Genomics Data 9 (2016) 130–136

Contents lists available at ScienceDirect

Genomics Data

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

Context based computational analysis and characterization of ARS consensus sequences (ACS) of *Saccharomyces cerevisiae* genome

Vinod Kumar Singh, Annangarachari Krishnamachari *

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

ARTICLE INFO

Article history:
Received 26 March 2016
Received in revised form 27 June 2016
Accepted 6 July 2016
Available online 19 July 2016

Keywords:
Saccharomyces cerevisiae
Origin of replication
Autonomous replicating sequences
Computational methods

ABSTRACT

Genome-wide experimental studies in *Saccharomyces cerevisiae* reveal that autonomous replicating sequence (ARS) requires an essential consensus sequence (ACS) for replication activity. Computational studies identified thousands of ACS like patterns in the genome. However, only a few hundreds of these sites act as replicating sites and the rest are considered as dormant or evolving sites. In a bid to understand the sequence makeup of replication sites, a content and context-based analysis was performed on a set of replicating ACS sequences that binds to origin-recognition complex (ORC) denoted as ORC-ACS and non-replicating ACS sequences (nrACS), that are not bound by ORC. In this study, DNA properties such as base composition, correlation, sequence dependent thermodynamic and DNA structural profiles, and their positions have been considered for characterizing ORC-ACS and nrACS. Analysis reveals that ORC-ACS depict marked differences in nucleotide composition and context features in its vicinity compared to nrACS. Interestingly, an A-rich motif was also discovered in ORC-ACS sequences within its nucleosome-free region. Profound changes in the conformational features, such as DNA helical twist, inclination angle and stacking energy between ORC-ACS and nrACS were observed. Distribution of ACS motifs in the non-coding segments points to the locations of ORC-ACS which are found far away from the adjacent gene start position compared to nrACS thereby enabling an accessible environment for ORC-proteins. Our attempt is novel in considering the contextual view of ACS and its flanking region along with nucleosome positioning in the *S. cerevisiae* genome and may be useful for any computational prediction scheme.

1. Introduction

In a typical DNA replication process, each parent DNA is duplicated precisely and entirely before cell division takes place. DNA unwinding and loading of replication machinery happens at specific sites called origin of replication (ORI) or replication sites. ORI differs by number, architecture, and replication proteins among the three domains of life: the Bacteria, the Eukaryotes and the Archaea [1]. Identification of ORI plays an important role to understand the replication behavior and its coordinated processes during cell division. Most bacterial genomes are circular and have single ORI site. Computational methods employed to detect these origin sites exploit the property of asymmetrically biased nucleotide composition, and this strategy is found to be successful in bacterial genomes [2–6]. Replication origin firing plays a significant role in cell cycle regulation. Hence locating these sites may be useful in drug development process to combat diseases caused by bacteria, virus, and parasites [4].

Eukaryotic chromosomes are linear and have multiple ORIs which fire in a stochastic manner and carry out the complete duplication of the large genome within a limited period [7,8]. Thus, ORIs prediction in eukaryotes is more challenging and complicated compared to prokaryotes [9] and hence generally used skew based methods are not suitable. Structure and location of replication origins and their functions are more or less well characterized in the eukaryotic model organism *Saccharomyces cerevisiae* [10–12]. Origin activity in yeast depends on cis-acting replicator sequence called autonomous replication sequences (ARS). ARS contains an essential 11 bp ACS (ARS consensus sequences) element and three non-essential elements i.e. B1, B2, and B3. Additionally, ACS and B1 are recognized as the binding sites for origin recognition complex (ORC) proteins, and their motif is well defined. B2 is probably DNA unwinding element (DUE) or helicase loading site, and B3 is transcription factor ARS-binding factor 1 (Abflp) binding site that acts as a replication enhancer element [13–15].

S. cerevisiae genome comprises >12,000 potential ACS sites and approximately 400 origins [16]. A match to the ACS sequence pattern is essential for replicating sites, but it is not sufficient for functional origin. This indicates that presence of some additional unknown functional sequences, chromatin and conformation structures [17]. A study on ORC-ACS and nrACS data show significant differences in nucleosome occupancy status in their flanking regions [12]. Moreover, ORC-ACS are flanked by asymmetric well-positioned nucleosomes on both sides [12]. Nucleosome-free region (NFR) of ORIs are permissive sites for

* Corresponding author.
E-mail address: chari@jnu.ac.in (A. Krishnamachari).

http://dx.doi.org/10.1016/j.gdata.2016.07.005

multiprotein assemblies, i.e., ORC like protein that play critical role in regulating key DNA-templated processes [18]. A study on yeast has also shown that chromatin structure and nucleotides flanking both sides of ACS + B1 play a critical role in origin efficiency along with ARS activity [19].

Carrying out research on DNA replication process is an active area of investigation and it can be divided into three broad categories: 1) replication mechanism and ORC binding sites [19–21], 2) hidden intrinsic characterization of ARS at sequence level [22] and their location in genome [23] and 3) physical properties of DNA like cleavage intensity and DNA bending [24]. Collective study on all these areas, is useful for understanding the regulatory mechanisms of replication process. The present study focus on chosen set of replicating and non-replicating sites using in silico analysis. A recent study on yeast replication sites by Li et al. focused on some nucleotide compositional and DNA conformational properties of ORI regions for their computational prediction experiments [25]. Hence, we have considered the sequence context around ORC-ACS and nrACS motif and its possible influence on conformation and stability of DNA. Our present study attempts to address the issue using two well demarcated datasets of replicating and non-replicating sequences having the ACS like motifs. This may shed light on sequence make of the said sequences.

