Expert Systems with Applications 37 (2010) 5946–5950

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Support vector machine method on predicting resistance gene against *Xanthomonas oryzae* pv. *oryzae* in rice

Xia Jingbo [a,*], Hu Xuehai [a], Shi Feng [a], Niu Xiaohui [a], Zhang Chengjun [b]

[a] College of Science, Huazhong Agricultural University, Wuhan, Hubei 430070, PR China
[b] National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research, Huazhong Agricultural University, Wuhan, Hubei 430070, PR China

ARTICLE INFO

Keywords:
Support vector machine
Prediction
Gene
Sequence
Chaos games representation

ABSTRACT

Motivation: Identification of disease-resistant genes in the rice is a tough work in various experimental studies. *Xanthomonas oryzae* pv. *oryzae* (*Xoo*) which causes bacterial blight are considered to be the most devastating diseases in most rice-growing regions. However, currently there is no existing method for the prediction of disease-resistant genes from sequence data. Accurate prediction of *Xoo* from protein sequences is illuminating for gene finding projects.

Results: We propose a novel machine-learning approach based on the method of support vector machine (SVM) and chaos game representation (CGR), to assess the chance of a protein in rice to be *Xoo* resistant. We choose 13 already cloned genes for positive data and 48 selective gene in rice for negative data, the average accuracy achieves 100% in resubstitution test, 95.08% in jackknife test, and the Matthews correlation coefficient achieves 0.8509. The successful application of SVM + CGR approach in this study suggests that it should be more useful in quantifying the protein sequence–structure relationship and predicting the structural property profiles from protein sequences.

1. Introduction

Rice is the main food for plenty of people in the world. Although the third time of increasing rice harvest is initiated for decade with the goal of "super rice" or "super hybrid rice", still a number of challenges have to be met to achieve the goal of increasing rice production in a sustainable manner. The biggest challenge is the increasing occurrence of diseases in almost all of the rice-producing areas causing great yield loss (Zhang, 2007a).

It becomes an accepted fact that the best idea to resolve the problem is rapid development of Green Super Rice (GSR) which requires a gradual reduction in application of pesticides, fertilizers, and water while still achieving continuous yield increase and quality improvement, of course including adequate resistances to major diseases and insects. Therefore, identification of Genes for Disease Resistance is one of most important step in GSR (Zhang, 2007b).

Xanthomonas oryzae pv. *oryzae* (*Xoo*) which causes bacterial blight second only to the Pyricularia grisea which causes rice blast are considered to be the most devastating diseases in most rice-growing regions. Now six gene Xa1, xa5, xa13, Xa21, Xa3/Xa26, and Xa27 have been reported to be isolated for bacterial blight resistance (Gu et al., 2005; Iyer & McCouch, 2004; Ponciano, Yoshikawa, Lee, Ronald, & Whalen, 2006; Ronald & Song, 2006; Song et al., 1995; Sun, Yang, Wang, & Zhang, 2003; Yoshimura et al., 1998). The Xa21 gene is the first-cloned bacterial blight-resistant gene which encodes a leucine-rich repeat (LRR) receptor-like protein kinase and the research suggestive of a role in cell surface recognition of a pathogen ligand and subsequent activation of an intracellular kinase leading to a defence response (Song et al., 1995). Along with xa21, some other genes are cloned, e.g., Xa26, which also encodes a LRR receptor kinase-like protein (Sun et al., 2004). Meanwhile, Xa1 is induced by pathogen infection and wound which encodes a cytoplasmic receptor-like protein with NBS domain and nucleotide-binding LRR domain (Yoshimura et al., 1998). The recessive gene xa5 is a general eukaryotic transcription factor which encodes the gamma subunit of transcription factor IIA (TFIIAγ) (Iyer & McCouch, 2004). The fully recessive gene xa13 encodes a novel plasma membrane protein that plays a key role in both disease resistance and pollen development (Chu et al., 2006). Xa27 induction will occur only in the immediate vicinity of infected tissue which encodes identical proteins to avrXa27 whose product is a nuclear localized type-III effector (Gu et al., 2005).

