BioSystems 107 (2012) 52–55

Contents lists available at SciVerse ScienceDirect

BioSystems

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

Nucleotide correlation based measure for identifying origin of replication in genomic sequences

Kushal Shah*, Annangarachari Krishnamachari

School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

ARTICLE INFO

Article history:
Received 13 July 2011
Received in revised form 30 August 2011
Accepted 10 September 2011

Keywords:
Computational biology
Genome analysis
Sequence analysis
Bioinformatics
Computational prediction
Biological features

ABSTRACT

Computational prediction of the origin of replication is a challenging problem and of immense interest to biologists. Several methods have been proposed for identifying the replicon site for various classes of organisms. However, these methods have limited applicability since the replication mechanism is different in different organisms. We propose a correlation measure and show that it is correctly able to predict the origin of replication in most of the bacterial genomes. When applied to *Methanocaldococcus jannaschii*, *Plasmodium falciparum* apicoplast and *Nicotiana tabacum* plastid, this correlation based method is able to correctly predict the origin of replication whereas the generally used GC skew measure fails. Thus, this correlation based measure is a novel and promising tool for predicting the origin of replication in a wide class of organisms. This could have important implications in not only gaining a deeper understanding of the replication machinery in higher organisms, but also for drug discovery.

1. Introduction

The most fundamental process in the cell cycle is the replication of DNA and a growing cell must copy the genomic DNA before the cell division. DNA replication is a complex process which includes the selection of the initiation site, unwinding of the helix and assembly of the replication machinery. Due to its central role in the cell cycle, identification of the origin of replication in various organisms is also important for discovery of new drugs for treatment of various diseases (McFadden and Roos, 1999; Soldati, 1999; Raghuram et al., 2007).

It is general practice now to make use of various computational approaches to identify biological features and verify the same with rigorous experimental procedures. Right from primer design to gene identification, computational route is preferred as a starting point for any research investigation (Lobry, 1996a). Researchers employ context based measures along with pattern search algorithms to narrow down the search space and this approach is yielding results as one tries to analyze sequence data related to bacteria, archea and eukaryotic genomes.

In bacteria, a protein called DnaA binds to the DNA in a sequence specific manner and initiates replication process (Mott and Berger, 2007). In circular bacterial chromosomes, the replication begins from the replication origin "ori" and this process elongates bidirectionally till the replication terminus region "ter". During replication process, the leading and lagging strands are subject to mutational pressures which introduces an asymmetrically biased nucleotide composition (Lobry and Sueoka, 2002). This asymmetry can be easily captured by taking the ratio of $(C-G)/(G+C)$ with a sliding window of few kilobases and the resulting graphic pattern depicts two transition regions or points which is indicative of the ori and ter region (Mrazek and Karlin, 1998; Lobry, 1996b). If we do similar analysis by considering the A and T bases, the resultant skew measure does not show prominent transitions as in GC skew. Hence, the GC skew measure has become a prominent computational method to identify origin of replication in bacterial genomes. However, GC skew fails to capture ori in some bacterial, many archeal and almost all eukaryotic genomes. Several approaches have been suggested in this regard (Zhang and Zhang, 2005; Song et al., 2003; Chew et al., 2007; Mackiewicz et al., 2004) and yet a general computational tool is still elusive. This could be due to the fact that DNA replication is initiated from numerous loci in eukaryotic chromosomes and the mechanism is more complex than that of bacteria.

In *Saccharomyces cerevisiae*, the autonomous replicating sequence (ARS) contains a specific consensus sequence called EACS i.e. WWWWTTTAYRTTTWGTT where W = A or T, Y = C or T, and R = A or G (Chang et al., 2011). The underlined 11 base pair element is generally referred to as the A1 element. Origin recognition complex binds to this consensus sequence and initiates the replication process. In addition to the ACS, other response elements B1, B2,

* Corresponding author.
E-mail addresses: kkshah@mail.jnu.ac.in (K. Shah), chari@mail.jnu.ac.in (A. Krishnamachari).

