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

Paper Summary:

Recombination hotspot: Recombination hotspots are the regions within the genome where the rate, and the frequency of recombination are optimum.

Hotspot: 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.

In this paper, 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.

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) 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.

Individual members opinions on the dataset:

Anika Tabassum (011 161 150):

The dataset used in this paper is a sequence of DNA segments of yeast. A number of 490 positive and 591 negative samples are in the dataset. These represent the hotspots and coldspots. The dataset is a little bit imbalanced as the number of negative samples or the coldspots are smaller. The sequence of A, T, C, G which are the basic DNA symbols are present in the dataset.

Ashak Mahmud ID:011 161 144:

Here in this dataset ,we are working with DNA sequence and the DNA sequence contains A-adenine, G-guanine, C-cytosine, and T-thymine.
The dataset used here consists of DNA sequences of nucleotides with both positive and negative instances.The positive instances are denoted as hotspot and negative are as coldspots.
It can be formulated as following:

$$S= S+ \cup S-$$

Where,
S=DNA sequences of nucleotides.
S+=The set of positive instances or recombination hotspots.
S-=The set of negative instances or recombination coldspots.

Nasif Ishtiaque Islam ID (011171223):

Here on this paper, we have two datasets. One is hotspot and the other is coldspot. Here CD-HIT used to reduce redundancy. Each dataset contains DNA sequences.

Shayed Ashraf (011 171 202):

In this paper, we have proposed iRSpot-SF a method for prediction of recombination hotspots using sequence based features. So we used two datasets for running the model. In this dataset we have two types of instances, one is coldspot and another is hotspot.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.

Md. Shahadat Hossen(011 171 243):

The dataset is consisting of two types of instances. One of these two instances is positive instances or Recombination hotspots with 478 sample sequences and other instance is negative instances or coldspots with 572 sample sequences. Dataset is imbalanced. The DNA sequence consists of four alphabetical symbols (A, C, G, T). To extract the feature, this paper used “Nucleotide Composition”, “g-gapped Di-nucleotide Composition”, and “TF-IDF of K-mers” &“Reverse Complement Composition method. From the total 1212 feature, Feature selected Algorithm selects 17 most effective features. To reduce number of feature, this paper used “KNN”, “SVM (linear kernel)”, “SVM (RBF kernel)” & RF algorithms where SVM (RBF kernel) gives the best accuracy which is 84.58%. This paper measures the performance by using ‘Accuracy’, ‘Sensitivity’, ‘Specificity’ and ‘Precision’