A Novel Approach to Predict the Origin of Replication

Mashiyat Alam Promi

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of

MSc in Computer Science & Engineering

June 2020

Abstract

In the genome of every species, there exists an origin, known as the origin of replication (ORI), from where the genome starts to replicate itself during the process of cell division. Finding out this origin; is therefore a very prime and demanding problem in bioinformatics research, as this is the main responsible key-factor for the replication process of DNA. In this study, we start off by choosing a benchmark dataset of a yeast named Saccharomyces cerevisiae, generate simple and inexpensive sequence based features, label and prepare the features for computation, feed them to 10 basic machine learning algorithms, compare the results, and finally propose a novel approach, to help predict the Origin of Replication by achieving **98.15%** of accuracy by implementing Logistic Regression classifier with 10 fold cross validation. Here in this study, we also represent a comparison table containing the results for all 10 experimented classifiers, to showcase the clear distinction and success of our proposed approach, from that of others.

I dedicate this dissertation to my parents who never gave up on me and helped me in all things, great and small.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor Dr. Swakkhar Shatabda, a wonderful human being who introduced me to the world of bioinformatics, inspired me to work hard and learn all the experimental strategies and research methodologies.

I would also like to thank my parents and my husband, who encouraged me throughout the whole time of this research, with their continuous support, kindness and motivation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The information stored in the DNA, plays a vital role in every life. When a cell dies or needs to get replaced, it is important to copy and reserve the information of that cell. Replication is the process that is responsible for this act. So finding out the origin of replication is a problem, that needs to be solved with a smart computational approach.

1.1 Problem Statement

When a cell is being divided, it replicates its genome. In a DNA sequence, replication begins in a region called ORI (Origin of Chromosome Replication). These regions are well known for preserving hidden messages, which contain commands, telling the cells to start the replication process [1]. This is why, it is very important to find out the ORIs in the DNA sequences, so that we can find out the exact key string, removing which, we can stop the bacterial genomes from replicating themselves. Moreover, to understand the gene expression regulation and cell-division cycle, proper identification of ORIs is a must for the researchers so that they can figure out the key factors of genetic diseases [2].

1.2 Motivation

Finding ORIs in laboratories, is a quite challenging task to do. Conventionally, they try to find out the ORIs by manually cutting down the DNA into small fragments and experiment if the genome can replicate without it or not, which is a very time consuming, extravagant and strenuous method. Whereas, we can apply machine learning-based

computational methods to minimize this struggle and successively find out ORIs in the sequences. Various machine learning approaches and computational strategies can be implemented to persuasively work with the DNA sequences to find out the ORIs.

1.3 Objectives of the Thesis

The goal of this study is to implement different classification algorithms and extract efficient features to predict the accuracy of detecting ORIs in bacterial genomes and figuring out the best one among those, by comparing the outcomes.

1.4 Thesis Contributions

This study focuses on preparing the dataset properly for iterative computation and extracting the features of an ideal length, which eventually helps in achieving the best outcomes when experimented along with the different types of classification algorithms. Also, this study intends to contribute to the problem by choosing the best combination of feature and classification method, (features of length 7 and Logistic Regression with 10 fold cross validation) to help predict the origin of replication more accurately and confidently.

1.5 Organization of the Thesis

The thesis is organised as follows:

Chapter 2 provides related works.

Chapter 3 presents the proposed method.

Chapter 4 discusses the results and experimental analysis.

Chapter 5 presents the conclusions, summarizes the thesis contributions, and discusses the future works.

Chapter 2

Background

2.1 Biological Preliminaries

Before we move to further intensive discussion of the problem, we need to go through some basic biological preliminaries related to our research, which are briefly discussed in the following:

2.1.1 DNA

DNA stands for Deoxyribonucleic Acid and it is the hereditary element in almost all living organisms. The information stored in DNA consists of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of these letters (or bases A,C,G,T) is the information available for making and maintaining an organism. The most important task of DNA is to replicate itself and help in cell division.[3] DNA carries all kinds of genetic information for the cell to grow, to absorb nutrients, and to propagate.[4]

2.1.2 RNA

RNA stands for Ribonucleic Acid and it is one of the three major biological macro elements that are essential for all living organisms. The information stored in RNA, also consists of four chemical bases: adenine (A), cytosine (C), guanine (G), and uracil (U). The main task of RNA is to copy the genetic information from DNA and transfer it to the proteins. There are three types of RNAs: mRNA, rRNA and tRNA. mRNA is the messenger RNA that temporarily helps in copying the genetic information from

DNA. rRNA stands for ribosomal RNA, which is a component that helps to form the structure of the ribosome. Finally, tRNA is

the transfer RNA that helps to translate and transfer the information of mRNA to the proteins.[4]

2.1.3 Protein

Proteins are the power forces of the cell. They work as enzymes, structural elements, signal transferring and a lot more. They are complex molecules that play vital roles in the body. They do most of the required tasks, for a cell's structure and function and regulate the tissues and body organs. Proteins are made of amino acids. There are 20 different types of amino acids that can be combined to make a protein. Each protein element has its unique 3-dimensional structure and function determined by the sequence of amino acids they are made of. [5]