Meanwhile, some other genes are finding out to be connected with Xa21 gene. XB3 is an E3 ubiquitin ligase, as a substrate for the XA21 Ser and Thr kinase will bind to XA21. XB3 contains an ankyrin repeat domain and a RING finger motif, the research indicate that Xb3 is necessary for full accumulation of the XA21 protein and for Xa21-mediated resistance (Ronald & Song,

* Corresponding author. Tel.: +86 02787282425; fax: +86 02787282133.
E-mail address: xjb@mail.hzau.edu.cn (J. Xia).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.010

2006). OsPRI genes PR1a, PR1b and PR1c were cloned that at the juvenile and adult stages will induce a resistance response to a wildtype *Xoo* strain when the Xa21 locus exists (Ponciano et al., 2006).

By the literatures mentioned earlier, we note that in usual way, one used to find disease-resistant gene by experimental methods, which is a trial and error procedure. With the finish of the whole genome sequencing of the rice, sequence-based prediction using bioinformatic method sheds light on the gene recognition research field.

A support vector machine (SVM) is an effective tool for classification and prediction, which has been used in various fields related to protein function prediction, such as prediction of thermal protein (Hu et al., 2009) and soluble protein (Susan, Abhijit, Bhaskar, Valadi, & Petety, 2006). However, it is still a fresh idea for *Xoo* prediction.

During our research, an SVM method is used in order to accelerate finding the *Xoo* disease-resistant genes and make some prediction to the fine-mapped genes. A key issue in manipulating SVM is what mathematical features can one extract from the data itself, since different features chosen might lead to a big difference in the accuracy of prediction.

It is widely accepted that the functions of proteins are affected by their structure. In common, during the research in protein function prediction, residue composition, dipeptide composition and tripeptide composition play important role in feature extraction during protein function prediction and have got some successes. Unfortunately, these methods cannot generate a good visual representation (Hao, Lee, & Zhang, 2000). Jeffrey (1990) proposed the chaos game representation (CGR) of DNA sequences, which performs the pattern hiding in sequences. In fact, CGR is an iterative mapping technique that processes a given sequence into a picture (see Section 2) with fractal structure, visually revealing previously unknown structure.

Furthermore, the CGR of DNA sequences has been extended to represent protein sequences and to study protein structure (Basu, Pan, Dutta, & Das, 1997; Fiser, Tusnady, & Simon, 1994; Yu, Anh, & Lau, 2004). In order to discriminate patterns of protein sequences belonging to different functional classes, Basu et al. (1997) used CGR algorithm to generate protein sequence using a 12-sided regular polygon with each vertex representing a group of amino acid residues leading to conservative substitutions. The authors claim that CGR has the potential to reveal the evolutionary and functional relationships even between the proteins having no significant sequence homology, which is the fundamental character of sequence alignment.

In this paper, we present a novel prediction technique that combined the method of SVM and CGR, to assess the chance of a protein in rice to be *Xoo* resistant. The average accuracy achieves 100% in resubstitution test and 95.08% in jackknife test.

2. Materials and methods

2.1. Data set

The proteins for the analysis are chosen based on literature (Chu & Wang, 2007; Gu, Sangha, Li, & Yin, 2008; Sun et al., 2003; Wu, Li, Xu, & Wang, 2008; Zhang, 2007a) reports on the *Xanthomonas* resistance, and the sequences are collected from NCBI (http://www.ncbi.nlm.nih.gov/) using the key words (gene names) mentioned in the above-mentioned papers.

All we get from the NCBI comprised 48 proteins (Wu et al., 2008) including a few alleles and partial CDS. In order to reduce the redundancy, the homologues with more than 95% similarity are eliminated by CD-HIT (Li, Jaroszewski, & Godzik, 2002). By this way, thirteen proteins left, which are listed in Table 1. These proteins are all proved resistance to the *Xoo* directly or indirectly.

