PLoc-Euk: An Ensemble Classifier for Prediction of Eukaryotic Protein Sub-cellular Localization

Rajkamal Mitra, Piyali Chatterjee, Subhadip Basu, Mahantapas Kundu and Mita Nasipuri

Abstract Protein Sub-Cellular Localization is very important information as they play a crucial role in their functions. Thus, prediction of protein Sub-Cellular Localization has become very promising and challenging problem in the field of Bioinformatics. Recently, a number of computational methods based on amino acid compositions or on the functional domain or sorting signal. But, they lack of contextual information of the protein sequence. In this paper, an ensemble classifier, PLoc-Euk is proposed to predict sub-cellular location for the eukaryotic proteins which uses multiple physico-chemical properties of amino acid along with their composition. PLoC-Euk aims to predict protein Sub-Cellular Localization in eukaryotes across five different locations, namely, *Cell Wall, Cytoplasm, Extracellular, Mitochondrion,* and *Nucleus.* The classifier is applied to the dataset extracted from http://www.bioinfo.tsinghua.edu.cn/~guotao/data/ and achieves 73.37% overall accuracy.

Keywords Sub-cellular localization · Physico-chemical properties of amino acid · Ensemble classifier

R. Mitra
Rate Integration Software Technologies Pvt. Ltd.,
213 A, A.J.C. Bose Road, Kolkata 20, India
e-mail: rajkamal.mitra@evolving.com

P. Chatterjee (✉)
Department of Computer Science & Engineering,
Netaji Subhash Engineering College, Garia 152, Kolkata, India
e-mail: chatterjee_piyali@yahoo.com

S. Basu (✉) · M. Kundu · M. Nasipuri
Department of Computer Science & Engineering,
Jadavpur University, Kolkata 700032, India
e-mail: subhadip@cse.jdvu.ac.in

M. Kundu
e-mail: mahantapas@gmail.com

M. Nasipuri
e-mail: mitanasipuri@yahoo.com

S.C. Satapathy et al. (eds.), *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, Advances in Intelligent Systems and Computing 516, DOI 10.1007/978-981-10-3156-4_12

1 Introduction

With the deluge of gene products in the post genomic age, the gap between the newly found protein sequences and their cellular location is growing larger. To use these newly found protein sequences for drug discovery it is desired to develop an effective method to bridge such a gap. In real life, it is found that proteins may simultaneously exist at or move between two or more different Sub-Cellular locations. Thus, localization of proteins is very challenging problem in Bioinformatics. The annotations of protein Sub-Cellular localization can be detected by various biochemical experiments such as cell fraction, electron microscopy and fluorescent microscopy. These accurate experimental approaches are time consuming and expensive which necessitates the computational techniques to predict protein Sub-Cellular Localization which will be useful for protein function prediction. A number of in-silico Sub-Cellular Localization methods have been proposed. Most of the prediction methods can be classified into various categories which are based on the recognition of protein N-terminal sorting signals, amino acid composition, functional domain, homology and fusion. Sorting signals are short sequence segments that localize proteins to intra or extra cellular environments. These include signal peptides, membrane-spanning segments, lipid anchors, nuclear import signals and motifs that direct proteins to organelles such as Mitochondria, Lysosomes etc. [1]. Nakai and Kanehisa [2] took pioneering attempt to propose a computational method, named PSORT, based on sequence motifs and amino acid composition by exploiting a comprehensive knowledge of protein sorting. Reinhardt and Hubbard [3] used amino acid composition information to predict protein subcellular location in neural network based system. Chou and Elrod [4, 5] also used amino acid composition in prediction of subcellular location applying covariant discriminant algorithm. They got better prediction accuracies when they used correlations of residue pairs and acid composition. A work based on Signal based information [6] has been proposed by Emanuelsson and co-authors where individual sorting signals e.g. signal peptides, mitochondrial targeting peptides chloroplast transit peptides are identified [14]. Then they proposed an integrated prediction system using neural network based on the prediction of individual sorting signals. The reliability of the method is based on the quality of the genes 5′-region or protein N-terminal sequences assignment. However, the assignment of 5′-regions are usually not reliable using gene identification methods. Inadequate information of signals may give inaccurate results which results in low accuracy. Hua and Sun [7] used a radial Basis kernel SVM based prediction system using Amino Acid composition. Another voting scheme based work using amino acid composition for prediction of 12 Sub-Cellular locations is done by Park and Kanehisa [8] where a set of SVMs was trained based on its amino acid, amino acid pair and gapped amino acid pair compositions. MultiLoc [9] is an SVM based approach which integrates N-terminal targeting sequences, amino acid composition and protein sequence motifs. It predicts eukaryotic proteins very well. Hortron et al. [10] proposes extension to PSORT-II which is a sorting signal composition based

