Oryza Vol. 49. No. 4, 2012 (239-245)

Computational approach for the prediction of ERF and DREB proteins in indica rice using support vector machine

N Hemalatha N*, MK Rajesh and NK Narayanan

**Aloysius Institute of Management and Information Technology, St. Aloysius College, Beeri, Mangalore, Karnataka*
[1]Central Plantation Crops Research Institute, Kasaragod 671124, Kerala.
[2]School of Information Science and Technology, Kannur University, Kannur, Kerala

ABSTRACT

Drought and salt stress are considered to be major impediments in rice production systems. To understand the genetics of tolerance to these abiotic stresses and develop drought/salt tolerant cultivars, genomic regions influencing yield and its response to water deficit have to be identified. A method for predicting two drought tolerant proteins viz. dehydration-responsive element binding proteins (DREB) and ethylene responsive factor (ERF) in the genome of indica rice has been described. The proposed method, ERFDREBSVMPRED, was developed using support vector machine and a prediction accuracy of 89% for DREB and 81% for ERF was achieved. The developed tool could predict DREB protein with 100% specificity at a 71% sensitivity rate and ERF protein with 100% specificity at a 60% sensitivity rate.

Key words: *rice, ERF, DREB, protein, support vector machine*

Rice is a staple food for a majority of the world's population mainly in Asia, Africa and India and, it accounts for more than half of the calories consumed. The population in these regions is expected to double during the next 50 years. However, increase in rice productivity is becoming more and more difficult because of the effect of biotic/abiotic stresses like pests, diseases, drought and salinity. There is an urgent need to develop improved rice cultivars which are tolerant to such stresses.

Intensive research has been undertaken in the past few decades to identify drought and salt-responsive mechanisms in plants, both from a biological and genetic perspective. Transcription factors have been found to play a significant role in regulating abiotic-stress responsive gene expression (Sakuma *et al*., 2002). Dehydration-responsive element-binding proteins (DREBs) and ethylene-responsive element (ERE) binding factors are two major subfamilies of the AP2/ ethylene-responsive element-binding protein family, which are known to play decisive roles in the regulation of abiotic- and biotic-stress responses.

DREB (proteins) can regulate the expression of many stress-inducible genes in plants and hence plays a critical role in improving abiotic stress tolerance of plants by interacting with specific cis-acting element named DRE/CRT, which is present in the promoter region of various abiotic stress-related genes. Fine-tuning of ethylene production is significant in developmental processes and in plant responses to stress. Ethylene response factors (ERFs) are plant transcriptional regulators that mediate ethylene-dependent gene expression via binding to the GCC motif found in the promoter region of ethylene-regulated genes. In an earlier work, we had identified 23 novel signature sequences related to ERF family and 21 sequences related to DREB family by carrying out a genome-wide analysis in *Oryza sativa* spp. *indica* (Hemalatha *et al.,* 2011)**.**

Development of a genome-wide prediction tool for ERF and DREB genes will significantly advance rice genome annotation whereby, function can be assigned for a potential gene(s) in the raw sequence(s). In the recent years, adoption of discriminative machine

learning techniques has given a boost to computational gene prediction. Therefore, machine learning algorithms were to predict ERF and DREB genes from whole genomes. We propose a novel gene prediction tool, named ERFDREBSVM$_{PRED}$, for predicting ERF and DREB genes. Models for creating this prediction tool were developed using support vector machine (SVM). Statistical accuracy and prediction of this tool was tested using an independent data test and jackknife validation was carried out for testing whether the data used was biased or unbiased.

MATERIALS AND METHODS

We have utilized 23 ERF and 21 DREB sequences belonging to *indica* rice, based on results obtained from our earlier work (Hemalatha *et al.,* 2011), and also already annotated ERF and DREB sequences from NCBI. Twenty ERF and 19 DREB genes were randomly selected from the original set for creating the positive dataset/training set and the remaining ERF and DREB genes for the creation of negative dataset/test set. For training and testing, we used independent data test, where sequences in the training set and test set are entirely different. For generating features, different window lengths were generated with respect to all four nucleotides (A, T, C and G) (Anwar *et al.,* 2008). The aim of generating window length is to transform the variable length of nucleotide sequences to fixed length feature vectors. This is an important and most crucial step during classification using machine learning techniques because they require fixed length patterns. For generating the fixed length feature vectors, the frequency of 64 features (3-mer), 256 features (4-mer), 1024 features (5-mer) and 4096 features (6-mer) in the given dataset were obtained.