2.1.4 Replication

Replication is the process, where the DNA makes a same identical copy of itself. Whenever a cell gets divided, it makes a copy of its genome and the new cell contains the same information as the parent cell, so that it can work or behave as a template of its parent cell. Replication is the most important feature of every living organism. The replication starts at a point in the DNA, called the origin of chromosomal replication (ORIc), where the DNA double helix opens up. At this point, a little segment of RNA creates the starting point for the new DNA synthesis. This is the time when the DNA starts to replicate itself with the help of an enzyme called DNA polymerase. In this process, the cell reviews the newly synthesized DNA to make sure that there are no mistakes or alteration. When the replication is complete, the new cell can replicate itself again.[6]

2.1.5 Transcription

The process of transcription takes place in the nucleus and it is the process of copying the information of the DNA strands to an RNA(mRNA) molecule. This process is executed in three steps: Initiation, elongation and termination. The first step of the transcription process is called initiation. In this step, the enzyme RNA polymerase and

proteins bind to a small region of the DNA, named the promoter. Promoter works to send signals to the DNA to unwind. Upon unwinding, the enzyme becomes eligible to read the base of a DNA strand and it acquires the necessary information to make a strand of mRNA depending on the same base of information of the mother DNA. The next step is called elongation. In this step, the mRNA strands are built and they get their nucleotides. Here, for a very short period of time, the newly formed mRNA gets attached to the unwound DNA, and the adenine (A) of the DNA strand binds with the uracil (U) of the RNA strand. And the last step is called termination, where the new RNA polymerase comes across an information of the DNA, which leads it to stop the process. And this is where the process of transcription ends and the new mRNA strand is complete and it separates itself from the mother DNA strand.[7]

2.1.6 Translation

Translation is apparently the process of producing proteins and it takes place in the nucleus. The instruction from mother DNA is extracted from the mRNA and with the help of this information, a specific series of amino acids leads to build the protein. Translation has three major steps as well: initiation, elongation and termination.

At initiation, the mRNA transfers its information to tRNA (transfer RNA) which has the amino acids, and this step is called the "initiation complex" which is the actual preparation for starting the process of translation.

In the next step, elongation, the amino acid chain becomes longer and the mRNA reads out the information of the codon (a sequence of three DNA/RNA nucleotide) one by one, and the amino acid keeps matching the codons and adding them to the protein chain. The last step is called termination and it begins when the mRNA comes across a "STOP" information from the codon. This instruction of stopping the process, stimulates the separation of the protein chain from the tRNA and the new protein chain makes its way out of the ribosome.[8]

2.1.7 The Origin of Replication

In a DNA sequence, the replication process starts at a specific region. This region contains hidden messages and commands for the cell to start the replication process. This region is called the origin of chromosomal replication or oriC. [9]

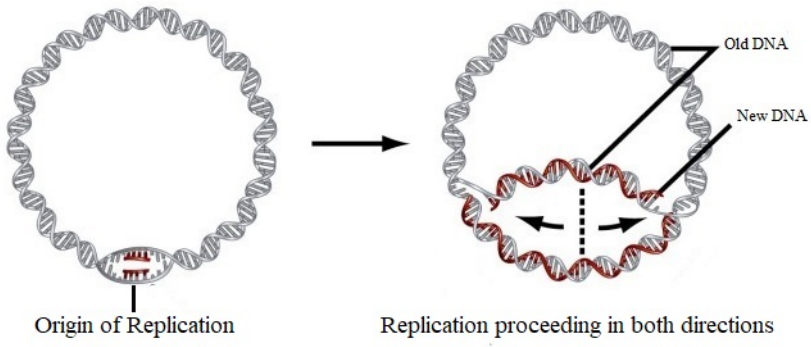

Figure 2.1: DNA, Origin of Replication, Process of Replication [10]

2.1.8 Laboratory Methods for Finding OriC

In laboratories, the process of finding out the origin of replication is manually cutting the DNA sequence into small chunk or fragments and check if the genome can replicate without that fragment or not. This cutting and checking process is repeated again and again until the key origin is found. If we consider the length of a DNA sequence, finding the origin of replication in a laboratory can be a super lengthy process and this is why, machine learning computational methods are needed to the rescue.

2.2 Literature Review

In this day and age, with the availability of highly enriched, clean and computable datasets, it has become very convenient for us to implement and play around with the computational methods and solve any kind of bioinformatics related problems. Researchers have been trying to figure out lot of strategies to identify ORIs in Genomes since time.

