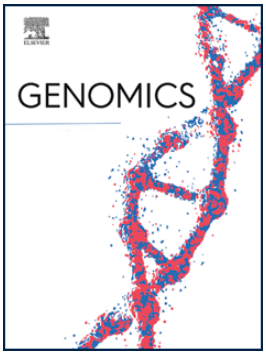

iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components

Lichao Zhang, Liang Kong

PII: S0888-7543(18)30135-6
DOI: doi:10.1016/j.ygeno.2018.03.003
Reference: YGENO 8993

To appear in: *Genomics*

Received date: 22 October 2017
Revised date: 27 February 2018
Accepted date: 3 March 2018

Please cite this article as: Lichao Zhang, Liang Kong, iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components, *Genomics* (2018), doi:10.1016/j.ygeno.2018.03.003

iRSpot-PDI: Identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components

Lichao Zhang[a,*], Liang Kong[b,*]

[a]*School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao, 066004, PR China*
[b]*School of Mathematics and Information Science & Technology, Hebei Normal University of Science & Technology, Qinhuangdao, 066004, PR China*

Abstract

Recombination spot identification plays an important role in revealing genome evolution and developing DNA function study. Although some computational methods have been proposed, extracting discriminatory information embedded in DNA properties has not been received enough attention. The DNA properties include dinucleotide flexibility, structure and thermodynamic parameter, which are significant for genome evolution research. To explore the potential effect of DNA properties, a novel feature extraction method, called iRSpot-PDI, is proposed. A wrapper feature selection method with the best first search is used to identify the best feature set. To verify the effectiveness of the proposed method, support vector machine is employed on the obtained features. Prediction results are reported on two benchmark datasets. Compared with the recently reported methods, iRSpot-PDI achieves the highest

*Corresponding author.
Email addresses: zhanglichaoouc@126.com (Lichao Zhang), kongliangouc@126.com (Liang Kong)

values of individual specificity, Matthew's correlation coefficient and overall accuracy. The experimental results confirm that iRSpot-PDI is effective for accurate identification of recombination spots. The datasets can be download from the URL: http://stxy.neuq.edu.cn/info/1095/1157.htm.

Keywords: Recombination hotspots, Property matrix, Diversity function, Support vector machine

1. Introduction

Caused by double-strand breaks(DSB), the genome is divided into two gametes for sexual reproduction, while diverse gametes combined together to form new genetic variations. Gene recombination is vital for cell division and is a key process to produce hereditary differences[1]. Therefore, gene recombination provides chances for the natural exchanges of genetic material[2] and makes genome produce new genetic variations and accelerates the process of biological evolution. The population geneticists have been interested in estimating the recombination rate[3–5], which has been documented that recombination rate is various among the different species, different chromosomes in the same species, and even in different regions within the same chromosomes for some species[6], whereas some single-stranded viruses have conserved recombination patterns[7]. Generally, regions with a high recombination rate are called as hotspots, and regions with a low recombination rate are called as coldspots[8].

Recombination correlates with various DNA sequence features, such as local GC content, dinucleotides bias, biased codon usage, palindromes [9, 10], and possible conserved sequences[11]. Apart from the DNA factors, many

epigenetic features, such as DNA methylation[12] and histone modification[13, 14], are also associated with recombination. In addition, PRDM9 protein, a meiosis-specific histone methyltransferase that trimethylates H3 at K4, is known to be a major determinant of meiotic recombination hotspots in human and mouse[15–17]. PRDM9 is a DNA-binding zinc finger protein, whose binding motif is enriched in hotspots. The rapid evolution of the zinc finger binding region of PRDM9 partially explains the hotspot-related genomic polymorphism[15] and the lack of conservation of hotspot position between human and chimpanzee[18].

With the development of bioinformatics, computational recombination spots identification methods have been developed by machine learning techniques due to that they can easily incorporate various sequence information[2, 8, 19–34]. RF-DYMHC[21] is constructed based on gapped dinucleotide composition. iRSpot-GAEnsC[29]is proposed by using nucleotide, dinucleotide and trinucleotide content. IDQD[22] uses k-mer approach and the increment of diversity. In addition, the weighted feature vector is given based on the dinucleotide properties[35]. However, the sequence order information is often ignored, which is proved to be beneficial for biology prediction by previous bioinformatics studies[19]. For this purpose, some methods are proposed based on DNA property to reflect the sequence order information, such as iRSpot-PseDNC[2], iRSpot-DACC-PCA[30], iRSpot-EL[23] and iRSpot-TNCPseAAC[19]. Some credible results have been obtained by above methods. To further excavate the sequence order information from DNA properties, this study develops a novel method to extract discriminatory information from the dinucleotide pairs at different positions along the

given sequence. A DNA property matrix is defined based on 15 DNA properties. Then, a formula called diversity function is designed based on the DNA property matrix to quantize the correlation of dinucleotide pairs at different positions along DNA sequence. After the features are extracted, a wrapper feature selection method with the best first search is applied to find the optimal feature set. These features are then used as inputs to support vector machine (SVM) classifier. The performance of the proposed method are evaluated on two benchmark datasets. The experimental results show that iRSpot-PDI is promising for recombination spots identification. Comparison of our method with recently reported methods shows that the proposed method provides a reliable and satisfying results. Meanwhile, the execution time results show that the proposed method has less time complexity after feature selection. iRSpot-PDI provides a novel way for using DNA property, which can be widely used for other tasks in bioinformatics such as DNA-binding protein identification, protein fold prediction, tumor classification and analysis, etc.

2. Materials and methods

In developing a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications [36–42], one should observe the Chou's 5-step rule [43]; i.e., (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the predic-

tion; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by-one.

2.1. Dataset

In order to obtain reliable prediction results and facilitate comparison with other existing methods, the datasets that are constructed by Jiang et al.[21] and Liu et al.[23] are selected to design and assess our method. For convenience, the first mentioned of two is expressed by S_1 and the second is expressed by S_2. The dataset S_1 can be expressed as

$$S_1 = S_1^+ \bigcup S_1^-$$

where S_1^+ is the subset of S_1 including 490 recombination hotspots with the relative hybridization ratios[44] higher than 1.5[21], S_1^- is the subset of S_1 including 591 recombination coldspots with the relative hybridization ratios[44] lower than 0.82[21], and $\bigcup$ is a mathematical operator representing union. Similar to S_1, dataset S_2 includes 478 hotspots and 572 coldspots with sequence similarity no more than 75%. The details of these two datasets are listed in Supplementary Materials S_1 and S_2.

