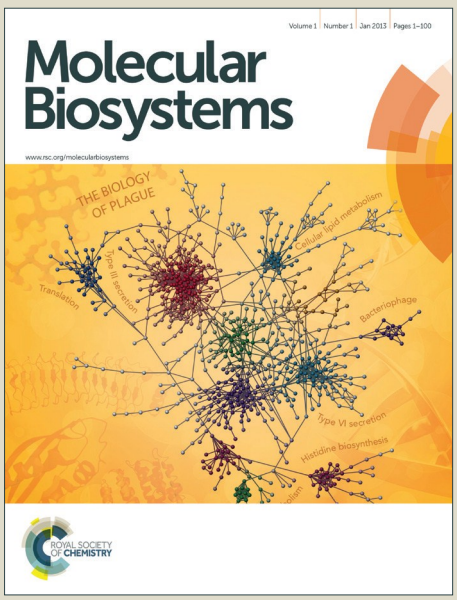
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Journal Name

ARTICLE

Combining the pseudo dinucleotide composition with the Z curve method to improve the accuracy of predicting DNA elements: a case study in recombination spots

Chuan Dong [a, b, c #], Ya-Zhou Yuan [a, b, c #], Fa-Zhan Zhang [a, b, c], Hong-Li Hua [a, b, c], Yuan-Nong Ye [d], Abraham Alemayehu Labena [a, b, c], Hao Lin [a, b, c], Wei Chen [e], and Feng-Biao Guo [a, b, c *]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Pseudo dinucleotide composition (PseDNC) and Z curve showed excellent performance in the classification issues of nucleotide sequences in bioinformatics. Inspired by the principle of Z curve theory, we improved PseDNC into the phase-specific PseDNC (psPseDNC). In this study, we used recombination spots prediction as a case to illustrate the capability of psPseDNC and also PseDNC fused with Z curve theory based on a novel machine learning method named large margin distribution machine (LDM). We verified that combining the two widely used approaches could generate better performance than only using the PseDNC with support vector machine based (SVM-based) model. The best Mathew's correlation coefficient (*MCC*) achieved by our LDM-based model was 0.7037 through the rigorous jackknife test and improved by ~6.6%, ~3.2%, ~2.4% compared with three previous studies. Similarly, the accuracy was improved by 3.2% compared with our previous iRSpot-PseDNC web server through independent data test. These results demonstrate that the joint use of PseDNC and Z curve has enhanced performance and can extract more information from a biological sequence. To facilitate researchers, we constructed a user-friendly web server for predicting hot/cold spots, HcsPredictor, which can be freely accessed from http://cefg.cn/HcsPredictor. In summary, we provided a united algorithm by integrating Z curve with PseDNC. We hope this united algorithm could be extended to other classification issue in DNA elements.

1. Introduction

Gene recombination and mutation in genomes are the most important driving forces in the process of biological evolution. Gene recombination in eukaryotes can lead to genetic information change, and the short contiguous DNA fragments can be also produced in bacteria through homologous recombination[1]. Therefore, recombination events can make genomes produce diversities even in the same species. Previous studies have also shown that the recombination rate has a large variation among different species, different chromosomes in the same species, and even in different regions within the same chromosomes for some species [2], whereas some single-stranded viruses have conserved recombination patterns [3]. Generally speaking, the regions with high recombination rate are called hot spots. In contrast, the

[a] *Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China. E-mail: fbguo@uestc.edu.cn; Tel: 86-28-83202351*
[b] *Center of Information in Biomedicine, University of Electronic Science and Technology of China, Chengdu, China.*
[c] *Key Laboratory for Neuro-information of the Ministry of Education, University of Electronic Science and Technology of China, Chengdu, China.*
[d] *School of Biology and Engineering, Guizhou Medical University.*
[e] *Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, China*
[#] Co-first authors
[*] Corresponding author.

Electronic Supplementary Information (ESI) available: [Supplementary information S1: Benchmark and independent dataset for *S. cerevisiae*; Supplementary information S2: Data of recombination spots in other species.].

regions with low recombination rate are called cold spots. The investigations of recombination event and identification hot spots have significance for understanding the genome evolution process. Traditionally, researchers used experimental and comparative genomics methods to determine recombination spots [2, 4, 5]. However, merely using experimental and comparative methods are both expensive and time-consuming in some cases. In addition, due to the vast amount of data, it is also unrealistic to determine those events by wet-lab experiment. As an alternative way, many researchers have focused on developing new computational methods to identify hot/cold spots[6-8]. Recently, Liu *et al*. introduced gapped k-mer to extract features from a sequence [9]. They used it to identify recombination spots. Better performance can be obtained by their method.

