Journal of Southeast University (English Edition) Vol. 22, No. 1, pp. 112 – 116 Mar. 2006 ISSN 1003—7985

Support vector machine for prediction of meiotic recombination hotspots and coldspots in *Saccharomyces cerevisiae*

Weng Jianhong Zhou Tong Sun Xiao Lu Zuhong

(State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China)

Abstract: A novel method for predicting hotspots and coldspots using support vector machine (SVM) based on statistical learning theory is developed. This method is applied to published 303 hot and 48 cold open reading frames (ORFs) in *Saccharomyces cerevisiae*. The sequence features of general dinucleotide abundance and dinucleotide abundance based on codon usage are extracted, and then the data sets are classified with different parameters and kernel functions combined with the method of two-fold cross validation. The result indicates that 87.47% accuracy can be reached when classifying hot and cold ORF sequences with the kernel of radial basis function combined with dinucleotide abundance based on codon usage.

Key words: meiotic recombination; hotspot; coldspot; dinucleotide abundance; support vector machine

Meiotic recombination is a fundamental biological feature about which we still know remarkably little. Recombination occurs more frequently in some regions of the eukaryotic genomes than in others, with variations of several orders of magnitude observed in frequencies of meiotic exchange per unit physical distance[1]. Hotspots are genomic regions with unusually high levels of meiotic recombination, and contrarily, coldspots are the regions with relatively low levels of meiotic recombination[2]. Although observations concerning individual hotspots and coldspots have given clues as to the mechanism of meiotic recombination initiation, our ability to predict hotspots and coldspots from DNA sequence information is very limited[2]. From studies in yeast, we know that genetically defined hotspots are associated with local double-strand DNA breaks (DSBs)[1,2]. 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[2-5]. Although experimental techniques can be applied for this purpose, they are laborious and time-consuming and, therefore, have become infeasible for large numbers of genomic sequences. Hence, efficient and reliable computational methods for discriminating hotspots from coldspots are required.

A suitable approach to this task employs statistical learning theory, such as the support vector machine (SVM), which is a type of supervised machine learning algorithm that can be integrated with prior knowledge based on investigation.

In this study, we develop a novel method for predicting hot and cold ORFs located in hotspots and coldspots by using dinucleotide abundance[6] combined with SVM to extract sequence features and to determine classification, respectively. Application of this method to published data sets demonstrates that the method can distinguish hot and cold ORFs with high accuracy.

1 Materials and Methods

1.1 Data set

Gerton et al. have estimated relative recombination rates for most of the *Saccharomyces cerevisiae* loci using DNA microarrays[2]. They detected 303 hot ORFs clustered into 177 hotspots whose recombination rates ranked in the top 12.5% and 49 cold ORFs clustered into 40 coldspots whose recombination rate ranked in the bottom 12.5%[2]. In this study, we extracted the 303 hot ORFs and 48 cold ORFs (one of the cold ORFs listed in Ref. [2] was not correct) from the GenBank database. So, the final data set for analysis comprised 351 sequences, which contained no ORFs shorter than 150 bp.

1.2 Dinucleotide abundance

Dinucleotide abundance was measured by frequencies of dinucleotides (FD). Each ORF was represented by a 16-dimensional vector with respect to the 16 dinucleotides. The FD value of the i-th dinucleotide was calculated by

$$f_{\mathrm{FD}}^{i} = \frac{o_i}{n_{\mathrm{T}}} \tag{1}$$

where o_i is the observed number of the i-th dinucleotide made up of two continuous nucleotides, and n_{T} is

Received 2005-06-16.

Biographies: Weng Jianhong (1981—), male, graduate; Lu Zuhong (corresponding author), male, doctor, professor, zhlu@seu.edu.cn.

the total number of dinucleotides in the ORF.

