I.J. Intelligent Systems and Applications, **2012, 11, 62-67**
Published Online October 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2012.11.07

Prediction of Drought Resistance Gene with Clustered Amino Acid Features

Xia Jingbo, Shi Feng*, Hu Xuehai, Li Zhi, Song Chaohong, Xiong Huijuan
College of science, Huazhong agricultural university, Wuhan, P.R. China
Institute of applied mathematics, Huazhong agricultural university, Wuhan, P.R. China
(* To whom correspondence should be addressed, shifeng@mail.hzau.edu.cn)

***Abstract*—**Drought resistant gene plays important role in molecular breeding while little is known for its genetic mechanism. By extracting the clustered amino acids features, crucial numerical features are inferred for the resistance property of the given gene. Support vector machine algorithm is used to testify the reliability of feature extraction method. After carefully parameters choosing, the accuracy of the predictor achieves 79.36% in Jack-knife test, and the Mathews correlation coefficient achieves 0.5636.

Index Terms— Support Vector Machine, Classifier, Amino Acid Composition, K-Means

I. Introduction

For crops, drought resistance refers to the ability of a crop plant to survive or keep sufficient yield in a water deficit environment. Drought resistant plants own complex expressions including various morphological physiological and biochemical traits. For instance, drought resistant plant owns special traits in leaf age, leaf area, leaf rolling, rooting system and plant height, it also make differences in the reduced transpiration, high water-use efficiency, stomatal closure and osmotic adjustment, etc[1, 2].

Traditionally, researchers selected proper drought resistant species by evaluating and screening the crucial multi-index of the above drought resistance traits [3]. Meanwhile, marker-assisted selection was involved to help breeders to improve drought-related traits. Analysis of sequence data and gene products should facilitate the identification and cloning of genes at target QTLs [4].

In general, the *phytohormone* abscisic acid (*ABA*) was known to be essential to stress resistance, which acted as a hormone in plant development, drought tolerance, and adaptive responses to environmental stresses [5]. It was believed that plant begin to conserve water under the influence of certain hormones. After unknown complex signaling procedure, their leaves closed microscopic pores to stop water loss, they slowed their own growth, and they signaled numerous genetic changes [6]. However, plant hormones regulated stress responses through complex molecular networks, and very little was known about the genetic mechanisms and *ABA* receptor recognition and signaling procedures [5].

A landmark result was obtained by Nishimura, who showed structures of *pyrabactin* resistance 1 (*PYR1*), a prototypical *PYR*/ *PYR1*-like (*PYL*)/ regulatory component of ABA receptor. Each copy of the *PYR1* molecule had an internal open space like the inside of a tin can, and when a hormone molecule comes along, it fitted neatly into one of the two spaces [6].

Unfortunately, these results are still insufficient to ensure wide scale genes finding. With the consistently accumulation of gene data and the development of the bioinformatics methods, sequence based prediction methods shed light on the candidate genes prediction.

Generally, statistic information of amino acid composition is used to explorer the correlation between protein sequence and the function it manifested [7, 8]. After that, a more comprehensive work could be done by using information of dipeptide composition [9, 10]. Actually, the utilization of the sequence based information is far away from perfection. For this, the most illuminative work could be traced to Kuo-chen Chou [11], who proposed Pseudo amino acid composition (PseAAC). In PseAAC, each amino acid sequence was converted to a numerical vector, representing abundant physical and chemical information, including hydrophobicity, polarity, acidity, etc.

Up until now, few work focus on the prediction of drought resistant gene. The motive of this paper is to propose a novel prediction tool so as to evaluate the possibility of the unknown protein to be drought resistant or not. Training data are retrieved from public dataset, Swiss-prot and Pfam. Numerical features, including Amino acid, dipetide and tripeptide compositions, are screened to reveal the properties of each protein. By using support vector machine, a classifier is built for prediction.

After carefully parameters choosing, the accuracy of the predictor achieves 79.36% in Jackknife test, and the Mathews correlation coefficient achieves 0.5636.

II. Materials and Methods

2.1 Data

We obtained 101 drought tolerant protein sequences from Swiss-Prot database (http:// expasy.org/sprot/). All of the sequences are not fragment and with sequence length more than 100 amino acids. These proteins are set as the positive dataset.

