Acta Technica *62* (2017), No. 5A, 17–22

The method for selecting genomic prime pairs in tissue culture based on SVM

Tang Shou Guo[1], Li Yong[1], Zhang Zhi Kun[1]

Abstract. In molecular and biochemical markers technology, genomic primer pairs were used to detect variants in plant tissue culture by polymerase chain reaction. This paper presents a method for primers selection problem using Support Vector Machine (SVM). The gene sequences of plant are taken as two parts, one is a set of obtained genomic primer pairs from experiments, another is a set of sample genes. The Euclidean distance between obtained primers are used as training SVM classifier for filtering the all samples genes, and then a candidate subset is generated. The proposed method was tested by experiment in selecting the primer pairs of tissue cultured seedlings of kiwifruit genome. The result demonstrated the proposed method is reasonable.

Key words. Tissue culture, primer selection, euclidean distance, support vector machine1..

1. Introduction

In plant science research, tissue culture is extensively employed in the production, conservation and improvement of plant resources. Plants regenerated from tissue culture possess an array of genetic and epigenetic changes. The specific genetic changes are usually analyzed by molecular marker technology with polymerase chain reaction (PCR), such as the restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP) and so on[1][2]. PCR is highly dependent on primers or primer pairs that are small arbitrary ologo-nucleotides used to amplify a set of DNA fragments. Based on amplification products with high clear bands are selected, the number of polymorphism bands is counted for detecting whether the tissue cultured seedling mutated. Thus, the primers selection is a critical step in tissue culture. But the selected primers used in experiments differ with each other greatly in experience of researchers what could be called the primer selection bias. The biased primers produce a highly variable number of other primers by PCR. In information technol-

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650550, China

ogy, selecting a subset of relevant features to be used in model construction is called Feature Selection. It is obviously that primers selection is a typical usage of Feature Selection. A method using SVM[3-5] for selecting primer pairs in tissue culture is discussed for resolving the problem of primer selection bias in the study.

2. Implementation of the proposed model

2.1. Data preparation

In cluster analysis of gene expressions, the data of gene expression needed to be pre-treated for making sure the comparability among gene expressions. Data pre-treatment contains two types, one is standardizing original data, another is calculating logarithms of ratios of gene expressions. The regular cluster analysis algorithm is based on the similarity of individuals to evaluate the similarity of two expression patterns, such as Pearson correlation coefficients, Euclidean distance measure, Covariance analysis, Cosine correlation coefficients, and so on. For different cluster analysis methods, data transform and similarity could also be different. It is a critical point to select the appropriate data pre-treated method and similarity for obtaining more accurate clustering results[6]. Based on the scoring function, the genes with high similarity could be classified as a same type. Calculating similarity or divergence between individuals is called similarity measure. The definition of similarity is the key to data analysis, which can greatly affect outputs of clustering algorithms. In practical calculations, the similarity measure is transformed into calculating the distance between the two groups of data. The smaller the distance, the more similar the expression pattern; on the contrary, the expression pattern is more different. Supporting two microarray data are $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$, and the distance function $d(X, Y)$, they must meet the following conditions:

$$\begin{cases} d(X,Y) \geq 0 \\ d(X,Y) = d(Y,X) \\ d(X,Y) = 0,\ if\ X = Y \\ d(X,Y) = d(X,Z) + d(Z,Y) \end{cases} \quad (1)$$

Euclidean distance is used to calculate the similarity between genes in this study, which is defined as $d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (2)

Taking Euclidean distance as the similarity measure function, it is relative largely impacted by the amplitude changes of gene expression profiles, therefore, it should be standardized. In biological sciences, the DNA four bases, adenine, guanine, cytosine, and thymine are denoted with A, G, C, and T respectively. And the gene expression profiles of primers are normally encoded with the 4 bases A, G, T, and C, i.e. the 4 nucleotide combinations resulting in the formation of DNA molecules. Mathematically, it is represented with a character set containing four characters $\sum = (A, G, C, T)$. By using the four characters, the initial generation group is represented with a N number of arbitrary DNA chain of individuals. And the four characters are needed to be coded for computer processing. This study uses a binary encoding, that is, 00,01,10 and 11 represent A, T, G, and C respectively.

