Journal of Intelligent & Fuzzy Systems 41 (2021) 1309–1317
DOI:10.3233/JIFS-210213
IOS Press

iRspot-DCC: Recombination hot/ cold spots identification based on dinucleotide-based correlation coefficient and convolutional neural network

Wang Guo*, Xingmou Liu, You Ma and Rongjie Zhang
Chongqing Key Laboratory of Complex Systems and Bionic Control, Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. The correct identification of gene recombination cold/hot spots is of great significance for studying meiotic recombination and genetic evolution. However, most of the existing recombination spots recognition methods ignore the global sequence information hidden in the DNA sequence, resulting in their low recognition accuracy. A computational predictor called iRSpot-DCC was proposed in this paper to improve the accuracy of cold/hot spots identification. In this approach, we propose a feature extraction method based on dinucleotide correlation coefficients that focus more on extracting potential DNA global sequence information. Then, 234 representative features vectors are filtered by SVM weight calculation. Finally, a convolutional neural network with better performance than SVM is selected as a classifier. The experimental results of 5-fold cross-validation test on two standard benchmark datasets showed that the prediction accuracy of our recognition method reached 95.11%, and the Mathew correlation coefficient (MCC) reaches 90.04%, outperforming most other methods. Therefore, iRspot-DCC is a high-precision cold/hot spots identification method for gene recombination, which effectively extracts potential global sequence information from DNA sequences.

Keywords: Recombination spots, correlation coefficient, DNA property matrix, support vector machines, convolutional neural network

1. Introduction

Meiotic recombination is one of the most important steps in genetic inheritance and evolution. As the main driving force of gene evolution, it provides a new genetic variation combination [1], thereby accelerating the evolution of sexual reproduction organisms. It is thus of great significance for the recognition of the sites of gene recombination [2]. Research indicated that the probability of recombination in different genomes was different [3, 4]. Generally speaking, the regions with high recombination rate are called hot spots, while the regions with low recombination rate are called cold spots.

In the recent years, some methods regarding distinguishing cold or hot spots of recombination have been proposed. For instance, Zhou et al. [5] proposed a SVM method based on codon composition. Jiang et al. [6] Classified the recombination points by extracting the composition characteristics of gap dinucleotide and by using random forest method. All the above three methods have ignored the physical

*Corresponding author. Wang Guo, Chongqing Key Laboratory of Complex Systems and Bionic Control, Chongqing University of Posts and Telecommunications, Chongqing, China. E-mail: guowangcq@qq.com.

ISSN 1064-1246/$35.00

property information contained in dinucleotide pairs and only considered the frequency information of the nucleobase in the DNA sequence. Therefore, a feature extraction method based on the pseudo dinucleotide composition and the dinucleotide's physical properties [8] was proposed. Then, IR-SF [14] further improved the recognition accuracy by fusing frequency features and physical property features in dinucleotide pairs, but none of the above feature extraction methods fully considered the global sequence information hidden in DNA, which has an indispensable place in recombination recognition. To solve this problem, some feature extraction methods based on DNA attribute matrices have been proposed [9–11]. Liu et al. [12] proposed a self-crossing feature extraction method based on dinucleotide. Zhang et al. put forward iRspot-PDI employing the diversity of dinucleotide [13], and a feature selection method based on SVM [14]. With the rapid development and widespread use of deep learning, more and more deep learning classification algorithms have been proposed [15–17], and some deep learning algorithms have been applied to identify recombination points. For instance, iRSpot-DTS [18] uses sparse automatic coders to reconstruct the DNA sequence feature, iRSpot-SPI [19] adopts deep neural networks as classifiers. Although all of them have effectively improved the prediction accuracy, most of the existing prediction methods are still at a low level of accuracy.

On the basis of previous studies, a predictor was proposed to improve the recognition accuracy of reorganization points in the presented paper. The presented paper focused on the issues including how to extract more effective information from DNA sequences, how to select more representative features to distinguish cold or hot spots of recombination, and how to select appropriate classifiers to improve prediction accuracy.

Firstly, a feature extraction method based on correlation coefficient (DCC) of dinucleotide spatial was proposed, by which, the hidden information in DNA sequences can be more effectively excavated; Secondly, 234 representative features were selected based on the weight value of each feature calculated by SVM. The prediction accuracy of other feature extraction methods was compared with that of the feature extraction and screen methods reported herein under the same selection of SVM as the classifier with the purpose of exhibiting the superiority of the features obtained by the above two steps. Finally, a novel predictor iRspot-DCC was proposed based on the excellent ability of recognition and classification of convolutional neural network. The cross test experiments based on two data sets demonstrated that the reorganization points could be well identified by the reported method with the prediction accuracy reaching 95.11%, which is more precise than most of the reported predictors. The main contributions of this article are as follows:

- Based on dinucleotide correlation coefficients, a more efficient feature extraction method is proposed to identify DNA recombination cold/hot spots. This method takes into account the global sequence features of DNA more fully than the previous methods.
- Combining convolutional neural network and svm weight calculation, a computational predictor called iRSpot-DCC was proposed to substantially improve the accuracy of DNA recombination cold/hot spot identification.

