Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2016, 12, 2893

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

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

Received 13th May 2016,
Accepted 1st July 2016

DOI: 10.1039/c6mb00374e

www.rsc.org/molecularbiosystems

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 to give the phase-specific PseDNC (psPseDNC). In this study, we used the prediction of recombination spots 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 compared to only using PseDNC with a 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%, and ~2.4% compared with three previous studies. Similarly, the accuracy was improved by 3.2% compared with our previous iRSpot-PseDNC web server through an independent data test. These results demonstrate that the joint use of PseDNC and *Z* curve enhances performance and can extract more information from a biological sequence. To facilitate research in this area, 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 issues 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 a change in genetic information, and short contiguous DNA fragments can also be 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 shows 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, regions with a high recombination rate are called hot spots. In contrast, regions with a low recombination rate are called cold spots. Investigations of recombination events and identification of 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, they merely used experimental and comparative methods, which 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 experiments. As an alternative way, many researchers have focused on developing new computational methods to

[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, Guiyang, China*
[e] *Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, China*

† Electronic supplementary information (ESI) available: S1: benchmark and independent dataset for *S. cerevisiae*; S2: data of recombination spots in other species. See DOI: 10.1039/c6mb00374e

‡ Co-first authors.

identify hot/cold spots.[6–8] Recently, Liu *et al.* introduced a gapped *k*-mer to extract features from a sequence.[9] They used it to identify recombination spots. Better performance can be obtained by using their method.

In 2013, we proposed a novel feature vector 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 the jackknife test showed a better classification performance than the 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 have been successfully used in the issues of predicting recombination spots,[10] nucleosome position,[13] splice sites,[14] translation initiation sites,[15] sigma-54 promoters,[12] methylation sites,[16] *N*-6-methyladenosine sites,[17] replication origins,[18] enhancers,[19] microRNA precursors[20] and so on. To facilitate research in this area, 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 the 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] As the *Z* curve variables contain many forms from single to multi-nucleotides, a great deal of information can be reflected by this. 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 whether we could improve the classifying accuracy through their joint use. We adopted two joint forms, one is using phase-specific pseudo dinucleotide composition (psPseDNC) and the other is to fuse the *Z* curve variables and pseudo dinucleotide composition directly. The recombination spot prediction issue is chosen as a case study to show the power of combining 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 beginning for predicting hot/cold spots in multiple species and we hope this united algorithm could be extended to other classification issues in DNA elements.

2. Materials and methods

2.1. The recombination spot 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 a benchmark data set. It contains 490 recombination hot spots and 591 cold spots. The trading-off parameters of LDM-based models were determined by 5-fold cross-validation. Gerton *et al.* even 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 dataset, and the remaining DNA sequences were sorted in descending 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 bases except 'A, T, G, C', so there were 575 genes in the final independent dataset. All of the sequences described above were downloaded from the *S. cerevisiae* genome database (http://www.yeastgenome.org/). Both the benchmark and independent datasets can be obtained from the ESI,† S1.

2.2. The recombination spot 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 frames). Firstly, the recombination rate in the above mentioned species was downloaded from the 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 rates 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 the sequence cannot 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 the recombination rate. We used mean D_i values, which were defined in a 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 the ESI,† 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 the margin distribution through maximizing the margin mean and minimizing the margin variance simultaneously.[35] Considering that this algorithm optimizes margin distribution, it is called large margin distribution machine (LDM), which is inspired by the above idea. The LDM optimizes the margin distribution through first and second-order statistics, so it may have the advantage of being more robust than classifiers only optimizing the margin. For example, the LDM is not very sensitive to the changing of LDM trading-off parameters. There are two solvers in the 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 the 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 from the LAMDA group website (http://lamda.nju.edu.cn/code_LDM.ashx). We used it to perform classification. Due to the reason that the number of sequences in our dataset is not too large, we use the dual coordinate descent method in the present work. There are four parameters (C, λ_1, λ_2, g) that need to be optimized. C is a penalty parameter for measuring the losses of instances. λ_1 and λ_2 are parameters for trading-off the margin variance. g is a parameter in the RBF (Radial Basis Function) kernel. In order to obtain the best performance, we used an 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. Consider the bases A, C, G and T occurring in an ORF or a fragment of the 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 eqn (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 eqn (2) is used to generate the phase-specific dinucleotide Z curve variables.

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

where X = A, C, G, T in the above equation. Both phase-specific single nucleotides and phase-specific dinucleotide 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.

