 2011
Revised 21 February 2012
Accepted 21 February 2012
Available online 28 February 2012

Edited by Gianni Cesareni

Keywords:
Origin of replication
Bendability
Cleavage intensity
Support vector machine

ABSTRACT

In this study, we introduced two DNA structural characteristics, namely, bendability and hydroxyl radical cleavage intensity to analyze origin of replication (ORI) in the *Saccharomyces cerevisiae* genome. We found that both DNA bendability and cleavage intensity in core replication regions were significantly lower than in the linker regions. By using these two DNA structural characteristics, we developed a computational model for ORI prediction and evaluated the model in a benchmark dataset. The predictive performance of the jackknife cross-validation indicates that DNA bendability and cleavage intensity have the ability to describe core replication regions and our model is effective in ORI prediction.

1. Introduction

Replicon hypothesis is a useful description of replication in prokaryotes, but the situation in eukaryotes is more complex. In eukaryotes, initiation of DNA replication occurs at specific genomic loci called origin of replication (ORI) [1–3]. Accurate replication of the whole genome in every cell division is essential for maintaining genomic stability. Thus, determining the location of ORI is important for understanding how such origins are specified and utilized in various developmental situations such as the mitotic and meiotic cell cycles.

As an important eukaryotic model organism, the budding yeast has the best characterized replication origins. Budding yeast replication origins (called autonomously replicating sequences or ARS elements) are almost exclusively intergenic and consist of approximately 200 bp sequences that can be divided into A and B domains [4]. The A domain contains an essential ARS consensus sequence (ACS), which is essential for binding of the origin recognition complex (ORC) [5–7]. The B domain tends to be helically unstable and additionally contains a number of short sequence motifs that contribute to origin activity, such as B1, B2 and B3 elements. The B1 element, found in every ARS, is adjacent to the ACS and is part of the ORC binding site [8,9]. The B2 element presents in most but not all ARSs and frequently overlaps with DNA unwinding elements [10,11]. The B3 element is a transcription factor binding site found in some ARSs that influences nucleosome positioning [12].

Despite their common function of domains from different replication origins, no consensus motif with predictive value has been found yet. Therefore, it is impossible to accurately identify a real ORI from the vast intergenic region using sequence information. Thus, several resource-intensive techniques have been proposed for ORI identification, such as ARS assays, two-dimensional gels and microarray-based detection of origin activity. However, none of these experimental methods is practical on genomic scale.

In this study, we conducted a computational analysis of two structural characteristics, namely, DNA bendability and cleavage intensity around ORIs in the *Saccharomyces cerevisiae* genome. We found that both DNA bendability and cleavage intensity in core replication regions were significantly lower than those in surrounding regions. Based on this finding, we developed a support vector machine (SVM) based model for ORI prediction and achieved a high predictive accuracy under the jackknife cross-validation.

2. Materials and methods

2.1. Dataset

As September of 2009, the OriDB database [13] collected 732 *S. cerevisiae* ORIs. Each origin site in the database is assigned a status

* Corresponding authors. Address. Department of Physics, Hebei United University, Tangshan 063000, China. Fax: +86 315 3725715.
E-mail addresses: chenwei_imu@yahoo.com.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

0014-5793/$36.00
doi:10.1016/j.febslet.2012.02.034

(confirmed, likely, or dubious) which expresses the level of confidence that the site genuinely corresponds to an origin. Origin sites labeled "confirmed" are those that have been experimentally validated. In order to prepare a high quality dataset, we picked out the 322 experimentally confirmed ORIs from the OriDB. According to the information of the *Saccharomyces* Genome Database (http://www.yeastgenome.org/), 1000 bp long sequence (500 bp upstream and 500 bp downstream of ORI, respectively) was extracted from each of the 322 ORIs.

In order to train an ORI prediction model, we constructed a benchmark dataset that containing 322 positive samples (core replication region ranging from 0 bp to +250 bp [13]) and 966 negative samples (regions ranging from −500 bp to −250 bp, −250 bp to 0 bp and +250 bp to + 500 bp) extracted from the 322 experimentally confirmed ORIs. No two sequences in the benchmark dataset share more than 28% sequence identity.