2.2.1 Machine learning based Methods

A well known tool for finding ORIs in bacterial genomes is OriFinder [11], which can work with unannotated genome sequences based on an integrated strategy using DNA boxes and the Z-curve method. Later, Ori-Finder 2 [12] was published, as an improved version of this tool, to predict ORICs in the archaeal genomes. This tool also followed

the Z-curve method along with a better analysis of base composition asymmetry, focusing on DNA distribution boxes and frequently occurring genes near the ORIs. A predictor named iORI-PseKNC 2.0 [13] was developed, based on the feature named PseKNC (pseudo k-tuple nucleotide composition), having 90 physicochemical properties to formulate positive and negative ORI samples. They used a two-step feature selection strategy to get rid of all the redundant information and noise. As a result, 88.53% accuracy was achieved in the 5-fold cross validation test by using support vector machine (SVM). This was basically an improved study of iORI-PseKNC [14] where they incorporated 6 local structural properties to encode the DNA and achieved 83.72% accuracy from jackknife cross-validation test. Both of the studies were based on a benchmark dataset of Saccharomyces cerevisiae genome and the outcomes were astonishing. A web tool iRO-3wPseKNC [15] was also developed on a three window based pseudo k-tuple nucleotide composition with a rigorous cross validation, which was successful in predicting the entire replication origin of some yeast species. Other than these, we can brag about a discrete feature extraction technique, named pseudo-tri-nucleotide composition (PseTNC) based on which, an efficient ensemble model iNuc-ext-PseTNC [16] was developed pretty recently, to identify the nucleosome positioning in genomes with 88.60% to 94.3% of accuracies using six-fold cross-validation. Also, an interesting tool named iOri-Human [17], was developed by implementing dinucleotide physicochemical properties into pseudo nucleotide composition to identify origin of replication in human genome. Also, there exists a tool named BOriS [18], which can identify ORI sequences from some gammaproteobacterial chromosomal fragments, even from full chromosomes by implementing motif based DNA classification methods.

2.2.2 Other Methods

Apart from some machine learning approaches, we also look for few other approaches and researches for determining ORIs. We find a study trying to identify nucleosome occupancy and the relation of it to figure out the distribution of ORIs in the DNA sequences [19]. We find a study developing a replicable oriC vector to find origin of replication[20]. We encounter a study trying to predict origin of replication in bacterial genomes by using Correlated Entropy Measure (CEM)[21]. Last but not the least, we come across a study, that use the properties of DNA segment like Multi-view Ensemble Learning (MEL)[22] to predict the origin of replication. All these studies mentioned

above, shows significant accomplishment in finding ORI related problems and we expect to learn about more of these approaches in future to expand our knowledge.

2.3 Observations

Browsing the recent research works and everything related to the problem of finding the origin of replication in DNA sequences, we can say that this is a popular and demanding problem which needs constant improvement and meticulous attention of the researchers and there are lots of scopes to work it.

2.4 Summary

So here in this section, we briefly discuss all the important biological factors and their roles in finding the ORIs. Also we note a lot of other studies and their innovative approaches to solve the same problem.

Chapter 3

Materials and Methods

Finding ORIs in genomes, is basically a classification problem, which can be solved by persuasively focusing on generating effective features and implementing the classifiers. This study begins with extracting features from the benchmark dataset of S. cerevisiae [14], using a tool named PyFeat [23]. Using the same tool, we feed these features to some machine learning-based classifiers to predict the accuracy of identifying ORIs and then compare the results. In figure 3.1 we have a flow chart to show how our system works in a nutshell, step by step.

3.1 Datasets

In this study, we consider a benchmark dataset of a yeast named S. cerevisiae, containing 405 positive ORIs and 406 negative ORIs. We obtain the dataset from the study named iORI-PseKNC [14], where they perform some filtering and cleansing procedures on the primary dataset of 740 S. cerevisiae ORIs (collected from a biological database of confirmed and predicted DNA replication origins called OriDB [24]) to build a solid and computable dataset. Firstly, they get rid of some annotated ORIs which lacked confidence because of having the "likely" and "dubious" kind of tags. Then they figure out 410 confirmed ORIs and 410 confirmed non- ORIs of 300bp length. Finally, they use the CD-HIT [25] software to get rid of redundancy and bias, and obtain this benchmark dataset of total 811 sequences of ORI samples (405 positive and 406 negative) which can be freely downloaded from their website. [1]

