International Journal of Life science & Pharma Research

STUDY OF PROTEIN SUBCELLULAR LOCALIZATION PREDICTION: A REVIEW

SHALINI KAUSHIK[1]*, USHA CHOUHAN[2] AND ASHOK DWIVEDI[3]

*[*1,2,3] Department of Mathematics, Bioinformatics and Computer applications, Maulana Azad National Institute of Technology, Bhopal, India.*

ABSTRACT

Protein subcellular localization, an important study on cytobiology, proteomics and drug design, directly relates to the functions of proteins at their prescribed cellular positions. Prediction of the subcellular localizations based on the machine learning has shown a great interest. This article focuses on the current research on extraction of protein sequence, machine learning algorithms and methods based on sequence and annotation. It was observed that features such as gene ontology, functional domains could improve the accuracy of prediction. Study of cells proteins, proteomics provides the annotations between the interaction groups and their associated functions. Knowing the localization of individual protein is very vital. Transport across the eukaryotic cells, comprising of subcellular compartments, organelles is very highly regulated and complex. In-silico subcellular localization has been an area of active research for years. The openly available methods that are of importance diverge in four aspects the underlying biological motivation, the computational method used, localization coverage, and reliability. This review has a study on the main events in the protein sorting process and widely used methods.

Keywords: Subcellular compartments, Gene Ontology, Combined features, machine learning

INTRODUCTION

Cells are highly ordered structure and contain various subcellular compartments that ensure the normal function operation of the entire cell. Subcellular organelles are bathed by cytosol and include – nucleus, mitochondria, endoplasmic reticulum, ribosomes, Golgi apparatus, lysosomes, peroxisomes, cytoskeleton etc. Different types of localization are present in different type of cells such as some localization are present in some type of cells while lack in different and also some localizations or compartments are common in all cell types. Two different types of cells named as eukaryotic cell and prokaryotic cell are illustrated in figure 1 and figure 2, which are labeled with the subcellular compartments present in them.

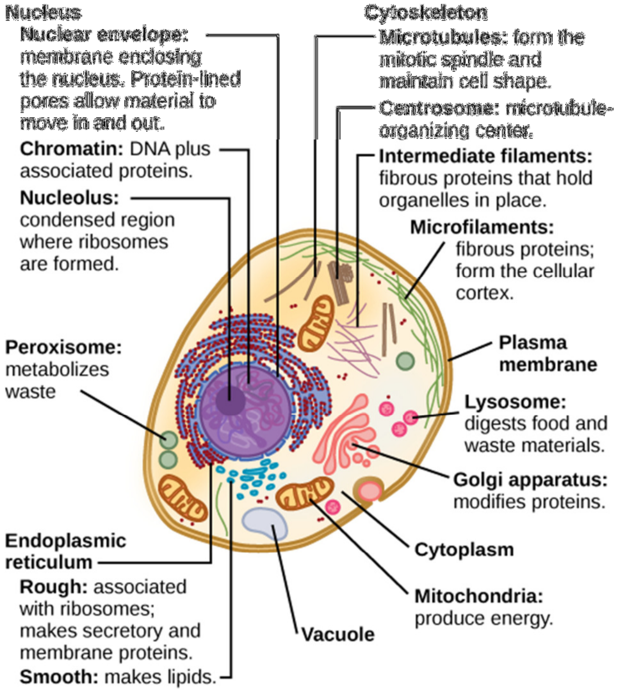
Nucleus
Nuclear envelope: membrane enclosing the nucleus. Protein-lined pores allow material to move in and out.
Chromatin: DNA plus associated proteins.
Nucleolus: condensed region where ribosomes are formed.
Cytoskeleton
Microtubules: form the mitotic spindle and maintain cell shape.
Centrosome: microtubule-organizing center.
Intermediate filaments: fibrous proteins that hold organelles in place.
Microfilaments: fibrous proteins; form the cellular cortex.
Peroxisome: metabolizes waste
Plasma membrane
Lysosome: digests food and waste materials.
Golgi apparatus: modifies proteins.
Cytoplasm
Endoplasmic reticulum
Rough: associated with ribosomes; makes secretory and membrane proteins.
Smooth: makes lipids.
Vacuole
Mitochondria: produce energy.

