Classification of cereal proteins related to abiotic stress based on their physicochemical properties using support vector machine

Manju Mary Paul, Anil Rai and Sanjeev Kumar*

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi 110 012, India

Abiotic stress factors severely limit plant growth and development as well as crop yield. There is a great need to develop understanding of plant physiological responses to abiotic stresses in order to improve crop productivity through crop improvement programmes. Proteins play a central role in plant adaptations under stress and hence their identification is important to the biologist. Identification of such proteins by wet lab experimentation is sometimes expensive and time-consuming. In such a situation, *in silico* approaches can be used to narrow down this search. In this study, classification of cereal proteins subjected to four different stresses, namely, extreme temperature, drought, salt and abscisic acid (ABA) was undertaken. Classification models were built using support vector machine (SVM) to predict the function of proteins under these abiotic stresses on the basis of 34 physicochemical features extracted from the protein sequence. Specific features of the protein sequence that are highly correlated with certain protein functions were selected by stepwise logistic regression, a feature selection method. SVM was trained using different kernel functions and cross-validated using 10-fold cross-validation technique. Prediction precision was assessed through different measures such as sensitivity, specificity and accuracy. The accuracy of protein function prediction using SVM with different kernel functions ranges from 60% to 100%.

Keywords: Abiotic stress, cross-validation, physicochemical properties, proteins, support vector machine.

ABIOTIC stress has negative impact on growth and productivity of crops. Abscisic acid (ABA), drought, heat and salinity are among major abiotic stresses of plants[1]. Abiotic stress causes series of morphological, physiological, biochemical and molecular changes which are not favourable for plant growth. These stress conditions are interrelated and induce cellular damage in plant either independently or in combination. ABA is the central regulator of many plant responses to environmental stress, particularly stress related to osmotic regulation. ABA is produced in plants under water deficit and high salinity conditions and plays an important role in stress response[2]. Worldwide, drought is one of the most serious abiotic stresses to agricultural crops. It is associated with reduced water availability and cellular dehydration in plants. Therefore, there are changes in cellular metabolism associated with an osmotic adjustment. Heat stress is associated with an enhanced risk of improper protein folding and denaturation of several intracellular protein and membrane complexes. It leads to reduction in the duration of developmental phases causing development of fewer organs/smaller organs, reduced light perception over the shortened life cycle and perturbation of processes related to carbon assimilation which are responsible for significant yield losses in cereals[3]. Salt stress is responsible for low agricultural production in several hot and dry semi-arid regions, where, agriculture is dependent on irrigation[4]. Plants, as sessile organisms, often have to cope with multiple environmental stresses and in order to mitigate these stresses, most plants employ complex regulatory mechanisms to trigger effective responses against various abiotic stresses.

Plants have special physiological mechanisms by which tolerance against different stresses is expressed. These mechanisms are regulated by a number of genes or proteins. Hence, it is important to identify these genes/proteins involved in various plant stress responses. However, identification of genes/proteins, which are important for these abiotic stresses, by wet lab experimentation is expensive and time-consuming. Therefore, *in silico* approaches are used to narrow down this search and then wet lab experimentations are used for validation.

Computational approaches are used for predicting and classifying unknown proteins into their functional groups in a cost-effective way[5]. *In silico* classification of proteins based on their functional trait can be done using their physicochemical properties derived from the sequences. Large number of physicochemical properties can be derived from a protein sequence. These are often interrelated; therefore, it is important to select only important properties (i.e. features) through an appropriate feature selection procedure. Feature selection reduces the dimensionality of data by selecting only a subset of measured features (predictor variables) to develop a

*For correspondence. (e-mail: sanjeevk@iasri.res.in) classification model. Selection criteria usually involve the minimization of a specific measure of predictive error for models fit to different feature subsets. Algorithms search for a subset of predictors that optimally model measured responses, subject to some constraints. A good feature subset is one that contains features highly correlated with the response class, yet uncorrelated with each other. Also, classification and prediction performance can be improved by avoiding overfitting through feature selection procedure. There are different methods for feature selection such as correlation-based feature selection, Markov blanket filter, fast correlation-based feature selection, sequential forward selection (SFS), sequential backward elimination (SBE)[6], step-wise logistic regression, etc.

