Computational prediction of replication sites in DNA sequences using complex number representation

Shubham Kundal,[1] Raunak Lohiya,[2] Hritik Bansal,[1] Shreya Johri,[3] Varuni Sarwal,[3] and Kushal Shah[4,*]

[1]*Department of Electrical Engineering, Indian Institute of Technology (IIT) Delhi, Hauz Khas, New Delhi - 110016, India.*
[2]*Department of Mathematics, Indian Institute of Technology (IIT) Delhi, Hauz Khas, New Delhi - 110016, India.*
[3]*Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology (IIT) Delhi, Hauz Khas, New Delhi - 110016, India.*
[4]*Department of Electrical Engineering and Computer Science, Indian Institute of Science Education and Research (IISER), Bhopal - 462066, Madhya Pradesh, India.*

Computational prediction of origin of replication (ORI) has been of great interest in bioinformatics and several methods including GC-skew, auto-correlation etc. have been explored in the past. In this paper, we have extended the auto-correlation method to predict ORI location with much higher resolution for prokaryotes and eukaryotes, which can be very helpful in experimental validation of the computational predictions. The proposed complex correlation method (iCorr) converts the genome sequence into a sequence of complex numbers by mapping the nucleotides to $\{+1, -1, +i, -i\}$ instead of $\{+1, -1\}$ used in the auto-correlation method (here, i is square root of -1). Thus, the iCorr method exploits the complete spatial information about the positions of all the four nucleotides unlike the earlier auto-correlation method which uses the positional information of only one nucleotide. Also, the earlier auto-correlation method required visual inspection of the obtained graphs to identify the location of origin of replication. The proposed iCorr method does away with this need and is able to identify the origin location simply by picking the peak in the iCorr graph.

I. INTRODUCTION

DNA replication is a complex biological process by which the genome/chromosome of an organism creates a copy of itself during cell division. The segment of DNA sequence where the process of replication initiates is called origin of replication (ORI). The ability to computationally predict ORI location is important to understand the statistical features in a DNA sequence and in the future, could also provide information for the development of new drugs for treatment of diseases.

Prokaryotic organisms are usually found to have a single origin of replication from where two replication forks transmit in contrary directions [1–3]. More evolved organisms are found to contain multiple sites from which replication initiates and this helps to speed up the process [4, 5]. Experimental detection of ORI locations is very challenging and so far has been completed only for a very few archaea, eubacteria and eukaryotic genomes [6]. Here, computational prediction can play a significant role by considerably reducing the search space which can save a large amount of experimental time, effort and resources. Computational prediction of ORI rests on the general hypothesis that the origin location and its flanking regions have different statistical properties as compared to rest of the genome. Motivation for this hypothesis comes from the fact that the replication process of the leading and lagging strands takes place through a slightly different set of proteins which can leave certain statistical signatures at the origin location [7, 8].

Different computational methods have been earlier developed to predict origin of replication in DNA sequences including GC-skew [7–10], Z-curve [11], CGC Skew [12], AT excursion [13], Shannon entropy [14–16], wavelet approach [17], auto-correlation based measure [18], correlated entropy measure (CEM) [19], GC profile [20] and few others. All methods use the fundamental property of identifying differences in statistical properties in the upstream and downstream side of replication origin to account for mutational pressures developed in the opening and ending strands of ORI [21, 22].

In the GC-skew [9] and auto-correlation method [18], the entire genome is divided into overlapping segments/windows and the value of a certain statistical measure is calculated for each window. For bacterial genomes, usually the window size is chosen to be around one-hundredth of the genome size and two consecutive windows have an overlap of four-fifths of the window size. So, only one-fifth of the genome sequence is changed per window which helps to reduce the noise produced by sharp variations of correlation measure in adjacent windows.

In the GC-skew method, the number of G and C nucleotides is counted for each segment/window and the GC-skew value,

$$S_{GC} = \frac{C - G}{C + G} \tag{1}$$

is plotted against the window number. An ORI (or TER) is then predicted to be present at the location where the GC-skew value crosses the zero line from above (or below). The auto-correlation method (henceforth, called gCorr) goes a step further and uses the positional information of the G nucleotides in each window and hence, is informationally richer than the GC-skew method. Predictions of the gCorr method for ORI location of chromosome 1 and 10 of *P. falciparum* have been recently ex-