2.2. The method of SVM

The basic idea of SVM classification can be summarized as: firstly, the input sample space is mapped to a feature space through a linear decision boundary or a nonlinear decision boundary, such as quadratic, polynomial and radial basis function. And then, an optimal hyperplane is constructed in the feature space to make sure two types of samples (which can be extended to multi-types of samples) can be segmented in the feature space. The feature mapping is only related to low dimensional input vector and the vector dot product in the feature space, where the dot product could be replaced of a kernel function. It could avoid the curse of dimensionality and solve the high-dimensional eigenvalue problem. The discriminant function of support vector machine is: $f(x) = sgn[\sum_{i=1}^{n} a_i^* y_i K(x_i, x) + b^*]$ (3)

In Eq. (3), $K(x_i, x)$ is called kernel function, and the selection of kernel function should make it be a vector dot product in feature space, that is, there exists function $\phi(x)$ and $\phi(x_i) \bullet \phi(x) = K(x_i, x)$ .. The selection of an appropriate kernel function is important, since the kernel function defines the feature space in which the training set examples will be classified. SVM classifier is adopted in this work due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and flexibility in modeling diverse sources of data. By using SVM as a classifier, a method for identifying primer pairs is proposed in the study. Firstly, a prime pairs set is settled with partly observed primer pairs from experiments, and the other genes data of plants is taken as sample data set. The Euclidean distance among the data in the primer pair set is calculated as reference. Secondly, the Euclidean distance among the sample data set and the primer pair set is calculated. The sample data far from the primer pairs is collected in a non-primer pair set. Based on the primer pairs set and non-primer pairs set, a model is trained from the two sets with SVM, and used for sample data set to learn. At last, the feature genes are learned, and the desired primer pairs are obtained with further optimization.

3. Experimental results

3.1. Data source

The selection of genomic primer pairs of kiwifruit is used as an example in the study with java programming. The data source is from the Key Laboratory of Southwest Forestry University and related research results and divided into two parts. The first part is the partial original genes of kiwifruit, which called sample set, and 1677 groups of data are obtained according to six bases as one group:

TTGCCATTAAGGACTCCTCTAGCTAATGAGTATCAGTTGGGCCTCAAG
ATGGGGGCTGTATGAGCTTGTGACTAGATGCCCGAGGAACTTGGGCCG
CTGGATGGACCTTCCGTAATCCGCGTGTTTGGCAAAGGCTAAGGAGAC
GTGAGTTTGGCTGGAGGATATGGGAAGTTAGGACACAAGCACTTCGGG
GAAAACGAAGTGGTCTGACTTATTCCTTGATTTTCTTTGACTTACATT
ATGATCTTCTAACTTGAGGCGGAGCATGGCGAGGATGTAGGTCGACCT
TGTTCCTACGGTGTACATCGACTTGGTCATCTGGACCATTCATTTTGA

TTAGAGTATGGAGCGCCTTCTGGACCATTCATTTAAAAAAATAAGTGA
CCGTCCTAAACTTGGCATGACCGTTGGGCATGCATTTGCACCTGCACC
AATTTGGTCAATGAAAGGGGCTCACAATTGAATACGGGCTCCTGCACT
GTACTTTCGCCTCAAAATAAGTGGGACCCGCCTACATTGGTTGCTGTG
GCCCCGGGTGTGCTCGTGTTGCCCTAGAGATCCGAGGTTTGTGTCTCG
TGAGAAAATTGGGCCGTGGGGAGTTTGGCTGGGGCAGCCGGTGATGAG
ATTTTTGGGCAGAGTTGAGTTTGGGTTTTTGAGCAAGAATGGAGGCAT
TGAGCTCATTGGCTGAATCGATCTCTAAGATATTCCATGAAGAACCGT
CATAACCAAGTGAATCTGAACCATTATTAGTTGACTGGGACAGGGCCT
GCTCTAACACTCTATGATGTTGTGGAATAAAAAATTTATCGTTAAAAG
TGCTGACCGGAGGTTGGTGGCCGTTGATTATGTTTCTATATGAGGATA
GACAACGAGACGATCAGGACTGTGGCGCCTAATGTTAAGACGGGCTCG
TTCTGACTTTTTAATTTTGGCCCTCCGGCGTGGCGATAGAATCATCTG
GGACTATATATGGTCATCTTCTCTGCTCTGTGGCCTCAGGATAAAATG
ACCATTTGTTGTGTTATCATGAGACTTTCAATAGTCTAAGTGTAAACT.