Many linear and nonlinear statistical techniques are available for binary classification such as discriminant analysis (DA), logit or probit models, neural networks, random forest, etc. However, support vector machine (SVM) is a promising nonlinear, non-parametric classification technique, which has already been applied to number of scientific datasets for classification. It has theoretical advantages over other machine learning methods as it is based on fewer assumptions and is capable of discovering nonlinear separating boundaries between classes. SVM was first proposed by Vapnik[7] and attracted a high degree of interest in the machine learning research community. Further, SVMs simultaneously minimize the empirical classification error and maximize the geometric margin and works on the principle of structural risk minimization (SRM). It maps input vector to a higher dimensional space, where a maximal separating hyperplane is constructed to separate the data. SVM has been employed in the classification of genes in various microarray experiments[8,9] and also used in the classification of proteins based on various physicochemical properties[5]. Though primarily it has been designed for binary classification, subsequently other variations were developed to extend this to the problem of multi-class classification[10,11]. To build confidence on the classifier, estimation of the error is a critical step. There are number of cross-validation techniques used for this purpose such as re-substitution validation, hold-out validation, leave-one-out cross-validation, 10-fold cross-validation, bootstrap cross-validation (BCV), leave-one-out bootstrap (LOOBT), BT632, BT632+, etc.[12].

In this study, classifiers were built for *in silico* classification of proteins related to major abiotic stresses of cereals such as ABA, drought, heat and salt using SVM. These SVM classifiers were trained using three different kernel functions, i.e. polynomial, radial and sigmoid. The prediction accuracy of classifiers, with respect to abiotic stresses and kernel functions, was estimated using 10-fold cross-validation technique. Performance of classifiers for classification of proteins related to each abiotic stress was found to be quite satisfactory.

Materials and methods

Protein sequences from the Poaceae family which are responsible for regulation of four different stresses, i.e. ABA, drought, heat and salt were downloaded from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/). These proteins were either upregulated or downregulated in response to abiotic stresses undertaken. The upregulated and down-regulated protein sequences were named as positive and negative proteins respectively, and considered as two classes in this analysis for each of the abiotic stresses. These two classes of the protein sequences, under each stress, were further subdivided into two parts randomly. The first subpart, two-thirds of the sequence, have been considered as training set and rest one-thirds of the sequences are considered as test set. Sample size for these sub-categories is given in Table 1.

Feature selection from protein sequences

Physicochemical properties of the protein sequences are useful in providing insight into the structural and functional behaviour of a molecule. In order to extract physicochemical features, Protparam tool was used. It computes various physicochemical properties from protein sequences. The parameters computed by ProtParam include molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). To reduce the dimensionality of the data, feature selection procedure was carried out using stepwise logistic regression (LR). Logistic regression model is given by

$$\pi(x) = \frac{\exp[\beta_0 + \boldsymbol{\beta} x]}{1 + \exp[\beta_0 + \boldsymbol{\beta} x]}, \tag{1}$$

where $\pi(x)$ = probability of outcome, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ are the parameters of the logistic model. $\mathbf{X}$ is an $n \times p$ matrix pertaining to n proteins; whereas, p is total number of features included in the model. LR computes maximum

Table 1. Sample size of protein sequences for positive and negative regulation of different abiotic stresses