In 2013, we proposed a novel feature vectors named pseudo dinucleotide composition (PseDNC)[10], which considered six local DNA structural properties and used it to predict recombination spots in the genome of *Saccharomyces cerevisiae*. The results of 5-fold cross-validation and jackknife test showed a better classification performance than previous method [8]. PseDNC considered the sequence-order information and also the global composition information existing in nucleotide sequences[10, 11]. Based on PseDNC, our collaborators proposed pseudo k-tuple nucleotide compositions [12]. PseDNC or pseudo k-tuple nucleotide compositions has been successfully used in the issues of predicting recombination spots[10], nucleosome position[13], splice sites[14], translation initiation site[15], sigma-54 promoter [12], methylation sites[16], N 6-methyladenosine sites[17], replication origins[18], enhancers[19], microRNA precursors[20] and so on. To facilitate the use of researchers, our collaborators have constructed one online web-server and one standalone tool to generate various modes of pseudo nucleotide composition[21].

On the other hand, the Z curve feature has also shown excellent performance in classification issue of nucleotide sequences. In a graphical way, Z curve can transform a DNA sequence into a unique three-dimensional curve according to its special format[22, 23]. Owing to Z curve variables contain many forms from single to multi-nucleotides, a great deal of information can be reflected by it. This theory has been widely used in protein-coding gene recognition[24-27], exon and intron recognition[28], promoter recognition[29,30], translation start recognition[31] and nucleosome position mapping[32].

Encouraged by the success of PseDNC and the Z curve in classifying nucleotide sequences, in the present work we want to investigate that whether we could improve the classifying accuracy through joint use of them. We adopted two joint forms, one is using the phase-specific pseudo dinucleotide composition (psPseDNC) and the other is to fuse the Z curve variables and pseudo dinucleotide composition directly. Recombination spots prediction issue is chosen as a case study to show the power of combing the two feature extracting methods. Based on a novel method large margin distribution machine (LDM) we also build a user-friendly web server called HcsPredictor, which can be accessed from http://cefg.cn/HcsPredictor. HcsPredictor can be used to recognize hot/cold spots not only for *S. cerevisiae*, but also for other organisms such as *Homo sapiens*, *Mus musculus*, and *Escherichia coli*. This could be a start for predicting hot/cold spots in multiple species and we hope this united algorithm could be extended to other classification issue in DNA elements.

2. Materials and Methods

2.1. The recombination spots datasets in the genome of *S. cerevisiae*

We used the recombination spots in the genome of *S. cerevisiae* constructed by Liu *et al*.[8] as benchmark data set. It contains 490 recombination hot spots and 591 cold spots respectively. The trading-off parameters of LDM-based models were determined by 5-fold cross-validation. Gerton *et al*. ever estimated the recombination rate at a single gene level for *S. cerevisiae* using DNA microarray technology[5]. From this, we constructed an independent dataset through the following processes: excluding the genes overlapping with the benchmark data set, and the retaining DNA sequences were sorted in descent order according to their recombination rate. The top 288 and lowest 288 rank genes were selected as hot and cold spots, respectively. There was a sequence containing unusual base except 'A, T, G, C', so there were 575 genes in the final independent dataset. All of the sequences described above were downloaded from *S. cerevisiae* genome database (http://www.yeastgenome.org/). Both the benchmark and independent dataset can be obtained from Supplementary Information S1.

2.2. The recombination spots datasets in the genome of other species

We surveyed the cold/hot spots in Liu *et al*.'s study [8] and found that all of the sequences are genes. Therefore when we constructed the hot/cold dataset of *H. sapiens*, *M. musculus*, and *E. coli*, we also selected genes or ORFs (open reading frame). Firstly, the recombination rate in the above mentioned species was downloaded from ReDB database (http://www.bioinf.seu.edu.cn/ReDatabase/index.html) [33]. All of these data are from Jensen-Seaman M. I. *et al.* [2]. According to their recombination rate, the CDS (Coding DNA Sequence) regions with high and low recombination rate were downloaded from Ensembl (http://uswest.ensembl.org/info/data/ftp/index.html?redirect=no). Then some sequences were further excluded if they met any one of the following conditions: (1) 'N' appears in the sequences; (2) the length of sequence can't be divided by three; (3) genes are located in the negative chains. The highest 400 and lowest 400 genes in *H. sapiens* and *M. musculus* were obtained according to recombination rate. We used mean D_i values, which was defined in previous study, [1] located in the same locus of *E. coli* to measure their recombination rate. The highest 50 and lowest 50 genes were regarded as hot and cold spots. Those datasets can be obtained from Supplementary Information S2.

2.3. Large Margin Distribution Machine (LDM)

The margin distribution has a crucial influence on the performance of classifiers[34]. The generalization performance can be improved by optimizing margin distribution through maximizing the margin mean and minimizing the margin variance simultaneously [35]. Considering this algorithm optimizes margin distribution, it is called large margin distribution machine (LDM), which is leaded by the above idea. LDM optimizes the margin distribution through the first and second-order statistics, so it may have the advantage of more robust than classifiers only optimizing the margin. For examples LDM is not very sensitive to the changing of the LDM trading-off parameters. There are two solvers in LDM. The dual coordinate descent method can solve the dual problem, whereas the average stochastic gradient descent (ASGD) method is used to solve the classification of a large dataset. Generally speaking, two steps are needed when using LDM to implement classification. Firstly, the feature vectors are mapped into a high-dimensional space; secondly a hyper-plane, which maximizes the margin mean and minimizes the margin variance simultaneously, is then calculated to separate the samples easily. The LDM package can be downloaded at LAMDA group website (http://lamda.nju.edu.cn/code_LDM.ashx). We used it to perform classification. Due to the number of sequences in our dataset is not too large, we use dual coordinate descent method in the present work. There are four parameters (C, λ_1, λ_2, g) need to be optimized. C is the penalty parameter for measuring the losses of instances. λ_1 and λ_2 are the parameters for trading-off the margin variance. g is the parameter in RBF (Radial Basis Function) kernel. In order to obtain the best performance, we used exhaustive search to determine the four parameters via 5-fold cross-validation.