method called WOLF PSORT where amino acid content, sequence length, sorting signals are used. The use of feature sets increased the prediction accuracy of PSORT II with the same classifier k-nearest neighbor. In the work of Chou and Shen [11] proposed an ensemble classifier with kNN basic classifier which uses the concept of pseudoAA (pseAA) composition. Mer and his co-author proposed a novel approach [12] exploiting amino acid composition and different levels of amino acid exposure. The concept was based on that differently exposed residues have different evolutionary pressures to mutate towards specific amino acid types whose side chains have physicochemical properties that agree to the Sub-Cellular location where the protein performs its better activity. To predict singleplex or multiplex protein siLoc-Euk [13] uses multi-label classifier over 22 location sites. APSLAP [14] uses adaptive boosting technique empowered with physicochemical descriptor, Amino acid composition and CTD. From the above mentioned methods, it can be observed that some predictors have experimented with different feature sets for a particular classifier [2–7] or some predictors have taken a voting scheme or ensemble classifier from set of classifiers [8, 11]. In this work, these facts motivate us to use multiple physico-chemical properties weighted by AAC and ensemble classifier of different classifier.

2 Materials and Methods

In this work, an attempt has been taken to use combination of amino acid composition and their physicochemical properties for prediction of five different eukaryotic Sub-Cellular locations, i.e. Cell wall, Cytoplasm, Mitochondrion, Extracellular and Nucleus. Here, whole experiment is conducted in two stages. In the first stage, four different types of classifiers, namely, PART, Multi-Layer Perceptron (MLP), Adaboost and RBF neural network are taken and their performance are observed for prediction. In the second stage of experiment, an ensemble classifier is constructed on the basis of two well performed classifier (in this case, PART and Adaboost Classifier) to achieve better prediction accuracy.

2.1 The Feature Set

The Amino Acid Composition (AAC) of a protein specifies the occurrence (sometimes percentage) for each of the 20 amino acids. AAC of a protein for location is based on the hypothesis that differences in AAC associate with different locations [12]. On the other hand, use of appropriate physico-chemical properties of amino acids also determines its location of activity. Relevant physico-chemical properties of amino acids can be mentioned in this respect, namely, hydropathy, charge, solubility, pKa value, LP value, hydrophilicity and Isoelectric point value. According to the theory of Lim (1974), amino acid residue hydrophilic patterns

incline to occur in secondary structure of a protein sequence. The hydrophobic value of amino acid residue represents the major driving force behind protein folding and protein has activity only in specific folding pattern. As proteins take different functions in different part of cellular location it can be concluded that the Hydropathy and Hydrophilicity feature of amino acid have a great influence in protein Sub-Cellular localization. Charge is also important in this field, e.g., it has been seen that the most nucleus protein consists of much more amino acid residues which are positively charged [15]. On the other hand, LP [16] values of amino acids are basically used for protein function prediction as the function and location of a protein is highly correlated, LP value can be used as a feature for protein Sub-Cellular Localization. Studies say that the solubility of a protein is highly related with its function [3] and is a major property of proteins that determines their function and location within a cell. Isoelectric points or pKa value of amino acids are changed according with their location environment. So proteins which reside in particular location of a cell may have identical isoelectric point and pKa value.