0303-2647/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2011.09.003

B3 are also essential and found to be closely linked to the replication machinery which includes a helicase such as MCM (Bell and Dutta, 2002; Marahrens and Stillman, 1992; Chang et al., 2011; Lee and Bell, 1997; Xu et al., 2006; Breier et al., 2004). Experimental research on the origin of replication in other archea (Klyne and Kelly, 1995; Liacho et al., 2010) has revealed that the architecture and mechanism can vary a lot across various organisms, which further complicates the computational prediction of replication region.

In this paper, we are suggesting a correlation based approach which directly considers the spatial positioning of a specific base in a genomic region and may provide vital clues about the functional features embedded within. The correlation property has been exploited earlier for finding coding segments and other important biological features (Voss, 1992; Hod and Keshet, 2004; Li and Kaneko, 1992; Li et al., 1994; Bernaola-Galvan et al., 2002; Karlin and Brendel, 1993; Munson et al., 1992; Tiwari et al., 1997). This promising measure is also well suited for analyzing symbolic sequences like DNA. The usefulness of this measure is demonstrated with few example whose origin of replication has been known and verified by experimental methods.

2. Materials and Methods

The primary method for a computational prediction of the origin of replication in a particular organism is to divide the entire genome of that organism into subsequences of equal lengths and then study some property associated with each subsequence of nucleic acids. The origin of replication is then identified by a predefined change in value of this property as we analyze the subsequences in a sequential manner along the entire length of the genome sequence.

2.1. GC Skew Method

The GC skew method was the first computational method proposed for identification of origin of replication in genomes (Lobry, 1996a,b; Mrazek and Karlin, 1998). For a given sequence of nucleic acids, the GC skew measure is given by

$$S = \frac{n_C - n_G}{n_C + n_G}$$

where n_C and n_G are the number of occurrences of Cytosine (C) and Guanine (G).

In this method, the origin of replication is said to be at the position where S undergoes an abrupt transition across $S = 0$.

2.2. Correlation Measure

The auto-correlation function, $C(k)$, of a discrete sequence, $\{a_i : i = 1, 2, \ldots, N\}$ with $a_i \in \{+1, -1\}$, is defined as (Beauchamp and Yuen, 1979; Cavicchi, 2000)

$$C(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} a_j a_{j+k} \tag{1}$$

Using Eq. (1), the correlation measure, C_G, can now be defined as the average of all correlation values in Eq. (1),

$$C_G = \frac{1}{N-1} \sum_{k=1}^{N-1} |C(k)| \tag{2}$$

where the subscript “G” refers to “genome”. The value for C_G ranges from 0 to 1 and is independent of the length of the sequence. Lower value of C_G corresponds to lower correlation strength embedded in that sequence and vice-versa. The value of C_G for a typical random sequence will be zero and a highly correlated sequence will approach unity.

To use Eq. (2), we need to convert the nucleic acid sequence into a discrete sequence of bits. Since a DNA sequence is made up of four bases, we can generate a string of bits for the A base by assigning a value of +1 to every occurrence of A and −1 to all other positions (similarly for T, G, C). For example, a DNA sequence *ATGTTCAG* gives rise to four different discrete sequences $\{1, -1, -1, -1, -1, -1, 1, -1\}$, $\{-1, 1, -1, 1, 1, -1, -1, -1\}$, $\{-1, -1, 1, -1, -1, -1, -1, 1\}$ and, $\{-1, -1, -1, -1, -1, 1, -1, -1\}$ corresponding to the four bases A, T, G, C respectively. Thus, a given DNA sequence gives rise to four different bit strings and four different values of correlation strength (i.e. auto-correlation values) corresponding to each of the four bases, A, T, G, C. However, for the purpose of identifying the origin of replication, the correlation values of the Guanine residue give the best results.

In this method, the origin of replication can be identified by an abrupt change in the value of correlation measure, C_G. Unlike the GC skew measure, the characteristic signature is not so well defined for the correlation measure. However, this method is still very useful in identifying the origin of replication, as shown in the next section.