2.2. DNA property Matrix

Several DNA properties have been used in recombination spots identification research [2, 45–49]. These properties include value of dinucleotide flexibility, structure and thermodynamic parameters, which are list in Table 1. To take advantage of these DNA properties, the DNA sequence needs to

be converted to a dinucleotide sequence (see Figure 1). The element of the dinucleotide sequence consists of two adjacent nucleotides. The order of a dinucleotide is defined by the order of the first nucleotide in the dinucleotide. According to the DNA properties listed in Table 1, a DNA sequence can be converted into a matrix $P = (p_{i,j})_{(L-1)\times 15}$, where L is the length of the sequence, and 15 is the total number of properties. $p_{i,j}$ denotes the jth property value of the ith dinucleotide pair which is composed of two adjacent nucleotides in the sequence. Here, the parameters of each DNA property are normalized by the following formula

$$\frac{X-\mu}{\sigma} \tag{1}$$

where the X is the original property value, μ and σ are the mean and standard deviation of property values, respectively. The matrix P with normalized DNA properties is called DNA property Matrix. It is obvious that the size of property matrix is diverse with different DNA sequences.

2.3. Feature extraction method

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 [50]. 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 (PseAAC)[51] was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics (see, e.g., [40, 52–55] as well as a long list of references cited in[56]. Because it has been widely and increasingly used, recently three powerful open access soft-wares, called PseAAC-Builder, propy, and PseAAC-General, are established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC[43], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as Functional Domain mode, and Sequential Evolution or PSSM mode. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC[46] has been developed for generating various feature vectors for DNA/RNA sequences. Particularly, recently a very powerful web-server called 'Pse-in-One' [57] and its updated version 'Pse-in-One2.0' [58] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences, and of course including the current ones. In this study, we are to use property diversity information (PDI) to define Chou's pseudo components for analyzing the recombination spots.

Similar to the general form of PseKNC[46, 59] for DNA/RNA sequences, the feature vector (FV) can be formulated as

$$\mathrm{FV} = [\psi_1, \psi_2, ..., \psi_u, ..., \psi_\Omega]^{\mathrm{T}} \tag{2}$$

where T is a transpose operator, the components ψ_u, $u = 1, 2, ..., \Omega$ depends on how to extract the desired information from statistical samples concerned, while the subscript Ω is an integer representing the dimension of feature vector FV.

Stimulated and encouraged by the success of PseKNC[46, 59], a novel formula is defined as

$$\psi_{g,k,j} = \frac{1}{L-2g-1} \sum_{i=g+1}^{L-g-1} \Theta_{i,g,k,j} \tag{3}$$

where $k, j = 1, 2, ..., 15$, $g = 1, 2, ..., G$. g is a distance factor which is defined by the difference between the orders of the corresponding dinucleotide pair. The distance factor reflects the degree of separation between dinucleotide pairs along the sequence (see Figure 1). G is the maximum of g. In order to reflect the correlation of various position dinucleotide along a DNA sequence, $\Theta_{i,g,k,j}$ called diversity function is defined by the following mathematical equations

$$\Theta_{i,g,k,j} = \theta_1 + \theta_2 \tag{4}$$

$$\theta_1 = (\frac{p_{i-g,k} - p_{i,j}}{2})^2 \tag{5}$$

$$\theta_2 = (\frac{p_{i,j} - p_{i+g,k}}{2})^2 \tag{6}$$

θ_1 describes the differentiation between the ith and the $i+g$th dinucleotide according to properties j and k. Similarly, θ_2 expresses the differentiation between the ith and the $i-g$th dinucleotide according to properties j and k. $\Theta_{i,g,k,j}$ reflects the diversity information of dinucleotide pairs on different DNA properties with the distance g. Specifically, $\Theta_{i,g,k,j}$ reflects the property differentiation between the dinucleotide pairs in g interval along the sequence as well as the forward and backward order information. Due to containing the content of dinucleotide, the arrangement of dinucleotide, the length of DNA sequence and the property differentiation of dinucleotide pairs, we think that Eq.3 could reflect diversity information of a given DNA sequence. According

to Eq.3, a given sequence can be converted into a feature vector with uniform size of $15 \times 15 \times G$.

2.4. Feature selection algorithm

With G increasing, the dimension of feature vector rises rapidly. The dimension is 225 with $G = 1$ and it rises up to 14625 with $G = 65$. However, the prediction results are not always improved with G increasing. As shown in Figure 2 and Figure 3, the values of overall accuracy and Matthew's correlation coefficient reach peak with $G = 5$, and these accuracies are not further improved with G reaching up to 13. Therefore, $G = 5$ is selected and a 1125 dimensional feature vector can be formulated as

$$\mathrm{FV}_{1125} = [\psi_{1,1,1}, \ldots, \psi_{1,15,15}, \ldots, \psi_{5,1,1}, \ldots, \psi_{5,15,15}]^{\mathrm{T}} \tag{7}$$

To further reduce the complexity and redundant features, feature selection is the process of identifying and removing as many irrelevant and redundant features as possible. This can help obtain a more efficient model and speed up the computational analysis. Many feature selection methods have been used in a wide range of bioinformatics studies[60]. Two main directions have been developed for feature selection: filter and wrapper. Filter approaches use statistical properties of the features to filter out poorly informative ones. This approach does not take into account the biases of the induction algorithms and selects feature subsets that are independent of the induction algorithms. Wrapper approaches use the target learning algorithm as a black box to estimate the worth of attribute subsets by measuring accuracy estimates. Feature wrappers often achieve better results than filters due to the fact that they are tuned to a specific prediction method.

Moreover, the wrapper feature selection approach provides some protection against overfitting because of the internal cross-validation function used for accuracy estimation. In this study, a wrapper feature selection process with SVM classifier is performed. Specifically, sequential forward selection strategy is adopted, in which features are sequentially added one by one to an empty candidate set until the addition of further features does not increase the criterion. The criterion is defined by the overall accuracy obtain from SVM with 5-fold cross validation. Finally, a 6 dimensional feature vector is obtained, and can be formulated as

$$\mathrm{FV}_6 = [\psi_{3,15,11}, \psi_{3,5,9}, \psi_{2,12,1}, \psi_{3,14,7}, \psi_{3,2,8}, \psi_{1,6,13}]^{\mathrm{T}} \tag{8}$$

For convenience, the feature extraction method obtaining the 6 features and the corresponding classifier are denoted as iRSpot-PDI.