	Training set		Test set		
Protein class	Positive	Negative	Positive	Negative	Total
ABA	1,737	2,093	869	1,046	5,745
Drought	369	469	184	235	1,257
Heat	57	41	28	21	147
Salt	3,063	2,093	1,532	1,046	7,734
Total	5,226	4,696	2,613	2,348	14,883

likelihood estimates of parameters of the logistic model. Stepwise LR enters independent (predictor) features in a stepwise manner. Feature selection was carried out by stepwise logistic regression using 'proc logic' of SAS ver. 9.3 software. Independent variables were, in stepwise manner, introduced into the model, evaluated, and then retained or discarded based on their significance in the overall model. Features are added to the logistic regression equation one at a time based on statistical criterion of reducing the –2 log Likelihood (–2 log L) errors for the included features. Details for calculating –2 log L is given in eq. (2). The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1), the likelihood of obtaining the data evaluated at the maximum likelihood estimate (MLE) of the parameter over the maximized value of the likelihood function for the simpler model (L_0), and the likelihood of obtaining the data when the parameter is zero. After adding each feature, the model is tested for its inclusion/exclusion or exclusion of other features which are present in the model. This process of inclusion and exclusion of features stop at the point when it is not possible to reduce –2 log L statistically. This procedure helps in identification of important physicochemical features of proteins related to a particular abiotic stress.

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)]. \tag{2}$$

After identification of important features of proteins related to a particular abiotic stress, classifier has been developed based on SVM.

Support vector machine

A support vector machine constructs a hyperplane which separates two different groups of feature vectors with a maximum margin. The hyperplane is constructed through learning from the training datasets by minimizing $||\mathbf{w}||^2$ and estimating the model parameters **w** and b that satisfy the following conditions

$$\mathbf{x}_\mathrm{i} \cdot \mathbf{w} + b \geq 1 \quad \text{for} \quad y_i = +1, \tag{3}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad \text{for} \quad y_i = -1, \tag{4}$$

where y_i is the class index which indicates the two classes of proteins (upregulated and downregulated) for each stress. Here, the values of y_i are taken as 1/–1 for positive/negative protein classes respectively, $\mathbf{x}_i$ represents feature vector with physicochemical descriptors of a protein as its elements. **w** is a vector normal to the hyperplane, $|b|/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin and $||\mathbf{w}||^2$ is the Euclidean norm of **w**. With the estimates of **w** and b, a given vector **x** can be classified by

$$f_{\mathrm{w,b}}(x) = \mathrm{sgn}(\mathbf{x}_i \cdot \mathbf{w} + b). \tag{5}$$

Further, SVM maps the input variable into a high-dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ in case of nonlinear relationship between predictor and response variables. SVM is applied to this feature space and then the decision function can be written as

$$f(x) = \mathrm{sgn}\left(\sum_i \alpha_i y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x})_i + b^*\right), \tag{6}$$

where the coefficients α_i and b^* are determined by maximizing the following Langrangian expression

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i \cdot \mathbf{x}_j) \tag{7}$$

under conditions

$$\alpha_i \geq 0 \quad \text{and} \quad \sum_i \alpha_i y_i = 0.$$

A positive or negative value from eqs (5) or (6) indicates that **x** belongs to the positive or negative class respectively.

Kernel selection of SVM

The training vectors $\mathbf{x}_i$ of observations are mapped into a higher (maybe infinite) dimensional space by the function ϕ. The SVM then finds a linear separating hyperplane with the maximal margin in this higher dimension space, with a penalty parameter $C > 0$ of the error term. $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is called the kernel function. Here, following popular kernel functions have been used to fit the nonlinear SVM.

Polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + r)^d \qquad \gamma > 0 \tag{8}$$

RBF kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \,||\, \mathbf{x}_i - \mathbf{x}_j \,||^2) \qquad \gamma > 0 \tag{9}$$

Sigmoid kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tan h(\gamma \mathbf{x}_i^T \mathbf{x}_j + r)\ \gamma > 0. \tag{10}$$

CRAN package: e1071 version 1.6-2 of R software has been used for modelling SVM and a computer program has been developed in R[14,15]. The training of the support vector machine has been done by using subroutine 'svm' of e1071 package on training data set. Function ‘tune’, which tunes the parameters of kernel functions using a grid search over supplied parameter ranges, was used to obtain an optimal value of parameters of kernel functions. Different parameters for a kernel functions and grid space of search for optimal values are given in Table 2.