The negative data are randomly chosen from the protein data set download from KOME (ftp://cdna01.dna.affrc.go.jp/pub/data/20081001/INE_FULL_SEQUENCE_AMINO_DB.zip) by the standard that does not belong to the 48 positive data and the protein is longer than 100aa. Even though a huge number of proteins are in this data set and obviously have no evidence that they are *Xoo*-resistant gene, we just choose 48 out of it so as to keep the balance of positive and negative data set size.

2.2. SVM

A support vector machine (SVM) is a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the better the generalization error of the classifier.

Suppose we are given a set of samples, i.e. a series of input d-tuple vectors $X_i \in R^d$ $(i = 1, 2, \ldots, N)$, with corresponding labels $y_i \in \{-1, +1\}$ $(i = 1, 2, \ldots, N)$, where +1 and −1 are used to stand, respectively, for the positive set and negative set. The goal here is to construct a classifier and derive one decision function from the available samples, which has small probability of misclassifying a future sample.

SVM performs a nonlinear mapping of the input vector X from the input space R_d into a higher dimensional Hilbert space, where the mapping is determined by the kernel function

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2).$$

Table 1
List of *Xoo*-resistant gene in rice.

	Gi number	Description
1	117655416	Indica pathogenesis-related protein PR1a(Pr1a) mRNA
2	117655418	Indica pathogenesis-related protein PR1b(Pr1b) mRNA
3	117655420	Indica pathogenesis-related protein PR1c(Pr1c) mRNA
4	2943741	Oryza sativa mRNA for XA1 (Yoshimura et al., 1998)
5	55585038	Indica transcription factor IIA gamma subunit (TFIIAy) mRNA (Iyer & McCouch, 2004)
6	89892339	Indica cultivar IR24 disease-resistant allele XA13 (Xa13) gene (Chu et al., 2006)
7	89892335	Indica cultivar IRBB13 disease-resistant allele XA13 (Xa13) gene (Chu et al., 2006)
8	94481122	Indica group Xa21 gene for receptor kinase-like protein: II you 8220
9	14279687	Receptor-like kinase Xa21-binding protein 3 (Xb3) mRNA
10	90018760	Indica bacterial blight-resistance protein XA26 (Xa26) gene (Xiang, Cao, Xu, Li, & Wang, 2006)
11	66735941	Indica xa27-IR24 allele (Gu et al., 2005)
12	66735943	Indica Xa27-IRBB27 allele (Gu et al., 2005)
13	68248527	Japonica XB3-related protein (XBOS36) mRNA

The function is called the RBF (radial basis function) kernel with one parameter γ. Finally, for the selected kernel function, the learning task amounts to solving the following convex quadratic programming (QP) problem,

$$\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j),$$

subject to

$$0 \le \alpha \le C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0,$$

where the form of the decision function is $f(x) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) + b)$.

For a given data set, only the kernel function and the regularity parameter C must be selected. A complete description to the theory of SVMs for pattern recognition is found in Vapnik (1998).

2.3. Features used in SVM training

The features used in this paper are residue compositions, dipeptide compositions and CGR features.

2.3.1. Residue and dipeptide composition

It is known that the protein sequence is formed by 20 different kinds of amino acids, so we can extract the residue composition as the desired information directly from protein sequence. Thus, we get 20 mathematical features by this means.

On the other hand, for a protein that contained n amino acids $P_1, P_2, \ldots, P_n$, it contained $n-1$ dipeptides, such as $P_1P_2, P_2P_3, \ldots, P_{n-1}P_n$. The composition of a dipeptide was defined as:

$$compostion(i) = dipeptide(i)/n - 1,$$

where i denotes the 400 dipeptides, $dipeptide(i)$ denotes the number of the ith dipeptide.

