BioSystems 128 (2015) 19–25

Contents lists available at ScienceDirect

BioSystems

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

Computational prediction of origin of replication in bacterial genomes using correlated entropy measure (CEM)

Harsh Parikh [a], Apoorvi Singh [b], Annangarachari Krishnamachari [c], Kushal Shah [b,*]

[a] Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India
[b] Department of Electrical Engineering, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India
[c] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

ARTICLE INFO

Article history:
Received 29 November 2014
Received in revised form
31 December 2014
Accepted 1 January 2015
Available online 7 January 2015

Keywords:
Bioinformatics
Information theory
Signal processing
Cell cycle
Genomic signal processing

ABSTRACT

We have carried out an analysis on 500 bacterial genomes and found that the de-facto GC skew method could predict the replication origin site only for 376 genomes. We also found that the auto-correlation and cross-correlation based methods have a similar prediction performance. In this paper, we propose a new measure called correlated entropy measure (CEM) which is able to predict the replication origin of all these 500 bacterial genomes. The proposed measure is context sensitive and thus a promising tool to identify functional sites. The process of identifying replication origins from the output of CEM and other methods has been automated to analyze a large number of genomes in a faster manner. We have also explored the applicability of SVM based classification of the workability of each of these methods on all the 500 bacterial genomes based on its length and GC content.

1. Introduction

DNA replication is one the most fundamental process in the cell cycle and also the first step in cell division cycle (Leonard and Mechali, 2013; Sernova and Gelfand, 2008). It is a complex process and involves several protein molecules and complexes. Identification of the origin of replication in an organism's genome can be very helpful in several ways. Firstly, it can help a lot in understanding the statistical properties and organizational features of the DNA sequence of the organism. Secondly, it can also help in the discovery of new drugs for treatment of various diseases (McFadden and Roos, 1999; Soldati, 1999; Raghuram et al., 2007). Thirdly, in some cases like *Plasmodium falciparum* (causes malaria), discovery of the replication site can be very helpful in developing vectors for mass production of the parasitic cell which is very essential for further research studies.

All computational methods developed so far to predict replication origin rely on the fact that the leading and lagging strands are subject to different mutational pressures during the replication process, which leads to differences in the statistical properties of the DNA sequence on two sides of the replication site (Lobry and Sueoka, 2002). One of the first methods proposed to capture this asymmetry in bacterial genomes is known as the GC-skew method (Mrazek and Karlin, 1998; Lobry, 1996; Touchon, 2005). In this method, the entire genome is divided into windows and the value of GC-skew, $(C-G)/(G+C)$, is calculated for each window. The value of this skew is then plotted against the window number and the replication origin is believed to be at the window where the skew changes sign from positive to negative. The window size is usually taken to be 1/100th of the genome and each window is then shifted by about 1/5th of its size. Thus, the overlap between two windows is usually 4/5th of the window-size. This is, however, just a rough estimate and the actual choice of window size and shift may vary for each genome. A similar strategy making use of A and T bases may be useful in some cases (Touchon, 2005) but is not as widely used as GC skew. At least for the case of bacterial genomes, the GC skew measure is usually considered to be a de-facto computational method to identify origin of replication. However, as we will see in Section 3, the GC skew method fails to predict the replication origin for several bacterial genomes.

Another method that was proposed to predict the replication origin was based on the auto-correlation function (Shah and Krishnamachari, 2012). This method was found to be useful for both bacterial as well as eukaryotic genomes. When applied to a large set of bacterial genomes, this method has a success rate comparable to that of the GC skew method. In this paper, we propose

* Corresponding author. Tel.: +91 11 26591102.
E-mail addresses: harsh081193@gmail.com (H. Parikh), apoorvisingh@gmail.com (A. Singh), chari@mail.jnu.ac.in (A. Krishnamachari), kkshah@ee.iitd.ac.in (K. Shah).

http://dx.doi.org/10.1016/j.biosystems.2015.01.001
0303-2647/

a new method called correlated entropy measure (CEM) which is a combination of auto-correlation (Beauchamp and Yuen, 1979; Cavicchi, 2000) and Shannon entropy (Shannon, 1948; Schneider, 2010, 1997). We have carried out analysis of 500 diverse bacterial genomes (with widely varying GC content: 16–74%) and found that this new entropy-cum-correlation based measure is able to predict the origin location for all these genomes. We have also carried out a classification of all these bacterial genomes based on the workability of each of these methods to recognize patterns in bacterial genomes based on its length and GC content.