In this work, every protein sequence is represented by seven elements vector where each element in the vector represents a particular physicochemical property weighted by AAC. It is mathematically represented as $P = [P_1, P_2, P_3, P_4, P_5, P_6, P_7]$ of any protein P refers to occurrence of any residue a_i of 20 amino acids and is calculated using the Eq. 1. Finally it is normalized in the range [0, 1].

$$AACa_i = \frac{Occurrence\ of\ a_i}{length\ of\ protein\ sequence}. \tag{1}$$

The feature indices of Charge, Hydrophilicity, LP value, Hydropathy were taken from AAindex dataset [17]. The physicochemical properties are weighted by AAC using Eqs. 2–8.

$$P_1 = \sum_{n=1}^{20} AAC_i \times hydropathy\,(a_i) \tag{2}$$

$$P_2 = \sum_{n=1}^{20} AAC_i \times charge\,(a_i) \tag{3}$$

$$P_3 = \sum_{n=1}^{20} AAC_i \times solubility\,(a_i) \tag{4}$$

$$P_4 = \sum_{n=1}^{20} AAC_i \times isoelectricpoint\,(a_i) \tag{5}$$

$$P_5 = \sum_{n=1}^{20} AAC_i \times pK\,(a_i) \tag{6}$$

$$P_6 = \sum_{n=1}^{20} AAC_i \times hydrophilicity\,(a_i) \tag{7}$$

$$P_7 = \sum_{n=1}^{20} AAC_i \times LP\,(a_i) \tag{8}$$

2.2 Design of the Classifier

As previously mentioned, four different classifiers, namely, PART, RBF NN, Adaboost and MLP are taken and their individual performance is observed. Prediction decisions of two well performed classifiers are combined to construct an ensemble classifier PLoc-Euk to boost up its prediction accuracy. The basis of ensemble classifier is to accept prediction decision from one of its component classifier which classifies a protein at higher confidence. Ensemble classifier PLoc-Euk is constructed from two component classifiers PART and Adaboost as they are found to have better prediction accuracy compared to MLP and RBFNN.

2.3 Experimentation and Results

Data Set. We have taken 1001 Eukaryotic protein sequences with five Sub-Cellular locations extracted from (http://www.bioinfo.tsinghua.edu.cn/~guotao/data/) where 750 protein sequences serve as training data and remaining 251 sequences act as test data. For training data 150 protein sequences are taken from each Sub-Cellular location and 50-51 protein sequences are taken as test data for every five locations.

Performance Measure. The performance of classifiers is evaluated using two performance measures: Matthews Correlation Coefficient and Accuracy which is described as follows:

Matthews Correlation Coefficient (MCC)

It is used in machine learning as a measure of quality of binary (two class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of different sizes. The MCC is a correlation coefficient between the observed and predicted binary classifications. It returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and −1 an inverse prediction. Here, when considering a particular class as positive (here location i.e., cell wall) then all other locations are considered to be negative class. Thus, TP, FP, TN, FN for every class or location are calculated and used in computation of MCC.

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{((TP+FN)(TP+FP)(TN+FP)(TN+FN))}}. \tag{9}$$

Accuracy

It is calculated to measure the performance of a predictor system and defined by

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}. \tag{10}$$

where, TP, TN, FP, FN have their usual meanings.

Performance Evaluation. The whole experiment is conducted in two stages. Initially, four classifiers are applied for the prediction of Sub-Cellular location of test proteins. In the second stage, two best classifiers are taken as component classifier for constructing an ensemble classifier PLoc-Euk. As two classifiers are taken as component classifier, so PLoc-Euk takes prediction decisions from them which classify the test sample at higher confidence. In this section, performance of four classifiers, namely, PART, RBFNN, MLP and Adaboost classifier are observed in prediction of subcellular location i.e., cell wall, extracellular, mitochondrion, nucleus and cytoplasm. Tables 1, 2, 3 and 4 show MCC scores and Accuracy measures of our classifiers. From this table, it is evident, the average accuracies of PART classifier and Adaboost classifier are comparatively better than MLP and RBFNN. Finally, Table 5 shows performance of PLoc-Euk where in most of the cases it performs well compared to component classifiers. The comparison of the performances of PLoc-Euk and its component classifiers are graphically presented in Fig. 1.