2. Materials and methods

In this section, we provide a detailed explanation of the proposed method iRSpot-DCC, including an introduction to the selected benchmark dataset, a description of the process of extracting DNA features by DCC. The feature selection method and classification algorithm of iRSpot-DCC are also given. Finally, we describe the methods and indicators for evaluating the performance of the predictor.

The system architecture of iRSpot-DCC is given in detail as shown in Fig. 1. Specifically, in iRSpot-DCC, DCC is used to generate DNA feature vectors on a benchmark dataset, followed by feature selection through svm-based weight calculation, and finally, a convolutional neural network is used as the classifier.

2.1. Datasets

The selection of data sets has a certain degree of influence on the recognition accuracy of gene recombination spots. In order to show the superiority of our prediction method, two public benchmark data sets S_1 and S_2 are selected from Jiang [6] and liu [9] respectively. Both benchmark data sets contain different numbers of recombinant hot spots and recombinant cold spots. S_1 has 591 recombinant cold spots and 490 recombinant hot spots, while S_2 has 572 recombinant cold spots and 478 recombinant hot spots are con-

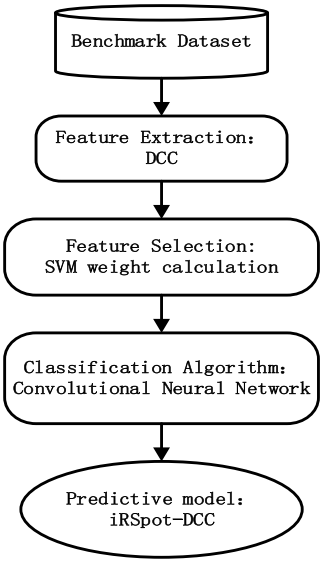

Fig. 1. A system structure of iRSpot-DCC.

tained. S_1 and S_2 can be expressed by the following equation:

$$\begin{cases} S_1 = S_1^+ \cup S_1^- \\ S_2 = S_2^+ \cup S_2^- \\ S = S_1 \cup S_2 \end{cases} \tag{1}$$

where S_1^+ and S_2^+ are the hot spot of gene recombination, and S_1^- and S_2^- are the cold spot of gene recombination [20], $\cup$ represents the union of the two parts. S is the set of S_1 and S_2 sample points, which is also the data set used in this paper. The datasets can be download from the URL: https://github.com/Guowang1999/IRspot-dcc.

2.2. Feature extraction method

2.2.1. DNA property matrix

DNA properties include physical and thermodynamic parameters [14], which play a significant role in the identification of cold hot spots in gene recombination. Therefore, we selected 15 DNA attributes in Table 1 to identify recombinant cold hot spots. These properties will be normalized by the following equation:

$$\frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{2}$$

where Y is the original value of the property, and Y_{min} and Y_{max} are the minimum and maximum value of the attribute respectively. In this study, a DNA sequence will be converted into a matrix $M = \left(p_{i,j}\right)_{(L-1)\times 15}$, where 15 is the total number of selected DNA attributes, and L is the length of the DNA sequence, $p_{i,j}$ represents the j-th attribute value of the i-th dinucleotide pair composed of adjacent nucleotides in the sequence. Obviously, different DNA sequences have different property matrices M.

2.2.2. Dinucleotide-based Correlation coefficient

At present, most recognition methods ignore the hidden information of DNA global sequence, in order to better extract the hidden information in DNA sequences, we propose a feature extraction method based on the correlation coefficient of dinucleotide, which can more fully mine the global DNA sequence information. Suppose a DNA sequence is represented as follows:

$$R = D_1 D_2 D_3 D_4 \bullet\bullet\bullet D_L \tag{3}$$

where L is the length of the DNA sequence, and $D_i (i = 1, 2\ldots, \mathrm{L})$ represents the i-th nucleotide base in the sequence. We use correlation coefficients to extract global sequence information in dinucleotides, as shown in the following equation:

$$C_{s,t}^G = \frac{\frac{1}{L-g-1}\sum_{i=1}^{L-g-1}\left(p_{is} - \overline{p_s}\right)\left(p_{i+g,t} - \overline{p_t}\right)}{\sqrt{\frac{\sum_{i=1}^{L-g-1}\left(p_{is}-\overline{p_s}\right)^2}{L-g-1}}\sqrt{\frac{\sum_{i=1}^{L-g-1}\left(p_{i+gs}-\overline{p_t}\right)^2}{L-g-1}}} \tag{4}$$

where s and t are the local property indexes of dinucleotides, and $\overline{p_s}$ and $\overline{p_t}$ are the average values of the local property indexes s and t respectively. $p(i,s)$ is the value of the property index s of the dinucleotide($D_i D_{i+1}$) at position i. When $s = t$, it represents the autocorrelation between the same attributes in the DNA sequence, whereas when $s \neq t$, it represents the cross-correlation between different attributes in the DNA sequence, g represents the distance between dinucleotides at two different positions, and its maximum value is G.

