SUB-CELLULAR LOCALIZATION PREDICTION USING MACHINE LEARNING APPROACH

Ashutosh Kumar Singh[1], S S Sahu[2], Ankita Mishra[3]
[1,2,3]Birla Institute of Technology, Mesra, Ranchi
Email: [1]ashutosh.4kumar.4singh@gmail.com, [2]sssahu@bitmesra.in,
[3]find.ankitamishra@gmail.com

Abstract— Prediction of sub cellular locations of protein has been a challenging issue as its location determines the function. Conventional methods based on lab experiments are expensive and time consuming. Hence computational prediction of the subcellular location gained as importance in the last two decades. In this paper we propose a machine learning approach based on neural network to predict the locations of the proteins form the sequence information. Several feature of the protein has been explored to accurately predict the sub cellular location. Performance of the proposed method demonstrates that it is competitive with the existing methods.

I. INTRODUCTION

A cell is the functional and structural unit in all living organisms. The cell is a three-dimensional space separated into different compartments. It contains the protein molecules after translation in different compartment or organelles. These cellular compartments have different function and physicochemical environment. Proteins need to be present at specific cellular compartments to make the cells to function properly. The different organelles includes Cell membrane, Cell wall, Cytoplasm, Endoplasm, Extracellular space, Golgi apparatus, Mitochondrion, Nucleus, Peroxisome, Plastid, Vacuole etc. Each protein has a special role to play in these organelles. Proteins are responsible for all biochemical reactions inside an organism. Hence, proper sub-cellular localization is a pre-requisite to identify protein functions. Thus, there is a demand for accurate and complete localization of all proteins. The information about the sub cellular localization can be extracted by using various biochemical methods. But predicting subcellular location of protein only by biochemical methods is unpractical because of the high costs and time complexities. In recent years, computational prediction of protein subcellular localization has gained tremendous attention to replace the time-consuming and laborious wet-lab methods. Predictors based on machine learning and pattern recognition are becoming increasingly popular. Several machine learning approaches such as support vector machine (SVM) [1], K-nearest neighbor approach [2].

In this paper we propose an efficient predictor based on neural network using the sequence features only. It is very simple to implement.
In this paper we use four different feature extraction methods: Amino acid composition, Dipeptide composition, Pseudo-Amino acid composition and NCC (N terminal- C centre- C terminal). We will use these features individually and also the combination of these features for prediction. We will use Machine learning to develop the predictor for sub-cellular localization of proteins. The above features will be extracted from the known protein sequences and will be used to train the neural network. This

network will be used as the predictor for prediction of sub-cellular localization of proteins.

II. METHODS

Dataset generation

The data were extracted from the SwissProt database (release 2013_02) all the plant proteins with known subcellular localization by parsing the 'SUBCELLULAR LOCATION' section of the COMMENT field. Sequences whose annotations were marked as 'PROBABLE', POSSIBLE', 'BY SIMILARITY' and 'FRAGMENT' were discarded. We ended up with 16089 sequences of proteins, annotated according to 11 different subcellular localizations as detailed in Table 1. We discarded proteins endowed with multiple localizations and those are length less than 50 amino acids.

However, the sequence number drastically reduced to 6149 after we put a sequence identity cutoff of <30% (Table 1) on each of them using BlastClust. To avoid homology bias in machine learning, a 25 or 30% sequence identity cutoff threshold is needed to guarantee that none of the proteins included in the benchmark datasets has greater than this threshold identity to any other sequences in the dataset. This was done within class as well as across the classes. Further, about 10% of the data was kept aside for later independent testing of the models. Finally, 5557 sequences are taken as Training dataset and 592 sequences as Independent Test dataset.

Feature extraction of protein

For prediction of sub-cellular localization of protein it is necessary to extract the desired features from the protein sequences. The following diverse features were extracted from the protein sequences for use in a machine learning framework for developing prediction models

Amino acid composition (AAC)

In this type of representation, each protein is defined by a 20-dimensional feature vector in Euclidean space where the co-ordinates are given by the occurrence frequencies of the 20 constituent amino acids [3,4]. For a query protein x, let $f(x_i)$ represents the occurrence frequencies of its 20 constituent amino acids. Hence the composition of the amino acids (P_x) in the query protein is given by,

$$P(x_i) = \frac{f(x_i)}{\sum_{i=1}^{20} f(x_i)} \qquad i = 1,2,3 \ldots\ldots 20 \quad (1)$$

Hence, the protein x in the composition space is defined as: $P(x) = [P_1(x), P_2(x), \ldots, P_{20}(x)]$.

