Application of Machine Learning Algorithms to Identify Recombination Spots

Sajid Ahmed, Student id: 011 131 047
Mehedi Hassan, Student id: 011 131 056
Bulbul Ahmed, Student id: 011 131 038
Jannatul Ferdous, Student id: 011 131 008

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of

BSc in Computer Science & Engineering

September 2018

Abstract

Meiotic recombination is a mechanism by which a cell promotes correct segregation of homologous chromosomes and repair of DNA damages. But it does not occur randomly across the whole genome. Relatively high frequencies meiotic recombination regions are identified as hotspots and relatively low frequencies meiotic recombination regions are identified as cold spots. But the accurate prediction of hot/cold spots is still an open challenge. Here, Recombination hotspots in a genome which are unevenly distributed. Again, hotspots are regions in a genome which show higher rates of meiotic recombination. Computational methods for recombination hotspot prediction often use sophisticated features which are derived from physio-chemical or structure-based properties of nucleotides. In this study we have taken a DNA data set. In this work, we have shown the uses of sequential based features which are computationally cheaper to generate. For this data set we used gapped k-mar composition. The data set which we have taken is a string data set. To do our work easier we have rearranged our string data set. Then we applied different algorithms on our data set to predict the result. It is also mentionable that we have tested our algorithm on standard benchmark dataset. Again, we also used 5-fold and 10-fold cross-validation in our dataset. Our analysis shows that compared to other methods, our work is able to produce significantly better results in terms of accuracy. For 5-fold cross-validation among all the algorithms SVM gives the best sensitivity and it is 0.7707. And, for 10-fold cross-validation, both LR and ANN gives best result of sensitivity and it is 0.7622. Here, the result of sensitivity for SVM is quite impressive and it is 0.7601.

Acknowledgements

First of all, we all would like to express our gratitude to the Almighty to enable us to complete this report on Application of Machine Learning Algorithm to Identify Recombination Spots.

We convey our deepest appreciation and gratitude to our Academic supervisor Dr. SwakkharShatabda, Associate Professor and Undergraduate Program Coordinator, Dept. of CSE, United International University for giving us the opportunity to work on this topic. It would never be possible for us to take our thesis work at this stage without his relentless support and encouragement.

Contents

List of Figures

Chapter 1

Introduction

It is Eukaryotic organisms who have regions of high that are known as hot-spots and low that are known as cold-spots. There is no similarityrelationship between biological and physical map distances. Some reasons are there thats why it is important to learn recombination hot spots and cold spots.It is not possible to understand the mechanism of initiating recombination without having proper knowledge of the reasons that regulate hot-spots and cold-spots. Again, it is also addition able that number of recombination events per DNA is relevant to the accuracy of DNA segment. Also, the distribution of exchanges may influence the probability of assembling new configurations of physically linked genes during evolution.Being having the understanding of hot spots and cold spots will be relevant to comprehending other DNA-related processes such as transcription and replication. It is the meiotic recombination that describes the process of alleles between homologous chromosomes during the meiosis process. It can easily provide material for natural selection by producing diverse gametes. Moreover, it might have contribution to the evolution of the genome via gene conversion or mutagenesis. Though it is not clear yet where recombination exactly happens in the genome and mechanism of recombination, it is very much assured that recombination plays an important role in promoting genome evolution. Thats why it is one of the most interested research fields for the scientists. Recombination presents a non-uniform distribution across the genome.

1.1 Motivation

It is very urgent to develop more methods to identify the recombination spots since the number of sequenced genomes showed explosive growth. Before our approached method each of the aforementioned methods has its own advantage, and did play a role in stimulating the development of this important area. At the same time, they also have many more disadvantages. Since none of the methods allows users to set the desired parameters for prediction, and hence it is difficult for them to optimize the predict system according to the need of their focus.

1.2 Objectives

The main objective is to solve the prediction problem focusing on the following steps-

1. Using string dataset to detect the meiotic recombination hotspots and cold spots.
2. Using sequence-based feature extraction.
3. Methodical approach to select features from the selected string dataset.
4. Establishing an effective classification method.
5. Comparison among results applying algorithms.

1.3 Contribution of the Thesis

Our study work is initiated in an attempt to address these shortcomings by developing a more powerful predictor for identifying DNA recombination spots. In this work basically, we have done two related works, finding the novel features for Recombination Spots and proposing the best algorithm to predict result among all algorithms. Our work consists:(i) benchmark dataset, (ii) sample representation, (iii) operation algorithm, (iv) validation.All of them are presented separately in different sections in each Chapter.

1.4 Organization of the Report

Rest of the report is organized as following: Chapter 2 briefly presents a literature review of the related work; Chapter 3 describes the methodology and materials proposed in this paper; experimental results are shown in Chapter 4 with a discussion; the paper conclude in Chapter 5

Chapter 2

Background

In this chapter, we present the necessary preliminaries to understand the problem. We also address the literature review.

2.1 Biological Preliminaries

In this section, we present a brief about few biological entities.

2.1.1 DNA

DNA is the shortform of Deoxyribonucleic acid. It is a heredity material in the human being and almost all other organisms. DNA, a molecule composed of two chains that coil around each other to form a double helix carrying genetic instructions used in growth, development, and reproduction of all known organisms. Again, DNA stores information as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Hereditary information of an organism is encoded through the sequences of these four nucleotides within a DNA. By using this information, we can easily decode the basic function and malfunction of the organism.