Comparison of PLoc-Euk with existing Predictors. We have taken Cello v 2.5 [18] and WOLF-PSORT [10] as existing methods for comparison because they are freely available though they are not too recent but they are based on machine learning method. To compare the performance of the present work PLoc-Euk, 251 test proteins are tested with Cello v2.5 and WOLF -PSORT. From Fig. 2. It can be explained that for cytoplasmic protein, mitochondrion and Nucleus proteins PLoc-Euk performs better than WOLF-PSORT. In case of mitochondrion protein it performs better than two predictors. But, for extracellular proteins, PLoc-Euk does not achieve well.

Table 1 Performance measures of PART classifier

Location	MCC	Accuracy (%)
Cell wall	0.6536	86
Cytoplasm	0.5846	58
Extra cellular	0.5312	50
Mitochondrion	0.7252	78
Nucleus	0.794	90
Average	0.66	72.51

Table 2 Performance measures of MLP classifier

Location	MCC	Accuracy (%)
Cell wall	0.393	66
Cytoplasm	0.5629	70
Extra cellular	0.3824	44
Mitochondrion	0.6276	62
Nucleus	0.74	68.6
Average	0.5412	62.154

Table 3 Performance measures of RBFNN classifier

Location	MCC	Accuracy (%)
Cell wall	0.47	64
Cytoplasm	0.47	66
Extra cellular	0.46	48
Mitochondrion	0.52	52
Nucleus	0.72	78
Average	0.53	61.753

Table 4 Performance measures of Adaboost classifier

Location	MCC	Accuracy (%)
Cell wall	0.6326	78
Cytoplasm	0.5233	66
Extra cellular	0.49	54
Mitochondrion	0.7510	72
Nucleus	0.704	78
Average	0.62	69.32

Table 5 Performance measures of PLoc-Euk classifier

Location	MCC	Accuracy (%)
Cell wall	0.65	84
Cytoplasm	0.58	60
Extra cellular	0.61	64
Mitochondrion	0.72	74
Nucleus	0.78	86
Average	0.67	73.37

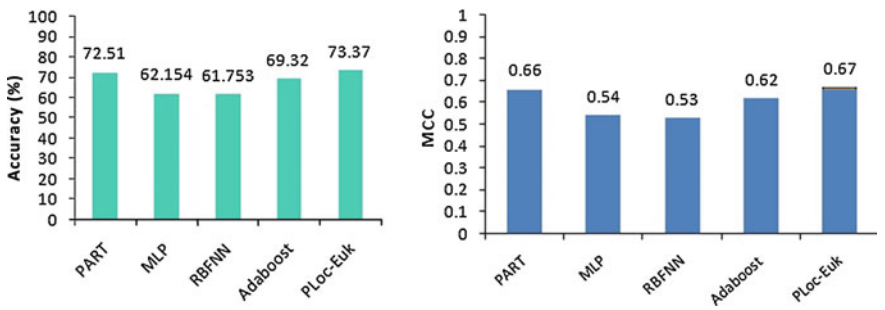

Fig. 1 Performance comparison of ensemble classifier PLoc-Euk and other classifiers

Conclusion. Sub-Cellular localization information of any protein gives proper insight of its function. Thus it has become very challenging task in Bioinformatics. Previously signal based, amino acid composition based, structural based approaches were taken for computation prediction approach. In this work we have combined

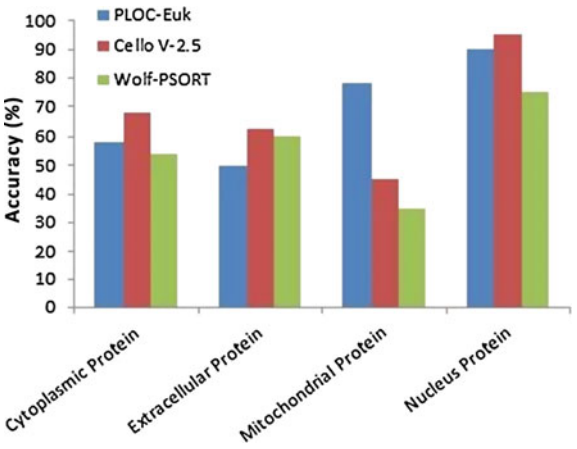

Fig. 2 Comparison of PLoc-Euk, CelloV-2.5 and Wolf-PSORT on test proteins of five locations

