Genomics xxx (xxxx) xxx–xxx

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 Aveneue, Satarkul, Badda, Dhaka 1212, Bangladesh

A B S T R A C T

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.

1. Introduction

Meiotic recombination plays an important role in human genome evaluation. However, the distribution of recombination along the genome is uneven. Hotspots are small regions in a genome where recombinations are more clustered or where recombination rates are high [2]. On the other hand, the regions with lower rates of recombinations are known as cold spots. In vitro method to find recombination hotspots uses DNA binding array to find initiation sites for recombination [2]. However, a large number of computational methods are proposed in the literature in recent years.

Most of the computational methods in the literature of recombination hotspot prediction have formulated the problem as a supervised learning problem of binary classification. Many classification algorithms are found to be used in the literature. They include Random Forest (RF) [35], Support Vector Machines (SVM) [47], Neural Network [36], Ensemble Methods [42], etc. Most of these methods make use of two types of features: sequence based features and derived features based on structure or other properties. Among sequence based features are gapped dinucleotide compositions [35], hexamer positions [59], k-mer frequencies [47], etc. Among other derived features are pseudo nucleotide composition [4,37], pseudo-amino acid compositions [51], dinucleotide based auto-cross covariance (DACC) [46], etc.

One of the first computational methods was proposed in [35]. They introduced the benchmark dataset for hotspot prediction and used gapped dinucleotide composition features with Random Forest to develop their method called RF-DYMHC. The increment of diversity combined with quadratic discriminant analysis (IDQD) was used in [47] with k-mer frequencies. Chen et al. [4] proposed iRSpot-PseDNC which used local structural properties of DNA intro pseudo dinucleotide composition and SVM as classification algorithm. Qui et al. used tri-nucleotide composition and pseudo amino acid components and SVM to develop their method named iRSpot-TNCPseAAC [51]. Li et al. [37] used a fusion of pseudo nucleic acid composition and SVM. Genetic algorithm based ensemble model iRSpot-GAEnsC was proposed in [36] that used di and tri nucleotide composition with probabilistic neural networks. An ensemble based method iRSpot-EL was proposed in [42]. A comparative study of different classification algorithms to solve hotspot prediction problem was presented in [26].

In a recent work by Liu et al. [45] an ensemble based method SVM-EL was proposed that used three types of features: k-mer frequencies, DACC and pseudo dinucleotide composition. In a subsequence work, they proposed iRSpot-DACC [46] that used the same set of features and SVM as classification algorithm. Application of z-curve method was

* Corresponding author.
E-mail address: swakkhar@cse.uiu.ac.bd (S. Shatabda).

https://doi.org/10.1016/j.ygeno.2018.06.003
Received 4 April 2018; Received in revised form 9 June 2018; Accepted 13 June 2018

presented as a case study to show improvements in predicting hotspot prediction in [25]. Another recent method was proposed in [62] called iRSpot-ADPM. They used feature selection techniques to select an optimal set of 85 features for effective prediction of hotspots. The same authors have proposed iRSpot-PDI [63] using diversity information of di-nucleotides. Another recent work in this area is iRSpot-Pse6NC [59], where the authors have used general PseKNC as features to identify recombination hotspots in *Saccharomyces cerevisiae*.

In this paper, we propose iRSpot-SF, a novel recombination hotspot prediction method using sequence based features. We use several sets of sequence based features that are computationally easy to generate. Recursive feature elimination technique is used to select only a small set of effective features. We have selected Support Vector Machine with RBF kernel as our classification algorithm. On a standard benchmark dataset, iRSpot-SF is able to produce significantly better results in terms of standard performance metrics. We have made our source codes available for other researchers to use at: https://github.com/abdlmaruf/iRSpot-SF.

2. Materials and methods

In this section, we present our proposed method, iRSpot-SF. This section is presented as suggested in [16] for attribute prediction of biological sequences. Firstly, we describe the benchmark dataset selected for this problem, followed by a description of the feature extraction process and different features generated for iRSpot-SF. Then we present feature selection and classification algorithm used by iRSpot-SF followed by a description of the performance measures and evaluation techniques. The tool implemented for this paper is described in the results and discussion section.

A simple system diagram of iRSpot-SF is given in Fig. 1. It starts with a dataset for recombination hotspot detection consisting negative and positive instances. Sequence based features are generated for each of these instances using a feature extraction module. Extracted features are then fed into a feature selection module where only effective features are selected. Using the selected features, the reduced dataset is then fed into a classification algorithm and a predicted model is stored for testing and prediction phase. The predicted model is output after cross-validation and saved for further use. The tool generates only the selected features for any test instance and using the stored model predicts whether the given sequence is a recombination hotspot or cold spot.

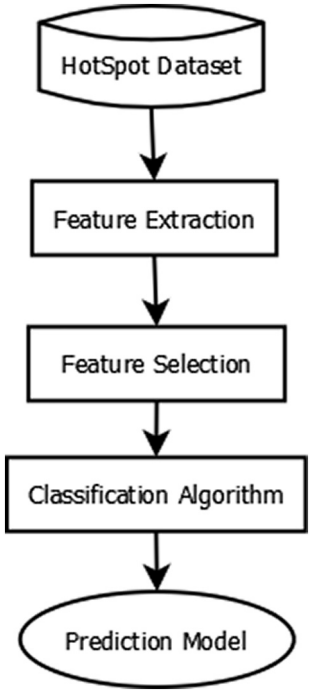

Fig. 1. A system diagram for iRSpot-SF.