2.5. Feature analysis

According to different distance factors g of Eq.3, the 1125 features of FV_{1125} can be divided into 5 feature subsets. The prediction result of each feature subset is shown in Table 2. From Table 2, we can see that the feature subsets corresponding to $g = 2$ and $g = 4$ perform better than other feature subsets. When $g = 3$, the sensitivity is the best. When $g = 2$, the specificity, overall accuracy and the Matthew's correlation coefficient are the highest. This illustrates that the feature subset of $g = 2$ is the most important to identify recombination spots. It can be concluded that the diversity information between the adjacent dinucleotide pairs with no intersection is significant to identify recombination spots. According to Eq.8, the

optimized features include 4 features with $g = 3$. These features reflect information between the adjacent ORFs. It reveals that the more sensitive with closer ORF to identify recombination spots just like Table 2 shown. As to the DNA properties, the selected features reflect the differentiation between the 15th and 11th properties, the 5th and 9th properties, the 12th and 1nd properties, the 14th and 7th properties, the 2nd and 8th properties, the 6th and 13rd properties, respectively. Among the selected DNA properties, two features reflect the differentiation information between flexibility and structure, two features reflect the differentiation information between flexibility and thermodynamic, and also one feature expresses the differentiation information between structure and thermodynamic. The differentiation information between the DNA properties may be related with some epigenetic factors which correlate with recombination, such as DNA methylation[12], histone modification[13, 14] and so on. Especially, the transcription factor PRDM9 is a major trans-regulator of recombination hotspots in the human and mouse genomes[15].

In order to show the affection of feature selection on identifying recombination spots, the results corresponding to FV_{1125} and FV_6 are listed in Table 3. As shown in Table 3, it is obvious that FV_6 performs better than FV_{1125}. The overall accuracy, the sensitivity, specificity and Matthew's correlation coefficient improve by 1.57%, 0.81%, 2.2% and 3.46%, respectively. The feature selection process is necessary to reduce the irrelevant and redundant information in all the 1125 features.

In order to analyze the power of each feature, the result of each feature in FV_6 is shown in Table 4. From Table 4, it is obvious that the feature $\psi_{3,5,9}$

is the most powerful feature of all the selected features. This feature reflects the differentiation sum of F-shift and twist property with the sequential three dinucleotide along the DNA sequence. The reason may be that the property such as F-shift and twist are more important to identify recombination spots. The results are similar to the previous study[23]. The parameters may correlate with the epigenetic factors. Another experiment is performed by adding features one by one, and the results are listed in Table 5. From Table 5, it is can be seen that the various accuracies keep increasing with adding features. Since the best performance is obtained by all the 6 features, it is concluded that the selected features are complementary.

2.6. Support vector machine

Although a wide range of classification algorithms have been proposed, support vector machine (SVM) which is always the most popular and best-performing for various biological problems. By using what is called the kernel trick, SVM can implicitly map the input data into a higher-dimensional feature space and then separate classes with a hyperplane. Among the four basic kernel functions including linear function, polynomial function, sigmoid function and Gaussian radial basis function (RBF), we select the RBF for its better performance than other kernel functions. Here, the publicly available software package LIBSVM [61] is employed to enforce the SVM classifier. The best combination of penalty parameter C and kernel parameter γ are selected by 5-fold cross validation with a simple but effective grid search strategy. The parameters C and γ are searched exponentially in the ranges of $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{5}]$ to probe the highest classification rate. Finally, the best overall accuracy is obtained by parameters $C = 4.5948$ and $\gamma = 24.2515$.

2.7. Prediction assessment

In statistical prediction, the following three methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test. However, among the three test methods, the jackknife test is deemed the least arbitrary (most objective) as elucidated in [62, 63]. Therefore, the jackknife test has been increasingly adopted or recognized by investigators to test the power of various prediction methods (see, e.g., [62]). Accordingly, the jackknife test is also adopted here to examine our method. Moreover, 5-fold cross validation is also employed to directly compare our results with the state of arts found in the literature. For comprehensive evaluation, the individual sensitivity (Sens), the individual specificity (Spec) and Matthew's correlation coefficient (MCC) over hotspots and coldspots, as well as the overall prediction accuracy over the entire dataset are reported.

To provide a more intuitive and easier-to-understand method to measure the prediction quality, here the criteria proposed in[64] is adopted. According to that criteria and demonstrated by recent publications(see, e.g., [37, 39, 42, 65–67]), the rates of correct predictions for the recombination hotspots and recombination coldspots are respectively defined by

$$\begin{cases} \Lambda^+ = (N^+ - N_-^+)/N^+, \text{ for the recombination hotspots} \\ \Lambda^- = (N^- - N_+^-)/N^-, \text{ for the recombination coldspots} \end{cases} \tag{9}$$

where N^+ is the total number of the recombination hotspots investigated while N_-^+ is the number of the recombination hotspots incorrectly predicted as the coldspots; N^- is the total number of the recombination coldspots investigated while N_+^- is the number of the recombination coldspots incorrectly

predicted as the hotspots. The overall success prediction rate is given by

$$\Lambda = \frac{N^+\Lambda^+ + N^-\Lambda^-}{N^+ + N^-} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{10}$$

It is obvious from Eqs.9 and 10 that, if and only if none of the recombination hotspots and the recombination coldspots are mispredicted, i.e., $N_-^+ = N_+^- = 0$ and $\Lambda^+ = \Lambda^- = 1$, we have the overall success rate $\Lambda = 1$. Otherwise, the overall success rate would be smaller than 1. On the other hand, it is instructive to point out that the following equation set is often used in literatures for examining the performance quality of a predictor

$$\begin{cases} \text{Sens} = \frac{\text{TP}}{\text{TP+FN}} \\ \text{Spec} = \frac{\text{TN}}{\text{TN+FP}} \\ \text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \\ \text{OA} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \end{cases} \tag{11}$$

where TP, TN, FP, FN represents the true positives, true negatives, false positives, false negatives, respectively. Sens, Spec, MCC and OA represent the sensitivity, specificity, Matthew's correlation coefficient and the overall accuracy. The relations between the symbols in Eq.10 and those in Eq.11 are given by:

$$\begin{cases} \text{TP} = N^+ - N_-^+ \\ \text{TN} = N^- - N_+^- \\ \text{FP} = N_+^- \\ \text{FN} = N_-^+ \end{cases} \tag{12}$$