*Electronic address: kushals@iiserb.ac.in

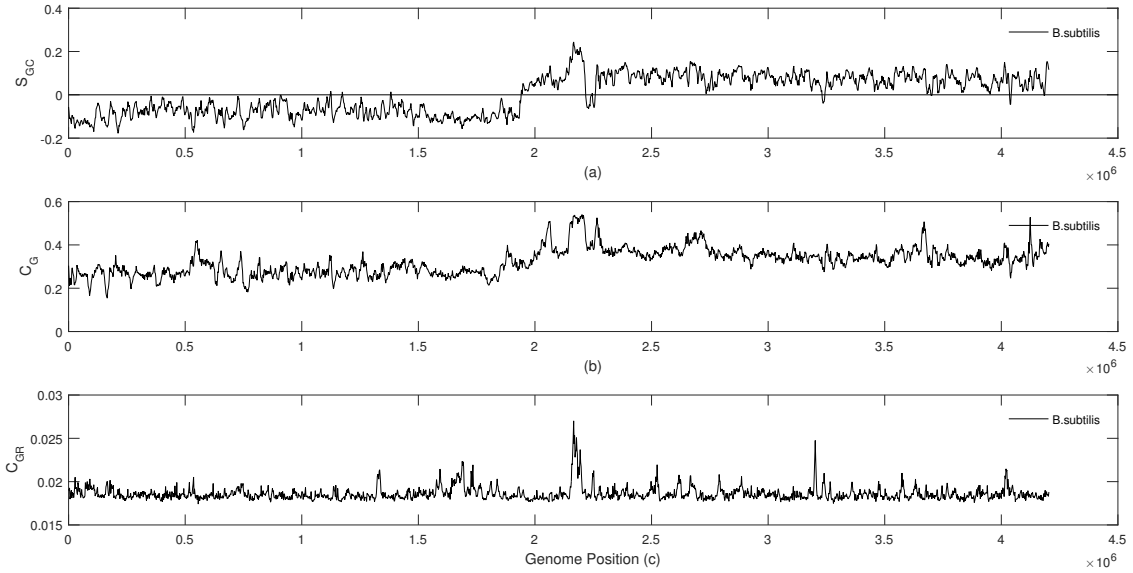

Figure 1: Plot of the (a) GC skew (S_{GC}), (b) gCorr (C_G) and (c) iCorr (C_{GR}) values for *B. subtilis* subsp. subtilis str. 168 (NC_000964). As can be seen, for *B. subtilis*, all three methods predict the TER location at nearby genome positions. The window size is 10000 and shift size is 2000.

perimentally verified [23]. It has also been shown earlier that variations of the auto-correlation method are able to predict the origin location of several more genomes as compared to the GC-skew method [19]. However, in the auto-correlation method, currently there is no clear way to differentiate between ORI/TER and the predicted location could be either of the two for prokaryotic genomes. However, eukaryotic genomes are linear and do not have a separate TER location. Hence, the prediction of auto-correlation method invariably corresponds to the ORI location in eukaryotes.

The auto-correlation method mainly has three limitations. Firstly, the ORI location is predicted in this method by visually inspecting the correlation profile which creates room for human error. Secondly, the window size required in this method is quite large, which becomes a problem for experimental validation. Thirdly, the auto-correlation method uses the positional information of only the G nucleotide and ignores the statistical properties of other nucleotides. In this paper, we propose a modification of this method which addresses all these limitations. The proposed complex correlation method (iCorr) uses four numbers $\{+1, -1, +i = \sqrt{-1}, -i\}$ and thus is able to represent the positions of each of the four nucleotides, unlike the auto-correlation method which uses only real numbers $\{+1, -1\}$. In the iCorr method, there is no need for visual inspection and the ORI/TER region is given by either the location of the peak value of the computed function. This method requires a much smaller window size as compared to the auto-correlation method and thus, leads to a much higher resolution. We have also developed an algorithm to optimize this resolution in order to get the best results with a fairly low window size. This mapping of DNA nucleotides to four unique complex numbers instead of two integers could also be very effective in solving many other problems of interest in computational biology [24–26].

We describe the iCorr method and the proposed algorithm in Sec. II, present the results in Sec. III and finally end with conclusions in Sec. IV.

II. METHODS

The primary computational approach for prediction of origin of replication is to divide the entire genome into overlapping windows/segments of equal length, and analyse each window to measure some statistical property using information theory and/or signal processing techniques. The values thus obtained are plotted against the window number. The origin of replication is predicted to be present in the window where a significant change is observed. This abrupt change can manifest in different ways depending on the actual statistical property being measured.

In the gCorr method, the G (Guanine) nucleotide of each location of the window/segment is denoted by $\{+1\}$ and all other nucleotides by $\{-1\}$. This helps in converting the symbolic sequence to a discrete number sequence thereby making it conducive for statistical analysis. We calculate the auto-correlation value of this discrete se-