Figure 1
Illustration to show the 10 subcellular locations.

(nucleus, peroxisomes, endoplasmic reticulum, vacuole, mitochondria, Golgi apparatus, lysosomes, plasma membrane, cytoskeleton, and cytoplasm) of eukaryotic proteins and functions related to each subcellular compartment in the cell.

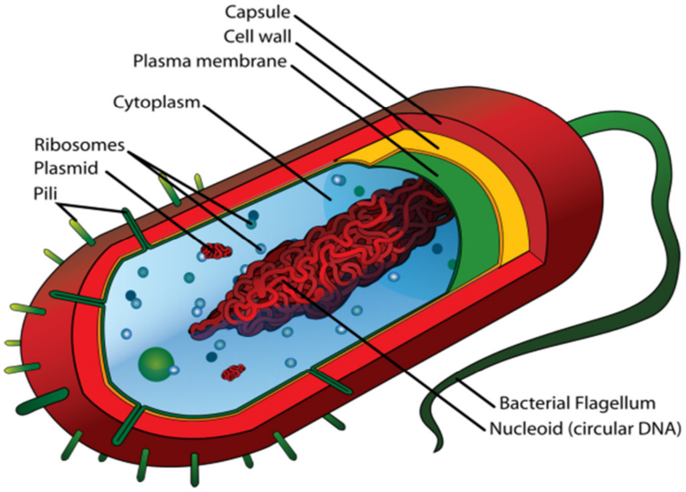

Figure 2
Structure of typical prokaryotic cell.

Table 1
Subcellular compartments with their functions.

Subcellular compartments	Cell	Function
Cell Wall	Plant, Fungi, Bacteria	Support Protection Allow H2O, O2, CO2 to diffuse in and out of cell.
Cell membrane	All cells	Support and protection. Barrier between cell and environment Maintain homeostasis.
Nucleus	All cells except prokaryotes	Control cell activities Contains hereditary material of cell.
Cytoplasm	All cells	Supports and protect cell organelles
Endoplasmic reticulum	All cells except prokaryotes	Carries materials through cell Aids in making proteins.
Ribosome	All cells	Synthesizes protein.
Mitochondrion	All cells except prokaryotes	Breaks down glucose molecules to release energy Site of aerobic cellular respiration
Vacuole	Plant cell have a single, large vacuole. Animal cells have small vacuoles.	Store water, food, metabolic, and toxic wastes.
Lysosome	Plant – uncommon Animal – common	Breaks down larger food molecules into smaller molecules. Digest old cell parts.
Chloroplast	Plant and Algae	Photosynthesis and release oxygen.
Golgi apparatus	All cells except prokaryotes	Modify proteins made by cells Package and export proteins.
Centrioles	Animal cells	Separate chromosome pairs during mitosis.

Table 1 explains various subcellular compartments present in the cell of the organism in the living system and it also describe the functions of subcellular localizations. The number of protein sequences deposited in pubic databases are increasing in a great extend because of explosive growth of biological data which in return need to be annotated experimentally for their function. Cell fractionation, electron microscopy, fluorescence microscopy are the methods applied to experimental validation of protein subcellular localization which is time consuming, laborious and costly. To overcome with this problem, computational methods are used to predict subcellular localization of protein. Computational methods of prediction subcellular localization of protein are much more reliable which produce subcellular localization as an output by taking some input information about protein. The input information that we are talking about are the related features of the particular protein that's why these features are also known as the fingerprints of the proteins. The general biological features and compartment specific features are explained as follows –