A native idea is taking tripeptide for future considerable feature, but unfortunately this feature seldom makes sense. The possible reason for this is that many tripeptides are not represented at all, owing to the small length of the proteins.

2.3.2. CGR algorithm and CGR features

In order to generate visually identifiable distinct patterns of protein sequence, Basu et al. (1997) classified 20 kinds of amino acids to 12 different groups according to their different function (Dayhoff, 1978). The reduced groups are listed below:

[Isoleucine(I), Leucine(L), Valine(V), methionine(M)], [Arginine(R), Lysine(K)], [Aspatic acid(D), Glutamic acid(E)], [Asparagine(N)], [Glutamine(Q)], [Histidine(H)], [Serine(S), Threonine(T)], [Proline(P)], [Alanine(A), Glycine(G)], [Cysteine(C)], [Phenylalanine(F), Tyrosine(Y)], [Tryptophan(W)].

Each group represents a vertex of 12-vertex polygon. Furthermore, Basu et al. claim that the following 12-vertex CGR algorithm is optimum for generation of distinct patterns for different protein families:

Step 1. Draw a 12-sided regular polygon, and each vertex represents a kind group of amino acids;
Step 2. Set the center point as the initial point;
Step 3. Given a protein sequence with length N, we draw N points in the polygon by the following way: in turn we read alphabet from the protein sequence, since each read belongs to one group of amino acids, then we determine a certain vertex of polygon and we draw the midpoint of initial point and the chosen vertex. After finishing draw one point, we set it to be the new initial point, and we can draw N points with such iteration. The procedure is illustrated in Fig. 1. More precisely, if we let $P_1(x_1, y_1), \ldots, P_{12}(x_{12}, y_{12})$ be the coordinates of 12 vertex of the polygon, then we can get coordinates of every points in CGR by the following formula:

$$CGR_i(x) = (CGR_{i-1}(x) + P_j(x_j))/2,$$
$$CGR_i(y) = (CGR_{i-1}(y) + P_j(y_j))/2,$$
$$i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, 12,$$

where $CGR_i(x, y)$ means the coordinate of the ith point drawn in CGR, and $P_j(x_j, y_j)$ represents the coordinate of chosen vertex by the jth read (each read determines a certain vertex of polygon).

Step 4. The 12-sided polygon is divided into 24 segments (grids) as shown in Fig. 1, and the segments are labeled serially with numbers 1–24 (not shown in Fig. 1). For each segment, namely, S_k, we count the number of points fall in it, and denote as L_k. (The points falling on boundaries should be counted in any one of the neighboring segments.) Then set

$$G_k = L_k/N, \quad k = 1, 2, \ldots, 24,$$

where N is the length of the protein sequence.

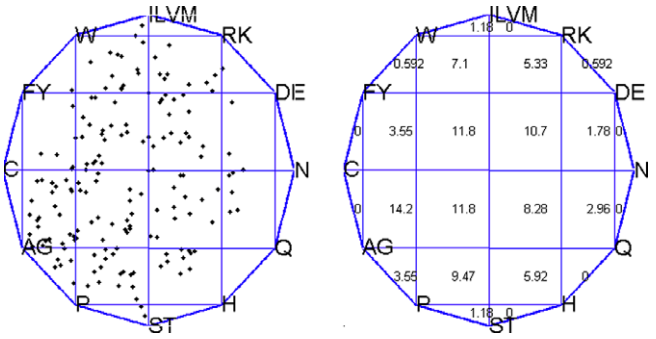

Fig. 1. Generation of vectors in the CGR of proteins.