The DCC feature extraction process is shown in Fig. 2. By DCC, when the value of g is fixed to a certain value, the 15 physical property correlation coefficients of the dinucleotides at different positions generate a 225-dimensional (15*15) feature vector of a DNA sequence when the value of g is changed, and the feature vector of a DNA sequence is 15*15*G

Table 1
Values of original dinucleotide properties

	AC/GT	AG/CT	AT	CG	CA/TG	TA	CC/GG	CG	AA/TT	GA/TC
F-rise	21.98	17.48	24.79	14.66	14.51	14.24	14.25	14.66	21.34	18.41
F-twist	0.06	0.05	0.07	0.05	0.05	0.05	0.06	0.05	0.07	0.06
F-slide	6.80	3.47	9.61	2.71	2.00	1.85	2.99	2.71	6.69	4.27
F-roll	0.06	0.04	0.05	0.04	0.04	0.03	0.04	0.04	0.04	0.05
F-shift	2.91	2.80	4.66	3.02	2.88	4.11	2.67	3.02	6.24	3.58
F-tilt	0.07	0.06	0.1	0.06	0.06	0.07	0.06	0.06	0.08	0.07
energy	–1.44	–1.28	–0.88	–2.17	–1.45	–0.58	–1.84	–2.17	–1.00	–1.30
entropy	–22.40	–21.00	–20.40	–27.20	–22.70	–21.30	–19.90	–27.20	–21.30	–22.20
enthalpy	–8.40	–7.80	–7.20	–10.60	–8.50	–7.20	–8.00	–10.60	–7.60	–8.20
twist	31.53	32.29	30.72	33.67	35.43	36.94	33.54	33.67	35.02	35.67
rise	3.24	3.32	3.21	3.29	3.37	3.39	3.36	3.29	3.25	3.30
tilt	0.33	–1.66	0.00	0.00	0.14	0.00	–0.77	0.00	–1.26	1.44
slide	–0.59	–0.22	–0.68	0.44	0.48	0.04	–0.17	0.44	–0.18	–0.05
shift	–0.02	–0.02	0.00	0.00	0.01	0.00	0.03	0.00	0.01	–0.01
roll	2.01	3.60	0.61	6.02	5.60	3.50	4.68	6.02	1.05	2.44

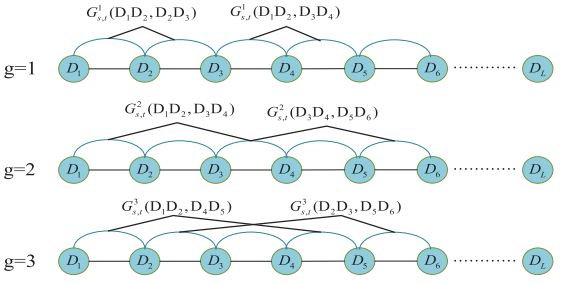

Fig. 2. Schematic diagram of the process of generating feature vectors through DCC. (1) The first tier depicts correlation information with g = 1(g is the gap between the different dinucleotide). (2) The second tier depicts correlation information with g = 2 (3) The third tier depicts correlation information with g = 3.

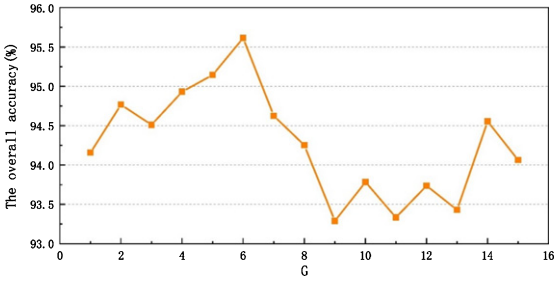

Fig. 3. The overall accuracy changes with increasing g from 2 to 15 on the S dataset. (1) As g changes from 2 to 15, the prediction accuracy fluctuates between 93.23% and 95.61%. When G = 6, the prediction accuracy reaches the maximum.