2. Materials and methods

2.1. Datasets

Two published data sets of *S. cerevisiae* genome were used in this study: 1). ORC-ACS data: ACSs to which ORC binds and show replication activity, and 2) nrACS: ACSs to which ORC doesn't bind and are also found not close to any known replication sites [12]. The original dataset consists of 251 ORC-ACS and 251 nrACS coordinates (genomic locations), and we have not included sequence segments from the chromosome ends (of size 10 kb). Thus we are left with 225 ORC-ACS and 230 nrACS cordinates from the original data. For the sequence based study, 10 kb flanking DNA sequence on both sides of ACS were extracted. The majority of ARS sequences have length ranging from 56 to 2000 bp and ACS motif lie within this segment. Hence, we considered 1000 bp flanking region on both sides for the study and the context is considered. Moreover, the flanking region may contain information and contribute to signal buildup, i.e., the order of sequence makeup may play a role in the protein binding process. The annotation file, nucleosome occupancy signal and ORC binding data were obtained from GEO database (GSE16926) and genome data for *S. cerevisiae* were downloaded from Saccharomyces Genome Database SGD [26]. The values of dinucleotide properties of interest are taken from the DiProDB [27]. The consensus motif of ORC-ACS and nrACS are more or less the same but the sequence makeup of the flanking region are different [12].

2.2. Jensen-Shannon divergence (JSD)

Genomic DNA data can be studied using information theoretic measures such as Shannon entropy and relative entropy for finding conserved protein binding sites in DNA and other genomic features [28,29]. JSD is one of the information theoretic based symmetric measures that captures the divergence between any two given probability distributions [28]. This measure exploits the compositional bias in symbolic DNA data and has been applied to distinguish different genomic regions or segments [28,30]. We have introduced a novel way of computing the JSD from two multiple aligned sets. The purpose of this measure in the present study is to see the divergence pattern arising from the two datasets mentioned earlier.

The adopted procedure to compute JSD is as follows, Centre (ACS) aligned two datasets of ACS sequences were taken together as one set and JSD measure is computed in a sliding overlapping window fashion, but the two required probability distributions are calculated from the ORC-ACS and nrACS datasets independently. Let two datasets namely *r* and *s*, are centrally aligned sequences w.r.t. ACS and the number of sequences having a length (L) are denoted by n_r and n_s respectively. For the chosen window of size w, Shannon entropy between the nucleotide distribution of each column of ORC-ACS (r) and nrACS (s) sequences is computed separately, averaged over the window size, and subtracted from combined entropy of the of all columns of the chosen window from the two datasets together (Fig. 1A). JSD can be calculated as,

$$\mathrm{JSD}\ (X_r;X_s) = H\left(\frac{X_r+X_s}{2}\right) - \frac{1}{2}\ H(X_r) - \frac{1}{2}\ H(X_s) \qquad (1)$$

$$\text{Where, } H(\tfrac{X_r+X_s}{2}) = -\tfrac{1}{w}\sum_{j=1}^{w}\sum_{i\in A,T,G,C}\left(\frac{p_j^r(i)+p_j^s(i)}{2}\right)\ \log_2\ \left(\frac{p_j^r(i)+p_j^s(i)}{2}\right)$$

.

$$H(X_r) = -\frac{1}{w}\sum_{j=1}^{w}\sum_{i\in A,T,G,C} p_j^r(i)\ \log_2\ \left(p_j^r(i)\right),$$

$$H(X_s) = -\frac{1}{w}\sum_{j=1}^{w}\sum_{i\in A,T,G,C} p_j^s(i)\ \log_2\ \left(p_j^s(i)\right)$$

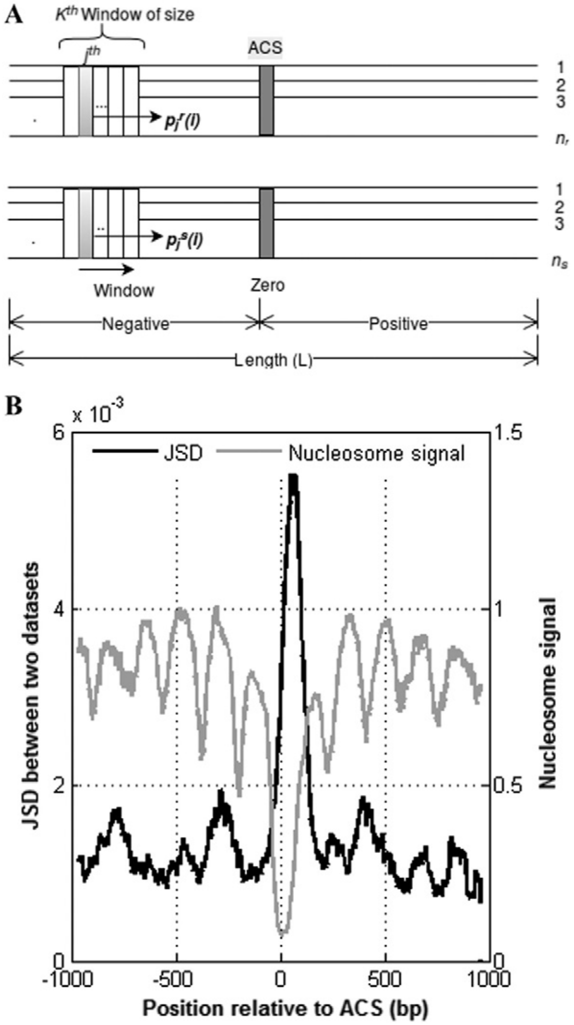

Fig. 1. A. **Scheme for JSD calculation** JSD for of *k*th window of two datasets *r* and *s* is calculated. The window is slided by 1 bp, and the whole process is repeated. Positions marked as positive or negative with respect to ACS. **B**. **Jensen-Shannon divergence between ORC-ACS and nrACS** Y1-axis represents the JSD (dark black) between ORC-ACS and nrACS sequences with respect to ACS location (x-axis). Y2-axis represents nucleosome occupancy signal (gray) along ORC-ACS.

Where $H(x)$ is Shannon-entropy for probability distribution of x, $p_j^r(i)$ and $p_j^s(i)$ is the probability of event i (or nucleotides' in our case) in jth column of kth overlapping window of size w in centrally aligned sequences of given dataset r and s respectively.

