Predicting protein subcellular location with network embedding and enrichment features

ABSTRACT

The subcellular location of a protein is highly related to its function. Identifying the location of a given protein is an essential step for investigating its related problems. Traditional experimental methods can produce solid determination. However, their limitations, such as high cost and low efficiency, are evident. Computational methods provide an alternative means to address these problems. Most previous methods constantly extract features from protein sequences or structures for building prediction models. In this study, we use two types of features and combine them to construct the model. The first feature type is extracted from a protein–protein interaction network to abstract the relationship between the encoded protein and other proteins. The second type is obtained from gene ontology and biological pathways to indicate the existing functions of the encoded protein. These features are analyzed using some feature selection methods. The final optimum features are adopted to build the model with recurrent neural network as the classification algorithm. Such model yields good performance with Matthews correlation coefficient of 0.844. A decision tree is used as a rule learning classifier to extract decision rules. Although the performance of decision rules is poor, they are valuable in revealing the molecular mechanism of proteins with different subcellular locations. The final analysis confirms the reliability of the extracted rules. The source code of the propose method is freely available at https://github. com/xypan1232/rnnloc

1. Introduction

Eukaryotic cells can be divided into many different compartments with distinct functions and morphological structures. These separated compartments are defined as various types of organelles on the basis of their specific composition and function, and all these organelles are necessary for normal cell survival. The cell nucleus, which is a representative organelle, is the main space for DNA storage and can coordinate the cellular activities, including growth, protein synthesis, and cell division [1]. The cell membrane, which is a double layer of lipids and proteins that surrounds the cell, acts as the barrier for cell protection and can control the movement of substances in and out of the cell [2]. The distinct roles of each type of organelles are partially attributed to different protein complexes located in the specific position and play specific functions. In particular, certain mechanisms are found to ensure that the required protein components are present at the corresponding sites, and the proteins must be localized in the appropriate compartment to ensure proper function [3]. Although the complex mechanism of protein trafficking has not been completely revealed, some features involved in protein trafficking have been identified, such as the presence of signal sequences within proteins. Proteins are recognized by corresponding receptors and span the membranes and translocate to the destination by relying on these specific signals [4]. The subcellular location can affect protein function because it determines the physiological environment that influences the properties of proteins and affect the subsequent interaction targets that result in specific functional pathways. The aberrant localizations of proteins are implicated in the pathogenesis that causes some diseases, including cancer and neurodegenerative diseases. Mutations in the nucleoporin NUP155 cause the reduction in nuclear envelope permeability, influencing the transport of mRNAs and proteins and leading to genetic disorders called familial atrial fibrillation [5]. The loss function of the peroxin-7 gene that encodes a peroxisomal import receptor causes rhizomelic chondrodysplasia punctata because of thedysregulation of protein trafficking [6]. Recent reports have suggested that the mislocalization of tau to dendritic spines contributes to a synaptic dysfunction that is related to neuronal injury and causes neurodegeneration [7]. Many publications have proposed that the mislocalization of nuclear proteins to the cytoplasm is a generalized mechanism for the inactivation of tumor suppressors [8–10]. Targeting protein localization has become a promising therapeutic strategy for several human diseases. Therefore, understanding the subcellular localization is crucial in elucidating the molecular function and valuable in developing the novel treatment of diseases. Traditional experimental methods, such as immunofluorescence, immunocytochemistry, and other tagged signal techniques, can provide solid evidence for characterizing the subcellular localization of proteins. However, these expensive and time-consuming methods only reveal extremely small parts of vast proteomes. In recent years, many efforts have been exerted to create the bioinformatics predictors of localization using computational methods. Various protein characteristics and different algorithms have been applied for predicting protein subcellular localization. As early in 1994, researchers have presented a method to discriminate the intracellular and extracellular proteins on the basis of the overall protein amino acid composition [11]. Similarly, neural networks have been applied to predict the subcellular location of proteins that depends on the amino acid composition, showing a prediction accuracy of 81% to three distinct subcellular locations in prokaryotic cell [12]. In addition to the methods based on sequence analysis, several homology-based predictors have been constructed to achieve improved performance in scope and the breadth of coverage [13,14]. A support vector machine-based method called ESLpred was developed using the combined features of dipeptide composition and blast scores with a relatively high accuracy [15]. The integrated data, including sequence motifs, isoelectric points, and transcript expression, were applied to build a Bayesian prediction model for predicting the localization of proteins [16]. Hum-mPLoc 3.0 integrates functional domain information based on Gene Ontology (GO) to predict protein locations using machine learning [17]. SubCons is an ensemble method for accurately predicting protein locations [18]. DeepLoc develops a hybrid deep learning model with attention mechanism to predict protein locations only from sequences [19]. node2loc predicts protein locations using the learned node embeddings from a protein–protein interaction (PPI) network with a recurrent neural network (RNN) [20]. These approaches have good performance for specific organisms and for certain localization categories. In this study, we construct a novel computational method based on the PPI network and gene functional annotation content. Many protein sequences are obtained from Swiss-Prot with the annotation of subcellular location, and all these proteins can be classified into 16 categories. We utilize this validated dataset to identify the most weighted features that are characterized by the PPI network, GO terms, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways through feature selection procedures and then apply these selected features to build an optimum classifier with RNN as the prediction engine. The optimum RNN classifier shows powerful ability to classify proteins into one of 16 categories of subcellular localization. We present decision rules for the discrimination of each given protein that can indicate the molecular mechanisms of proteins with different locations. These features used for constructing rules indicate the linkage between the function and localization of proteins.