2.1.2 Recombination spots

Recombination spots are areas in genomes which displayhigh rates of recombination related with neutral prospect. Recombination rate in hotspots may be numerous times of surrounding areas. Recombination hotspots results by high breaking of DNA processrelated to thoseareas, and also it is appliedfor both the cells which are known

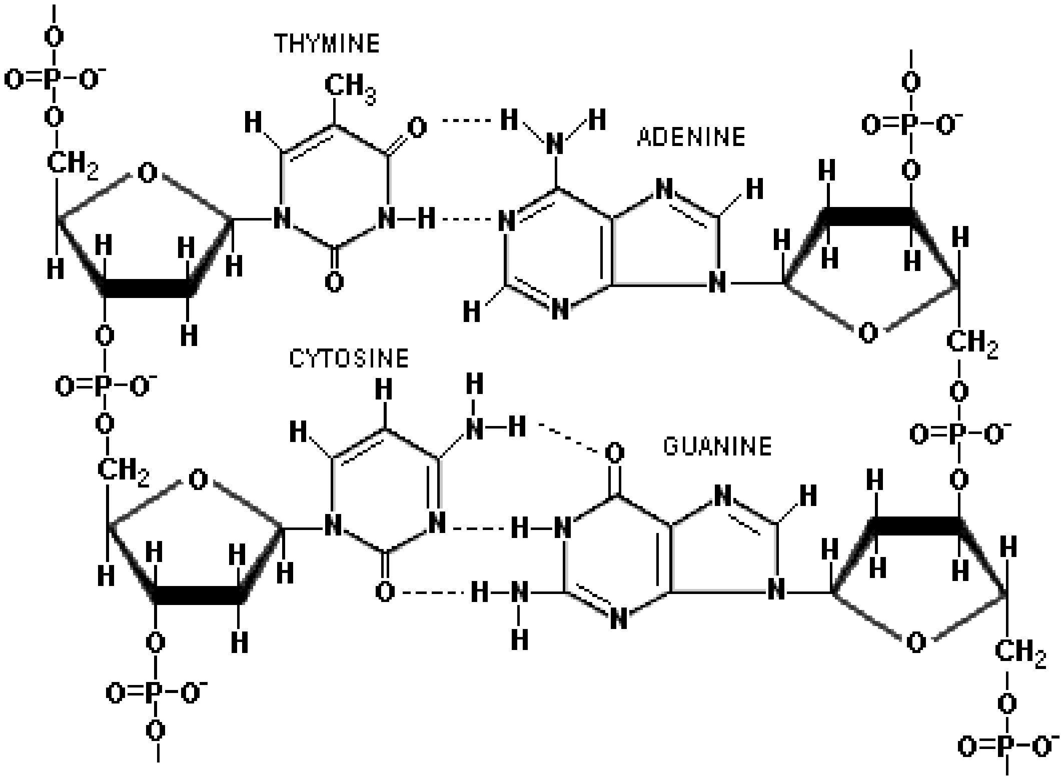


Figure 2.1: Chemical structure of DNA hydrogen bonds shown as dotted lines

as mitotic and meiotic. This title refers to recombination occurrences resultedby the jaggedformat of programmed meiotic breaks. It is DNA which have ”fragile sites” in its sequence that are bent to recombination. Those fragile sites are connectedto the following repeats: CCG-CGG, CTG-GAG, TTC-GAA, and NGC-GCN. Those fragile sites are reserved in mammals and in yeast, which suggest that something inherent to the molecular structure of DNA cause impermanence and is connectedto DNA-repeat impermanence. These fragile sites are thought to form hairpin structures on the lagging strand during replication from single-stranded DNA base-pairing with itself in the trinucleotide repeat region. These hairpin structures cause DNA breaks that lead to a higher frequency of recombination at these sites. Recombination hotspots are also thought to arise due to higher-order chromosome structure that make some areas of the chromosome more accessible to recombination than others.

2.1.3 Hotspots

Genomic regions that present relatively higher frequencies of recombination are called hotspots. Recombination hotspots are regions in a genome that can exhibit elevated rates of recombination relative to a neutral expectation. Recombination hotspots result from higher DNA break formation in these regions, and apply to both mitotic and meiotic cells.

2.1.4 Coldspots

Genomic regions that present relatively lower frequencies of recombination are called recombination coldspots.Recombination coldspots result from lower DNA break formation.

2.2 Types of Approaches

There are two types of approaches: laboratory methods and computational methods.

2.2.1 Laboratory Method

A laboratory method is an approach where work is being done in the laboratory. It is a planned learning activity dealing with original or raw data in the solution of the problem. Moreover, it is used to designate a teaching procedure in the physical science that uses experimentation with apparatus. Its aim is to provide firsthand experience to the experimenters. Also, it gives us the opportunity to participate in original research and to develop our skills by using the laboratory equipment. Moreover, it is the area where we can make use the power of observation and reasoning. But in this method, we cannot verify all knowledge through experiments. Moreover, this method is usually time-consuming. It is also an expensive method.