JSD at a given location k measures the heterogeneity between two probability distributions. JSD $= 0$ if only if $X_r = X_s$ i.e., lower bound is for identical distribution. JSD is plotted against the positions, to view the divergence embedded in the datasets. Higher value of JSD at a given location k indicates the region of interest and taken up for further study.

2.3. Average sequence measure (ASM)

Measures such as correlation, AT skew or GC skew have been used to capture desired genomic features like ORI, repeats, etc. [2]. Correlation-based methods are also used for sequence alignment, detection of local similarities in DNA and ORI detection in prokaryotes [5,31,32]. Based on these studies, we introduced the Average sequence measure (AS_k) for the sliding overlapping window of size w and at kth position of a given set of n centrally aligned sequences having equal length L was calculated as:

$$AS_k = \frac{1}{n}\sum_{j=1}^{n} S_k^j \tag{3}$$

Where, S_k^j is the value of chosen sequence measure for the kth window of size w in jth sequence. In simple words, a window is selected from the aligned sequences of ORC-ACS dataset, and finally value of the signal is computed. The window is then slided by 1 bp and computation gets repeated till the last possible window.

The S_k^j value for three type of sequence measures were considered in this study and calculated as follows:

1. Correlation measure: Steps for computing correlation value for DNA bases. First, DNA sequence of the chosen window is converted to four numeric strings. For example, base A is replaced with 1, while all other bases i.e., T, G, and C are replaced by −1. In an identical fashion, similar numeric strings were generated for other bases. Finally, the correlation was calculated for each of these numeric strings as described in Shah et al. [5].
2. A GC-Skew measure of the desired window sequence is defined as [33].

$$GC_{Skew} = \frac{f_G - f_C}{f_G + f_C}, \quad GC_{Skew} \in \begin{cases} (0,1] & \text{if} \quad f_G > f_C \\ [-1,0) & \text{if} \quad f_G < f_C \\ 0 & \text{if} \quad f_G = f_C \end{cases} \tag{4}$$

Where f_G and f_C are frequencies of bases guanine, and cytosine in the chosen sequences.

3. AT-skew calculation is similar to GC-skew calculation i.e. replace f_G by f_A and f_C by f_T in Eq. (4).

2.4. Average DNA structure conformation and thermodynamic profile

Protein binding sites in replication and promoter regions of the genome are less flexible (i.e. more rigid) and it helps in their recognition by the regulatory proteins [24,34]. These segments also show higher free energy compared to other parts of the genome that favor DNA unwinding [35]. These properties have been widely used for prediction of protein binding sites. The average profile value (APV_k) calculated using a sliding overlapping window of size w at location k in a set of n sequences of equal length L bps can be calculated as:

$$APV_k = \frac{1}{n.w}\sum_{j=1}^{n}\sum_{i=k}^{k+w-2} PV\left(N_i^j, N_{i+1}^j\right) \tag{5}$$

Where, $PV(N_i^j, N_{i+1}^j)$ is the value of given dinucleotide property at the ith location for jth sequence.

2.5. Software tools used

We have used MEME software suite to identify the embedded motifs in the datasets. MEME utilizes expectation maximization algorithm [36] for discovering novel recurring (ungapped and fixed length) motifs in the given set of sequences [37]. MAST tool was used to scan the putative sites. Positional conservations of the discovered motif were represented in graphical form [38]. For statistical comparison of medians of two independent distributions, MATLAB non-parametric Wilcoxon-ranksum test (Mann-Whitney test) was used. MATLAB codes were written to compute JSD, GC-skew and correlation measure.

3. Results and discussion

The primary objective of the study was to characterize the sequences of *S. cerevisiae* which have ACS like motif and why some of them bound to ORC proteins while others don't. This aspect was studied making use of experimentally verified and published datasets. Aforementioned, ORC-ACS and nrACS showed similarity regarding their motif, but there exists a difference when we map the nucleosome occupancy data [12]. This work is focused on computational based study of compositional, thermodynamic and structural properties of these said sequences and to understand how replicating machinery recognizes ORC-ACS over nrACS. For this study, sequences around each ACS datasets are processed so as to display ARS oriented in the same direction followed by their central (ACS) alignment. It is to be noted that ACS ends are located at position zero. Desired properties were calculated in an overlapping sliding window fashion at each bp step for all given sequences and averaged over the sequences of the respective class.

3.1. Divergence in the nucleotide distribution of ORC-ACS and nrACS sequences within nucleosome-free region (NFR)

Despite the similarity in sequences, ORC-ACS and nrACS differ in nucleosome occupancy around ACS matches. Nucleosomes are conserved, well positioned and asymmetric in ORC-ACS [12]. Some regions of DNA sequences have a relatively higher affinity for nucleosomes [39,40]. Hence, we hypothesized that distribution of bases in ORC-ACS should differ from nrACS sequences at conserved nucleosome positions. With the aim of finding diverged regions at nucleosomes positions of the given two sequence datasets, JSD was calculated (see Methods). Nucleosome wraps approximately 147 bps of DNA. Hence, sliding window equal to half of its size i.e., 75 bps was considered for study [12]. The overall divergence between the distribution of bases at nucleosome locations were not observed, but a marked divergence was seen around +60 position between mentioned datasets (Fig. 1B). Such a pattern of divergence might have occurred due to some conserved signal at this location of ORC-ACS sequences. This prompted us to study further a few sequence-based features for the identified region mentioned above.

3.2. Autocorrelation values (A and T bases) of ORC-ACS shows abrupt rise and fall within NFR

Analysis was further explored using nucleotide-based autocorrelation measure. This measure considers spatial locations of a particular base in a chosen genomic fragment and provides a clue about the nature of embedded signals within sequences [5]. Highly correlated sequence will give unity while the random will give zero for this measure. Correlation strengths of all four numeric strings for each of the sequences were computed independently for a sliding window of size 75 with step size 1 bp and the ensemble average was considered. For the nucleotides A and T base correlation strength for ORC-ACS sequences fall within NFR and shows a sudden change at around +60 position when compared to nrACS (Fig. 2) (Fig. S1). This clearly indicates that this region has some spatial correlation or hidden pattern of A or T base in ORC-ACS. This type of abrupt change in correlation strength due to T-rich ACS motif and A-rich B2 elements. These elements have been reported in prior studies [41]. Results are in conformation with the divergence pattern observed in the previous section.

