iR-Spot : Predicting sequence based recombination hotspot using CONV-1D

Nasif Ishtiaque Islam [1] and Shayed Ashraf[1] and Ashak Mahmud[1] and
Anika Tabassum[1] and Md. Rakibul Haque[2]

***Abstract*— Recombination is the process where two DNA molecules exchange nucleotide sequences with each other. The existence of recombination hotspots offers a way to learn what other processes are associated with recombination. The objective of our work is to find a better predicting model for recombination hotspot. iRSpot starts with DNA sequences for given hotspot and coldspot dataset. We use three feature extraction technique to find important features. Recursive feature elemination and XGboost both are used for feature selection. Model gives 77% accuracy after applying 1D neural network.**

I. INTRODUCTION

Recombination hotspots are the regions within the genome where the rate, and the frequency of recombination are optimum. Hotspot are the regions in a genome that has more clustered recombination. In other ways, highly recombination rates in the region the genome is a hotspot. The opposite is coldspot. Recombination hotspots experience intensely high levels of recombination compared to the genomic background. Recombination provides knowledge about DNA sequence variation and patterns along human chromosomes and this may help to map the position of alleles that cause various diseases. Recombination hotspot gives useful insights into the basic function of inheritance and the study of genetic diversity. Recombinant DNA enables the creation of multiple copies of genes and the insertion of foreign genes into other organisms to give them new traits, such as antibiotic resistance or a new colour.

The objective of our work is to find the optimal way of predicting recombination hotspot using sequence.

In this paper, we are proposing a prediction method iRSpot. This work is using sequence based features. We used several feature elimination techniques to find the optimal features. Xgboost and recursive fearure elemination (RFE) both techniques are used for feature section. The classification technique we used is 1D convolutional neural network.

II. LITERATURE REVIEW

Various methods are used in finding recombination hotspot.

In [1], four methods are used for feature extraction: Nucleotide k-mer composition, Gapped Di-nucleotide composition, TF-IDF of k-mers, Reverse complement k-mer composition. The number of features are 84, 128, 320 and 680. From total 1212 features, after feature selection 17 features are selected. For testing the significance of the feature adaboost algorithm is also performed. 10-fold cross validation is performed on the dataset and then SVM with linear kernel to compute feature set accuracy. SVM(linear kernel) gives 83.82% accuracy where SVM (RBF kernel) gives 84.58% accuracy. KNN, Random forest algorithms are also used to compare performance. Accuracy, sensitivity, specificity, precision and Mathew's Correlation Coefficient are used for performance measure. All the programs and algorithms are in python language using the sci-kit learn library and performed 10 times each. In terms of sensitivity iRSpot-SF achieves a value of 84.57% which is 7.38% improved.

Another work [2], Hexamer(6-mer) distribution is used for different DNA fragmentation. 5-fold cross-validation was adopted (namely K=5) with SVM as the prediction engine. The DNA sequence is represented by a set of 4096 features. The SVM with 5-fold cross-validation was adopted to examine the accuracies of 4096 feature subsets. The result is 84.08% which is almost 13% more improved.

III. METHODS

We present the methodology of our system. We extracted features from our dataset using three different feature extraction method. Then we merged all the features from the extracted features. In feature selection part, we used Xgboost. Conv 1D is used on the total feature set for the accuracy calculation.

The flowchart of the entire methodology is given ub the following diagram 1:

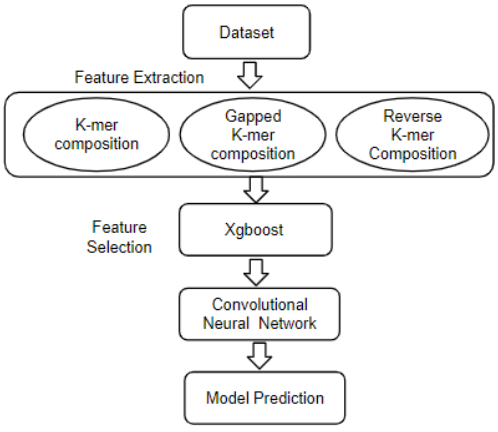

Fig. 1. Model diagram

*This project is a part of Pattern Recognition Laboratory
[1]Student of United International University
[2] Faculty member of United International University

A. Feature Extraction

We used three different types of feature extraction techniques.

1) k-mer composition: K-mer is the substring of any length k in a sequence. Counting K-mer is an essential technique in many bioinformatics methods. It also helps on error corrections of sequence reads. We can understand our dataset better using K-mer. Different combinations of data gives us different insights about the dataset. Space capacity is a matter of concern in K-mer. We calculated upto 4 mer here. We extracted 340 features using k-mer composition.

2) Gapped k-mer composition: In order to find a tradeoff between the sparse feature space problem and more sequence composition information, the gapped k-mer has been proposed. Gapped k-mer allows several gaps to exist in k-mers. Therefore, it cannot only significantly reduce the length of the resulting feature vectors, but also takes the evolutionary process into consideration. Experimental results show that this feature is able to obviously improve the performance for enhancer identification. We used gapped k-mer composition till 5 gap and gives 80 features.