Meanwhile, non-resistant protein sequences are downloaded from seed proteins of the Pfam database (http:// pfam.jouy.inra.fr/). Pfam is a large collection of multiple sequence alignments and hidden Markov models covering 9318 protein families, based on the Swiss-Prot and SP-TrEMBL databases.

In order to delete the homology of the dataset, the sequences with 90% sequence similarity are removed by using CD-HIT [12]. After that, the positive dataset consist of 87 drought resistant genes. While the negative dataset is constructed with 129 seed proteins from different protein family in Pfam dataset, which are manually downloaded and checked to be unrelated to drought resistant and environment stress.

2.2 K-means Clustering Algorithm

K-means clustering is a simple unsupervised learning algorithm which aims to partition n samples into k clusters in which each samples belongs to one cluster in the shortest distance [13]. The algorithm works as follow steps:

Algorithm (K-means clustering)

1. *Place k points as the initial centroid of each group.*
2. *Repeat*
3. *Assign each object to the group that has the closest centroid.*
4. *Recalculate the positions of each centroids.*
5. *Until convergence/ centroids stop moving.*

By using k-means clustering algorithm, numerical features with high dimension could be greatly reduced.

2.3 The Proposed Feature Selection Method

Generally, numerical features of the protein sequence are constituted by a numeric vector with high dimensions. In detail, the dimension of residue composition corresponds to 20, dipeptide corresponds to 400, and tripeptide corresponds to 8000. In sum, a numerical vector with 8420 dimension is constructed by this way.

However, dataset with high dimensions does not perform well in classifier or regression, since it not only hinders the effectiveness of the feature selection in protein sequences, but also significantly decreases the efficiency of the algorithm. Hence, it is a natural idea that the above features could be clustered into k groups according the physical and chemical properties. By doing this, the size of dataset is reduced into a relative small scale.

Due to the above consideration, a large scale feature reduction is carried on by clustering algorithm. As seen from Table 1, each residue owns a unique hydrophilicity value [14], which in turn reflects an important property for the protein structure. Denote x_i as the hydrophilicity value of the corresponding residue, a 2-tuple vector (x_1, x_2) is assigned for each dipeptide (R_1R_2). By doing this each dipeptide is presented as a 2-tuple numerical vector and could be prepared to be clustered. After implementing K-means clustering algorithms, 400 dipeptides are clustered into 16 groups. Similarly, a 3-tuple vector (x_1, x_2, x_3) is assigned for each dipeptide $(R_1R_2R_3)$ and 8000 tripeptides are grouped into 26 groups.

By this approach, each protein sequence could be mapped into a 62-dimensional vector, $(x_1,...,x_{26},x_{27,},...,x_{62})$, where the first 26 coordinates correspond to the composition of 26 chosen tripeptides, the followed 16 dimensions correspond to the composition of 16 chosen dipeptides, and the last 20 ones correspond to the amino acid compositions.

Table 1: List of hydrophilicity value of each amino acid

Amino acid	Hydrophilicity value
Arginine	3
Aspartic acid	3
Glutamic acid	3
Lysine	3
Serine	0.3
Asparagine	0.2
Glutamine	0.2
Glycine	0
Proline	0
Threonine	-0.4
Alanine	-0.5
Histidine	-0.5
Cysteine	-1
Methionine	-1.3
Valine	-1.5
Isoleucine	-1.8
Leucine	-1.8
Tyrosine	-2.3
Phenylalanine	-2.5
Tryptophan	-3.4

2.4 Support Vector Machine

Support vector machine (SVM) is a popular prediction tool based on statistical learning, which has been widely used in classification and regression in various fields. It works by selecting a proper hyper plane with minimum distance between different group data. SVM has shown its robustness and efficiency in function gene recognition, while in this paper, it is also selected to be the standard tool to classify the protein to be drought resistance and also evaluate the effectiveness of the proposed feature extraction method. For implicit information of SVM theory, refers to Cortes and Vapnik [15].

The SVM toolbox used in this study is osusvm (http://www.ece.osu.edu/~maj/osu_svm/), a tool kit developed in Matlab circumstance.

III. A Improved Particle Swarm Optimization Algorithm Model And Its Convergence Analysis

3.1 Evaluation Criteria

A generally used criterion for classifier evaluation is Jack-knife test. During Jack-knife testing, one sample is left out to be the testing dataset while the remaining is used for training. As a comprehensive cross validation, Jack-knife test is regarded as the most objective criterion.