2.2.2 Computational Approach

Computational approach mainly means to emphasize the difference between the approach and the verbal theory related to this topic. Vision of this approach requires a list of assumptions and obligations. But this assumptions and obligations needs to be formatted in such a way that they can be easily inputted in a computational method and can be able to predict a result. Compare to the laboratory approach the computational approach is faster. But it may give less accuracy than laboratory approach. Computational approach is a cost-efficient method. This is the main thing that researchers need to be focused on at present.

2.3 Literature Review

By collecting information from many renowned resources, we come to know that recombination plays an important role is genetic engineering mostly in genetic evolution.It is genetic evolution which mainly describes the exchange of genetic information during the time of each generation in diploid organisms.Many new combinations of genetic validations are provided by recombination and moreover it is a very important resource for biodiversity. It is biodiversity which can mainly do the work of accelerating the procedure of biological evolution. Again, it can be added that with having proper knowledge about recombination spots may also help in understanding the reproduction and growth of cells. At present, it is a time demanding approach to generate computational methods to predict recombination spots.

Many of us have tried previously to do some better works in this field. As an example, it can be said that Jiang et al. [1] had developed a predictor based on the gapped dinucleotide composition features to do this kind of work. Name of Jiang el al. predictor is RF-DYMHC. Again, using the same procedure Liu et al. [2], use the kmar gap approach and also the increment of diversity which is combined with quadratic discriminant analysis. He developed the predictor named as IDQD to do the similar kind of work of Jiang et al. But there are some problems in their predictors. First of all, their predictors only done their work with DNA sequence information. Again, their predictors had some more limitations. Two new predictors had been developed to overcome their limitations named as iRSpot-PseDNC [3] and iRSpot-TNCPseAAC [4]. Previous two approaches used DNA local sequence properties and pseudo dinucleotide composition as their base whereas the next two approaches used DNA trinucleotide composition [5] and the similar pseudo amino acid components [6].

All of these methods have some advantages and some disadvantages as well. They all plays a vital role in stimulating the development of this important area. Since, all of them have some disadvantages, they can be said as:

1. Users are not allowed to insert their desired parameters to do predictions in all these methods,

2. All the predictors cannot be directly used to do the work of genome-wide analysis except the RF-DYMHC [1], but it is also added that RF-DYMHC [1] predictor was also not so accurate since the window size is arbitrary in it.

To solve all of these problems which were faced before, a new we-server predictor has been proposed to identify the recombination spots. It is built by fusing [7] variable modes of pseudo nucleotide components [8–11] and di-nucleotide based auto-cross validation [12] into an ensemble classifier called iRSpot-EL. The new predictor does the works of allowing the users to select their desired parameters and also helps more in natural to conduct the genome-wide analysis due-to its built-in flexible sliding window method [7]. Again, it can be said that identifying recombination spots with an ensemble learning approach is also more accurate and stable but particularly [13], compared with the existing predictors Application of Machine Learning Algorithm to Identify Recombination Spots works more precisely.

Chapter 3

Proposed Method

In this chapter, we present the methodology of our paper. A system diagram of the overall method is given in Figure 3.1.

3.1 Dataset Description

Construction or selected of a benchmark dataset is the most significant part of supervised machine learning-based computational methods and the datasets consists of positive and negative instances for any binary classification task as recombination hotspot prognosis task. With the following formula it can be defined as:

$$\mathbb{S} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \tag{3.1}$$

Here, $\mathbb{S}$ is the total dataset that consists DNA sequences as strings of nucleotides. The set of positive $\mathbb{S}^{+}$ represents instances of recombination hotspots where $\mathbb{S}^{-}$ represents the set of negative instances or recombination coldspots. The dataset used in this paper are appropriated from [13] and been used extensively in the literature of recombination hotspot prediction [14–18]. As a whole, 490 DNA segments of hotspot samples as well 591 DNA segments of coldspots were selected for the dataset. Selection of these instances as per the suggestions recommended in [19]. Using CD-HIT [20] 75% similarity of the sequences was removed that helped to reduce the effect of homology and redundancy of similar sequence in the datasets. Following, the derived dataset from the result consists of 478 sequences that are positive samples or hotspots and 572 sequences that are coldspots. Below given a summary of the dataset in Table 3.1.

