ORIS: An interactive software tool for prediction of replication origin in prokaryotic genomes

Urminder Singh[1], Kushal Shah[2], Suman Dhar[3], Vinod Kumar Singh[1], Annangarachari Krishnamachari[1*],
1 School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India
2 Department of Electrical Engineering and Bharti School of Telecommunication Technology and Management, IIT Delhi, India
3 Special centre for Molecular Medicine, Jawaharlal Nehru University, New Delhi, India
*** E-mail: Co-corresponding authors : AK - chari@mail.jnu.ac.in, KS - kkshah@ee.iitd.ac.in**

Methods implemented in ORIS

This document explains the various computational methods implemented in ORIS. These methods can either be applied to the entire genome sequence or at a time or by dividing the genome into smaller windows. A windowing analysis system divides the whole sequence into fixed sized windows and then proceeds with the analysis on each of the window. Generally, sliding window technique is used in which a fixed sized window slides over the whole sequence and the sequence it covers at any time is analysed. The window moves with a fixed increment value, which is less than the size of the window. This style of genomic analysis has been already demonstrated to be very useful in origin finding [1, 2].

GC-skew and AT-skew method

In most of the bacterial genomes, during replication the leading and lagging strands are subjected to mutational pressures [2–4] and that results in asymmetry in base composition. This asymmetry can be easily captured by taking the ratio of $(C-G)/(C+G)$. A sliding window of the size of few kilo-bases is used. C and G means numbers of C's and G's respectively in that window. After calculating the ratio $(C-G)/(C+G)$, a graph of $S=(C-G)/(C+G)$ vs. window number can be plotted to analyse the results. The origin of replication could then be predicted around the position where S undergoes an abrupt transition across $S=0$.

The AT-skew method is similar to GC-skew. The only difference is that here the ratio$S=(A-T)/(T+A)$ is taken. The procedure and interpretation of results is same as that of GC-Skew. Although, from the perspective of predicting origin sites, AT-skew is not of much use in most of the genomes but, for some genomes it might be helpful. We have also provided a module which computes MK-skew as well.It is upto the user to pick and choose the appropriate measure deemed fit for the genome in question.

Cummulative Skew

A cumulative skew score is calculated by adding the skew values for each window, starting from 1st window till the last [5].

User-Defined Skew

Some genomes are GC rich and some are AT rich and this module will help biologists to custom build a new skew measure, e.g. MK-skew, RY-skew, to suit their needs. This module make use of the counts of the bases and to the best of our knowledge it is a new tool for the users.

Auto-correlation and cross-correlation Measure

The symbolic DNA sequence can be mapped to a binary sequence by assigning $\{1, 0\}$ or $\{+1, -1\}$ to each nucleotide base. A choice of any other number leads to numerical interference. Further, in this one dimensional symbolic sequence each base position is treated as a discrete time step. Hence, one can easily convert any DNA sequence into binary sequence but with strict preservation of the order. It is now possible to apply powerful DSP based algorithms to extract the embedded information from the signal sequence i.e. binary sequence.

While mapping DNA sequence to a binary sequence for the purpose of calculating the auto-correlation or cross-correlation, the following procedure is adopted: $+1$ corresponds to the presence of particular base(s) and -1 corresponds to absence of that base(s).

The auto correlation function$C(k)$, of discrete sequence, $\{a_i : i = 1, 2, ...N\}$with $a_i \in \{+1, -1\}$ is defined as [1, 6, 7]

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

The correlation measure [1] can now be defined as the average of all auto-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 Data'. The C_G measure can take values from 0 to 1. Completely correlated sequence give unity while a truly random sequence will give zero. It is possible to use our tools to compute the correlation value C_G for any nucleotide i.e A, T, G or C. It has been found earlier that the C_G measure with respect to the G residue usually gives better results, but ORIS supports the calculation of C_G for any residue. When C_G is plotted against the window number, the position of replication origin is given by a sharp peak or dip in the graph. It must be noted that the summation in Eq. (2) is computationally intensive and for the purpose of finding replication origin, taking a sum up to $k = 10$ suffices.