2.4.Z curve formulation

Now let us briefly describe the phase specific Z curve theory. Considering the bases A, C, G and T occurring in an ORF or a fragment of DNA sequence. Their frequencies at positions 1, 4, 7,…; 2, 5, 8, …, and 3, 6, 9, …, are denoted by a_1, c_1, g_1, t_1; a_2, c_2, g_2, t_2; a_3, c_3, g_3, t_3, respectively, then we can use the following equation (1) to calculate Z curve variables:

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i) \\ y_i = (a_i + c_i) - (g_i + t_i) \\ z_i = (a_i + t_i) - (g_i + c_i) \end{cases} \quad (1)$$
$$x_i, y_i, z_i \in [-1,1], i = 1,2,3$$

Therefore, the Z curve format transforms nucleotide sequences into three distributions with definite biological significance: purine versus pyrimidine, amino versus keto, weak hydrogen bonds versus strong hydrogen bonds. It also transforms the natural sequences into three groups of variables according to codon positions. In addition, phase specific dinucleotides occurring at the codon positions 1-2, 2-3, 3-1 were also taken into account. The following equation (2) is used to generate the phase specific dinucleotides Z curve variables.

$$\begin{cases} x_k^X = [p_k(XA) + p_k(XG)] - [p_k(XC) + p_k(XT)] \\ y_k^X = [p_k(XA) + p_k(XC)] - [p_k(XG) + p_k(XT)] \\ z_k^X = [p_k(XA) + p_k(XT)] - [p_k(XG) + p_k(XC)] \end{cases} \quad (2)$$
$$X = A, C, G, T; k = \text{1-2}, \text{2-3}, \text{3-1}$$

Where X=A, C, G, T in the above equation. Both phase-specific single nucleotide and phase-specific dinucleotides Z curve variables were considered in this study. Therefore 45 Z curve variables (9 variables for phase-specific single nucleotide Z curve and 36 variables for phase specific dinucleotides) can be obtained to characterize a DNA sequence.

2.5.PseDNC and psPseDNC theory

Inspired by the similar idea of Z curve theory, we improved PseDNC method and got phase specific PseDNC (psPseDNC). psPseDNC can reflect composition bias among three codon positions. A schematic illustration about psPseDNC is shown in Figure 1.

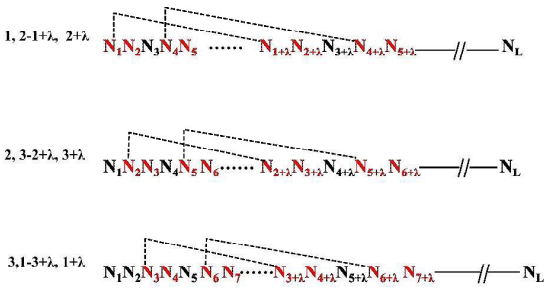


Figure 1 A schematic illustration to show the correlations of dinucleotides located in different phases in a DNA sequence.

In this Figure 1, the 1, 2, 3 on the left side represent the first, second, third phase in a DNA sequence, respectively. In psPseDNC algorithm, the nucleotides in 1-2, 2-3, 3-1 positions can interact with the dinucleotides located behind the 1+λ-2+λ, 2+λ-3+λ, 3+λ-1+λ positions. Their interactive relationships are represented by dotted lines. λ is an integer, which can represent the phase specific highest λ-tier rank of the correlation. The following equation (3) adopt similar form with PseDNC:

$$\begin{cases} \theta_{1-2,\lambda} = mean(\sum_{i\in 1} \Theta(N_i N_{i+1}, N_{i+\lambda} N_{i+1+\lambda})) \\ \theta_{2-3,\lambda} = mean(\sum_{i\in 2} \Theta(N_i N_{i+1}, N_{i+\lambda} N_{i+1+\lambda})) \\ \theta_{3-1,\lambda} = mean(\sum_{i\in 3} \Theta(N_i N_{i+1}, N_{i+\lambda} N_{i+1+\lambda})) \end{cases} \quad (3)$$

where $\theta_{1\text{-}2,\lambda}$, $\theta_{2\text{-}3,\lambda}$ and $\theta_{3\text{-}1,\lambda}$ are phase-specific order-correlated factors and reflect the sequence-order correlation. For details

about correlation function you can refer to our previous work[10]. In this study, the sequence feature vectors of each DNA can be calculated using PseDNC by incorporating to $\theta_{1\text{-}2,\lambda}$, $\theta_{2\text{-}3,\lambda}$, $\theta_{3\text{-}1,\lambda}$. There are three phases in a DNA sequence, therefore a DNA sequence is now represented by a (16+λ)×3 dimensional vector.