Moreover, the performance of each experiment is evaluated by other criteria including accuracy (Accu), sensitivity (Sens), specificity (Spec) and Matthew's correlation coefficient (MCC):

$$Sens = \frac{TP}{TP + FN}, \quad Spec = \frac{TN}{TN + FP},$$

$$Avc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

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

where TP means true positive, FN represent false negatives, TN means true negatives, and FP is false positives.

3.2 Comparison with Three Feature Extraction Methods

As depicted in Material and Methods, three feature extractions are used here. First is AAC, second is the clustered dipeptides composition and the third the clustered tripetides compostion.

During the following experiment, the combined feature selection strategies are utilized, i.e., amino acid composition, dipeptides composition and tripeptide composition. The three categories of features are added gradually into the classifier, and obtain increased results. The implicit results could be found in table 2.

Table 2: Comparison of different feature selection methods

	AAC	**AAC +dipeptides**	**AAC +dipeptides +tripeptides**
Dimension	20	36	62
γ	185	130	30
C	5	5	5
TP	54	55	59
TN	109	112	114
FP	20	17	15
FN	29	28	24
Specificity	84.50%	86.82%	88.37%
Sensitivity	65.06%	66.27%	71.08%
Accuracy	74.77%	76.61%	79.36%
MCC	0.4653	0.5038	0.5636

Results in table 2 show that the last classifier achieves the highest accuracy and MCC as well. In detail, the accuracy is 79.36%, and the MCC is 0.5636. These results are reasonable and prove the effectiveness of the classifier.

3.3 Analysis of the Feature Extraction Method

In order to evaluate the rationality of the proposed feature extraction method, a structure based analysis is done by following steps:

First, Anthepro 5.0 [16], a protein analysis soft, is used to analyse the structure of the 87 drought tolerant protein sequences. By doing this, a conserved region compromising 51 peptides from position 755 to 805 is found. This region could be found in Fig. 1.

Second, 16 chosen dipeptides and 26 chosen tripeptide features in conserved region are observed. For dipeptides, over 50% dipeptides come from five groups (the representative elements of these five groups are VV, AV, AA, VA, WV). While for tripeptide, over

50% tripeptides come from eight groups (the representative elements of these eight groups are RAA, AAA, VAA, AAR, VPR, ARA, RAA). These show that features in conserved region present a consistent regularity.

Third, 42 chosen features in conserved region and the whole sequence are compared numerically. For the i-th sequence with whole length, the numerical feature is ($x_{i1}, x_{i2}, ..., x_{i42}$), $i = 1, 2, ..., 87$, variance in the tripeptide features for each sample is computed by

$$\mathrm{var}_i = \sqrt{\frac{\sum_{j=1}^{26}(x_{ij} - Mean_i)}{26}}, \quad (2)$$

where $Mean_i = \frac{\sum_{j=1}^{26} x_{ij}}{26}$. For all of the 87 drought resistant protein sequence with full length, the average of var_i is 0.001743. On the other hand, the average of variance for the 87 short peptides in the conserved region is 0.003759.

Similarly, variance in the dipeptide features is

$$\mathrm{var}_i = \sqrt{\frac{\sum_{j=27}^{42}(x_{ij} - Mean_i)}{16}}, \quad (3)$$