3.3. Asymmetry in distribution of nucleotide bases

GC skew is a well-known computational measure for detecting replication origin sites in bacterial genomes [33]. To find out whether a GC-skew like pattern is present in eukaryotic replication sites, the average GC-skew analysis was done on ORC-ACS sequences and nrACS (using Eqs. (3) & (4)). The average GC-skew pattern of ORC-ACS shifts its polarity from positive to negative on moving from upstream of NFR to downstream (Fig. 3A), whereas in the case of nrACS no such distinct change was observed (Fig. 3B).

In bacterial genomes, GC-skew shows a transition in polarity at replicating or terminus sites due to the preference of G over C in the leading strand of DNA. Thus, one can easily say that replicating nature of ORC-ACS region may have caused this asymmetry and in *S. cerevisiae* replication mechanism [2,33,42] may be similar to that of bacteria in the same fashion for each origin sites. This study finds two interesting observations:

1. Both ORC-ACS and nrACS datasets showed significantly higher GC skew signal on upstream of NFR w.r.t to downstream with p-value < 10^{-247} and p-value < 10^{-165} (Wilcoxon rank sum test) respectively. In summary, Compositional bias on both sides of NFR was seen in both datasets, but the transition of GC-skew polarity on both sides of NFR was only visible in ORC-ACS. Significantly high value of GC-skew at upstream of nrACS suggests that nrACS matches may be evolving as replicating ACS or have very low chances of ORC binding to them. Hence there is a possibility that nrACS sites could not be detected as ORC-ACS sites in any of the experiment carried out till date.
2. The GC-skew value goes back to nearly zero around 1000 bp away from ORC-ACS whereas average inter ORC-ACS distance in yeast was about 40 kb (Fig. S2). Influence of mutational effect due to replication slowly reduces with distance and need not be stretched to 20 kb. The cause of this observation may be due to combined effect of variability in the location of fork-convergence sites, the stochastic firing of ARS, the asynchronous departure of two forks from the ARS, and the difference in rates of fork progression [7,42,43].

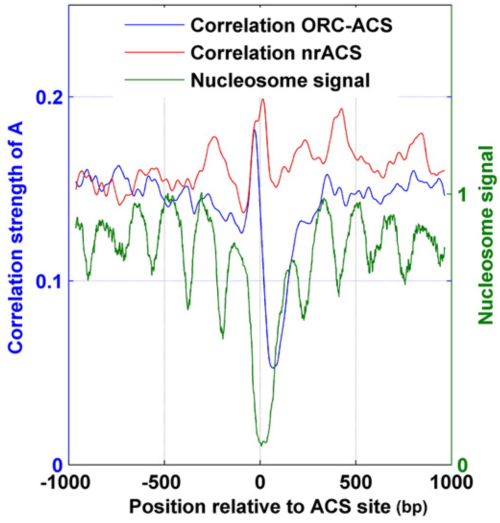

Fig. 2. Correlation strength of ORC-ACS and nrACS sequences Y1-axis represents correlation strength of 'A' for ORC-ACS (blue) and nrACS (red) sequences using windows of size 75 bp and step size 1 bp. Y2-axis represents nucleosome occupancy signal (green) along ORC-ACS sequences. X-axis represents position relative to ACS.

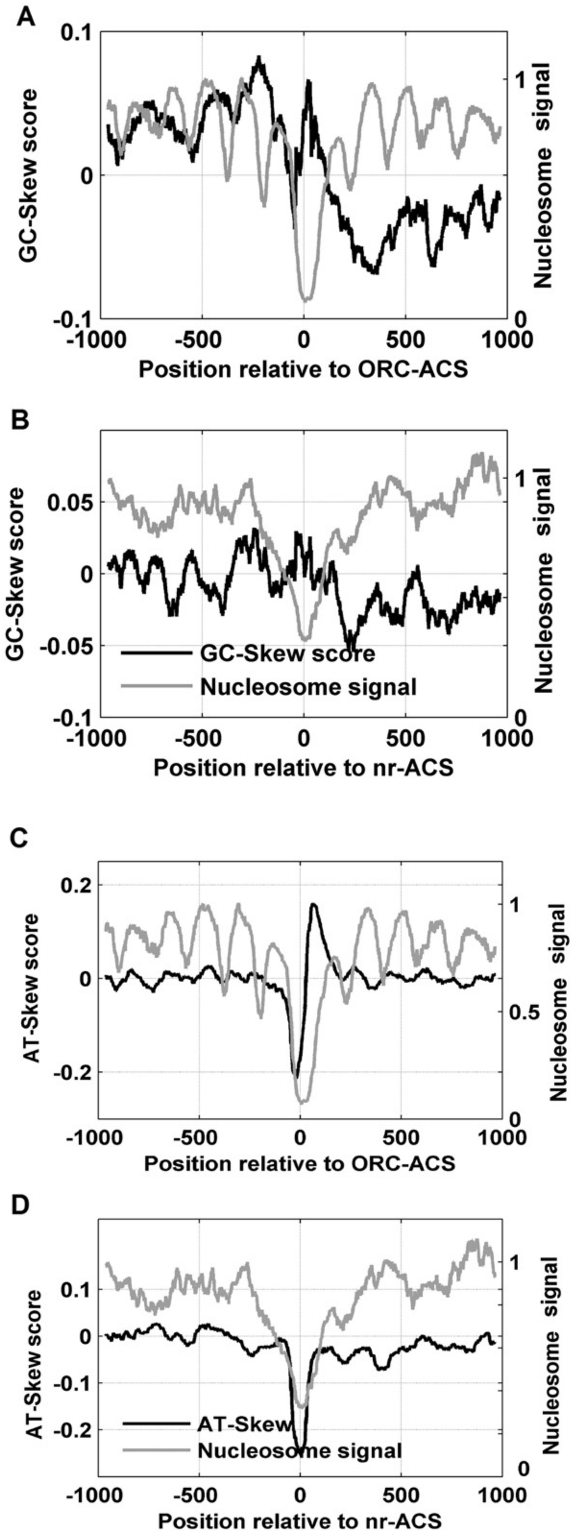

Fig. 3. Compositional skew plots Y1-axis represents (**A**) GC-Skew along ORC-ACS (**B**) GC-Skew along nrACS (**C**) AT-Skew along ORC-ACS (**D**) AT-Skew along nrACS sequences using sliding window of size 75 bp and step size 1 bp. Y2–axis represents nucleosome signal (gray) of corresponding sequences. X-axis represents position relative to ACS.