2.3. Feature selection method

As G increases, the dimension of the feature vector also increases rapidly. When $G = 1$, its dimension is 255, and when $G = 10$, the dimension is 2550. However, the prediction accuracy does not increase with the increase of G. as shown in Fig. 3, when $G = 6$, the prediction accuracy reaches its peak. Therefore, $G = 6$ is our choice, and the feature vector of 1350 dimension can be formulated as:

$$FV_{1350} = [\varphi_1, \varphi_2, ...\varphi_m, ..., \varphi_{1350}]^T \qquad (5)$$

Since the dimension of feature vector is too large, to improve the prediction accuracy and speed, it is necessary to select feature selection algorithm to remove noise and redundancy. First, sort all 1350 features according to the weight values calculated by svm, and then select the top k features with high weight values as the new feature vector [11]. For the

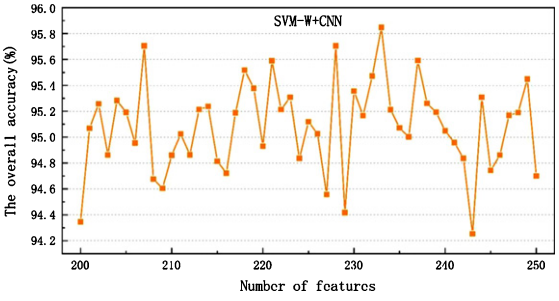

Fig. 4. The value of the overall accuracy under different number of features on the S dataset. (1) As the number of features changes from 200 to 250, the prediction accuracy fluctuates between 94.345 %and 95.70%, and the maximum value of prediction accuracy is obtained when the number of features is 234.

sake of illustration, we named this feature selection method SVM-W. As shown in Fig. 4, when k selects 234, the prediction accuracy reaches its peak. The new 234 dimensional feature vector can be expressed as:

$$FV_{234} = [\varphi_1, \varphi_2, ...\varphi_m, ..., \varphi_{234}]^T \quad (6)$$

2.4. Convolutional neural network

Convolutional Neural Network (CNN) is a classic deep neural network structure. Compared with traditional methods, it has the advantages of weight sharing and automatic feature extraction of the training data [21–23]. With SVM-W, we directly select the 234-dimensional feature vector from the original 1350-dimensional feature vector and do not reconstruct the new feature vector to eliminate useless information. Thus, the reduced dimensional feature vector still contains a lot of noise. Through the convolutional layer and pooling layer of the convolutional neural network, the main information can be further extracted from the original feature vectors. The noise can be eliminated to improve the prediction accuracy. Considering that feature screening performed by SWM-W does not completely eliminate noise, CNN is believed to be a classifier with obvious advantages. In this study, a 6-layer convolutional neural network structure was proposed.

As can be seen from Fig. 5, an input layer, two convolution layers, a pooling layer, a full connection layer and an output layer were included, which can be expressed as INPUT-C1-S2-C3-F4-OUTPUT. Input is a 1×234 eigenmatrix composed of 234 dimensional feature vectors. There are 16 convolution kernels of 1×3 in C1 layer. The input data is convoluted with each convolution kernel. After offsetting, the convolution layer C1 is obtained through the activation function, and 16 characteristic graphs with the size of 1×232 are generated. The formula of convolution is as follows:

$$x_j^{(l)} = \sum_{i \in N_j} a_j^{(l-1)} k_{ij}^l + b^l \quad (7)$$

$$a_i^{(l)} = f\left(x_j^{(l)}\right) \quad (8)$$

In the above formula, $x_j^{(l)}$is the j-th channel of the l-th feature map.b is the offset value, k is the convolution kernel, N_j is the set of input characteristic graphs, $f$is the activation function, $a_i^{(l)}$is the i-th channel of layer l after the action of the activation function of $x_j^{(l)}$. In this paper, The ReLU activation functions with fast computational performance was employed, and the expression is as follows:

$$f\left(x_j^{(l)}\right) = \max\left(0, x_j^{(l)}\right) \quad (9)$$

The input feature vector is first convolved with the convolution kernel, There were 16 convolution kernels with size of 1×2 and step size of 1 in C1 layer, and 16 characteristic graphs with size of 1×131 were output. After the action of the activation function f, the negative part of the output $x_j^{(l)}$ is changed to zero, and the final output result is $x_i^{(l)}$. The output part $a_i^{(l)}$ is then pooled using an average pooling layer with a sampling window of 1×2. C3 has 16 convolution kernels and their size is 1×2. F4 is the fully connected layer, which has 48 neurons. All neurons in C3 layer of each neuron were connected [21, 22]. The number of neurons in the output layer was set to 2, and the softmax function was taken as the activation function.