Estimation of prediction errors

Performances of developed models were assessed through numbers of available measures such as sensitivity, specificity and accuracy, which are defined as

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}), \tag{11}$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}), \tag{12}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{13}$$

where, true positives (TP) and the true negatives (TN) were correct predictions for proteins, which belong to the stress class and proteins which do not belong to that stress class respectively. A false positive (FP) occurs, when a protein belonging to non-stress class is predicted in stress class and a false negative (FN) occurs, when, a protein belonging to stress class is predicted in non-stress class. Sensitivity measures the proportion of actual positives, which are correctly identified as such, i.e. it is defined as the proportion of the proteins belonging to the class and is predicted rightly. Also, specificity measures the proportion of negatives which are correctly identified. However, it does not provide information about a protein which actually belongs to a class, it is predicted under non-class. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data.

Further, prediction errors are estimated using 10-fold cross-validation. In 10-fold cross-validation, the training data set for each stress was randomly partitioned into 10 subsamples. Out of 10 subsamples, a single subsample is retained for validation of the model, and the remaining 9 subsamples are used for training the models. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The above statistics were calculated for each fold and results were averaged to produce a single estimate.

Table 2. Parameters of kernel functions and grid space of search

Degree	[1 : 5]	Parameter needed for kernel of type polynomial (*d*)
Gamma	[–1 : 1]	Parameter needed for all kernels except linear (γ)
Coef ( )	[0 : 2]	Parameter needed for kernels of type polynomial and sigmoid (*r*)
Cost	[0–5]	Constant of the regularization term in the Lagrange formulation (*C*)

Results and discussion

In this study, important physicochemical features of proteins for each abiotic stress were selected through stepwise logistic regression. Summary of statistics of stepwise logistic regression for salt, drought, ABA and heat stress is given in Tables 3–6 respectively. It was observed from Table 3 that the values of –2 log *L* decrease with inclusion of new feature at every step in stepwise logistic regression until the 22nd step, where it was found to be 5780.468, which is higher than the previous step and there was further decrease in the value of –2 log *L* with inclusion of new features. Therefore, with the removal of composition of *Q*, the process of feature selection was terminated at this stage and only 20 important features were included in the model. Similarly, pursuing Tables 4–6 reveals that until step 17, 11 and 5, the values of –2 log *L* decreaseed with inclusion of new features. Thus, process of feature selection was terminated after steps 17, 11 and 5 and total of 13, 9 and 5 important features were retained for drought, ABA and heat respectively. Support vector classifiers for four abiotic stresses were trained on training data sets with respective selected features by using three different kernel functions. These trained models are nothing but the two parallel hyperplanes obtained through process of optimization under constrained conditions. The training data points which lie on these hyperplanes are called support vectors. The number of support vectors in each case is given in Table 7.

Performance assessment with different kernel functions

Estimates of prediction error (%) of the SVM-based classifiers, with three different kernel functions, for all the four stresses were obtained through 10-fold cross-validation. Results are presented in Table 8 and it can be seen that prediction error for predicting ABA stress is minimum (around 0.4%) for all kernel functions followed by drought and salt stresses (around 5%). The highest prediction error was found in case of heat stress (from 12–14%), which may be due to small size of sample.

Performance of the support vector classifiers with three different kernel functions was further evaluated on test data set. Measures of evaluation such as accuracy, sensitivity and specificity were calculated based on the values of TP, TN, FP and FN. Calculated values of these measures were given in Table 7 and used for comparative evaluation. It can be seen from Table 7 that in case of salt, performance of radial is best followed by polynomial and sigmoid; whereas in case of ABA, radial and polynomial function performs equally well and are better than sigmoid kernel. For drought stress, overall performance of radial and polynomial was better than sigmoid function. In case of heat, sigmoid kernel performs better than