From the above 12-vertex CGR algorithm, we find that each protein sequence will induce a 24-tuple vector $(G_1, G_2, \ldots, G_{24})$. We put the first protein in Table 1 as an example, where sequence is as below:

>GI number: 117655416
MASSSSRLSCCLLVLAAAAMAATAQNSAQDFVDPH
NAARADVGVGPVSWDDTVAAYAESYAAQRQGDCK
LEHSDSGGKYGENIFWGSAGGDWTAASAVSAWVSE
KQWYDHGSNSCSAPEGSSCGHYTQVVWRDSTAIGC
ARVVCDGDLGVFITCNYSPPGNFVGQSPY

and the vector got from the CGR of the protein is $(G_1, G_2, \ldots, G_{24})$ = (0.0118, 0, 0.00592, 0.071, 0.0533, 0.00592, 0, 0.0355, 0.118, 0.107, 0.0178, 0, 0, 0.142, 0.118, 0.0828, 0.0296, 0, 0.0355, 0.0947, 0.0592, 0, 0.0118, 0).

3. Results

SVM classifiers are applied to discriminate between *Xoo*-resistant or non-resistant proteins. Usually, a predictive method is evaluated by two different approaches, the resubstitution test and the jackknife test.

By the test of resubstitution, the structural class of each protein in a training data set is predicted using the rules derived from the same set. Although this test gives somewhat optimistic error esti-

mate because the same proteins are used to derive the prediction rules and to predict themselves, the resubstitution test is absolutely necessary due to its ability to reflect the self-consistency of a given method.

On the other hand, a cross-validation test for an independent testing data set is also needed because it can reflect the extrapolating effectiveness of a predictive method. Jackknife test is thought to be the most reliable one among cross-validation tests, and it is also called the test of leave-one-out. Moreover, in the jackknife test, each protein is singled out as an independent sample and used to examine the predictive method.

To start with, the SVM classifier is trained with 420 features comprising 20 residues and 400 dipeptides. In the resubstitution test, the prediction accuracy achieves 100%, it shows the self-consistency as one expects. While in the jackknife test, the test accuracy of the prediction is 93.44%, where $\gamma = 50$, $C = 90$.

Secondly, SVM classifier is trained with 24 features got from CGR. It also amazingly achieves 100% in resubstitution test. In the jackknife test, the smaller size of features slightly decrease the prediction accuracy to 88.52%, where $\gamma = 300$, $C = 400$.

Finally, we combine the above 444 features altogether in SVM classifier, and the accuracy in resubstitution remains 100%. While in the jackknife test, the classifier makes out 7 true data out of 13 positive data and recognizes all of 48 true negative data. Therefore, the accuracy achieves 95.08%, where $\gamma = 50$, $C = 90$.

Alternatively, we delete 400 dipeptide composition from above 444 features, so 44 features remain including residue composition and CGR features. Amazingly, with parameters chosen as $\gamma = 90$, $C = 10$, the classifier not only performs quick, but also achieve the same performance as the above-mentioned one.

The experimental result is shown in Table 2, where the Matthews Correlation Coefficient (MCC) is used in machine learning as a measure of the quality of binary classifications. It performs well even if the classes are of very different sizes. A returning value of +1 represents a perfect prediction, 0 an average random prediction and −1 an inverse prediction.

Another performance comparison between three kinds of different experiments results with 24, 420, and 444 features in classifiers are listed in Figs. 2–4. Here, ROC is a graphical plot of the sensitivity vs. (1-specificity) for a binary classifier system by varying its discrimination threshold value, and ROC value denote the area under ROC (Receive Operating Characteristic) curve. The bigger ROC value represents the higher accuracy of the classifier.

Meanwhile, ROCCH value denotes the area under convex hull of ROC curve which is also useful for the evaluation of machine-learning techniques.

The ROC and ROCCH value in Figs. 2–4 indicate that three SVM classifier developed in this paper are reasonable. Moreover, classi-

Table 2
Prediction result of jackknife test via SVM.

Feature	Sensitivity[a] (%)	Specificity[b] (%)	Accuracy[c] (%)	MCC[d]
420	90.91	76.92	93.44	0.7972
24	75	69.23	88.52	0.6488
444/44	100	76.92	95.08	0.8509

[a] Sensitivity is defined as TP/(TP + FP), where TP and FP are the numbers of correctly and incorrectly classified positive data, respectively.

