INDIAN INSTITUTE OF TECHNOLOGY DELHI

Summer Undergraduate Research Award Project Proposal

Computational Prediction of Origin of Replication in Prokaryotes and Primitive Eukaryotes

Shantanu KUMAR
Department of Electrical Engineering
2013EE10798
Ph: 9999659413
CGPA: 9.586

Barun PATRA
Department of Computer Science and Engineering
2013CS10773
Ph: 9560112674
CGPA: 9.564

Facilitator:
Kushal K. SHAH
Asst. Professor
Department of Electrical Engineering

.....
Dr. Kushal K. Shah
Facilitator

.....
HOD: Department of Electrical Engineering

1 Objectives

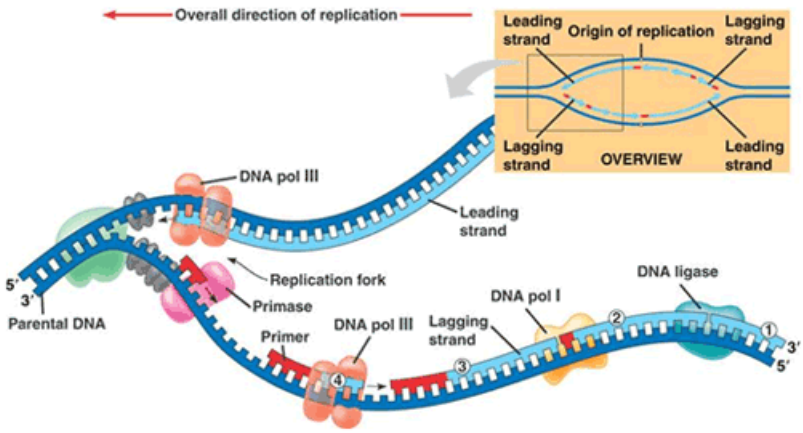
1. To make the Correlated Entropy Measure (CEM) method for finding the origin of replication in bacterial genomes, more robust, by classifying bacterial genomes into different groups with fixed CEM parameters

2. To find better classification factors for bacterial genomes, to define the workability of different statistical methods like entropy measure and auto correlation on them

3. To extend the Correlated Entropy Measure (CEM) method for finding the origin of replication to Eukaryotic cells and to define the workability of CEM for Eukaryotic cells

2 Motivation and Overview

DNA replication is the most fundamental process in the cell cycle. A cell must copy the genomic DNA before the cell division *(See figure)*.



Due to its central role in the cell cycle, identification of the origin of replication in various organisms is important for several reasons. It can help in understanding the statistical properties and organisational features of the DNA sequence of the organism. Secondly, it can also help in the discovery of new drugs for treatment of various diseases. Also, the discovery of the

replication site can be helpful in developing vectors for mass production of the parasitic cell, which is very essential for further research studies. Experimentally determining the origin of replication is not feasible. Hence, computational prediction of the origin of replication in an organism's genome is a important.

In essence, a genomic sequence is a stream of data, from which we intend to extract some meaningful patterns, and classify them based on certain parameters, using machine-learning tools like SVM, ANN and clustering algorithms. The aim is to treat the problem as a problem of data analysis and understanding, using signal processing techniques like auto-correlation of entropy, super entropy etc, to find regions of statistical anomalies, where we believe the origin of replication to lie.

3 Previous Work

Many computational methods have been developed in the past to locate the origin of replication in the bacterial genome. One of the first methods proposed to capture differences in statistical properties on the two sides of the replication site in bacterial genomes was the GC-skew method[1][2][3]. 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. This value 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. 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, the GC skew method fails to predict the replication origin for several bacterial

[1] Mrazek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. 95, 3720-3725.

[2] Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacte- ria. Mol. Biol. Evol. 13, 660-665

[3] 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.

genomes.

Another method that was proposed to predict the replication origin was based on the auto-correlation function[4]. 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.

A new method called Correlated Entropy Measure (CEM) was proposed by H. Parikh et. al.[5] It was a combination of auto-correlation and Shannon entropy. They carried out the analysis of 500 diverse bacterial genomes with widely varying GC content and found that this new entropy-cum-correlation based measure is able to predict the origin location for all these genomes. They also carried out a classification of all these bacterial genomes based on the workability of each of the earlier methods.