Table 3. Feature selection through stepwise logistic regression for salt stress

	Effect						
Step	Entered	Removed	DF*	Number in	Score chi-square	Pr **> ChiSq	–2 log *L*
1	Instability index		1	1	1538.6396	<0.0001	8492.989
2	Theoretical pI		1	2	977.6641	<0.0001	7445.325
3	Composition of *S*		1	3	503.2934	<0.0001	6923.619
4	Number of sulphur atoms		1	4	306.5713	<0.0001	6578.392
5	Number of nitrogen atom		1	5	124.3727	<0.0001	6456.570
6	Composition of *I*		1	6	103.9651	<0.0001	6350.993
7	Number of carbon atoms		1	7	227.4993	<0.0001	6131.524
8	Half life		1	8	98.8909	<0.0001	6026.403
9	Composition of *K*		1	9	48.5245	<0.0001	5977.178
10	Composition of *W*		1	10	38.1635	<0.0001	5938.713
11	Composition of *F*		1	11	29.3935	<0.0001	5909.103
12	Composition of *Q*		1	12	16.5212	<0.0001	5892.588
13	Number of positive amino acid		1	13	16.3233	<0.0001	5878.072
14	Composition of *M*		1	14	14.3775	0.0001	5863.935
15	Composition of *D*		1	15	16.2535	<0.0001	5847.539
16	Composition of *C*		1	16	18.2926	<0.0001	5829.332
17	Composition of *V*		1	17	11.4089	0.0007	5817.940
18	Composition of *Y*		1	18	13.5471	0.0002	5803.911
19	Composition of *A*		1	19	10.3869	0.0013	5793.558
20	Composition of *P*		1	20	8.5904	0.0034	5784.764
21	Composition of *E*		1	21	6.9706	0.0083	5777.793
22		Composition of *Q*	1	20	2.6703	0.1022	5780.468

*DF, Degrees of freedom; **Pr, Probability.

Table 4. Feature selection through stepwise logistic regression for drought stress

	Effect						
Step	Entered	Removed	DF*	Number in	Score chi-square	Pr **> ChiSq	–2 log *L*
1	Composition of *N*		1	1	307.2251	<0.0001	1355.148
2	Number of positive amino acid		1	2	161.9058	<0.0001	1172.919
3	Instability index		1	3	188.4156	<0.0001	960.916
4	Composition of *Y*		1	4	56.8022	<0.0001	901.223
5	Composition of *P*		1	5	37.2754	<0.0001	859.729
6	Composition of *Q*		1	6	33.8179	<0.0001	830.219
7	Composition of *W*		1	7	23.3684	<0.0001	805.164
8	Composition of *L*		1	8	20.5085	<0.0001	783.793
9	Aliphatic index		1	9	22.2672	<0.0001	761.930
10	Composition of *R*		1	10	15.3094	<0.0001	744.983
11	Composition of *T*		1	11	17.7495	<0.0001	728.240
12	Number of negative amino acid		1	12	7.8604	0.0051	720.372
13	Composition of *D*		1	13	6.4826	0.0109	713.895
14		Number of positive amino acid	1	12	0.3604	0.5483	714.255
15		Composition of *Q*	1	11	3.1613	0.0754	717.298
16	Half life		1	12	5.0289	0.0249	712.215
17	Composition of *A*		1	13	4.5254	0.0334	707.664

radial and polynomial. It can be seen from the Table 7 that sensitivity of radial kernel ranges from 79% to 100% for different stresses. For polynomial kernel, this range varies from 79% to 99.5%; whereas sensitivity for sigmoid kernel ranges from 53% to 93%. Among these three kernel functions, radial kernel has highest sensitivity. A model with 100% sensitivity means that it is capable of predicting all actual positives. Further, from this table, it can be seen that radial function has specificity ranging from 80% to 98%. Specificity of polynomial function ranges from 49% to 99.5%; while for sigmoid kernel function, it ranges from 20% to 100%. Accuracy of radial function ranges from 84% to 99.4%. For polynomial kernel function, accuracy ranges from 74% to 99.5%. Sigmoid kernel accuracy ranges from 39% to 96%. Therefore, the performance of radial function with respect to specificity is the best as it can predict actual negatives more accurately. Overall accuracy of radial kernel function is found