[b] Specificity is defined as TP/(TP + FN), where FN is the number of incorrectly classified negative data, which provides a high enough specificity that its predictions can be experimentally verified at a reasonable cost.

[c] Prediction accuracy = (TP + TN)/n, where TN is the number of correctly classified negative data and n is the size of the data sets.

[d] MCC equals to $(TP \times TN - FP \times FN)/\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}$.

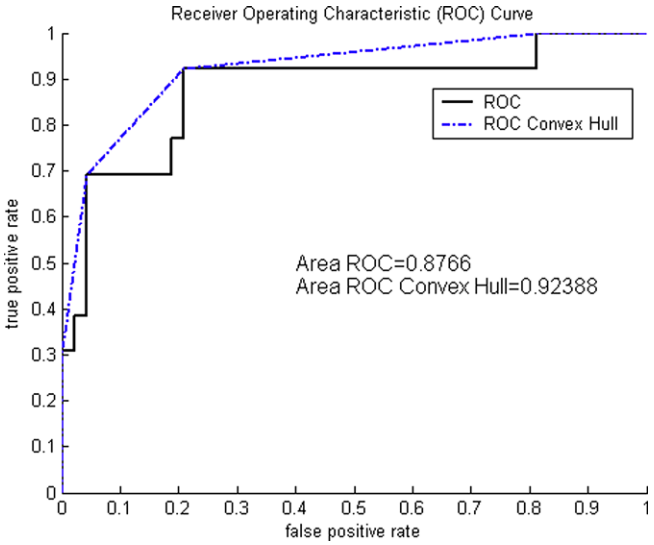

Fig. 2. ROC and ROCCH curve for SVM classifier (with 420 features).

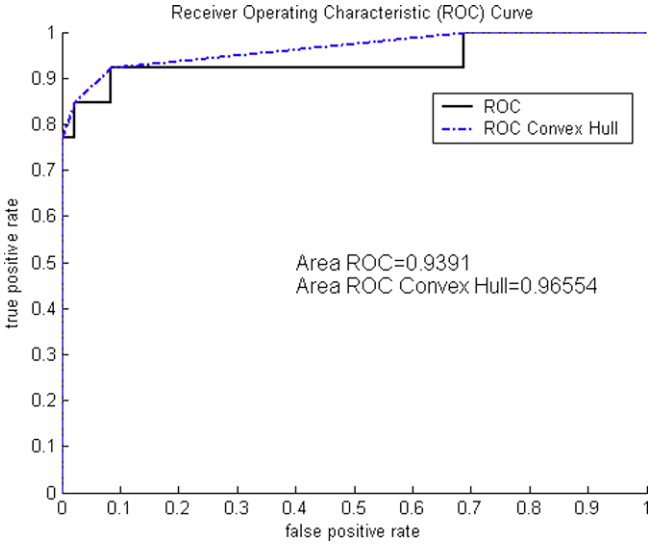

Fig. 3. ROC and ROCCH curve for SVM classifier (with 24 features).

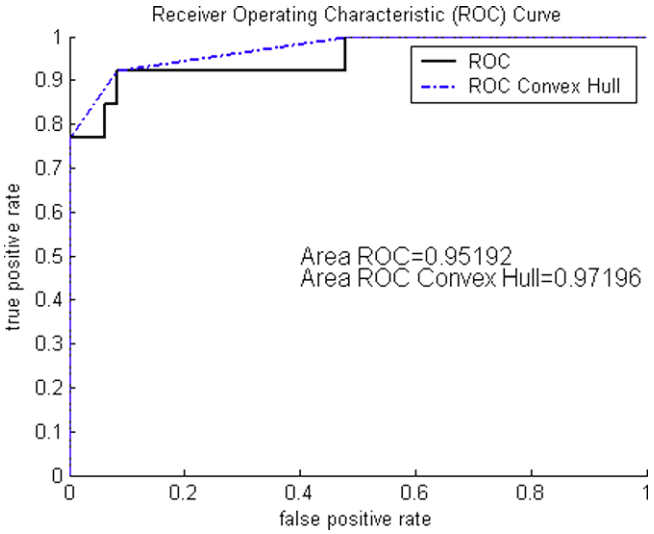

Fig. 4. ROC and ROCCH curve for SVM classifier (with 444 or 44 features).