Considering the codon composition in the ORFs[7], we extended the definition of DA by considering the composition of dinucleotides in each codon position for the selected ORFs. The first two nucleotides in the codon could be defined as one kind of dinucleotide. At the same time, the last two nucleotides in the codon and the first and third nucleotides in the codon were also defined as two kinds of dinucleotides. The extended dinucleotide abundance (EDA) value of the *i*-th dinucleotide for the *j*-th definition of dinucleotides was calculated by

$$f_{\mathrm{EDA}}^{ij} = \frac{o_{ij}}{n_{\mathrm{T}}^{j}} \tag{2}$$

where *j* may be 0, 1, 2, 3; each indicates one kind of dinucleotide. 0 means general dinucleotide made up of two continuous nucleotides; 1 means the dinucleotide made up of the first two nucleotides in the codon; 2 means the dinucleotide made up of the last two nucleotides in the codon; 3 means the dinucleotide made up of the first and third nucleotides in the codon.

1.3 Support vector machine

Support vector machine is a relatively new type of supervised learning algorithm for two- or multi-class classification based on linear decision rules[8, 9]. SVM takes as input i. i. d. (independent and identically distributed) training samples $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i represents the sample attributes and $y_i \in \{-1, +1\}$.

SVM will then find a hyperplane separating the training instances by their classes and maximizing the distance from the closest examples to the hyperplane (maximum-margin hyperplane). The classification of a sample will be determined by the sign of the function:

$$f(x) = \langle w, x \rangle + b \tag{3}$$

where *w* and *b* are the parameters of the hyperplane; a point *x* is classified as positive (negative) if $f(x) > 0$ ($f(x) < 0$). The examples closest to the hyperplane are called support vectors and are crucial for training.

For many training sets it will not be possible to separate samples by a linear function in the original feature space, so training instances are mapped into a higher dimensional space by a function Φ. SVM will then find a linear maximum-margin hyperplane in this higher dimensional space. In order to solve this problem, it is not necessary to directly define the mapping into higher dimensional space, but it is sufficient to give the dot product of two instances in this space[9].

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{4}$$

Eq. (4) is called a kernel function. Commonly used kernel functions comprise linear, polynomial and radial basis functions.

In this paper, we used SVM$^{\mathrm{light}}$ version 6.01 (http://svmlight.joachims.org/) written by Joachims for data training and classifying. All the kernel parameters were kept constant except for regulatory parameters *B* and *C*. Different kernel functions were used in our experiments, including linear function, polynomial function and radial basis function. The best results were obtained by using the radial basis function kernel with $\gamma = 150$. The values of regulatory parameters *B*, *C* and *J* were optimized to 0 to 200, respectively.

As in other statistical learning studies, SVM prediction accuracy can be described by means of the classification accuracy, precision and recall.

$$a = \frac{n_{\mathrm{TP}} + n_{\mathrm{TN}}}{n_{\mathrm{TP}} + n_{\mathrm{TN}} + n_{\mathrm{FP}} + n_{\mathrm{FN}}} \tag{5}$$

$$p = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}}} \tag{6}$$

$$r = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FN}}} \tag{7}$$

where n_{TP}, n_{TN}, n_{FP} and n_{FN} represent true positive, true negative, false positive, and false negative numbers, respectively.

2 Results and Discussion

2.1 Prediction accuracy

The basis of our approach is to describe hot and cold ORFs as vectors in a multi-dimensional feature space. We used two kinds of mapping methods (FD and EDA) to extract features from ORF sequences. Then we subjected the feature vectors representing training sequences to a supervised machine learning algorithm SVM.

To estimate the performance of the complete procedure, two-fold cross-validation was used. Two-fold cross-validation consisted of splitting the data set of hot and cold ORFs randomly into two parts and then alternatively using one part for testing and the remainder for training. When we applied the FD, each ORF was described as a 16-dimensional vector. We could classify the test data set with 85.47% accuracy by using the kernel of radial basis function. If we combined these four kinds of dinucleotides together, each ORF was described as a 64-dimensional vector. We could classify the test data sets with 87.47% accuracy by using the kernel of radial basis function (see Tab. 1). The results indicate that the sequence features derived from EDA are better than from FD. The first column in Tab. 1 indicates the kernel function used in SVM.