Substituting Eq.12 into Eq.11 and also considering Eq.10, we obtain:

$$\begin{cases} \text{Sens} = \frac{1-N_-^+}{N^+} \\ \text{Spec} = \frac{1-N_+^-}{N^-} \\ \text{MCC} = \frac{1-\frac{N_-^+}{N^+}-\frac{N_+^-}{N^-}}{\sqrt{(1+\frac{N_+^- - N^+}{N^+})(1+\frac{N_-^+ - N^-}{N^-})}} \\ \text{OA} = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \end{cases} \tag{13}$$

From the above equation, we can see: when $N_-^+ = 0$, meaning none of the recombination hotspots is mispredicted to be a coldspot, we have the sensitivity Sens = 1; while $N_-^+ = N^+$, meaning that all the recombination hotspots are mispredicted to be the coldspots, we have the sensitivity Sens = 0. Likewise, when $N_+^- = 0$, meaning none of the recombination coldspots is mispredicted, we have the specificity Spec = 1; while $N_+^- = N^-$, meaning all the recombination coldspots are incorrectly predicted as the hotspots, we have the specificity Spec = 0. The Matthew's correlation coefficient is usually used for measuring the quality of binary (two-class) classifications. When $N_-^+ = N_+^- = 0$, meaning that none of the recombination hotspots and coldspots is mispredicted, we have MCC = 1; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have MCC = 0, meaning on better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$, we have MCC = -1, meaning total disagreement between prediction and observation. When $N_-^+ = N_+^- = 0$, meaning that none of the recombination hotspots and coldspots is incorrectly predicted, we have the overall accuracy OA = $\Lambda = 1$; while $N_-^+ = N^+$ and $N_+^- = N^-$, meaning that all the recombination hotspots and coldspots are mispredicted, we have the overall accuracy OA = $\Lambda = 0$. As we can see from the above discussion, it is much more intuitive and easier-to-understand when using Eq.13 to examine

the predictor for its sensitivity, specificity, Matthew's correlation coefficient and overall accuracy. It's worth noting that 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 [38, 68–72], system medicine[73, 74] and biomedicine [75], a completely different set of metrics as defined in [76] is needed.

3. Results and discussion

In this section, the performance of iRSpot-PDI is compared with other state-of-the-art methods on two benchmark datasets. The results are listed in Table 6 and Table 7. From Table 6, it can be seen that the highest sensitivity is obtained by iRSpot-DACC-PCA which is based on auto-cross covariance. The reason for its good performance on the sensitivity may be that the auto-cross covariance is sensibility to identify recombination spots. However, the values of specificity, MCC and overall accuracy obtained by iRSpot-PDI are 5.92%, 3% and 1.11% higher than iRSpot-DACC-PCA. In addition, although the principal component analysis (PCA) is employed to reduce the feature vector dimension, the features used by iRSpot-DACC-PCA (173 features) are still much more than those used by iRSpot-PDI (6 features). As to iRSpot-GAEnsc, which performs best on overall accuracy among all the compared methods, ensemble classification technique is adopted with 84 features. The overall accuracy of iRSpot-PDI is improved by 0.37% compared with iRSpot-GAEnsc. Compared with the previous methods on dataset S_2, iRSpot-PDI is evaluated by 5-fold cross validation. From the Table 7, it can be seen that the highest sensitivity is obtained by the model IDQD which use the k-mer

approach and the increment of diversity combined with quadratic discriminant analysis. However, its overall accuracy is 2.39% lower than iRSpot-PDI. Besides, the values of specificity, MCC and overall accuracy obtained by iRSpot-PDI are 3.75%, 1.48% and 0.51% higher than the previous best method iRSpot-EL.

Execution time of our method on two benchmark datasets is also measured to compare the computational cost of classifiers based on different feature sets. The jackknife test is performed on dataset S_1, while 5-fold cross validation is performed on dataset S_2. The proposed methods are tested on Intel R CoreTM i7 3.6GHz computer with 16.0 GB RAM using 64-bit OS. Figure 4 illustrates the total execution time for each classifier based on both feature set FV_{1125} and FV_6. The execution time includes time for feature extraction, training the classifier and testing the classifier in jackknife test or 5-fold cross validation. From Figure 4, it can be seen that the total execution time for classifier with feature set FV_6 is much less than the total execution time for classifier with feature set FV_{1125}. These results show the efficiency of the feature selection process in reducing the execution time. It is worth to mention that we are unable to provide a comparison report with existing methods in terms of computational complexity since the computational complexity of the existing methods are not reported in their papers.

From the classification results and execution time results, it can be seen that the proposed method could extract effective sequence information to identify recombination spots. This reveals that the diversity information between the various position dinucleotide is benefit to distinguish recombination spots. It is worth noting that the differentiation between various properties

according to iRSpot-PDI may correlate with many epigenetic factors. This may be helpful and beneficial to epigenetic research. As demonstrated in a series of recent publications(see, e.g., [37–39, 42, 66, 70]) in developing new prediction methods or bioinformatics tools, user-friendly and publicly accessible web-servers represent the future's trend [77] in developing new predictors and will significantly enhance their impacts [56], we shall make efforts in our future work to provide a web-server for iRSpot-PDI presented in this paper as well.

4. Conclusions

Considering the importance of diversity information based on DNA properties. A novel feature extraction method is constructed in our study. The proposed method iRSpot-PDI reveals the following 3 aspects. Firstly, the distance between dinucleotide pairs is related to identify recombination spots. Secondly, the performance is satisfied when the distance is 1 and 4 between the dinucleotide pairs. Thirdly, properties of F-shift and twist are important to identify recombination spots. The optimized SVM classifier and the cross validation tests are employed to predict and evaluate the propsed method on two widely used benchmark datasets. The results show that iRSpot-PDI outperformed in the underlying data sets. Although iRSpot-PDI provides a promising tool for recombination spots identification, some accuracies such as sensitivity is still not high enough. This is partly because that the extracted features do not cover sufficient sequence information to characterize the DNA sequence. In addition, due to that inference of recombination rates, hotspots from genome-wide SNP data and pedigree analysis[78–80] are important in

recombination research, the relations between biology and mathematics will be taken into consideration in our future work to obtain a better predictor.