Cross-correlation method is based on the auto-correlation function and is defined when we have two different discrete sequences, $\{a_i : i = 1, 2, ...N\}$ and $\{b_i : i = 1, 2, ...N\}$.

$$C(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} a_j b_{j+k} \tag{3}$$

These two discrete sequences can be generated either from the same genome sequence or from two different genome sequence. In ORIS, we have implemented only the case where $\{a_i\}$ and $\{b_i\}$ are generated by the same genome sequence. For example, $\{a_i\}$ could be generated by mapping G to $\{+1\}$ and $\{b_i\}$ could be generated by mapping C

to {+1} resulting in G-C cross-correlation. The cross-correlation measure C_{GG} can now be defined as the average of all correlation values in Eq. (3),

$$C_{GG} = \frac{1}{N-1} \sum_{k=1}^{N-1} |C(k)| \tag{4}$$

Z-curve Method

We implemented this module as described in Zhang et al [8]. The Z-curve plot was proposed to show multiple origins embedded in the genomic data [9, 10]. This measure is extensively used for archaeal genome data and considered to be more informative than GC skew measure [11].

Pattern Search

DNA sequence motifs are elucidated by researchers after carrying out precise and careful experiments [12]. These sequence patterns may be strict or loose depending on the degree of conservation. Searching with consensus sequence or using a position weight matrix is a routine procedure or strategy to identify more examples of that class. For example, one can search for the DnaA boxes for predicting origin of replication in bacterial genomic data [13]. Further the search may have constraints such that results can have mismatches at few positions.

In the context of eukaryotic model organism *S. cerevisiae*, a motif called ACS (ARS consensus sequence) has been found to be limited to origin like sites. It is a 11 base-pair long motif and is present abundantly in the genome of S. cerevisiae. Researchers uses this motif as a seed to search for origin of replication in other species. In addition a separate module is included to enable the user to key in or develop a weight matrix of a defined length for searching patterns in the genome sequence data.

Shannon and Renyi entropy

Information theory quantifies uncertainty in data. Intuitively, if an event has two possible outcomes with equal probability, then the uncertainty is quite high. If the probability of one of the outcomes is much higher than the other, then the amount of uncertainty is quite low [14]. Entropy is a measure of this uncertainty and is linked to the amount of information. Shannon entropy [14] is the measure of average information of an outcome and gets maximized when all possible outcomes of an event are equally likely. Mathematically Shannon's entropy $H(X)$ is given by

$$H(X) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{5}$$

This entropy measure gives an idea about the sequence variability.

A generalized form of Shannon's entropy was proposed by Alfred Renyi [15]. It is mostly used in quantifying diversity, uncertainty or randomness of a system. For an order α, it is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} log(\sum_{i=1}^{n} p_i^\alpha) \tag{6}$$

Redundancy plot

A measure called Redundancy(R) is limited to information content and takes value from 0 to 1. R becomes unity when there is complete conservation and zero when all bases

are equally likely. The Redundancy is given by

$$R = 1 - \frac{H(X)}{H_{max}} \tag{7}$$

In case of DNA $H_{max} = 2$ hence,

$$R = 1 - \frac{H(X)}{2} \tag{8}$$

Composition and Statistics

Investigating the compositions of a genomic sequence is also very helpful to biologists. For example, how the composition of the four bases varies with the length of the sequence length? What is the percentage composition for a chosen base at a chosen region? How composition of group of bases, i.e. aminos, ketos, purines, pyrimidines etc. vary in the sequence? Answers to such questions may reveal interesting features about the genome organisation on a macro scale. These statistics can be exploited to identify biologically useful features.

DNA Bending Analysis

The origin of replication may also be identified by analysing the structural characteristics of the DNA [16]. One such structural feature is DNA bending or bendability. The non parallel characteristics of consecutive base pairs in DNA is represented by DNA bendability [17]. The bendability parameter is defined for three consecutive base pairs. In the context of origin of replication, the bendability changes abruptly in the origin of replication region in some genomes.