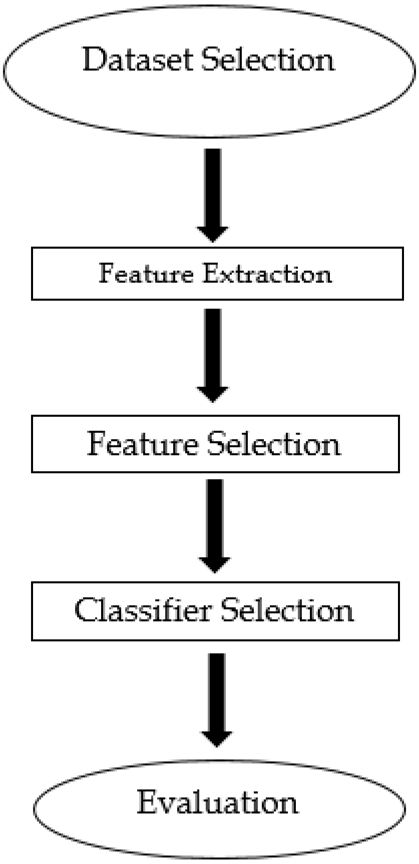
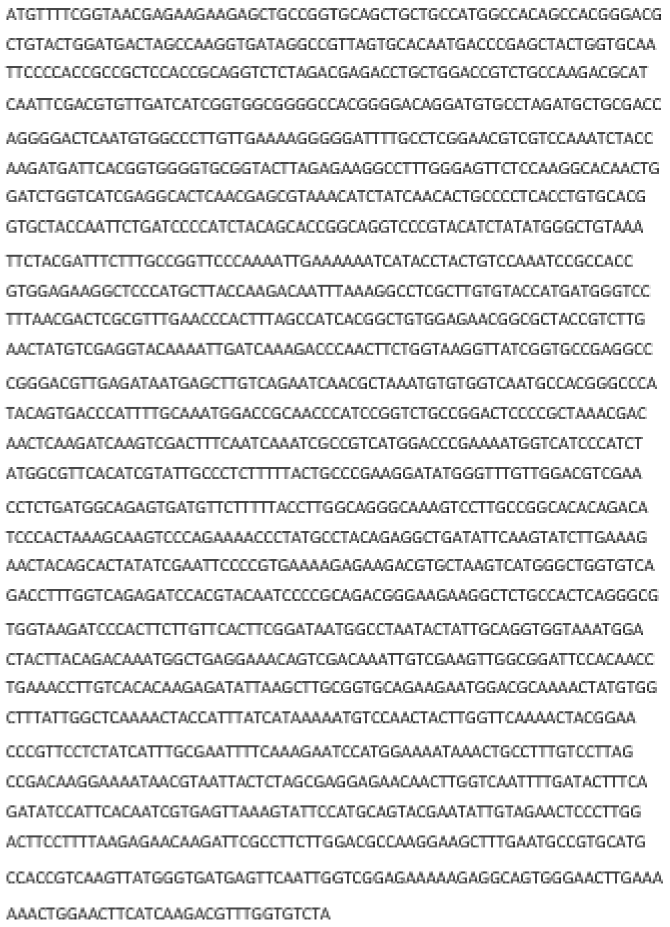


Figure 3.1: A system diagram to Identify Recombination Spots.

Class	Number of Instances	Relative Hybridization Ratio
Hotspot	478	>1.5
Clodspot	572	<0.82

Table 3.1: Summary of the dataset.



```
ATGTTTTCGGTAACGAGAAGAAGAGCTGCCGGTGCAGCTGCTGCCATGGCCACAGCCACGGGACG
CTGTACTGGATGACTAGCCAAGGTGATAGGCCGTTAGTGCACAATGACCCGAGCTACTGGTGCAA
TTCCCCACCGCCGCTCCACCGCAGGTCTCTAGACGAGACCTGCTGGACCGTCTGCCAAGACGCAT
CAATTCGACGTGTTGATCATCGGTGGCGGGGCCACGGGGACAGGATGTGCCTAGATGCTGCGACC
AGGGGACTCAATGTGGCCCTTGTTGAAAAGGGGGATTTTGCCTCGGAACGTCGTCCAAATCTACC
AAGATGATTCACGGTGGGGTGCGGTACTTAGAGAAGGCCTTTGGGAGTTCTCCAAGGCACAACTG
GATCTGGTCATCGAGGCACTCAACGAGCGTAAACATCTATCAACACTGCCCCTCACCTGTGCACG
GTGCTACCAATTCTGATCCCCATCTACAGCACCGGCAGGTCCCGTACATCTATATGGGCTGTAAA
TTCTACGATTTCTTTGCCGGTTCCCAAAATTGAAAAAATCATACCTACTGTCCAAATCCGCCACC
GTGGAGAAGGCTCCCATGCTTACCAAGACAATTTAAAGGCCTCGCTTGTGTACCATGATGGGTCC
TTTAACGACTCGCGTTTGAACCCACTTTAGCCATCACGGCTGTGGAGAACGGCGCTACCGTCTTG
AACTATGTCGAGGTACAAAATTGATCAAAGACCCAACTTCTGGTAAGGTTATCGGTGCCGAGGCC
CGGGACGTTGAGATAATGAGCTTGTCAGAATCAACGCTAAATGTGTGGTCAATGCCACGGGCCCA
TACAGTGACCCATTTTGCAAATGGACCGCAACCCATCCGGTCTGCCGGACTCCCCGCTAAACGAC
AACTCAAGATCAAGTCGACTTTCAATCAAATCGCCGTCATGGACCCGAAAATGGTCATCCCATCT
ATGGCGTTCACATCGTATTGCCCTCTTTTTACTGCCCGAAGGATATGGGTTTGTTGGACGTCGAA
CCTCTGATGGCAGAGTGATGTTCTTTTTACCTTGGCAGGGCAAAGTCCTTGCCGGCACACAGACA
TCCCACTAAAGCAAGTCCCAGAAAACCCTATGCCTACAGAGGCTGATATTCAAGTATCTTGAAAG
AACTACAGCACTATATCGAATTCCCCGTGAAAAGAGAAGACGTGCTAAGTCATGGGCTGGTGTCA
GACCTTTGGTCAGAGATCCACGTACAATCCCCGCAGACGGGAAGAAGGCTCTGCCACTCAGGGCG
TGGTAAGATCCCACTTCTTGTTCACTTCGGATAATGGCCTAATACTATTGCAGGTGGTAAATGGA
CTACTTACAGACAAATGGCTGAGGAAACAGTCGACAAATTGTCGAAGTTGGCGGATTCCACAACC
TGAAACCTTGTCACACAAGAGATATTAAGCTTGCGGTGCAGAAGAATGGACGCAAAACTATGTGG
CTTTATTGGCTCAAAACTACCATTTATCATAAAAATGTCCAACTACTTGGTTCAAAACTACGGAA
CCCGTTCCTCTATCATTTGCGAATTTTCAAAGAATCCATGGAAAATAAACTGCCTTTGTCCTTAG
CCGACAAGGAAAATAACGTAATTACTCTAGCGAGGAGAACAACTTGGTCAATTTTGATACTTTCA
GATATCCATTCACAATCGTGAGTTAAAGTATTCCATGCAGTACGAATATTGTAGAACTCCCTTGG
ACTTCCTTTTAAGAGAACAAGATTCGCCTTCTTGGACGCCAAGGAAGCTTTGAATGCCGTGCATG
CCACCGTCAAGTTATGGGTGATGAGTTCAATTGGTCGGAGAAAAAGAGGCAGTGGGAACTTGAAA
AAACTGGAACTTCATCAAGACGTTTGGTGTCTA
```