Thus, our efforts are in two parts. The first one focuses on the prediction of "putative" origin sites in the chosen genomes using five different methods. The second part attempts to see any classification possible using SVM with the results obtained in the first part.

2. Methods

As mentioned in Section 1, the primary method used for computational prediction of replication origin is to divide the entire genome into windows of equal lengths, calculate the value of some statistical property for each window and plot these values versus the window number. This overlapping window search procedure is a common strategy used in prediction experiments. The origin of replication is predicted to lie in the window where the value of this property changes sign or undergoes an abrupt change in its value. It is important to note that we are not claiming that our prediction will always match with experimental results. What we are offering is just a possible location for replication origin that needs to be further tested experimentally. Such a prediction is very useful since it is not experimentally feasible to carry out a search for replication origin sites over the entire length of the genome. By giving a much smaller search window, these computational methods decrease the experimental burden by a significant amount and reduce the search space.

2.1. GC skew method

The GC skew method is one of the most widely used methods for identification of origin of replication in bacterial genomes (Lobry, 1996, 1996; Mrazek and Karlin, 1998). For a given DNA sequence, the GC skew measure is given by

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

where n_C and n_G are the number of occurrences of cytosine (C) and guanine (G). In this method, the origin of replication is said to be at the position where S undergoes an abrupt transition from positive to negative.

2.2. Auto-correlation measure

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

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

where $\mu_a = 0$ is the mean of the random variable a_i and $\sigma = 1$ is the standard deviation. Using Eq. (1), the correlation measure, C_G, can now be defined as the average of all correlation values in Eq. (1),

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

where the subscript "G" refers to "genome". The value for C_G ranges from 0 to 1 and is independent of the length of the sequence. Lower value of C_G corresponds to lower correlation strength (nucleotides, in our case) in that sequence and vice-versa. For a typical random sequence, the value of C_G will be zero and will approach unity for a highly correlated sequence. It is important to note that the main idea behind taking the summation over all $C(k)$ is to capture the correlation strengths at all length-scales.

We need to convert the DNA sequence into a discrete sequence of bits in order to use Eq. (2). Since a DNA sequence is made up of four bases, we can generate a string of bits for the A base by assigning a value of +1 to every occurrence of A and −1 to all other bases (similarly for T, G, C). For example, a DNA sequence *AAGTTCAG* gives rise to four different discrete sequences $\{1, 1, -1, -1, -1, -1, 1, -1\}$, $\{-1, -1, -1, 1, 1, -1, -1, -1\}$, $\{-1, -1, 1, -1, -1, -1, -1, 1\}$ and, $\{-1, -1, -1, -1, -1, 1, -1, -1\}$ corresponding to the four bases A, T, G, C respectively. Thus, a given DNA sequence gives rise to four different bit strings and four different values of correlation strength (i.e. auto-correlation values, C_G) corresponding to each of the four bases, A, T, G, C. However, for the purpose of identifying the origin of replication, the correlation values of the Guanine residue (G) is good enough to obtain best results (Shah and Krishnamachari, 2012).

In this method, the origin of replication is considered to be at the window where the value of C_G undergoes an abrupt change.

2.3. Cross-correlation measure

The cross-correlation function can be defined in a similar manner to Eq. (1),

$$C_c(k) = \frac{1}{(N-k)\sigma_a\sigma_b} \sum_{j=1}^{N-k} (a_j - \mu_a)(b_{j+k} - \mu_b) \tag{3}$$

where $\{b_i : i = 1, 2, \ldots, N\}$ is a discrete sequence with $b_i \in \{+1, -1\}$, $\mu_b = 0 = \mu_a$ and $\sigma_a = 1 = \sigma_b$. And the cross-correlation measure is then defined as:

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

For our purpose, we take the a_i from locations of guanine residue and b_i from locations of cytosine.