2.2. DNA bendability

DNA bendability reflects the non-parallel tendency of consecutive base pairs in a DNA sequence and has been successfully applied to promoter prediction [14,15]. This index is defined as the summation of bendability parameter profile associated with every trinucleotide in a given DNA sequence. For example, DNA bendability of sequence AGCTA is 0.124 (0.017[AGC] + 0.017[GCT] + 0.090[CTA] = 0.124). Detailed descriptions on DNA bendability and the bendability parameters for each trinucleotide pattern can be found in the previous literature [16].

2.3. Cleavage intensity

Cleavage intensity indicates the likelihood of DNA cleavage by hydroxyl radicals and provides a map of local variation in the shape of DNA backbone [17–19]. It can be calculated from parameters for a set of tetranucleotide patterns in a given DNA sequence. The parameters of tetranucleotides were derived by Greenbaum et al. from experiments in which DNA sequences were exposed to hydroxyl radicals [17]. However, this type of hydroxyl radical cleavage pattern considers properties of a single DNA strand of the double helix, and therefore could not measure the minor groove width, which depends on both DNA strands and is an important recognition element for protein binding [20–22]. Recently, Bishop et al. [23] developed the ORChID2 algorithm (http://dna.bu.edu/orchid/) to predict cleavage intensity by incorporating hydroxyl radical cleavage information from both strands of the DNA duplex. The algorithm can calculate the cleavage intensity for each nucleotide in a DNA sequence.

2.4. Support vector machine

Support vector machine (SVM) is an effective method for supervised pattern recognition and has been widely used in the realm of bioinformatics [24–29]. The basic idea of SVM is to transform the data into a high dimensional feature space and then determine the optimal separating hyperplane. In this work, the SVM implementation was based on the freely available package LibSVM 2.81 written by Chang and Lin [30]. The radial basis kernel function (RBF) was used to obtain the best classification hyperplane. The regularization parameter C and the kernel width parameter γ were optimized using a grid search approach.

2.5. Performance evaluation

The performance of the prediction model was evaluated using sensitivity (*Sn*), specificity (*Sp*) and accuracy (*Acc*), which are expressed as follows:

$$Sn = \frac{TP}{TP + FN} \quad (1)$$

$$Sp = \frac{TN}{TN + FP} \quad (2)$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \quad (3)$$

TP, *TN*, *FP* and *FN* represent the number of the correctly recognized ORI, the number of the correctly recognized non-ORI, the number of non-ORI recognized as ORI and the number of ORI recognized as non-ORI sequences, respectively. Meanwhile, the quality of a classifier can be evaluated by measuring the area under the receiver operating characteristic curve (auROC). The value of auROC score ranges from 0 to 1, with a score of 0.5 corresponding to a random guess and a score of 1.0 indicating perfect separation.

3. Results and discussion

3.1. DNA structural profiles in replication origin regions

To investigate the structural properties in replication origin regions, we firstly analyzed the bendability of DNA sequences surrounding ORI. We calculated DNA bendability for every trinucleotide in genomic regions from −500 bp to +500 bp relative to ORI and plotted the average DNA bendability profiles using a sliding window approach with a window size of 50 bp and a step size of 1 bp (Fig. 1). We found that bendability scores within core replication regions (0 to +250 bp) were significantly lower than those within surrounding regions ($P < 3.2e\text{-}16$, Mann–Whitney U-test) and this structural difference is independent of the window size for smoothing (Supplementary Fig. S1).

We next calculated DNA cleavage intensity for all the DNA positions from −500 bp to +500 bp relative to ORI using the ORChID2 algorithm [23]. Using a 50 bp sliding window with 1 bp offset, the average DNA cleavage intensity score was plotted in Fig. 2. We found that cleavage intensity in the core replication regions was statistically lower than did the linker regions on both sides ($P < 4.3e\text{-}11$, Mann–Whitney U-test) and the intensity of the observed cleavage intensity signal is also independent of the window size for smoothing (Supplementary Fig. S2).