Automatic Prediction

ORIS is capable of finding putative origins of replication sequences in whole genome data. ORIS uses the above mentioned methods along with an annotation file to find and report coordinates of possible origin of replication. The annotation file can be .ptt file for annotated genomes or an output file produced by GLIMMER software [18] if the genome is unannotated. Though, ORIS automatically determines parameters like window size and increment, it gives some flexibility to the user by taking additional parameters, for the analysis, from the user. The user can choose a species specific or enter his/her own dnaA box motif to search. Each putative origin of replication identified by ORIS is then searched for this dnaA box motif and its reverse complement. User can specify a minimum number of dnaA box motifs required in order to find the putative origin of replication sequence.

Case Study

Experimental investigations about biological processes and organisms are carried out at the molecular level to get deeper insights and understanding of the problem in question.NextGen sequencing and other high throughput technologies are generating massive amount of gene, transcript,and genomic data and provides a golden opportunity for computational biologists to make useful inferences and novel findings. Hence, theoretical (or) computational approaches are making extensive use of genomic data to decipher the embedded statistical regularities, patterns and organizational makeup of the organism under investigation. Analysis are routinely carried out on a large scale

taking genomic data as a starting point and further extended to unravel the hidden information in the long one dimensional symbolic sequences.

Diversity gets manifested in the genetic make up of the organism and the exploitation of their genomic data may provide vital clues about biological features and functional sites. Genomes of species can be GC rich or AT rich and further they may be circular or linear. Thus complexity arises due to genome organisation, compositional heterogeneity and diversity factors. For a long time, GC skew method has been used as the de-facto method for computational prediction of replication origins in bacterial genomes. These skew measuress when plotted shows abrupt changes in its values or marked fluctuations at the point of replication origin. It was shown that there are some very important bacterial-like DNA sequences where the GC skew method does not work and an alternate method based on the auto-correlation function was proposed [1]. Though both these methods work for many bacterial genomes but, they are not sufficient. Still there exists a scope for new measures especially when we consider archea and eukaryotic genomes as well.

Hence a need to provide biologists a computaional tool which offer an ensemble of measures to help them in identifying putative origin sites. Keeping this in mind a software tool named "ORIS" has been developed and most of the useful measures mentioned in software sections were implemented.

Fig. 1 shows the output of ORIS for the four measures i.e GC skew, Cumulative GC skew, Auto-correlation and Shannon's entropy with respect to the *E. Coli* genome.

It is clearly seen from the Fig. 1 that GC and cumulative GC skew measure distinctly shows the origin of replication of *E. Coli* i.e where the values make an abrupt transition. Auto-correlation and entropy measure shows a noisy picture and the region of interest is not distinctly seen. Thus for this *E. Coli* genome skew measures are appropriate and useful.One can improve the graphic plots by suitable choosing the window parameters.

ORIS has few more modules which can enable biologists to search for sequence elemnets like "dnaA" box and they can also construct a customised skew measure for a specific species. Further one can choose the desired segment and perform the analysis with a click of a button. In some circular bacterial genomes where the replication origin is located in the starting or at the end of the sequence, it is desirable to rotate the sequence in order to capture the origin site. For example, consider the sequence AAAGGTTCCTAAAT. After rotating it clockwise by 5 bp we will get the following sequence TAAATAAAGGTTCC. Fig. 2 shows the results of GC skew, Correlation and distribution of DnaA box(TTATCCACA) without mismatches on orignal and rotated sequence of *B. subtilis*. After rotating the sequence clockwise by 500000 bp, methods in ORIS are able to capture the correct origin of replication site. ORIS gives user the option of doing such analysis by allowing them to rotate any given sequence clockwise or anticlockwise. However, such rotations are meaningful only in case of circular genomes. Thus ORIS provides biologists a dedicated software tool rich with independent modules to compute the chosen measure.