Table 1 Distribution of subcellular localization

Subcellular location	# sequences retrieved	Training dataset (sequences length >50)	Independent dataset (sequences length >50)
Plastid	11302	2468	248
Cytoplasm	739	351	40
Extracellular	237	140	14
Nucleus	734	568	63
Mitochondrion	759	447	52
Cell Membrane	1256	829	92
Golgi Apparatus	277	204	23
Endoplasmic Reticulum	393	280	29
Vacuole	260	176	20
Peroxisome	80	57	06
Cell Wall	52	37	05
Total	**16089**	**5557**	**592**

classes for All-Plant data from UniProt release 2013_02 in training dataset and Independent testing dataset.

Dipeptide composition (DIPEP)

In this representation, the occurrence frequencies of each dipeptide in the sequence is computed producing a fixed pattern length of 400 (20 × 20) for the query protein. Thus, the composition of the dipeptide is given as:

$$P(x_i,x_j) = \frac{f(x_i,x_j)}{\sum_{i=1}^{20}\sum_{j=1}^{20} f(x_i,x_j)}$$ i, j = 1, 2, 3 ..... 20 (2)

where $P(x_i,x_j)$ is the fraction of each (x_i,x_j) dipeptide and $f(x_i,x_j)$ is the frequency of occurrence of (x_i,x_j) dipeptides, and the denominator represents the total number of all possible dipeptides.

Pseudo amino acid composition (PseAAC)

In composition based methods, protein sequence order and length information are completely lost, which in turn may affect the prediction accuracy of the model. To include all the details of its sequence order and length, Chou [7] proposed an effective way of representing known proteins as pseudo amino acid compositions (PseAAC) in his seminal study. In this representation, the protein character sequence is coded by some of its physicochemical properties. The PseAAC feature extraction process has been followed form [10].

Terminal-based N-Center-C (NCC) amino acid composition In this method, the amino acid composition of the N-terminal region, the C-terminal region, and the remaining center portion of protein sequence is computed separately and then concatenated together to represent a sample protein. In this technique, a protein sample is represented as: P (x) = [$AAC_{N-terminal}$ $AAC_{Center\ region}$ $AAC_{C-terminal}$]. The AAC for each segment is computed using (1). Hence, a 60 dimensional feature vector is used to represent a protein. In an empirical study, the residue length of 25 was found to be the best compromise.

In this paper, several hybrid features also implemented by combining these individual feature. These hybrid features are AACDIPEP, NCCDIPEP, PseAACDIPEP and PseAACNCCDIPEP.

Artificial Neural Network

Machine learning is evolved from the study of pattern recognition and computational learning. The basic concept behind Machine learning is to generate an algorithm which learn from the available data and make predictions on unknowns. In Machine Learning, artificial neural networks (ANN) are a family of models which is inspired by biological neural networks. They constitute of a set of interconnected neurons. The connections have numeric weights which can be tuned according to the input data, thus making the network capable of learning and adaptive to inputs.

The network used for classification of protein sequences is shown in Fig. 1.

In this work, the multi class problem is divided into multi binary class scenario by one vs rest method. In this way we will develop eleven separate neural network model for the protein sequences of each organelle. Finally, all the networks will be combined to develop the final model for prediction.

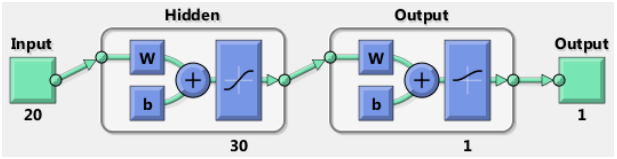

Fig. 1 Neural Network used for classification

Training / testing schema: The training data was transformed into a five-fold cross-validation scheme, where the dataset is divided into five different parts. Four parts are combined to form one training set and the models developed from this set are then tested on the fifth part (called testing set). This process is repeated five times changing the training / testing set each time, and is thus called five-fold cross-validation. In addition, we have also tested the performance of our models on independent test datasets, those that have not been used in any kind of machine learning.

Evaluation parameters: The performance of models developed is evaluated based on the overall accuracy defined as:

Accuracy (Acc) $= \frac{TP+TN}{TP+TN+FP+FN} \times 100$

TP: True positive
TN: True negative
FP: False positive
FN: False negative

III. RESULT AND DISCUSSION

To show the distinguishing capability of the features extracted from the protein, the amino

acid composition feature is shown by Andrews plot in Figure 2.