Acknowledgement

The authors thank the anonymous referees for many valuable suggestions that have improved this manuscript. This work is supported by the National Natural Science Foundation of China (Grant No. 61602100), the Natural Science Foundation of Hebei Province (Grant No. F2016407082), the Youth Foundation of Hebei Educational Committee (Grant No. QN2015131), Doctoral Foundation of Northeastern University at Qinhuangdao (Grant No. XNB201613).

References

[1] Prosenjit Paul, Debjyoti Nag, Supriyo Chakraborty. Recombination hotspots: Models and tools for detection. DNA Repair, 2016, 40: 47-56.

[2] W. Chen, P.M. Feng, H. Lin, K.C. Chou. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res., 2013, 41(6):e68.

[3] Richard R. Hudson. Two-Locus sampling distributions and their application. genetics, 2001, 159:4,1805-1817.

[4] Kao Lin, Andreas Futschik, Haipeng LiA. Fast Estimate for the Population Recombination Rate Based on Regression. Genetics, 2013, 194:2,473-484.

[5] S Sheehan, YS Song. Deep Learning for Population Genetic Inference. PLoS Computational Biology, 2016, e1004845

[6] M.I. Jensen-Seaman, T.S. Furey, B.A. Payseur, Y. Lu, K.M. Roskin, C.F. Chen, M.A. Thomas, D. Haussler, H.J. Jacob. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Res., 2004, 14, 528-538.

[7] P. Lefeuvre, J.M. Lett, A. Varsani, D. Martin, J. Virol.. Widely Conserved Recombination Patterns among Single-Stranded DNA Viruses. Journal of Virology, 2009, 83, 2697-2707.

[8] C. Dong, Y. Yuan, F. Zhang, H. Hua, Y. Ye, A.A. Labena, H. Lin, W. Chen, F.B. 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., 2016,12, 2893-2900.

[9] Lobachev, K.S., Shor, B.M., Tran, H.T., Taylor, W., Keen, J.D., Resnick, M.A., Gordenin, D.A.. Factors affecting inverted repeat stimulation of recombination and deletion in Saccharomyces cerevisiae. Genetics, 1998, 148, 1507-1524.

[10] Nasar, F., Jankowski, C., Nag, D.K.. Long palindromic sequences induce double-strand breaks during meiosis in yeast. Mol. Cell Biol., 2000, 20, 3449-3458.

[11] Myers, S., Freeman, C., Auton, A., et al.. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. , 2008,40, 1124-1129.

[12] Maloisel, L., Rossignol, J.L.. Suppression of crossing-over by DNA methylation in Ascobolus. Genes Dev., 1998, 12, 1381-1389.

[13] Cesarini, E., DAlfonso, A., Camilloni, G.. H4K16 acetylation affects recombination and ncRNA transcription at rDNA in Saccharomyces cerevisiae. Mol. Biol. Cell, 2012, 23, 2770-2781.

[14] Yamada, S., Ohta, K., Yamada, T.. Acetylated Histone H3K9 is associated with meiotic recombination hotspots, and plays a role in recombination redundantly with other factors including the H3K4 methylase Set1 in fission yeast. Nucleic Acids Res., 2013, 41, 3504-3517.

[15] Myers, S., Bowden, R., Tumian, A., et al.. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science, 2010, 327, 876-879.

[16] Parvanov, E.D., Petkov, P.M., Paigen, K.. PRDM9 controls activation of mammalian recombination hotspots. Science, 2010, 327, 835.

[17] Baudat, F., Buard, J., Grey C., J., et al.. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science, 2010, 327, 836-840.

[18] Auton, A., Fledel-Alon, A., Pfeifer, S., et al.. A fine-scale chimpanzee genetic map from population sequencing. Science, 2012, 336, 193-198.

[19] W.R. Qiu, X. Xiao, K.C. Chou, iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci., 2014, 15(2), 1746-1766.

[20] H. Jiang, N. Li, V. Gopalan, M.M. Zilversmit, S. Varma, V. Nagarajan, J. Li, J. Mu, K. Hayton, B. Henschen, M. Yi, R. Stephens, G. McVean, P. Awadalla, T.E. Wellems, X. Su. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. Genome Biol., 2011, 12, R33.

[21] P. Jiang, H. Wu, J. Wei, F. Sang, X. Sun, Z. Lu. RF-DYMHC: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features. Nucleic Acids Res., 2007, 35, W47-W51.

[22] G. Liu, J. Liu, X. Cui, L. Cai. Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. J. Theor. Biol., 2012, 293, 49-54.

[23] B. Liu, S. Wang, R. Long, K.C. Chou. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics, 2017, 33(1), 35-41.

[24] T. Zhou, J. Weng, X. Sun, Z. Lu. Support vector machine for classification of meiotic recombination hotspots and coldspots in Saccharomyces cerevisiae based on codon composition. BMC Bioinf., 2006, 7, 223.

[25] G. Liu, H. Li. The correlation between recombination rate and dinucleotide bias in Drosophila melanogaster. J. Mol. Evol., 2008, 67, 358-367.

[26] G. Liu, H. Li, L. Cai. Processed pseudogenes are located preferentially in regions of low recombination rates in the human genome. J. Evol. Biol., 2010, 23, 1107-1115.

[27] L. Hansen, N.K. Kim, L. Mario-Ramrez, D. Landsman. Analysis of biological features associated with meiotic recombination hot and cold spots in Saccharomyces cerevisiae. PLoS One, 2011, 6, e29711.

[28] L.Q. Li, S.J. Yu, W.D. Xiao, Y.S. Li, L. Huang, X.Q. Zheng, S.W. Zhou, H. Yang. Sequence-based identification of recombination spots using pseudo nucleic acid representation and recursive feature extraction by linear kernel SVM. BMC Bioinf., 2014, 15, 340.

[29] Muhammad Kabir, Maqsood Hayat. iRSpot-GAEnsC: identifng recombination spots via ensemble classifer and extending the concept of Chous PseAAC to formulate DNA samples. Mol Genet Genomics, 2016, 291:285-296.

[30] B. Liu, Y. Liu, X. Jin, X. Wang, B. Liu. iRSpot-DACC: a computational predictor for recombination hot/cold spots identifcation based on dinucleotide-based auto-cross covariance. Sci. Rep., 2016, 6, 33483.

