International Journal of Agriculture Sciences
ISSN: 0975-3710&E-ISSN: 0975-9107, Volume 9, Issue 5, 2017, pp.-3749-3751.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000217

Research Article

ARTIFICIAL NEURAL NETWORKS (ANNS) APPROACH FOR CLASSIFICATION OF SEED STORAGE PROTEINS OF VARIOUS NUTRITIONALLY SUPERIOR CEREAL CROPS

AVASHTHI HIMANSHU[1,2], JHA RICHA[3], SHARMA MUGDHA[4], YADAV ARVIND KUMAR[5], MISHRA A.K.[1*], RAMTEKE PRAMOD WASUDEO[6] AND KUMAR ANIL[7]

[1]ICAR - Agricultural Knowledge Management Unit, Indian Agricultural Research Institute, Pusa Campus, New Delhi, 110012, India
[2]Department of Computational Biology & Bioinformatics, JSBB, Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, 211007, Uttar Pradesh
[3]Department of Biotechnology, Uttaranchal Institute of Technology, Uttaranchal University, Arcadia Grant, Dehradun, 248007, Uttarakhand, India
[4]Department of Bioscience and Biotechnology, Banasthali University, Banasthali, 304022, Rajasthan, India
[5]ICAR-National Research Centre on Plant Biotechnology, Pusa Campus, New Delhi, 110012, India
[6]Department of Biological Sciences, School of Basic Science, Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, 211007, Uttar Pradesh
[7]Department of Molecular Biology & Genetic Engineering, CBSH, G. B. Pant University of Agriculture & Technology, Pantnagar, 263145, Uttarakhand, India
*Corresponding Author: Email- akmishra@iari.res.in , him.awasthi1989@gmail.com

Received: December 20, 2016; Revised: January 13, 2017; Accepted: January 14, 2017; Published: January 30, 2017

Abstract- Seed storage proteins comprise a key part of the protein content and play pivotal role to maintain the quality of seed. The composition of storage proteins are very essential because they determine the total protein content of the seed and show their effect on nutritional quality of the seed as well as functional properties of food processing. Therefore, classification is required to categorize these proteins and for the development of crops with improved nutritional superior varieties. Bioinformatics tools and techniques are extensively employed in the arena of agriculture to annotate the biological data. Annotation uncovers the structural and functional characteristics of genes as well as proteins also. In present study seed storage proteins of five major cereal crops were categorized into four classes i.e. albumins (12), globulins (42), glutelins (11) and prolamins (68) using six physicochemical properties (number of amino acid, molecular weight, theoretical pI (isoelectric point), aliphatic index, instability index and hydropathicity) by employing Artificial Neural Networks (ANNs) approach.

Keywords- Seed Storage Proteins, Physicochemical properties, Classification, Artificial Neural Network, Machine learning algorithm.

Citation: Avashthi Himanshu, et al., (2017) Artificial Neural Networks (ANNs) Approach for Classification of Seed Storage Proteins of Various Nutritionally Superior Cereal Crops. International Journal of Agriculture Sciences, ISSN: 0975-3710 & E-ISSN: 0975-9107, Volume 9, Issue 5, pp.-3749-3751.

Academic Editor / Reviewer: Dr Shambhavi Yadav, Dr Mamta Pandey

Introduction

Seed Storage Proteins (SSPs) accumulate in high levels during the late stages of seed development [1-2]. Nutritionally, composition of seed storage proteins is very important because it determines the food quality of seed [3-4]. For instance, the characteristics of high quality cereals are characterized by the quality of grains protein which is not yet determined by the composition of proteins. These storage proteins determine the total protein content of the seed and also its quality for various uses [5]. In cereals, storage proteins are comprised of about 50% of the total protein in mature grains and thus have a vital role in nutritional quality for humans and livestock and in functional properties in food processing [6].