2.5. Prediction result assessment

In the aspect of statistical prediction, it is very important to evaluate the performance of the predictor in an objective and correct way. Here, we adopt the criteria mentioned in reference [26] and select widely recognized cross-validation to evaluate the performance of the predictor. According to the selected criteria, the correct prediction rates of recombination hot spots (data set *S*+) and recombination cold spots (data set S-) are defined as:

$$\begin{cases} \Lambda^+ = \frac{N^+ - N_-^+}{N^+} \\ \Lambda^- = \frac{N^+ - N_+^-}{N^-} \end{cases} \quad (10)$$

The Λ^+ and Λ^- represent the prediction accuracy of recombinant hot spots and recombinant cold spots respectively, N^+ and N^- represent the total number of recombinant hot spots and recombinant cold spots respectively, N_-^+ and N_+^- represent the wrong prediction number of recombinant hot spots and recombinant cold spots. The overall prediction accuracy can be obtained by combining the reorganized cold spots and hot spots, which can be expressed by the following equation:

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N_+ + N_-} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \quad (11)$$

where Λ represents the accuracy of the overall prediction. From the above equation, it can be analyzed that when $\Lambda = 1$, it means that all recombina-

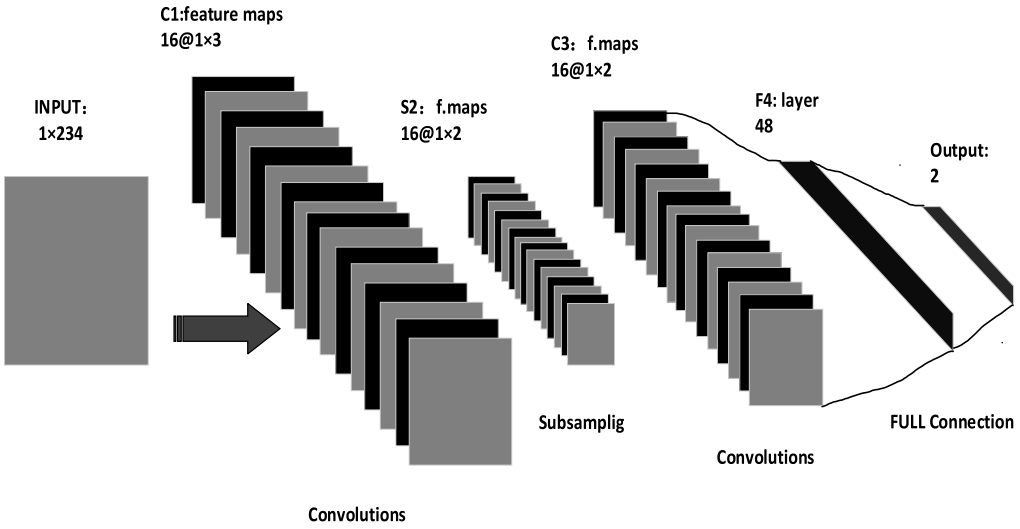

Fig. 5. The structure of a convolutional neural network.

tion points are predicted correctly, that is $N_{-}^{+} = N^{+}$, $N_{+}^{-} = N^{-}$. The following equation is widely used in evaluating the prediction results of the two categories:

$$\begin{cases} Sn = \frac{TP}{TP+FN} \\ Sp = \frac{TN}{TN+FP} \\ MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\ Acc = \frac{TP+TN}{TP+TN+FP+FN} \end{cases} \tag{12}$$

where *TP* (true negative) and *FN* (false negative) respectively represent the number of correct predictions of recombinant hot spots and the number of wrong predictions of recombinant hot spots. Similarly, *TN* (true negative) and *FP* (false positive) respectively represent the number of correct predictions of recombinant hot spots and the number of wrong predictions of recombinant hot spots, MMC is Marseus correlation coefficient. we can obtain the following relationship:

$$\begin{cases} TP = N^{+} - N_{-}^{+} \\ TN = N^{-} - N_{+}^{-} \\ FP = N_{+}^{-} \\ FN = N_{-}^{+} \end{cases} \tag{13}$$

Substituting Equation (13) into Equation (14), the following equation can be obtained:

$$\begin{cases} Sn = 1 - \frac{N_{-}^{+}}{N^{+}} \\ Sp = 1 - \frac{N_{+}^{-}}{N^{-}} \\ MCC = \frac{1-\left(\frac{N_{-}^{+}}{N^{+}} + \frac{N_{+}^{-}}{N^{-}}\right)}{\sqrt{\left(1+\frac{N^{+}}{N_{-}^{+}-N_{+}^{-}}\right)\left(1+\frac{N^{-}}{N_{+}^{-}-N_{-}^{+}}\right)}} \\ Acc = \Lambda = 1 - \frac{N_{-}^{+}+N_{+}^{-}}{N^{+}+N^{-}} \end{cases} \tag{14}$$

It can be seen from the above equation that $\Lambda^{+} = Sn$, $\Lambda^{-} = Sp$, they respectively represent the correct recognition rate of recombination hot spots and the correct recognition rate of recombination cold spots. MCC is Marseus correlation coefficient, which is usually used to evaluate the quality of the two classification results. When all recombinant points are correctly identified, MCC = 1, and when MCC = 0, it means that half of the recombinant cold points and half of the recombinant hot spots are identified incorrectly, i.e. $N_{-}^{+} = 1/2N^{+}$, $N_{+}^{-} = 1/2N^{-}$. On the contrary, when $N_{-}^{+} = N^{+}$, $N_{+}^{-} = N^{-}$, which means that all the reorganized cold spots and hot spots are incorrectly identified, the overall prediction accuracy ACC = 0. Through Sn, SP, MCC and ACC, we can make an objective and correct evaluation of the prediction results of the two classifications, which can help us to understand the results of the two classifications more intuitively.