[31] S.H. Guo, L.Q. Xu. Recombination spots prediction using DNA physical properties in the saccharomyces cerevisiae genome. AIP Conference Proceedings, 2012,1479, 1556-1559.

[32] R. Wang, Y. Xu, B. Liu. Recombination spot identifcation Based on gapped k-mers. Sci Rep., 2016, 6: 23934.

[33] C. Li, M. Han, Y. Yang, W. Fei, X. Zheng, D. Zhang, S. Yi, J. Zhu, C. Wang, J. Li. Identification of Meiotic Recombination Spots Based on Phase-Specific Sequence and Burrows-Wheeler Transform. Journal of Computational and Theoretical Nanoscience, 2016, 13(5), 4131-4135.

[34] A.K. Dwivedi1, U. Chouhan. Comparative study of artificial neural network for classification of hot and cold recombination regions in Saccharomyces cerevisiae, Neural Comput Applic,DOI 10.1007/s00521-016-2466-6.

[35] G. Liu, Y. Xing, L. Cai. Using weighted features to predict recombination hotspots in Saccharomyces cerevisiae. Journal of Theoretical Biology, 2015, 382, 15-22. eudo amino acid composition. J. Theor. Biol., 2014, 344, 12-18.

[36] W. Chen, P. Feng, H. Yang, H. Ding. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget, 2017, 8, 4208-4217.

[37] P. Feng, H. Ding, H. Yang. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Molecular Therapy - Nucleic Acids, 2017, 7, 155-163.

[38] X. Cheng, X. Xiao. pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics. doi:10.1016/j.ygeno.2017.08.005 (2017).

[39] L.M. Liu, Y. Xu. iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med Chem., 2017, 13, 552-559.

[40] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci Rep., 2017, 7, 42362.

[41] W.R. Qiu, B.Q. Sun, X. Xiao. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Molecular Informatics. 36 (2017) UNSP 1600010.

[42] Y. Xu, C. Li. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med Chem., 2017, 13, 544-551.

[43] K.C. Chou. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology, 2011, 273, 236-247.

[44] J.L. Gerton, J. DeRisi, R. Shroff, M. Lichten, P.O. Brown, T.D. Petes. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc. Natl. Acad. Sci.USA, 2000, 97, 11383-11390.

[45] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, K.C. Chou, iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal. Biochem., 2014, 462, 76-83.

[46] W. Chen, H. Lin, K.C. Chou, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. BioSyst., 2015,11, 2620.

[47] W. Chen, H. Lin, P. M. Feng, C. Ding, Y. C. Zuo, K.C. Chou, iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties, PLoS One, 2012, 7, e47843.

[48] P. Feng, W. Chen, H. Lin, Prediction of CpG island methylation status by integrating DNA physicochemical properties. Genomics, 2014, 104, 229-233.

[49] 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, 2015, 31, 119-120.

[50] K.C. Chou, Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry.2015, 11, 218-234.

[51] K.C. Chou. Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics, 2001, 43, 246-255.

[52] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J. Theor. Biol., 2015, 364, 284-294.

[53] M. Behbahani, H. Mohabatkar, M. 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., 2016, 411, 1-5.

[54] H. Huo, T. Li, S. Wang, Y. Lv, Y. Zuo, L. Yang. Prediction of presynaptic and postsynaptic neurotoxins by combining various Chou's pseudo components. Sci Rep., 2017, 7, 5827.

[55] P. Tripathi, P.N. Pandey, A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. J Theor Biol.,2017, 424, 49-54.

[56] K.C. Chou. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Current Topics in Medicinal Chemistry, 2017, 17, 2337-2358.

[57] B. Liu, F. Liu, X. Wang, J. Chen. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research,2015, 43, W65-W71.

[58] B. Liu, H. Wu. Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein Sequences. Natural Science, 2017, 9, 67-91.

[59] Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., Lin, H., Chen, W., Chou, K.C. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics, 2014, 30, 1522-1529.

[60] Y. Saeys, I. Inza, P. Larranaga. A review of feature selection techniques in bioinformatics. Bioinformatics, 2007, 23, 2507-2517.

[61] C.C. Chang, C.J. Lin. LIBSVM: A Library for Support Vector Machines, 2001.

[62] K.C. Chou, H.B. Shen. Recent progress in protein subcellular location prediction, Anal. Biochem., 2007, 370, 1-16.

[63] H.B. Shen, J. Yang, X.J. Liu, K.C. Chou. Using supervised fuzzy clustering to predict protein structural classes. Biochem. Biophys. Res. Commun., 2005, 334, 577-581.

[64] Y. Xu, X.J. Shao, L.Y. Wu. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ, 2013, 1, e171.

[65] J. Jia, Z. Liu, X. Xiao, B. Liu, pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of Theoretical Biology, 2016, 394, 223-230.

[66] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics, 2016, 32, 362-369.

[67] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics. 10.1016/j.ygeno.2017.10.008 (2017).

[68] X. Cheng, X. Xiao, pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC. Molecular BioSystems, 2017, 13, 1722-1727.

[69] X. Cheng, X. Xiao. pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene, 2017, 315-321.

[70] X. Cheng, S.G. Zhao, W.Z. Lin. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics, 2017, 33, 3524-3531.

[71] X. Xiao, X. Cheng, S. Su. pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. Natural Science. 2017, 9, 331-349.

[72] X. Cheng, X. Xiao, pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics. doi:10.1016/j.ygeno.2017.10.002 (2017).

[73] X. Cheng, S.G. Zhao, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics, 2017, 33, 341-346.

[74] X. Cheng, S.G. Zhao, X. Xiao. iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget, 2017, 8, 58494-58503.

[75] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. iPTM-mLys: identifying multiple

lysine PTM sites and their different types. Bioinformatics, 2016, 32, 3116-3123.

[76] K.C. Chou. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. Molecular Biosystems, 2013, 9, 1092-1100.

[77] H.B. Shen. Recent advances in developing web-servers for predicting protein attributes. Natural Science, 2009,1, 63-92.

[78] Simon Myers, Leonardo Bottolo, Colin Freeman, Gil McVean, Peter Donnelly, A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome, Science, 2005, 310, 321-324.