[1] http://lin.uestc.edu.cn/server/iOriPseKNC/data.html

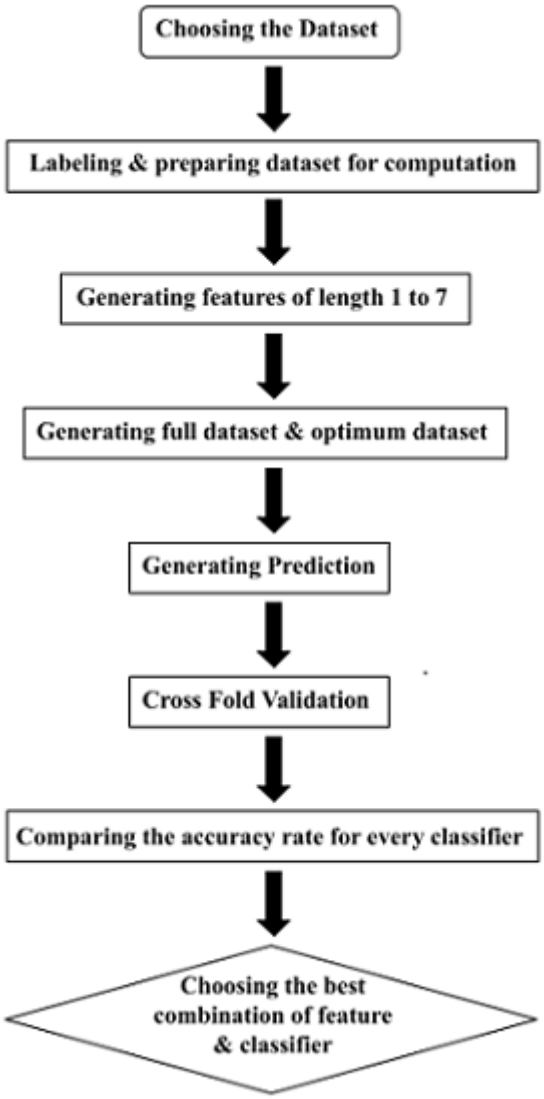

Figure 3.1: System Diagram

3.2 Feature Description

As, DNA sequences have only 4 types of nucleotides: A,C,G and T respectively standing for Adenine, Cytosine, Guanine and Thymine, our features are simply constructed with the all possible combinations of these 4 types of nucleotides, of different lengths. Considering the length, we construct our features as Monograms, Bigrams, Trigrams, and so on. Monograms; refer to the nucleotides of length 1 (i.e: A,C,G,T), Bigrams; refer to the nucleotides of length 2 (i.e: AA, AC, AG, AT....), Trigrams; refer to the nucleotides of length 3 (i.e : AAA, ACA, ACG, ACT....) and just like that, we generate the features of some more combinations of bigger lengths.

3.3 Feature Selection Techniques

In our last study [26], we worked with the k-mers, upto the length k=(1,2,3,4,5) and considering the DNAs containing only 4 kinds of nucleotides: A,C,G,and T; there were total $4^k = (4^1 + 4^2 + 4^3 + 4^4 + 4^5) = 1364$ features, and using the classifier SVM with cross fold validation, we achieved 84.36% of accuracy.

To improve the result in this study, we come up with the idea of increasing the value of k, upto 7, and we generate total $4^k = (4^1 + 4^2 + 4^3 + 4^4 + 4^5 + 4^6 + 4^7) = 21844$ features. Here, we use the tool PyFeat [23] to generate all 21844 features from our dataset, using the Linux-based commands following their user manual and finaly 2 datasets (full dataset and optimum dataset) are generated in csv format.

3.4 Classification Algorithms

Again, using the same tool [23], we generate predictions implementing the following ten machine learning algorithms/classifiers: Logistic Regression, K Nearest Neighbours, Decision Tree, GaussianNB, Bagging Classifier, Random Forest, AdaBoost Classifier, Gradient Boosting Classifier, Support Vector Machine and Linear Discriminant Analysis with 10 fold cross validation.

As we start working with these classification algorithms, We briefly describe them in the following:

LR (Logistic Regression): Logistic regression is a predictive analysis which is mostly preferred when the dependent variable (result or expected output) is binary (1=yes/0=no).

It is used to describe data and explain the relationship of one dependent binary variable with other experimental variables.[27]

SVM (Support Vector Machine): SVM is an intensive machine learning algorithm, great for both classification and regression problems. SVM is a preferred option when we work with complex relationships. It is designed to transform complex experimental data into optimum formats with a very innovative approach called the kernel trick. SVMs kernel trick surely is a magical thing and does excellent job in classification. However, the resultant output from SVM can be very hard to recondite in some cases.[28]

LDA (Linear Discriminant Analysis): LDA is commonly used in the pre-processing and pattern classification projects in machine learning and it is widely popular for its dimension reduction ability. It can compress high-dimensional data into low-dimensional attributes, while keeping a proper distinction between classes and reducing the cost of computing.[29]

GNB (Gaussian Naive Bayes): Gaussian Naive Bayes is a supervised highly efficient approach, specially used when the data or features have continuous values. GNB is generally based on the Bayes theorem but it is slightly different from actual Naive Bayes classifier as it uses Gaussian normal distribution to support spontaneous flow of data.[30]

KNN (K Nearest Neighbors): KNN is a very simple yet powerful method as it does the classification by sorting out similar data points or feature vectors from the training data and gives a generalize prediction. KNN is the best option if the experimental data and features are very few in number and can be defined with simple relationships.[31]

GB (Gradient Boosting): Gradient Boosting is a powerful prediction tool which works with the combination of an optimized loss function, decision tree and a weak learner based prediction. Gradient boosting makes an additive model based on its predictions and gradually minimizes the loss function.[32]

ABC (AdaBoost Classifier): AdaBoost is known as the first boosting algorithm out there. It works with single split decision trees as weak learners to make a prediction. AdaBoost focuses on weights by their individual observation and accuracy. AdaBoost is preferred for binary classification.[32]

RF (Random Forest): When we work with a large number of data with low correlation, Random Forest is a wise choice. Random forest works with a large number of decision trees having random splits and finds a predicted class sorted out from highest

votes and that class leads to the model's prediction. When a large number of data works as a whole, even in uncorrelated models, the prediction starts to move into the right direction and this is the advantage of using Random Forest.[33]