Improvement in nutrient composition of seeds is a major target of molecular breeding and a lot of work has been performed on seed storage proteins to provide a base for improving the nutritional and processing properties of crops using genetic engineering methods. Hence, classification of seed storage proteins is very crucial for the development of superior crop varieties with improved nutritional quality [7]. In a recent study, neural networks approach has been used to classify the seed storage proteins of rice into four classes such as albumins, globulins, glutelins and prolamins [8]. Thus in this study, Artificial Neural Networks (ANNs) approach is employed to classify the seed storage proteins of various cereal crops such as rice, maize, foxtail millet, finger millet [9], sorghum and maize.

The aim of this research is to provide understanding of possible benefits of ANNs approach within a context of bioinformatics [10]. ANNs belongs to a group of machine learning techniques and became very popular to solve complex problems in areas of agriculture, business, engineering, medicine, and bioinformatics and systems biology [11]. It solves problems originating from linear and non linear data of above given areas. This approach might be applied for classification of data sets with high complexity. The focus within this research is on classification of seed storage proteins based on their physiochemical properties [12] using ANNs approach. ANNs predict the outcome of new independent input data and also to identify correlated patterns between input data sets and related target values. Thus this approach is ideally suited for classification of agricultural data which are known to be complex and often non-linear.

In 1924, T. B. Osborne classified the seed storage proteins into a variety of groups based on their solubility in a series of solvents *i.e.* in water (albumins), in dilute saline (globulins), in alcohol or ether mixtures (prolamins) and in dilute acid or alkali (glutelins). Although, nowadays “Osborne fractionation” is still widely used and more usual to classify seed proteins into various groups such as: storage

proteins, structural and metabolic proteins and protective proteins [13].

Materials and Methods

Retrieval of protein sequences: The amino acid sequences of seed storage proteins of foxtail millet (*Setaria italica*) were retrieved from foxtail millet database (http://foxtailmillet.genomics.org.cn/) and sequences of rice (*Oryza sativa*), maize (*Zea mays*), sorghum (*Sorghum bicolor*) and finger millet transcriptome data [14] of developing spikes (*Eleusine coracana*) were retrieved from (NCBI) National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

Database creation and homology search: Offline nucleotide database was formed using UGENE software and search homologous sequences in transcriptome data using tblastn (https://blast.ncbi.nlm.nih.gov/Blast.cgi? PAGE_ TYPE=BlastDocs&DOC_TYPE=Download) program because fewer information is available in relation to finger millet sequences in the databases. It search translated nucleotide databases using a protein query sequence. Seed storage protein sequences of sorghum were taken as query sequences to fetch the seed storage nucleotide sequences from developing spikes transcriptome data. These nucleotide sequences were further translated into protein sequence using Transeq (http://www.ebi.ac.uk/Tools/st/emboss_transeq/) tool of (EBI) European Bioinformatics Institute.

Physicochemical characterization: Physical and chemical parameters of seed storage proteins were computed using Expasy's ProtParam tool (http://web.expasy.org/protparam/). This tool provides various properties but in the present study we are now focused only six major properties *i.e.* number of amino acid, molecular weight, theoretical pI (isoelectric point), aliphatic index, instability index and hydropathicity. These parameters were further used to classify the proteins [15].

Attribute normalization and selection: To minimize the redundancy data normalization was performed. Without loss of information all the data sets are organized according to the relation between attributes and tuples. For the seed storage proteins data set, physicochemical properties were found to be influential attributes.

Data analysis, pre-processing and network design: To analyze and identify errors or missing values, input dataset of seed storage proteins with their six physiochemical properties from excel spreadsheet were loaded into Alyuda Neurointelligence (http://www.alyuda.com/neural-networks-software.htm) software. Different steps were followed to pre-process data and network design. Before fed to a neural network data numeric values transformed into scaling numeric values (-1 to 1), it makes suitable to design neural network.

Results and Discussion:

Back propagation Artificial Neural Network (BANN) analysis: For classification of seed storage proteins of various cereal crops three layered back-propagation training algorithm was used in the form of Alyuda Neuro-intelligence (AN). The normalized data of physicochemical properties was imported into Alyuda Neurointelligence software for data analysis, pre-processing and network optimization. The results of data processing [Fig-1] proved that if the sequence contained prolamin functional domain or related properties then it shows by value 1 in prolamin column and in others columns it shows 0, similarly for albumin, globulin and glutelin respectively. ANN recognizes the patterns and generates results in the form of 0 and 1 value.