where $Mean_i = \frac{\sum_{j=27}^{42} x_{ij}}{16}$. For sequence with full length, the average of var_i is 0.0056, while for conserved region, it is 0.0085.

```
Matrix PAM100 % of identity for delay 0
Gap opening penalty  10    Hydrophilic gaps    Gap extension penalty  .05
Number of perfect matches 0 identity : . %
 100%  >= 75 >= 50                      106 in sequence N° 1
                      750       760       770       780       790       800       810       820       830       840       850       860
                      |         |         |         |         |         |         |         |         |         |         |         |
   1 UniRef90_A ----------------------GCMSTGAHFNPNNMTHGAPED-----------------------------------------------------------EVRHAGDLGNIIAN-
   2 UniRef90_Q ----------------------GCMSTGPHFNPAGKEHGAPGD-----------------------------------------------------------DNRHAGDLGNITVG-
   3 UniRef90_C -----------MPSFTSKEWSNGSSPSSSSHIDDMLELP-EIDDRCFALPRVN-------------------------------------------------SLQHEEKLTLGGTGN
   4 UniRef90_C ------GVVSSSEQYATQYSNGSSSNSSSSHLDEVLESLPEIDDRCFALPRVNSLRALQQQRH-----------------------------------HQEDTKVGLLQQQQQQGLVAGTG
   5 UniRef90_A ----------------GILHRSGSSSSSSSSEDDGMGGRRKKGMKEKIKEKLP---------------------------------------------------GGHKDNQQHTATG-
   6 UniRef90_P ----------------GILHRSG-SSSSSSSEDDGMGGRRKKGIKEKIKEKLP---------------------------------------------------GGHKDDQHATAT--
   7 UniRef90_Q ----------------RGILHRSGSSSSSSSEDDGMGGRRKKGIKEKIKEKLP---------------------------------------------------GGHGDQQHTGGTYR
   8 UniRef90_Q ----------------GILQRSGSSSSSSSSEDDGMGGRKKKGIKDKIKEKLP---------------------------------------------------GGHGDQQQTAG---
   9 UniRef90_O ------AGVGGTGTGTGILHRSGSSSSSSSEDDGQGGRRKKKGVMQKIKEKLP---------------------------------------------------GGHRQEEQYQSQT-
  10 UniRef90_P ----------------EASDKCQREKRKTINGDDLLWAMATLGFEDYIEPLKV---------------------------------------------------YLQKYREMEGDSK-
  11 UniRef90_Q ----------------EASDKCQKEKRKTVNGDDLLWAMATLGFEDYLEPLKI---------------------------------------------------YLARYRELEGDNK-
  12 UniRef90_O ----------------AVVFPTHIFVANCGDSRAVLCRGKTPLALSVDHKPDRDDEAARIEAA-----------------------------------GGKVIRWNGARVFGVLAMSRSIG
  13 UniRef90_P ----------------AVVFPSHIFVANCGDSRAVLCRGKTALPLSVDHKPDREDEAARIEAA-----------------------------------GGKVIQWNGARVFGVLAMSRSIG
  14 UniRef90_P ----------------AGLAPCLPYLQGRGPLG-GCCGGVKNLLGSAKTTADRKTACTCLKSA-----------------------------------ANAIKGIDLNKAAGIPSVCKVNI
  15 UniRef90_Q ----------------STLAPCLPYLMNRGPLG-GCCGGVKGLLGQAQTTVDRQAACACLKSA-----------------------------------ASSFTDLDLGKAASLPSTCNVNI
  16 UniRef90_Q ----------------SSLAPCISYLTKGGAVPPACCSGVKSLNSAAKTTPDRQAACGCLKSA-----------------------------------YNSISGVNAGNAASFPGKCGVSI
  17 UniRef90_Q ----------------SSLAPCIAYLQKGGTPAPGCCNGVKSLVGAAQTTADKQAVCNCLKGA-----------------------------------AGQVPGLNNQYAQSLPSLCGVNI
  18 UniRef90_Q ----------------SAVGPCLTYARGGAGPSAACCSGVRSLKAAASTTADRRTACNCLKNA-----------------------------------ARGIKGLNAGNAASIPSKCGVSV
  19 UniRef90_Q ----------------EAGGPKLPMKYGRVDVSGPEQCPEEGRLPDAGPPSPADHLRQVFYRM-----------------------------------GLNDKEIVALSGAHTLGRARPDR