Interestingly, both bendability and cleavage intensity demonstrated a strong signal at approximately 180 bp downstream of ORI. To demonstrate whether the signal corresponds to some

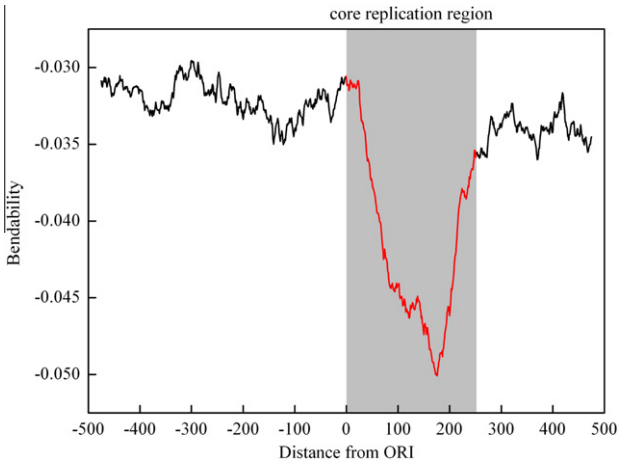

Fig. 1. DNA bendability profile in core replication region and surrounding regions. DNA bendability was smoothed using a 50 bp sliding window with 1 bp step. The horizontal axis represents the nucleotide position, which ranges from −500 bp to +500 bp relative to ORI (denoted as 0). The vertical axis represents DNA bendability score.

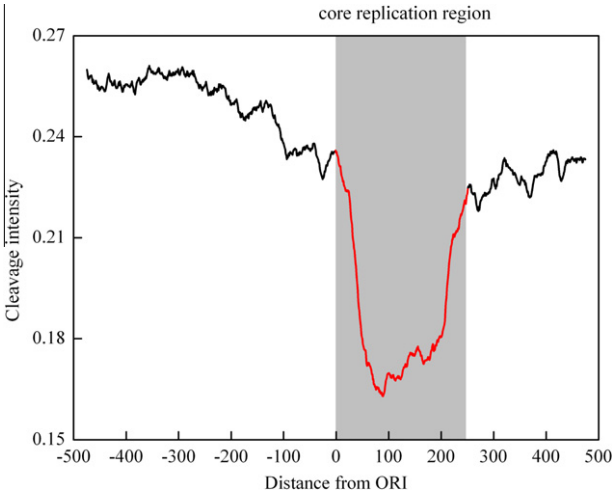

Fig. 2. DNA cleavage intensity profile in core replication region and surrounding regions. DNA cleavage intensity was smoothed using a 50 bp sliding window with 1 bp step. The horizontal axis represents the nucleotide position, which ranges from −500 bp to +500 bp relative to ORI (denoted as 0). The vertical axis represents DNA cleavage intensity score.

known motifs, we analyzed the region ranging from 160 bp to 200 bp downstream of ORI using MEME [31] and found a sequence consensus AA[AG]CA[TA]AA[AG][AT]. It has been proved that this motif can be recognized by ORC1–5 proteins, which depend on unusual DNA topology for binding [32].

3.2. Effect of AT content

Functional dissection analyses have demonstrated that core replication regions are AT-rich and contain several A+T-rich elements [33–37]. Thus, it is necessary to perform a control analysis and investigate whether the lower bendability and cleavage intensity scores in core replication regions were consequences of high AT content or not. By preserving dinucleotide composition, we shuffled the core replication regions in the benchmark dataset and generated 322 (equal to the number of experimentally confirmed ORIs) random sequences with the same length of 250 bp. By using a 50 bp sliding window with 1 bp displacement, we plotted the average bendability and cleavage intensity scores for random sequences (Fig. 3).

We found that both bendability and cleavage intensity profiles for random sequences are markedly different from natural sequences, and do not show significantly lower bendability or cleavage intensity score in core replication regions. These findings strongly demonstrate that the lower bendability and cleavage intensity scores in core replication regions are non-random events and can not be explained by AT content alone.