Table 5. Feature selection through stepwise logistic regression for ABA stress

	Effect						
Step	Entered	Removed	DF*	Number in	Score chi-square	Pr **> ChiSq	–2 log *L*
1	Number of negative amino acid		1	1	1747.9651	<0.0001	5464.306
2	Molecular weight		1	2	3732.4232	<0.0001	1148.043
3	Composition of *E*		1	3	378.5988	<0.0001	710.850
4	Composition of *D*		1	4	152.7613	<0.0001	536.503
5	Number of positive amino acid		1	5	115.1702	<0.0001	426.639
6	Amino acid number		1	6	236.2700	<0.0001	346.525
7		Molecular weight	1	5	0.2028	0.6525	346.738
8	Composition of *C*		1	6	29.7063	<0.0001	323.389
9	Composition of *G*		1	7	15.9303	<0.0001	304.973
10	Composition of *T*		1	8	13.3642	0.0003	292.156
11	Composition of *M*		1	9	6.2321	0.0125	287.821

Table 6. Feature selection through stepwise logistic regression for heat stress

	Effect						
Step	Entered	Removed	DF*	Number in	Score chi-square	Pr **> ChiSq	–2 log *L*
1	Gravy		1	1	42.6819	<0.0001	151.019
2	Composition of *W*		1	2	24.2972	<0.0001	123.383
3	Composition of *Y*	–	1	3	13.7552	0.0002	110.346
4	Composition of *D*		1	4	5.9772	0.0145	103.911
5	Composition of *P*		1	5	7.2325	0.0072	95.979

Table 7. Performance evaluation of support vector classifiers using different kernel functions

				Test data set						
			Number of	Positive		Negative		Accu-	Sensiti-	Specifi-
Abiotic stress	Kernel function	Penalty and kernel function parameters	support vectors	TP	FN	TN	FP	racy (%)	vity (%)	city (%)
Salt	Radial	$C = 1$, $\gamma = 0.05$	2062	1453	79	833	213	87	95	80
	Polynomial	$C = 1$, degree = 3, $\gamma = 0.05$, coef. 0 = 0	3153	1395	137	514	532	74	91	49
	Sigmoid	$C = 1$, $\gamma = 0.05$, coef. 0 = 0	3331	809	723	208	838	39	53	20
ABA	Radial	$C = 1$, $\gamma = 0.111$	435	869	0	1035	11	99.4	100	98
	Polynomial	$C = 1$, degree = 3, $\gamma = 0.111$, coef. 0 = 0	1142	865	4	1041	5	99.5	99.5	99.5
	Sigmoid	$C = 1$, $\gamma = 0.111$, coef. 0 = 0	462	756	113	973	73	90	87	93
Drought	Radial	$C = 1$, $\gamma = 0.077$	287	162	22	197	38	86	88	84
	Polynomial	$C = 1$, degree = 3, $\gamma = 0.077$, coef. 0 = 0	462	145	39	224	11	88	79	95
	Sigmoid	$C = 1$, $\gamma = 0.077$, coef. 0 = 0	324	127	57	198	37	78	69	84
Heat	Radial	$C = 1$, $\gamma = 0.2$	44	22	6	19	2	84	79	90
	Polynomial	$C = 1$, degree = 3, $\gamma = 0.2$, coef. 0 = 0	54	26	2	11	10	75.5	93	52
	Sigmoid	$C = 1$, $\gamma = 0.2$, coef. 0 = 0	46	26	2	21	0	96	93	100

Coef., Coefficient.