2.1. Dataset description

The most important part of supervised machine learning based computational methods is the construction or selected of a benchmark dataset. For any binary classification task as recombination hotspot prediction task, the dataset consists of positive and negative instances. It can be formulated as following:

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \quad (1)$$

Here, $\mathbb{S}$ is the total dataset that consists DNA sequences as strings of nucleotides. The set of positive instances or recombination hotspots is denoted by $\mathbb{S}^+$ and the set of negative instances or recombination coldspots is denoted by $\mathbb{S}^-$. The dataset selected in this paper is taken from [35] and widely used in the literature of recombination hotspot prediction [4,47,51,42,46]. In total, 490 DNA segments of hotspot samples and 591 DNA segments of coldspots were selected for the dataset. The recommendation for selecting these instances was following the suggestions in [29]. To reduce the effects of homology and redundancy of similar sequences in the datasets, CD-HIT [27] was used to remove sequences with > 75% similarity. The resulting dataset consists of 478 sequences that are positive samples or hotspots and 572 sequences that are coldspots. A summary of the dataset is given in Table 1.

2.2. Feature extraction

A number of different features are used in the literature of recombination hotspot prediction. Among them are dinucleotide based auto cross-covariance [46], pseudo k-tuple nucleotide composition [42], etc. It is important to note that all these methods despite their effectiveness and success in prediction of hotspots in DNA sequences are different forms of Chou's general form [8] for DNA/RNA features. Also, note that they depend on different structural properties of the nucleotide compositions.

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elucidated in a comprehensive review [18]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition or PseAAC [15] was proposed. Ever since the concept of PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics [23,3,49,54,19]. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its concept has been extended to deal with DNA/RNA sequences by the PseKNC (Pseudo K-tuple Nucleotide Composition) [5, 6, 44] and proved very useful as well. Particularly, recently a very powerful web-server called ˜Pse-in-One [41] and its updated version Pse-in-One [43] have been

Table 1
Summary of the dataset.

Class	Number of instances	Relative hybridization ratio
Hotspot	478	> 1.5
Coldspot	572	< 0.82
Total	1050	

Table 2
Summary of features used in this paper.

Feature Group	Nnumber of Features
Nucleotide Composition	84
g-gapped Di-nucleotide Composition	128
TF-IDF of k-mers	320
Reverse Complement Composition	680

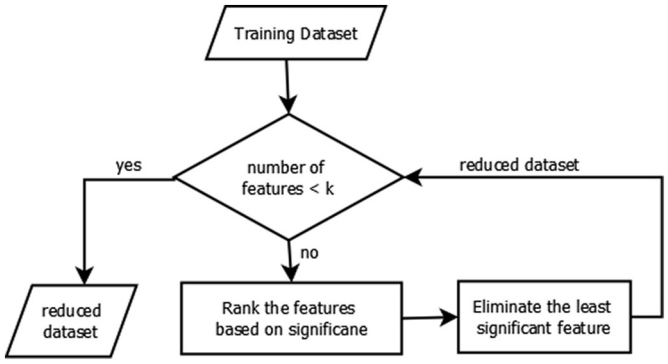

Fig. 2. Block diagram of recursive feature elimination procedure.

Table 3
Summary of selected features.

No	Feature	Feature Group
1	T***A	Gapped dinucleotide
2	C****C	Gapped dinucleotide
3	G****C	Gapped dinucleotide
4	C*****C	Gapped dinucleotide
5	G******G	Gapped dinucleotide
6	RC(CGCC)	Reverse complement k-mer composition
7	RC(CTAA)	Reverse complement k-mer composition
8	RC(AAAAG)	Reverse complement k-mer composition
9	RC(AGATA)	Reverse complement k-mer composition
10	RC(CCCAC)	Reverse complement k-mer composition
11	RC(CGCAC)	Reverse complement k-mer composition
12	RC(CTAAG)	Reverse complement k-mer composition
13	RC(GGCAC)	Reverse complement k-mer composition
14	RC(GGCCA)	Reverse complement k-mer composition
15	RC(TATAA)	Reverse complement k-mer composition
16	RC(TATCA)	Reverse complement k-mer composition
17	RC(TATGA)	Reverse complement k-mer composition

Table 4
Comparison of the performance by selected 17 features by RFE with other feature selection method and full feature set without any feature selection.

Classifiers	Acc(%)	P_c(%)	S_n(%)	S_p(%)	MCC	AUC
No Feature Selection	80.18%	80.43%	80.17%	78.01%	0.6027	0.886
Ranking (AdaBoost)	81.72%	82.13%	81.72%	74.70%	0.634	0.8946
RFE	84.58%	85.27%	84.57%	75.97%	0.6941	0.8937

established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users studies. PseKNC have been successfully used in prediction of nucleosome position [31], promoters [39, 40], replication origin region [61] and epigenetic modification sites [7]. In the current study, we are to use some most effective features to define pseudo components for analyzing the recombination spots.