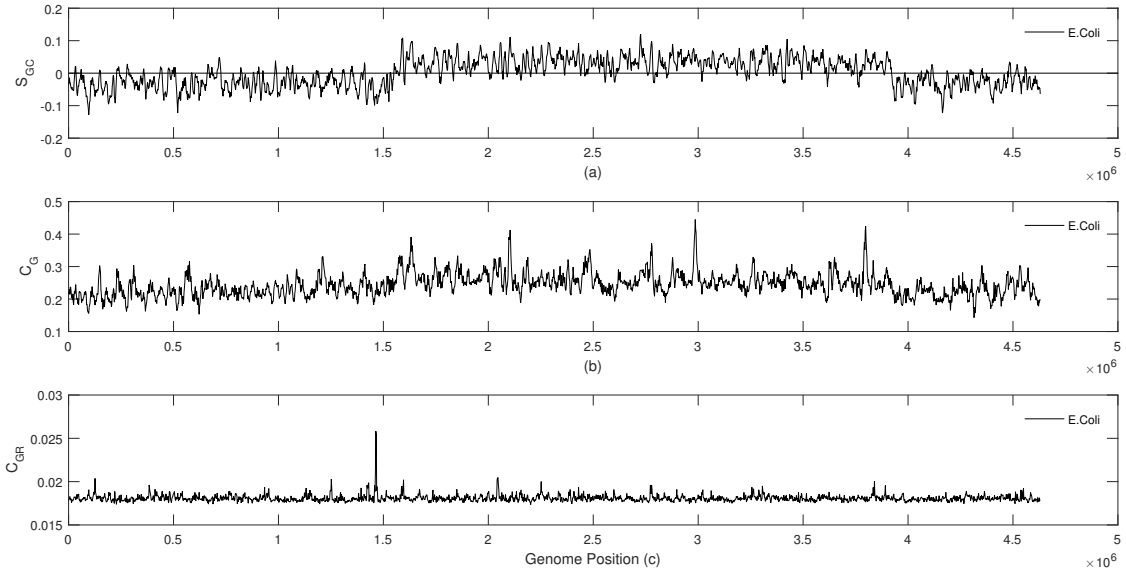

Figure 2: Plot of the (a) GC skew (S_{GC}), (b) gCorr (C_G) and (c) iCorr (C_{GR}) values for *E. coli* str. K-12 substr. MG1655 (NC_000913). As can be seen, for *E. coli*, all three methods predict the TER location at nearby genome positions which also matches with the experimentally known TER location. However, the graph of iCorr is lot less noisy as compared to GC skew and gCorr, thereby substantially reducing the ambiguity. The window size is 10000 and shift size is 2000.

quence using the function [27, 28],

$$C(k) = \frac{1}{(N-k)\sigma^2} \sum_{j=1}^{N-k} (a_j - \mu_a)(a_{j+k} - \mu_a) \qquad (2)$$

where $k = 1, 2, 3, \ldots, N$, $a_i \in \{+1, -1\}$ denotes the value at the ith position of the discrete sequence, N is the window size, $\mu_a = 0$ and $\sigma = 1$ are the means and standard deviation of the random variable a_i. The auto-correlation measure, C_G, is then defined as the average of all correlation values in Eq. (2) [18],

$$C_G = \frac{1}{N-1} \sum_{k=1}^{N-1} |C(k)| \qquad (3)$$

where the subscript "G" refers to "genome". C_G ranges from 0 to 1 and is independent of the length of the sequence. The value of C_G is a good indicator of the correlation strength between the positions of the G nucleotide. Thus, a sequence with $C_G = 0$ corresponds to a lack of correlation and one with $C_G = 1$ to a highly correlated sequence.

Since a DNA sequence is made up of four bases, we can generate a string of bits for the A (Adenine) base by assigning a value of $\{+1\}$ to every occurrence of A and $\{-1\}$ to all other positions (similarly for T and C). In the above method, only the G-track is chosen for analysis since it gives much better results as compared to the other three discrete sequences [18]. Though this method has been found to work better than the GC-skew method in some situations, it has an inherent limitation of assigning the same value of $\{-1\}$ to T, A and C. Due to this, it does not capture the rich variations produced by the four bases present in DNA sequence.

In this paper, we propose a complex correlation (iCorr) method which extends the above method to complex states and thereby completely eliminates the most fundamental limitation in gCorr and other computational methods for ORI prediction. Furthermore, we propose an algorithm for automatically optimising the window size and shift size for a given genome. We use $\{+1, -1, +i = \sqrt{-1}, -i\}$ for multi-variate classification of the four bases present in a DNA sequence. A DNA sequence made up of ATGC base pairs can give rise to 24 different discrete sequences using the $\{+1, -1, +i = \sqrt{-1}, -i\}$ mapping as opposed to only 4 sequences provided by the $\{+1, -1\}$ mapping. Out of these 24 sequences, only 3 were found to be independent, namely, $\{A \to +i,\ G \to -1,\ T \to -i,\ C \to +1\}$, $\{A \to -i,\ G \to +i,\ T \to -1,\ C \to 1\}$ and $\{A \to +i,\ G \to +1,\ T \to -i,\ C \to -1\}$. The remaining 21 are equivalent to one of these 3 mappings. After analysing all these various complex number sequences for ORI prediction, we have found that the following mapping gives the best results: $\{A \to +i,\ G \to +1,\ T \to -i,\ C \to -1\}$. The signals obtained from the other 2 choices are very noisy. We calculate the auto-correlation of this generated discrete sequence using the same formula given in Eqs. (2) and