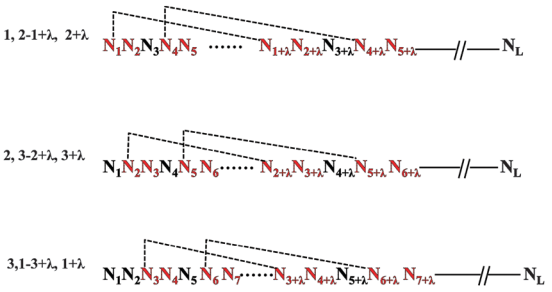

Fig. 1 A schematic illustration to show the correlations of dinucleotides located in different phases in a DNA sequence.

2.5. PseDNC and psPseDNC theory

Inspired by a similar idea of Z curve theory, we improved the PseDNC method to give phase-specific PseDNC (psPseDNC). psPseDNC can reflect composition bias among three codon positions. A schematic illustration of psPseDNC is shown in Fig. 1.

In Fig. 1, 1, 2, 3 on the left side represent the first, second, third phase in a DNA sequence, respectively. In the psPseDNC algorithm, the nucleotides in the 1-2, 2-3, 3-1 positions can interact with the dinucleotides located behind the $1 + \lambda - 2 + \lambda$, $2 + \lambda - 3 + \lambda$, $3 + \lambda - 1 + \lambda$ 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 eqn (3) adopts a similar form to PseDNC:

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

where $\theta_{\text{1-2},\lambda}$, $\theta_{\text{2-3},\lambda}$ and $\theta_{\text{3-1},\lambda}$ are phase-specific order-correlated factors and reflect the sequence-order correlation. For details of the 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 into $\theta_{\text{1-2},\lambda}$, $\theta_{\text{2-3},\lambda}$, $\theta_{\text{3-1},\lambda}$. There are three phases in a DNA sequence; therefore a DNA sequence is now represented by $(16 + \lambda) \times 3$ dimensional vectors.

2.6. Mixed variables

As mentioned above, PseDNC and the 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), there are a total of $16 + \lambda$, $16 + \lambda + 9 + 36 = 61 + \lambda$, and $(16 + \lambda) \times 3 = 48 + 3 \times \lambda$ variables for PseDNC,

PseDNC fused with Z curve and psPseDNC, respectively. All of the variables were scaled to [0, 1] using the following equation:

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

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

2.7. Cross-validation and the jackknife test

N-Fold cross-validation technology, the bootstrap test, the independent dataset test and the jackknife test are often used to assess the performance of classification methods. *N*-Fold cross-validation refers to the fact that the datasets are randomly partitioned into N subsets, then the $N - 1$ subsets are used for the training model and the remaining one is used as the testing dataset. Every random subset was used as the testing dataset in turn among the N folds, therefore the program was performed N times in the N-fold cross-validation process. Every sample is used for testing and others are used for building a model if N is equal to the number of samples. This is also called the jackknife test. In this work, the trading-off parameters of the LDM were obtained *via* 5-fold cross-validation. Due to the uniqueness of the jackknife test and the independent dataset test, there is no evaluation bias using the two methods. Herein, we used 5-fold cross-validation, the jackknife test and the 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 performed 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:

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

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 correctly predicted percentage of positive and negative samples. MCC has a range from -1 to 1. For the range of MCC from 0 to 1, it means prediction results are better than random prediction, otherwise they are worse than random prediction ($-1 < \mathrm{MCC} < 0$). 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 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 eqn (5) we know that recall and Sn have the same mathematical style; therefore we only listed Sn in tables among the two evaluators.

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 combined with the LDM. In order to evaluate the performance of our method, we performed ten times 5-fold cross-validations. The mean values of Sn, Sp, Acc, MCC, precision, AUC are summarized in Table 1.

Not only for SVM-based models but also for LDM-based models, psPseDNC and PseDNC fused with Z curve always showed better performance compared with PseDNC. For LDM-based models, the MCC of using variables PseDNC fused with the Z curve, psPseDNC are improved by 5.4% and 3.8% compared to 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% compared to only using PseDNC, respectively. In addition, PseDNC fused with the Z curve, psPseDNC, can obtain a higher AUC score compared to merely using PseDNC. The improved results hold both in the LDM and SVM based models, illustrating that better results could be obtained after adding the Z curve variables or using its phase-specific idea. Therefore, we can conclude safely that the 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 PseDNC fused with the

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

Machine learning methods	Methods	Sn (%)	Sp (%)	Acc (%)	MCC	Precision (%)	AUC
SVM-model	PseDNC	68.98	91.29	81.17	0.6249	86.78	0.8720
	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
LDM-model	PseDNC	70.37	90.78	81.53	0.6307	86.36	0.8752
	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 $(C, \lambda_1, \lambda_2) = 3, 2^5, 2^8$ and $g = 0.4$ for PseDNC; $(C, \lambda_1, \lambda_2) = 2, 2^7, 2^7$, and $g = 0.8$ for PseDNC + *Z*; $(C, \lambda_1, \lambda_2) = 3, 2^1, 2^8$, and $g = 0.6$ for psPseDNC. In SVM-based models $C = 8$, $\gamma = 0.125$ for PseDNC; $C = 2$, $\gamma = 2$ for PseDNC + *Z*; $C = 2$, $\gamma = 0.5$ for psPseDNC.