General Biological Features

The general biological features comprises of various features such as amino acid composition, dipeptide composition, relative solvent accessibility etc. In the amino acid composition, the prediction based on the n-peptide compositions has cited to be effective in PSL prediction. Suppose if n=1 then the n-peptide composition refers to the amino acid composition generating 21 dimensional feature vector (20 amino acid and a symbol X, for the others), indicating the frequency of amino acids in the sequence. If n=2, it refers to the di-peptide composition that gives a constant length of 21*21 di-peptides, indicating the frequency of the amino acid pairs in the sequence. The proteins present in the various compartments have different residue compositions and relative solvent accessibility. For example, CP proteins have balanced acidic and basic surface residues, while EC proteins have the acidic area in excess. The amino acid composition for both buried and exposed residues are considered with a cut off of 25% to represent the results obtained from SABLEII. There are basically two secondary structure elements encoding schemes. The secondary structure encoding scheme 1(SSE1) studies that the transmembrane a-helices are the most observed ones in the IM proteins whereas β-barrels are found in OM proteins. The secondary structural elements are vital for prediction of IM and OM localizations. But the SSE1 alone couldn't characterize the protein that are similar with the SSE compositions but localized in various subcellular compartments. OM proteins that are characterized by the β-strand might be similar to the proteins in the other compartments, repeat throughout the transmembrane domains. To depict the properties of protein even further three properties composition, transition, and distribution, are used to encode predictions of HYPROSP II. Composition gives us the global composition of SSE type in a protein, Transition studies the percentage of a specific SSE type followed by another. The distribution decodes the chain length in which the first 25, 50, 75 and 100% of the amino acids are located.

Compartment-specific Biological Features

The signal peptides (SIG), is one of the peculiar compartment specific biological features that defines the n-terminal peptides, between 15-40 amino acids long. They target the proteins for the translocation through the conventional secretory pathway. It is reported that if there is presence of the signal peptides, it indicates that the protein doesn't reside in the CP and various methods have been developed for the prediction. SignalP 3.0, a neural network-hidden Markov model-based method, to study the presence and location of signal peptide cleavage sites. Transmembrane a-helices (TMA) and transmembrane β-barrels (TMB) study about the IM and OM proteins. In TMA, the IM proteins are characterized by a-helices, a chain of 20-25 amino acids which traverse the IM. The presence of the protein in the IM is confirmed if there is one or more transmembrane a-helices are found. TMHMM 2.0, a hidden Markov model-based method helps to identify potential transmembrane a-helices while in TMB a greater number of the proteins that are located in the OM are characterized by the β-barrel structures. TMB-Hunt, a method that uses a k-nearest neighbor algorithm, is applied to differentiate transmembrane β-barrels from non-transmembrane β-barrels. Twin-arginine translocase (TAT) motifs export the proteins from CP to PP. The proteins translocated by twin-arginine take a unique twin-arginine motif useful to differentiate PP and non-PP proteins. TatP1.0, a neural network-based method is used to prediction of twin-arginine translocase motifs. Non-classical protein secretion (SEC) is one of the compartment specific biological features, in which the n-terminal peptide was very vital to export to an extracellular space. EC proteins can be secreted without classical N-terminal signal peptide. SecretomeP 2.0, a non-classical protein secretion prediction method, is incorporated in the method. Sequence and structure conservation: The localization sites of homologous sequences that are known could be very helpful for identifying the exact location of a protein. Both the sequence and structural homology approaches to identify the localization. PSLseq, based on pairwise sequence alignment of clustalW is used for the sequence homology modeling. In this approach we use secondary similarity comparison (PSLsse). Based on secondary structure elements predicted by HYPROSP II, SSEA carries the pairwise secondary structure alignment. In the approaches like sequence and structural homology, the known localization of the top-rank aligned protein is assigned to the query protein as its predicted localization.