3. Results and discussions

In this section, the advantages of iRSpot-DCC are clarified by comparing with other identification methods. After feature extraction and feature screening, DNA sequences can be transformed into representative DNA feature vectors. And the prediction results can be obtained by substituting these feature vectors into the classifier for recognition.

The prediction accuracy is determined by the DNA feature vectors produced by different methods and the classifier. Therefore, we compare the proposed method with other recognition methods from the two perspectives of DNA feature vector and selected classifier. Specifically, when using SVM as a classifier, the prediction accuracy of the DNA feature vector generated by previous methods is compared with that of the proposed method. In addition, in terms of the feature vectors generated by the proposed method, the prediction performance of CNN and SVM is compared.

3.1. Compared with the previous method when using SVM

In previous studies [9–14], most of them employed SVM as classifier. To reflect the features of DNA sequences generated by our feature extraction and screening method are superior, we also employed SVM as a classifier and compared the results with other studies using SVM.

As shown in Table 2: The experimental accuracies of iRSpot-ADPM [14] and iRSpot-DACC [12] on the dataset S2 were only 84.57% and 68.36%, respectively. Their prediction accuracies improved to 90.68% and 88.65% when S1 and S2 were used as the experimental benchmark dataset, respectively. The expansion of data volume is beneficial to improve the prediction accuracy.

When using SVM as a classifier, iRSpot-DCC outperforms iRSpot-ADPM and iRSpot-DACC by 2.49% and 4.52% on the benchmark dataset, respectively. MMC was 4.18% and 8.82% higher than them, respectively. All four prediction performances of iRSpot-DCC outperformed iRSpot-ADPM. Since they both used SVM-W for feature selection and the same classification algorithm, the source of the prediction performance gap can only be the feature extraction methods DCC and ADPM. Therefore, DCC is a feature extraction method that is more efficient than ADPM and takes into account the global DNA sequence more fully. Similarly, it can be concluded that the performance of DCC is also better than that of DACC. Previous studies iRSpot-TNCPseAAC and iRSpot-PseDNC did not consider the DNA global sequence in feature extraction and extracted only a small number of features. Hence, their prediction accuracy was much lower than that of iRSpot-ADPM and even lower than that of iRSpot-DCC, illustrating the importance of DNA global sequence information and DCC's effectiveness.

Another reason for the improvement in prediction accuracy is feature selection. Using SVM as a classifier and DCC to extract features, the prediction accuracy for feature selection using SVM-W and PCA was 93.17% and 91.64% respectively. The data demonstrate that SVM-W filters out more representative features compared to PCA. If the SVM classification is used directly without feature filtering, the prediction accuracy is 91.12%, which is similar to the accuracy of PCA and 2% lower than that of SVM-W. The above analysis shows that SVM-W can also effectively improve the prediction accuracy.

Table 2
Performance comparison of different feature extraction and screening methods when using svm classification (by 5-fold cross-validation test)

Predictor	dataset	FE+SM	Sn(%)	Sp(%)	Mcc(%)	Acc(%)
iRSpot-TNCPseAAC	S2	TNCPseAAC+NO	76.56	70.99	47.37	73.52
iRSpot-PseDNC	S2	PseDNC+NO	71.75	85.84	58.30	79.33
iRSpot-ADPM	S2	ADPM+SVM-W	77.19	90.73	69.05	84.57
iRSpot-DACC	S2	DACC+PCA	68.99	67.54	35.80	68.36
iRSpot-ADPM	S1 + S2	ADPM+SVM-W	81.53	94.1	81.94	90.68
iRSpot-DACC	S1 + S2	DACC+PCA	87.08	90.98	77.83	88.85
iRSpot-DCC(NO+SVM)	S1 + S2	DCC+NO	85.45	93.16	83.83	91.12
iRSpot-DCC(PCA+SVM)	S1 + S2	DCC+PCA	86.53	93.31	84.32	91.64
iRSpot-DCC(SVM)	S1 + S2	DCC+SVM-W	90.08	95.71	86.11	93.17

FE+SM: feature extraction and (+) screening methods. (a) TNCPseAAC, ADPM represent the feature extraction methods in the iRSpot-TNCPseAAC, iRSpot-ADPM articles respectively. (b) SVM-W, PCA are two different feature selection methods, NO means no feature selection.