Pattern search

DNA sequence motifs are elucidated by researchers after carrying out precise and careful experiments. These sequence patterns may be strict or loose depending on the degree of conservation. Searching with consensus sequence or using a position weight matrix is a routine procedure or strategy to identify more examples of that class or family. For example, one can search for the DnaA boxes for predicting origin of replication in bacterial genomic data. Further the search may have constraints such that results can have mismatches at few positions. Further a sequence logo can be generated from the search hits which shows the probability of occurrences of each base at the

corresponding positions. Using ORIS it is possible to perform motif search of a defined length with or without mismatches in the entire genome data.

Computational identification of origin like sites in eukaryotic organisms is very difficult as they have multiple origin sites embedded in their genome. However, in the context of eukaryotic model organism *S. Cerevisiae*, a motif called ACS (ARS consensus sequence) has been found to be linked to origin like sites. It is an 11 base-pair long motif and is present abundantly in the genome of *S. Cerevisiae*. Fig. 3 shows the ACS sequence logo generated by ORIS software by searching the chromosome 4 of *S. Cerevisiae*. Further a separate module is also included to enable the user to key in or develop a weight matrix of a defined length for searching purposes.

Comparative Evaluation

We have compared general features of ORIS with that of available software tools either online or offline for identifying origin sites. For bacterial origin prediction Oriloc [19] uses skew methods and DNA walk model. Orifinder [11] uses skew methods and Z-curve method and distribution of DnaA boxes to predict the origin of replication. Oris incorporates methods like auto-correlation, cross-correlation, pattern search, sequence logo generation, entropy measures along with the popular methods used in the above mentioned tools.

We have compared the performance of ORIS with Ori Finder using five bacterial genomes. Table 1 shows the predicted origin of replication for each bacterial genome, by both tools. We also compared our results with DORIC [20] database and found that results produced by ORIS are in accordance with the data given in the DORIC database. It is clear from the results that ORIS is not only comparable to Ori Finder, but performs better in some cases. ORIS is able to find more putative origins of replication by a number of methods which are not present in Ori Finder. Moreover, user can tune search parameters like minimum number of dnaA box matches, which dnaA box to search, which methods to use and a depth of search parameter, which defines the extent of search in the genome. Fig. 4 shows the results produced by ORIS' automatic prediction method.

It is evidently clear that ORIS is rich in providing many measures for the purpose of identifying probable origin like sites in genomic sequences with a better user friendly GUI. ORIS gives flexibility of analysis and thus it can also be used as a generic tool for basic whole genome analysis.

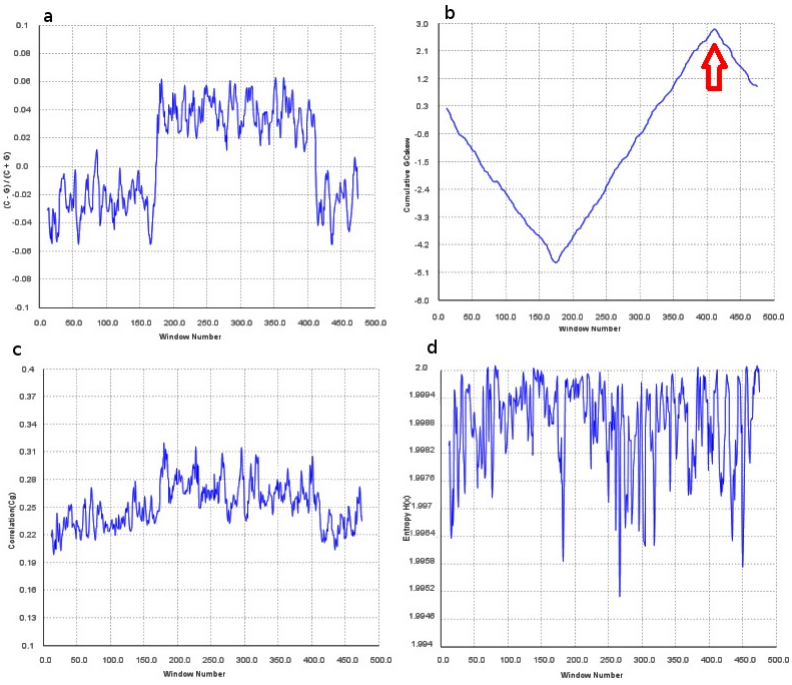

Figure 1. ORIS graphic plots for the four measures for the genome of *E. Coli* [GenBank:CP000948.1]. The window size chosen for analysis was 50k bp with increment of 10k bp. The x-axis depicts window number and the y axis represents the value of corresponding measure applied in each case. The region of interest is marked by an arrow. (a) GC skew. (b) Cummulative GC skew. (c) Correlation measure. (d) Shannon's Entropy