Table 8. Estimates of error (%) using 10-fold cross-validation technique

Stress	Radial	Polynomial	Sigmoid
Salt	5.0	5.2	5.1
ABA	0.1	0.2	0.4
Drought	3.7	4.1	3.8
Heat	12.1	13.2	12.0

to be reasonably satisfactory. This clearly shows that proteins pertaining to ABA stress can be predicted with quite high accuracy; whereas, predictions of drought, salt and heat stresses are reasonably good by using the developed classifiers. The range of accuracy, sensitivity and specificity in this study was found comparable to other classification studies of proteins using SVM[5,16].

Conclusion

Plants have special physiological mechanisms of stress tolerance where proteins play the central role. Identification of these proteins by wet lab experimentation is expensive and time-consuming. Therefore, an *in silico* approach is advocated to narrow down this search prior to wet lab validation. Classification models were built to predict the function of proteins of cereals under four abiotic stresses using specific features of the protein sequence that are highly correlated with certain protein functions. In this study, features were selected through stepwise linear regression and the classification models, based on SVM, were trained using different kernel functions. The estimates of errors were obtained through 10-fold cross-validation techniques and comparative performances of the models were assessed on test data sets through different measures such as sensitivity, specificity and accuracy. In case of salt, performance of radial was found to be the best followed by polynomial and sigmoid; whereas in case of ABA, radial and polynomial function performs equally well and are better than sigmoid kernel. In case of drought stress, performances of all the three kernel functions were almost same. Thus, by using the developed classifiers, proteins pertaining to ABA, drought and salt stress can be classified with high accuracy level whereas heat stress protein can be classified with reasonably good accuracy. Heat stress protein can be predicted with reasonably good accuracy. These developed models can be used to develop a Web-based server for reliable prediction of functions of protein sequences with respect to these abiotic stresses which may further be validated through wet laboratory experiments. This may lead to considerable saving of cost and time.

1. Mahajan, S. and Tuteja, N., Cold, salinity and drought stresses: an overview. *Arch. Biochem. Biophys.*, 2005, **444**(2), 139–158.
2. Skriver, K. and Mundy, J., The gene expression in response to abscisic acid and osmotic stress. *Plant Cell*, 1990, **2**, 503–512.
3. Stone, P., The effects of heat stress on cereal yield and quality. In *Crop Responses and Adaptations to Temperature Stress* (ed. Basra, A. S.), Food Products Press, Binghamton, NY, 2001, pp. 243–291.
4. Kosova, K., Vítamvas, P., Prasil, I. T. and Renaut, J., Plant proteome changes under abiotic stress – contribution of proteomics studies to understanding plant stress response. *J. Proteomics*, 2011, **74**(8), 1302–1322.
5. Lee, B. J., Shin, M. S., Oh, Y. J., Oh, H. S. and Ryu, K. H., Identification of protein functions using a machine-learning approach based on sequence-derived properties. *Proteome Sci.*, 2009, **7**, 27.
6. Saeys, Y., Inza, I. and Larranaga, P., A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007, **23**, 2507–2517.
7. Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
8. Brown, M. P. *et al.*, Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 1999, **97**(1), 262–270.
9. Furey, T. S., Cristianianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 2000, **16**, 906–914.
10. Crammer, K. and Singer, Y., On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2001, **2**, 265–292.
11. Ding, C. H. Q. and Dubchak, I., Multi-class protein recognition using support vector machines and neural networks. *Bioinformatics*, 2001, **17**, 349–358.
12. Fu, W., Carroll, R. J. and Wang, S., Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, 2005, **21**, 1979–1986.
13. Burges, C. J. C., A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.*, 1998, **2**(2), 1–47.
14. R Development Core Team, 2011, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0; http://www.R-project.org/
15. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A., 2011, e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6; http://CRAN.R-project.org/package=e1071
16. Cai, C. Z., Wang, W. L., Sun, L. Z. and Chen, Y. Z., Protein function classification via support vector machine approach. *Math. Biosci.*, 2003, **185**, 111–122.

Received 27 September 2013; revised accepted 12 August 2014