BC (Bagging Classifier): A Bagging classifier is an estimation tool that forms random subsets of classifiers from the given dataset, and generates individual estimations out of those subsets. Apparently, it uses the power of combined average estimations of many to improve the estimation of one final meta-estimation.[34]

DT (Decision Tree): Decision Tree belongs to the foundation level of Machine Learning approaches. It works with nodes and branches, where the data and features are evaluated at every node and these nodes train the tree to follow a certain and intuitive path to make a prediction. It can handle both classification and regression related problems, providing both numerical values and binary categorized values as predictions.[35]

3.5 Performance Evaluation

To strengthen the precision of our approach and to ensure maximum accuracy, here we use 10 fold cross validation. Cross validation is an important step and it is necessary to prove that the proposed approach not only performs better for our own dataset, but also fits into any other challenging or real world dataset. Here, we evaluate the performance of our experiment considering the following 7 parameters, with 10 fold cross validation within the mentioned ranges:

$$0.0\% \leq Accuracy \leq 100.0\%$$

$$0.0 \leq auROC \leq 1.0$$

$$0.0 \leq auPR \leq 1.0$$

$$0.0 \leq F1Score \leq 1.0$$

$$-1.0 \leq MCC \leq 1.0$$

$$0.0\% \leq Sensitivity \leq 100.0\%$$

$$0.0\% \leq Specificity \leq 100.0\%$$

The definitions of these measures for evaluations [36] are in the following section, where we will be reviewing some basic machine learning terms and formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1Score = \frac{2TP}{2TP + FP + FN}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Here, TP, TN, FP, FN respectively stands for True Positive, True Negative, False Positive and False Negatives; and all 4 of these can have values from 0 to 1, where lower values indicates less successive predictors and greater values indicates more successive predictors. auROC and auPR presents the areas under two curves having the values ranging from 0 to 1. Mathew's correlation coefficient MCC values may range from -1 to +1 where -1 indicates a bad classifier and +1 indicates a good classifier. Sensitivity stands for the measure of clear positive cases that were predicted to be True Positive, and Specificity stands for the measure of clear negative cases that were predicted to be True Negative. Finally, there comes Accuracy and F1 Score. Accuracy; is the measure of correct predictions divided by the total number of predictions. And F1 Score; is the measure counted by taking the false positives and false negatives into account, and by calculating the weighted average of Precision and Recall.

3.6 Summary

So here in this study, we work with the same old benchmark dataset, generating 21844 features, considering several machine learning algorithms and classifiers and lastly evaluating their individual performances with 10 fold cross validation and other necessary parameters, we are about to find out which approach leads us to get a better and improved prediction.

Chapter 4

Experimental Analysis

In this chapter, we demonstrate the experimental analysis and comparison table of classification algorithms implemented in this study.

4.1 Experimental Setup

To use the tool PyFeat [23], we install python 3.7 and Linux on a Lenovo Ideapad 310 computer Core i5 7th Gen with 8 GB RAM, which runs on a 64 bit Windows 10 operating system.

4.2 Effect of Feature Selection

As previously, in [26], we composed simple k-mer based features up to length 5, here in this study we try to increase the length a little bit. First of all, we label the positive 405 ORI sequences with label 1 and negative 406 ORI sequences with label 0. Then using PyFeat [23], we compose the features of length 1 to 7, and these features are extracted in two datasets : full and optimum. Again, using PyFeat, we experiment with these datasets multiple times implementing the classifiers with 10 fold cross validation. Finally, comparing the resulted outcomes, we discover that the features of length 7 for the classifier Logistic Regression provides the best possible prediction among all of the features and classifiers.

4.3 Comparison of Classification Algorithms

In the table 4.1, we can see the accuracy for all 10 classifiers: LR (Logistic Regression) 98.15%, SVM (Support Vector Machine) 97.78%, LDA (Linear Discriminant Analysis) 92.85%, GNB (Gaussian NB) 88.66%, KNN (K Nearest Neighbors) 85.31%, GB (Gradient Boosting) 83.73%, ABC (AdaBoost Classifier) 81.24%, RF (Random Forest) 80.40%, BC (Bagging Classifier) 79.80% and DT (Decision Tree) 71.51%. We can also see that the classifier **Logistic Regression** provides the highest accuracy of **98.15%**, which is remarkably high from others. Figure 4.1 and 4.2 respectively represents the accuracy and auROC (area under receiver operating characteristic precisely, true positive rate and false positive rate ratio) values for all 10 predicting classifiers.

4.4 Summary

As in our last study [26], we achieved 84.36% of accuracy with the same dataset and features of length 5, using Support Vector Machine with cross fold validation. Another similar study iori-pseknc [14] achieved 83.72% of accuracy having the same benchmark dataset, using Pseudo k-tuple nucleotide composition method along with jackknife cross-validation. Finally, we can say that our approach achieves (98.15%) of accuracy, with the features of length 7 and Logistic Regression, which is a significantly better outcome compared to the mentioned others.