3.3. Analyzing the relationship between DNA bendability and cleavage intensity

The distribution patterns of DNA bendability and cleavage intensity demonstrate a similarity with respect to genomic position relative to ORI (Figs. 1 and 2). To determine whether the similar distribution patterns were independent of or associated with each other, we compared trinucleotide bendability scores and DNA cleavage intensity scores of tetranucleotides. For each of the trinucleotides having DNA bendability scores, we calculated the average DNA cleavage intensity of all tetranucleotides containing the corresponding trinucleotide in the core replication region. We found that the Pearson's correlation coefficient between the trinucleotide bendability scores and their corresponding average

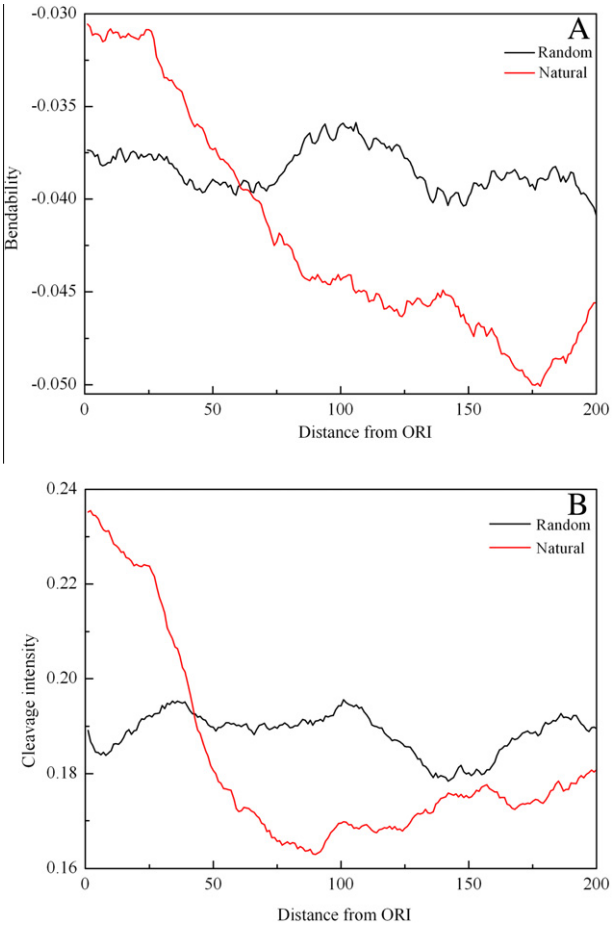

Fig. 3. DNA bendability (A) and cleavage intensity (B) profiles for random sequences with the same dinucleotide composition and length (250 bp) as core replication sequences. DNA bendability and cleavage intensity were all smoothed using a 50 bp sliding window with 1 bp step. The horizontal axis represents the nucleotide position relative to ORI. The vertical axis represents DNA bendability (A) or cleavage intensity (B) score.

tetranucleotide cleavage intensity scores was 0.266, which suggests no direct relationship between DNA bendability and cleavage intensity.

3.4. Identification of ORI using structural features

The above results have demonstrated that DNA structural features are dramatically different between core replication regions and surrounding regions. Thus, we proposed a computational model to identify ORIs based on DNA structural features. In the prediction experiments, we converted each sequence in the benchmark dataset into a vector using DNA bendability and cleavage intensity. To this end, we calculated DNA structural features using the sliding window approach and found that a window size of 50 bp with a step of 50 bp yields the best predictive performance. Thus, each sequence in the benchmark dataset was encoded by a 10-dimensional vector and served as the SVM input. The first five elements of the vector were bendability indexes and the rests were cleavage intensities, which were all calculated in a 50 bp sliding window with 50 bp displacement.

The jackknife cross-validation test was employed to evaluate the performance of our model. For the jackknife cross-validation, each sample in the benchmark dataset is in turn singled out as an independent test sample and all the rule parameters are calculated based on the remaining samples without including the one being identified. To demonstrate the application of our proposed

Table 1
Performance comparison of ORI prediction models in the benchmark dataset.

Parameters	*Sn* (%)	*Sp* (%)	*Acc* (%)	*auROC*
Structural features	85.38	86.17	85.86	0.848
k-mer (*k* = 3, 4)	66.51	78.73	75.62	0.752

model to genome wide ORI prediction, the model was validated in the benchmark dataset and compared with a model based on local word contents of *k*-mer (*k* = 3, 4). The predictive performances are presented in Table1 in terms of sensitivity, specificity, accuracy and auROC. From Table 1, we found that the ORI prediction model based on DNA structural features perform better than the pure sequence-based model. This indicates that DNA bendability and cleavage intensity are informative for ORI prediction.