In this paper, we use only sequence based features which is cheap to generate and effective in use. In this section, we describe the sequence based features used in this paper. Let's assume a DNA sequence in the dataset $\mathbb{S}$ is denoted by D. Given the length of this sequence is L, formally it can be represented as following:

$$D = N_1 N_2 \cdots N_L \tag{2}$$

here, $N_i \in \Sigma$ are nucleotides of the DNA sequence and the alphabet consists four symbols, $\Sigma = \{A, C, G, T\}$. All the features presented in this paper are based solely on this sequence. A summary of the featurres used in this paper is given in Table 2.

2.2.1. Nucleotide k-mer composition

A k-mer is a string of k literals taken from the DNA alphabet $\Sigma = \{A, C, G, T\}$. In this paper, we have considered k-mers, with $k = 1, 2, 3$. Nucleotide k-mer composition can be formally defined as following:

$$\text{k-mer composition}(s_i) = \frac{1}{L} count(s_i), \forall\ k = 1, 2, 3, s_i \in \Sigma^k \tag{3}$$

here, $Count(s_i)$ is the number of occurrences of the k-mer $s_i \in \Sigma^k$ in the given input DNA sequence D. Composition is found after normalizing it by the length of the DNA sequence or segment.

2.2.2. Gapped dinucleotide composition

Gapped di-nucleotide compositions were first proposed in the literature in [30] for regulatory sequence prediction and later used for recombination hotspot prediction in [57]. However, it is to be noted that the concept of gapped composition was earlier in literature of protein attribute prediction [55, 24, 38] by using gapped dipeptide compositions. In this paper, we use the composition of gapped di-nucleotides with gaps, $g = 1, 2, \cdots, 8$. A g-gapped di-nucleotide (g-gapped 2-mer) is a string of the form uvw, where $u \in \Sigma$ and $v \in \Sigma$ are nucleotides that compose the di-nucleotide and $v \in \Sigma^g$ is any string of length g over the alphabet. This feature can be formally defined as following:

$$\text{g-gapped 2-mer composition}(s^g) = \frac{1}{L} count(s^g), \forall\ g = 1, 2, \cdots, 8 \tag{4}$$

2.2.3. TF-IDF of k-mers

Term frequency(TF) and inverse document frequency (IDF) are used in information retrieval [1] and recently been used in genome sequence attribute prediction [28]. We have used k-mers with $k \in \{3, 4\}$. TF is the normalized frequency or composition of a string in a given sequences (defined by Eq. 3). However, IDF searches for k-mers that are sparsely distributed in the sequences. It is formally defined as below:

$$IDF(s_i) = log\left(\frac{|\mathbb{S}|}{\sum_{D \in \mathbb{S}} d(s_i, D)}\right) \tag{5}$$

Here, $|\mathbb{S}|$ is the number of sequences in the dataset $\mathbb{S}$ and $d(s_i, D)$ is a function that returns 1 if and only if the k-mer $s_i \in \Sigma^k$ is present in the sequence D, and otherwise 0. TF-IDF is the multiplication of TF and IDF for each of the k-mers. To the best of our knowledge, this is the first application of TF-IDF features for recombination hotspot prediction.

2.2.4. Reverse complement composition of k-mers

Its often observed in genome sequences that hidden patterns and their reverse complements are both present. Such patterns often provides important regulatory information [48]. Reverse complement composition of k-mers are the normalized count of k-mer and its reverse complements present in the DNA sequence.

$$\text{k-mer RC composition}(s_i) = \frac{1}{L} count_R(s_i), \forall\ k = 3, 4, 5, s_i \in \Sigma^k \tag{6}$$

The difference of this feature with that defined in Eq. 3 is that here, $count_R(s_i)$ with count all the occurrences of the k-mer and its reverse complement.

2.3. Feature selection

We merged all the features generated using our feature generation methods and created a huge feature set. As we extract these features