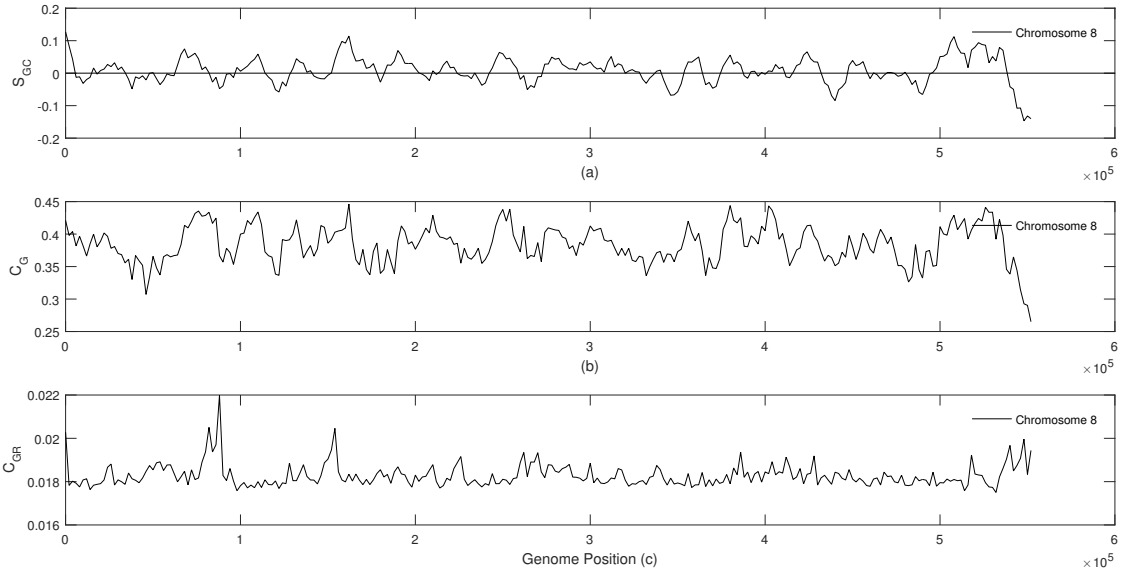

Figure 3: Plot of the (a) GC skew (S_{GC}), (b) gCorr (C_G) and (c) iCorr (C_{GR}) values for chromosome 8 of *S. cerevisiae*. As can be seen, for this chromosome, the graphs of GC skew and gCorr are very noisy and do not make any clear predictions. On the other hand, the iCorr method gives clear peaks which are close to the experimentally known ORI locations. The window size is 10000 and shift size is 2000.

(3) (using $\mu_a = 0,\ \sigma = 1$), but now $C(k)$ comes out to be a complex number. However, the final auto-correlation value obtained is still a real number since the RHS of Eq. (3) uses the absolute value (or magnitude) of these complex $C\,(k)$ values. This is the iCorr value (denoted by C_{GR}) and is plotted against the window number (or genome length). The graph produces a sharp peak at one or more locations. We propose that the genome position(s) corresponding to these peak value(s) contain the origin of replication or the termination position (applicable to prokaryotes which have circular genomes).

Further, we develop a algorithm with the primary objective of allowing the user to be able to get the best possible results without the need to set the window size and shift size beforehand. The algorithm is initially run on 10 different sets of window sizes and shift sizes, namely: (1000,1000), (2000,1000), (3000,1000), (4000,1000), (5000,1000), (1000,500), (2000,500), (3000,500), (4000,500) and (5000,500) to obtain the iCorr curve. For each of the iCorr values obtained, we perform normalisation of the values obtained given by using the formula:

$$X_i = \frac{x_i - \mu}{\sigma} \tag{4}$$

where $i \in \{1, 2, ..., N\}$ is the window number, N is the total number of windows, x_i is the iCorr value for the ith window, μ is the mean of all N iCorr values and σ is the standard deviation. Out of all the N values of $\{X_i\}$, the set $\{Y_j\}$ contains the values which represent the peaks corresponding to the replication sites along the genome. Elements of $\{Y_j\}$ are all those elements of $\{X_i\}$ which lie between X_L and X_H, where X_H is the maximum among all $\{X_i\}$ and X_L is calculated as follows:

$$X_L = \begin{cases} \max\left(X_H/3, 6\right), & \text{if } X_H > 10.5 \\ 6, & \text{if } 10.5 \geq X_H \geq 6 \\ X_H, & \text{if } X_H < 6 \end{cases}$$

which implies that if $X_H < 6$, then $\{Y_i\}$ is an empty set and our algorithm cannot predict any replication site locations along the genome. The numbers above were chosen based on visual inspection of the resulting data.