Support vector machine (SVM), a strong machine learning technique for classification, was used in this study. The SVM approach, which was originally introduced by Vapnik and coworkers about two decades ago, is based on the statistical and optimization theory and has been successfully applied in a number of classification and regression problems coworkers (Cortes and Vapnik, 1995; Vapnik 1995). One big advantage of SVM is the scantiness of the solution *i.e.* it separates the hyperplane solely based on the support vectors and not on the complete data set, thereby making it less prone to over-fitting than other classification methods such as the artificial neural networks (Byvatov and Schneider, 2003).

In this study to implement SVM, SVM^{light} package (Joachims, 1999) has been used which allows the user to choose a number of parameters and kernels (*e.g.* linear, polynomial, radial basis function, and sigmoid) or any pre-defined kernel with the assumption that there exists a number of patterns X_i [R^d(i_1, 2, . . . n) with corresponding target values y_i [{target value}. Here the target value is either +1 (representing an ERF/ DREB gene) or -1 (for non ERF/ DREB gene). SVM maps the input vectors x_i into higher dimensional space with minimum error on the training set. The decision function is implemented by SVM using the Equation 1.

$$F(x) = \text{sign}\ (\Sigma y_i\ \alpha_i\ K(x_i x_j + b)) \quad \text{.....(1)}$$

The value of α_i is given by the task of quadratic programming, thus maximizing the subject to 0d"α_id"C where C is the regulatory parameter that controls the trade-off between the margin and the training error, and b is the threshold for defining the hyperplane. The selection of kernel is very important in SVM and is analogous as choosing architecture in artificial neural network. In this study, learning was carried out using three kernels: linear, polynomial, and radial basis function.

In sequence similarity search basic local alignment search tool (BLAST) was used which allows comparing a set of data against a database of sequences and informs if the set of data matches any of the sequences in the database (Altschul *et al*., 1990). Similarity search is conducted by this tool for predicting the function of a given sequence against a database of annotated sequences. In this work, we have conducted a 5-fold cross validation for predicting both ERF and DREB genes, result of which are analyzed in the results section.

In statistical prediction, three methods which are often used to examine a predictor for its effectiveness are independent dataset test, cross validation test and jackknife test (Joachims, 1999). In the independent dataset test, although none of the data to be tested occurs in the training dataset used to train the predictor, the selection of data for the testing dataset could be quite arbitrary. For the cross validation test, 5-fold, 8-fold or 10-fold cross-validation is usually

preferred. The problem with the cross validation is the number of possible selections in dividing a target dataset even for a very simple dataset (Burset and Guigo, 1996, Altschul *et al*., 1990). Therefore, any result by the cross-validation test represents one of many possible results only, and cannot avoid the uncertainty either and hence is used to estimate generalization error. In the jackknife validation, each of the data in the benchmark dataset is in turn singled out as a tested one and the predictor is trained by the remaining ones. During the jackknifing process, both the training dataset and testing dataset are actually open, and a data will in turn move from one to the other. This validation excludes the memory effects during entire testing stage and hence the outcome thus obtained is always unique for a given benchmark dataset. Therefore, of the above three examination methods, the jackknife test is considered the most objective (Tukey,1958; Quenouille,1949) and has been widely recognized and used by investigators to examine the accuracy of various predictors (Chou and Shen, 2008; Chen *et al*., 2008; Chen *et al*., 2009; Chou and Shen, 2010a; Chou and Shen, 2010b). Accordingly, the jackknife test has been used in this study to evaluate our method.