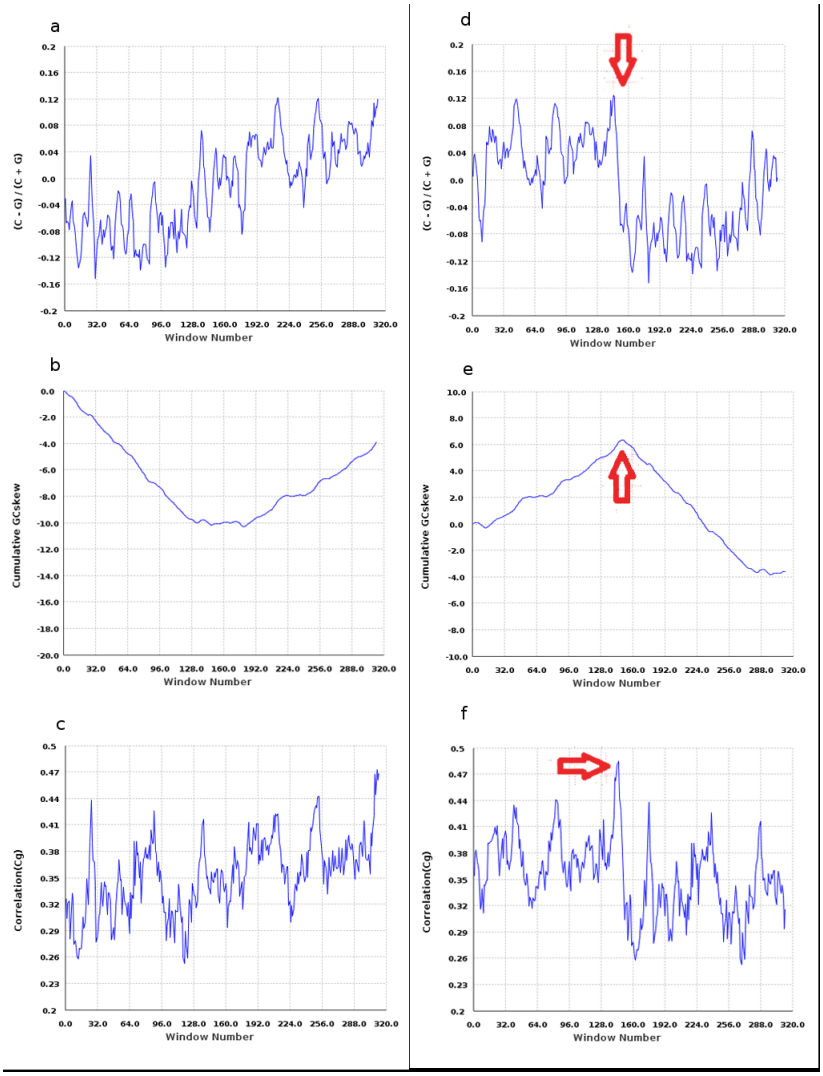

Figure 2. Results for the three methods before and after rotation of the genome sequence for the genome of *B. subtilis* [RefSeq:NC_000964.3]. The window size chosen for analysis was 50k bp with increment of 50k bp. The region of interest is marked by an arrow. (a) Cummulative GC skew. (b) Correlation Method. (c) Distribution of DnaA box(TTATCCACA) without mismatch. Then the sequence was rotated clockwise by 500000 bp. (d) Cummulative GC skew. (e) Correlation Method. (f) Distribution of DnaA box(TTATCCACA) without mismatch. The results after rotation gives the correct prediction about the replication origin in *B. subtilis*