In the case of AT-skew plot, only ORC-ACS shows transition behavior within NFR (Fig. 3C, and D). It is interesting to note that, AT-polarity

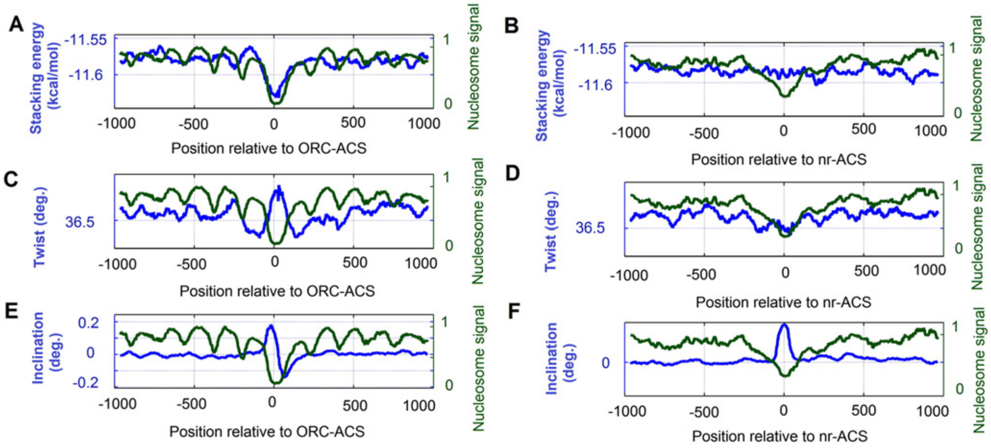

Fig. 4. Stacking energy and conformational variation along ORC-ACS and nrACS sequences Y1-axis represents blue curve represents stacking energy of (**A**) ORC-ACS (**B**) nr-ACS sequences, helical twist angle of (**C**) ORC-ACS (**D**) nr-ACS sequences, inclination angle of (**E**) ORC-ACS (**F**) nr-ACS sequences, Y2-axis represents (Green curve) nucleosome signal. X-axis represents position relative to ACS.

transition was only visible within NFR. This finding was slightly different to an earlier study done by Aiger [42] on the genic and intergenic sequences of leading and lagging strand. They reported transition in AT-polarity on moving from upstream replication termination region of ARS to downstream replication terminal region [42]. However, this study shows that property exists only within NFR because of A and T rich regions. This might have occurred due to consistency in the signal attained by aligning ARS sequences by their ACS of the same orientation. This is clear from the analysis of correlation and skew measure, that the context of the sequence makeup near the vicinity of NFR is important and may play a significant role in the replication process.

3.4. Motif discovery and their localization within NFR

Analysis of sequences in this study indicated the presence of A-rich pattern around +50 location. Earlier work by Eaton et al. [41] reported that motif downstream of ACS is A-rich but did not figure out the pattern. Hence, the downstream region i.e. +20 to +180 (within NFR) of ORC-ACS was searched for any conserved motif using MEME suit [37]. The study discovered an A-rich motif (in 88 out of 225 ORC-ACS sequences) (Fig. S3A). PSSM of the discovered motif was used to search similar motifs in downstream of nrACS sequences i.e. between +20 to +180 using MAST suit with p-value < 10^{-4} and E-value < 10 [44].

The frequency distribution of discovered motif shows that downstream region of ORC-ACS has a higher number of discovered motifs up to +60 bp when compared to nrACS (Fig. S3B & S3C, where 1 bin represents 5 bps). The frequency plot of the previously discovered B2 motif (5′-ANWWAAAT-3′) [19] (Fig. S4) matches with frequency plot of motif discovered in this study. A-rich motifs discovered in both studies shows abundance up to +65 in ORC-ACS. The length of the newly discovered motif is bit longer compared to the previously known B2 motif, and the chances of random occurrence of a long motif in a given DNA sequence are very low. B2 elements are not present in all ARS but play a significant role in enhancing ARS efficiency and DNA unwinding [45] by loading MCM2 helicase or possibly act as a second site for ORC binding [46,47].

3.5. Variability in DNA structural and thermodynamical properties within region of interest in ORC-ACS sequences

DNA replication process starts with unwinding of DNA double helix and the thermal stability of the double helix is contributed by Watson and Crick base pairing and base stacking [48]. Base pairing such as GC and AT content plays a significant role in DNA helix stability of replication regions [49].

Base stacking is the stacking of one base over the other in a DNA single strand. Stacking energy is energy required to destack two consecutive stacking bases in a single strand. Hence to examine the effect of base stacking energy at replication sites, average base stacking energy profile for the above datasets, based on Eq. (5), was calculated using dinucleotide stacking values obtained from DiProDB database [27,50]. ORC-ACS data shows lower stacking base energy compared to nrACS, when considered within NFR (Fig. 4A & B). This implies that NFR region needs less energy to destack bases of a single strand. Furthermore, earlier published study also reported the destacking of bases play a major role in DNA unwinding during replication process [51], and our findings were in conformity with that study.