To assess the performance of gene prediction tool, the standard prediction measures by Burset and Guigo were applied (Burset and Guigo, 1996). The following is a brief description of these parameters: (i) The sensitivity or percent coverage of ERF/DREB gene is the percentage of ERF/DREB gene correctly predicted (ii) The specificity or percent coverage of non- ERF/DREB gene is the percentage of non- ERF/DREB gene correctly predicted. (iii) The accuracy is the total number of predictions that were correct. (iv) Precision is the proportion of the predicted positive cases that were correct. These parameters can be calculated using Equations 2–5,

$$\text{Sensitivity} = \frac{TP}{TP+FN}\times 100 \quad \text{.....(2)}$$

$$\text{Specificity} = \frac{TN}{FP+TN}\times 100 \quad \text{.....(3)}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\times 100 \quad \text{.....(4)}$$

$$\text{Precision} = \frac{TP}{TP\ \ FP}\ \ 100 \quad \text{.....(5)}$$

where TP, TN are truly or correctly predicted positive (ERF/DREB) gene and negative (non- ERF/DREB) gene, respectively (Fig. 1) and FP, FN are falsely or wrongly predicted ERF/DREB and non ERF/DREB genes, respectively.

Mathew Correlation Coefficient (MCC) is considered to be the most robust parameter of any class prediction method. An MCC equal to 1 is regarded as a perfect

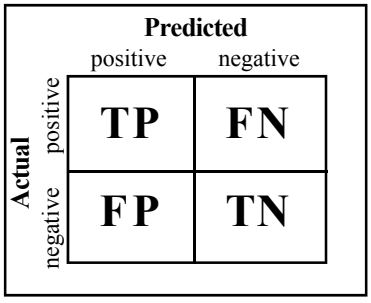

Fig. 1. Criteria of classification of a prediction into true positive (*TP*), true negative (*TN*), false positive (*FP*), or false negative (*FN*). If a positive sample is predicted as positive then it is classified under true positive prediction and vice versa for true negative prediction. But if a positive sample is predicted as negative class and vice versa then it is classified as false negative and false positive prediction, respectively.

prediction where as 0 is for a completely random prediction. The value of MCC ranges from -1 to 1, and a positive MCC value stands for better prediction performance. MCC can be calculated using the Equation 6.

$$\text{MCC}=\frac{(TP\ \ TN)-(FP\ \ FN)}{\sqrt{(TP\ \ FP)(TP\ \ FN)(TN\ \ FP)(TN\ \ FN)}} \quad \text{.....(6)}$$

The prediction performance of ERFDREBSVM$_{PRED}$ for ERF and DREB using independent data test are graphically represented using ROC curves which are used for intuitively visualizing prediction performance. ROC curves plots the true positive rate (TPR) as function of the false positive rate (FPR) which is equal to (1-specificity). The area under the ROC curve is the average sensitivity over all possible specificity values which can be used as a measure of prediction performance at various threshold

values. ROC curves of random predictors will be around the diagonal line from bottom left to top right of the graph with scores around 0.5, while a perfect predictor will produce a curve along the left and top boundary of the square and will receive a value of one.

RESULTS AND DISCUSSION

BLAST which conducts similarity search is one of the common practices for predicting the function of a new sequence against a data base of well annotated sequences. In this study we used BLAST for predicting ERF/DREB genes using 5-fold cross-validation where four sets of ERF/DREB and non- ERF/DREB proteins were used to create a BLAST data base, and ERF/DREB genes of the corresponding test set were searched against this BLAST data base (Table 1, Table 2). This process was repeated five times so that BLAST search was performed once for each ERF/DREB gene. This demonstrates that BLAST alone cannot predict all ERF/DREB genes and hence is not a good method for the annotation of ERF/DREB genes.

The prediction accuracy of the SVM based classifier was assessed by two distinct approaches: cross-validation test and the independent data set test.

Table 2. Result of BLAST search on data set of DREB genes used for $ERFDREBSVM_{PRED}$

Data set	No. of DREB genes	Summary of BLAST hits		Accuracy
		No. of hit	Total hits	
Test1	4	4	0	0
Test2	4	4	0	0
Test3	4	4	0	0
Test4	4	4	0	0
Test5	5	5	0	0
Average/Total	21	21	0	0

Table 1. Result of BLAST search on data set of ERF genes used for $ERFDREBSVM_{PRED}$

Data set	No. of ERF genes	Summary of BLAST hits		Accuracy
		No. of hit	Total hits	
Test1	5	5	0	0
Test2	5	3	2	40%
Test3	5	4	1	20%
Test4	4	5	0	0
Test5	4	5	1	20%
Average/Total	23	22	3	13%