Figure 3.2: Data Sequence of Benchmark Dataset.

3.2 Example Data Sequence

An example data sequence from the dataset is given here in Figure 3.2.

3.3 Representation of DNA Sequences

We want to reduce the risk of genetic mutations from transcription errors. We see that our approach can do the exact same thing that we require. Moreover, our approach can find out cancer DNA which is very beneficial in medical science. Thats why we have chosen this.

Each DNA sequence is a sequence of a nucleotide of length 300 with a DNA alphabet

Feature Extract	Number of Features
A-A	160
AA-A	640
AAA-A	2560

Table 3.2: Feature Extraction Table.

$= A, C, G, T$.

3.4 Feature List

In total four types of features were used in our work. They are as follows:

1. Two Combination
2. Three Combination
3. Four Combination
4. K-mer gap

A summary of the features are given in Table 3.2.

3.5 Algorithm Selection

To process our dataset, we have used cross-validation. In cross-validation, we used K-fold and satisfied K-fold validation. Moreover, we have used scaling to process our data. Again, we have also used data transformation.

Five different algorithms were used in our work. They are as follows:

1. Linear regression[LR]
2. Support vector machine [SVM]
3. Decision tree [DST]
4. Artificial neural network [ANN]
5. K nearest neighbor [KNN]
6. Random forest [RF]

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

$$Sensitivity(S_n) = \frac{TP}{TP + FN}$$

$$Specificity(S_p) = \frac{TN}{TN + FP}$$

$$Precision(P_c) = \frac{TP}{TP + FP}$$

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

3.6 Performance Evaluation

It is very important to fix a sampling method on the datasets to access the performance of classifier algorithms. There are three most popular techniques to do this kind of works. They are: independent test set, cross-fold validation, and jack-knife test.We have used the cross-fold validation technique in order to compare the performance of our method to identify recombination spots with the existing methods.

We have used some performance measures to do our work. They are: accuracy, sensitivity, specificity, precision and Mathews Correlation Coefficient (MCC). Confusion matrix gives four different measures for a binary classifier problem: i) true positive (TP) or number of positive samples in the dataset that are correctly predicted by the classifier; ii) true negative (TN) or number of negative samples in the dataset that are correctly predicted by the classifier; iii) false positive (FP) or number of negative samples in the dataset that are wrongly predicted by the classifier; and iv) false negative (FN) or number of positive samples in the dataset that are wrongly predicted by the classifier. The following equations are used to measure different values in our dataset:

All the measures have the values in the range [0, 1] except MCC. In this, a maximum value of 1 will mean the best classifier with 100% accuracy or precision or other measures and 0 means the worst classifier. MCC has a value in the range [-1, +1]. In this, maximum value of +1 indicates a perfect prediction algorithm. Many classifiers are probabilistic and uses a threshold to cut-off the negative samples from positive samples and thats why the confusion matrix becomes dependent on the threshold. This set of metrics is valid only for the single-label systems. For multi-label systems whose existence has become more frequent in system biology [21–25] and system medicine

[26, 27] and biomedicine [28], a completely different set of metrics as defined in [29] is needed.

Chapter 4

Experimental Analysis

In this chapter, we present the experimental analysis and discussion on our work.

4.1 Results

At first, we present the 5-fold cross validation results from our experiments using the different algorithms on the benchmark datasets.

Table 4.1 shows results of cross fold validation for different algorithms. We also show a bar chart of the accuracies of the different algorithms in Figure 4.1.

At first, we used the linear regression (LR) method. We get accuracy 81.08% when we use 5-fold cross-validation. Secondly, we use support vector machine (SVM) method. We get accuracy 82.43% when we use 5-fold validation. Thirdly we use decision tree (DST) method. We get accuracy 72.53% when we use 5-fold validation. Fourthly we use an artificial neural network (ANN) method. We get accuracy 81.55% when we use

Model	**Accuracy**	**auROC**	**Aupr**	**F1-score**	**MCC**	**Sensitivity**	**Specificity**
LR	81.08%	0.8846	0.8897	0.7855	0.6185	0.7558	0.8569
SVM	82.43%	0.8990	0.9038	0.8006	0.6462	0.7707	0.8694
DST	72.53%	0.7240	0.6298	0.7023	0.4499	0.7091	0.7388
ANN	81.55%	0.8931	0.9005	0.7875	0.6295	0.7537	0.8676
KNN	77.86%	0.8439	0.8243	0.6934	0.5860	0.5520	0.9696
RF	78.25%	0.8600	0.8412	0.7478	0.5621	0.7049	0.8479

Table 4.1: 5-fold Cross-Validation Results.