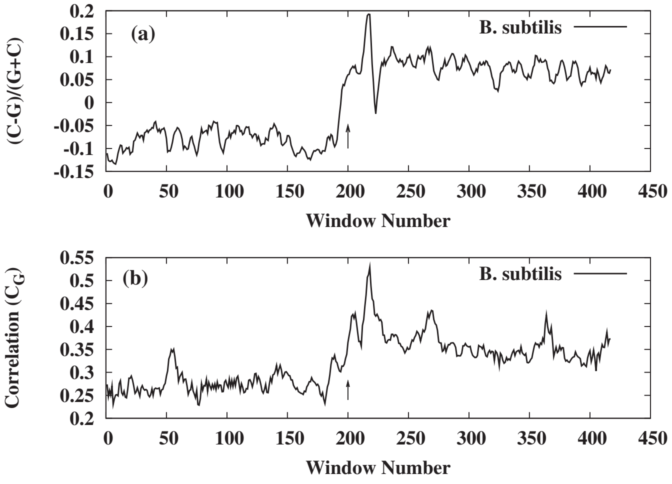

Fig. 1. Prediction of origin of replication position for *B. subtilis* (NC_000964). As indicated by the arrow, both the GC skew and correlation measure are able to correctly predict the origin of replicatoin of *B. subtilis*.

Genome data of all the species were downloaded from NCBI FTP website and the PlasmoDB website.

3. Results

Though the well known GC skew measure has been shown to work for many bacterial genomes, it does fail to correctly predict the origin in many other bacterial genomes. The correlation based measure proposed in this paper, is benchmarked against this GC skew measure (Mrazek and Karlin, 1998) using the whole genome data of few species.

Both GC skew and our method searches the whole genome in an overlapping window fashion for putative origins. In the former method, the transition from one polarity to the other i.e. positive to negative values of GC skew or vice-versa, is considered as a clue to identify the origin of replication. In our correlation based measure, a pronounced jump from an average value either from low to high or vice versa is considered as a clue to the replicon site. Results should be viewed in that perspective. To demonstrate the usefulness of our measure, we have considered four examples, namely (i) *Bacillus subtilis* (ii) *Methanocaldococcus jannaschii* (iii) *Plasmodium falciparum*apicoplast and (iv) *Nicotiana tabacum* plastid.

Fig. 1 depicts the plot of GC skew as well as correlation values (due to the “G” base) for the bacterial organism *B. subtilis*. It can be easily seen that both the measures capture the region of replication origin quite nicely. Fluctuations seen in the graph are due to the choice of window size and one can minimize the noise by trial and error. We have seen the same pattern in many other bacterial genomes, indicating that both the methods are in close agreement with each other for these genomes (data not shown). For this plot, the window size was chosen to be 50 kb with an increment of 10 kb.

In the case of the archeal genome *M. jannaschii*, the GC skew does not depict its characteristic transition about the zero value as evident from Fig. 2 and the pattern also looks too noisy. Our proposed measure shows a distinct transition at one point as shown by the arrow in Fig. 2b and is found to perform better than GC skew in this particular case. The origin of replication shown by the arrow in Fig. 2b is also in conformity with the results obtained with Z-curve method (Zhang and Zhang, 2004). For this plot, the window size was chosen to be 50 kb with an increment of 10 kb.

Considering the importance of the parasitic organism *P. falciparum*, we tried out the plasmid genome apicoplast as a test case

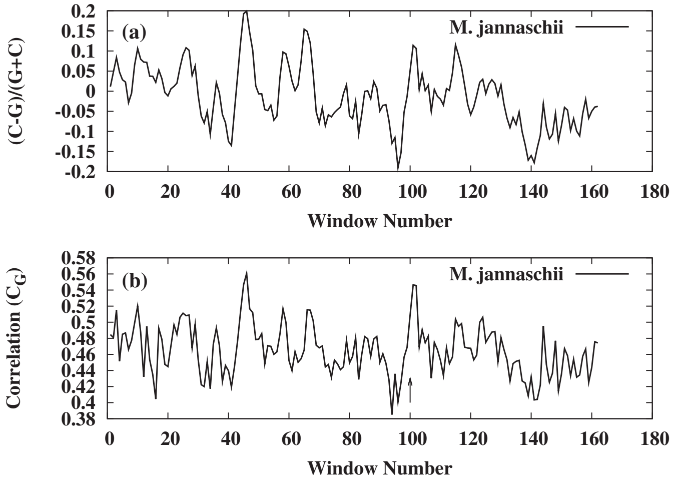

Fig. 2. Prediction of origin of replication position for *M. jannaschii* (NC_000909). As can be seen, correlation measure is able to predict one of the origins of replication in this organism, while the GC skew measure fails to predict the origin position.

whose origin has been worked out in detail (Singh et al., 2003). Fig. 3 shows the GC skew plot which fails to show the point of transition at the experimentally identified origin of replication. In contrast, a plot of the correlation measure accurately points out the origin of replication, as indicated by an arrow in Fig. 3b. This result is also in conformity with experimental evidence (Singh et al., 2003). Thus, the correlation measure shows a promising tool to identify replication origin even in cases where GC skew fails. For this plot, the window size was chosen to be 10 kb with an increment of 100 bases.