  20 UniRef90_Q ----------------EAGGPKLPMKYGRVDVSGPEQCPEEGRLPDAGPPSPADHLRQVFYRM-----------------------------------GLNDKEIVALSGAHTLGRARPDR
  21 UniRef90_Q ----------------VTGGPTIDFVPGRKDS---KVSPKEGGLPDAKQG--VSHLRDIFYRM-----------------------------------GLTDREIVALSGGHTLGRAHPDR
  22 UniRef90_Q --------------ETAARAYDQAAVLMNGQSAKTNFPVIKSNG--SNSLEINSALRSPKS-------------------------------------LSELLNAKLRKNCKDQTP
  23 UniRef90_Q --------------EAAARAYDQAALLMNGQNAKTNFPVVKSEEGSDHVKDVNSPLMSPKS-------------------------------------LSELLNAKLRKSCKDLTP
  24 UniRef90_Q --------------EEAARAYDEAAVLMSGRNAKTNFPLNNNNTGETSEGKTDISASSTMSSS-----------------------------------TSSSSLSSILSAKLRKCCKSPSP
  25 UniRef90_A --------------DMAARAHDVAALTIKGSSAYLNFPELADELPRPASTSPKDIQAAAAK-------------------------------------AAALDFGHRSHDAESELSR
  26 UniRef90_C --------------VAAARAYDTAVFHLRGPSARLNFPELLSQDDDVSTQQQGKMSADSIR-------------------------------------KKATQVGARVDALQTALQ-
  27 UniRef90_A --------------EMAARAHDVAAIALRGRSACLNFADSASRLPVPAT----AEARDIQK-------------------------------------AAAEAAEAFRPGKDDDAVA
  28 UniRef90_A --------------EIAARAHDAAMLALAAGDACLNFADSAELLAVPASYRNLAEVRHAVT-------------------------------------EAVEDFERRQELGEKDALS
  29 UniRef90_D --------------EDAARAYDEAARAMYGALARTNFPAHPAQAPAVALPAAIEGVVRGAS-------------------------------------ASCESTTTSTNHSDVASNL
  30 UniRef90_Q --------------EDAARAYDEAARAMYGALARTNFPVHPAQAPAVAVPAAIEGVVRGAS-------------------------------------ASCESTTTSTNHSDVASSL
  31 UniRef90_O --------------MSEDDLPYQLKITQTGDLQTVGRYDFDGQLKSAMIAHPKLDPVTKELHA-----------------------------------LSYDVVKKPYLKYFRFSPDGVKS
  32 UniRef90_Q --------------MSEDDLPYQVRVTDNGDLETIGRFDFDGQLSSAMIAHPKIDPVTKELFA-----------------------------------LSYDVVKKPYLKYFKFSPEGEKS
  33 UniRef90_Q --------------MSEDDLPYQVQITPNGDLKTVGRFDFDGQLESTMIAHPKVDPESGELFA-----------------------------------LSYDVVSKPYLKYFRFSPDGTKS
  34 UniRef90_Q --------------MSEDDLPYHVRITSNGDLTTVGRYDFNGQLNSTMIAHPKLDPVNGDLHA-----------------------------------LSYDVVQKPYLKYFRFSADGVKS
  35 UniRef90_Q --------------MSEDDLPYHVKVTQTGDLETSGRYDFDGQLKSTMIAHPKIDPETRELFA-----------------------------------LSYDVVSKPYLKYFRFTSDGEKS
  36 UniRef90_Q --------------MSEDDLPYQVKIDGQGDLETIGRFGFDDQIDSSVIAHPKVDATTGDLHT-----------------------------------LSYNVLKKPHLRYLKFNTCGKKT
  37 UniRef90_O --------------LGESDLPYAVRLTESGDIETIGRYDFDGKLAMSMTAHPKTDPITGETFA-----------------------------------FRYGPVP-PFLTYFRFDSAGKKQ
  38 UniRef90_B GRVLKRTELAEGLILEKTSSPLGVLLIVFESRPEALVQIASLAIRSGNGLLLKGGKEAKRSNA-----------------------------------ILHKVITEAIPDSVGEKLIGLVT
  39 UniRef90_B GRVLKKTQLADGLILEKTSSPLGVLLIVFESRPDALVQIASLAIRSGNGLLLKGGKEAKRSNA-----------------------------------ILHKVITEAIPDIVGGKVIGLVT
  40 UniRef90_Q -----------ELTKDAALDPTGLVAVAVAHAFALFVGVSVAANISGG-----------------------------------------------------HLNPAVTFGLAVGGN-
  41 UniRef90_Q -----------KLTDGGATTPSGLVSAALAHAFGLFVAVSVGANISGG-----------------------------------------------------HVNPAVTFGAFIGGN-
  42 UniRef90_B ----------------SPLLPSDPYDRAQARFWADFVDKKVADAASGVWTKKGEELETAK-----------------------------------KDFIAVLKQLQEV-LGEKPYFGGDNF
```