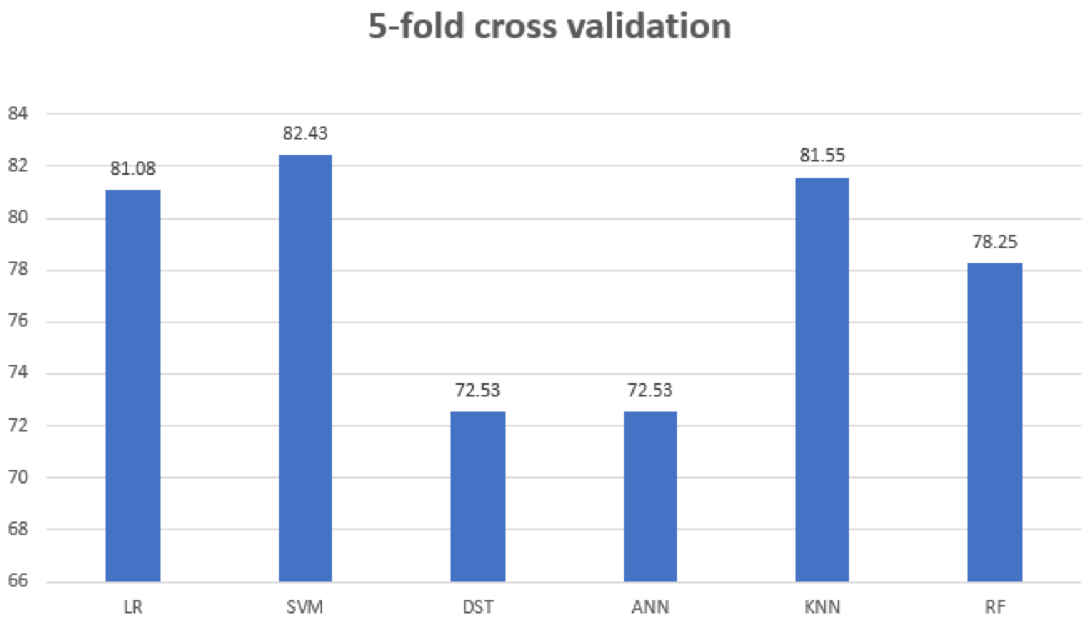


Figure 4.1: Bar Chart of 5-fold cross-validation.

5-fold validation. Fifthly we use k-nearest neighbor (KNN) method. We get accuracy 77.86% when we use 5-fold validation. Thirdly we use a random forest (RF) method. We get accuracy 78.25% when we use 5-fold validation. We get the best result for the SMV method when we use 5-fold cross-validation.

We also present the results from 10-fold cross fold validation in Table 4.2. Figure 4.2 shows a bar chart of the accuracies of the different algorithms used in 10-fold cross fold validation.

At first, we used the linear regression (LR) method. We get accuracy 82.03%

Model	Accuracy	auROC	aupr	F1-score	MCC	Sensitivity	Specificity
LR	82.03%	0.8860	0.8893	0.7946	0.6392	0.7622	0.8694
SVM	81.16%	0.9009	0.9049	0.7866	0.6217	0.7601	0.8551
DST	72.23%	0.7222	0.6239	0.7026	0.4444	0.7219	0.7227
ANN	81.55%	0.8982	0.9036	0.7909	0.6298	0.7622	0.8605
KNN	77.28%	0.8428	0.8290	0.6901	0.5707	0.5563	0.9553
RF	79.81%	0.8679	0.8530	0.7615	0.5967	0.7070	0.8748

Table 4.2: 10-fold cross-validation Results.