Table 3
Performance comparison of different methods on datasets by 5-fold cross-validation test

Predictor	dataset	Sn(%)	Sp(%)	Mcc(%)	Acc(%)
iRSpot-ADPM	S1 + S2	81.53	94.1	81.94	90.68
iRSpot-DACC	S1 + S2	87.08	90.98	77.83	88.85
iRSpot-DCC(SVM)	**S1 + S2**	**90.03**	**95.71**	**86.11**	**93.17**
iRSpot-DCC(CNN)	**S1 + S2**	**94.01**	**96.08**	**90.04**	**95.11**

iRSpot-DCC combines DCC and SVM-W to achieve higher prediction accuracy, which not only has the highest prediction accuracy of 93.17% among all methods using SVM as a classifier, but also has MCC and Sn over 90%. Finally, compared with the previous prediction method [9, 11, 13] that did not use SVM as the classifier, our prediction accuracy has improved by about 5%.

3.2. Compare the prediction performance of CNN and SVM

Table 3 indicates that the overall accuracy reached 95.11% when employed CNN as classifier, while Sn, Sp and MCC were 94.01%, 96.08%, 90.04%, respectively. Compared with employed SVM as classifier, the overall accuracy is improved by about 2%, while Sn and MCC increased by 3.98% and 3.93%. The above analysis demonstrated that CNN exhibited better prediction performance than SVM when using the feature extraction and screening methods in this paper.

The above experimental results show that the more representative feature vectors generated by DCC and SVM-W can effectively improve cold/hot spot identification accuracy. The prediction performance of CNN is better than that of SVM. it can be concluded that the combination of CNN and more representative feature vectors makes the prediction accuracy of iRspot-DCC higher than most other predictors.

4. Conclusion

The size of the data volume will influence the prediction results to some extent. Data sets S1 and S2 were combined to be taken as the benchmark data set to increase the data volume. In the aspect of feature extraction, more abundant DNA sequence information was obtained by employing the feature extraction method based on the spatial correlation coefficient. Then, 234 dimensional features were screened out based on the weight calculation of SVM. It was found that the as-proposed method exhibited superiority as compared to other feature extraction and screening methods. Finally, the proposed CNN classification method possessed better prediction performance when being compared with the SVM classifier.

Although iRSpot-DCC has achieved high accuracy in DNA recombination cold/hot spot identification, with the development of deep learning, many algorithms with better performance than CNN have emerged [27–30]. Therefore, it is one of our future tasks to combine different feature extraction methods and optimization algorithms [31–35] to further improve the prediction accuracy. The user-friendly web server reported in a series of recent publications [36, 37] plays a role in promoting the rapid development of medical science, and it is still urgent to provide a network server for the prediction of DNA recombination points to the public in future work.

Acknowledgments

The work is financially supported by National Natural Science Foundation of China Project (61803059) and (61703347), InnovationTeam Project of Chongqing Education Committee (CXTDX201601019).

References

[1] P. Paul, D. Nag and S. Chakraborty, Recombination hotspots: Models and tools for detection, *DNA Repair* **40** (2016), 47–56.
[2] M.J. Lercher and L.D. Hurst, Human SNP variability and mutation rate are higher in regions of high recombination, *Trends in genetics* **18**(7) (2002), 337–340.
[3] M.I. Jensen, T.S. Furey and Y. Lu, Comparative recombination rates in the rat, mouse, and human genomes, *Genome Research* **14**(4) (2004), 528–538.
[4] M. Eugenio, B. Richard, B. Alessandro, H. Wolfgang and L.M. Steinmetz, High-resolution mapping of meiotic crossovers and non-crossovers in yeast, *Nature* **454** (2008), 479–485.
[5] T. Zhou, J. Weng, X. Sun and Z. Lu, Support vector machine for classification of meiotic recombination hotspots and coldspots in Saccharomyces cerevisiae based on codon composition, *Bio Med Central* **7**(1) (2006), 223.

[6] J. Peng, W. Haonan, W. Jiawei, S. Fei, S. Xiao and L. Zuhong, RF-DYMHC: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features, *Nucleic Acids Research* **35**(Web Server issue) (2007), W47–51.

[7] G. Liu, J. Liu, X. Cui and L. Cai, Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae, *Journal of Theoretical Biology* **293** (2012), 49–54.

[8] S. Ranka, T. Kahveci and M. Singh, Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, *International Conference on Bioinformatics* (2012), 279–304.

[9] L. Bin, W. Shanyi, L. Ren and C. Kuo-Chen, iRSpot-EL: identify recombination spots with an ensemble learning approach, *Bioinformatics (Oxford, England)* **33**(1) (2017), 35–41.

[10] Q. Wang-Ren, X. Xuan and C. Kuo-Chen, iRSpot-TNCPseAAC: identify recom-bination spots with trinucleotide composition and pseudo amino acid components, *International Journal of Molecular Sciences* **15**(2) (2014), 1746–1766.

[11] M.A.A. Maruf and S. Shatabda, iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components, *Genomics* **111**(4) (2019), 966–972.

[12] 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* **6**(1) (2016), 33483.

