Recombination spots prediction using DNA physical properties in the saccharomyces cerevisiae genome

Shou-Hui Guo, Li-Qin Xu, Wei Chen, Guo-Qing Liu, and Hao Lin

Citation: AIP Conference Proceedings **1479**, 1556 (2012); doi: 10.1063/1.4756460
View online: http://dx.doi.org/10.1063/1.4756460
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1479?ver=pdfcov
Published by the AIP Publishing

Recombination Spots Prediction Using DNA Physical Properties in the Saccharomyces Cerevisiae Genome

Shou-Hui Guo[a], Li-Qin Xu[a], Wei Chen[b*], Guo-Qing Liu[c] and Hao Lin[a*]

[a]*Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[b]*Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China*
[c]*School of Mathematics, Physics and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China*

Abstract. The prediction of meiotic recombination is difficult and current available methods are limited. In this study, we propose a novel method for discriminating between recombination hotspots and coldspots using support vector machine(SVM) with the DNA physical properties. Results of optimized pseudo-tetranucleotide show overall accuracy of 83.1% by using 5-fold cross-validation. High predictive successful rate exhibit that this model can be applied for discriminating between recombination hotspots and coldspots.

Keywords: Meiotic recombination; Support vector machine; Pseudo-tetranucleotide
PACS: 87.14.gk; 87.14.gf; 87.18.wd;

INTRODUCTION

Meiotic recombination is an important biological process. As a main driving force of evolution, recombination provides natural new combinations of genetic variations[1]. Recombination does not occur randomly along the genome. In a genome, regions that exhibit elevated rates of recombination relative to a neutral expectation are called recombination hotspots, while regions with low rates of recombination are called recombination coldspots.

Several global mapping studies have been performed to map DSB sites on chromosomes in yeast to determine whether they share common DNA sequences and/or structural elements[3-5]. They found that meiotic recombination events concentrate in 1–2.5 kilobase regions and do not share a consensus sequence. Thus, it is very difficult to predict hotspots and coldspots only by DNA sequence information[2]. Since experimental techniques are laborious and time-consuming, it is necessary to develop novel efficient and reliable computational approaches to predict recombination hot/cold spots.

In this study, we propose a model to predict hotspots and coldspots by using support vector machine (SVM) based on pseudo-tetranuclceotide character and physical structure character. The results demonstrate that the method can distinguish between hotspots and coldspots with high accuracy.

MATERIALS AND METHODS

Data sets

According to the method that Gerton reported in his paper, the data of recombination rates that they generated were estimated using the frequency of DSBs formation[2]. In another study[6], Liu et al. (2011) used the increment of diversity combined with quadratic discriminant analysis (IDQD) method to predict hot/cold spots based on the dataset including 490 hotspots and 591 coldspots records. In order to make a comparison with their result, our model was also trained and tested in the same dataset.

* Corresponding Author: hlin@uestc.edu.cn (H. L.), chenwei_imu@yahoo.com.cn (W. C.)

Numerical Analysis and Applied Mathematics ICNAAM 2012
AIP Conf. Proc. 1479, 1556-1559 (2012); doi: 10.1063/1.4756460

Pseudo-nucleotide composition

Nucleotide composition can represent DNA sequence effectively. In this paper, we take tetranuclceotide as characteristic parameters of DNA sequences. Though the multi-nucleotide can represent the information coded in DNA sequences, it can not reflect the difference of physical structure of DNA sequences. Therefore, a series of physicochemical properties have been introduced[7]. Based on the physicochemical properties parameter, the pseudo-nucleotide characters appeared. Based on Zhou's method[8] and pseudo-amino acid composition[9], we define the pseudo-tetranucleotide composition character.