[79] Gilean A. T. McVean, Simon R. Myers1, Sarah Hunt, Panos Deloukas, David R. Bentley, Peter Donnelly, The Fine-Scale Structure of Recombination Rate Variation in the Human Genome, Science, 2004, 304, 581-584.

[80] Michael P. H. Stumpf, Gilean A. T. McVean, Estimating recombination rates from population-genetic data. Nature Reviews Genetics, 2003, 4, 959-968.

Table 1: The original dinucleotide property.

Property index	AA/TT	AC/GT	AG/CT	AT	CA/TG	CC/GG	CG	GA/TC	GC	TA
F-roll	0.04	0.06	0.04	0.05	0.04	0.04	0.04	0.05	0.05	0.03
F-tilt	0.08	0.07	0.06	0.10	0.06	0.06	0.06	0.07	0.07	0.07
F-twist	0.07	0.06	0.05	0.07	0.05	0.06	0.05	0.06	0.06	0.05
F-slide	6.69	6.80	3.47	9.61	2.00	2.99	2.71	4.27	4.21	1.85
F-shift	6.24	2.91	2.80	4.66	2.88	2.67	3.02	3.58	2.66	4.11
F-rise	21.34	21.98	17.48	24.79	14.51	14.25	14.66	18.41	17.31	14.24
Roll	1.05	2.01	3.60	0.61	5.60	4.68	6.02	2.44	1.70	3.50
Tilt	-1.26	0.33	-1.66	0.00	0.14	-0.77	0.00	1.44	0.00	0.00
Twist	35.02	31.53	32.29	30.72	35.43	33.54	33.67	35.67	34.07	36.94
Slide	-0.18	-0.59	-0.22	-0.68	0.48	-0.17	0.44	-0.05	-0.19	0.04
Shift	0.01	-0.02	-0.02	0.00	0.01	0.03	0.00	-0.01	0.00	0.00
Rise	3.25	3.24	3.32	3.21	3.37	3.36	3.29	3.30	3.27	3.39
Energy	-1.00	-1.44	-1.28	-0.88	-1.45	-1.84	-2.17	-1.30	-2.24	-0.58
Enthalpy	-7.60	-8.40	-7.80	-7.20	-8.50	-8.00	-10.60	-8.20	-9.80	-7.20
Entropy	-21.30	-22.40	-21.00	-20.40	-22.70	-19.90	-27.20	-2.20	-24.40	-21.30

Note: The index in the first 6 lines denote dinucleotide flexibility parameters. The second 6 lines denote dinucleotide structure parameters, and the bottom 3 lines denote thermodynamic parameters.

Table 2: Prediction results corresponding to different distance factors on dataset S_1 by jackknife test.

Distance	Sens(%)	Spec(%)	MCC(%)	Overall accuracy (%)
1	69.39	91.71	63.37	81.59
2	70.20	92.05	64.50	82.15
3	70.82	89.85	62.33	81.22
4	70.61	91.71	64.43	82.15
5	70.00	90.86	62.85	81.41

Table 3: Prediction results corresponding to FV_{1125} and FV_6 on dataset S_1 by jackknife test.

Feature set	Sens(%)	Spec(%)	MCC(%)	Overall accuracy (%)
FV_{1125}	70.82	91.71	64.60	82.24
FV_6	71.63	93.91	68.06	83.81

Table 4: Prediction results corresponding to each feature in FV_6 on dataset S_1 by jackknife test.

Feature	Sens(%)	Spec(%)	MCC(%)	Overall accuracy (%)
$\psi_{3,5,11}$	63.27	92.89	59.66	79.46
$\psi_{3,5,9}$	66.53	90.52	59.42	79.65
$\psi_{2,12,1}$	10.20	96.45	13.37	57.35
$\psi_{3,14,7}$	20.61	92.89	19.84	60.13
$\psi_{3,2,8}$	66.53	86.29	54.28	77.34
$\psi_{1,6,13}$	9.39	89.00	-2.64	52.91

Table 5: The jackknife test results on dataset S_1 with adding feature one by one.

Feature	Sens(%)	Spec(%)	MCC(%)	Overall accuracy (%)
$\{\psi_{3,5,11}, \psi_{3,5,9}\}$	66.94	93.06	62.99	81.22
$\{\psi_{3,5,11}, \psi_{3,5,9}, \psi_{2,12,1}\}$	69.18	92.39	64.04	81.87
$\{\psi_{3,5,11}, \psi_{3,5,9}, \psi_{2,12,1}, \psi_{3,14,7}\}$	70.00	93.40	66.03	82.79
$\{\psi_{3,5,11}, \psi_{3,5,9}, \psi_{2,12,1}, \psi_{3,14,7}, \psi_{3,2,8}\}$	71.84	92.05	65.90	82.89
$\{\psi_{3,5,11}, \psi_{3,5,9}, \psi_{2,12,1}, \psi_{3,14,7}, \psi_{3,2,8}, \psi_{1,6,13}\}$	71.63	93.91	68.06	83.81

Table 6: Performance comparison of different methods on dataset S_1 by jackknife test.

Method	Sens(%)	Spec(%)	MCC(%)	Overall accuracy (%)
iRSpot-PseDNC	73.06	89.49	63.8	82.04
iRSpot-DACC	75.71	88.16	64.7	82.52
iRSpot-DACC-PCA	**76.33**	87.99	65.1	82.70
iRSpot-GAEnsC	73.77	79.92	54.0	83.44
iRSpot-PDI	71.63	**93.91**	**68.1**	**83.81**

Table 7: Performance comparison of different methods on dataset S_2 by 5-fold cross validation.