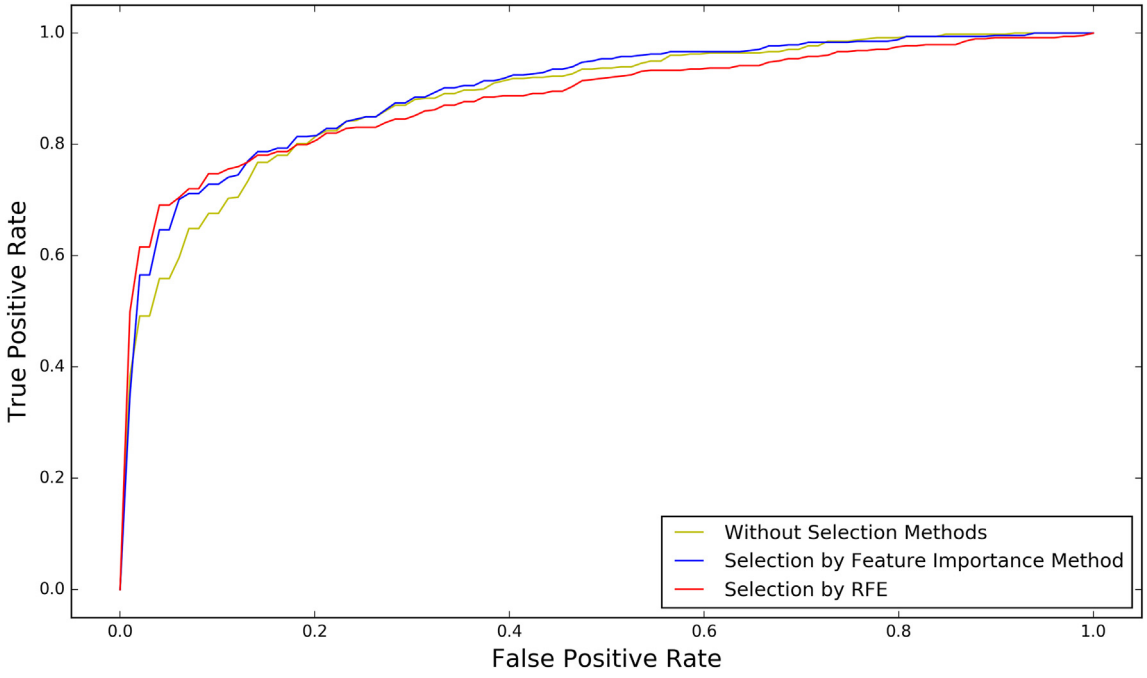

Fig. 3. Receiver Operating Characteristic (ROC) curve for different feature selection methods and no feature selection.

Table 5
Comparison of performances of different classification algorithms on the benchmark dataset using the selected feature.

Classifiers	Acc(%)	P_c(%)	S_n(%)	S_p(%)	MCC	AUC
kNN	82.19%	84.07%	82.18%	66.97%	0.6553	0.8685
SVM(Linear kernel)	83.82%	84.34%	83.81%	75.76%	0.6773	0.8828
SVM(RBF kernel)	84.58%	85.27%	84.57%	75.97%	0.6941	0.8937
RF	82.58%	83.06%	82.57%	75.77%	0.6522	0.8796

directly from the sequences of DNA, they are very easy to generate compared to physico-chemical properties or structure based features and pseudo-nucleotide compositions. But after merging the feature sets to a unified feature set, it has 1212 features which are very time-consuming to learn. We have used a recursive feature elimination technique. Recursive feature elimination (RFE) [32] depicted in Fig. 2 recursively selects and eliminates the features until it reaches its maximum performance in accuracy. During each iteration, the dataset is assessed with cross-validation of 10 folds and then SVM with linear kernel is used to compute the accuracy of the feature set. After some iteration, our feature selection algorithm gives us maximum accuracy from our feature set with just 17 features. The details of the features

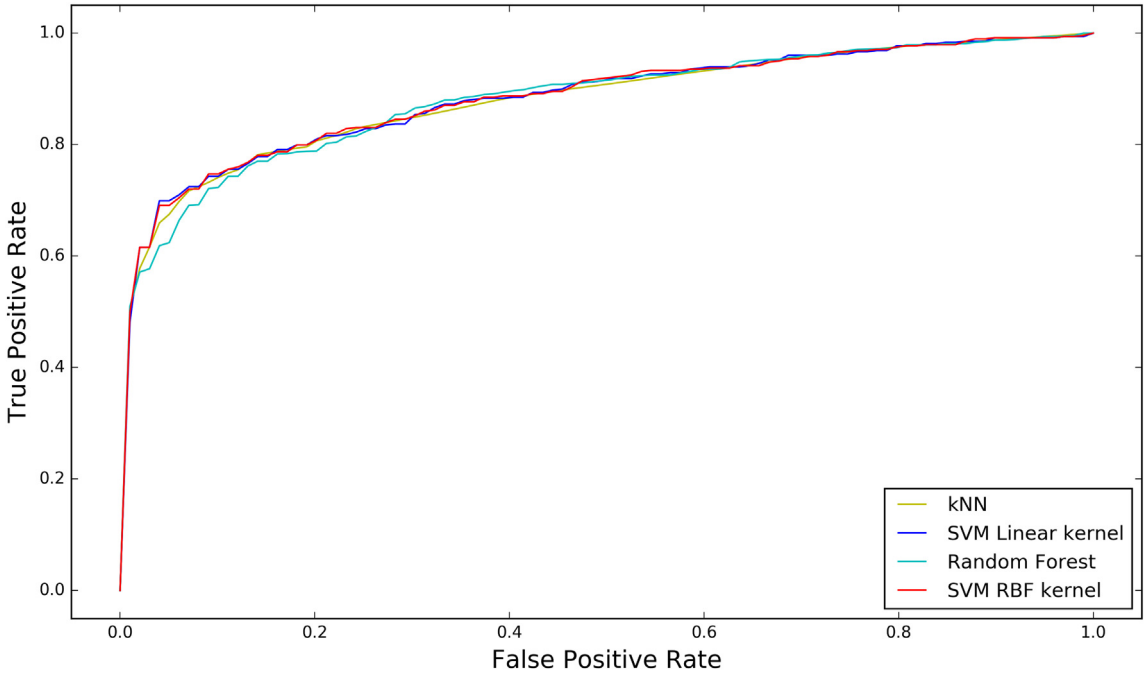

Fig. 4. Receiver Operating Characteristic (ROC) curve for different classification algorithms on the selected optimal feature set.

Table 6
Comparison of performance of iRSpot-SF with other state-of-the-art predictors using cross-fold validation.