Given a DNA sequence P that is composed by L nucleotides:

$P = R_1 R_2 R_3 R_4 R_5 R_6 R_7 ... R_L$

where R_1 represents the first nucleotide of the sequence, R_2 represents the second nucleotide, …., R_L represents the L-th nucleotide, and they each belong to one of the four nucleotide (A, C, G, and T). Thus a DNA fragment can be represented by the pseudo- tetranucleotide composition, a vector of 256+λ dimensions as following:

$$X = [P_1, ...P_{256}, P_{256+1}, \cdots, P_{256+\lambda}]^T \qquad (\lambda < L) \tag{1}$$

The first 256 components reflect the effect of the tetranucleotide composition, while the components from 256+1 to 256+λ reflect the effect of DNA sequence physical and chemical properties.

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{256} f_i + \omega \sum_{k=1}^{\lambda} t_k}, & 1 \le u \le 256 \\ \dfrac{\omega t_{u-256}}{\sum_{i=1}^{256} f_i + \omega \sum_{k=1}^{\lambda} t_k}, & 256+1 \le u \le 256+\lambda \end{cases} \tag{2}$$

where f_i represents the occurrence frequency of ith tetranucleotide, t_k represents the k-tier sequence correlation factor computed according to the following Eqs (3-4), ω represents the weight factor for the DNA sequence order effect.

$$t_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k} \qquad (k < L) \tag{3}$$

$$J_{i,i+k} = [H(R_{i+k}) - H(R_i)]^2 \tag{4}$$

In Eq.(4), t_1 reflects the first tier correlation factor, it reflects the most contiguous bases' order correlation along DNA sequence, while t_2 reflects the second most contiguous bases' order correlation, and so forth. $H(R_i)$ reflects the physical and chemical property of the ith base, it was calculated as follows:

$$H(R_i) = \frac{H^0(R_i) - ave(H^0)}{SD(H^0)} \tag{5}$$

where *ave* represents the average value of the physical and chemical property values of the bases, *SD* reflects standard deviation, $H^0(R_i)$ reflects the salvation free energy of the ith base[7]. We use the 5-fold cross validation to optimize the parameters ω and λ.

Sequence physical structure

Six physical structure parameters based on molecular dynamics was introduced, they are double helix structure parameters of base pair called Twist, Tilt, Roll, Shift, Slide and Rise, respectively[10]. They may influence the combination of the protein of a certain DNA sequence, and then affect the DNA recombination.

Thus, we take the physical structure properties as another character to classify recombination coldspots/hotspots. According to the Eqs.(3)-(6), for each of the six structural properties, we can calculate the m-tier structural correlation factor t'_m according to the follows:

$$t'_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k} \qquad (k < L) \tag{6}$$

$$J'_{i,i+k} = [P(D_{i+k}) - P(D_i)]^2 \tag{7}$$

$$P(D_i) = \frac{P^0(D_i) - ave(p^0)}{SD(P^0)} \quad (8)$$

where, D_1 represents the first dinucleotide of the sequence, D_2 represents the second dinucleotide, ..., D_{L-1} represents the (L-1)-th dinucleotide. m denotes structural correlation length. We use the six physical structure parameters to calculate $P(D_i)$, thus, we finally obtain $6m$ characteristics. The 5-fold cross validation is used to optimize the parameters m.

Support vector machine

Support vector machine (SVM) is a supervised learning algorithm for classification based on linear decision rules[11].The main idea of SVM theory is to transform the original data to a higher dimensional space by using a kernel function. Then SVM will find a hyperplane to classify the training sample into different classes. In this paper, we use libsvm 2.86 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) to train and test sample datasets.

Parameter optimization and feature selection

The feature selection is an important step for eliminating the redundant features and improving the efficiency and performance of prediction. Here we use the fselect.py, a libsvm tool, to optimize features. The basic idea of this algorithm is to rank each feature according to a score as described in the literature[12]. The ranked feature with a higher score indicates that it is a more highly relevant one for prediction. Based on the score of features, we add features into feature set one by one from higher to lower rank. Then we can obtain a series of feature sets. Subsequently, we perform the 5-fold cross validation to examine the overall accuracies of feature sets and achieve the best feature set with the highest accuracy.