Tab. 1 Prediction accuracy of SVM classification of hot and cold ORFs

Kernel	Feature mapping	n_{TP}	n_{TN}	n_{FP}	n_{FN}	p/%	r/%	a/%
Linear function	FD	287	8	40	16	87.77	94.72	84.05
	EDA	276	21	27	27	91.09	91.09	84.62
Polynomial function	FD	283	14	34	20	89.27	93.40	84.62
	EDA	269	28	20	34	93.08	88.78	84.62
Radial basis function	FD	265	35	13	38	95.32	87.46	85.47
	EDA	274	33	15	29	94.81	90.43	87.47

Considering the difference between the number of hotspots and the number of coldspots used in our experiment, we calculated the prediction accuracy in hotspots and coldspots, respectively. Concerning the EDA feature, the accuracies in hot and cold groups were 90.43% ($n_{TP}/(n_{TP}+n_{FN})$) and 68.75% ($n_{TN}/(n_{TN}+n_{FP})$), respectively, based on the radial basis kernel. According to Tab. 1, the prediction accuracy in the cold group is lower than in hot group because the number of cold ORFs (48) is much less than that of hot ORFs (303).

2.2 Effect of distribution of dinucleotides

In this study, we showed that this novel method could predict hot and cold ORFs with high accuracy, which suggested that frequencies of dinucleotides were sufficient to predict hot and cold ORFs. In order to evaluate the difference between the two classes of sequences, we calculated the distribution of nucleotides in hot and cold ORFs (see Fig. 1). Interestingly, from Fig. 1, we find that the distributions of nucleotides in hot and cold ORFs are different and the average abundance indicates that the frequencies of C and G in hot ORFs are higher than in cold ORFs. Fig. 2 shows the