weighted physicochemical based properties of amino acids and their composition as input vector. We have taken 7 relevant physicochemical properties and represented them according to their amino acid composition. Thus weighted properties indicate their intensity and dominance over the protein thereby making the predictor to predict their Sub-Cellular location properly. In addition to these physicochemical properties the performance of the different classifiers has been observed and it is found that we get good performance in PART and Adaboost classifier and also from PLOC-Euk classifier which was designed upon PART and Adaboost classifier. We also compare our work with some existing prediction system. Signal based information can be added with the physicochemical properties to strengthen the prediction power of this classifier. Individual physicochemical properties also have its own influence on a protein to be in a particular location within the cell. So, a number of physicochemical properties can be taken and any feature optimization technique can be employed to reduce the dimension of the input feature vector physicochemical properties, more cellular location also can be included to increase the number of classes and it will also make our system reliable. From further analysis of our work, we can also create a relationship between the Sub-Cellular location and Protein-Protein Interaction [19, 20] and domain information [21] of protein which may be a further research of Bioinformatics.

References

1. R. Mott, J. Schultz, P. Bork, C.P. Ponting, Predicting protein cellular localization using a domain projection method. Genome Res. **12**, 1168–1174 (2002)
2. K. Nakai, M. Kanehisa, A knowledge Base for predicting protein localization sites in Eucaryotic cells. Genomics **14**, 897–911 (1992)
3. A. Reinhardt, T. Hubbard, Using Neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. **26**, 2230–2236 (1998)

4. K.C. Chou, D.W. Elrod, Protein subcellular location prediction. Protein Eng. **12**, 107–118 (1999)
5. K.C. Chou, D.W. Elrod, Prediction of membrane types and subcellular locations. Proteins **34**, 137–153 (1999)
6. O. Emanuelsson, H. Nielson, S. Brunak, G. von Heijne, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. **300**, 1005–1016 (2000)
7. S. Hua, Z. Sun, Support vector machine approach for protein subcellular localization prediction. Bioinformatics **17**, 721–728 (2001)
8. K. Park, M. Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics **19**, 1656–1663 (2003)
9. A. Hoglund, P. Donnes, T. Blum, H.W. Adolph, O. Kohlbacher, MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics **22**, 1158–1165 (2006)
10. P. Horton et al., WoLF PSORT: protein localization predictor. Nucleic Acids Res. **35**, W585–W587 (2007)
11. K.C. Chou, B. Shen, Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc. **3**, 153–162 (2008)
12. A.S. Mer, M.A. Andrade-Navarro, A novel approach for protein subcellular location prediction using amino acid exposure. BMC Bioinformatics **14** (2013), doi:10.1186/1471-2105-14-342
13. K.C. Chou, Z.C. Wu, X. Xiao, iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLOS One **6**, e18258 (2011)
14. V. Saravanan, P.T. Lakshmi, APSLAP: an adaptive boosting technique for predicting subcellular localization of apoptosis protein. ActaBiotheoretica **61**, 481–497 (2013)
15. H. Nakashima, K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. **238**, 54–61 (1994)
16. M.A. Andrade, S.I. O'Dnoghue, B. Rost, Adoption of protein surfaces to sub-cellular locations. J. Mol. Biol. **276**, 517–525 (1998)
17. S. Kawashima, H. Ogata, M. Kanehisa, AAindex: amino acid index database. Nucleic Acids Res. **27**, 368–369 (1999)
18. C.S. Yu, Y.C. Chen, C.H. Lu, J.K. Hwang, Prediction of protein subcellular localization. Proteins **64**, 643–651 (2006)
19. P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, D. Plewczynski, PPI_SVM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. Cell Mol. Biol Lett. **16**, 264–278 (2011). doi:10.2478/s11658-011-0008-x
20. S. Saha, P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, Funpred-1: protein function prediction from a protein interaction network using neighborhood analysis cell. Mol. Biol. Lett. (2014). doi:10.2478/s11658-014-0221-5
21. P. Chatterjee, S. Basu, J. Zubek, M. Kundu, M. Nasipuri, D. Plewczynski, PDP-CON: prediction of domain/linker residues in protein sequences using a consensus approach. J. Mol. Model. (2016). doi:10.1007/s00894-016-2933-0