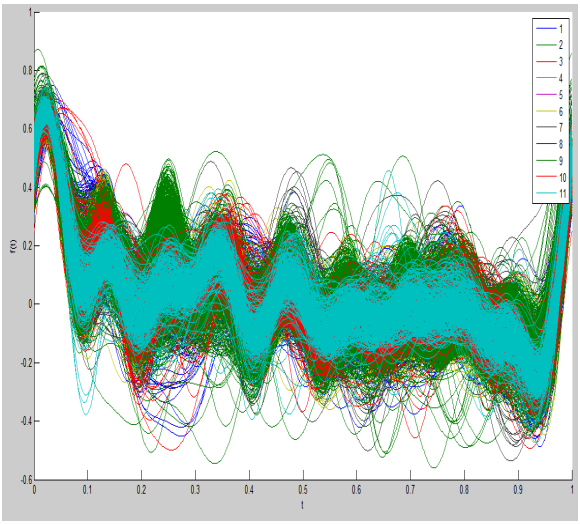

Fig. 2 Andrews plot of amino acid composition feature among 11 different subcellular locations

From the variations in the plot we deduce that the feature is capable of discriminating the different classes. Thus, artificial neural network can be applied to classify the protein sequences of different organelles.

The performance of the neural network in the training dataset is listed in Table-2. From the table it is elucidated that the PseAAC feature has the capability of predicting the classes in 75 % overall accuracy. Again the performance of the models has been assessed in the independent dataset and the results shown in Table 3. The efficiency of the predictor to identify protein location is compared with the existing web tools such as YLoc+ and Eku-mPLoc. The performance is listed in Table 4. From the comparison it is evident that our proposed method based on PseAACNCCDIPEP feature is superior than the existing methods.

Table-2 Performance comparison of the feature extraction methods in artificial neural network on training dataset.

Table-3 Performance comparison of the feature extraction methods in artificial neural network on test dataset.

	AAC	DIPEP	PSE AAC	NCC	AAC DIPEP	NCC DIPEP	PSE AAC DIPEP	PSEAAC NCC DIPEP
Cellmemb	62.6	67.8	75.6	64.4	73.3	79	74.9	76.8
Cellwall	16.2	59.5	24.3	21.6	48.7	43.2	5.41	27
Endoplasm	8.21	36.1	14.6	14.6	31.8	41.8	30.7	48.2
Extracell	28.6	44.3	33.6	42.9	49.3	57.9	35.7	59.3
Golgi	29.4	63.2	30.9	9.8	63.7	53.4	43.6	75
Mito	13.9	30.2	17.7	17.5	17	43.9	22.8	34
Nucl	69.2	68.5	69	69.2	72.7	77.6	77.1	79.6
Peroxi	0	0	1.75	0	14	14	7.02	0
Plastid	92.4	91	91.6	92.9	92.7	91.7	90.4	93.7
Vacu	0.57	9.66	25	6.25	32.4	27.3	42.1	30.7
Cyto	4.56	22.2	14	35.3	32.8	39	35.3	37.9
Overall	**61.2**	**67.3**	**65**	**64.1**	**69.6**	**73.2**	**68.8**	**74.2**

Table-4. Comparison Table

	Overall Efficiency
YLoc+	34.53%
Eku-mPloc	53.5%
PSEAACNCCDIPEP	58.8%

IV. CONCLUSION

In this work, we have developed a machine learning framework based on artificial neural network to accurately predict the subcellular locations of protein targeting eleven different locations. From the empirical study we found that the hybrid feature PseAACNCCDIPEP is capable of discriminating the protein sub-cellular locations efficiently. The proposed approach has shown its superiority over existing methods.

V.REFERENCE

1. C. Yu, C. Lin, and J. Hwang, "Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions," Protein Science, vol. 13, no. 5, pp. 1402–1406, 2004
2. G. A. Arango-Argoty, J. F. Ruiz-Munoz, J. A. Jaramillo-Garzon, and C. G. Castellanos-Dominguez "An adaptation of Pfam profiles to predict protein

sub-cellular localization in Gram positive bacteria" 34th Annual International Conference of the IEEE EMBS,2012

3. Kaundal R, Saini R, Zhao PX: Combining Machine Learning and Homology-Based Approaches to Accurately Predict Subcellular Localization in Arabidopsis. Plant Physiology 2010, 154:36-54.
4. Sahu SS, Panda G: A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational Biology and Chemistry 2010, 34:320-327.
5. Kaundal R, Raghava GPS: RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. Proteomics 2009, 9(9):2324-2342.
6. Garg A, Bhasin M, Raghava GPS: Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. Journal of Biological Chemistry 2005, 280:14427-14432.
7. Chou KC: Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 2001, 43:246-255.
8. Jiang X, Wei R, Zhang TL, Gu Q: Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Peptide Lett 2001, 15:392-396.
9. Zhang TL, Ding YS, Chou KC: Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 2008, 250:186-193.
10. Rakesh Kaundal, Sitanshu S. Sahu, RuchiVerma and Tyler Weirick, "Identification and characterization of plastid-type proteins from sequence-attributed features using machine learning", BMC Bioinformatics 14(S14): S7, 2013