Stacking energy also influences the conformational structure of DNA, e.g., twist angle between stacking base pairs that cause helical repeats in DNA. Helical twist angle showed higher value within NFR of ORC-ACS as compared to nrACS data (methods Eq. (5); Fig. 4C & D). Such a correlation between base stacking energy and the helical twist angle was also observed by Swart and Cooper et al. [52,53]. ORC-ACS in the NFR shows higher twist angle in the ORC binding site compared to the neighboring region. nrACS within NFR has helical twist angle similar to its neighboring region. This unique feature, i.e., higher twist angle may have a role in DNA-protein interaction during replication process [49].

The orientation of a base pair on either overall or a local helix axis is known as base Inclination angle; a major factor in determining DNA structure. ORC-ACS data shows (Fig. 4E & F) switching of inclination from +ve to −ve within NFR. In the DNA structure, this implies a sudden change from the right-hand rotation to left-hand rotation about a vector from the helical axis towards the major groove [54] within ORC-ACS NFR. This signifies the role of A-rich region in shaping DNA structure for ORC binding.

Table 1
Distribution of ACS matches within three types of intergenic regions.

	ACS	
Intergenic region	ORC-ACS (225)	nrACS (230)
Tandem (one promoter)	98	73
3064 (48.4% of total)	(3.2% of tandem)	(2.4% of tandem)
Convergent (no promoter)	66	60
1671 (26.4% of total)	(3.9% of convergent)	(3.5% of convergent)
Divergent (two promoter)	45	33
1593 (25.2% of total)	(2.8% of divergent)	(2.1% of divergent)
Total	209/6328	166/6328
6328	(3.3% of intergenic)	(2.6% of intergenic)

Note: Convergent intergenic regions for both datasets have a high density of ACS within them followed by tandem and convergent.

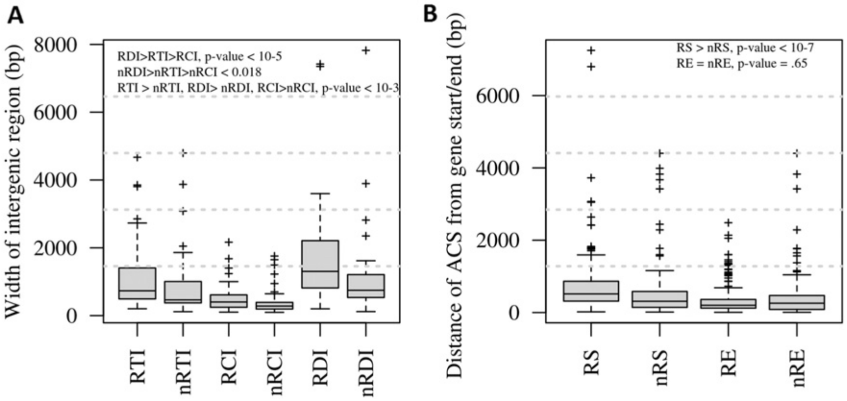

Fig. 5. Distribution of intergenic regions containing ACS (A) Boxplot of width (in base pairs) of three types of intergenic regions embedding ORC-ACS and nrACS: RTI - ORC-ACS within the tandem intergenic region, nRTI - nrACS within the tandem intergenic region. RCI - ORC-ACS within the convergent intergenic region, nRCI - nrACS within the convergent intergenic region. RDI - ORC-ACS within the divergent intergenic region, nRDI - nrACS within the divergent intergenic region. (B) Distance (in base pairs) of ORC-ACS and nrACS from gene start or end position: RS - Distance from ORC-ACS to nearest gene start position, RS - Distance from nrACS to nearest gene start position, RE - Distance from ORC-ACS to nearest gene end position, nRE - Distance from nrACS to nearest gene start position.

In this study, we figured out A-rich B2 elements are abundantly present within NFR of ORC-ACS. Our analysis suggests a new possible role of B2 element on DNA conformation and thermodynamics that favors protein binding to these sites and DNA unwinding in the NFR of ORC-ACS. Previous studies also showed the effect of AT-tract of B2 element on the free energy contribution for DNA unwinding [45]. Junction of poly(A) tracts and mixed base sequence also plays a role in influencing DNA structural properties and nucleosome organization [55,56].

3.6. Distribution of ACS in the non-coding segment of the genome

In the human genome, the replicating regions are surrounded by abundant genes and replication fork progression is co-oriented with transcription [57]. Most of these replicating sites are not randomly distributed. They instead overlap with promoter regions [58]. To examine the relationship between transcription segments and ACS sites, three types of intergenic regions namely tandem, divergent and convergent were extracted from SacCer3 annotation file. Distribution of ORC-ACS and nrACS in above mentioned intergenic regions were analyzed and tabulated (see Table 1.)

The median of the intergenic width of divergent, tandem, and convergent are 1305, 753 and 403 bp respectively for ORC-ACS data while for nrACS data median of intergenic widths are 749, 469, and 282 bp respectively (Fig. 5A). The analysis of Table 1 and Fig. 5A gives two observations: 1) the convergent intergenic regions have the highest percentage of ACS (i.e. 3.9% for ORC-ACS and 3.5% for nrACS) and have the lowest median of intergenic width (p-values < 10^{-5} for ORC-ACS and p-values < 0.018 for nrACS, Wilcoxon rank sum right tail test) followed by tandem and divergent, 2) the width of intergenic regions embedding ORC-ACS is significantly broader as compared to nrACS sites (p-value < 10^{-3}, Wilcoxon rank sum test). It seems convergent intergenic region is highly preferred followed by tandem and divergent. This analysis also validates the possible relation between transcription and replication [58].

The locations of the nearest gene start site or TSS from ORC-ACS site are further as compared to nrACS (p-values < 10^{-7}, Wilcoxon rank sum right tail test, Fig. 5B). In our view, this may facilitate to accomplish replication proteins [59]. Hence small distance in case of nrACS may be playing a crucial role by limiting DNA accessibility for ORC to bind despite its close similarity to ORC-ACS motifs. Thus, the analysis showed that the replicating ACSs prefer to maintain sufficient distance from promoter regions where transcription machinery binds.