2.6. Mixed Variables

As mentioned above, PseDNC and Z curve method were used to generate the identified variables. In total, three groups of variables were considered, including PseDNC, PseDNC fused with Z curve directly and psPseDNC. Because we used Z curve variables only for single nucleotides (9 variables) and dinucleotides (36 variables), so there are a total of 16+λ, 16+λ+9+36=61+λ, and (16+λ)×3=48+3×λ variables for PseDNC, PseDNC fused with Z curve and psPseDNC, respectively. All of the variables were scaled to [0, 1] using the following equation,

$$fv = \frac{fv^{(0)} - \min(fv^{(0)})}{\max(fv^{(0)}) - \min(fv^{(0)})} \quad (4)$$

where $fv^{(0)}$ represents the initial feature vector, and min($fv^{(0)}$), max($fv^{(0)}$), fv represent minimal value, maximum value, scaled feature vector in this equation, respectively. It was observed via preliminary trials that when the variables λ, ω of PseDNC and psPseDNC are 3, 0.05, the proposed method yields the best predictive results for the identification of recombination spots.

2.7. Cross-validation and jackknife test

N-fold cross-validation technology, bootstrap test, independent dataset test and jackknife test are often used to assess the performance of classification methods. N-fold cross-validation refers to that the datasets are randomly partitioned into n subsets, then the n-1 subsets are used for training model and the remaining one is used as testing dataset. Totally N times were performed in turn in n-fold cross-validation process. Every sample will be used for testing and others are used for building a model if n equal to the number of samples. This is also called jackknife test. In this work, the trading-off parameters of LDM were obtained via 5-fold cross-validation. Due to the uniqueness of jackknife test and independent dataset test, there is no evaluating bias using the two methods. Herein, we used 5-fold cross-validation, jackknife test and independent test to evaluate the performance of our classifier. If we randomly separate the training data into five sub-samples, there are many possibilities. In order to avoid the bias of estimation, we did totally 10 times 5-fold cross-validation and used the average performance of them as the final result.

2.8. Performance evaluation

We used specificity (*Sp*), sensitivity (*Sn*), accuracy (*Acc*), the Mathew's correlation coefficient (*MCC*), precision and recall to evaluate the performance of our methods. They are often formulated using the following equations,

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \\ Sp = \dfrac{TN}{TN + FP} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \\ Acc = \dfrac{TP + TN}{TP + TN + FP + FN} \\ Precision = \dfrac{TP}{TP + FP} \\ Recall = \dfrac{TP}{TP + FN} \end{cases} \quad (5)$$

where *TP* represents the number of true positive samples in our prediction result and *TN*, *FP*, *FN* represent the number of true negative, false positive and false negative samples, respectively. Therefore, *TP*+*FN*， *TN*+*FP* represent the number of positive and negative samples. That means *Sn* and *Sp* can reflect the correctively predicted percentage of positive and negative samples. *MCC* has the range from -1 to 1. For the range of *MCC* from 0 to 1, it means prediction results are better than random prediction, otherwise worse than random prediction (-1<*MCC*<0). Precision represents the percentage occupied by real hot spots in those predicted as hot spots. Meanwhile, the ROC (receiver operating characteristic) curve was also used to evaluate the performance of the current method, where its vertical coordinate is for the true positive rate (sensitivity) and the horizontal coordinate for the false positive rate. The best possible prediction method would yield a point with the coordinate (0, 1) representing 100% sensitivity and 0 false positive rate or 100% specificity. Therefore, the (0, 1) point is also called a perfect classification. A completely random guess would give a point along a diagonal from the point (0, 0) to (1, 1). The area under the ROC curve, called AUC (area under the curve of ROC), is often used to indicate the performance quality of a binary classifier: the value 0.5 of AUC is equivalent to random prediction, while 1 of AUC represents a perfect one. In fact, from equations (5) we know that recall and *Sn* have same mathematic style, therefore we only listed *Sn* in tables among the two evaluators.

Journal Name

ARTICLE

3. Results and Discussion

3.1. Recombination hot/cold spots in the genome of *S. cerevisiae*

Previously, we have proposed a SVM-based method to identify recombination hot/cold spots of *S. cerevisiae* by using PseDNC. Here, we try to improve the performance by using psPseDNC, PseDNC fused with Z curve combining with LDM. In order to evaluate the performance of our method, we performed ten time 5-fold cross-validations. The mean values of *Sn*, *Sp*, *Acc*, *MCC*, precision, *AUC* are summarized in Table 1.

Table 1 Results of different methods from 5-fold cross-validation test on *S. cerevisiae* benchmark.