We carried out independent data test with different kernels and parameters of support vector machine (SVM) with window length varying from 3 to 6. Testing of SVM on independent data test for ERF gene resulted in the achievement of 81% accuracy with an MCC value of 0.67 using linear and polynomial kernel with window length 3 where sensitivity is 60% and specificity is 100%. On carrying out independent data test for DREB gene, we obtained an accuracy of 89% with an MCC value of 0.78 using polynomial kernel with window length 3 and having sensitivity and specificity as 71%

Table 3. Classification accuracy of three kernels using SVM^{light} with independent data set (ERF)

Algorithm	Window length	Independent data test					Jackknife validation				
		Sn	Sp	Acc	Prec	MCC	Sn	Sp	Acc	Prec	MCC
Linear	3	60	100	81	100	0.67	100	100	100	100	100
	4	20	100	63	100	0.35	100	100	100	100	100
	5	40	100	72	100	0.52	97	100	98	100	98
	6	40	100	72	100	0.52	100	100	100	100	100
Polynomial	3	60	100	81	100	0.67	100	100	100	100	100
	4	20	100	63	100	0.35	100	100	100	100	100
	5	20	100	63	100	0.35	100	100	100	100	100
	6	0	100	55	0	0	100	100	100	100	100
RBF	3	100	0	45	45	0	100	100	100	100	100
	4	100	0	45	45	0	100	100	100	100	100
	5	100	0	45	45	0	100	100	100	100	100
	6	100	0	45	45	0	100	100	100	100	100

and 100%, respectively (Table 3, Table 4). Hence it can be concluded that for independent data test window length 3 is optimal with polynomial kernel as the classifier. 8-fold and 10-fold cross validation was carried out on the same dataset for window length varying from 3 to 6. It shows average 8-fold and 10-fold cross validation result for DREB gene with polynomial kernel and a window length of 5 as the best with accuracy of 84.5%. Similarly, average cross validation result of 8-fold and 10-fold for ERF gene shows that polynomial kernel and a window length of 4 as the best result with accuracy of 66.5% (Tables 5, 6). The performance comparison of both the approach (Fig. 2 and 3).

On performing jack-knife validation for the datasets used for cross validation and independent data tests for ERF and DREB genes; it was observed that all values were greater than 95%. These results shows that data used for the training are totally unbiased.

A plot of ROC curve is a measure that depicts the relationship between sensitivity and specificity of a given class. To evaluate the best classifier for ERF and DREB, we plotted ROC curves for both proteins on the independent test performance. Figure 4 shows the ROC curve drawn for DREB protein for the best classifier's performance which was obtained for

Table 4. Classification accuracy of three kernels using SVM^{light} with independent data set (DREB)

Algorithm	Window length	Independent data test					Jackknife validation				
		Sn	Sp	Acc	Prec	MCC	Sn	Sp	Acc	Prec	MCC
Linear	3	71	92	84	83	0.65	100	100	100	100	100
	4	71	58	63	50	0.28	100	100	100	100	100
	5	57	58	58	44	0.14	97	100	98	100	98
	6	57	58	58	44	0.14	100	100	100	100	100
Polynomial	3	71	100	89	100	0.78	100	100	100	100	100
	4	71	58	63	50	0.28	100	100	100	100	100
	5	42	58	52	38	0.01	100	100	100	100	100
	6	0	100	63	0	0	100	100	100	100	100
RBF	3	100	0	37	36	0	100	100	100	100	100
	4	100	0	36	36	0	100	100	100	100	100
	5	100	0	36	36	0	100	100	100	100	100
	6	100	0	36	36	0	100	100	100	100	100

Table 5. Prediction performance of ERFDREBSVM$_{PRED}$ on ERF proteins with different kernels using cross validation

Algorithm	Window Length	8- fold validation					10- fold validation					Avg Acc
		Sn	Sp	Acc	Prec	MCC	Sn	Sp	Acc	Prec	MCC	
Linear	3	70	30	54	57	0	81	50	65	71	0.3	59.5
	4	75	54	65	59	0.3	67	56	62	71	0.26	63.5
	5	77	65	71	73	0.4	79	56	69	73	0.38	70
	6	71	56	65	72	0.26	83	69	76	78	0.56	70.5
Polynomial	3	75	56	67	67	0.3	75	50	60	58	0.2	63.5
	4	67	46	58	55	0.1	71	81	75	88	0.57	66.5
	5	44	77	59	58	0.2	48	88	66	75	0.38	62.5
	6	0	91	40	0	-0.08	0	94	44	0	-0.07	42
RBF	3	100	0	60	60	0	100	12	60	58	0.1	60
	4	95	0	58	59	-0.5	100	12	61	59	0.1	59.5
	5	100	31	55	54	0.04	94	12	55	55	0.08	55
	6	100	0	56	56	0	100	0	51	51	0	53.5