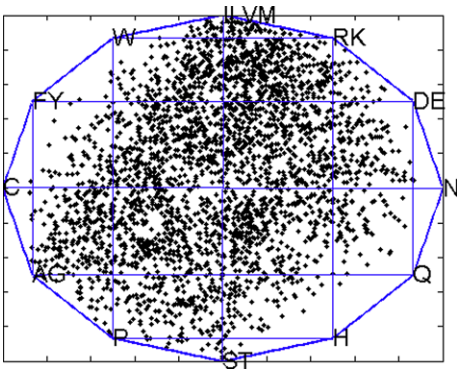

Fig. 5. *Xoo*-resistant gene sequence.

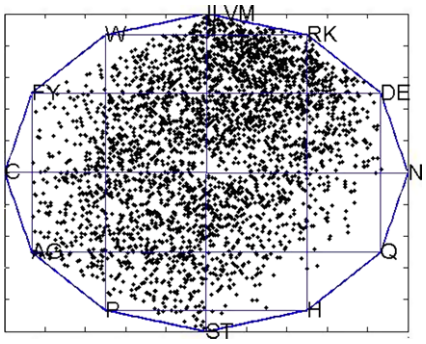

Fig. 6. Non-resistant gene sequence.

fier with 44 features is the most efficient and reliable one among them.

4. Conclusion

In comparison with above-mentioned previous works, we can conclude that our CGR + SVM method have some advantages:

(1) Combining residue, dipeptide and CGR as the whole features in SVM classifier is a new idea.
(2) The dimension of data set is considerable. Especially, one can note that CGR plays an important role among them. As we know, the small dimension of data set naturally leads to fewer store space and faster operation speed.
(3) Instead of the amino acid composition used widely in the previous structural class prediction work, the chaos games' figures are used for the structural class prediction. The CGR algorithm can generate a picture of every protein sequence, which can provide visual help.

Taking two pictures for visual example, we draw two CGR map in Figs. 5 and 6. Picture in Fig. 5 is CGR map drawn from a long protein sequence assembled by several *Xoo*-resistant genes, and picture in Fig. 6 is related to non-resistant gene. Visual comparison shows that they do have some difference in dots' position, density and clustering trend.

We expect that applying our SVM-CGR-based classification approach helps to obtain high recognition rates for the detection of *Xoo*-resistant genes.

Acknowledgments

This work was partly supported by Discipline-crossing Research Foundation and Doctor Foundation of Huazhong Agricultural University.