Similar results were obtained for the genome of *N. tabacum* plastid. As shown in Fig. 4a, the GC skew method shows no transition at the known location of origin of replication, which should be around 130 kb and 137 kb (Krishnan and Rao, 2009; Kunnimalaiyaan and Nielsen, 1997). Our proposed correlation measure is able to correctly identify the characterized location at 130 kb. Fig. 4b also predicts the presence of another origin or replication at around 100 kb, which needs to be verified by experiments. For this plot, the window size was chosen to be 10 kb with an increment of 1 kb.

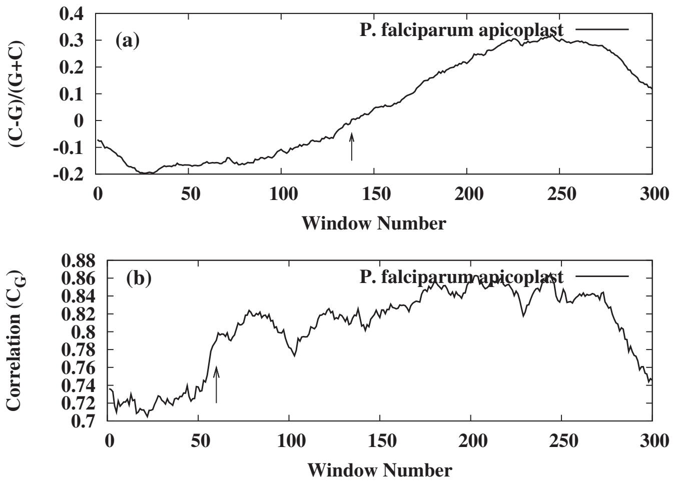

Fig. 3. Prediction of origin of replication position for *P. falciparum* apicoplast (PFC10_API_IRAB). The origin of replication position predicted for this genome by the correlation measure agrees with the experimental findings, whereas the position predicted by GC skew measure in incorrect (Singh et al., 2003). This figure clearly shows that the correlation measure is a better indicator of the origin of replication position.

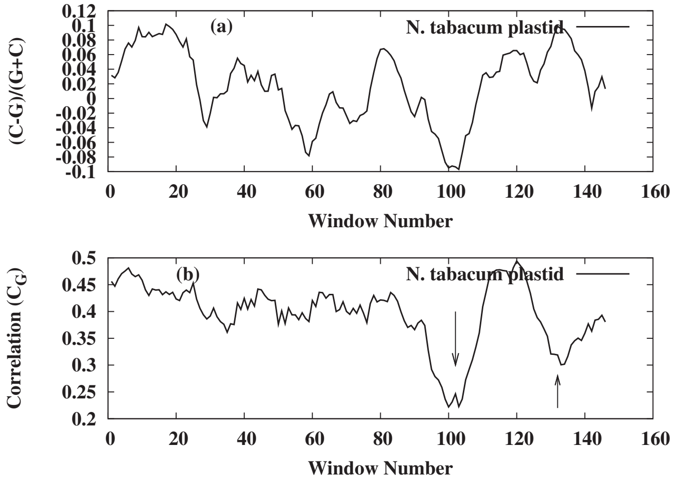

Fig. 4. Prediction of origin of replication position for *N. tabacum* plastid (NC_001879). The origin of replication position predicted for this genome by the correlation measure agrees with the experimental findings, whereas the GC skew measure in unable to make a clear prediction (Krishnan and Rao, 2009; Kunnimalaiyaan and Nielsen, 1997). Like Fig. 3, this figure also clearly shows that the correlation measure is a better indicator of the origin of replication position.