The second part is 8 prime pairs that obtained from experiments, which named as primer pairs set, AGGCAT, ACTCAG, AGGCAG, ACGCTG, AAGCTG, ACGCAA, AAGCAA, ACGCAG. The 8 primer pairs are encoded with binary code as example as following: 010000110110, 011110110100, 010000110100, 011100111000, 010100111000, 011100110101, 010100110101, and 011100110100.

3.2. Selecting genomic primers of kiwifruit

The Euclidean distance of 8 primer pairs was calculated as Table 1:

Table 1 The Euclidean distance of 8 primer

β_1	$\beta = 1.0, \quad \beta_2 = 0.0$				$\beta = 1.0, \quad \beta_2 = 0.6$			
	$\alpha = 0.0$		$\alpha = 0.4$		$\alpha = 0.0$		$\alpha = 0.4$	
	first mode	second mode	first mode	second mode	first mode	second mode	first mode	second mode
0.0	19.4709	107.265	18.7791	101.095	26.0933	123.201	25.1949	116.538
0.2	19.7373	113.570	18.9443	106.221	26.4164	129.764	25.3721	121.700
0.4	20.2631	121.759	19.3353	113.098	27.1212	138.738	25.8797	129.097
0.6	21.0210	131.299	19.9260	121.230	28.1703	149.458	26.6828	138.112
0.8	21.9820	141.826	20.6916	130.278	29.5210	161.459	27.7451	148.317
1.0	23.1173	153.085	21.6079	140.009	31.1284	174.413	29.0304	159.409

Then calculate the Euclidean distance between the sample data set and the prime pairs set. The maximum Euclidean distance of gene expression is 6 from encoding ways. And the 8 known prime pairs are about 2 from Table 1. Therefore, the genes that its' Euclidean distance is greater than 3 is as non-primer pairs set as following:

CCTGGC, TATGCC, CTGGCT, GGCTCC, CTTGCC, CTTGAT, GATGCT, TCTGTC, CACCTC, TCCACC, AGTGCA, TGCAGC, AGCGTT, CAGGTC, GTCGCC, GATGAC, ATCAGC, TTTGTT, GATGGT, GTCATC, CACGGT,

GATGCA, TACGCA, CGCAGG, GTCGAC, CGCGAA, ATTGCT, CGTGTG, GCTGCC, TTTGTC, AGCGAT, CTGGTT, ATCGCA, CGCACT, CTAGTC, TTAGGC, CGGGCA, TTGGCT, GGCATC, ATCATC, CTTGGC, GATGGC, TGCATA, TTTGGT, TACGGA, TGTGCC, AACTCC, TGTGGA, ATCGAC, CTTGTT, ACTGCC, TTCAAC, GTTGTC, ATCGTT, TACAGC, CATGAT, ACAGCC, ATTGAC, TTCGTT, CGCACC, AACGCC, TTTGGC, GGCGGA, GCTGGC, CACTGC, CTCCTC, CTCGTC, GCCTCC, TGCACA, CACACG, CACGCA, CATGCC, TTCATC, TATGCT, TTTGAC, TTCGTC, CGTGAG, CGAGTA, CTGGGC, GCTGAC, GTCGTT, ATAGTC, GACTCC, TTGGGC, CTAGAT, GATGCC, CGAGGA, CGAGCT, CGTGTT, CGCGTG, CACTTC, CAAGCA, AGCACT, TCTGAC, GTGGTC, GGCGAG, CATGGC, TATGTT, TACGGT, TACATC, GGTGTA, CTTGGT, CTGGAC, AGCGCC, TCCGTC, CATGAC, TGCACC, TTTGCA, CCTGCA, TTGGTC, CACAAT, CGGGCT, TACGGG, TGCACT, CACTCC, TTCGCC, CCCGCC, GTTGCT, GTTGCC, CTCGTG, TGTGCT, GCCGGT, CATGAA, ATTGGC, CAGGGC, GTTGAC, AACACT, CACTCT, CTGACC, GCCGTT, GGTGGC, GTGGCC, GACGAT, TGTGGC, GACGGG, TTGGCC, TCCGGC, GGCGAT, GGCGTG, CGTGGC, GGTGGA, CTCTGC, TCTGCT, AATGGC, CATGAG, CGCCCT, AATGCT, TACGTG, CGTGAT, TCTGGC, GTCGGT, CATGCT, CTAGCT, CCCGTC, GTTGGT, TATGGC, CAAGTT, GACATC, CTCGTT, CACACT, GGTGCC, GTAGGC, CCTGAT, CCCGGC, CCGGCC, GATGTT, TTCGAT, CTAGTT, TGTGTA, TATGAC, GTCAAC, AACGGT, GTCGTC, ATCGGT, GACGAC, CCTGAC, GGCGGC, GGCGTC, GACAGC, GGCGCC, GACGGC, CACAGC, TACGTA, GACGTA, CGTACT, CTTAGC, TATGCA, TCAGGC, CAGGCT, GGCACT, CTTGGA, CGCAGT, AGCGGC, GCGGCC, TTCAGC, AGCGGA, CTTGAC, GTGGCT, GCAGCC, GGCACC, GACGGA, GTAGCC.