1. Materials and methods

We first collect a benchmark dataset for protein localization from Swiss-Prot. Then, two types of features derived from the PPI network and gene functional annotation content (GO terms and KEGG pathways) are extracted to represent the proteins. These features are analyzed and ranked with some powerful feature selection methods. An incremental feature selection (IFS) [21] with three classifiers are evaluated to select the optimum features and construct the optimum classifier. Decision trees (DTs) [22] are used to extract the classification rules from the ranked features. The entire process is illustrated in Fig. 1. 2.1. Dataset In this study, we first use the dataset consisting of 5497 proteins from 16 subcellular locations that are retrieved from Swiss-Prot (http:// cn.expasy.org/release 54.0). This benchmark dataset has been used in our previous study [20]. The protein sequences in this set have similarity less than 0.7 through CD-HIT [23], and only sequences longer than 50 amino acids or shorter than 5000 amino acids are included. Proteins without the GO and KEGG pathway information are discarded, resulting in 4986 proteins. The data details are given in Table 1. 2.2. Feature extraction We extract two types of features to represent proteins. The first type is the embedding features learned from a PPI network using node2vec [24], and the second type is enrichment features derived from the functional annotation content, including GO terms and KEGG pathways. 2.2.1. Embedding features learned by node2vec from a PPI network Interacting proteins may share similar locations. Thus, we extract features under the context of the PPI network. node2vec, a powerful network embedding method [25,26], is applied to learn node embedding features from the PPI network that is downloaded from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING version 9.1) [27]. The PPI network consists of 2,425,314 interactions from 20,770 proteins. Here, the embedding dimension is set to 500, and other default parameters of node2vec (https://snap.stanford.edu/ node2vec/) are used. Thus, we extract a 500-D feature vector for each protein from the PPI network. 2.2.2. Enrichment features derived from GO terms and KEGG pathways Embedding features can abstract the relationship of proteins, and the essential properties of proteins should be included to construct a feature vector containing additional information on proteins. Different from sequence- or structure-based features used in several previous studies, we adopt the GO terms and KEGG pathways to represent the existing functions of proteins. These features are obtained by applying enrichment theory [28] on each protein. Thus, these features are called enrichment features. Given a protein and one GO term or KEGG pathway, let H1 be the set consisting of protein and its direct neighbor in the PPI network in STRING, and H2 be the set containing proteins annotated by the GO term or KEGG pathway. The enrichment feature of such protein on the GO term or KEGG pathway is defined as the hypergeometric test P value of H1 and H2. Thus, 297 KEGG pathways induce 297 enrichment features, and 20,681 GO terms produce other 20,681 enrichment features. A total of 20,978 enrichment features are obtained for each protein. For latter descriptions, the enrichment features are divided into two subtypes, namely, GO enrichment and KEGG enrichment features. 2.3. Feature selection Many features are used to represent investigated proteins, especially the enrichment features. Considering the efficiency of the classifier, a feature selection procedure is necessary. In this study, we first use Boruta feature selection [29] to rapidly identify relevant features that are evaluated through minimum redundancy maximum relevance (mRMR) [30] feature selection, resulting in a feature list. We then implement IFS [21] integrated with a supervised classifier on the basis of such list to select the optimum features and construct the optimum classifier. 2.3.1. Boruta feature selection A main component of Boruta [29] is the random forest (RF) classifier. It first creates copies of original data that are shuffled to obtain shuffled data. A RF is trained on the combined data from the original and shuffled data, and the feature importance score of the RF measures the feature importance. Second, the standardization of feature importance score for each feature is used as Z-score. Third, the maximum Z score of shadow features is selected as the MZSF. Finally, the original features with importance score greater than the MZSF are considered important, otherwise they are unimportant. The above process is repeated until all features are involved. Here, we use the Boruta program downloaded from <https://github.com/scikit-learn-contrib/boruta_py> and execute it with default parameters. The features selected by Boruta are analyzed through mRMR. 2.3.2. mRMR feature selection Some features are closely related to output labels, and some features are redundant because they are related to each other. mRMR [30] is proposed on the basis of the above idea. Mutual information is used to measure the relevance of features and labels and the redundancy between the features themselves. Features with high relevance and low redundancy are considered important and are highly ranked in the output feature list. mRMR sorts the features selected by Boruta in a feature list. The mRMR program is retrieved from http://home.penglab. com/proj/mRMR/index.htm. Default parameters are adopted. 