In neural network architecture, input layer size represents the number of input layers fed to the neural network. Physicochemical parameters of 250 cereal's seed protein sequences were taken as input layer with six nodes. Three nodes (node -1, node 0 and node +1) were generated in scaling range of -1 to 1 in the form of hidden layer after estimation of several permutations. The output layer (nodes) size depends on the number of categories into which the input dataset has to be classified. Input data were classified into four classes, *viz.*, albumins, globulins, glutelins and prolamins. The best optimized architecture of the neural network selected was generated *i.e.*6-3-4 having 6 input, 3 hidden and 4 output nodes as shown in [Fig-2].

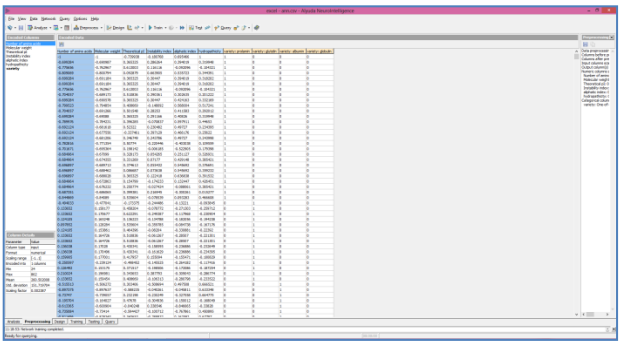

Fig-1 Pre-processed data based on presented functional domain in the form of numeric values 0 and 1.

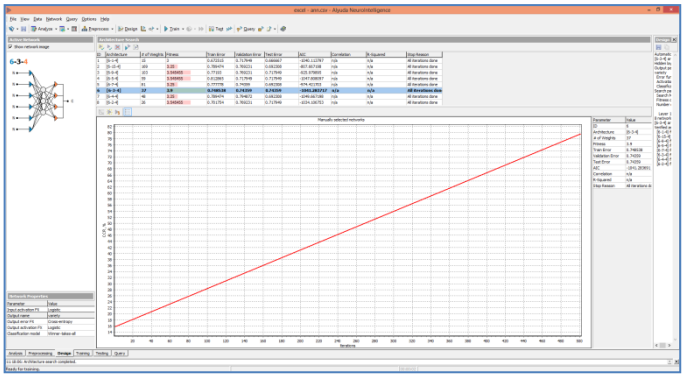

Fig-2 Selection of network in blue colour horizontal line and three layered neural network of seed storage proteins (six input layer nodes in blue colour, three hidden layer nodes in white colour and four output layer nodes in orange colour).

Learning rate and momentum rate parameters were selected as default with a value of 0.1. To enable valuable parameter and accomplishing advantageous classification rates the original input dataset of the seed storage proteins was divided into three sections: 60% for training, 20% for validation and the remaining 20% for testing the optimized network.

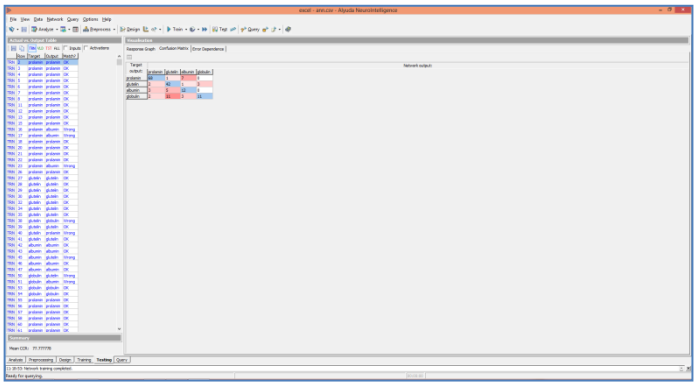

Fig-3 Illustrating Confusion matrix of accurately (blue colour) and in accurately (pink colour) classified seed storage proteins.