Classifiers	Acc	auROC	auPR	F1	MCC	SN	SP	CM
LR	**98.15%**	**0.9929**	**0.9970**	**0.9814**	**0.9634**	98.27%	**98.03%**	398 8 7 398
SVM	97.78%	0.9909	0.9950	0.9782	0.9562	**98.77%**	96.80%	393 13 5 400
LDA	92.85%	0.9791	0.9841	0.9283	0.8584	92.84%	92.86%	377 29 29 376
GNB	88.66%	0.9308	0.9108	0.8832	0.7751	86.42%	90.89%	369 37 55 350
KNN	85.31%	0.9198	0.8979	0.8483	0.7099	82.96%	87.68%	356 50 69 336
GB	83.73%	0.9067	0.9104	0.8383	0.6765	84.44%	83.00%	337 69 63 342
ABC	81.24%	0.8739	0.8732	0.8143	0.6272	81.98%	80.54%	327 79 73 332
RF	80.40%	0.8732	0.8545	0.796	0.6108	76.79%	83.99%	341 65 94 311
BC	79.80%	0.8779	0.8627	0.7914	0.5985	76.79%	82.76%	336 70 94 311
DT	71.51%	0.715	0.6551	0.7171	0.4322	72.35%	70.69%	287 119 112 293

Table 4.1: Comparison table of the results for different classifiers

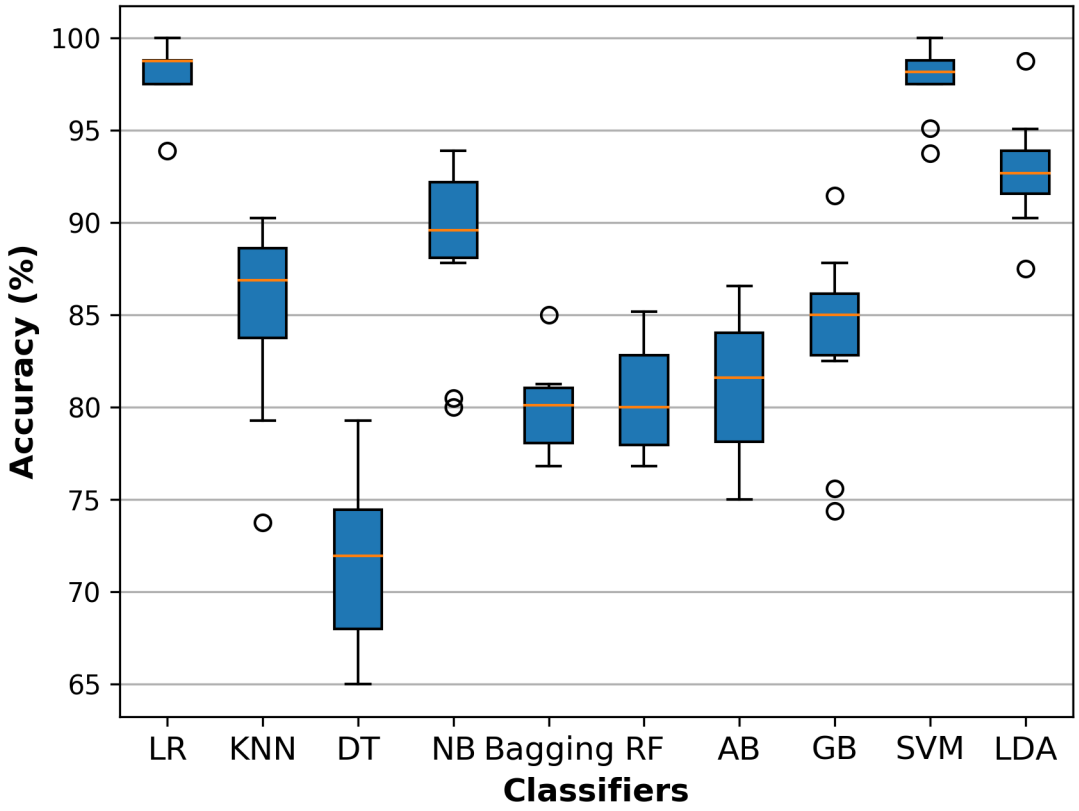

Figure 4.1: Accuracy using different classifiers

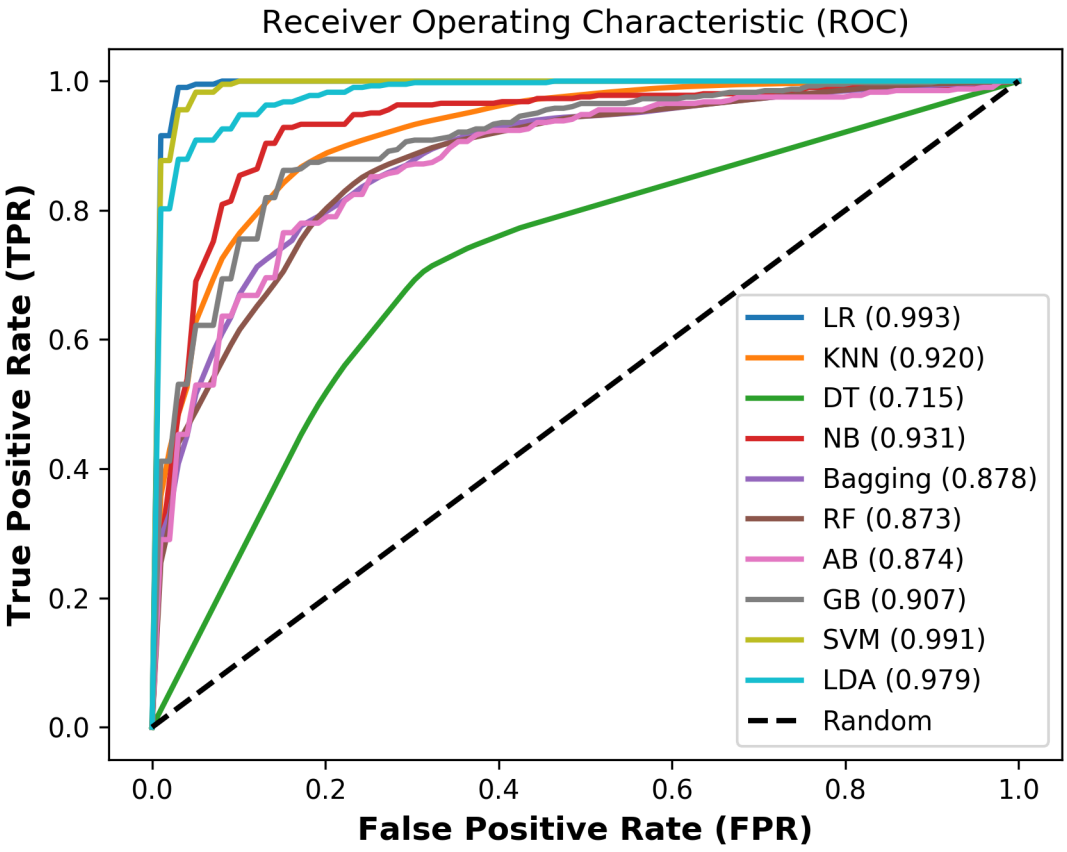

Figure 4.2: auROC graph for different classifiers

Chapter 5

Summary, Conclusions, and Future Work

5.1 Summary

In this study, we discover a combination of feature and classifiers, which works best together to provide us a comparatively better result to predict ORIs in genomes. Our proposed approach achieves overall **98.15%** accuracy with simple k-mer features of length 7 using Logistic Regression with 10 fold cross validation.

5.2 Conclusions

We acknowledge the fact that, in this study we use a pre-built tool and work with only one species (Saccharomyces cerevisiae). But with that specific species and set of data, our method achieves significantly astonishing results under all the standard evaluation metrics. The bright side of our proposed approach is the simplicity of it's feature generation process and easy execution of the experiments.

5.3 Future Work