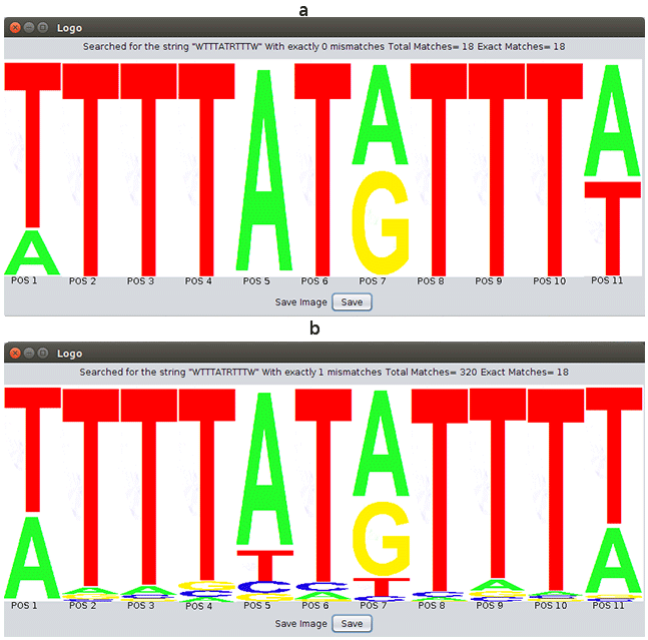

Figure 3. Sequence logo showing results of ACS pattern(WTTTAYRTTTW) search in the genome of *S. Cerevisiae* [GenBank:BK006935.2]. (a) ACS Pattern search with no mismatch. (b) ACS Pattern search with one mismatch

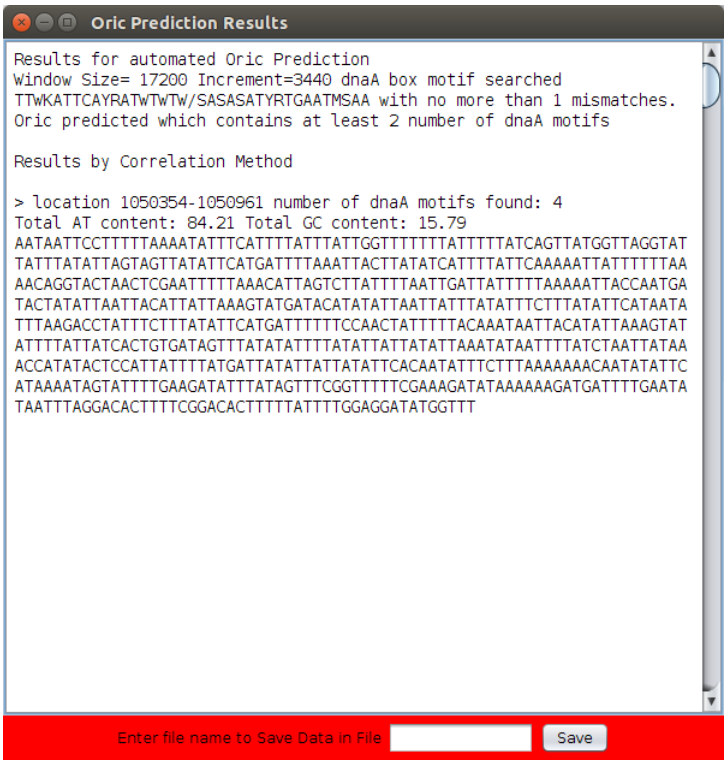

Figure 4. Results produced by ORIS' automatic prediction module for the genome of *B. subtilis* [RefSeq:NC_000964.3].

Table 1. Comparision of Ori Finder and ORIS results. We generated list of possible coding regions using GLIMMER v3.02 (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi) to show the applicability of ORIS to unannotated genome sequences. ORIS takes the GLIMMER output to automatically predict putative origin or replication sequences.