4. Discussion and Conclusion

DNA sequences with the same frequency counts of the four nucleic acids (A, T, G and C) may not necessarily be the same from the functional or structural point of view. This is because the precise positions of each base in the DNA sequence plays an important role in determining various processes that take place within the cell. Hence, a computational prediction tool that is based purely on the frequency counts may not be precise and robust. This explains why methods like GC skew (Mrazek and Karlin, 1998) give good results only for a limited set of lower organisms. For higher organisms, a robust prediction tool must make use of the complete information present in the genomic sequence. The correlation based measure proposed in this paper is a significant development in that direction. As shown in this paper, the correlation based measure gives good results not only in higher organisms, but also in lower organisms where the conventional GC skew method is applicable. If we add pattern searching features of binding site patterns, such as A1, B1, B2 elements, to the proposed algorithm, it will definitely be a very useful tool for experimental biologists in reducing the search space in identifying the origin of replication.

Both the measures (GC skew and correlation) could correctly predict the origin of replication in most of the bacterial genomes. In the case of an archeal genome, *M. jannaschii*, GC skew measure could not identify the origin of replication whereas correlation measure could accurately predict the same. Similarly, in the case of *P. falciparum* apicoplast and *N. tabacum* plastid, it has been clearly demonstrated that the correlation measure could delineate the origin of replication accurately and is found to be better than GC skew. The later predicted the origin at an incorrect position at least in these two organisms. This implies that the correlation among the positions of the nucleic acids is very important and has to be considered while developing any computational tool. This correlation among the base positions seems to be a common feature across the tree of life.

It is interesting to note that the correlation measure, C_G, due to the Guanine residue is able to identify the origin of replication whereas the correlation measure of other bases does not give any useful information. We also calculated the cross-correlations among A, T and G, C, but these values did not give anything meaningful. Similar to the GC skew like measure, a calculation of

(A − T)/(A + T) is unable to correctly identify the origin of replication. Thus, the Guanine residue seems to be playing a fundamental role in one of the most important biological processes that take place inside a cell. This could possibly be due to some important chemical property of Guanine which is not present in Adenine, Thymine or Cytosine and needs to be further explored through experiments.

The advantage of correlation measure over GC skew measure is not an accident or by chance. GC skew measure is based on compositional differences between G and C or, in other words, purely on the counts of these two bases. It is possible that even a random sequence may have similar compositional differences. Correlation takes into account the order of the bases i.e. how the bases are spatially arranged along the DNA chain in terms of positions. It is very unlikely that a random sequence will have a correlation value similar to that of a real DNA sequence. Correlation in the genome is arising out of intrinsic biological constraints (Yin and Yau, 2005).

Our approach is novel in the context of predicting origin of replication and shows great promise for its applicability in higher organisms.

Funding

KS would like to thank the SCIS (JNU) and Center of Excellence of the Department of Biotechnology (Government of India) for financial support.

Acknowledgement

We would like to thank Suman Dhar (SCMM, JNU) for interesting discussions.

References

Bell, S., Dutta, A., 2002. DNA replication in eukaryotic cells. Annual Review of Biochemistry 71, 333–374.

Bernaola-Galvan, P., Carpena, P., Roman-Roldan, R., Oliver, J.L., 2002. Study of statistical correlations in DNA sequences. Gene 300, 105–115.

Beauchamp, K.G., Yuen, C.K., 1979. Digital Methods for Signal Analysis. Geroge Allen and Unwin, London.

Breier, A.M., Chatterji, S., Cozzarelli, N.R., 2004. Prediction of *Saccharomyces cerevisiae* replication origins. Genome Biology 5, R22.

Cavicchi, T.J., 2000. Digital Signal Processing. John Wiley & Sons, New York.

Chang, F., May, C.D., Hoggard, T., Miller, J., Fox, C.A., Weinreich, M., 2011. High resolution analysis of four efficient yeast replication origins reveals new insights into the ORC and putative MCM binding elements. Nucleic Acids Research (Advanced access).

Chew, D.S.H., Leung, M., Choi, K.P., 2007. Excursion: a new approach to predict replication origins in viral genomes by locating AT-rich regions. BMC Bioinformatics 8, 163.