In future, we intend to work along with various species, datasets and the features having gapped nucleotides. Also, we aspire to build a free web tool, as an extension to this study, where we will be able to input FASTA formats of genome sequences, and it will provide us the predicted ORI regions as output, so that experimenting with a large scale

of data becomes much easier. Moreover, we hope to explore demanding challenges in solving ORI finding related problems and contributing more to this great community of Bioinformatics, through our hard work and research grind.

Bibliography

[1] P. Compeau and P. Pevzner, "Bioinformatics algorithms: an active learning approach," *La Jolla, California:Active Learning Publishers, vol. 1*, August 2015. 1

[2] F.-Y. Dao, H. Lv, F. Wang, and H. Ding, "Recent advances on the machine learning methods in identifying dna replication origins in eukaryotic genomics," *Frontiers in Genetics*, December 2018. 1

[3] "Help Me Understand Genetics what is dna," https://ghr.nlm.nih.gov/primer/basics/dna, accessed: 2020-04-02. 3

[4] "RNA Society what is rna," https://www.rnasociety.org/what-is-rna, accessed: 2020-04-02. 3, 4

[5] "Help Me Understand Genetics what are proteins and what do they do?" https://ghr.nlm.nih.gov/primer/howgeneswork/protein, accessed: 2020-04-02. 4

[6] "Scitable nature education," https://www.nature.com/scitable/definition/replication-33/, accessed: 2020-04-02. 4

[7] "lumenlearning steps of genetic transcription," https://courses.lumenlearning.com/wm-biology1/chapter/reading-steps-of-genetic-transcription/, accessed: 2020-06-06. 5

[8] "ibiologia dna translation —introduction, steps daigram," https://ibiologia.com/dna-translation/#DNA_Translation, accessed: 2020-06-06. 5

[9] D. Kornberg and D. TA, "Replication," *San Francisco: W H. Freeman*, 1980. 5

[10] Pearson Prentice hall inc, "4 dna replication animation," 2005, [Image; accessed January 27, 2021]. [Online]. Available: https://pulpbits.net/4-dna-replication-animation/ 6