2.4. Entropy measure

Entropy is a mathematical quantity widely used in both information theory (Shannon, 1948; Cover and Thomas, 1991) and statistical mechanics (Reif, 1965). Though the physical origin of this concept is quite different in these two domains, the final mathematical formula is exactly the same. If a discrete random variable takes on L values with probabilities p_i with $i = 1, 2, \ldots, L$, the entropy of this random variable is given by

$$H = -\sum_{i=1}^{L} p_i \log_2 p_i \tag{5}$$

There is nothing sacrosanct about the choice of the base of the logarithm in Eq. (5) and it can be chosen to be anything. However, it is usually chosen to be either 2 or e. If we choose base 2 for the logarithm, a binary sequence ($L = 2$) can have a maximum entropy of $H = 1$ and minimum $H = 0$.

For the case of DNA sequence, $L = 4$, and each value of p_i represents the frequency of occurrence of each nucleotide. For example, the DNA sequence *AAGTTCAG* used earlier has these values for p_i:

$$p_A = \frac{3}{8} \quad p_T = \frac{2}{8} \quad p_G = \frac{2}{8} \quad p_C = \frac{1}{8}$$

The corresponding value of entropy is 1.906. For a DNA sequence, the maximum possible value of entropy is $H_{max} = 2 (= \log_2 L \text{ for } L = 4)$. Like the correlation method, for this method also the replication origin is considered to be at the window where the value of entropy undergoes an abrupt change.

Though all these three methods directly or indirectly uses the counts of occurrences of a particular base (or bases), there is also a clear distinction between them. The GC-skew method uses only information about the G and C counts whereas the entropy measure uses all bases in a non-Markovian sense. However, the correlation uses only the G nucleotide (in principle, it can consider all bases) and includes both the short as well as long range memory. The value of correlation reflects in some way the order or non-randomness in the occurrences of the bases in the given genomic sequence.

2.5. Correlated entropy measure (CEM)

In computational study, the genomic sequence is modeled as a one dimensional symbolic sequence (representing bases) and the order of these symbols is dictated by biological reasons. Hence, the context i.e. the composition and the order, is important and this provides clues for identifying biologically significant functional sites such as promoters, coding segments, repeats, protein binding sites and origin of replication. It is to be noted that these functional sites have a specific range and any detection method should take this aspect into consideration.

In 2011, a variant of the entropy method known as super-information was proposed to analyze the exons and introns of genome sequences (Bose and Chouhan, 2011). This method, however, does not work for the problem of predicting replication origins. Super-information is the entropy of entropy. In order to calculate this, the given DNA sequence is divided into several sub-sequences and the entropy of each of these sub-sequences is calculated (see Section 2.4). The super-information is basically the entropy of all these entropy values. In our method, we calculate the auto-correlation of all these entropy values using Eq. (2). We would like to call this method correlated entropy measure (CEM).

In the proposed method, we divide the genome into windows as mentioned at the beginning of this section. Each of these windows is further divided into sub-windows of size nearly 1/400th of the window size. For these sub-windows, we evaluate their entropy (Section 2.4). With the entropy value for each window, use auto-correlation measure (Section 2.2) to find value of correlated entropy for that window. The ratio of subwindow size to window size has been chosen empirically so that number of subwindows and size of subwindows are adequate to calculate entropy of sub-window and auto-correlation of calculated entropy values.

Genome data of all the 500 bacterial species were downloaded from NCBI FTP website: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria.

2.6. Prediction of replication origin site

In the earlier paper on auto-correlation (Shah and Krishnamachari, 2012), the replication origin was identified through a visual inspection of the correlation graph. However, if one needs to analyze a large number of genomes, this approach is not practical. Also, visual inspection may lead to some unintentional errors of judgment. In order to address this issue, we have used a simple algorithm to be able to automatically predict the actual location of the replication origin without the need for any visual inspection. Here are the steps of the algorithm:

1. For a given sequence of output from any prediction method described earlier, take the derivative of this sequence (as a convention we are using left hand derivative) and center it by subtracting mean. Let this new derivative sequence obtained after subtracting the mean be $\{d_i\}$, where $i \in \{1, 2, 3, \ldots, N-1\}$ and N is the number of windows (length of the original output of our prediction method).
2. Calculate the standard deviation of this sequence, d_i, and denote it by σ_d.
3. Find each value in the derivative sequence, d_i, that is either more than 2.5 times standard deviation or less than -2.5 times standard deviation, i.e. if $|d_i| > 2.5\sigma_d$. Let the collection of all such points be d'_j, where $j \in \{1, 2, 3, \ldots, M\}$ and obviously, $M < N$.
4. Check if the total number of such points, d'_j, obtained above is less than 15% of the total number of windows, ($M < 0.15N$), to check the sanity of collection and also to ensure that the sequence is not noisy.
5. If there is at least one point in the collection, d'_j, and fulfills the above criteria ($M < 0.15N$), then there is certainly a definite peak or dip or shift. Return True.
6. Else, the sequence is noisy and has no definite peak or dip. Return False.