4. Conclusion

It is difficult to unambiguously identify ORIs from a given DNA sequence because of their degeneracy and frequent organization in several independent modules across the intergenic regions [36,38]. Thus, it is necessary to incorporate more features for ORI prediction. In this study, two DNA structural features around ORI were computationally analyzed in replication regions. The sequence around ORI overall had a lower DNA bendability score and a lower DNA cleavage intensity score than did the linker regions on both sides. DNA replication is thought to be a highly regulated process referring to interactions between regulatory proteins and DNA sequences [39]. Structural divergences between core replication regions and surrounding regions may facilitate DNA unwinding, protein binding and replication fork progression during the process of genome replication.

By combining DNA bendability and cleavage intensity, we proposed a SVM model to predict ORIs. The predictive performance demonstrated that our model is helpful for ORI recognition and that DNA bendability and cleavage intensity may be the hidden codes in replication regions. We expect that DNA bendability and cleavage intensity will shed light on genome-wide ORI prediction and provide novel insights into regulatory mechanisms of DNA replication.

Acknowledgments

We wish to express our gratitude to the editor and anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this article. We thank Prof. Xiyin Wang for his kindly help in language correction. This work was supported by Grants from The National Nature Scientific Foundation of China (No. 61100092), The Scientific Research Startup Foundation of North China Coal Medical University (No. 10101115), Foundation of Scientific and Technological Department of Hebei Province (No. 11275532), The Scientific Research Foundation of Sichuan Province (No. 2009JY0013) and The Fundamental Research Funds for the Central Universities (No. ZYGX2009J081).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2012.02.034.