Method	Sens(%)	Spec(%)	MCC(%)	Overall accuracy (%)
iRSpot-EL	75.29	88.81	65.10	82.65
iRSpot-TNCPseAAC	76.56	70.99	47.37	73.52
iRSpot-PseDNC	71.75	85.84	58.30	79.33
IDQD	**79.52**	81.82	61.60	80.77
RF-DYMHC	73.01	86.56	60.49	80.40
iRSpot-PDI	71.84	**92.56**	**66.58**	**83.16**

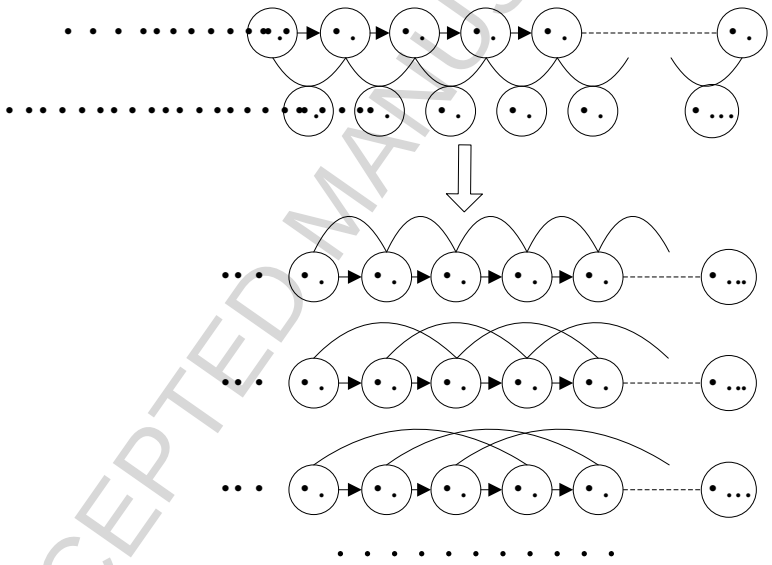

Figure 1: A schematic illustration to show the relation of relation of DNA sequence and dinucleotide sequence, and the distance between dinucleotide pairs.

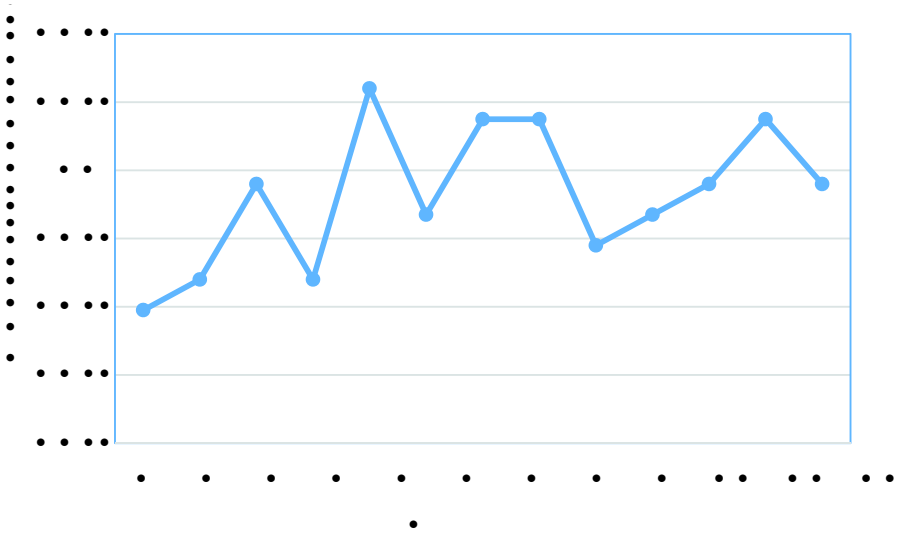

Figure 2: The curvilinear figure shows that the overall accuracies vary with different G.

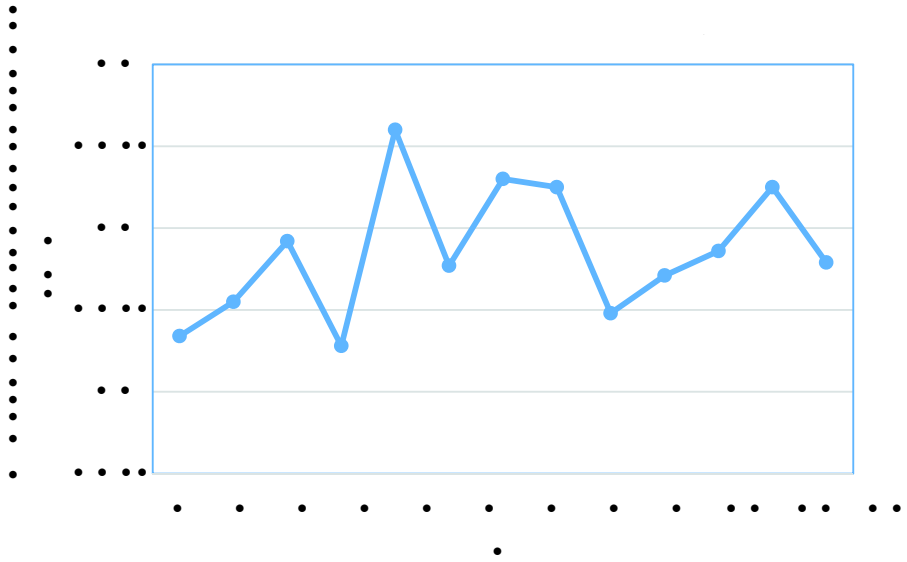

Figure 3: The curvilinear figure shows that the Matthews correlation coefficients vary with different G.

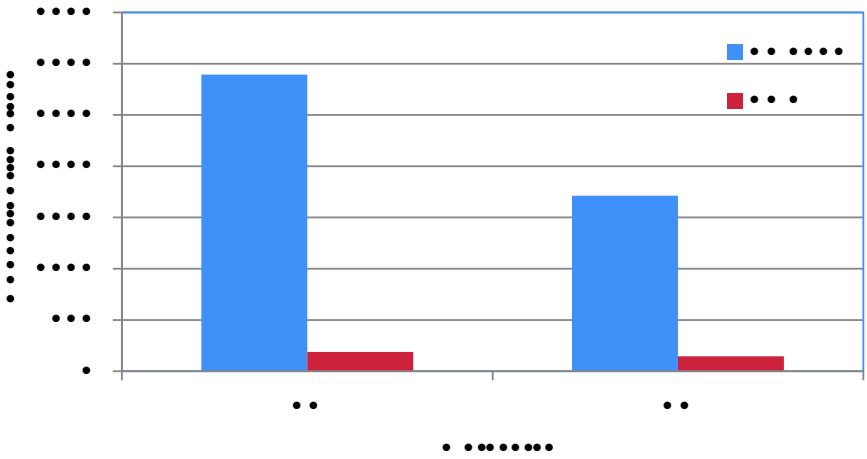

Figure 4: The total execute time of the proposed method based on SVM classifier for both feature set FV_{1125} and FV_6.

Highlights

- Features are extracted from DNA property to represent difference information.
- Difference information features can improve prediction significantly.
- Experimental results show that our feature extraction method is very promising.