4. Conclusion

In *S. cerevisiae*, out of a large number of available ACS sites, replication protein complex binds only to some of them as demonstrated by many genome-wide experiments. Hence, there was a need to understand the sequence context in which these motifs are occurring. Our study focused on content and contextual analysis of ORC binding ACS and ORC free ACS sites. We could see profound changes in the conformation related features such as DNA helical twist, inclination angle and stacking energy in ORC-ACS compared to nrACS. Our analysis suggests a new possible role of A-rich B2 elements along with T-rich ACS, which may be providing structurally and thermodynamically favorable environment for ORC to bind DNA and carry forward the replication process. This study also showed that nrACS are closer to the nearest transcription start site position as compared to ORC-ACS, and this may be a limiting factor for DNA accessibility to ORC proteins.

Even though both datasets have similar sequences, our study confirm the fact that context features within the nucleosome-free region differ. This may be the reason for the ORC-ACS to act as replicating sites. Our attempt is novel in considering ARS consensus sequences and its flanking region with nucleosome positioning to get contextual insights in *S. cerevisiae* genome. Thus, our context-based computational study is bit comprehensive in the analysis of the replication origin sequences of *S. cerevisiae* and may be useful for biologists and bioinformaticians to plan further studies and will be helpful to develop origin prediction tools.

Author contribution

AK conceived and designed the study. VKS performed the analysis. VKS and AK wrote the manuscript. Both the authors have read and approved the manuscript.

Conflict of interest

We declare that there is no conflict of interests for this work.

Acknowledgment

VKS would like to thank JNU, UGC, DST, DBT-COE, and CSIR, India for the fellowship and support. Authors would also like to thank Dr. David MacAlpine, Duke University for his useful comments and kindly providing ORC-ACS and nrACS data for our study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2016.07.005.