The summary of the feature extraction is in the following table I:

TABLE I
FEATURES SUMMARY

Feature group	number of features
K-mer	340
Reverse K-mer	340
Gapped k-mer	80

3) Reverse k-mer: Reverse k-mercomplement is the the reverse complement of DNA sequence. If the sequence is AATCG, then the complement will be CGATT. Reverse complement k-mer often gives hidden and important information from DNA sequence. Reverse composition is just the reverse form of k-mer composition. Here we used reverse k-mer for extract the features from dataset. Reverse k-mer composition gives 340 features.

B. Feature selection

We used two types of feature selection methods. We combined all the features found from the feature extraction techniques.

1) Recursive Feature elimination: Recursive feature elimination (RFE) is a feature selection method that removes the insignificant feature. RFE only chooses features that are applicable for the prediction. This is an iterative process until the desired number of features are achieved. At first it ranked all the dataset according to their importance and then removed them recursively. From different feature subsets, best features with the highest values are selected.

2) Xgboost: XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.It is a feature selection technique to reduce unnecessary features.In our implementation, we used Xgboost for more faster. Its really fast when compared to other implementations of gradient boosting.

The summary of the feature selection techniques are in the following table II

TABLE II
FEATURE SELECTION METHODS

Selection model	Acc(%)	Sn(%)	Sp(%)	MCC(%)	Pc(%)
Xgboost	74	60	86	49	79
RFE	69	63	74	37	68

C. Classification

We applied 1D convolutional neural network. The model extracts features from sequences data and maps the internal features of the sequence. Convolutional 1D allows to use larger filter sizes. To reduce overfitting we've used Dropout and Regularization. We've also used Batch Normalizer to normalize our data on every step. Finally, Softmax function gives us result on probability format.

We applied 1D CNN on our model after feature selection. Our model gives 77% accuracy after applying 1D convolutional neural network on 760 features.

IV. DATASET

The dataset used here is a yeast dataset consisting of DNA sequences of nucleotides with both positive and negative instances. The positive instances are denoted as hotspot and negative are as coldspots. Dataset has 490 DNA segments of hotspot samples(positive) and 591 DNA segments(negative) of coldspot samples. The basic symbols of DNA sequences are A, T, C, G. This dataset represents the set of these sequences. Dataset is slightly imbalanced with less number of positive samples.

The paper used a yeast DNA sequence dataset. In the paper, CD-HIT is used to reduce the effect of redundancy of similar sequences. After this, the dataset contains a total 1050 samples where 478 are hotspot samples and 572 are coldspot samples.

V. RESULTS AND DISCUSSION

A. Results

Various models are used to find the result. Different kind of classifiers such as : Support vector classifier (SVC), Gaussian Naive bias (NB), Random forest, adaBoost, Lgistic regression (LR), k neighbours, decision tree and Conv1D are performed.

The summary of the classifiers result is in the following table V-A

TABLE III

RESULT OF DIFFERENT ALGORITHMS

Classifier	Sn(%)	Sp(%)	MCC	Pc(%)	Acc(%)
SVC	57.05	**91.48**	0.52	85.00	75.69
NB	60.40	82.39	0.44	74.49	72.31
RF	57.72	90.91	0.52	84.31	75.69
AdaBoost	59.06	81.25	0.41	72.73	71.08
LR	**61.74**	71.02	0.33	64.74	66.77
KNN	55.70	79.55	0.36	69.75	68.62
DT	58.39	63.07	0.21	57.24	60.92
CNN-1D	60.04	89.97	**0.53**	**85.55**	**76.41**

B. Discussion and analysis

The confusion matrix is used to evaluate accuracy. The confusion matrix gives an output matrix and provides the description of performance of the system. The samples are from two classes: 'True' and 'False,' and the implementation of the confusion matrix can be done. There are 4 terms: 1) True positive 2) True negative 3) False positive 4) False negative. The equations of the evaluation matrix are given bellow:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP} + FP \tag{4}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

DT was the poor performer in terms of MCC and accuracy. But SVC and RF performs well in terms of Sp and Sn respectively. These two algorithms also performs well in terms of MCC.However, we selected CNN-1D as the best performing classifier and suitable for our method as its accuracy was highest of 76.41% and superior to RF in terms of MCC and Pc which are 0.53 and 85.55respectively.

We compared the performance of iRecSpot-CNN-1D with other state-of-the-art predictors who used the same datasets.

Performance comparison of our iRSpot with difffertent existing model is given in the following table IV:

From the results shown in table IV, it is evident that our proposed method could not outperforms all these methods in terms of all the evaluation metrics considered for the experiments. Classification methods and feature selection approaches employed by other state-of-the-art pre-dictors is also given in table IV.