Machine learning methods	Methods	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	*Precision (%)*	*AUC*
	PseDNC	68.98	91.29	81.17	0.6249	86.78	0.8720
SVM-model	PseDNC+Z	80.35	85.79	83.32	0.6629	82.42	0.9061
	psPseDNC	77.39	88.76	83.60	0.6689	86.00	0.9125
	PseDNC	70.37	90.78	81.53	0.6307	86.36	0.8752
LDM-model	PseDNC+Z	77.82	89.78	84.36	0.6846	86.33	0.9087
	psPseDNC	78.22	88.05	83.60	0.6685	84.42	0.9080
QD-model	IDQD(*Liu et al.* [8])	79.40	81.00	80.30	0.6030	-	-

The trading-off parameters in LDM-based models are (C, λ_1, λ_2) = 3, 2^5, 2^8 and g=0.4 for PseDNC; (C, λ_1, λ_2) = 2, 2^7, 2^7, and g=0.8 for PseDNC+Z; (C, λ_1, λ_2) = 3, 2^1, 2^8, and g=0.6 for psPseDNC. In SVM-based models C=8, γ=0.125 for PseDNC; C=2, γ=2 for PseDNC+Z; C=2, γ=0.5 for psPseDNC.

Not only for SVM-based models but also for LDM-based models, psPseDNC and PseDNC fused with Z curve always had better performance compared with PseDNC. For LDM-based models, the *MCC* of using variables PseDNC fused with Z curve, psPseDNC are improved by 5.4% and 3.8% than only using PseDNC, respectively. For SVM-based models the *MCC* of using variables PseDNC fused with Z curve, psPseDNC were improved by 3.8% and 4.4% than only using PseDNC, respectively. In addition, PseDNC fused with Z curve, psPseDNC can obtain higher AUC score compared with merely using PseDNC. The improved results hold both in LDM and SVM based models, illustrating that better result could be obtained after adding the Z curve variables or using its phase specific idea. Therefore, we can conclude safely that Z curve can also reflect more information about recombination events and can be used to predict recombination spots or other DNA elements. In addition, we find that the method of PseDNC fused with Z curve has the best performance for the LDM-based models, *i.e.*, an *MCC* of 0.6846 with an accuracy of 84.36% were obtained by our method. The *Acc* and *MCC* in our best result were improved by 4.06% and 8.1%, respectively than previous study [8].

In order to carry out an objective evaluation for our method, we did rigorous jackknife test on our dataset using the parameters determined by 5-fold cross-validation test. The result details of jackknife test are listed in Table 2. In order to give a more objective evaluation, we also listed another two jackknife results, which can be obtained from two previous studies[36,37].

Table 2 Results of jackknife test based on different methods and models.

Machine learning methods	Method	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	Precision *(%)*	*AUC*
	PseDNC (Chen *et al* [10])	73.06	89.49	82.04	0.6380	-	-
SVM-model	PseDNC+Z	81.63	86.97	84.55	0.6878	83.86	0.9126
	psPseDNC	77.76	89.34	84.09	0.6789	85.81	0.9158
	PseDNC	71.02	90.86	81.87	0.6374	86.57	0.8780
LDM-model	PseDNC+Z	78.78	90.69	85.29	0.7037	87.53	0.9118
	psPseDNC	77.96	88.83	83.90	0.6750	85.27	0.9132
SVM-model	iRSpot-TNCPseAAC [36]	87.14	79.59	83.72	0.6710	-	-
SVM-model	Li *et al.* [37]	76.12	90.69	84.09	0.6800	-	-

The trading-off parameters in LDM-based models are (C, λ_1, λ_2) = (3, 2^5, 2^8), and g=0.4 for PseDNC; (C, λ_1, λ_2) = (2, 2^7, 2^7), and g=0.8 for PseDNC+Z; (C, λ_1, λ_2) = (3, 2^1, 2^8), and g=0.6 for psPseDNC. In SVM-based models C=2, γ=2 for PseDNC+Z; C=2, γ=0.5 for psPseDNC.

As shown in the table, the PseDNC fused with Z based on LDM has the best performance with an *MCC* value of 0.7037 and accuracy of 85.29%. Meanwhile, PseDNC fused with Z curve based on LDM can also obtain better AUC and precision. It is much better than our previous method. The ROC curves shown in Figure 2 can further demonstrate the better performance of our new methods. Compared with other two available studies *MCC* and *Acc* were improved as well [36,37].

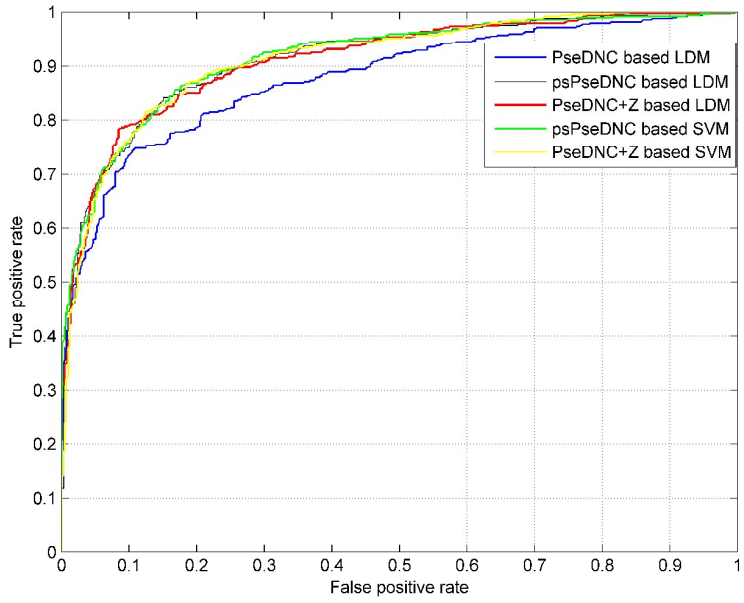


Figure2 ROC curves of different methods for identifying recombination hot/cold spots.