MCC - Mathew correlation co-efficient, Sn - Sensitivity Percentage, Sp-Specificity, ACC - Accuracy precision

Table 6. Prediction performance of ERFDREBSVM$_{PRED}$ on DREB proteins with different kernels using cross validation

Algorithm	Window Length	8- fold validation					10- fold validation					Avg Acc
		Sn	Sp	Acc	Prec	MCC	Sn	Sp	Acc	Prec	MCC	
Linear	3	100	0	56	56	0	62	31	48	49	-0.06	52
	4	48	42	45	43	-0.07	78	56	68	64	0.35	56.5
	5	67	52	59	68	0.2	81	88	84	88	0.69	71.5
	6	85	50	69	71	0.38	69	63	64	68	0.36	66.5
Polynomial	3	100	0	56	56	0	62	50	56	50	0.13	56
	4	52	39	47	42	-0.06	59	69	67	55	0.3	57
	5	63	75	80	48	0.44	100	83	89	81	0.82	84.5
	6	0	79	38	0	-0.21	0	75	40	0	-0.27	39
RBF	3	100	0	56	56	0	100	0	51	51	0	53.5
	4	100	12	58	56	0.15	100	0	53	53	0	55.5
	5	100	19	61	59	0.22	100	19	63	59	0.22	62
	6	100	12	61	60	0.15	100	0	52	52	0	56.5

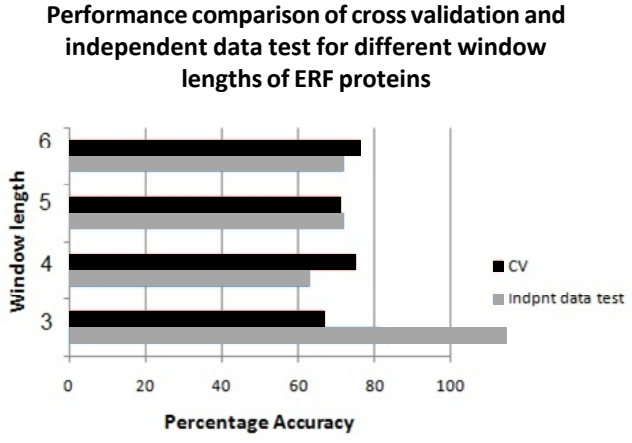

Fig. 2. Performance comparison of ERF proteins with respect to cross validation and independent data test

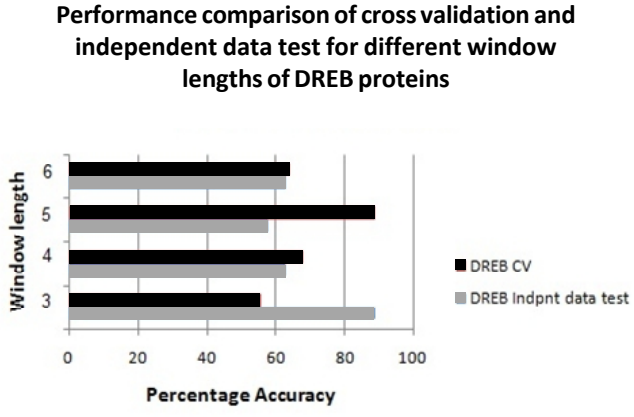

Fig. 3. Performance comparison of DREB proteins with respect to cross validation and independent data test

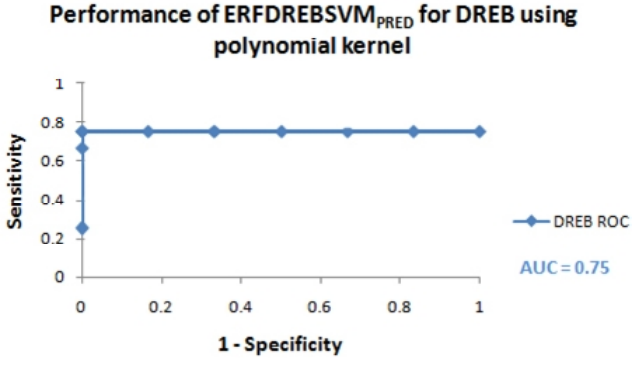

Fig. 4. ROC curve for ERFDREBSVM$_{PRED}$ for DREB protein validated using independent data test