Methods	S_n(%)	S_p(%)	MCC	Acc(%)
RF-DYMHC	73.01%	86.56%	0.6049	80.40%
IDQD	79.52%	81.82%	0.6160	80.77%
iRSpot-PseDNC	71.75%	85.84%	0.5830	79.33%
iRSpot-TNCPseAAC	76.56%	70.99%	0.4737	73.52%
iRSpot-EL	75.29%	88.81%	0.6510	82.65%
iRSpot-ADPM1575	74.88%	90.04%	0.6613	83.14%
iRSpot-ADPM	77.19%	90.73%	0.6905	84.57%
iRSpot-PDI	71.48%	92.56%	0.6658	83.16%
iRSpot-SF	84.57%	75.76%	0.6941	84.58%

Table 7
Comparison of performance of iRSpot-SF with other state-of-the-art predictors using jack knife test.

Methods	S_n(%)	S_p(%)	MCC	Acc(%)
iRSpot-PseDNC	73.06%	89.49%	0.6380	82.04%
iRSpot-DACC	75.71%	88.16%	0.6470	82.52%
iRSpot-GAEnSc	73.77%	79.92%	0.6620	83.44%
iRSpot-ADPM1575	74.08%	90.69%	0.6620	83.16%
iRSpot-ADPM	75.51%	90.52%	0.6730	83.72%
iRSpot-PDI	71.63%	93.91%	0.6810	83.81%
iRSpot-SF	75.31%	91.08%	0.6774	83.91%

selected are given in the next section. The reduced set of features then used for learning and predicting the recombination hotspots.

2.4. Classification algorithm

We have selected support vector machines (SVM) [22] as our classification algorithm for iRSpot-SF. SVM tries to separate two classes in a binary classification problem using a maximum margin. The problem of finding the decision boundary is an optimization problem. SVM's often use kernel functions that convert the feature space and enables non-linear classification. We have tried linear and RBF kernels for our method. Two other important parameters for SVM are γ and C. They were selected using grid search. We have also used K-Nearest Neighbor algorithm (KNN) and Random Forest (RF) [33] algorithms in order to compare the performance of SVM with them.

2.5. Performance evaluation

In order to assess the performance of classification algorithms, it is very important to fix a sampling method on the datasets. Three of the most popular techniques are: independent test set, cross-fold validation and jack-knife test. In order to compare the performance of iRSpot-SF with the other state-of-the-art methods, we have used cross-fold validation technique. To make sure that we are not overfitting our model on the dataset, we have performed the feature selection method nested within the loops of the cross-fold validation.

Several performance measures are used in this paper. They are: accuracy, sensitivity, specificity, precision and Mathew's Correlation Coefficient (MCC). For a binary classification problem, the confusion matrix gives four different measures: i) true positives (TN) or number of positive samples in the dataset that are correctly predicted by the classifier; ii) true negatives (TN) or number of negative samples in the dataset that are correctly predicted by the classifier; iii) false positives (FP) or number of negative samples that are wrongly predicted as positives by the classifier and iv) false negatives (FN) or number of positive samples that are wrongly predicted as negatives by the classifier. Based on these values different measures are calculated as given in following equations:

$$Accuracy\,(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Sensitivity\,(S_n) = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity\,(S_p) = \frac{TN}{TN + FP} \tag{9}$$

$$Precision\,(P_c) = \frac{TP}{TP + FP} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

Except MCC all the measures have the values in the range [0,1]. Here a maximum value of 1 will mean the best classifier with 100% of accuracy or precision or other measures and 0 means the worst classifier. However, MCC has a value in the range $[-1, +1]$. Here a maximum value of $+1$ indicates a perfect prediction algorithm. Many of the classifiers are probabilistic and uses a threshold to cut-off the negative samples from positive samples and thus the confusion matrix becomes dependent on the threshold. In such cases area under receiver operating characteristic (ROC) curve is preferred. ROC curve is the plot of true positive rate against false positive rate for different threshold values. Area under ROC (AUC) has the values in the range [0,1]. Here 0.5 value indicates a random classifier and + 1 is the maximum value indicating a predictor with the maximum true positive rate. The set of metrics is valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology [10, 11, 12, 58, 9] and system medicine [13, 14] and biomedicine [50], a completely different set of metrics as defined in [17] is needed.

3. Results and discussion

In this section, we present the experimental results achieved for iRSpot-SF. All the programs and the algorithms were developed in Python language using the sci-kit learn library. All experiments were run 10 times and only the average results are reported.

3.1. Feature selection

We have generated four set of features in our method and performed recursive feature elimination method on the training dataset. To avoid overfitting of the data, we have put the feature selection nested within the loop of the cross-fold validation. The feature selected algorithm provided us with 17 most effective features. A summary of the selected features is given in Table 3. We could notice that most powerful of all the features are gapped dinucleotide compositions [5] and newly proposed reverse complement k-mer compositions [12].

To test the significance of the selected features and the effectiveness of the feature selection method, we have compared the performance of the selected 17 features with that of two other combinations. We have used ranking based feature selection using an AdaBoost classification algorithm that selected 135 most important features using significance ranking. We have also used the whole set of features. For these three sets, we have used the SVM classifier with RBF kernel and reported different performance metrics using 10-fold cross-validation method. The results are reported in Table 4.

From the reported values in Table 4, note that the optimal set of features selected by our method is able to produce better results in terms of accuracy, precision, sensitivity and MCC. The area under ROC curve is very similar to the ranking method. The ROC analysis curve is given in Fig. 3. Note that among these two methods, our method produces only 17 features compared to 135 features by the ranking based feature selection method.