In the above scheme, the number 15% of N is chosen by trial and error. Choosing a small value like 5% caused the failure in detection of peaks in smaller bacteria, while larger values like 20–25% resulted in too many peaks/dips.

2.7. Classification of genomes

As mentioned in Section 1, none of the conventional methods can predict the replication origin for all bacterial genomes on their own. One reason for this could be that the huge diversity in the statistical properties of various genomes and in their mutational probabilities. Hence, it becomes important to be able to further classify the organisms for which a particular method is able to predict the replication origin. With this aim in mind, we have chosen two parameters for doing this classification: GC content and genome length. It will be interesting if just based on these two numbers for a genome, we can directly say whether a given method will be able to predict the replication origin or not. The following steps were followed in doing this classification:

- Take GC content on the x-axis and genome length in y-axis.
- For each genome, put a dot on this graph. Genomes for which a particular method is able to predict the replication origin are colored blue and others are colored red.
- Using support vector machines (SVM), draw a line on this graph so as to segregate dots of different colors in different regions.
- This is done separately for each measure.

SVM's are supervised-learning based machine learning algorithms which have been used for analysis and classification of data (Cortes and Vapnik, 1995). For the purpose of classification of data, SVM maps finite dimensional data set to a higher-dimensional data space. It makes data points separation easier in larger dimension by constructing hyper-planes and their sets. The mappings used by SVM models are defined in terms of a function called kernel function. We employed non-linear classification meaning that chosen kernel function was non-linear. As a result, hyper-planes constructed by SVM may be nonlinear to allow for maximum-margin

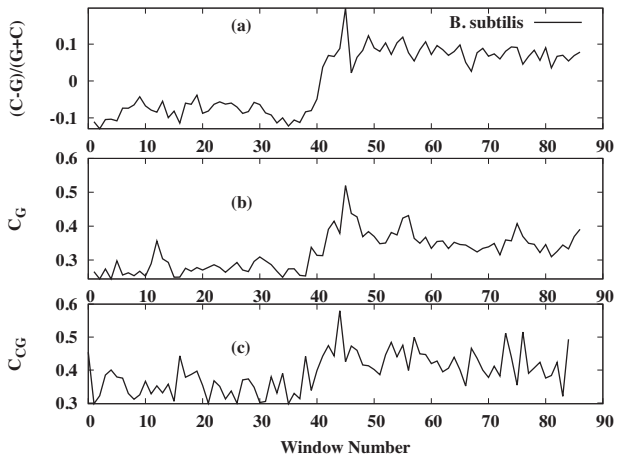

Fig. 1. Plot of (a) GC-skew (b) auto-correlation (C_G) (c) cross-correlation (C_{CG}) methods used for prediction of replication origin for *B. subtilis*. The window size is chosen to be 50 kb and shift is 48993. As can be clearly seen, all the three methods are able to predict the replication origin for this organism. Also, there is not much difference between the results of auto-correlation and the cross-correlation method.

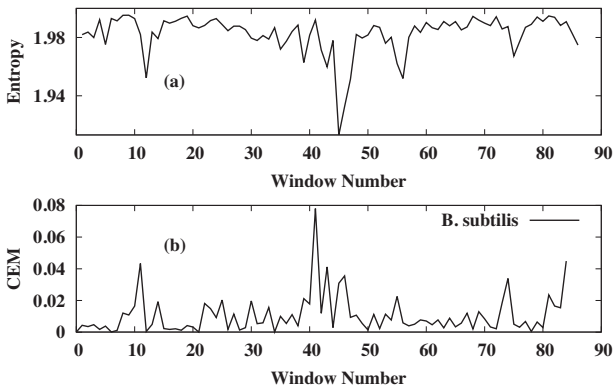

Fig. 2. Plot of the (a) entropy and (b) correlated entropy measure (CEM) used for prediction of replication origin for *B. subtilis*. The window size is chosen to be 50 kb and shift is 48993 bases. As can be clearly seen, both these methods are able to predict the replication origin for this organism. Though there is a small difference in the window where the entropy values take a dip and the window where the CEM shows a peak, we choose to neglect this since our main focus is on the qualitative features of the various methods.