Our independent dataset contains 287 positive samples and 288 negative samples. In the saving file all the 287 positive samples are listed ahead. To construct unbalanced test set, we submitted additional 100 samples to HcsPredictor and iRSpot-PseDNC web server every time according to the storing order in the independent dataset. For example in the first time we submitted 100

sequences and they are all positive samples; in the second time we submitted 200 sequences and they are all positive samples too; in the third time we submitted 300 sequences and they are 287 positive samples and 13 negative samples. We repeated this process until all of the sequences were submitted. Finally we obtained an accuracy of 67% among 100 firstly submitted positive samples, while iRSpot-PseDNC obtained 61%. Our new method got an accuracy of 67% when we submitting 200 positive samples, while iRSpot-PseDNC obtained 59.5%. Our method achieved accuracies of 66.3% and 71.5%, respectively when we submitting 300 and 400 simples with unbalanced number, whereas iRSpot-PseDNC obtained 59.3% and 66.25%, respectively. After submitting all of them into the two web servers, we obtained an accuracy of 76.52% on independent dataset, and the Acc was improved by 3.2% compared with iRSpot-PseDNC. These results further illustrated that PseDNC fused with Z curve has better classification performance than PseDNC as we expect. Improved results may give the credit to the following two points. Firstly, LDM optimizes the margin distribution, but SVM merely optimizes the single margin. Secondly, we combined PseDNC and Z curve variables as the input vectors. Since Z curve and PseDNC are two different algorithms, they can reflect different information in a DNA sequence. Z curve can transform a natural sequence into three groups of variables according to codon positions. And the three group features from Z curve can represent three independent distributions such as purine/pyrimidine, amino/keto and strong-H bond/weak-H bond bases, respectively. In addition, the considerable sequence feature, especially for local and global information, can be contained by PseDNC, therefore if representing a DNA sequence according to Z curve and PseDNC, more information can be reflected. If we adopt feature eliminate technology, the performance can be further improved, however, we did not do this, because our main aim is not to improve the performance of predicting recombination spots, but rather to prove that the classification performance could be improved by combining the Z curve and PseDNC methods, and recombination spot prediction serves only a case study. Furthermore, another attempt was to introduce LDM into the bioinformatics field. For the same feature vectors, *MCC* can be improved by 1~2% compared LDM with SVM-based models on this issue. Because of its better performance than SVM, LDM has potential to be used as a supplementary tool in other classifying issues of DNA elements.

3.2. Recombination hot/cold spots in the genomes of other species

Because the united form of PseDNC and Z curve variables based on LDM gave the best *MCC* and accuracy in predicting the recombination spots in the genome of *S. cerevisiae*, we extended our method to the genomes of *H. sapiens*, *M. musculus*, and *E. coli*. Table 3 summarized the results obtained from jackknife test on those species.

Table 3 Predictive results for recombination hot/cold spots using jackknife test in other species' genomes

Species	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	Precision *(%)*	*AUC*
H.sapiens	84.00	72.25	78.13	0.5664	75.17	0.8450
M. musculus	76.25	74.50	75.38	0.5076	74.94	0.8263
E. coli	80.00	58.00	69.00	0.3895	65.57	0.6872

C=6, $\lambda_1=2^{-6}$, $\lambda_2=2^{-1}$, g=0.7 for *H. sapiens*; C=5, $\lambda_1=2^{-8}$, $\lambda_2=2^{8}$, g=1 for *M. musculus*; C=1, $\lambda_1=2^{-8}$, $\lambda_2=2^{5}$, g=0.1 for *E. coli*.

Comparing Table 1 and Table 2, it is obvious that the result is still higher than random prediction though it is not as good as in *S. cerevisiae*. This suggested that recombination may be a complex event, and species from different domains may adopt different mechanism and signals for recombination event. In addition, we also performed across organism prediction and used the model from *S. cerevisiae* to predict hot/cold spots in *H. sapiens* and *E. coli*. A very poor performance was obtained. Most of the input sequences were predicted as hot spots. This may result from the distantly phylogenetic relationship between them. Inversely, we also used the LDM-based models of *S. cerevisiae*, *H. sapiens*, *M. musculus*, and *E. coli* with their best trading-off parameters, to predict the independent dataset of *S. cerevisiae*. Consequently, accuracies with 76.52%, 61.22%, 52.35% and 41.91% were obtained respectively. Given that the first three genomes are eukaryotes, and the last one *E. coli* is one of prokaryotes. It can be concluded that recombination mechanism/signal, or at least the prediction model, is related to phylogenetic distance. Therefore, each genome may need specific model to predict recombination spots accurately. Aiming to this, we build an LDM-based web server called HcsPredictor to identify recombination spots in each of the four genomes. It adopts the united form of PseDNC and Z curve. HcsPredictor can be freely available from http://cefg.cn/HcsPredictor/.