Fig.1: Screenshot from Anthepro 5.0, a conserved region in the drought resistant sequence

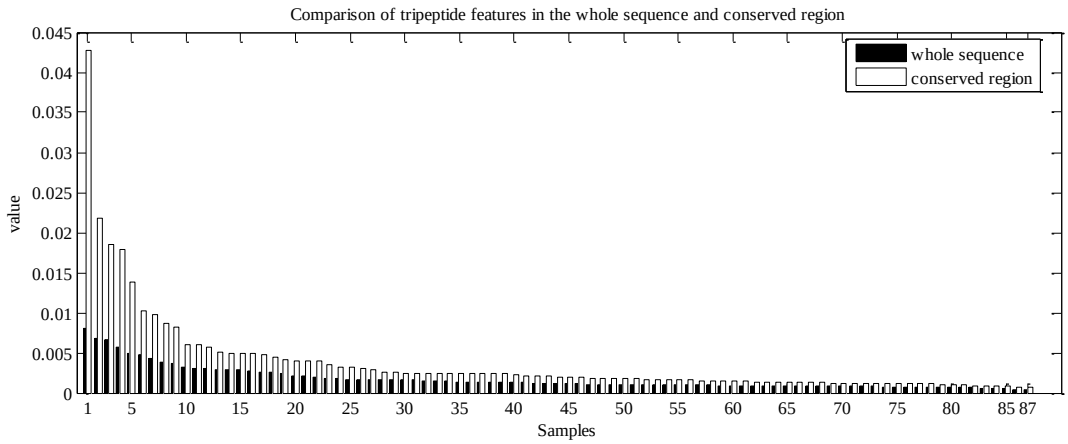


Fig. 2: Comparison of tripeptide features in the whole sequence and conserved region

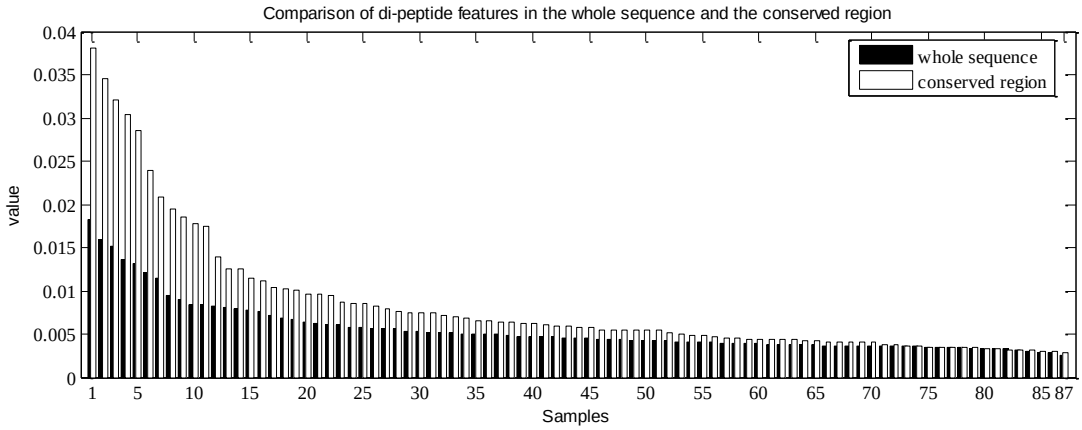


Fig. 3: Comparison of dipeptide features in the whole sequence and conserved region

Furthermore, comparison of the numerical value of the features is presented in Fig. 2 and Fig. 3. Feature values in both figures are sorted in decent order and the value in the conserved region is apparently higher than those with whole length.

All of the comparative results show that numerical features in the conserved region reflect higher consistency and clearer tendency than the whole sequence does. Take into account that the conserved region, as a whole, really should have its unique characteristics, the difference between the two circumstances inturn convince the rationality of the 42 numerical features.

IV. Conclusion

Drought resistance is a complex trait for crops, which might be dominated by a number of factors. Amino acid based analysis is aimed to find the key factor for drought tolerance with a sequence based approach. The focus of this paper is just put here.

With clustering algorithm by using k-means, 62 important features are extracted and used to build an efficient classifier, which achieves a high prediction rate for the specified sample database. The method developed in this research is rooted in the utilization of amino acid information, which in turn proves the effectiveness of the sequence based approach.