References

[1] N.P. Robinson, S.D. Bell, Origins of DNA replication in the three domains of life. FEBS J. 272 (2005) 3757–3766.
[2] J. Mrázek, S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 3720–3725.
[3] J.R. Lobry, N. Sueoka, Asymmetric directional mutation pressures in bacteria. Genome Biol. 3 (2002) (RESEARCH0058).
[4] E.V.S. Raghu Ram, A. Kumar, S. Biswas, A. Kumar, S. Chaubey, M.I. Siddiqi, et al., Nuclear gyrB encodes a functional subunit of the *Plasmodium falciparum* gyrase that is involved in apicoplast DNA replication. Mol. Biochem. Parasitol. 154 (2007) 30–39.
[5] K. Shah, A. Krishnamachari, Nucleotide correlation based measure for identifying origin of replication in genomic sequences. Bio Systems 107 (2012) 52–55.
[6] F. Gao, Recent advances in the identification of replication origins based on the Z-curve method. Curr. Genomics 15 (2014) 104–112.
[7] P.K. Patel, B. Arcangioli, S.P. Baker, A. Bensimon, N. Rhind, DNA replication origins fire stochastically in fission yeast. Mol. Biol. Cell 17 (2006) 308–316.
[8] C. Peng, H. Luo, X. Zhang, F. Gao, Recent advances in the genome-wide study of DNA replication origins in yeast. Front. Microbiol. 6 (2015) 117.
[9] N.P. Robinson, I. Dionne, M. Lundgren, V.L. Marsh, R. Bernander, S.D. Bell, Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. Cell 116 (2004) 25–38.
[10] T.J. Kelly, G.W. Brown, Regulation of chromosome replication. Annu. Rev. Biochem. 69 (2000) 829–880.
[11] D.M. Gilbert, Making sense of eukaryotic DNA replication origins. Science 294 (2001) 96–101.
[12] M.L. Eaton, K. Galani, S. Kang, S.P. Bell, D.M. MacAlpine, Conserved nucleosome positioning defines replication origins. Genes Dev. 24 (2010) 748–753.
[13] C.S. Newlon, J.F. Theis, The structure and function of yeast ARS elements. Curr. Opin. Genet. Dev. 3 (1993) 752–758.
[14] L. Zou, B. Stillman, Assembly of a complex containing Cdc45p, replication protein A, and Mcm2p at replication origins controlled by S-phase cyclin-dependent kinases and Cdc7p-Dbf4p kinase. Mol. Cell. Biol. 20 (2000) 3086–3096.
[15] G.M. Wilmes, S.P. Bell, The B2 element of the *Saccharomyces cerevisiae* ARS1 origin of replication requires specific sequences to facilitate pre-RC formation. Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 101–106.
[16] C.a. Nieduszynski, Y. Knox, A.D. Donaldson, Genome-wide identification of replication origins in yeast by comparative genomics. Genes Dev. 20 (2006) 1874–1879.
[17] K. Shirahige, T. Iwasaki, M.B. Rashid, N. Ogasawara, H. Yoshikawa, Location and characterization of autonomously replicating sequences from chromosome VI of *Saccharomyces cerevisiae*. Mol. Cell. Biol. 13 (1993) 5043–5056.
[18] O.J. Rando, H.Y. Chang, Genome-wide views of chromatin structure. Annu. Rev. Biochem. 78 (2009) 245–271.
[19] F.J. Chang, C.D. May, T. Hoggard, J. Miller, C.a. Fox, M. Weinreich, High-resolution analysis of four efficient yeast replication origins reveals new insights into the ORC and putative MCM binding elements. Nucleic Acids Res. 39 (2011) 6523–6535.
[20] H. Rao, B. Stillman, The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 2224–2228.
[21] K. Yoshida, A. Poveda, P. Pasero, Time to be versatile: Regulation of the replication timing program in budding yeast. J. Mol. Biol. 425 (2013) 4696–4705.
[22] I. Liachko, E. Tanaka, K. Cox, S.C.C. Chung, L. Yang, A. Seher, et al., Novel features of ARS selection in budding yeast *Lachancea kluyveri*. BMC Genomics 12 (2011) 633.
[23] J. Bechhoefer, N. Rhind, Replication timing and its emergence from stochastic processes. Trends Genet. 28 (2012) 374–381.
[24] W. Chen, P. Feng, H. Lin, Prediction of replication origins by calculating DNA structural properties. FEBS Lett. 586 (2012) 934–938.
[25] W.-C. Li, Z.-J. Zhong, P.-P. Zhu, E.-Z. Deng, H. Ding, W. Chen, et al., Sequence analysis of origins of replication in the *Saccharomyces cerevisiae* genomes. Front. Microbiol. 5 (2014) 574.
[26] J.M. Cherry, E.L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E.T. Chan, et al., Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res. 40 (2012) D700–D705.
[27] M. Friedel, S. Nikolajewa, J. Sühnel, T. Wilhelm, DiProDB: a database for dinucleotide properties. Nucleic Acids Res. 37 (2009) 37–40.
[28] J.V. Braun, H.-G. Muller, Statistical methods for DNA sequence segmentation on. JSTOR 13 (1998) 142–162.
[29] T.D. Schneider, A brief review of molecular information theory. Nano Commun. Networks 1 (2010) 173–180.
[30] W. Li, P. Bernaola-Galván, F. Haghighi, I. Grosse, Applications of recursive segmentation to the analysis of DNA sequences. Comput. Chem. 26 (2002) 491–510.
[31] E.A. Cheever, G.C. Overton, D.B. Searls, Fast Fourier transform-based correlation of DNA sequences using complex plane encoding. Comput. Appl. Biosci. (CABIOS) 7 (1991) 143–154.
[32] M. Curilem Saldías, F. Villarroel Sassarini, C. Muñoz Poblete, A. Vargas Vásquez, I. Maureira Butler, Image correlation method for DNA sequence alignment. PLoS One 7 (2012) 1–11.
[33] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13 (1996) 660–665.
[34] X.-Q. Cao, J. Zeng, H. Yan, Structural properties of replication origins in yeast DNA sequences. Phys. Biol. 5 (2008) 036012.
[35] A. Kanhere, M. Bansal, A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinf. 6 (2005) 1.
[36] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings/… International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology, 2, 1994, pp. 28–36.
[37] T.L. Bailey, M. Boden, F.a. Buske, M. Frith, C.E. Grant, L. Clementi, et al., MEME suite: tools for motif discovery and searching. Nucleic Acids Res. 37 (2009) 202–208.
[38] T.D. Schneider, R.M. Stephens, Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18 (1990) 6097–6100.
[39] K. Chen, Q. Meng, L. Ma, Q. Liu, P. Tang, C. Chiu, et al., A novel DNA sequence periodicity decodes nucleosome positioning. Nucleic Acids Res. 36 (2008) 6228–6236.
[40] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, et al., A genomic code for nucleosome positioning. Nature 442 (2006) 772–778.
[41] M.L. Eaton, K. Galani, S. Kang, S.P. Bell, D.M. MacAlpine, Conserved nucleosome positioning defines replication origins. Genes Dev. 24 (2010) 748–753.
[42] N. Agier, G. Fischer, The mutational profile of the yeast genome is shaped by replication. Mol. Biol. Evol. 29 (2012) 905–913.
[43] Y.I. Pavlov, C.S. Newlon, T.A. Kunkel, N. Carolina, Yeast Origins Establish a Strand Bias for Replicational Mutagenesis. 10, 2002 207–213.
[44] T.L. Bailey, M. Gribskov, Combining evidence using *p*-values: application to sequence homology searches. Bioinformatics 14 (1998) 48–54.
[45] J.F. Theis, C.S. Newlon, Domain B of ARS307 contains two functional elements and contributes to chromosomal replication origin function. Mol. Cell. Biol. 14 (1994) 7652–7659.
[46] S.P. Bell, B. Stillman, ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. Nature 357 (1992) 128–134.
[47] J.a. Belsky, H.K. MacAlpine, Y. Lubelsky, A.J. Hartemink, D.M. MacAlpine, Genome-wide chromatin footprinting reveals changes in replication origin architecture induced by pre-RC assembly. Genes Dev. 29 (2015) 212–224.
[48] P. Yakovchuk, E. Protozanova, M.D. Frank-Kamenetskii, Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res. 34 (2006) 564–574.
[49] A. Uzman, H. Lodish, A. Berk, L. Zipursky, D. Baltimore, Molecular cell biology. Biochemistry and Molecular Biology Education, fourth ed., Section 1.2The Molecules of Life, 29, 2000 (New York, NY, 2000, ISBN 0-7167-3136-3).
[50] J. Sponer, H.a. Gabb, J. Leszczynski, P. Hobza, Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. Biophys. J. 73 (1997) 76–87.
[51] E.T. Kool, Hydrogen bonding, base stacking, and steric effects in DNA replication. Annu. Rev. Biophys. Biomol. Struct. 30 (2001) 1–22.
[52] M. Swart, T. van der Wijst, C. Fonseca Guerra, F.M. Bickelhaupt, Pi-pi stacking tackled with density functional theory. J. Mol. Model. 13 (2007) 1245–1257.
[53] V.R. Cooper, T. Thonhauser, A. Puzder, E. Schröder, B.I. Lundqvist, D.C. Langreth, Stacking interactions and the twist of DNA. J. Am. Chem. Soc. 130 (2008) 1304–1308.
[54] A. Lebrun, R. Lavery, Modelling extreme stretching of DNA. Nucleic Acids Res. 24 (1996) 2260–2267.
[55] C. Yoon, G.G. Privé, D.S. Goodsell, R.E. Dickerson, Structure of an alternating-B DNA helix and its relationship to A-tract DNA. Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 6332–6336.
[56] E. Segal, J. Widom, Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr. Opin. Struct. Biol. 19 (2009) 65–71.
[57] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. D'Aubenton-Carafa, A. Arneodo, et al., Human gene organization driven by the coordination of replication and transcription. Genome Res. 17 (2007) 1278–1285.
[58] A. Necsulea, C. Guillet, J.-C. Cadoret, M.-N. Prioleau, L. Duret, The relationship between DNA replication and human genome organization. Mol. Biol. Evol. 26 (2009) 729–741.
[59] 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) 1–13.