				Ori Finder		ORIS	
S. No	Name	Ref. Seq	dnaA box searched	Coordinates of Oric	# dnaA box matches	Coordinates of Oric	# dnaA box matches
1	*E. Coli*	CP000948	TTATCCACA	4021241..4021618	4	4021241..4021618	4
2	*Helicobacter pylori*	NC_000915.1	TCATTCACA	1608998..1609149	3	1608998..1609149	3
3	*Bacillus subtilis*	NC_000964.3	TTATCCACA	Gives Error	NA	1751..1938 4134916..4135350 4212890..4213199	3 5 4
4	*Deinococcus Radiodurans*	NC_001263.1	TTATCCACA	0..0	0	1183..1903 2645500..2645697 121731..121943	13 3 3
5	*Bacillus licheniformis*	NC_006322.1	TTATCCACA	1652..1827 4176153..4176301 4219817..4220150 4222098..4222577	3 3 4 8	1652..1827 4176153..4176232 4176392..4176589 60590..60893	3 3 2 2

References

1. Kushal Shah and Annangarachari Krishnamachari. Nucleotide correlation based measure for identifying origin of replication in genomic sequences. *BioSystems*, 107(1):52–55, 2012.

2. JAN Mrázek and Samuel Karlin. Strand compositional asymmetry in bacterial and large viral genomes. *Proceedings of the National Academy of Sciences*, 95(7):3720–3725, 1998.

3. JR Lobry. Origin of replication of mycoplasma genitalium. *Science*, 272(5262):745–746, 1996.

4. Jean R Lobry and Noboru Sueoka. Asymmetric directional mutation pressures in bacteria. *Genome biology*, 3(10):research0058–1, 2002.

5. Andrei Grigoriev. Analyzing genomes with cumulative skew diagrams. *Nucleic acids research*, 26(10):2286–2290, 1998.

6. Kenneth G Beauchamp and CK Yuen. *Digital methods for signal analysis*. Routledge, 1979.

7. Kushal Shah and Annangarachari Krishnamachari. On the origin of three base periodicity in genomes. *Biosystems*, 107(3):142–144, 2012.

8. Chun-Ting Zhang, Ren Zhang, and Hong-Yu Ou. The z curve database: a graphic representation of genome sequences. *Bioinformatics*, 19(5):593–599, 2003.

9. Ren Zhang and Chun-Ting Zhang. Identification of replication origins in the genome of the methanogenic archaeon, methanocaldococcus jannaschii. *Extremophiles*, 8(3):253–258, 2004.

10. Ren Zhang and Chun-Ting Zhang. Identification of replication origins in archaeal genomes based on the z-curve method. *Archaea*, 1(5):335–346, 2005.

11. Feng Gao and Chun-Ting Zhang. Ori-finder: a web-based system for finding oric s in unannotated bacterial genomes. *BMC bioinformatics*, 9(1):79, 2008.

12. L Slonczewski, W Foster, and M Gillen. Microbiology an evolving science‖ ww norton company. *Inc., New York*, 2009.

13. Melissa L Mott and James M Berger. Dna replication initiation: mechanisms and regulation in bacteria. *Nature Reviews Microbiology*, 5(5):343, 2007.

14. Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

15. Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

16. Ivan Brukner, Roberto Sanchez, Dietrich Suck, and Sandor Pongor. Sequence-dependent bending propensity of dna as revealed by dnase i: parameters for trinucleotides. *The EMBO journal*, 14(8):1812–1818, 1995.

17. Wei Chen, Pengmian Feng, and Hao Lin. Prediction of replication origins by calculating dna structural properties. *FEBS letters*, 586(6):934–938, 2012.

18. Arthur L Delcher, Douglas Harmon, Simon Kasif, Owen White, and Steven L Salzberg. Improved microbial gene identification with glimmer. *Nucleic acids research*, 27(23):4636–4641, 1999.

19. AC Frank and JR Lobry. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics*, 16(6):560–561, 2000.

20. Feng Gao, Hao Luo, and Chun-Ting Zhang. Doric 5.0: an updated database of oric regions in both bacterial and archaeal genomes. *Nucleic acids research*, 41(D1):D90–D93, 2012.