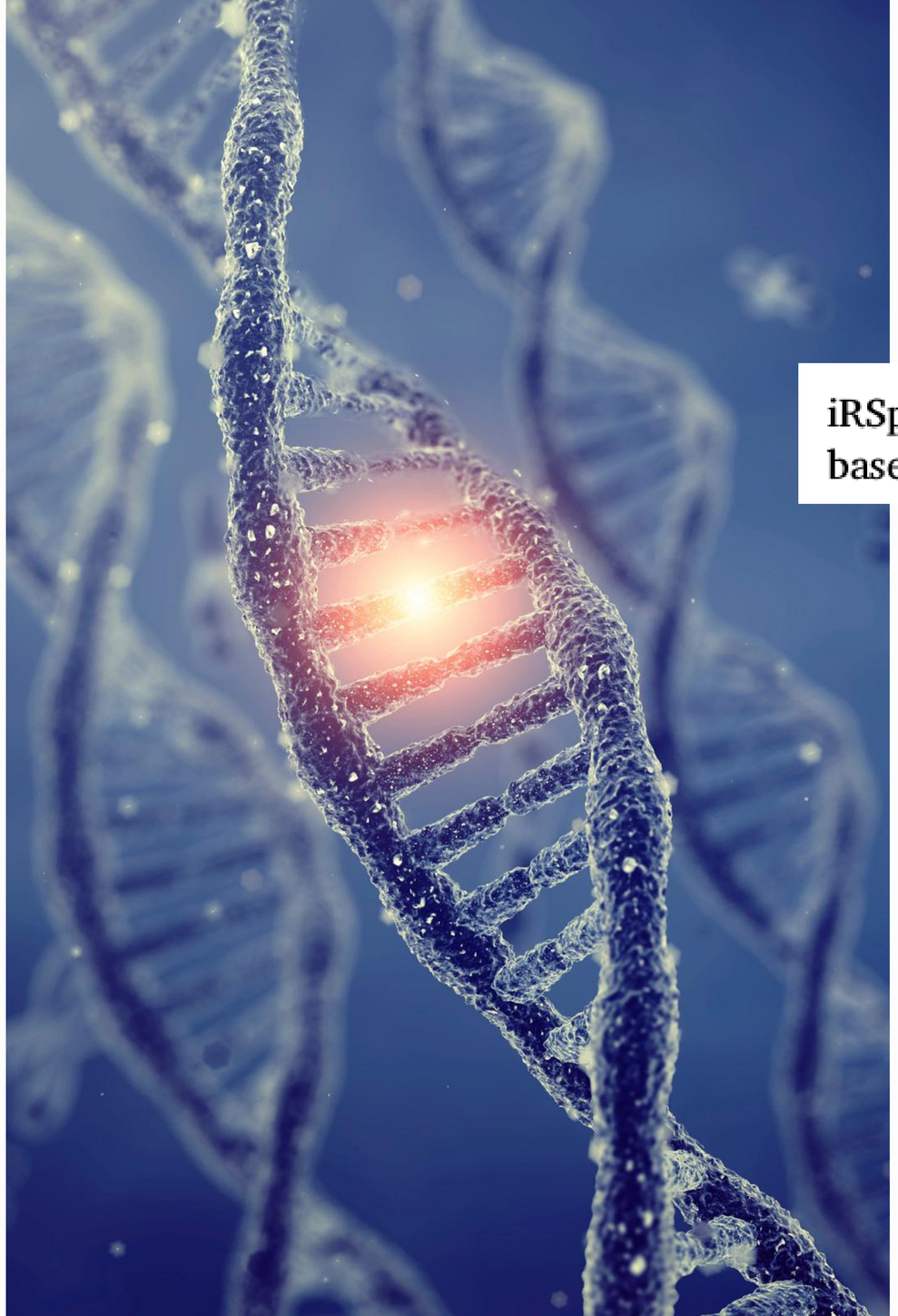


CSI416 - PATTERN RECOGNITION LAB

Team name: black_crow

Topic

iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components

Team Members

Anika Tabassum (011161150)
Ashak Mahmud (011161144)
Shayed Ashraf (011171202)
Nasif Ishtiaque Islam (011171223)