Hod, S., Keshet, U., 2004. Phase transition in random walks with long-range correlations. Physical Review E 70, 015104(R).

Karlin, S., Brendel, V., 1993. Patchiness and correlations in DNA sequences. Science 259, 677–680.

Klyne, R.K., Kelly, T.J., 1995. Genetic analysis of an ARS element from the fission yeast *Schizosaccharomyces pombe*. The EMBO Journal 14, 6348–6357.

Krishnan, N.M., Rao, B.J., 2009. A comparative approach to elucidate chloroplast genome replication. BMC Genomics 10, 237.

Kunnimalaiyaan, M., Nielsen, B.L., 1997. Fine mapping of replication origins (ori A and ori B) in *Nicotiana tabacum* chloroplast DNA. Nucleic Acids Research 25, 3681–3686.

Lee, D.G., Bell, S.P., 1997. Architecture of the yeast origin recognition complex bound to origins of DNA replication. Molecular and Cellular Biology 17, 7159–7168.

Li, W., Kaneko, K., 1992. Long-range correlations and partial 1/f a spectrum in a noncoding DNA sequence. Europhysics Letters 17, 655.

Li, W., Marr, T.G., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. Physica D 75, 392–416.

Liacho, I., Bhaskar, A., Lee, C., Chung, S.C.C., Tye, B., Keich, U., 2010. A comprehensive genome-wide map of autonomously replicating sequences in a naive genome. PLoS Genetics 6, e1000946.

Lobry, J.R., 1996. Origin of replication of Mycoplasma genitalium. Science 272, 745–746.

Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular Biology and Evolution 13, 660–665.

Lobry, J.R., Sueoka, N., 2002. Asymmetric directional mutation pressures in bacteria. Genome Biology 3, 0058.1.

Mackiewicz, P., Zakrzewska-Czerwińska, J., Zawilak, A., Dudek, M.R., Cebrat, S., 2004. Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Research 32, 3781–3791.

Marahrens, Y., Stillman, B., 1992. A yeast chromosomal origin of DNA replication defined by multiple functional elements. Science 255, 817–823.

McFadden, G.I., Roos, D.S., 1999. Apicomplexan plastids as drug targets. Trends in Microbiology 7, 328–333.

Mott, M.L., Berger, J.M., 2007. DNA replication initiation: mechanisms and regulation in bacteria. Nature Reviews Microbiology 5, 343–354.

Mrazek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proceedings of the National Academic Sciences of the United States 95, 3720–3725.

Munson, P.J., Taylor, R.C., Michaels, G.S., 1992. DNA correlations. Nature 360, 636.

Raghuram, E.V.S., Kumar, A., Biswas, S., Kumar, A., Chaubey, S., Siddiqui, M.I., Habib, S., 2007. Nuclear gyrB encodes a functional subunit of the *Plasmodium falciparum* gyrase that is involved in apicoplast DNA replication. Molecular and Biochemical Parasitology 154, 30–39.

Singh, D., Choubey, S., Habib, S., 2003. Replication of the *Plasmodium falciparum* apicoplast DNA initiates within the inverted repeat region. Molecular and Biochemical Parasitology 126, 9–14.

Soldati, D., 1999. The apicoplast as a potential therapeutic target in toxoplasma and other apicomplexan parasites. Parasitology Today 15, 5–7.

Song, J., Ware, A., Liu, S., 2003. Wavelet to predict bacterial ori and ter: a tendency towards a physical balance. BMC Genomics 4, 17.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Computer Applications in the Biosciences 13, 263–270.

Voss, R.F., 1992. Evolution of long range fractal correlations and 1/*f* noise in DNA base sequences. Physical Review Letters 68, 3805–3808.

Yin, C., Yau, S., 2005. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. The Journal of Comparative Biology 12, 1153–1165.

Xu, W., Aparicio, J.G., Aparicio, O.M., Tavaré, S., 2006. Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in *S. cerevisiae*. BMC Genomics 7, 276.

Zhang, R., Zhang, C., 2004. Identification of replication origins in the genome of the methanogenic archaeon, *Methanocaldococcus jannaschii*. Extremophiles 8, 253–258.

Zhang, R., Zhang, C., 2005. Identification of replication origins in archeal genomes based on the Z-curve method. Archaea 1, 335–346.