The output confusion matrix was generated for accurately and inaccurately classified proteins out of 250 seed storage proteins into four categories *i.e.* albumins (12), globulins (11), glutelins (42) and prolamins (68). These results illustrated that seed storage proteins contain large proportion of prolamin and glutelin in comparison to other seed storage proteins.

Earlier, a research was conducted by Marla *et al.,* (2010) regarding classification

of seed storage proteins on rice. In their study, for classification 170 seed storage protein sequences of rice were taken into account whereas in present study 250 seed storage protein sequences from five cereal crops were taken and results showed that rice seeds contain large proportion of prolamins and glutelins than other storage proteins. Same results were also obtained in present study which revealed that higher proportion of prolamin and glutelin in comparison to other seed storage proteins.

Conclusion

Neural network approach made easy to classify the seed storage proteins. In present study we focused on back propagation algorithm to classify the seed storage proteins of five major cereal crops. These five crops were taken into account with their physicochemical properties as an input dataset. The results of Alyuda Neurointelligence (6-3-4) showed 6 input (physicochemical parameters), 3 hidden (functional domain) and 4 output layers (categories of proteins) respectively. A large number of proteins (250) were correctly and successfully classified into four classes *i.e.* albumins (12), glutelins (42), globulins (11) and prolamins (68). These results illustrated that seed storage proteins contain large proportion of prolamin and glutelin than other storage proteins. Further, similar approaches may be utilised for *In silico* classification of other biomolecules into their functional categories, their relativity and genetic diversity. Related proteins and their genetic mining may prove to be a boon for genetic engineering in the field of agriculture and food industry in identifying related transgenes in the population and assisted genetic manipulations. Such nutrigenomic approaches will pave way for the incorporation and expression of nutritionally valuable traits into crop varieties and biofortification of desirable functional elements.

Acknowledgement

Authors are grateful and duly acknowledge to DIC Bioinformatics, Biotechnology Information System Network (BTISNet), Department of Biotechnology, Government of India, New Delhi for providing all necessary facilities for conducting quality research in area of Bioinformatics at Agricultural Knowledge Management Unit, Indian Agricultural Research Institute, New Delhi, India.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Alche J., et al. (2006) *J Agric Food Chem.,* 54(15), 5562-70.
[2] Toru Fujiwara., et al. (2002) *Storage Proteins. The Arabidopsis Book. American Society of Plant Biologists*. DOI: 10.1199/tab.0020.
[3] Shewry., et al. (1995) *American Society of Plant Physiologists,* 7, 945-956.
[4] Radhika., et al. (2015) *J Food Sci Technol.,* 52(7), 4246-4255.
[5] Shewry P. R. (2006) Improving the protein content and quality of temperate cereals: wheat, barley and rye. Impacts of Agriculture on Human Health and Nutrition-Volume II. Edited by Ismail Cakmak, Ross M. Welch.
[6] Shewry, et al. (2001) *Journal of Experimental Botany,* 53(370), 947–958.
[7] Nguyen, et al. (2012) *Journal of Experimental Botany,* 63(16), 5991–6001.
[8] Marla, et al. (2010) *J. Plant Biochemistry & Biotechnology,* 19(1), 123-126.
[9] Tiwari, et al. (2016) *Bioinformation,* 12(3), 156-164.
[10] Himanshu Avashthi, et al. (2014) *Int J Comput Bioinfo In Silico Model,* 3(4), 454-459.
[11] Mlambo, et al. (2016) *Int J Adv Res in Comp Sci Softw Eng.,* 6(3), 59-65.
[12] Garg, et al. (2016) *Bioinformation,* 12(2), 74-77.
[13] Bailey K., et al. (1942) *Biochem J.,* 36(1-2), 140–154.
[14] Anil Kumar, et al. (2015) *Int J Comput Bioinfo In Silico Model,* 4(6), 749-752.
[15] Richa Jha, et al. (2012) *ISRN Vet Sci.,* 2012, 512848