Acknowledgment

This paper is partly supported by Discipline Crossing Research Foundation of Huazhong Agricultural University

References

[1] Golbashy, M.; Ebrahimi, M.; Khorasani, S.K.; Choukan, R. Evaluation of drought tolerance of some corn (Zea mays L.) hybrids in Iran. African Journal of Agricultural Research. 2010, 5(19): 2714-2719.

[2] Kandaswamy, K.K.; Chou, K.C.; Martinetz, T.; Möller, S.; Suganthan, P.N.; Sridharan, S.; Pugalenthi, G. AFP-Pred: A random forest

approach for predicting antifreeze proteins from sequence-derived properties. Journal of Theoretical Biology. 2011, 270(1):56-62.

[3] Yuan, Z.M.; Tan, X.S. Nonlinear Screening Indexes of Drought Resistance at Rice Seedling Stage Based on Support Vector Machine. Acta Agronomica Sinica. 2010, 36(7): 1176−1182.

[4] Tuberosa, R.; Salvi, S. Genomics-based approaches to improve drought tolerance of crops. Trends in Plant Science. 2006, 11(8): 405-412.

[5] Yuan, G.F.; Jia, C.G.; Li, Z.; Sun, B.; Zhang, L.P.; Liu, N.; Wang, Q.M. Effect of brassinosteroids on drought resistance and abscisic acid concentration in tomato under water stress. Scientia Horticulturae. 2010, 126(2): 103-108.

[6] Nishimura, N.; Hitomi, K.; Arvai, A.S.; Rambo, R.R.; Hitomi, C.; Cutler, S.R.; Schroeder, J.I; Getzoff, E.D. Structural Mechanism of Abscisic Acid Binding and Signaling by Dimeric PYR1. Science. 2009, 326: 1373-1379.

[7] Salgado, J.C.; Rapaport, I.; Juan A. Asenjo Prediction of retention times of proteins in hydrophobic interaction chromatography using only their amino acid composition. Journal of Chromatography A, 2005, 1098, 1-2(9):44-54.

[8] Zuo, Y.C.; Li, Q.Z. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides, 2009, 30(10): 1788-1793.

[9] Liu, T.; Zheng, X.; Wang, J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. Biochimie. 2010, 92(10): 1330-1334.

[10] Zakeri, P.; Moshiri, B.; Sadeghi, M. Prediction of protein submitochondria locations based on data fusion of various features of sequences. Journal of Theoretical Biology. 2011, 269(1): 208-216.

[11] Shen, H. B.; Chou, K.C. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Analytical Biochemistry. 2008, 373: 386-388.

[12] Li, W.; Jaroszewski, L.; Odzik, G.A. Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics. 2011, 17, 282–283.

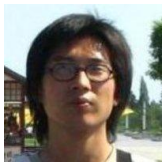
[13] MacQueen, J.B. Some methods for classification and analysis of multivariate observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967, 1: 281-297.

[14] Cortes, C.; Vapnik,V. Support vector networks. Machine Learning. 1995, 20(3): 273-297.

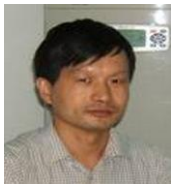
[15] Hopp, T.P.; Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. 1981, 78: 3824-3828.

[16] Delάage, G.; Geourjon, C. ANTHEPROT: A software to display and analyze 3D NMR structures. J. Trace and Microprobe Techniques. 1995, 13: 337-338.

Authors' Profiles



XIA Jing-bo (1979-), male, Wuhan, China, Associate professor, Ph.D., his research directions include data mining, bioinformatics, bioNLP and Thue equations.

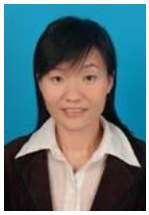


SHI Feng (1966-), male, Wuhan, China, Professor, Ph.D., his research directions include bioinformatics, intelligence computation and computational mathematics.

HU Xue-hai (1980-), male, Wuhan, China, Associate professor, Ph.D., his research directions include bioinformatics and fractal geometry.

LI Zhi (1977-), male, Wuhan, China, Associate Professor, his research directions include data mining and mathematical modeling.

SONG Chao-hong (1970-), female, Wuhan, China, Associate Professor, Ph.D., her research directions include bioinformatics, intelligence computation and computational mathematics.



XIONG Hui-juan (1982-), female, Wuhan, China, instructor, Ph. D., her research directions include numerical optimization and intelligence computation.