hyper-plane and better classification. There are various kernels that can be used for the analysis:

- Polynomial (homogeneous):

$$k(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j)^d$$

- Polynomial (inhomogeneous):

$$k(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + c)^d$$

- Gaussian radial basis function (GRBF):

$$k(\vec{x}_i, \vec{x}_j) = e^{-\gamma \|\vec{x}_i - \vec{x}_j\|^2} \quad \text{for } \gamma = \frac{1}{2\sigma^2} > 0$$

where σ is the variance of the distribution. In this paper, we have used the polynomial (homogeneous) kernel with dimension, $d = 2$.

3. Results

For a long time, GC skew method has been believed to be the de facto method for computational prediction of replication origin in bacterial genomes. In Shah and Krishnamachari (2012), it was shown that there are some very important bacteria-like DNA sequences where the GC skew method does not work and an alternate method based on the auto-correlation function was proposed. Though both these methods work for many bacterial genomes, there is still a large number where these measures do not work. In this paper, we propose a new method called correlated entropy measure (CEM) (Section 2.5) and show that this method can predict the replication origin for almost all bacterial genomes.

Fig. 1 shows the plot of the three methods (GC skew, auto-correlation and cross-correlation) for the *B. subtilis* genome. Fig. 2 shows a similar plot for two other methods (entropy and CEM) for the same genome. It can be clearly seen that all the five methods give a sharp change in values around the window number 40. Though there are small variations in the window where each method gives an abrupt change, we choose to ignore this since our main objective in this work is not to precisely identify the origin location but only to indicate whether a certain statistical method can be used for origin prediction or not. An accurate location corresponding to any method can be found out by suitably choosing the window and shift size. In this paper, the window size was taken to be 50 kb and the shift was 48993. We have chosen such a large value of the shift so as to reduce the computational time. The particular value of 48993 was chosen so as to minimize the trailing end at the last window. The sub-window size for the CEM method is chosen to be 125 nucleotides.

We have applied all these five methods to 500 bacterial genomes and found that the CEM method works for all these genomes and is far superior to all the other methods in terms of predictability. In comparison, the GC-skew method is able to predict the origin location for only 376, auto-correlation method for 389, cross-correlation method for 369 and entropy method for only 236 genomes (see Table 1). Among the 500 genomes, there are 164 for which all of these methods work.

Figs. 3–5 show the classification of various bacterial genomes based on the workability of the GC skew, auto-correlation and entropy method respectively. The classification plot for cross-correlation method looks very similar to the graph for auto-correlation and hence, is not shown separately. CEM is able to predict replication origin for all the 500 genomes and hence, it does not have the decision boundary due to which its SVM classification plot is not shown. This classification has been generated using the support vector machine (with a polynomial kernel of dimension, $d = 2$) described in Section 2.7. We have used polynomial classification instead of GRBF (Gaussian radial basis function) to find subtle planes separating the regions. GRBF gives us Gaussian surfaces around each point which ends up overfitting the hyperplane. To avoid that and to get a linear hyper plane, we used polynomial

Table 1
Comparison table of the workability of various methods when applied to 500 genomes. It is important to note that the False Positives and False Negatives shown in 3rd and 4th column in the above table pertains only to the success/failure of the SVM classification and not to the workability of our various methods. We are not able to estimate the true success/failure of our computational predictions due to lack of experimental data.