3.2. Strength of classification algorithm

We have compared the performance of SVM classifier with RBF kernel with three other configurations of algorithms to show the effectiveness of the algorithm used for iRSpot-SF. They are: K-nearest neighbor algorithm (KNN), Random Forest (RF) and SVM with linear kernel. We have used the same 17 selected features and performed 10-fold cross validation on the data and reported results in terms of accuracy, sensitivity, precision, specificity, MCC and AUC in Table 5. The best results in each criterion are shown in bold-faced font in the table.

From the values reported in Table 5, it is clear that SVM with RBF kernel outperforms all other classification algorithms used in the experiments. However, the closest competitor of that is the SVM with linear kernel. Note that, almost all classifiers are performing very similar to each other and that denotes the effectiveness of the selected features. The ROC analysis for different classifiers on the optimal feature set is also given in Fig. 4.

3.3. Comparison with previous methods

We have also compared the results achieved by iRSpot-SF with that of previous methods in the literature. We have used eight different methods for the sake of comparison. They are RF-DYMHC [35], IDQD [47], iRSpot-PseDNC [4], iRSpot-TNCPseAAC [51], iRSpot-DACC [46], iRSpot-EL [42], iRSpot-ADPM1575 [62], iRSpot-ADPM [62] and iRSpot-PDI [63]. We have reported sensitivity, specificity, MCC and accuracy achieved by different algorithms in Table 6. For the other algorithms, the results are taken as reported in the literature. The best values are shown in bold faced fonts.

As suggested in [16] and followed in many proposed methods in the literature [63, 56, 53, 60, 21], jackknife tests often report the robustness of the method and ensures that the classification algorithms are not over-fitting the data. We report the jack-knife test results of iRSpot-SF with a comparison to other methods in the literature in Table 7. Here too, note that iRSpot-SF is able to produce best results in terms of accuracy and comparable results in terms of other performance metrics in comparison to the previous methods. Note the increase in the specificity of our method in jackknife test compared to that found in cross-validation methods.

From the reported values in the table, we could notice that the accuracy of iRSpot-SF is highest among all the methods. The closest performer to iRSpot-SF is iRSpot-ADPM [62]. We notice that the accuracy achieved by different algorithms in the literature in past we slightly better than each other. It encourages us to look at the value of MCC achieved by our algorithm which is also highest among all and that is 0.6941. Here too, the next best performing method is iRSpot-ADPM which has a value of 0.6905 and significantly better than other methods. In terms of sensitivity our method, iRSpot-SF achieves a value of 84.57% which is 7.38% improved. Note that the dataset is slightly imbalanced with less number of positive samples and thus such increase in the sensitivity shows the effectiveness of finding positive samples of hotspots.

3.4. Web application

As pointed out in [20] and demonstrated in a series of recent publications [60, 53, 52, 34], user-friendly and publicly accessible web-servers represent the current direction for developing practically more useful prediction methods and computational tools. With a similar vision, we have implemented an web application based on the models learnt in this paper for iRSpot-SF. Our web application is redily available for use at: http://irspot.pythonanywhere.com/server.html. The website contains a guideline for users with a user friendly interface.

4. Conclusion

In this paper, we have proposed iRSpot-SF a method for prediction of recombination hotspots using sequence based features. Effective features were selected from sequence based features generated using recursive feature elimination. We have extracted four different set of features which produced state-of-the-art results using an SVM classifier with RBF kernel. We believe the reverse complement k-mer composition features have got potentials for further investigation. One of the limitations of our work was that we could not implement a web application for iRSpot-SF. However, we wish to develop a further improved version of our method and present it with an easy to use web server application.