Course Teacher

Md. Rakibul Haque
Lecturer, Dept of CSE, UIU

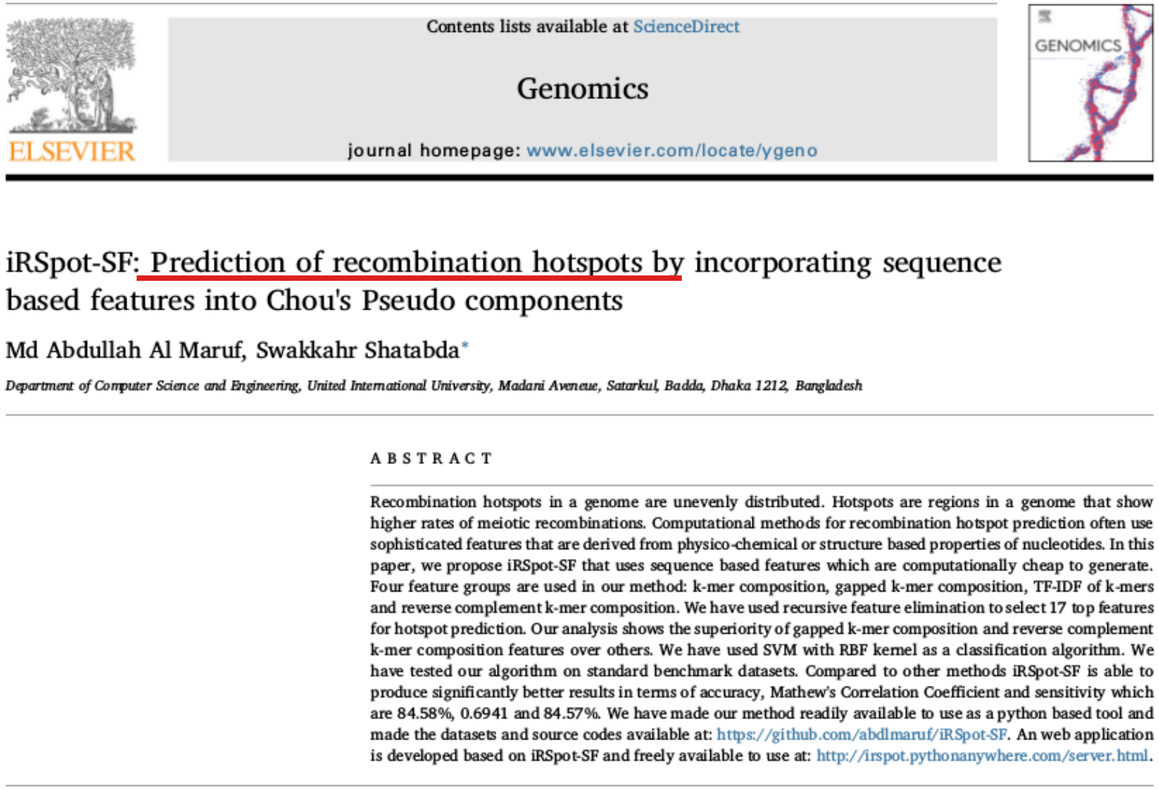
Introduction

What is recombination?

- Recombination is the process where two DNA molecules exchange nucleotide sequences with each other.

Importance of recombination:

- 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.



Contents lists available at ScienceDirect

Genomics

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

iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components

Md Abdullah Al Maruf, Swakkahr Shatabda*

Department of Computer Science and Engineering, United International University, Madani Avenue, Satarkul, Badda, Dhaka 1212, Bangladesh

ABSTRACT

Recombination hotspots in a genome are unevenly distributed. Hotspots are regions in a genome that show higher rates of meiotic recombinations. Computational methods for recombination hotspot prediction often use sophisticated features that are derived from physico-chemical or structure based properties of nucleotides. In this paper, we propose iRSpot-SF that uses sequence based features which are computationally cheap to generate. Four feature groups are used in our method: k-mer composition, gapped k-mer composition, TF-IDF of k-mers and reverse complement k-mer composition. We have used recursive feature elimination to select 17 top features for hotspot prediction. Our analysis shows the superiority of gapped k-mer composition and reverse complement k-mer composition features over others. We have used SVM with RBF kernel as a classification algorithm. We have tested our algorithm on standard benchmark datasets. Compared to other methods iRSpot-SF is able to produce significantly better results in terms of accuracy, Mathew's Correlation Coefficient and sensitivity which are 84.58%, 0.6941 and 84.57%. We have made our method readily available to use as a python based tool and made the datasets and source codes available at: https://github.com/abdlmaruf/iRSpot-SF. An web application is developed based on iRSpot-SF and freely available to use at: http://irspot.pythonanywhere.com/server.html.

Introduction

Hotspot and Coldspot?

In genomic regions:

Hotspots → higher frequencies of recombination

Coldspots → lower frequencies of recombination

Recombination hotspot plays a vital part in evolutionary development.