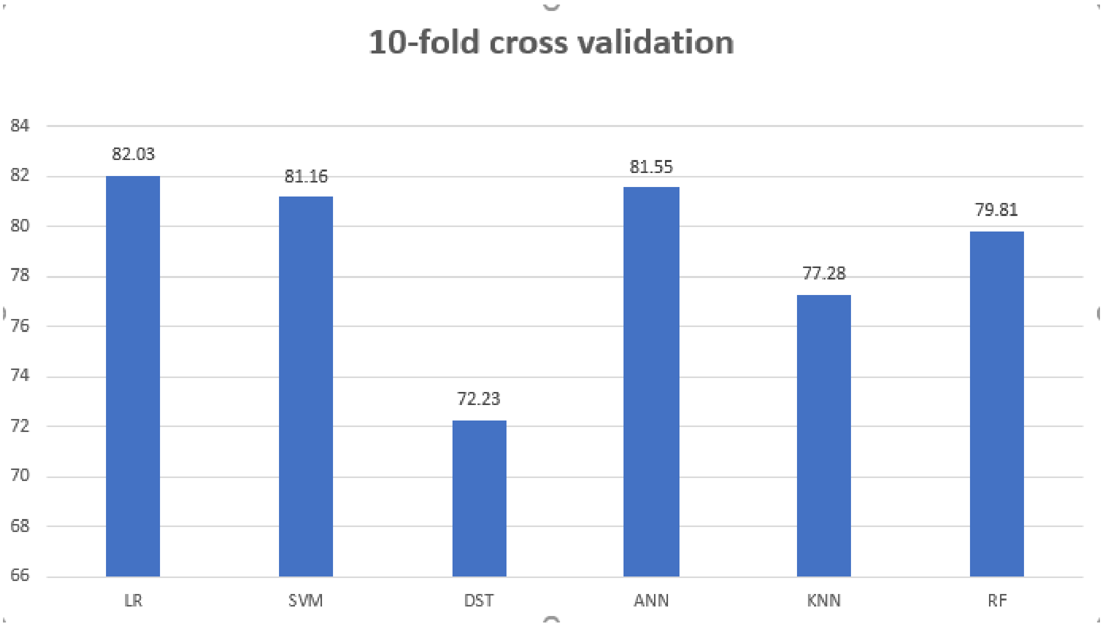


Figure 4.2: Bar Chart of 10-fold cross-validation.

when we use 10-fold cross-validation. Secondly, we use support vector machine (SVM) method. We get accuracy 81.16% when we use 10-fold validation. Thirdly we use decision tree (DST) method. We get accuracy 72.23% when we use 10-fold validation. Fourthly we use an artificial neural network (ANN) method. We get accuracy 81.55% when we use 10-fold validation. Fifthly we use k-nearest neighbor (KNN) method. We get accuracy 77.28% when we use 10-fold validation. Thirdly we use a random forest (RF) method. We get accuracy 89.81% when we use 10-fold validation. We get the best result for the LR method when we use 10-fold cross-validation.

4.2 Reliability to Take SVM

Support Vector Machine is commonly used in bioinformatics. It is a linear model which works for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. Basically, SVM is an extended version of linear classifier where it usually gives a high accuracy rate. For different datasets it can be said that they all needs different analysis. To find out the best machine learning approach we need to do cross-validation with the provided dataset. All machine learning algorithms have their own parameters, depending on it, people need to select the suitable algorithm their classifying dataset. Compare to all other machine learning algorithms it comes to the point that SVM is more accurate. Again, it is more robust; for doing optimal margin gap between separating hyperplanes, SVM could do predictions better than any other predictor with the test data. SVM is also computationally more effective since it uses Kernel trick in dual problem. It can run well on high dimension data. Moreover, it provides less risk of overfitting the model. SVM has the capability to solve global optima. In SVM, it can work with clumsy dataset. Again, it uses various Kernel. Trained classifier mainly uses a portion or works with the specific parts of the training dataset. At the same time other classifiers can work with all training dataset to define the decision function turning them more generalization. On the other hand, SVMs most and the major advantage is the less parameters to tune to make it operational and yielding high accuracy rates. Furthermore, SVM is the best algorithm to work with multivariate numeric data since it results from optimized problem. If the number of features is less, SVM works quite efficiently. But if the number of features is more, then Nave Bayes works more efficiently compare to SVM. It is true that SVM dont works more accurately every time in all situation. ANN-based classifier can provide more accuracy in some cases where SVM shows less accuracy. SVM is the best for linear separable cases. Again, its type of work completion (determining maximum-margin hyperplane) is one of the best for reducing the prediction errors compare to other classifiers.

SVM is faster in training, better in accuracy with stability. That is why we have selected SVM to do our thesis work.

Chapter 5

Conclusion

In this chapter, we discuss challenges, limitations and future work of our thesis.

5.1 Challenges

At first, we have faced the problem of selecting the proper feature list. Because the dataset we have chosen is so much clumsy. After solving this problem, a new problem arises and that is handling imbalanced dataset. There also arises a problem which absolves data. At the end of solving all these problems, we need to select a proper parameter for models. There we have faced the problems of inadequate metrics for train/test to compare models.

5.2 Limitations

At the time of doing our work, we faced some limitations. When we started the work, at first, we faced the problem of changing our dataset. Since, we have a dataset which has many non-numeric values. With whom it seemed difficult to find out an accurate result. Moreover, it made our way more difficult to use this dataset with our algorithms. So, we have change the dataset with some numeric values, which helped us to work with the dataset. Lacking of changing the dataset had made us slower at the first time when we started in our work. In addition, we can say that, we also faced the difficulty to work at the laboratory. If we done our work in the laboratory, we may get more better and accurate result instead of having algorithm-based result, which we have to predict in some cases. Since, we were unable to arrange proper laboratory facilities to do our work, so we cannot assure that our predicted results are 100% accurate and redundancy free. But we can assure the thing that our results are more accurate compared to other works related to our works. It is also mentionable that we do not use any ensemble method to do our working. All these are the limitations which we have faced while doing our work.

5.3 Future Plan

Though we have several limitations while our thesis work, we have decided in that time that in future we will improve our work and we will do more detailed work regarding this thesis work. Since, we faced several limitations and challenges at the time of doing our work, we decided that we will try to overcome our limitations. Correct identification of recombination spots can provide important clues to understand evolution mechanism. Mainly, laboratory approach can produce accurate information to determine recombination spots. Again, we do not use more features to do our work. We will try to improve the number of features which will help us to get more accurate result in future. We also plan to use an ensemble learning approach to improve our testing result. These are the future plan of us to improve our work.