[13] L. Zhang and L. Kong, iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components, *Genomics* **111**(3) (2018), 457–464.

[14] Z. Lichao and K. Liang, iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components, *Journal of Theoretical Biology* **441** (2018), 1–8.

[15] G. Aquino, J. De Jesus Rubio and P.J. Novel, Nonlinear Hypothesis for the Delta Parallel Robot Modeling, *IEEE Access* **8**(1) (2020), 46324–46334.

[16] J. De Jesus Rubio, SOFMLS:online self-organizing fuzzy modified least-squares network, *IEEE Transactions on Fuzzy Systems* **17**(6) (2019), 1296–1309.

[17] H.S. Chiang, M.Y. Chen and Y.J. Huang, Wavelet-Based EEG Processing for Epilepsy Detection Using Fuzzy Entropy and Associative Petri Net, *IEEE Access* **7** (2019), 103255–103262.

[18] S. Zhang, K. Yang, Y. Lei and K. Song, iRSpot-DTS: Predict recombination spots by incorporating the dinucleotide-based spare-cross covariance information into Chou's pseudo components, *Genomics* **111**(6) (2019), 1760–1770.

[19] A. Zaheer Ullah Khan, iRSpot-SPI: Deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components, *Chemometrics and Intelligent Laboratory Systems* **189** (2019), 169–180.

[20] W. Chen, H. Lin, P.-M. Feng, C. Ding, Y.-C. Zuo and K.-C. Chou, iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties, *PLos One* **7**(10) (2012), e47843.

[21] J.L. Gerton, J. Derisi and M. Lichten, Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae, *Proceedings of the National Academy of Sciences* **97**(21) (2000), 11383–11390.

[22] K. Eunhee, M. Junhong and Y.J. Chul, A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction, *Medical Physics* **44**(10) (2017), e360–e375.

[23] K.H. Cha, L. Hadjiiski, R.K. Samala, H.P. Chan, E.M. Caoili and R.H. Cohan, Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets, *Medical Physics* **43**(4) (2016), 1882–1896.

[24] J. De Jesus Rubio, Stability Analysis of the Modified Levenberg-Marquardt Algorithm for the Artificial Neural Network Training, *IEEE Transactions on Neural Networks and Learning Systems* (2020), 1–15.

[25] G. Hernandez and E. Zamora, Hybrid neural networks for big data classification, *Neurocomputing* **390** (2020), 327–340.

[26] R. Tkachenko and I. Izonin, Model and Principles for the Implementation of Neural-Like Structures Based on Geometric Data Transformations, *Advances in Computer Science for Engineering and Education* (2019), 578–587.

[27] I. Izonin and R. Tkachenko, Multiple Linear Regression Based on Coefficients Identification Using Non-iterative SGTM Neural-like Structure, *Advances in Computational Intelligence* (2019), 467–479.

[28] H. Shengfeng, W. Rynson, H. Lau, L. Wenxi and H. Zhe, SuperCNN: A Superpixelwise Convolutional Neural Network for Salient Object Detection, *International Journal of Computer Vision* **115**(3) (2015), 330–344.

[29] T.C. Theelen, Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images, *IEEE Transactions on Medical Imaging* **35**(5) (2016), 1273–1284.

[30] D. Zhen, Y. Wu, M. Pei and Y. Jia, Vehicle Type Classification Using Unsupervised Convolutional Neural Network, *IEEE Transactions on Intelligent Transportation Systems* **16**(4) (2015), 1–10.

[31] Z. Zhang, S. Ding and Y. Sun, A support vector regression model hybridized with chaotic krill herd algorithm and empirical mode decomposition for regression task, *Neurocomputing* **410** (2020), 185–201.

[32] Z. Zhang, S. Ding and W. Jia, A hybrid optimization algorithm based on cuckoo search and differential evolution for solving constrained engineering problems, *Engineering Applications of Artificial Intelligence* **85** (2019), 254–268.

[33] G.F. Fan, et al., Forecasting electricity consumption using a novel hybrid model, *Sustainable Cities and Society* **61** (2020), 102320.

[34] Z. Zhang and W.C. Hong, Electric load forecasting by complete ensemble empirical mode decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm, *Nonlinear Dynamics* **98**(4) (2019), 1107–1136.

[35] M.W. Li, et al., Periodogram estimation based on LSSVR-CCPSO compensation for forecasting ship motion, *Nonlinear Dynamics* **97**(4) (2019), 2579–2594.

[36] Z. Chang-Jian, T. Hua, L. Wen-Chao, L. Hao, C. Wei and C. Kuo-Chen, iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition, *Oncotarget* **7**(43) (2016), 69783–69793.

[37] C. Wei, D. Hui, F. Pengmian, L. Hao and C. Kuo-Chen, iACP: a sequence-based tool for identifying anticancer peptides, *Oncotarget* **7**(13) (2016), 16895–169.