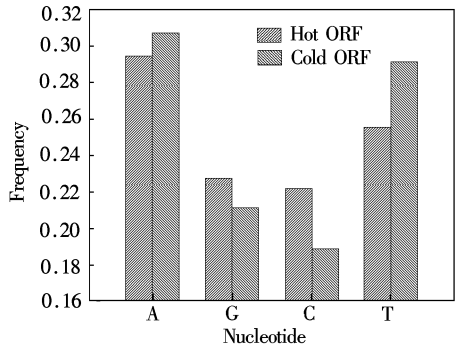

Fig. 1 Average abundance of nucleotides

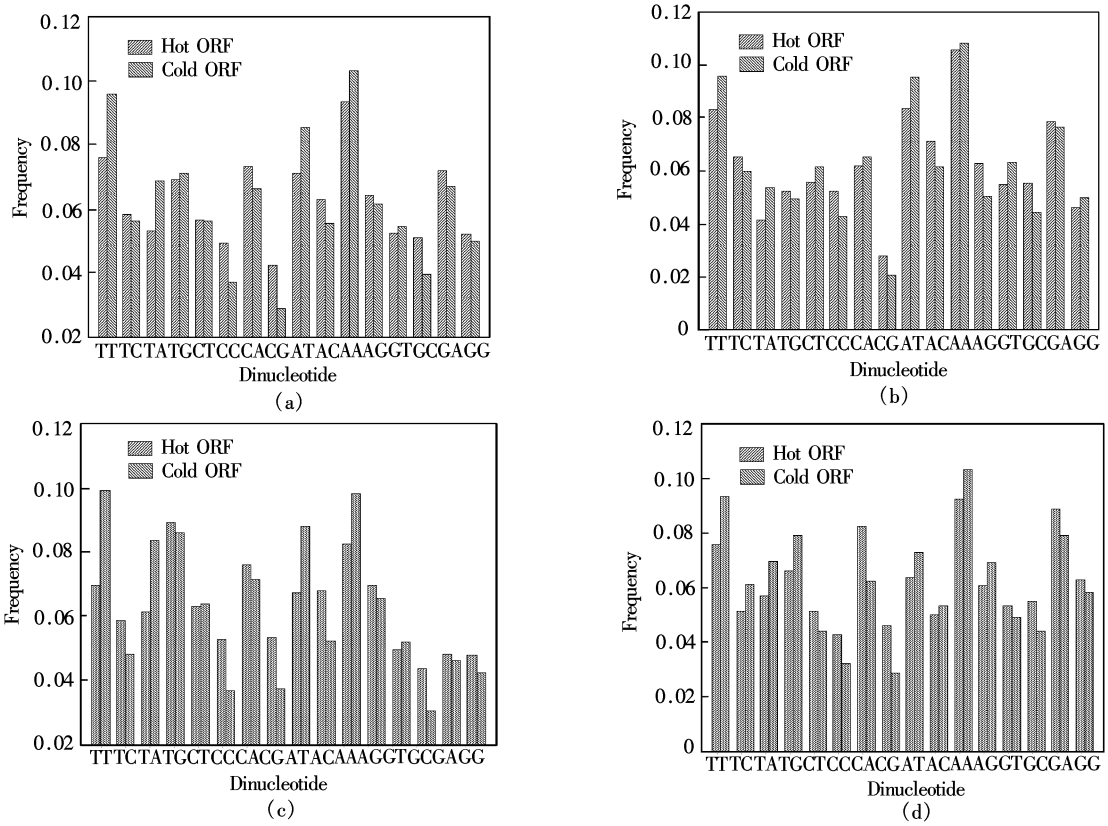

Fig. 2 Average abundance of the dinucleotides of type 0 to 3. (a) Type 0; (b) Type 1; (c) Type 2; (d) Type 3

distribution of four kinds of dinucleotides in 303 hot and 48 cold ORFs. From these figures, we can see that the average abundance of those dinucleotides including guanine or cytosine are higher in hot ORFs than in cold ORFs. It has been reported that the position of hotspots may be regulated by some features of chromosome structure related to GC-richness in *S. cerevisiae*[2-5,10]. This kind of positive correlation between G + C content and recombination rate was also observed in organisms, such as *Drosophila melanogaster*, mouse and human. In such eukaryotic organisms, the recombination machinery induces genetic conversion between parental chromosomes during meiosis. Experimental evidence in mammals suggests that genetic conversion associated with recombination favors the copy of the most GC-rich sequence over the other[11-15].

To assess the statistical differences between the hot and cold ORFs, the frequencies of dinucleotides in the hot and cold data sets were compared using an χ^2 test. The frequencies of 42 dinucleotides were found to be significantly different ($p < 0.05$) between the two groups (see Tab. 2), which supported the feasibility of our classification method again.

Tab. 2 P values generated by χ^2 test

Dinucleotides	P_0	P_1	P_2	P_3
TT	0	0.005	0	0
TC	0.313	0.114	0	0
TA	0	0	0	0
TG	0.296	0.302	0.341	0.002
CT	0.954	0.042	0.912	0.030
CC	0	0	0	0
CA	0.001	0.247	0.163	0
CG	0	0.004	0	0
AT	0	0.001	0	0.002
AC	0	0.002	0	0.181
AA	0.021	0.6	0.001	0.024
AG	0.242	0	0.253	0.005
GT	0.244	0.003	0.399	0.121
GC	0	0.001	0	0
GA	0.072	0.660	0.426	0.020
GG	0.338	0.188	0.037	0.136

Note: P_0, P_1, P_2, P_3 mean the p values in the four kinds of dinucleotieds, respectively.

In addition, Gerton et al. have found the correlations between the hot and cold ORFs and gene function. One interpretation of these correlations is that certain categories of genes are associated with a particular chromatin structure that is favorable (hotspots) or unfavorable (coldspots) for initiating meiotic recombination[2]. It has also been reported that there is a relationship between gene function and the codon usage pattern in eukaryotic organisms[16,17]. Our EDA attributes, to some extent, describe the codon bias of the ORFs in *S. cerevisiae*. Therefore, taken together, we conclude that with our representation in the SVM, the EDA attributes are deemed to perform accurate classifications well.

2.3 Potential improvements

The training data set is important to develop accurate SVM classification systems for hot and cold ORFs. At present, the only publicly available and validated 303 hot and 48 cold ORFs are used in this work. This data set may not be representative of all hotspots. Hence, further improvement in the prediction capability is expected if a more comprehensive training data is used. In this study, we extracted features just from the DNA sequences. Some experiments indicated that transcription elements were correlated with hotspots[1,2]. Hence, prediction of hot and cold ORFs by combining computer classification with additional information such as transcription factors and structure information is helpful in developing a better tool for predicting hot ORFs.

3 Conclusion

This study shows that the SVM classification system extracting features using DA from DNA sequences performs well, and the result from using EDA is better than that using FD. The study provides evidence that there is a kind of positive correlation between G + C content and recombination rate[2-5]. At the same time, as our EDA attributes describe the codon bias of the ORFs in *S. cerevisiae*, the high accuracy also suggests that there is a relationship between hot ORFs and the codon usage pattern in *S. cerevisiae*[16,17]. This method extracts features only from DNA sequences, so it can also be easily extended to other eukaryotic genomes.

