MANTHAN
International Journal of Scientific Research and Innovation
ISSN NO: 0974-6331 www.bbmanthan.in
2015 Volume:15 Pages: 5-9

Identification of Stress Responsible Genomic Sequences using Back Propagation Neural Networks

S. B. Lal[1] & S. P. Varma[2]
[1] Ph.D. (IT) Student, B.R.A.Bihar University Muzaffarpur
[2] University (PG) Department of Mathematics, B.R.A.Bihar University, Muzaffarpur
Email: sblall@iasri.res.in sp.varma2shailesh@gmail.com

ABSTRACT

Back propagation neural networks (BPNN) have found widespread use for classification and function approximation. It is one of the most widely used methods in bioinformatics such as gene structure prediction, protein structure prediction, gene expression data analysis etc. Keeping the capability of this technique in mind, the problem of abiotic stress in plants were addressed. It is well known that the abiotic stress factors severely limit plant's growth, development and yield. An efficient prediction method for identification of protein functions based on its physico-chemical properties has been developed using BPNN classification method. On the basis of 34 features extracted from the protein sequences, models were built to predict its functions where the selected protein sequences were related to various abiotic stresses namely heat, drought, salt and ABA. This method analyses and identifies specific features of the protein sequences that are highly correlated with certain protein functions. Different measures such as sensitivity, specificity, accuracy were computed. Prediction of function of an unknown protein is considered to be an important objective for overcoming the abiotic stresses among plants.

Keywords: Backpropagation neural network, physico-chemical properties, salt, heat, ABA, accuracy, sensitivity, specificity.

INTRODUCTION

Abiotic stress, as a natural part of every ecosystem, affects organisms in a variety of ways. It is the negative impact of non-living factors on the living organisms in a specific environment. It has been claimed by one study that abiotic stress causes the most crop loss of any other factor and that most major crops are reduced in their yield by more than 50% from their potential yield [15]. The non-living variables adversely affect the population performance or individual physiology of the organism. The major abiotic stresses that affect plant growth and productivity are drought, heat, cold and salinity. Abiotic stress often causes a series of morphological, physiological, biochemical and molecular changes that unfavorably affect plant growth, development and productivity [11].

Abscisic acid (ABA) is an important phytohormone and plays a critical role in response to various stress signals. The application of ABA to plant mimics the effect of a stress condition. As many abiotic stresses ultimately results in desiccation of the cell and osmotic imbalance, there is an overlap in the expression pattern of stress genes after heat, drought, high salt or ABA application. This suggests that various stress signals and ABA share common elements in the signaling pathway [14]. Main function of ABA seems to be the regulation of plant water balance and osmotic stress tolerance.

Drought is one of the most serious abiotic stress which is associated with reduced water availability and cellular dehydration in plants. Heat stress is associated with an enhanced risk of improper protein folding and denaturation of several intracellular protein and membrane

complexes. The most observed effect of heat stress on plants is the retardation of growth. As heat stress often occurs simultaneously with drought stress, the combination of drought and heat stress induce more detrimental effect on growth and productivity of crops than when each stress was applied individually [12].

Salinity is a major environmental stress and is a substantial constraint to crop production. Increased salinization of arable land is expected to have devastating global effects, resulting in 30% land loss within next 25 years and up to 50% by the middle of 21st century [16]. 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.

The stress signal transduces inside the nucleus to induce multiple stress responsive genes, the products of which ultimately lead to plant adaptation to stress tolerance directly or indirectly [9]. Overall, the stress response could be a coordinated action of many genes, which may cross-talk with each other [13]. The stress-induced gene products are also involved in the generation of regulatory molecules like ABA, salicylic acid and ethylene. Therefore, 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.

Determining the functions of unknown proteins in a cost effective manner is necessary because experimental methods are time consuming and costlier. Instead, computational approaches for predicting and classifying protein functions would be preferred for faster results. Approaches based on sequence and structure comparisons play an important role in predicting and classifying the function of unknown proteins [8].