Method	Origin prediction	SVM classification	
	Success	False +ves (%)	False −ves (%)
GC skew	376 (75.2%)	7.8	10.5
Auto-correlation	389 (77.8%)	6	13.4
Cross-correlation	369 (73.8%)	6.2	12
Entropy	236 (47.2%)	12.1	8
CEM	500 (100%)	0	0

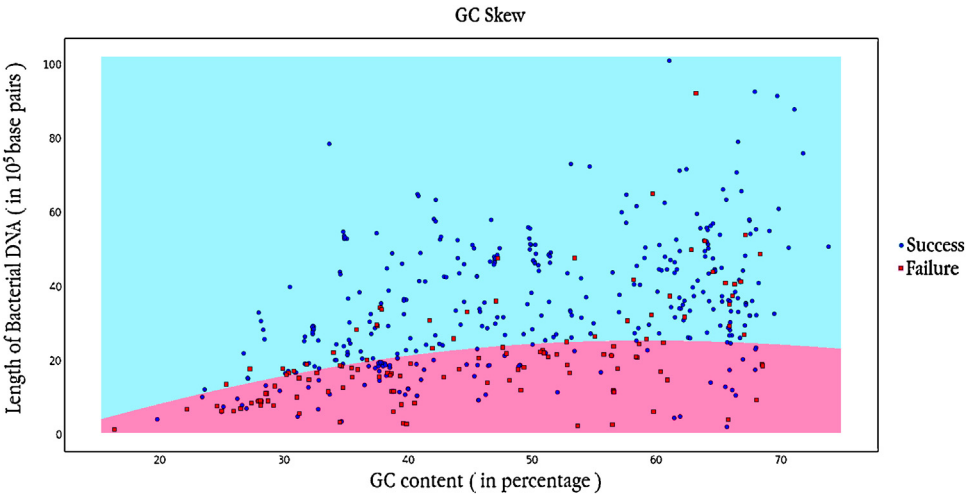

Fig. 3. Classification of bacterial genomes based on the workability of the GC skew method. The ****blue dots represent the 376 genomes for which the method works and the remaining red dots are the ones for which it does not work. The blue area denotes the SVM based classification of the genomes for which the GC skew method works.

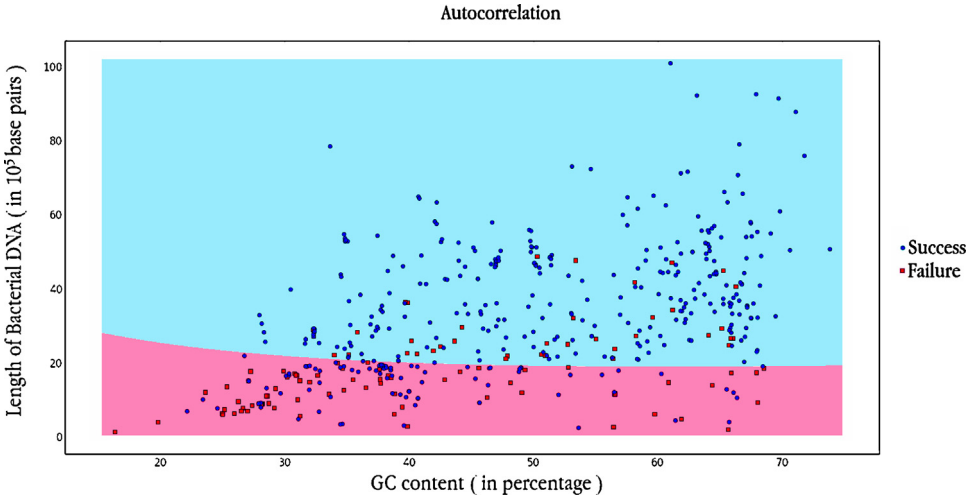

Fig. 4. Classification of bacterial genomes based on the workability of the auto-correlation method. The ****blue dots represent the 388 genomes for which the method works and the remaining red dots are the ones for which it does not work. The blue area denotes the SVM based classification of the genomes for which the auto-correlation method works. The classification plot for cross-correlation method looks very similar to this graph and hence, has not been shown separately.