References

[1] Lichten M, Goldman A S. Meiotic recombination hotspots [J]. *Annu Rev Genet*, 1995, **29**(1): 423 - 444.

[2] Gerton J L, DeRisi J, Shroff R, et al. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae* [J]. *Proc Natl Acad Sci USA*, 2000, **97**(21): 11383 - 11390.

[3] Baudat F, Nicolas A. Clustering of meiotic double-strand breaks on yeast chromosome Ⅲ [J]. *Proc Natl Acad Sci USA*, 1997, **94**(10): 5213 - 5218.

[4] Klein S, Zenvirth D, Dror V, et al. Patterns of meiotic double-strand breakage on native and artificial yeast chro-

mosomes [J]. *Chromosoma*, 1996, **105**(5): 276 - 284.
[5] Zenvirth D, Arbel T, Sherman A, et al. Multiple sites for double-strand breaks in whole meiotic chromosomes of *Saccharomyces cerevisiae* [J]. *EMBO J*, 1992, **11**(9): 3441 - 3447.
[6] Karlin S, Cardon L R. Computational DNA sequence analysis [J]. *Annu Rev Microbiol*, 1994, **44**(1): 619 - 654.
[7] Lin K, Kuang Y, Joseph, J S, et al. Conserved codon composition of ribosomal protein coding genes in *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics [J]. *Nucleic Acids Res*, 2002, **30**(11): 2599 - 2607.
[8] Vapnik V N. *Statistical learning theory* [M]. New York: Wiley, 1998: 375 - 440.
[9] Friedel C C, Jahn K H, Sommer S, et al. Support vector machines for separation of mixed plant-pathogen EST collections based on codon usage [J]. *Bioinformatics*, 2005, **21**(8): 1383 - 1388.
[10] Kliman R M, Irving N, Santiago M. Selection conflicts, gene expression, and codon usage trends in yeast [J]. *J Mol Evol*, 2003, **57**(1): 98 - 109.
[11] Kliman R M, Hey J. Reduced natural selection associated with low recombination in *Drosophila melanogaster* [J]. *Mol Biol Evol*, 1993, **10**(6): 1239 - 1258.
[12] Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes [J]. *Proc Natl Acad Sci USA*, 2001, **98**(10): 5688 - 5692.
[13] Marais G, Piganeau G. Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genome [J]. *Mol Biol Evol*, 2002, **19**(9): 1399 - 1406.
[14] Perry J, Ashworth A. Evolutionary rate of a gene affected by chromosomal position [J]. *Curr Biol*, 1999, **9**(17): 987 - 989.
[15] Fullerton S M, Bernardo Carvalho A, Clark A G. Local rates of recombination are positively correlated with GC content in the human genome [J]. *Mol Biol Evol*, 2001, **18**(6): 1139 - 1142.
[16] Ma J M, Zhou T, Gu W J, et al. Cluster analysis of the codon use frequency of MHC genes from different species [J]. *Biosystems*, 2002, **65**(2, 3): 199 - 207. (in Chinese)
[17] Richard J E, Lin K, Tan T. A functional significance for codon third bases [J]. *Gene*, 2000, **245**(2): 291 - 298.

基于支持向量机的酵母重组热点和冷点的预测

翁建洪　周　童　孙　啸　陆祖宏

(东南大学生物电子学国家重点实验室,南京 210096)

摘要:使用基于统计学习理论的支持向量机(SVM)方法,提出了针对重组热点和冷点分类预测的新方法. 对酵母基因组的 303 个重组热点开放阅读框(hot ORF)以及 48 个重组冷点开放阅读框(cold ORF),提取了序列的一般二联碱基丰度特征,以及基于密码子使用偏性的二联碱基丰度特征,然后使用二倍交叉验证方法,选择不同的核函数和对应参数,对数据集进行了训练和分类预测. 研究结果表明,当使用径向基核函数,并采用基于密码子使用偏性的二联碱基丰度特征时,预测准确率为 87.47%.

关键词:减数分裂重组;热点;冷点;二联碱基丰度;支持向量机

中图分类号:Q617