Many linear and nonlinear statistical techniques are available for binary classification such as discriminant analysis (DA), logit or probit models, random forest, and support vector machine etc. Hence, in this study classification of cereal proteins based on physico-chemical properties for four different stresses (heat, drought, salt, ABA) was undertaken using back propagation neural networks.

Background

Proteins are constructed from smaller molecules called amino acids. A protein is created when a series of amino acids are bound together. The arrangement of the amino acids can be described at 3 different levels. The primary structure is simply the order in which the amino acids are bound together. The secondary structure of a protein describes the way the string of amino acids fold. The specific bonding interactions among amino acids will determine if it is a coil or folding arrangement for these interactions to take place based on the primary structure. The tertiary structure of a protein describes the overall shape of the protein. It is a protein's shape which ultimately determines a protein's function. Therefore, changing a gene's sequence can change the primary structure of a protein, the amino acid sequence. This in turn may change a protein's secondary structure and then its shape, or tertiary structure. With an altered shape, the protein may function differently or may not function at all.

The function of a protein is determined by its amino acid sequence. The segments in the sequence of amino acids are known as motifs [1], they are crucial for their biological functions and can be used for their identification. Twenty different types of amino acids are combined in a linear sequence and have the necessary information to generate a unique 3-dimensional structure. There can be an infinite number of possible combinations of amino acids. The physical and chemical properties of the side chains of amino acids of a protein are important for the folding of the protein and its function. Proteins can be grouped into families and super-families according to their features such as hydrophobicity, length, 3-dimensional shape and electric charge, composition or structure which has common biological functions.

Due to the dramatic developments in molecular biology, a large number of data is constantly being generated through genome

sequencing projects. The information extraction from these data has been a crucial task for researchers. A rapid increase in the rate at which new protein structures are discovered has been greatly in need of new methods and algorithms generated for their classification and analysis.

The protein classification is not only necessary for predicting the function or the secondary / tertiary structure of a protein but also to classify a new protein belonging to a given family with previously known characteristics. Primary structures of most of the proteins are similar because many of them have a common evolutionary origin. Proteins of unrelated families can also have common structures. This kind of two fold nature of structures makes protein classification a complex task.

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 downregulated protein sequences were named as positive (+ve) and negative (-ve) 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.

Table.1 Sample size for positive and negative dataset selected for different abiotic stress

Protein class	Training set		Test set		Total
	+ve	-ve	+ve	-ve	
Drought	369	469	184	235	1257
Heat	57	41	28	21	147
Salt	3063	2093	1532	1046	7734
ABA	1737	2093	869	1046	5745
Total	5226	4696	2613	2348	14883

Inputs of BPNN for classification

ProtParam is a tool which allows the computation of various physical and chemical parameters for a given protein stored for a sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) [5]. These computed parameters with their values were utilized for formulation of input vector for BPNN.

After computation of important parameters of proteins related to a particular abiotic stress, classifier has been developed based using BPNN.

Back propagation neural networks

Back propagation neural network (BPNN) is a computational model that is inspired by the functioning of human brain by using a network of processing elements which operate on available data. Every processing element of BPNN represents a neuron that communicates with other elements. A weight is assigned to each connection between neurons and it represents the influence of one neuron on the other. Every neuron processes only local data received through its input connections and has a unique output that is fed to other neurons. The emergent intelligent behaviour of a BPNN comes from the interactions, between its processing units, that occur when it is presented with input data [4].

Weight values are determined by the iterative flow of training data through the network (i.e., weight values are established during a training phase in which the network learns how to identify particular classes by their typical input data characteristics). The training procedure is critical for BPNN to accomplish its purpose. Usually, training is a supervised procedure, in which a set of input-output patterns (a training set) is presented to the network and the computed result is compared to the actual value. The difference is used to update the weights of each layer using generalized delta rule. After several training epochs, when the error between the actual and the computed output is less than a previously specified value,

the network is considered to be trained. The knowledge learnt by the network is effectively represented by the set of weights, that is, the strength of the connections between neurons. Once trained, the neural network can be applied towards the classification of new data according to the acquired knowledge.