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

Machine learning methods	Method	Sn (%)	Sp (%)	Acc (%)	MCC	Precision (%)	AUC
SVM-model	PseDNC (Chen *et al.*[10])	73.06	89.49	82.04	0.6380	—	—
	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
LDM-model	PseDNC	71.02	90.86	81.87	0.6374	86.57	0.8780
	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 $(C, \lambda_1, \lambda_2) = (3, 2^5, 2^8)$, and $g = 0.4$ for PseDNC; $(C, \lambda_1, \lambda_2) = (2, 2^7, 2^7)$, and $g = 0.8$ for PseDNC + *Z*; $(C, \lambda_1, \lambda_2) = (3, 2^1, 2^8)$, and $g = 0.6$ for psPseDNC. In SVM-based models $C = 2$, $\gamma = 2$ for PseDNC + *Z*; $C = 2$, $\gamma = 0.5$ for psPseDNC.

Z curve shows the best performance for the LDM-based models, *i.e.*, an MCC of 0.6846 with an accuracy of 84.36% was obtained using our method. The Acc and MCC in our best result were improved by 4.06% and 8.1%, respectively than a previous study.[8]

In order to carry out an objective evaluation for our method, we did a rigorous jackknife test on our dataset using the parameters determined by the 5-fold cross-validation test. The result details of the 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]

As shown in the table, the PseDNC fused with the *Z* curve based on the LDM shows the best performance with an MCC value of 0.7037 and an accuracy of 85.29%. Meanwhile, PseDNC fused with the *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 Fig. 2 can further demonstrate the better performance of our new methods. Compared with another two available studies MCC and Acc were improved as well.[36,37]

Our independent dataset contains 287 positive samples and 288 negative samples. In the saved file all the 287 positive samples are listed ahead. To construct the unbalanced test set, we submitted an additional 100 samples to HcsPredictor and the iRSpot-PseDNC web server every time according to the storing order in the independent dataset. For example the first time we submitted 100 sequences and they are all positive samples; the second time we submitted 200 sequences and they are all positive samples too; 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

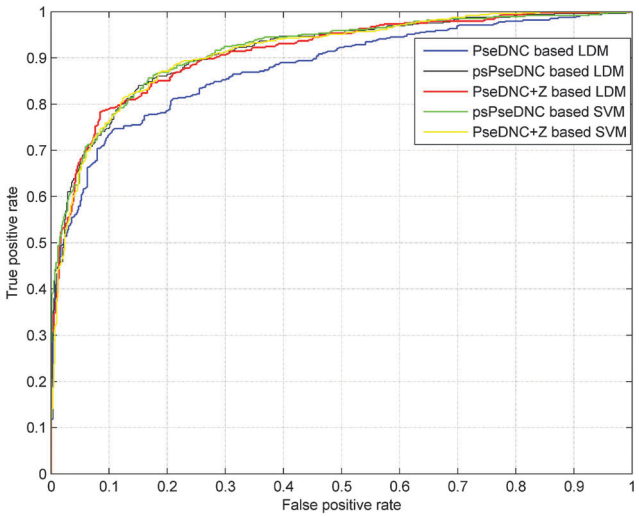

Fig. 2 ROC curves of different methods for identifying recombination hot/cold spots.