PREVIOUS WORK

The general belief, subcellular localization of the protein predicts its function is reliable as the domain of the protein provides some admissible information about the function. That couldn't be only source about the protein as many properties studied during the prediction help in deciding the function of the protein. Sequence based methods, : (a) sorting-signals based methods, such as PSORT[1], WoLF PSORT, TargetP[2] and SignalP[3] , predict the localization via the recognition of N-terminal sorting signals in amino acid sequences; (b) composition-based methods, such as amino- acid compositions (AA)[4], amino-acid pair compositions (PairAA)[5], gapped amino-acid pair compositions (GapAA)[6], and pseudo amino-acid composition (PseAA)[7]; and (c) homology-based methods, such as Proteome Analyst, PairProSVM[8] and some other predicators[9,10] and also annotation based methods, that generally uses the coherence between the annotations and the subcellular localizations are known as the traditional methods for the prediction. Annette Hoglund et.al, in 2006 proposed an integrated way for the prediction of subcellular localization that focuses on the N-terminal sequences, Amino acid composition and the specific protein sequence motifs from the entrenched motif databases. These features help predicting the localization for a set of SVMs by providing input which was used for enhancing the prediction systems TargetLoc and MultiLoc.[11] TargetLoc, using the N-terminal sequences, predicts the four plant and three non-plant localizations while MultiLoc looks at all the 11 eukaryotic subcellular locations. MultiLoc, which has an accuracy of 75% in a cross validation test wins over PSORT method that has <60% accuracy. PSORTb 3.0[12] is still the most accurate SCL predictor with a greater coverage and recall also for the prokaryotes. It serves both as an online server (with associated email client for greater job updates) and is also an open source with easy installation allowing it to be used for many diversified purposes in any existing bioinformatics analysis methods. PSORTb 3.0 can handle wider range of prokaryotes and their subcategory localizations. It predicts the bacteria with atypical cell morphological characters with the help of the added predictive ability of archaeal protein SCL prediction. PSORTb 3.0 stands out over the other SCL prediction tools in terms of precision, accuracy and recall for all the bacterial proteins that were shown by Nancy Y. Yu, et.al, in 2010. Emily Chia-Yu Su, et.al, that provides the information on subcellular localization derived from hybrid prediction technique for gram-negative bacteria that integrates one-versus-one support vector machines model and structural homology model which has an accuracy of 93.7% and 93.2% by ten-fold cross validation.[13] Results show that biological features from gram-negative bacteria have shown a significant improvement while a slight downfall in the performance of homologous sequences couldn't be identified.