TABLE IV

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART PREDICTORS

Methods	Sn(%)	Sp(%)	MCC	Acc(%)
iRSpot-TNCPseAAC	76.56	70.99	0.4737	73.52
iRSpot-PseDNC	71.75	85.84	0.5830	79.30
IDQD	79.52	81.82	0.6160	80.77
iRSpot-ADPM	77.19	90.73	0.6905	84.57
iRSpot-SF	84.57	75.76	0.6941	84.58
iRSpot-EF	94.35	95.80	0.9037	95.14
iRSpot-CNN-1D	**60.04**	**89.97**	**0.5345**	**76.41**

1) Roc curve: We made AUC-ROC curve. For mechine learning problems, AUC-ROC curves are one of the most important evolution metrics for checking classification model's performance. We use AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. The Roc curve is given in the following figure 2:

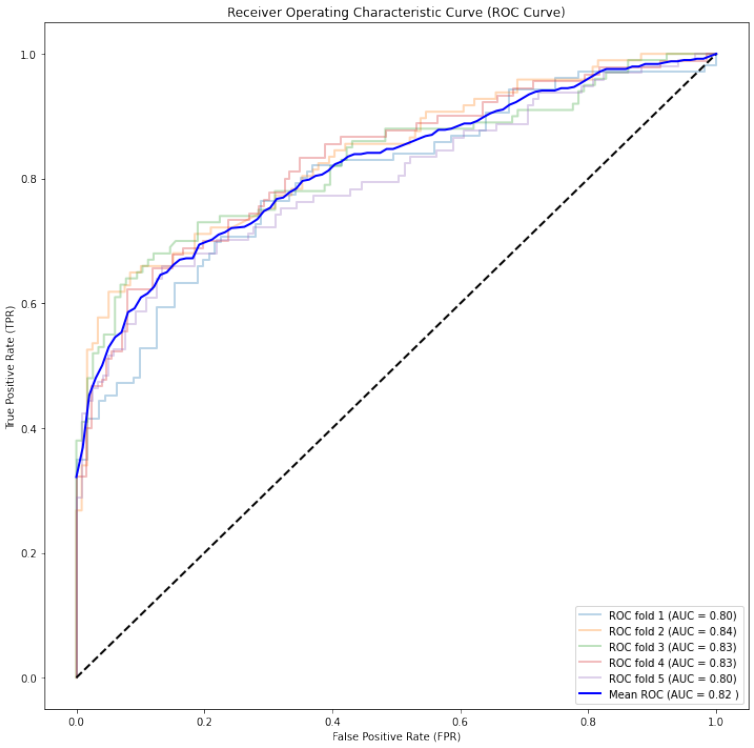

Fig. 2. Roc curve

VI. CONCLUSIONS

We design iRSpot that can predict recombination hotspots. We choose important features using feature elimination techniques. Two different feature slection methods are used to find best result. We use conv1D on total features as conv1D needs a huge number of features for better performance. We wish to improve our model in future.

ACKNOWLEDGMENT

We would like to thank Md Rakibul Haque Sir (faculty member, Department of CSE, United International University) for his mentorship, guidance, and support throughout this research project.

REFERENCES

[1] Al Maruf, Md Abdullah and Shatabda, Swakkhar, "iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components," in Genomics, vol. 111, no. 4, pages: 966–972, Elsevier, 2019.

[2] 2 Yang, Hui and Qiu, Wang-Ren and Liu, Guoqing and Guo, Feng-Biao and Chen, Wei and Chou, Kuo-Chen and Lin, Hao, "iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC" in International journal of biological sciences, vol. 14, no. 8, page: 883, Ivyspring International Publisher, 2018

[3] 3Jiang, Peng and Wu, Haonan and Wei, Jiawei and Sang, Fei and Sun, Xiao and Lu, Zuhong"'RF-DYMHC: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features"' in Nucleic acids research, Oxford University Press, 2007

[4] 4Chen, Wei and Feng, Peng-Mian and Lin, Hao and Chou, Kuo-Chen"'iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition"' in Nucleic acids research, vol. 41, Oxford University Press, 2013

[5] 5Qiu, Wang-Ren and Xiao, Xuan and Chou, Kuo-Chen "'iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components"' in International journal of molecular sciences, vol. 15, no 2, Multidisciplinary Digital Publishing Institute, 2014

[6] 6 Liu, Bin and Wang, Shanyi and Long, Ren and Chou, Kuo-Chen"'iRSpot-EL: identify recombination spots with an ensemble learning approach"' in Bioinformatics, vol. 33, no. 1, page. 35-41, Oxford University Press, 2017

[7] 7 Zhang, Lichao and Kong, Liang "'iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components "' in Journal of theoretical biology, vol. 441, page. 1-8, Elsevier, 2018

[8] 8 Zhang, Lichao and Kong, Liang "'iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components"' in Genomics,vol. 111, no. 3, Elsevier, 2019

[9] 9 Jani, Md Rafsan and Mozlish, Md Toha Khan and Ahmed, Sajid and Tahniat, Niger Sultana and Farid, Dewan Md and Shatabda, Swakkhar "'iRecSpot-EF: Effective sequence based features for recombination hotspot prediction"' in Computers in biology and medicine, vol. 103, page. 17-23, Elsevier, 2018