100 firstly submitted positive samples, while iRSpot-PseDNC gave 61%. Our new method showed an accuracy of 67% when we submitted 200 positive samples, while iRSpot-PseDNC gave 59.5% accuracy. Our method achieved accuracies of 66.3% and 71.5%, respectively when we submitted 300 and 400 samples with an unbalanced number, whereas iRSpot-PseDNC gave 59.3% and 66.25%, respectively. After submitting all of them into the two web servers, we obtained an accuracy of 76.52% on the independent dataset, and the Acc was improved by 3.2% compared with iRSpot-PseDNC. These results further illustrated that PseDNC fused with the *Z* curve shows better classification

performance than PseDNC as we expect. Improved results may give credit to the following two points. Firstly, the LDM optimizes the margin distribution, but the SVM merely optimizes the single margin. Secondly, we combined PseDNC and the *Z* curve variables as the input vectors. Since the *Z* curve and PseDNC are two different algorithms, they can reflect different information in a DNA sequence. The *Z* curve can transform a natural sequence into three groups of variables according to the codon positions. And the three group features from the *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 the *Z* curve and PseDNC, more information can be reflected. If we adopt feature elimination 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 as a case study. Furthermore, another attempt was to introduce the LDM into the bioinformatics field. For the same feature vectors, MCC can be improved by 1–2% comparing LDM with SVM-based models on this issue. Because of its better performance than the SVM, the 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 the *Z* curve variables based on the 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 the jackknife test on those species.

Comparing Tables 1 and 2, it is obvious that the result is still better 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 mechanisms and signals for recombination events. 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 of 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 the prokaryotes, it can be concluded that the recombination mechanism/signal, or at least the prediction model, is related to the phylogenetic distance. Therefore, each genome may need a specific model to predict recombination spots accurately. Aiming at 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 the Z curve. HcsPredictor is freely available from http://cefg.cn/HcsPredictor/.

Table 3 Predictive results for recombination hot/cold spots using the 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*.

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

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

HcsPredictor can not only predict recombination spots in the genome of *S. cerevisiae*, but also in *H. sapiens*, *M. musculus*, and *E. coli*. Because our models were trained using gene and open reading frames (ORFs) merely, the sequences under prediction should be ORFs. Based on the discussion in the above section, the performance of hot/cold classification was influenced by the phylogenetic distance. Therefore a model of the species, which has the closest evolutionary distance with the submitting sequences should be selected before using this web server to perform prediction. Two ways are provided to submit a 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 using the fasta format and must provide their email address simultaneously. In this way, the results will be returned to the mailbox after the web server completes the prediction. This could be a start of predicting recombination spots in multiple species and we hope

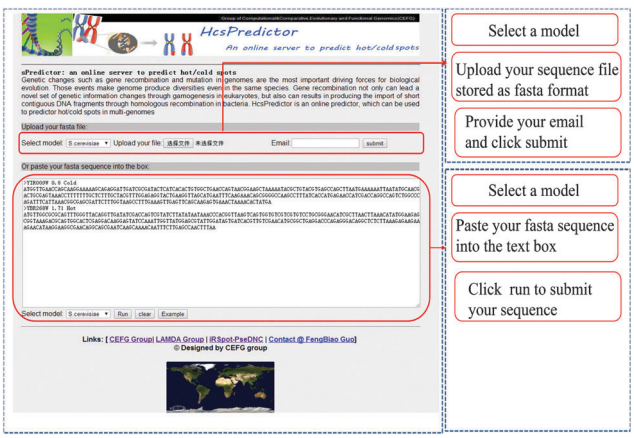

Fig. 3 A screenshot of the HcsPredictor web server.

it could lead to more novel computational models and feature selection techniques on this issue.

4. Conclusions

As a case study we used psPseDNC and PseDNC fused with the *Z* curve to predict recombination spots. We obtained much better performance than PseDNC (see Tables 1 and 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. Similarly, the accuracy was improved by 3.2% compared with our previous iRSpot-PseDNC web server through an independent data test. And also the results from cross species prediction demonstrated that species from different domains may adopt different mechanisms and signals for recombination events.

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 the receiver operating characteristic curve

Conflict of interest statement

The authors declare that they have no conflict of interest.

Author contributions

F. B. Guo conceived, designed, and coordinated the study. C. Dong analyzed the data. C. Dong, Y. Z. Yuan, F. Z. Zhang, H. L. Hua, and Y. N. Ye constructed the datasets. Y. Z. Yuan and F. Z. Zhang double checked the results. C. Dong, Y. Z. Yuan, A. A. Labena and F. B. Guo wrote the manuscript. W. Chen, H. Lin gave us much advice about this work. All of the authors read and approved this 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 the LDM algorithm, and 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]; the 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. U. S. A.*, 2000, **97**, 11383–11390.
6 T. Zhou, J. Weng, X. Sun and Z. Lu, *BMC Bioinf.*, 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**, 623149.
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, *Chemom. Intell. 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., Nonlinear, Soft Matter Phys.*, 2003, **67**, 061916.
29 J. Y. Yang, Y. Zhou, Z. G. Yu, V. Anh and L. Q. Zhou, *BMC Bioinf.*, 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 Bioinf.*, 2014, **15**, 340.