Recent Methodology

Among the annotation methods, Gene Ontology (GO), more attractive and informative, is a set of normalized data that annotates the function of gene and the gene products over various species and families. 'Ontology' basically refers to the systematic account of existence as the basic categories of being and their relations. GO annotations such as cellular location, molecular function, and biological process, of homologous proteins are often useful determining the functions of unascertained proteins in in-vivo.[14] Proteins, Nucleic acids, Membranes, and Organelles the cell components which are majorly located in the cells where as some located in the extracellular area. One or more arranged assemblies of molecular functions comprise sequence of events that are termed as biological functions. Molecular functions could be attained from the activity of the individual or the gene complexes at a molecular level. Gene Ontology Annotation (GOA) database provides annotations to non- redundant proteins of many species in UniProt Knowledgebase (UniProtKB) using normalized GO vocabularies.[15] The homogenization of the GO annotations and UniProtKB database could serve a source for the information of the subcellular localization. For a protein's accession number, GO terms could be redeemed from the GO annotation database file. The GO- based predictors can be classified into three categories: (a) using InterProScan for searching against dedicated protein databases[16,17] ; (b) to search against the GO annotation database such as Euk-OET-PLoc, applying the accession numbers of proteins, Hum-Ploc[18], Euk-mPLoc, "Euk-mPLoc 2.0"[19] a new predictor is generated by the information from the hybridization of gene ontology, functional domain and sequential evolution through three various types of pseudo amino acid composition. The overall jackknife success rate engineered by Euk-Ploc 2.0 is above 24% which is higher than the pairwise sequence identity of localized single and multiple location protein from the eukaryotic protein benchmark dataset of swiss-prot database which was not recorded ≥25% , Gneg-Ploc[20] and an integrated method; and (c) using the accession numbers of homologous proteins retrieved from BLAST to search against the GO annotation database, such as ProLoc- GO[21], iLoc-Virus[22], iLoc-Gneg[23] and Cell-PLoc 2.0.[24] GO annotation is said to be one of the effective method for the prediction of subcellular localization from the studies carried out over years. Multi label subcellular localization (SVM classifier with a new decision scheme, mGOASVM), on the semantic similarity among gene ontology (GO) features was projected for the formulation of semantic similarity vectors for classification.[25] Combination of the semantic contributors of their ancestors in the GO graph quoted to be a novel method which helped to encode a GO feature's semantics into a numerical value. This even helped inventing a new algorithm to measure the semantic similarity between two GO features. In 2011, a new method was developed for the subcellular localization, which integrates the homology based profile alignment methods and the functional domain Gene Ontology features.[26] The score of the feature vectors from these two methods is combined together to increase the performance. The paper also helps studying the different approaches for building GO vectors based on GO terms returned from the InterProScan. The results show that GO methods are parallel to profile alignment methods and are better than those based on the amino acid composition. It was also studied that these two methods could prove better results when combined than to the individual results. Shibiao Wan, in 2014, proposed HybridGO-Loc, a multi label subcellular localization predictor that dominates not only the GO term occurrences but also the inter term relationships.[27] This gives them an accuracy of 88.9% and 87.4% respectively, higher than the sophisticated predictors as iLoc-Virus (74.8%) and iLoc-Plant (68.1%). In 2012, Shibiao wan implemented the same method and found that for a given protein, the accession number of the homologs could be identified by the BLAST search. These, together with the original accession numbers are employed as the keys against the Gene annotation database to achieve a set of GO terms. For a given set of proteins, a set of T-GO terms is achieved by finding all the GO terms that are close to the training proteins from the GO database and then these closer terms form the base of the T-dimensional Euclidean space where GO vectors reside. Chin-Sheng Yu, et.al, in 2014, integrated the CELLO localization-predicting and BLAST homology-searching approaches, to study GO type categories including the subcellular localizations for the proteins queried. CELLO2GO, used for checking the correlation of two proteins with the same function has outperformed the PSORTb3.0[28] by 5% recording the recall and accuracy both with 96.5%. Xiao Wang, et.al, in 2016, used the GO information of apoptosis proteins and their homologous proteins revived from GOA database to calculate feature vectors and combined the distance weighted KNN algorithm. This helped them solving the data imbalance problem for the prediction of subcellular localization of apoptosis proteins. The prediction accuracy is directly proportional to the number of the homologous proteins. With the optimal conditions that are with the maximum number of the proteins the prediction accuracy was recorded as 96.8% by the jackknife test.[29]With the proteins appearing in various subcellular positions simultaneously and the present computational tools are updated with the obsolete data giving a chance of missing the latest databases. To overcome these issues, Xiaotong Guo, et.al, developed a multi-label classification algorithm to resolve first problem and combined several latest databases to improve prediction performance.[30] He proved that ensemble learning and feature reduction can improve the performance of weak learning problems by performing six experiments. As the first experiment, seven types of multi-labelled base classifiers that are, random forest (RF), decision tree (J48), k nearest neighbour (IBK), logistic regression for multi-label classification (IBLR_ML), k nearest neighbour for multi-label classification (MLkNN), lazy multi-label classification (BRkNN), and Hierarchy of multi-label learners (HOMER), are employed for a fivefold cross validation for 188 dimensional training set. IBLR_ML has highest AP value of the cross validation (59.37%), while HOMER has the lowest value (34.88%). For the second experiment, J48, IBLR_ML, MLkNN, and BRkNN, which have the higher AP values were combined using multi label ensemble classification and gave out the fivefold cross validation for the training sets. AP value comparison of three different ensemble classifiers is found to be 61.70% higher than the other two ensemble classifiers. Seven types of multi labelled base classifiers, employed in experiment 3 gives us the fivefold cross validation for PSSM-20_dimensional feature set. It results in a more efficient classification. IBLR_ML obtains the highest AP value of 62.01%. J48, IBLR_ML, MLkNN, and BRkNN are combined in the experiment 4 whose AP value was found to be 64.27%. Multi-labelled base classifiers, for the fivefold cross validation are employed for PseAAC-420 dimensional feature training set in the experiment 5 and AP value of IBLR_ML was found to be 56.36%. In the final experiment, fivefold cross validation was performed for the set of proteins using same method. The prediction of the protein subcellular localization, with the multi label features would still be complicate. The presence of the protein at various locations of their movement between the subcellular locations makes it complicated. Several methods were proposed to resolve the problem. Md. Al Mehedi Hasan, et.al, in 2017 proposed a method that focused on developing the efficient multi label protein subcellular localization, MKLoc. This multiple kernel learning (MKL) based SVM has shown better results than the other top systems (MDLoc, BNCs, YLoc+).[31] Hang Zhou, et.al, in 2016, proposed Hum-mPLoc 3.0[32], an amino acid based predictor, covering 12 human subcellular localizations. The sequences are from the multi view complementary features such as, context vocabulary annotation-based gene ontology (GO) terms, peptide-based functional domains, and residue-based statistical features. We propose a feature representation, HCM (Hidden Correlation Modelling) for determining the structural hierarchy of the domain knowledge databases. This creates more feature vectors by modelling the hidden correlations among their annotations. The experimental results have displayed that prediction accuracy of HCM has increased by 5-11% and F1 by 8-19%. Hum-Ploc 3.0 when applied on the whole human proteome reveals the protein's co-colonization preferences.