We carry out this procedure for all the 10 sets of window sizes and shift sizes mentioned above, and out of the 10 sets of $\{Y_i\}$ thus formed containing the topmost peaks, we select one set by comparing their X_H and X_L values with each other. If there is one non-empty set of $\{Y_i\}$ values with $X_H > 18$, we select this one to represent the replication sites. If there are more than one sets in this category, we select the set with the lowest value of X_H/X_L. This would ensure that the final window size and shift size values correspond to the graph, all of whose peaks are possibly distinguisable in terms of amplitude. If there are no such sets, we then check for non-empty sets with $18 \geq X_H > 10.5$ and if no set is found, then we check for $10.5 \geq X_H \geq 6$.

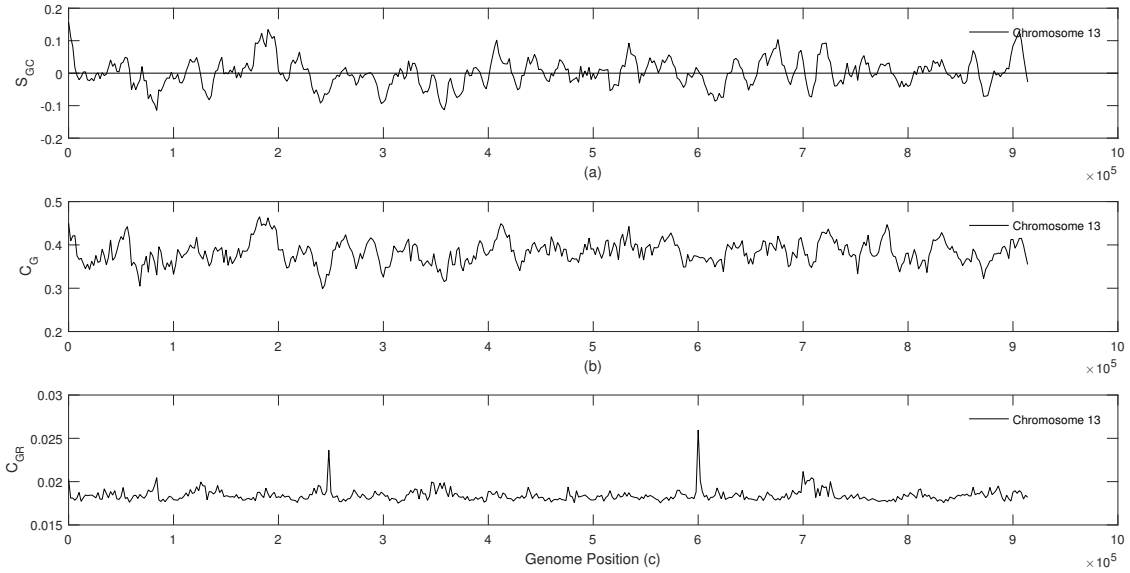

Figure 4: Plot of the (a) GC skew (S_{GC}), (b) gCorr (C_G) and (c) iCorr (C_{GR}) values for chromosome 13 of *S. cerevisiae*. As can be seen, for this chromosome, the graphs of GC-skew and gCorr are very noisy and do not make any clear predictions. On the other hand, the iCorr method gives clear peaks which are close to the experimentally known ORI locations. The window size is 10000 and shift size is 2000.

III. RESULTS

We have applied the method described in the previous section to two bacterial genomes obtained from NCBI [29] and 16 chromosomes of one eukaryote (*S. cerevisiae*) obtained from OriDB [30]. In this section, we describe the results obtained.

Figures 1 and 2 show a plot of the (a) GC skew (S_{GC}), (b) gCorr (C_G) and (c) iCorr (C_{GR}) values for two prokaryotic genomes, *B. subtilis* and *E. coli*, respectively. In GC skew method, the ORI (or TER) location is given by the point where the graph crosses the zero line from above (or below). In gCorr method, the ORI/TER location is given by the position where the graph undergoes a sharp jump (higher or lower). In iCorr method, the ORI/TER location is given by the position of peak values. As can be clearly seen, all these three methods correctly predict the TER location for *B. subtilis* and *E. coli*. However, the graphs for GC skew and gCorr are lot more noisy as compared to the graph for the iCorr method. In Fig. 1a and 2a, there are also several other points of zero-crossing which can be erroneously considered to the ORI/TER location thereby making the GC skew prediction quite ambiguous. And this problem only becomes worse as we further reduce the window size. The gCorr method predicts the presence of ORI/TER in a genome where a sudden transition is observed. The transition spans several windows and its detection depends on human judgement which reduces the accuracy in ORI prediction. In contrast, the iCorr method for prokaryotes predicts the location by finding peak in the graph. Peak is obtained at a single point which helps to narrow down our area of interest to a single window. In the case of *B. subtilis*, the gCorr predicts the ORI to be present in a genome segment whose length is around 100k nucleotides (see Fig.1b). In contrast, the iCorr method can bring down the range to 10k nucleotides which implies a 10 times higher resolution.