4 Approach

In our project, we aim to carry the idea of Correlated Entropy Measure (CEM) method forward and improve upon it.

4.1 CEM Method

In computational study, the genomic sequence is modelled as a one dimensional symbolic sequence representing bases and the order of these symbols is dictated by biological reasons. For computational prediction of replication origin the entire genome is divided into windows of equal lengths, the value of some statistical property is calculated for each window and these values are plotted versus the window number. In the CEM method, we divide the genome into windows as mentioned. Each of these windows is further divided into sub-windows. For these

[4]Shah, K., Krishnamachari, A., 2012. Nucleotide correlation based measure for iden- tifying origin of replication in genomic sequences. BioSystems 107, 52.

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

sub-windows, we evaluate their entropy *(1)*. With the entropy value for each window, use auto-correlation measure *(3)* to find value of correlated entropy for that window.

For a sub window, the entropy H is given as,

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

where p_i is the probability of occurrence of the i^{th} base.

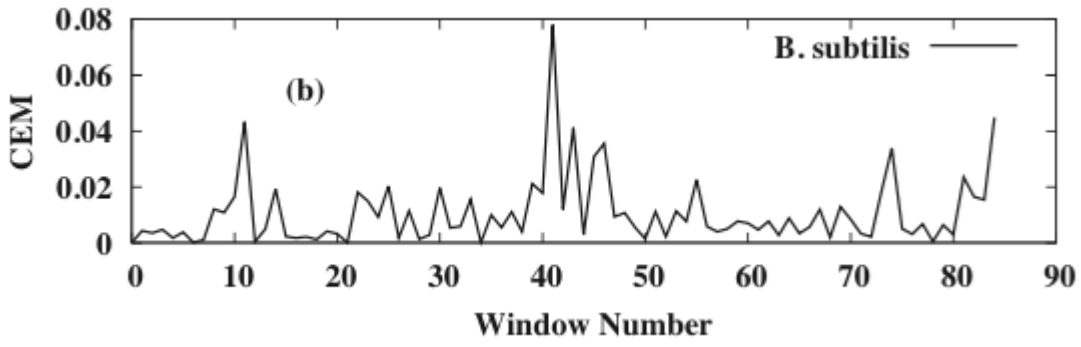
For a discrete sequence $\{a_i\}$, i$\in\{$1,2 ... N$\}$ the autocorrelation function C is given as,

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

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

where μ_a is the mean of the sequence and σ is the standard deviation.

A typical CEM versus window number plot is shown bellow. The replication of origin is believed to be around the peak.



Even though this method is successful in predicting the origin of replication for 500 species of bacteria, it is very sensitive to method parameters like the window and sub-window sizes. Varying these parameters over a range leads to the results becoming inconclusive. Moreover, the window and sub-window size needs to be adjusted for different bacterium. We intend to investigate the relationship between the method parameters (window and sub-window sizes)

with the actual results, i.e. how the plot of CEM versus window number varies with changes in the method parameters.

4.2 Methodology for Prokaryotes

We also aim at making the method more robust. We would use the fact that given any bacterial genome, there is a window and sub-window size that yields the desired result through CEM, to try and group bacteria based on these ranges of window and sub-window size using machine-learning tools like clustering algorithms, ANN and SVM. These groups can then be defined on the basis of common properties shared within the group. Thus given a new bacterial genome, we would be able to assign it a group using its inherent properties and use CEM with the window and sub-window range of the group to predict the origin of replication.

Classification of bacterial genomes on the basis of the workability of a particular method is also very useful to understand the limits of a method's functionality. In the classification proposed by Harsh et al, this was done on the basis of GC content and genome length. This classification can be made better by trying a combination of different parameters and choosing the best one that gives the least false positives and wrong negatives.

4.3 Methodology for Eukaryotes

The CEM method yields inconclusive results for primitive eukaryotes. We intend to try and use different concepts from information theory and statistics to find the origin of replication for them, or at least comment on whether the origin of replication can be found computationally.

5 Budget, Facilities and Duration

5.1 Budget

No budget is required for this project

5.2 Facilities

No special facilities are required for this project

5.3 Duration

We expect to reach a satisfactory state by the end of summer 2015.

– ∗ ∗ ∗ ∗ –