[11] F. Gao and C.-T. Zhang, "Ori-finder: A web-based system for finding oric s in unannotated bacterial genomes," *BMC bioinformatics, vol. 9, no. 1, pp. 79*, February 2008. 6

[12] H. Luo1, C.-T. Zhang, and F. Gao, "Ori-finder 2, an integrated tool to predict replication origins in the archaeal genomes," *Frontiers in microbiology, vol. 5, pp. 482*, September 2014. 6

[13] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, and H. Lin, "Identify origin of replication in saccharomyces cerevisiae using two-step feature selection technique," *ISCB*, November 2018. 7

[14] W.-C. Li, E.-Z. Deng, H. Ding, W. Chen, and H. Lin, "iori-pseknc: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition," *Chemometrics and Intelligent Laboratory Systems, vol. 141, pp. 100-106*, February 2015. 7, 9, 16

[15] L. Bin, W. Fan, H. De-Shuang, and C. Kuo-Chen, "iro-3wpseknc: identify dna replication origins by three-window-based pseknc," *Bioinformatics (Oxford, England)*, April 2018. 7

[16] M. Tahir, M. Hayat, and S. A. Khan, "inuc-ext-psetnc: an efficient ensemble model for identification of nucleosome positioning by extending the concept of chou's pseaac to pseudo-tri-nucleotide composition," *Springer Berlin Heidelberg*, October 2018. 7

[17] Z. Chang-Jian, T. Hua, L. Wen-Chao, L. Hao, C. Wei, and C. Kuo-Chen, "iori-human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition." *Oncotarget*, November 2014. 7

[18] T. Sperlea, L. Muth, R. Martin, C. Weigel, T. Waldminghaus, and D. Heider, "boris: Identification of origins of replication in gammaproteobacteria using motif-based machine learning supplementary information," *Cold Spring Harbor Laboratory*, April 2019. 7

[19] S. Yin, W. Deng, L. Hu, and X. Kong, "The impact of nucleosome positioning on the organization of replication origins in eukaryotes," *Biochem Biophys Res Commun*, July 2009. 7

[20] M. A. Shahid, M. S. Marenda, P. F. Markham, and A. H. Noormohammadi, "Development of an oric vector for use in mycoplasma synoviae," *J Microbiol Methods*, August 2014. 7

[21] H. Parikh, A. Singh, A. Krishnamachari, and K. Shah, "Computational prediction of origin of replication in bacterial genomes using correlated entropy measure (cem)," *Biosystems*, February 2015. 7

[22] V. K. Singh, V. Kumar, and A. Krishnamachari, "Prediction of replication sites in saccharomyces cerevisiae genome using dna segment properties: Multi-view ensemble learning (mel) approach," *Biosystems*, January 2018. 7

[23] M. R. Jani, S. Ahmed, D. M. Farid, S. Shatabda, A. Sharma, and I. A. Dehzangi, "Pyfeat: A python-based effective feature generation tool for dna, rna, and protein sequences," *Bioinformatics (Oxford, England)*, March 2019. 9, 11, 15

[24] C. C. Siow, S. R. Nieduszynska, C. A. Müller, and C. A. Nieduszynski, "Oridb: a dna replication origin database. nucleic acids research," *Nucleic Acids Research, Volume 40, Issue D1, 1*, January 2012. 9

[25] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." *Bioinformatics (Oxford, England)*, August 2006. 9

[26] S. Sangskriti, M. A. Promi, D. M. Farid, and S. Shatabda, "Prediction of origin of replication in genome using dna sequence based features," *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, December 2018. 11, 15, 16

[27] "StatisticSolutions logistic regression," https://www.statisticssolutions.com/what-is-logistic-regression/, accessed: 2020-06-06. 12

[28] "KDnuggets support vector machine," https://www.kdnuggets.com/2017/02/yhat-support-vector-machine.html, accessed: 2020-06-06. 12

[29] "digitalvidya everything you need to know about linear discriminant analysis," https://www.digitalvidya.com/blog/linear-discriminant-analysis/, accessed: 2020-06-06. 12

[30] "opengenus gaussian naive bayes," https://iq.opengenus.org/gaussian-naive-bayes/, accessed: 2020-06-06. 12

[31] "xnextcon knn explained," http://blog.xnextcon.com/?p=213, accessed: 2020-06-06. 12

[32] "machinelearningmastery a gentle introduction to the gradient boosting algorithm for machine learning," https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/, accessed: 2020-06-06. 12

[33] "TowardsDataScience understanding random forest," https://towardsdatascience.com/understanding-random-forest-58381e0602d2, accessed: 2020-06-06. 13

[34] "TowardsDataScience using bagging and boosting to improve classification tree accuracy," https://towardsdatascience.com/using-bagging-and-boosting-to-improve-classification-tree-accuracy-6d3bb6c95e5b, accessed: 2020-06-06. 13

[35] "TowardsDataScience decision trees explained," https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6, accessed: 2020-06-06. 13

[36] D. Powers, "Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation," *Journal of Machine Learning Technologies*, January 2008. 14