Compared to prokaryotic genomes, the computational prediction of ORI in eukaryotic genomes has been considerably much more challenging due to the rich and complex structure of DNA with multiple ORI being present in a single chromosome. And an added disadvantage is that experimentally verified ORI locations are available for only a few eukaryotes like *S. cerevisiae*. Figures 3 and 4 show the graph of the various methods for chromosome 8 and 13 of S. *cerevisiae* respectively. As can be seen, the plot of the GC skew and gCorr method in Figs. 3 and 4 is very noisy and thus, does not help in making any clear prediction. On the other hand, the iCorr method gives clear peaks which are actually close to the experimentally known ORI locations. Note that, unlike the prokaryotic genome which are circular, eukaryotic genomes are linear and do not have a separate TER region. Hence, its not clear whether all zeros of the GC skew plot correspond to ORI or only those where the graph crosses zero from above.

Details of the ORI prediction by iCorr method for all chromosomes of *S. cerevisiae* and the closest experimentally confirmed ORI are given in Table I. Column 1 of this table denotes the chromosome number. Column 2 de-

Chromosome Number	Window Size, Shift Size	ORI prediction of iCorr method	OriDB
chr01	5000,500	Window Number= 407 Region: 2,03,500-2,08,500 X_i: 7.51	Likely
chr02	1000,500	Window Number= 4 Region: 2,000-3,000 X_i: 29.93	Confirmed
chr03	2000,500	Window Number= 540 Region: 2,70,000-272000 X_i: 10.30	Confirmed
chr04	1000,1000	Window Number= 1309 Region: 13,09,000-13,10,000 X_i: 10.38	Likely
		Window Number= 350 Region: 3,50,000-3,51,000 X_i: 9.67	Confirmed
		Window Number= 1188 Region: 11,88,000-11,89,000 X_i: 8.33	Nothing
		Window Number= 384 Region: 3,84,000-3,85,000 X_i: 8.08	Dubious
		Window Number= 1427 Region: 14,27,000-14,28,000 X_i: 6.98	Dubious
		Window Number= 1391 Region: 13,91,000-13,92,000 X_i: 6.73	Confirmed
		Window Number= 442 Region: 4,42,000-4,43,000 X_i: 6.27	Likely
chr05	1000,1000	Window Number= 543 Region: 5,43,000-5,44,000 X_i: 11.72	Nothing
		Window Number= 2 Region: 2,000-3,000 X_i: 7.45	Confirmed
		Window Number= 171 Region: 1,71,000-1,72,000 X_i: 7.06	Confirmed
chr06	3000,500	Window Number= 356 Region: 1,78,000-1,81,000 X_i: 13.21	Nothing
chr07	1000,500	Window Number= 1504 Region: 7,52,000-7,53,000 X_i: 10.38	Dubious
		Window Number= 1614 Region: 8,07,000-8,08,000 X_i: 9.81	Likely
		Window Number= 1060 Region: 5,30,000-5,31,000 X_i: 9.01	Likely
		Window Number= 2178 Region: 10,89,000-10,90,000X_i: 8.79	Confirmed
		Window Number= 2103 Region: 10,51,500-10,52,500 X_i: 7.13	Nothing
		Window Number= 1442 Region: 7,21,000-7,22,000X_i: 6.81	Nothing
chr08	1000,1000	Window Number= 557 Region: 5,57,000-5,58,000 X_i: 16.93	Confirmed
chr09	4000,500	Window Number= 128 Region: 64,000-68,000X_i: 10.13	Nothing
		Window Number= 779 Region: 3,89,500-3,93,500X_i: 9.43	Nothing
chr10	1000,1000	Window Number= 715 Region: 7,15,000-7,16,000 X_i: 10.64	Confirmed
		Window Number= 3 Region: 3,000-4,000 X_i: 7.96	Confirmed
chr11	5000,1000	Window Number= 621 Region: 6,21,000-6,26,000 X_i: 14.25	Dubious
chr12	3000,1000	Window Number= 750 Region: 7,50,000-7,53,000X_i: 10.63	Dubious
		Window Number= 304 Region: 3,04,000-3,07,000 X_i: 9.72	Nothing
		Window Number= 104 Region: 1,04,000-1,07,000X_i: 6.79	Nothing
		Window Number= 123 Region: 1,23,000-1,26,000 X_i: 6.36	Nothing
chr13	1000,1000	Window Number= 410 Region: 4,10,000-4,11,000 X_i: 11.37	Nothing
		Window Number= 610 Region: 6,10,000-6,11,000 X_i: 10.45	Confirmed
		Window Number= 85 Region: 85,000-86,000 X_i: 7.24	Nothing
chr14	1000,1000	Window Number= 455 Region: 4,55,000-4,56,000X_i: 10.50	Dubious
		Window Number= 55 Region: 55,000-56,000 X_i: 7.48	Nothing
		Window Number= 705 Region: 7,05,000-7,06,000 X_i: 7.19	Dubious
chr15	3000,1000	Window Number= 30 Region: 30,000-33,000 X_i : 11.80	Confirmed
		Window Number= 345 Region: 3,45,000-3,48,000 X_i: 11.21	Likely
		Window Number= 895 Region: 8,95,000-8,98,000 X_i: 8.29	Likely
		Window Number= 236 Region: 2,36,000-2,39,000 X_i: 7.14	Dubious
		Window Number= 713 Region: 7,13,000-7,16,000 X_i: 6.39	Nothing
chr16	2000,1000	Window Number= 818 Region: 8,18,000-8,20,000 X_i: 11.83	Confirmed
		Window Number= 522 Region: 5,22,000-5,24,000 X_i: 8.51	Nothing