References

[1] Akiko Aizawa, An information-theoretic perspective of tf–idf measures, Inf. Process. Manag. 39 (1) (2003) 45–65.
[2] Frédéric Baudat, Jérôme Buard, Corinne Grey, Adi Fledel-Alon, Carole Ober, Molly Przeworski, Graham Coop, Bernard De Massy, Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice, Science 327 (5967) (2010) 836–840.
[3] Mandana Behbahani, Hassan Mohabatkar, Mokhtar Nosrati, Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of chou's general pseudo amino acid composition, J. Theor. Biol. 411 (2016) 1–5.
[4] Wei Chen, Peng-Mian Feng, Hao Lin, Kuo-Chen Chou, irspot-psednc: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (6) (2013) e68.
[5] Wei Chen, Tian-Yu Lei, Dian-Chuan Jin, Hao Lin, Kuo-Chen Chou, Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[6] Wei Chen, Hao Lin, Kuo-Chen Chou, Pseudo nucleotide composition or pseknc: an effective formulation for analyzing genomic sequences, Mol. BioSyst. 11 (10) (2015) 2620–2634.
[7] Wei Chen, Hui Yang, Pengmian Feng, Hui Ding, Hao Lin, idna4mc: identifying dna n4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (22) (2017) 3518–3523.
[8] Wei Chen, Xitong Zhang, Jordan Brooker, Hao Lin, Liqing Zhang, Kuo-Chen Chou, Pseknc-general: a cross-platform package for generating various modes of pseudo nucleotide compositions, Bioinformatics 31 (1) (2014) 119–120.
[9] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, ploc-mgneg: Predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general pseaac, Genomics (2017), http://dx.doi.org/10.1016/j.ygeno.2017.10.002.
[10] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, ploc-mplant: predict subcellular localization of multi-location plant proteins by incorporating the optimal go information into general pseaac, Mol. BioSyst. 13 (9) (2017) 1722–1727.
[11] Xiang Cheng, Xuan Xiao, Kuo-Chen Chou, ploc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal go information into general pseaac, Gene 628 (2017) 315–321.
[12] Xiang Cheng, Shu-Guang Zhao, Wei-Zhong Lin, Xuan Xiao, Kuo-Chen Chou, plocmanimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (22) (2017) 3524–3531.
[13] Xiang Cheng, Shu-Guang Zhao, Xuan Xiao, Kuo-Chen Chou, iatc-misf: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (3) (2016) 341–346.
[14] Xiang Cheng, Shu-Guang Zhao, Xuan Xiao, Kuo-Chen Chou, iatc-mhyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (35) (2017) 58494.
[15] Kuo-Chen Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2004) 10–19.
[16] Kuo-Chen Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247.
[17] Kuo-Chen Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. BioSyst. 9 (6) (2013) 1092–1100.
[18] Kuo-Chen Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (3) (2015) 218–234.
[19] Kuo-Chen Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (21) (2017) 2337–2358.
[20] Kuo-Chen Chou, Hong-Bin Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (02) (2009) 63.
[21] Shahana Yasmin Chowdhury, Swakkhar Shatabda, Abdollah Dehzangi, Idnaprot-es: Identification of dna-binding proteins using evolutionary and structural features, Sci. Rep. 7 (1) (2017) 14938.
[22] Corinna Cortes, Vladimir Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[23] Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, James Lyons, Kuldip Paliwal, Abdul Sattar, Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into chou's general pseaac, J. Theor. Biol. 364 (2015) 284–294.
[24] Hui Ding, Dongmei Li, Identification of mitochondrial proteins of malaria parasite using analysis of variance, Amino Acids 47 (2) (2015) 329–333.