References

[1] MacAlpine, D.M. and Bell, S.P. (2005) A genomic view of eukaryotic DNA replication. Chromosome Res. 13, 309–326.
[2] Necsulea, A., Guillet, C., Cadoret, J.C., Prioleau, M.N. and Duret, L. (2009) The relationship between DNA replication and human genome organization. Mol. Biol. Evol. 26, 729–741.
[3] Sequeira-Mendes, J., Díaz-Uriarte, R., Apedaile, A., Huntley, D., Brockdorff, N. and Gómez, M. (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet. 5, e1000446.
[4] Nieduszynski, C.A., Knox, Y. and Donaldson, A.D. (2006) Genome-wide identification of replication origins in yeast by comparative genomics. Genes Dev. 20, 1874–1879.
[5] Rowley, A., Cocker, J.H., Harwood, J. and Diffley, J.F. (1995) Initiation complex assembly at budding yeast replication origins begins with the recognition of a bipartite sequence by limiting amounts of the initiator ORC. EMBO J. 14, 2631–2641.
[6] Rao, H. and Stillman, B. (1995) The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. Proc. Natl. Acad. Sci. USA 92, 2224–2228.
[7] Lee, D.G. and Bell, S.P. (1997) Architecture of the yeast origin recognition complex bound to origins of DNA replication. Mol. Cell. Biol. 17, 7159–7168.
[8] Diffley, J.F. and Cocker, J.H. (1992) Protein–DNA interactions at a yeast replication origin. Nature 357, 169–172.
[9] Bell, S.P. and Stillman, B. (1992) ATP-dependent recognition of eukaryotic origins of DNA replication by a multi protein complex. Nature 357, 128–134.
[10] Matsumoto, K. and Ishimi, Y. (1994) Single-stranded-DNA-binding protein dependent DNA unwinding of the yeast ARS1 region. Mol. Cell. Biol. 14, 4624–4632.
[11] Natale, D.A., Umek, R.M. and Kowalski, D. (1993) Ease of DNA unwinding is a conserved property of yeast replication origins. Nucleic Acids Res. 21, 555–560.
[12] Lipford, J.R. and Bell, S.P. (2001) Nucleosomes positioned by ORC facilitate the initiation of DNA replication. Mol. Cell. 7, 21–30.
[13] Nieduszynski, C.A., Hiraga, S., Prashanth, A.K., Benham, C.J. and Donaldson, A.D. (2007) OriDB: a DNA replication origin database. Nucleic Acids Res. 35, D40–D46.
[14] Abeel, T., Saeys, Y., Bonnet, E., Rouze, P. and Van de Peer, Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res. 18, 310–323.
[15] Akan, P. and Deloukas, P. (2008) DNA sequence and structural properties as predictors of human and mouse promoters. Gene 410, 165–176.
[16] Brukner, I., Sánchez, R., Suck, D. and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. EMBO J. 14, 1812–1818.
[17] Greenbaum, J.A., Pang, B. and Tullius, T.D. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res. 17, 947–953.
[18] Tullius, T.D. and Greenbaum, J.A. (2005) Mapping nucleic acid structure by hydroxyl radical cleavage. Curr. Opin. Chem. Biol. 9, 127–134.
[19] Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D. and Margulies, E.H. (2009) Local DNA topography correlates with functional noncoding regions of the human genome. Science 324, 389–392.
[20] Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B. and Mann, R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131, 530–543.
[21] Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein–DNA recognition. Nature 461, 1248–1253.
[22] Churchill, M.E. and Travers, A.A. (1991) Protein motifs that recognize structural features of DNA. Trends Biochem. Sci. 16, 92–97.
[23] Bishop, E.P., Rohs, R., Parker, S.C., West, S.M., Liu, P., Mann, R.S., Honig, B. and Tullius, T.D. (2011) A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. ACS Chem. Biol. 6, 1314–1320.
[24] Chou, K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
[25] Zhu, L., Yang, J. and Shen, H.B. (2009) Multi label learning for prediction of human protein subcellular localizations. Protein J. 28, 384–390.
[26] Chen, W. and Lin, H. (2010) Prediction of midbody, centrosome and kinetochore proteins using gene ontology. Biochem. Biophys. Res. Commun. 401, 382–384.
[27] Lin, H. and Ding, H. (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J. Theor. Biol. 269, 64–69.
[28] Xiong, Y., Liu, J. and Wei, D.Q. (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79, 509–517.
[29] Zou, D., He, Z., He, J. and Xia, Y. (2011) Supersecondary structure prediction using Chou's pseudo amino acid composition. J. Comput. Chem. 32, 271–278.
[30] Chang, C.C., Lin, C.J. LIBSVM: a library for support vector machines, Software available at <http://www.csie.ntu.edu.tw/_cjlin/libsvm>.
[31] Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 34, W369–W373.
[32] Sun, J. and Kong, D. (2010) DNA replication origins, ORC/DNA interaction, and assembly of pre-replication complex in eukaryotes. Acta Biochim Biophys Sin (Shanghai). 42, 433–439.
[33] Clyne, R.K. and Kelly, T.J. (1995) Genetic analysis of an ARS element from the fission yeast *Schizosaccharomyces pombe*. EMBO J. 14, 6348–6357.

[34] Dubey, D.D., Kim, S.M., Todorov, I.T. and Huberman, J.A. (1996) Large, complex modular structure of a fission yeast DNA replication origin. Curr. Biol. 6, 467–473.
[35] Okuno, Y., Satoh, H., Sekiguchi, M. and Masukata, H. (1999) Clustered adenine/thymine stretches are essential for function of a fission yeast replication origin. Mol. Cell. Biol. 19, 6699–6709.
[36] Takahashi, T., Ohara, E., Nishitani, H. and Masukata, H. (2003) Multiple ORC binding sites are required for efficient MCM loading and origin firing in fission yeast. EMBO J. 22, 964–974.
[37] Reeves, R. and Beckerbauer, L. (2001) HMGI/Y proteins: flexible regulators of transcription and chromatin structure. Biochim. Biophys. Acta. 1519, 13–29.
[38] Cotobal, C., Segurado, M. and Antequera, F. (2010) Structural diversity and dynamics of genomic replication origins in *Schizosaccharomyces pombe*. EMBO J 29, 934–942.
[39] Segurado, M., de Luis, A. and Antequera, F. (2003) Genome-wide distribution of DNA replication origins at A+T-rich islands in *Schizosaccharomyces pombe*. EMBO Rep. 4, 1048–1053.