Table I: Results of iCorr optimisation algorithm applied to 16 chromosomes of *S. cerevisiae* and its comparison with experimental results.

notes the optimized window size and shift size given by our algorithm described in the previous section. Column 3 denotes the window number and region along the chromosome at which a peak was detected, and also the normalised value of the peak. Column 4 denotes whether that detected peak corresponds to a Confirmed, Likely or Dubious, experimental detection of ORI as given in the OriDB database [30]. Mention of Nothing in this column denotes that this genome location does not correspond to any experimentally detected ORI location. To compare our predictions with experiments, we took a region flanked by 5000 base pairs on both sides of our computational prediction. As can be seen in this table, the algorithm was able to predict at least one experimentally detected ORI location (Confirmed or Likely) for 11 out of the 16 chromosomes.

It is important to note that our iCorr method predicts only a few ORI locations in *S. cerevisiae* chromosomes which actually contain many experimentally verified ORI locations [30]. This clearly indicates that different ORI

locations in eukaryotes have different statistical properties which might require more sophisticated computational methods for their correct identification. Also, not all ORI locations may be predictable by using the same statistical measure. This holds for prokaryotes as well, where the ORI locations of different genomes can have different statistical properties, thereby requiring different computational tools for ORI prediction.

IV. DISCUSSION

In the past, several methods have been developed to predict ORI location for prokaryotes but most of them utilised only a limited amount of information present in the DNA sequence. The GC skew method [9] considered frequency counts of G and C nucleotides as the sole means to predict ORI location and neglected the importance of positioning of each base in a DNA sequence. The autocorrelation based gCorr method was developed to remove this inherent flaw of GC skew method by considering relative base positions of the G nucleotide. However, this method was unable to differentiate between A, C and T nucleotides. In an attempt to fully discover the rich variety of bases present in a sequence, we have extended the basic gCorr method to complex states. The iCorr method presented in this paper takes into consideration the relative base positioning of all the four nucleotides. This method has been found to significantly improve the resolution of ORI prediction of prokaryotes and has also been able to predict the ORI locations of *S. cerevisiae* to a good extent. The prediction of iCorr method currently does not match with the experimentally verified ORI locations for *P. falciparum* [23], but we hope that we will be able to validate and refine our methods as more experimental data becomes available in the future for this and other genomes.

Similar to all the previously existing computational methods, iCorr only suggests the ORI/TER location and does not guarantee its existence, which needs to be experimentally verified. With the advantages of pin-point peak detection and utilisation of rich structure present in DNA, the iCorr method is a significant progress in ORI prediction for prokaryotes and eukaryotes. Here it is important to note that the predictions made by these computational methods are significantly dependent on the choice of window/segment size into which the genome is divided for statistical analysis. If the window size is taken to be too large, then the meaningfulness of the predictions obviously goes down. And if the window size is taken to be too small, the graphs can be very noise and lead to decrease in accuracy and precision. Optimization of the window size and shift size for a given genome is an open problem which we have tried to tackle in this paper with the help of our proposed algorithm.

Acknowledgments

We would like to thank Ayush Jain for helping us with some part of the analysis. A part of this work was done while KS was at the Indian Institute of Technology (IIT) Delhi.