Classification performance

A standard classification methodology were adopted to assess classification performance as given in the classification literature [6], but sometimes this performance measure is misleading as this method does not discriminate between positive and negative cases. Two commonly used indicators in life and medical sciences are sensitivity and specificity. These measures are very frequently used in two-class problems. When using a system for classifying a protein of unknown class, depending on the class predicted by the system and on the actual class of protein, one of the following four types of result can be observed.

True positive (TP) – the system predicts that the protein belongs to a given class and the protein really does belong to that class.

False positive (FP) – the system predicts that the protein belongs to a given class but, in fact, it does not belong to it.

True negative (TN) – the system predicts that the protein does not belong to a given class, and indeed it does not belong to it.

False negative (TN) – the system predicts that the protein does not belong to a given class, but in fact, the protein does belong to it.

On the basis of these parameters, sensitivity, specificity and accuracy can be defined as given below:

Sensitivity = TP / (TP + FN) -(1)

Specificity = TN / (TN + FP) -(2)

Accuracy = (TP + TN) / (TP + TN + FP + FN) -(3)

Sometimes sensitivity, specificity and accuracy are called true positive rate and true negative rate, respectively. The sensitivity of a test is defined as the proportion of the proteins belonging to the class and is predicted rightly. Sensitivity measures the ability of the classifier system to correctly assign a protein to its real class. Sensitivity tells us nothing about whether or not some proteins which do not belong to the class would also be predicted into that class and, if so, in what proportion. On the other hand, specificity measures the ability of the system to reject a given protein as belonging to a class to which class it does not belong. A test with high specificity indicates that if protein is predicted as belonging to a class there is high probability of it is actually belonging to that class. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data.

Results and Discussion

In this study, physico-chemical properties of proteins for each abiotic stress were selected as input to the BPNN. BPNN classifier for four abiotic stresses were trained on training data sets with respective selected features by using sigmoid functions. Performance assessment for all the four stresses were obtained. Results are presented in Table 3 and it can be seen that prediction error for predicting ABA stress is minimum (around 0.4%) 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 BPNN classifier 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 are given in Table 3. It may also be noted that measure of accuracy in case of ABA is highest whereas lowest in case of drought. The measures of sensitivity and specificity were also obtained similarly. This experiment may be extended for other activation functions too.

Table 4 Estimates of accuracy, specificity and sensitivity obtained on test data

Abiotic stress class of protein	Accuracy (%)	Sensitivity (%)	Specificity (%)	Prediction error
ABA	90	87	93	0.4
Drought	78	69	84	3.8
Salt	39	53	20	5.1
Heat	96	93	100	12.0

Conclusions

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, physico-chemical properties of proteins, extracted using ProtParam, were selected for input to BPNN. The network was trained using different activation functions. The model was 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 function. In case of drought stress, performances of all the three activation 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. These developed models can be used to develop a web based application 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.

References

1. Attwood, T.K. Beck, M.E., Bleasby, A.J., Degtyarenko, K., Parry Smith, D.J. Progress with the PRINTS protein fingerprint database. Nucleic Acids Res. 1996. 24:182-8.
2. Fausett, L. 1994. Fundamentals of neural networks. Upper Saddle River: Prentice Hall, Inc.
3. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., BairochA.;Protein Identification and Analysis Tools on the ExPASy Server; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005). pp. 571-607
4. Hand, D.J. Construction and assessment of classification rules. New York: John Wiley and Sons. 1997.
5. 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.
   learning approach based Thomashow MF. Plant cold acclimation: Freezing tolerance genes and regulatory mechanisms. Annu Rev Plant Physiol Plant MolBiol 1999; 50:571‐99
6. Shinozaki K, Yamaguchi‐Shinozaki K.Molecular responses to dehydration and low temperature: Differences and
7. 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.
8. Prasad PVV, Staggenborg SA, and RisticZ(2009) Impacts of drought and/or heat stress on physiological, developmental, growth and yield processes of crop plants. In: Ahuja LH, Reddy VR, Saseendran SA and Yu Q (eds) Responses of Crops to Limited Water:Understanding and Modeling Water Stress Effects on Plant Growth
9. W. Wang, B. Vinocur, A. Altman, Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance, Planta 218 (2003) 1–1.