3.3 HcsPredictor: a web server for predicting recombination spots in multi-species

The home page (http://cefg.cn/HcsPredictor) of this web server was shown in Figure 3.

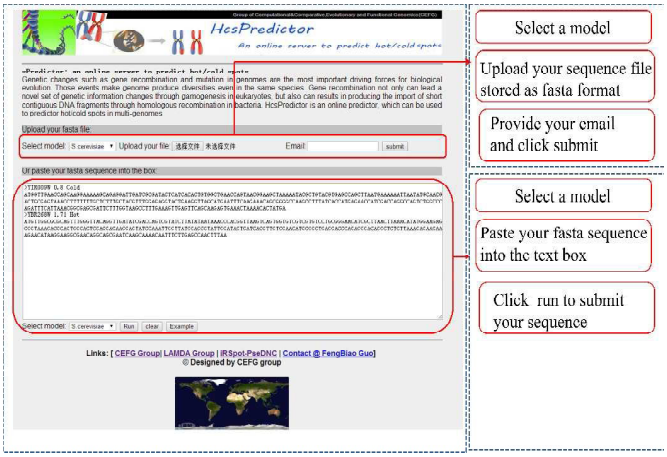


Figure 3 The screenshot of HcsPredictor web server.

HcsPredictor can not only predict recombination spots in *S. cerevisiae*'s genome, but also in *H. sapiens*, *M. musculus*, and *E. coli*. Due to our models were trained using gene and open reading frames (ORFs) merely, the sequences under prediction should be ORFs. Based on the discussion in above section, the performance of hot/cold classification was influenced by the phylogenetic distance. Therefore the model of species, which has closest evolutionary distance with the submitting sequences should be selected before using this web server to perform prediction. Two ways are provided to submit query sequence. The first way is that users can paste their sequences into the box and obtain the result from the web server directly. Alternatively, they can also upload a file with fasta format and must provide their email address simultaneously. In this way, the results will be returned to the mailbox after the web server completing the prediction. This could be a start of predicting recombination spots in multiple species and we hope it could arouse more novel computational models and feature selection techniques on this issue.

4. Conclusions

As a case study we used psPseDNC and PseDNC fused with Z curve to predict recombination spots. We obtained much better performance than PseDNC (see Table 1 and Table 2). The best Mathew's correlation coefficient (*MCC*) achieved by our LDM-based model was 0.7037 through the rigorous jackknife test and improved by ~6.6%, ~3.2%, ~2.4% compared with three previous studies, respectively. Similarly, the accuracy was improved by 3.2% compared with our previous iRSpot-PseDNC web server through independent data test. And also the results from cross species prediction demonstrated that species from different domains may adopt different mechanism and signals for recombination event.

Abbreviations

PseDNC, pseudo dinucleotide composition; psPseDNC, phase-specific PseDNC; LDM, large margin distribution machine; SVM, support vector machine; *MCC*, Mathew's correlation coefficient; *S.cerevisiae*, *Saccharomyces cerevisiae*; *H.sapiens*, *Homo sapiens*; *M.musculus*, *Mus musculus*; *E.coli*, *Escherichia coli*; *Sn*, sensitivity; *Sp*, specificity; *Acc*, accuracy; *AUC*, Area under the curve of ROC; ROC, The curves of receiver operating characteristic curve.

Conflict of interest statement.

The authors declare that they have no any conflict of interest.

Author contributions

FB Guo conceived, designed, coordinate the study. C Dong analyzed the data. C Dong, YZ Yuan, FZ Zhang, HL Hua, and YN Ye constructed the datasets. YZ Yuan and FZ Zhang double checked the results. C Dong, YZ Yuan, AA Labena and FB Guo wrote the manuscript. W Chen, H Lin gave us many advises about this work. All of authors read and approved this is the final manuscript.

Acknowledgements

We gratefully acknowledge Dr. Teng Zhang and Prof. Zhi-Hua Zhou for kindly providing open source codes of LDM and helping us to understand LDM algorithm, Dr. koji Yahara for providing *E. coli* recombination data. We are also indebted to thank the funding for the open access charge: The National Natural Science Foundation of China [31470068]; Sichuan Youth Science and Technology Foundation of China [2014JQ0051]; Fundamental Research Funds for the Central Universities of China [ZYGX2015Z006 and ZYGX2015J144].