RESULTS AND DISCUSSION

When we use pseudo-tetranucleotide character to build our prediction model, λ ranged from 1 to 10 by the step of 1, ω ranged from 0.1 to 1 by the step of 0.1. 5-fold cross-validated result is shown in Table 1. The best result was obtained when λ=4, ω=0.2. In other words, when it contains 260(256+4) features in this model, it can achieve the best accuracy.

TABLE (1). Prediction accuracies of coldspots/hotspots by using different parameters *

Method (number of feature)	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
Pseudo-tetranucleotide(260)	70.2	92.2	82.2	0.647
Physical structure(282)	73.5	89.3	82.1	0.641
pseudo-tetranucleotide and Physical structure (542)	73.9	89.3	82.3	0.644
Pseudo-tetranucleotide(feature selection) (47)	71.8	92.4	83.1	0.663
Physical structure(feature selection) (267)	73.3	90.0	82.4	0.647
Pseudo-tetranucleotide and Physical structure(feature selection) (408)	74.3	90.0	82.9	0.656
IDQD [6]	79.4	81.0	80.3	0.603

**Sn*, sensitivity; *Sp*, specificity; *Acc*, accuracy; *MCC*, Matthew's correlation coefficient.

$$Sn = \frac{TP}{TP + FN},\ Sp = \frac{TN}{TN + FP},\ ACC = \frac{TP + TN}{TP + FN + TN + FP},$$

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

TP, true positive; *FN*, false negative; *TN*, true negative; *FP*, false positive.

For physical structure character, we choose the parameter *m* from 1 to 70 by the step of 1. Finally, we obtain the best predict result (82.1%) when *m*=47, that is to say, there are 282(47×6) characteristics to represent the DNA sequences. When combining the 260 characteristics and 282 characteristics together, *Acc* improved a little bit (82.3%) as is shown in Table 1.

In order to further improve prediction accuracy, we use fselect.py to optimize pseudo-tetranucleotide features, physical structure features and the combination of pseudo-tetranucleotide and physical structure features, respectively. Obviously, the *Acc* get a little improved after feature selection for each of the three kinds of parameters (from 82.2% to 83.1%, from 82.1% to 82.4%, from 82.3% to 82.9%). We also list Liu's results[6] in the last line in Table 1. As we can see from Table 1, our method outperforms the IDQD method in terms of *Sp, Acc* and *MCC* for recombination spots prediction.

CONCLUSION

In this paper, we propose a novel method based on SVM classifier using DNA physical properties to predict meiotic recombination hotspots and coldspots. Our method achieves high accuracy in classifying hot/cold spots, outperforms the IDQD method [6], suggesting that feature selection is an effective way to improve prediction accuracy and optimize prediction model. High accuracies demonstrate that the DNA physical properties are important for meiotic recombination.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (11047180, 61102162), the Scientific Research Foundation of Sichuan Province (2009JY0013) and the Fundamental Research Funds for the Central Universities (ZYGX2009J081).

REFERENCES

1. J. Zheng et al., *Genome Biol.* **11**, R103 (2010).
2. J. L. Gerton et al., *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11383-11390 (2000).
3. F. Baudat and A. Nicolas, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5213-5218 (1997).
4. S. Klein et al., *Chromosoma* **105**, 276-84 (1996).
5. D. Zenvirth et al., *EMBO J.* **11**, 3441-3447 (1992).
6. G. Liu et al., *J. Theor. Biol.* **293**, 49-54 (2012).
7. M. Monajjemi et al., *Biochemistry* (Moscow) **71**, S1-S8 (2006).
8. X. Zhou et al., *Talanta* **85**, 1143-1147 (2011).
9. K. C. Chou, *Proteins* **43**, 246–255 (2001).
10. Y. C. Zuo and Q. Z. Li, *Physica A* **389**, 4217-4223 (2010).
11. V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 375-440 (1998).
12. I. Guyon et al., *Feature Extraction: Foundations and Applications*. Berlin Heidelberg: Springer, 2006, pp. 315-323.