References

Basu, S., Pan, A., Dutta, C., & Das, J. (1997). Chaos game representation of protein. *Journal of Molecular Graphics and Modelling, 15*, 279–289.
Chu, Z., & Wang, S. (2007). Isolation, structure, function relationship, and molecular evolution of disease resistance genes. In Q. Zhang (Ed.), *Genetics and improvement of resistance to bacterial blight in rice* (pp. 349–377). Beijing: Science Press.
Chu, Z., Yuan, M., Yao, J., Ge, X., Yuan, B., Xu, C., et al. (2006). Promoter mutations of an essential gene for pollen development result in disease resistance in rice. *Genes and Development, 20*(10), 1250–1255.
Dayhoff, M. (1978). *Atlas of protein sequence and structure*. Maryland: National Biomedical Research Foundation, Silver Spring.
Fiser, A., Tusnady, G. E., & Simon, I. (1994). Chaos game representation of protein structure. *Journal of Molecular Graphics, 12*, 302–304.
Gu, K., Sangha, J. S., Li, Y., & Yin, Z. (2008). High-resolution genetic mapping of bacterial blight resistance gene Xa10. *Theoretical and Applied Genetics, 116*, 155–163.
Gu, K., Yang, B., Tian, D., Wu, L., Wang, D., Sreekala, C., et al. (2005). R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature, 435*(7045), 1122–1125.
Hao, B. L., Lee, H. C., & Zhang, S. Y. (2000). Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons & Fractals, 11*, 825–836.
Hu, X. H., Xia, J. B., Niu, X. H., Ma, X., Song, C. H., & Shi, F. (2009). Chaos game representation for discriminating thermophilic from mesophilic protein sequences. In *The 3rd international conference on bioinformatics and biomedical engineering (iCBBE2009), Beijing*.
Iyer, A. S., & McCouch, S. R. (2004). The rice bacterial blight resistance gene xa5 encodes a novel form of disease resistance. *Molecular Plant–Microbe Interactions, 17*(12), 1348–1354.
Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research, 18*, 2163–2170.
Li, W., Jaroszewski, L., & Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics, 18*, 77–82.
Ponciano, G., Yoshikawa, M., Lee, J. L., Ronald, P. C., & Whalen, M. S. (2006). Pathogenesis related gene expression in rice is correlated with developmentally controlled Xa21-mediated resistance against *Xanthomonas oryzae* pv. *Oryzae*. *Physiological and Molecular Plant Pathology, 69*, 131–139.
Ronald, P. C., & Song, W. Y. (2006). Rice XA21 binding protein 3 is a ubiquitin ligase required for full Xa21-mediated disease resistance. *Plant Cell, 18*, 3635–3646.
Song, W. Y., Wang, G. L., Chen, L. L., Kim, H. S., Pi, L. Y., Holsten, T., et al. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene Xa21. *Science, 270*, 1804–1806.
Sun, X., Cao, Y., Yang, Z., Xu, C., Li, X., Wang, S., et al. (2004). Xa26, a gene conferring resistance to *Xanthomonas oryzae* pv. *oryzae* in rice, encodes an LRR receptor kinase-like protein. *Plant Journal, 37*, 517–527.
Sun, X., Yang, Z., Wang, S., & Zhang, Q. (2003). Identification of a 47 kb DNA fragment containing Xa4, a locus for bacterial blight resistance in rice. *Theoretical and Applied Genetics, 106*, 683–687.
Susan, I. T., Abhijit, J. K., Bhaskar, D. K., Valadi, K. J., & Petety, V. B. (2006). A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in *Escherichia coli*. *Bioinformatics, 22*(3), 278–284.
Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
Wu, X., Li, X., Xu, C., & Wang, S. (2008). Fine genetic mapping of xa24, a recessive gene for resistance against *Xanthomonas oryzae* pv. *oryzae* in rice. *Theoretical and Applied Genetics, 118*(1), 185–191.
Xiang, Y., Cao, Y., Xu, C., Li, X., & Wang, S. (2006). Xa3, conferringresistance for rice bacterial blight and encoding a receptor kinase-like protein, is the same as Xa26. *Theoretical and Applied Genetics, 113*(7), 1347–1355.
Yoshimura, S., Yamanouchi, U., Katayose, Y., Toki, S., Wang, Z. X., Kono, I., et al. (1998). Expression of Xa1 a bacterial blight-resistance gene in rice is induced by bacterial inoculation. *Proceedings of the National Academy of Sciences of the United States of America, 95*(4), 1663–1668.
Yu, Z. G., Anh, V., & Lau, K. S. (2004). Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *Journal of Theoretical Biology, 226*, 341–348.
Zhang, Q. (2007a). Genetics of quality resistance and identification of major resistance genes to rice bacterial blight. In Q. Zhang (Ed.), *Genetics and improvement of resistance to bacterial blight in rice* (pp. 130–177). Beijing: Science Press.
Zhang, Q. (2007b). Strategies for developing Green Super Rice. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 16402–16409.