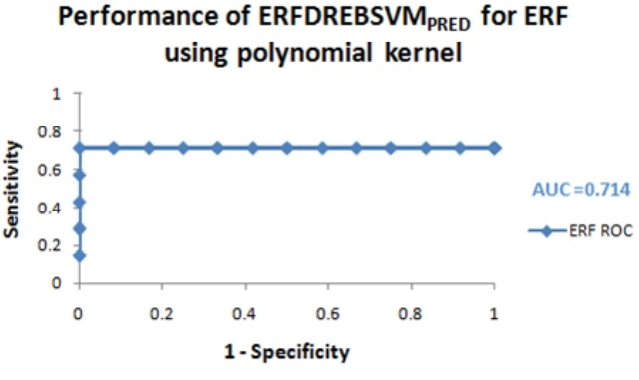

Fig. 5. ROC curve for ERFDREBSVM$_{PRED}$ for ERF protein validated using independent data test

polynomial kernel. The ROC curve drawn for an ERF protein which obtained an accuracy of 81% for polynomial kernel is depicted in the figure 5. Each point on the curve is plotted based on different threshold values. The ROC curve for a perfect classifier results in a straight line up to the top left corner and then straight to the top right corner. Accordingly for both DREB and ERF, the ROC depicted "good classification" with area under the curve (AUC) 0.75 and 0.714 respectively. The AUC specifies the probability that the decision function assigns a higher value to the positive than to the negative sample when one positive and a negative sample are drawn at random.

Thus, in this study, a novel and systematic method has been described for developing a tool for prediction of *ERF* and *DREB* genes using support vector machine. The performance of the developed tool was validated using various statistical parameters and it was found to be highly satisfactory. The method can be utilized for automatic annotation of genomic data.

REFERENCES

Altschul S, Gish W, Miller W, Myers E and Lipman D 1990. Basic local alignment search tool. J Mol Biol 215: 403-410.

Anwar F, Baker SM, Jabid T, Mehedi Hasan M, Sohail M, Khan H and Walshe R 2008. Pol II promoter prediction using characteristic 4 mer motifs: a machine learning approach. BMC Bioinf 9:414-418.

Burset M and Guigo R 1996. Evaluation of gene structure prediction programs. Genomics 34:353-367.

Byvatov E and Schneider G 2003. Support vector machine applications in bioinformatics. Appl Bioinf 2:67-77.

Chen C, Chen LX, Zou XY and Cai PX 2008. Predicting protein structural class based on multi-features fusion. J Theor Biol 253:388-392.

Chen C, Chen L, Zou X and Cai P 2009. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Peptide Lett 16:27-31.

Chou KC and Shen HB 2008. Cell-PLoc: a package of Web servers for predicting sub-cellular localization of proteins in various organisms. Nat Protoc 3:153-162.

Chou KC and Shen HB 2010. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites. PLoS ONE 5: e9931.

Cortes C and Vapnik V 1995. Support vector networks. Mach Learn 20: 273-293.

Hemalatha N, Rajesh MK and Narayanan NK 2011.Genome-wide Analysis of Putative ERF and DREB Gene Families in Indica Rice (O. sativa L. subsp. indica). Int J Mach Learn Comput 2(5):556-559.

Joachims T 1999. Making large-scale SVM learning practical. In pages 169-184 Advances in Kernel Methods - Support Vector Learning. Scholkopf B, Burges CJC and Smola, AJ. (Eds.). Cambridge MA; MIT Press.

Quenouille M 1949. Approximate tests of correlation in time series. J Roy Stati Soc : Series B 11:18-84.

Sakuma Y, Liu Q, Dubouzet JG, Abe H, Shinozaki K and Yamaguchi-Shinozaki K 2002. DNA-binding specificity of the ERF/AP2 domain of Arabidopsis DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. Biochem Biophys Res Commun 290:998-1009.

Tukey JW 1958. Bias and confidence in not-quite large samples. Ann Math Stat 29:614.

Vapnik V 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York.