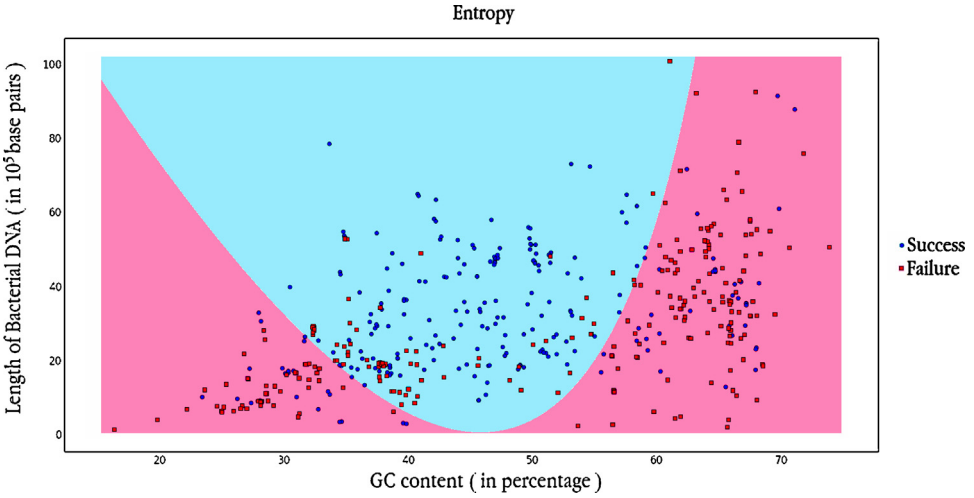

Fig. 5. Classification of bacterial genomes based on the workability of the entropy method. The ****blue dots represent the 238 genomes for which the method works and the remaining red dots are the ones for which it does not work. The blue area denotes the SVM based classification of the genomes for which the entropy method works.

classification in SVM. The class weights for failure were taken to be equal to (number of success/total bacteria) and the class weights for success were taken to be equal to (number of failure/total bacteria). Class weight is the weight of each point of the particular class. We give different weights to classes just to normalize it, so that even if one class has exceptionally large number of points compared to other, the smaller class is not undermined.

The efficiency of our classification can be gauged by measuring the percentage of false positives and false negatives for each method (see Table 1). For the case of CEM classification, the false negatives and positives are certainly zero since CEM was able to predict replication origin sites for all 500 genomes. In the case of SVM classification of auto-correlation method, the false positives are 6% and false negatives are 13.4%. For classification of cross-correlation results, the false positives are 6.2% and false negatives are 12%. For classification of GC skew method, the false positives are 7.8% and false negatives are 10.5%. For classification of entropy method, false positives are 12.1% and false negatives are 8%. The percentage of false positives is higher for classification of the entropy method mainly because this method does not work for a much larger set of genomes as compared to the other methods. It is important to note that the real purpose of doing this SVM analysis is just to add a separating hyperplane and the real classification into success/failure should be done based on the method described in Section 2.6. SVM is surely redundant for the CEM method but it could have served as a true classifier for other methods if the percentage of false positives and negatives would have been close to zero.

It can be seen in Figs. 4 and 5 that the GC skew and correlation based methods work well for genomes whose length is larger than a certain threshold. The workability of these methods does not seem to be affected much by the GC content. In contrast, the entropy method seems to perform poorly for genomes which have a high or low GC content but is not much affected by the genome length. This is quite reasonable since the entropy calculation will be more or less uniform across all windows if the genome is dominated by one or two nucleotides.

4. Discussions and conclusion

Experimental investigations about biological processes, organisms are nowadays carried out at the molecular level to get deeper insights into the problem in question. NextGen sequencing and other high throughput technologies are generating massive amount of gene, transcript, and genomic data and provides us a golden opportunity to study and understand the underlying biological reasons. Needless to mention about the usefulness and power of the computational framework to analyze these data and make meaningful inferences. Hence, theoretical (or) computational approaches are increasingly make use of genomic data and 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 sequence.

One of the critical steps in the life cycle of a cell (or) organism is the ability to completely duplicate the entire genetic information in a timely manner and this replication process is carried out in a precise sequence and regulated by protein complexes. The point of initiation of the DNA duplication event is very vital which is generally referred as origin of replication "OriC". The mechanism or mode of DNA replication is different in bacteria compared to Eukaryote and the rules of the process in the latter case is not well understood. During replication, mutational errors are caused which gives rise to DNA strand asymmetry. Computational approaches exploits this property to identify origin of replication. Identification of replication origin may help in the process of developing suitable drug or treatment for many bacterial diseases and at the same time may help in developing new vectors for recombinant experiments. Our current study is a small step in that direction.