[1] K. J. Marians. "Prokaryotic DNA replication", *Annual Review of Biochemistry* 61, 673 (1992)
[2] M. L. Mott and J. M. Berger. "DNA replication initiation: mechanisms and regulation in bacteria", *Nature Reviews Microbiology* 5, 343 (2007)
[3] E. P. Rocha, A. Danchin and A. Viari. "Universal replication biases in bacteria", *Molecular Microbiology* 32, 11 (1999)
[4] L. M. Kelman and Z. Kelman. "Multiple origins of replication in archaea", *Trends in microbiology* 12, 399 (2004)
[5] H. P. Nasheuer et al. "Initiation of eukaryotic DNA replication: regulation and mechanisms", *Prog. Nucleic Acid Res. Mol. Biol.* 72, 41 (2002)
[6] N. V. Sernova and M. S. Gelfand. "Identification of replication origins in prokaryotic genomes", *Briefings in Bioinformatics* 9, 376 (2008)
[7] J. R. Lobry. "Asymmetric substitution patterns in the two DNA strands of bacteria", *Molecular Biology and Evolution* 13, 660 (1996)
[8] J. R. Lobry. "Origin of replication of Mycoplasma genitalium", *Science* 272, 745 (1996)
[9] J. Mrazek and S. Karlin. "Strand compositional asymmetry in bacterial and large viral genomes", *Proc. Natl. Acad. Sci. USA* 95, 3720 (1998)
[10] M. Touchon, S. Nicolay, B. Audit, E. Brodie, Y. d'Aubenton-Carafa, A. Arneodo and C. Thermes. "Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins", *Proc. Natl. Acad. Sci. USA*, vol. 102, 9836 (2005)
[11] R. Zhang and C. T. Zhang. "Identification of replication origins in archaeal genomes based on the Z-curve method", *Archaea* 1, 335 (2005)
[12] A. Grigoriev. "Analyzing genomes with cumulative skew diagrams", *Nucleic Acids Research* 26, 2286 (1998)
[13] D. S. H. Chew, M. Y. Leung and K. P. Choi. "AT excursion: a new approach to predict replication origins in viral genomes by locating AT-rich regions", *BMC Bioinformatics* 8, 163 (2007)
[14] T. D. Schneider. "Information content of individual genetic sequences", *Journal of Theoretical Biology* 189, 427 (1997)
[15] T. D. Schneider. "A brief review of molecular information theory", *Nano Communication Networks* 1, 173 (2010)
[16] C. E. Shannon. "A mathematical theory of communication", *Bell Syst. Tech. J.* 27, 379 (1948)
[17] J. Song, A. Ware and S. L. Liu. "Wavelet to predict bacterial ori and ter: a tendency towards a physical balance", *BMC Genomics* 4, 17 (2003)
[18] K. Shah and A. Krishnamachari. "Nucleotide correlation based measure for identifying origin of replication in ge-

nomic sequences", *Biosystems* 107, 52 (2012)

[19] H. Parikh, A. Singh, A. Krishnamachari and K. Shah. "Computational prediction of origin of replication in bacterial genomes using correlated entropy measure (CEM)", *Biosystems* 128, 19 (2015)

[20] W. C. Li et al. "Sequence analysis of origins of replication in the Saccharomyces cerevisiae genomes", *Frontiers in Microbiology* 5, 574 (2014)

[21] J. R. Lobry and N. Sueoka. "Asymmetric directional mutation pressures in bacteria", *Genome Biology* 3, research0058.1-14, (2002)

[22] P. Mackiewicz, J. Zakrzewska-Czerwinska, A. Zawilak, M. R. Dudek and S. Cebrat. "Where does bacterial replication start? Rules for predicting the oriC region", *Nucleic Acids Research* 32, 3781 (2004)

[23] M. Agarwal, K. Bhowmick, K. Shah, A. Krishnamachari and S. K. Dhar. "Identification and Characterization of ARS like sequences as putative origin(s) of replication in human malaria parasite Plasmodium falciparum", *FEBS Journal* 284, 2674 (2017)

[24] K. Shah and A. Krishnamachari, "On the origin of three base periodicity in genomes", *BioSystems* 107,142 (2012)

[25] T. Hoang, C. Yin, H. Zheng, C. Yu, R. L. He, S. S.-T. Yau, "A new method to cluster DNA sequences using Fourier power spectrum", *Journal of Theoretical Biology* 372, 135 (2015)

[26] G. Mendizabal-Ruiz, I. Romen-Godenez, S. Torres-Ramos, R. A. Salido-Ruiz, J. A. Morales, "On DNA numerical representations for genomic similarity computation", *PLoS ONE* 12, e0173288 (2017)

[27] K. G. Beauchamp and C. K. Yuen, *Digital methods for signal analysis* (George Allen and Unwin, London, 1979)

[28] T. J. Cavicchi, *Digital Signal Processing* (John Wiley & Sons, New York, 2000)

[29] NCBI, Resource Coordinators. "Database resources of the National Center for Biotechnology Information", *Nucleic Acids Research* 44, D7-19 (2016)

[30] C. C. Siow et al. "OriDB, the DNA replication origin database updated and extended", *Nucleic Acids Research* 40 (Database issue), D682-D686 (2012)