Bibliography

[1] P. Jiang, H. Wu, J. Wei, F. Sang, X. Sun, and Z. Lu, “Rf-dymhc: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features,” *Nucleic acids research*, vol. 35, no. suppl_2, pp. W47–W51, 2007. 7, 8

[2] G. Liu, J. Liu, X. Cui, and L. Cai, “Sequence-dependent prediction of recombination hotspots in saccharomyces cerevisiae,” *Journal of theoretical biology*, vol. 293, pp. 49–54, 2012. 7

[3] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, “irspot-psednc: identify recombination spots with pseudo dinucleotide composition,” *Nucleic acids research*, vol. 41, no. 6, pp. e68–e68, 2013. 8

[4] W.-R. Qiu, X. Xiao, and K.-C. Chou, “irspot-tncpseaac: identify recombination spots with trinucleotide composition and pseudo amino acid components,” *International journal of molecular sciences*, vol. 15, no. 2, pp. 1746–1766, 2014. 8

[5] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, “Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition,” *Analytical biochemistry*, vol. 456, pp. 53–60, 2014. 8

[6] K.-C. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001. 8

[7] K.-C. Chou and H.-B. Shen, “Signal-cf: a subsite-coupled and window-fusing approach for predicting signal peptides,” *Biochemical and biophysical research communications*, vol. 357, no. 3, pp. 633–640, 2007. 8

[8] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, “repdna: a python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects,” *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2014. 8

[9] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, “Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences,” *Nucleic acids research*, vol. 43, no. W1, pp. W65–W71, 2015.

[10] S.-H. Guo, E.-Z. Deng, L.-Q. Xu, H. Ding, H. Lin, W. Chen, and K.-C. Chou, “inuc-pseknc: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition,” *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

[11] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, “ipro54-pseknc: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition,” *Nucleic acids research*, vol. 42, no. 21, pp. 12 961–12 972, 2014. 8

[12] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, “Exploiting three kinds of interface propensities to identify protein binding sites,” *Computational Biology and Chemistry*, vol. 33, no. 4, pp. 303–311, 2009. 8

[13] P. Jiang, H. Wu, J. Wei, F. Sang, X. Sun, and Z. Lu, “Rf-dymhc: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features,” *Nucleic acids research*, vol. 35, no. suppl_2, pp. W47–W51, 2007. 8, 9

[14] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, “irspot-psednc: identify recombination spots with pseudo dinucleotide composition,” *Nucleic acids research*, vol. 41, no. 6, pp. e68–e68, 2013. 9

[15] B. Liu, S. Wang, R. Long, and K.-C. Chou, “irspot-el: identify recombination spots with an ensemble learning approach,” *Bioinformatics*, vol. 33, no. 1, pp. 35–41, 2016.

[16] B. Liu, Y. Liu, X. Jin, X. Wang, and B. Liu, "irspot-dacc: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance," *Scientific reports*, vol. 6, p. 33483, 2016.

[17] G. Liu, J. Liu, X. Cui, and L. Cai, "Sequence-dependent prediction of recombination hotspots in saccharomyces cerevisiae," *Journal of theoretical biology*, vol. 293, pp. 49–54, 2012.

[18] W.-R. Qiu, X. Xiao, and K.-C. Chou, "irspot-tncpseaac: identify recombination spots with trinucleotide composition and pseudo amino acid components," *International journal of molecular sciences*, vol. 15, no. 2, pp. 1746–1766, 2014. 9

[19] J. L. Gerton, J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes, "Global mapping of meiotic recombination hotspots and coldspots in the yeast saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences*, vol. 97, no. 21, pp. 11 383–11 390, 2000. 9

[20] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "Cd-hit: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012. 9

[21] X. Cheng, X. Xiao, and K.-C. Chou, "ploc-mgneg: Predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general pseaac," *Genomics*, vol. 110, no. 4, pp. 231–239, 2018. 13

[22] ——, "ploc-mplant: predict subcellular localization of multi-location plant proteins by incorporating the optimal go information into general pseaac," *Molecular BioSystems*, vol. 13, no. 9, pp. 1722–1727, 2017.

[23] ——, "ploc-mvirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal go information into general pseaac," *Gene*, vol. 628, pp. 315–321, 2017.

[24] X. Cheng, S.-G. Zhao, W.-Z. Lin, X. Xiao, and K.-C. Chou, "ploc-manimal: predict subcellular localization of animal proteins with both single and multiple sites," *Bioinformatics*, vol. 33, no. 22, pp. 3524–3531, 2017.

[25] X. Xiao, X. Cheng, S. Su, Q. Mao, and K.-C. Chou, "ploc-mgpos: incorporate key gene ontology information into general pseaac for predicting subcellular localization of gram-positive bacterial proteins," *Natural Science*, vol. 9, no. 09, p. 330, 2017. 13

[26] X. Cheng, S.-G. Zhao, X. Xiao, and K.-C. Chou, "iatc-misf: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals," *Bioinformatics*, vol. 33, no. 3, pp. 341–346, 2016. 14

[27] ——, "iatc-mhyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals," *Oncotarget*, vol. 8, no. 35, p. 58494, 2017. 14

[28] K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013. 14

[29] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou, "iptm-mlys: identifying multiple lysine ptm sites and their different types," *Bioinformatics*, vol. 32, no. 20, pp. 3116–3123, 2016. 14