We have used five measures, i.e. GC skew, auto-correlation, cross-correlation, Shannon entropy and CEM to see their predictability when we consider large and diverse datasets of genomic sequences. Our analysis of 500 bacterial full genome data shows that the CEM method is by far the most reliable method that can be used to predict the OriC of all bacterial genomes. It is quite natural to ask whether the enhancement of success achieved by the CEM measure is merely an accident or does it have any basis in cell biology. An important feature of the CEM measure is that unlike the other methods like auto-correlation, entropy, etc. this method takes account of the DNA sequence at shorter length-scales by using the concept of a sub-window. This is very important since the actual biological mechanism of finding the replication origin in a cell does not work at the length-scale of thousands of kilobases. This feature of a shorter length-scale gives a distinct advantage to CEM over the other methods discussed in this paper. However, the super-information method (Bose and Chouhan, 2011) also uses this same feature of a shorter length-scale but still fails to predict replication origin. The reason for this is that super-information is purely based on entropy and it has been shown in this paper that auto-correlation is a much better measure when it comes to predicting replication origin in DNA sequences. Thus, we believe that it is this combination of shorter length-scale and auto-correlation which gives CEM measure a great advantage over any other method.

Bioinformatic approaches are helping biologists to identify and characterize functional sites and are extensively employed to annotate newly sequenced genomes. Our present study should be seen in that perspective. Identifying origin of replication may give us clues about the evolutionary trajectory and the associated genetic events such as horizontal gene transfer and chromatin remodeling. Besides these the identification will help us to design and develop efficient vectors for recombinant experiments. Needless to stress the point that computational methods reduce the search space for the biologists in quest of identifying and characterizing biologically important features such as origin of replication in a cost effective way and it has become now the starting step in any experimental investigations dealing with sequences and structures. Many human diseases which are fatal are caused by bacteria and clues about its origin of replication will help us to design and develop a strategy in the form of drug or treatment to combat the specific disease and its progression.

Along with the implications for prediction of replication origin, the CEM method can also have important implications for DNA sequence analysis in general. At the fundamental level, a DNA sequence is the repository of information in a cell and it is very important to understand how the cell stores information in its DNA. What makes this understanding complex is that there are multiple levels or scales at which information is stored in a DNA sequence. And each of these levels has a different characteristic signature. Thus, the finding of a new measure useful for prediction of replication origin opens up the possibility of gaining fresh insights into the information encoding in the DNA.

References

Beauchamp, K.G., Yuen, C.K., 1979. Digital Methods for Signal Analysis. George Allen and Unwin, London.
Bose, R., Chouhan, S., 2011. Alternate measure of information useful for DNA sequences. Phys. Rev. E 83, 051918.
Cavicchi, T.J., 2000. Digital Signal Processing. John Wiley & Sons, New York.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. John Wiley & Sons, New York.
Leonard, A.C., Mechali, M., 2013. DNA replication origins. Cold Spring Harb. Perspect. Biol. 5, a010116.
Lobry, J.R., 1996. Origin of replication of *Mycoplasma genitalium*. Science 272, 745–746.
Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.
Lobry, J.R., Sueoka, N., 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. 3, 0058.1.
McFadden, G.I., Roos, D.S., 1999. Apicomplexan plastids as drug targets. Trends Microbiol. 7, 328–333.
Mrazek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. 95, 3720–3725.
Raghuram, E.V.S., Kumar, A., Biswas, S., Kumar, A., Chaubey, S., Siddiqui, M.I., Habib, S., 2007. Nuclear gyrB encodes a functional subunit of the *Plasmodium falciparum* gyrase that is involved in apicoplast DNA replication. Mol. Biochem. Parasitol. 154, 30–39.
Reif, F., 1965. Fundamentals of Statistical and Thermal Physics. McGraw-Hill, New York.
Schneider, T.D., 1997. Information content of individual genetic sequences. J. Theor. Biol. 189, 427–441.
Schneider, T.D., 2010. A brief review of molecular information theory. Nano Commun. Netw. 1, 173–180.
Sernova, N.V., Gelfand, M.S., 2008. Identification of replication origins in prokaryotic genomes. Brief Bioinform 9, 376–391.
Shah, K., Krishnamachari, A., 2012. Nucleotide correlation based measure for identifying origin of replication in genomic sequences. BioSystems 107, 52.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423, 623–656.
Soldati, D., 1999. The apicoplast as a potential therapeutic target in toxoplasma and other apicomplexan parasites. Parasitol. Today 15, 5–7.
Touchon, M., et al., 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc. Natl. Acad. Sci. U. S. A. 102, 9836–9841.