[25] Chuan Dong, Ya-Zhou Yuan, Fa-Zhan Zhang, Hong-Li Hua, Yuan-Nong Ye, Abraham Alemayehu Labena, Hao Lin, Wei Chen, Feng-Biao Guo, Combining pseudo dinucleotide composition with the z curve method to improve the accuracy of predicting dna elements: a case study in recombination spots, Mol. BioSyst. 12 (9) (2016) 2893–2900.
[26] Ashok Kumar Dwivedi, Usha Chouhan, Comparative study of artificial neural network for classification of hot and cold recombination regions in *Saccharomyces cerevisiae*, Neural Comput. & Applic. 29 (2) (2018) 529–535.
[27] Fu Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, Weizhong Li, Cd-hit: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152.
[28] Vinay Gadia, Gail Rosen, A text-mining approach for classification of genomic fragments. In Bioinformatics and Biomeidcine Workshops, 2008. BIBMW 2008, IEEE International Conference on, Pages 107–108, IEEE, 2008.
[29] Jennifer L. Gerton, Joseph DeRisi, Robert Shroff, Michael Lichten, Patrick O. Brown, Thomas D. Petes, Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*, Proc. Natl. Acad. Sci. 97 (21) (2000) 11383–11390.
[30] Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, Michael A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features, PLoS Comput. Biol. 10 (7) (2014) e1003711.
[31] Shou-Hui Guo, En-Ze Deng, Li-Qin Xu, Hui Ding, Hao Lin, Wei Chen, Kuo-Chen Chou, inuc-pseknc: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (11) (2014) 1522–1529.
[32] Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.
[33] Tin Kam Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[34] Md Mofijul Islam, Sanjay Saha, Md Mahmudur Rahman, Swakkhar Shatabda, Dewan Md Farid, Abdollah Dehzangi, iprotgly-ss: Identifying protein glycation sites using sequence and structure based features, Proteins 86 (7) (2018) 777–789.
[35] Peng Jiang, Haonan Wu, Jiawei Wei, Fei Sang, Xiao Sun, Lu. Zuhong, Rf-dymhc: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features, Nucleic Acids Res. 35 (suppl_2) (2007) W47–W51.
[36] Muhammad Kabir, Maqsood Hayat, irspot-gaensc: identifing recombination spots via ensemble classifier and extending the concept of chou's pseaac to formulate dna samples, Mol. Gen. Genomics. 291 (1) (2016) 285–296.
[37] Liqi Li, Sanjiu Yu, Weidong Xiao, Yongsheng Li, Lan Huang, Xiaoqi Zheng, Shiwen Zhou, Hua Yang, Sequence-based identification of recombination spots using pseudo nucleic acid representation and recursive feature extraction by linear kernel svm, BMC Bioinform. 15 (1) (2014) 340.
[38] Hao Lin, The modified mahalanobis discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition, J. Theor. Biol. 252 (2) (2008) 350–356.
[39] Hao Lin, En-Ze Deng, Hui Ding, Wei Chen, Kuo-Chen Chou, ipro54-pseknc: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (21) (2014) 12961–12972.
[40] Hao Lin, Zhi-Yong Liang, Hua Tang, Wei Chen, Identifying sigma70 Promoters with Novel Pseudo Nucleotide Composition, IEEE/ACM transactions on computational biology and bioinformatics (2017).
[41] Bin Liu, Fule Liu, Xiaolong Wang, Junjie Chen, Longyun Fang, Kuo-Chen Chou, Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences, Nucleic Acids Res. 43 (W1) (2015) W65–W71.
[42] Bin Liu, Shanyi Wang, Ren Long, Kuo-Chen Chou, irspot-el: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (1) (2016) 35–41.
[43] Bin Liu, Hao Wu, Kuo-Chen Chou, Pse-in-one 2.0: an improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences, Nat. Sci. 9 (04) (2017) 67.
[44] Bin Liu, Fan Yang, De-Shuang Huang, Kuo-Chen Chou, ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc, Bioinformatics 34 (1) (2017) 33–40.
[45] Bingquan Liu, Yumeng Liu, Huang Dong, Recombination hotspot/coldspot identification combining three different pseudocomponents via an ensemble learning approach, Biomed. Res. Int. 2016 (2016).
[46] Bingquan Liu, Yumeng Liu, Xiaopeng Jin, Xiaolong Wang, Bin Liu, irspot-dacc: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance, Sci. Rep. 6 (2016) 33483.
[47] Guoqing Liu, Jia Liu, Xiangjun Cui, Cai Lu, Sequence-dependent prediction of recombination hotspots in saccharomyces cerevisiae, J. Theor. Biol. 293 (2012) 49–54.
[48] Philippe Lopez, Hervé Philippe, Hannu Myllykallio, Patrick Forterre, Identification of putative chromosomal origins of replication in archaea, Mol. Microbiol. 32 (4) (1999) 883–886.
[49] Prabina Kumar Meher, Tanmaya Kumar Sahu, Varsha Saini, Atmakuri Ramakrishna Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chou's general pseaac, Sci. Rep. 7 (2017) 42362.
[50] Wang-Ren Qiu, Bi-Qian Sun, Xuan Xiao, Zhao-Chun Xu, Kuo-Chen Chou, iptm-mlys: identifying multiple lysine ptm sites and their different types, Bioinformatics 32 (20) (2016) 3116–3123.
[51] Wang-Ren Qiu, Xuan Xiao, Kuo-Chen Chou, irspot-tncpseaac: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2) (2014) 1746–1766.
[52] Farshid Rayhan, Sajid Ahmed, Swakkhar Shatabda, Dewan Md Farid, Zaynab Mousavian, Abdollah Dehzangi, M. Sohel Rahman, idti-esboost: Identification of drug target interaction using evolutionary and structural features with boosting, Sci. Rep. 7 (1) (2017) 17731.
[53] Swakkhar Shatabda, Sanjay Saha, Alok Sharma, Abdollah Dehzangi, iphloc-es: Identification of bacteriophage protein locations using evolutionary and structural features, J. Theor. Biol. 435 (2017) 229–237.
[54] Jiangning Song, Yanan Wang, Fuyi Li, Tatsuya Akutsu, Neil D. Rawlings, Geoffrey I. Webb, Kuo-Chen Chou, iprot-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Brief. Bioinform. (2018), http://dx.doi.org/10.1093/bib/bby028.
[55] Hua Tang, Ping Zou, Chunmei Zhang, Rong Chen, Wei Chen, Hao Lin, Identification of apolipoprotein using feature selection technique, Sci. Rep. 6 (2016) 30441.
[56] Md Raihan Uddin, Alok Sharma, Dewan Md Farid, Md Mahmudur Rahman, Abdollah Dehzangi, Swakkhar Shatabda, Evostruct-sub: An accurate gram-positive protein subcellular localization predictor using evolutionary and structural features, J. Theor. Biol. 443 (2018) 138–146.
[57] Rong Wang, Yong Xu, Bin Liu, Recombination spot identification based on gapped k-mers, Sci. Rep. 6 (2016) 23934.
[58] Xuan Xiao, Xiang Cheng, Su Shengchao, Qi Mao, Kuo-Chen Chou, ploc-mgpos: incorporate key gene ontology information into general pseaac for predicting subcellular localization of gram-positive bacterial proteins, Nat. Sci. 9 (09) (2017) 330.
[59] Hui Yang, Wang-Ren Qiu, Guoqing Liu, Feng-Biao Guo, H. Lin, irspot-pse6nc: identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general pseknc, Int. J. Biol. Sci. 14 (8) (2018) 883–891.
[60] Rianon Zaman, Shahana Yasmin Chowdhury, Mahmood A Rashid, Alok Sharma, Abdollah Dehzangi, and Swakkhar Shatabda. Hmmbinder: Dna-binding protein prediction using hmm profile based features, BioMed. Res. Int. 2017 (2017).
[61] Chang-Jian Zhang, Hua Tang, Wen-Chao Li, Hao Lin, Wei Chen, Kuo-Chen Chou, iori-human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition, Oncotarget 7 (43) (2016) 69783–69793.
[62] Lichao Zhang, Kong Liang, irspot-adpm: Identify recombination spots by incorporating the associated dinucleotide product model into chou's pseudo components, Journal of Theo. Biol. 441 (2018).
[63] Lichao Zhang, Kong Liang, irspot-adpm: Identify recombination spots by incorporating the associated dinucleotide product model into chou's pseudo components, Journal of Theo. Biol. 441 (2018) 1–8.