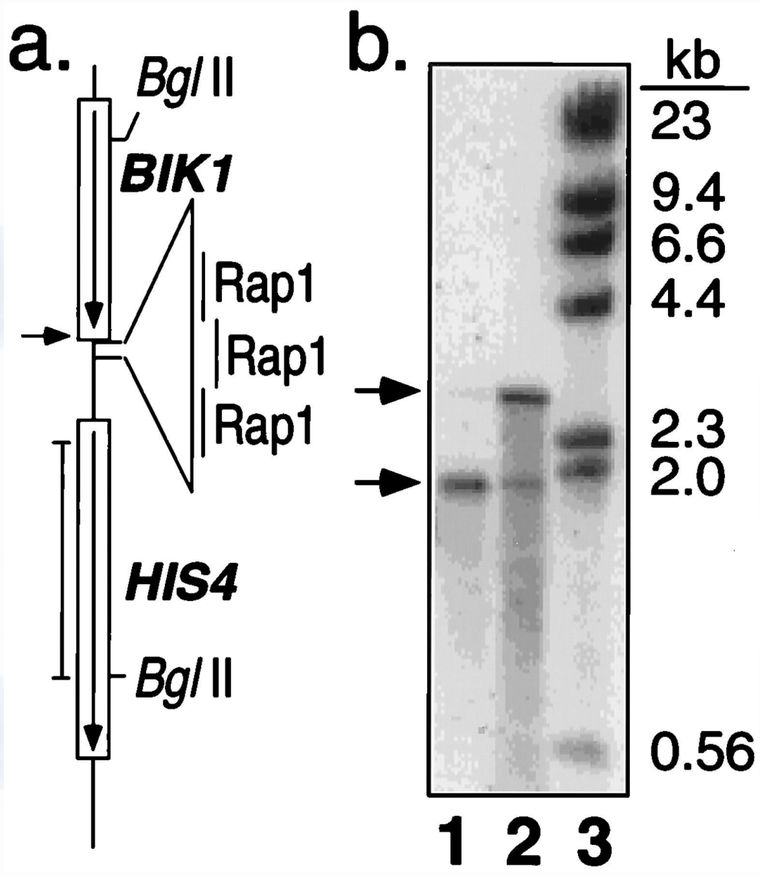


image from Internet

Objectives

- Understand the dataset that has been used in the paper.
- Analyze the benchmark dataset part of the following paper in order to gain a better understanding.
- Learn to map those sequences in vector form.
- Use feature elimination technique to reduce feature.
- Fit classification algorithm on the model.
- Try ANN to get better result.
- Analyze the predicted result of the model.

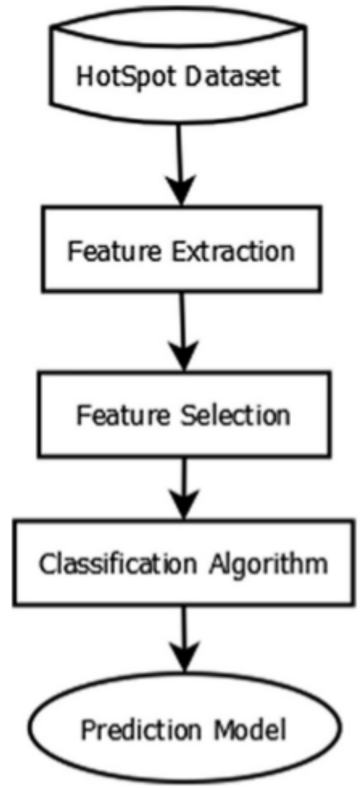


Figure 1: System Diagram

Real Life Application

- 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.

- This technology is also an important tool in agriculture, being used to improve plants' resistance to pests and increase crop yields.

- The more we learn DNA, we'll be more close to medical science.

Challanges

- Eliminating weak features
- Fitting ANN into our model
- Hyperparameter tuning
- Generating prediction model

Literature Review

- Four methods are used for feature extraction on our selected paper: Nucleotide k-mer composition, Gapped Di-nucleotide composition, TF-IDF of k-mers, Reverse complement k-mer composition.
- 10-fold cross validation is performed on the dataset and then SVM with linear kernel to compute feature set accuracy.
- KNN, Random forest algorithms are also used to compare performance.
- 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.

Methodology

- **Dataset Description:** 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) and591 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.

- **K-mer:** K-mer is the substring of any length k in a sequence. Counting K-mer is an essential technique in many bioinformatics methods.

- **Gapped-Nucleotide K-mer :** 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.

- **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.

Future Work

Predicted result using ANN

Improve the performance of overall system using parameter tuning

Implement a web application for our work

Submit our work in a journal

Conclusion

At first, we've selected the paper.
Then we understand the dataset.
In our selected paper, they used iRSpot-SF that uses sequence based features which are computationally cheap to generate.
We already review the literature, background studies, dataset understanding.
For feature extraction we used k-mer composition, trying to use gapped k-mer composition, and implemented recursive feature elimination technique.
We will use SVM with RBF kernel as a classification algorithm and finally using ANN we will predict the accuracy of this model.

It's question time

THANK YOU!