References

1. K. Yahara, X. Didelot, M. A. Ansari, S. K. Sheppard and D. Falush, *Mol. Biol. Evol.*, 2014, 31, 1593-1605.
2. M. I. Jensen-Seaman, T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C. F. Chen, M. A. Thomas, D. Haussler and H. J. Jacob, *Genome Res.*, 2004, 14, 528-538.
3. P. Lefeuvre, J. M. Lett, A. Varsani and D. Martin, *J. Virol.*, 2009, 83, 2697-2707.
4. J. Pan, M. Sasaki, R. Kniewel, H. Murakami, H. G. Blitzblau, S. E. Tischfield, X. Zhu, M. J. Neale, M. Jasin, N. D. Socci, A. Hochwagen and S. Keeney, *Cell*, 2011, 144, 719-731.
5. J. L. Gerton, J. DeRisi, R. Shroff, M. Lichten, P. O. Brown and T. D. Petes, *Proc. Natl. Acad. Sci. USA.*, 2000, 97, 11383-11390.
6. T. Zhou, J. Weng, X. Sun and Z. Lu, *BMC bioinformatics*, 2006, 7, 223.
7. P. Jiang, H. Wu, J. Wei, F. Sang, X. Sun and Z. Lu, *Nucleic acids Res.*, 2007, 35, W47-W51.
8. G. Liu, J. Liu, X. Cui and L. Cai, *J. Theor. Biol.*, 2012, 293, 49-54.
9. R. Wang, Y. Xu and B. Liu, *Sci. Rep.*, 2016, 6, 23934.
10. W. Chen, P. M. Feng, H. Lin and K. C. Chou, *Nucleic Acids Res.*, 2013, 41, e68.
11. W. Chen, H. Lin and K. C. Chou, *Mol. Biosyst.*, 2015, 11, 2620-2634.
12. H. Lin, E. Z. Deng, H. Ding, W. Chen and K. C. Chou, *Nucleic Acids Res.*, 2014, 42, 12961-12972.
13. S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen and K. C. Chou, *Bioinformatics*, 2014, 30, 1522-1529.
14. W. Chen, P. M. Feng, H. Lin and K. C. Chou, *Biomed. Res. Int.*, 2014, 2014.
15. W. Chen, P. M. Feng, E. Z. Deng, H. Lin and K. C. Chou, *Anal. Biochem.*, 2014, 462, 76-83.
16. Z. Liu, X. Xiao, W. R. Qiu and K. C. Chou, *Anal. Biochem.*, 2015, 474, 69-77.
17. W. Chen, P. Feng, H. Ding, H. Lin and K. C. Chou, *Anal. Biochem.*, 2015, 490, 26-33.
18. W. C. Li, E. Z. Deng, H. Ding, W. Chen and H. Lin, *Chemometrics Intellig. Lab. Syst.*, 2015, 141, 100-106.
19. B. Liu, L. Fang, R. Long, X. Lan and K. C. Chou, *Bioinformatics*, 2015, btv604.
20. B. Liu, L. Fang, F. Liu, X. Wang and K. C. Chou, *J. Biomol. Struct. Dyn.*, 2015, 1-13.
21. W. Chen, T. Y. Lei, D. C. Jin, H. Lin and K. C. Chou, *Anal. Biochem.*, 2014, 456, 53-60.
22. C. T. Zhang and R. Zhang, *Nucleic Acids Res.*, 1991, 19, 6313-6317.
23. R. Zhang and C. T. Zhang, *J. Biomol. Struct. Dyn.*, 1994, 11, 767-782.
24. C. T. Zhang and J. Wang, *Nucleic acids Res.*, 2000, 28, 2804-2814.
25. L. L. Chen, H. Y. Ou, R. Zhang and C. T. Zhang, *Biochem. Biophys. Res. Commun.*, 2003, 307, 382-388.
26. F. B. Guo, H. Y. Ou and C. T. Zhang, *Nucleic Acids Res.*, 2003, 31, 1780-1789.
27. Z. G. Hua, Y. Lin, Y. Z. Yuan, D. C. Yang, W. Wei and F. B. Guo, *Nucleic acids Res.*, 2015, W85–W90.
28. Y. Wu, A. W. Liew, H. Yan and M. Yang, *Phys. Rev. E. Stat. Nonlin. Soft Matter. Phys.*, 2003, 67, 061916.
29. J. Y. Yang, Y. Zhou, Z. G. Yu, V. Anh and L. Q. Zhou, *BMC bioinformatics*, 2008, 9, 113.
30. K. Song, *Nucleic Acids Res.*, 2012, 40, 963-971.
31. H. Y. Ou, F. B. Guo and C. T. Zhang, *Int. J. Biochem. Cell Biol.*, 2004, 36, 535-544.
32. X. Wu, H. Liu, H. Liu, J. Su, J. Lv, Y. Cui, F. Wang and Y. Zhang, *Gene*, 2013, 530, 8-18.
33. F. Sang, P. Jiang, W. Wang and Z. Lu, *Chin. Sci. Bull.*, 2010, 55, .3169-3173.
34. W. Gao and Z. H. Zhou, *Artif. Intell.*, 2013, 203, 1-18.
35. T. Zhang and Z. H. Zhou, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, 313-322, doi: 10.1145/2623330.2623710.
36. W. R. Qiu, X. Xiao and K. C. Chou, *Int. J. Mol. Sci.*, 2014, 15, 1746-1766.
37. L. Q. Li, S. J. Yu, W. D. Xiao, Y. S. Li, L. Huang, X. Q. Zheng, S. W. Zhou and H. Yang, *BMC bioinformatics.*, 2014, 15, 340.