Then the primer pairs and non-primer pairs set are trained with SVM for obtaining a model. At last, the expired primer pairs are learned from the sample data set based on the model.

3.3. Experiment results and discussion

The proposed approach is implemented with Java programming language and executed in a PC. Java programming language is used to implement the Euclidean distance calculation and SVM program for the experiment. Based on the original 8 primer pairs, the extension primer pairs are identified as following (excluding three consecutive bases):

GAATGG, GGATGG, GAGCTG, GCGTGT, AGGCTA, ATATGG, TGGCTG, ACACAA, AAGTGG, TGGCGA, GAGCAT, AGGCGG, GCATGG, GTACAT, GAGCGC, GTATGG, AAGTGA, GCATGA, ATACGG, GTGAGG, TTGCTG, GTGCTC, GTGTTG, GAGTTG, AGGCAT, GAACCG, GGACAG, ATGCTG, TGGTGG, TGGCCG, AGACGA, GGACTG, GATCAG, AGACGG, GCTCGA, GCGTGG, GGACTA, GCTCTG, GTGTTA, GTGTAA.

In addition, it can be observed that the sample data contains the prime pair

AGGCAT. A new prime pairs set is settled with excluding AGGCAT. For the constraints become relaxed, the non primer pairs set is settled with the sample genes that it's Euclidean distance are larger than 4. Based on the two new sets, new primer pairs are obtained. The experimental result contains the original primer AGGCAT. That is, by deleting the original primer pair of gene sequences, it can be identified with the proposed method.

4. Conclusion

This paper has proposed a SVM-based approach for selecting primer pairs in tissue culture process. Euclidean distance between genes is used to identify the informative genes in datasets. And the selection of tissue cultured kiwifruit primers was analyzed as an example. By deleting original primers of gene sequence, the method could also identify the primers from sample gene sequences obtained from experiments using DNA molecular markup technology. This result shows that the proposed method is effective, and provides a more reasonable way to select prime pairs besides DNA molecular markup technology.

References

[1] X. Gao, D. Yang, D. Cao, M. Ao, X. Sui, Q. Wang, J. N. Kimatu, L. Wang: *In vitro micropropagation of freesia hybrida and the assessment of genetic and epigenetic stability in regenerated plantlets.* Plant growth regulation *29* (2010), 257–267.

[2] Q. M Wang, Y. Z. Wang, L. L. Sun, F. Z. Gao, W. Sun, J. He, X. Gao, L. Wang: *Direct and indirect organogenesis of clivia miniata and assessment of DNA methylation changes in various regenerated plantlets.*Plant cell reports *31* (2012), No. 7, 1283–262.

[3] G. RaviKumar, G. A. Ramachandra, K. Nagamani: *An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets.* International journal of advanced research in computer science and software engineering *4* (2014), No. 2, 272–277.

[4] T. Abeel, T. Helleputte, Y. Vande Peer, P. Dupont, Y. Saeys: *Robust biomarker identification for cancer diagnosis with ensemble feature selection methods.* Bioinformatics *7* (2010), No. 3, 392–398.

[5] I. Guyon, J. Weston, S. Barnhill, V. Vapnik: *Gene selection for cancer classification using support vector machines.* Machine learning *46* (2002), No. 1, 389–422.

[6] K. Y. Yeung, D. R. Haynor, W. L. Ruzzo: *Validating clustering for gene expression data.*Bioinformatics *17* (2001) 309–318.

Received November 16, 2017