CONCLUSION

Experimental and Insilco techniques for the prediction and study of protein subcellular localization are an active area of research. Converting facts from the experiments to computational version and avoiding the complications at the same time would be challenging. Divination of proteins that shuffle among the compartments would be more complicated and many algorithms and biological motivations would be put up resolving this issue in the future. Prediction of subcellular localization computationally would surely help studying molecular biology in wider range. On analyzing the challenges in prediction, the functional characterization has shown a positive answer. GO annotations help improving the performance of prediction by appraising subcellular localization from various aspects.

ACKNOWLEDGEMENT

We shall ever, remain thankfully obliged to all those learned souls, our present and former teachers, known and unknown hand who directly or indirectly motivated us to achieve my goal and enlightened us with the touch of their knowledge and constant support. The author SK would like to acknowledge the Department of Biotechnology, New Delhi for Providing Bioinformatics Infrastructure facility at MANIT Bhopal to carry out this.

REFERENCES

1. Garg P, Sharma V, Chaudhari P, Roy N. SubCellProt: predicting protein subcellular localization using machine learning approaches. In silico biology. 2009 Jan 1;9(1, 2):35-44.
2. Emanuelsson O, Nielsen H, Brunak S, Von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of molecular biology. 2000 Jul 21;300(4):1005-16.
3. Emanuelsson O, Brunak S, Von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nature protocols. 2007 Apr 1;2(4):953-71.
4. Gao QB, Wang ZZ, Yan C, Du YH. Prediction of protein subcellular location using a combined feature of sequence. FEBS letters. 2005 Jun 20;579(16):3444-8.
5. Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. Journal of molecular biology. 1994 Apr 21;238(1):54-61.
6. Wang W, Geng X, Dou Y, Liu T, Zheng X. Predicting protein subcellular localization by pseudo amino acid composition with a segment-weighted and features-combined approach. Protein and peptide letters. 2011 May 1;18(5):480-7.
7. Wang W, Geng X, Dou Y, Liu T, Zheng X. Predicting protein subcellular localization by pseudo amino acid composition with a segment-weighted and features-combined approach. Protein and peptide letters. 2011 May 1;18(5):480-7.
8. Mak MW, Guo J, Kung SY. PairProSVM: protein subcellular localization based on local pairwise profile alignment and SVM. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2008 Jul 1;5(3):416-22.
9. Nair R, Rost B. Sequence conserved for subcellular localization. Protein Science. 2002 Dec 1;11(12):2836-47.
10. Wan S, Mak MW. Machine learning for protein subcellular localization prediction.

Walter de Gruyter GmbH & Co KG; 2015 May 19.

11. Höglund A, Dönnes P, Blum T, Adolph HW, Kohlbacher O. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. (2006): 1158-1165.
12. Nancy YY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010 Jul 1;26(13):1608-15.
13. Su CY, Lo A, Chiu HS, Sung TY, Hsu WL. Protein subcellular localization prediction based on compartment-specific biological features. InProceedings of IEEE Computational Systems Bioinformatics Conference (CSB'06): 14–18 August 2006; Stanford, California 2006 Aug (pp. 325-330).
14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA. Gene Ontology: tool for the unification of biology. Nature genetics. 2000 May 1;25(1):25-9.
15. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic acids research. 2006 Jan 1;34(suppl 1):D187-91.
16. Blum T, Briesemeister S, Kohlbacher O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC bioinformatics. 2009 Sep 1;10(1):274.
17. Mei S, Fei W, Zhou S. Gene ontology based transfer learning for protein subcellular localization. BMC bioinformatics. 2011 Feb 2;12(1):44.
18. Chou KC, Shen HB. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochemical and biophysical research communications. 2006 Aug 18;347(1):150-7.
19. Chou KC, Shen HB. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One. 2010 Apr 1;5(4):e9931.
20. Chou KC, Shen HB. Large-scale predictions of gram-negative bacterial protein subcellular locations. Journal of proteome research. 2006 Dec 1;5(12):3420-8.
21. Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY. ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC bioinformatics. 2008 Feb 1;9(1):80.
22. Xiao X, Wu ZC, Chou KC. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. Journal of Theoretical Biology. 2011 Sep 7;284(1):42-51.
23. Wang X, Zhang J, Li GZ. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC bioinformatics. 2015 Aug 25;16(12):S1.
24. Chou KC, Shen HB. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natural Science. 2010 Oct 29;2(10):1090.
25. Wang W, Geng X, Dou Y, Liu T, Zheng X. Predicting protein subcellular localization by pseudo amino acid composition with a segment-weighted and features-combined approach. Protein and peptide letters. 2011 May 1;18(5):480-7.
26. Wan S, Mak MW, Kung SY. Protein subcellular localization prediction based on profile alignment and Gene Ontology. InMachine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on 2011 Sep 18 (pp. 1-6). IEEE.
27. Wan S, Mak MW, Kung SY. Semantic similarity over gene ontology for multi-label protein subcellular localization. Engineering. 2013 Oct 16;5(10):68.
28. Wang W, Geng X, Dou Y, Liu T, Zheng X. Predicting protein subcellular localization by pseudo amino acid composition with a segment-weighted and features-combined approach. Protein and peptide letters. 2011 May 1;18(5):480-7.
29. Wang X, Li H, Zhang Q, Wang R. Predicting Subcellular Localization of Apoptosis Proteins Combining GO Features of Homologous Proteins and Distance Weighted KNN Classifier. BioMed research international. 2016 Apr 24;2016.

30. Guo X, Liu F, Ju Y, Wang Z, Wang C. Human Protein Subcellular Localization with Integrated Source and Multi-label Ensemble Classifier. Scientific Reports. 2016;6.

31. Hasan MA, Ahmad S, Molla MK. Protein subcellular localization prediction using multiple kernel learning based support vector machine. Molecular BioSystems. 2017;13(4):785-95.

32. Zhou H, Yang Y, Shen HB. Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. Bioinformatics. 2016 Dec 19:btw723.