2.3.3. IFS On the basis of the feature list yielded by mRMR, IFS is used to select the optimum features with a supervised classifier. We slightly change the original IFS to save computation time. First, various feature subsets with a step s are constructed, that is, feature subset 1 consists of the top s features, and feature subset 2 consists of the top 2 × s features, and so forth. A classifier is then trained on the samples consisting of the features from each feature subset, and the performance is evaluated using 10-fold cross-validation [31–35]. The features in the feature subset that produces the best performance are considered the optimum features, and the classifier with these features are termed the optimum classifier. 2.4. Synthetic minority over-sampling technique (SMOTE) As shown in Table 1, extreme class imbalance is found in the dataset, considerably affecting the model construction and performance. To resolve this issue, we apply SMOTE [36] to artificially synthesize the samples of minor classes. Thus, a balanced dataset is constructed and used for model training. For each class except the largest class, SMOTE randomly selects one sample in this class and calculates its Euclidean distance to all other samples in this class. The k nearest neighbors of this sample are then selected. Among these k neighbors, one neighbor is randomly selected to produce a new sample, which is defined as a linear combination of the sample and its selected neighbor and is placed into the class. With SMOTE, we obtain a dataset where each class contains the same number of samples. 2.5. Supervised classifiers integrated into IFS As mentioned in Section 2.3.3, IFS needs a supervised classifier for extracting its optimum features. In this study, three classifiers, namely, RF [37], instance-based learner (IBK) [38], and RNN [39], are used. 2.5.1. RF RF is a widely used classifier in tackling many biological data [40–43]. A RF consists of multiple DTs, and each DT is grown from a randomly selected feature subset and a bootstrap sample set. An advantage of RF is that it is a meta classifier that has good generalizability. 2.5.2. IBK IBK is a simple K-nearest neighbor (KNN) classifier. It can automatically select the K value through cross-validation and is an instancebased classifier compared with traditional KNN. One of its main output is a concept description with multiple stored instances and its previous performance. IBK contains a special scheme to reduce high memory requirements during model training. 2.5.3. RNN As previously demonstrated [20], RNN is a more powerful classifier for predicting protein locations than RF. Thus, we continue to use RNN as a supervised classifier in this study. RNN is a type of neural network with loop inside and is mainly developed for sequential data. 2.6. Rule learning using DTs Compared with the above black-box classifiers, an interpretable classifier is preferred to dig into the localization data. A DT [22] uses an optimized CART algorithm to construct binary trees using the features with the minimum Gini index. Each path from the root node (features) to a leaf node (output labels) forms a decision rule through conjunction when a tree is constructed. Decision rules are represented by IF-ELSE rules. In this study, DTs implemented in Scikit-learn are used. 2.7. Baseline method DeepLoc DeepLoc [19] is a deep learning-based method developed for predicting protein localizations from sequences alone. It uses RNNs for analyzing the entire sequence and attention mechanism for detecting important subsequences. DeepLoc is downloaded from https://github. com/ThanhTunggggg/DeepLoc. In this study, we also use 10-fold crossvalidation to evaluate the performance of DeepLoc with default parameters. 3. Results In this study, a 21,478-D feature vector for a protein is extracted. It consists of 500 embedding features and 20,978 enrichment features derived from GO terms and KEGG pathways. A rigorous feature selection procedure is developed to analyze these features. This section presents the detailed results of this procedure. 3.1. Results of Boruta and mRMR We first use Boruta to identify all relevant features from 21,478 features. A total of 4152 relevant features are accessed and ranked using mRMR in a feature list. Such list is provided in Supplementary Material S1. Among these 4152 features, 466 are embedding features, and others are enrichment features, including 3554 GO enrichment features and 132 KEGG enrichment features, as shown in Fig. 2(A). We also count the ranks of embedding and enrichment features in the list, as shown in Fig. 2(B). GO enrichment features receive higher ranks than the two other types of features. Most embedding features are assigned with low ranks, whereas, some of them have extremely high ranks. Embedding and enrichment features are all important for the prediction of protein subcellular location.

3.2. Results of IFS with black-box classifiers On the basis of feature list, we adopt IFS to select the optimum features for the three classifiers mentioned in Section 2.5. Step s is set to 5 for IBK and RF, whereas it is set to 10 for RNN because RNN is slower than IBK and RF. On each feature subset, three classifiers are trained and evaluated through 10-fold cross-validation. The performance, including the accuracies on 16 classes, overall accuracy, and Matthews correlation coefficient (MCC) [44], of each classifier on different numbers of top features are provided in Supplementary Material S2. For easy observation, three IFS curves are plotted in Fig. 3, indicating that RNN yields the highest MCC of 0.844 when using the top 550 features. RF yields the highest MCC of 0.810 when the top 2555 features are used, and IBK yields the highest MCC value of 0.826 when using the top 3830 features. The corresponding overall accuracies for RNN, RF, and IBK are 0.869, 0.840, and 0.853, respectively, as listed in Table 2. The accuracies on 16 classes yielded by the three classifiers are shown in Fig. 4. All results show that RNN can yield the best performance when using the smallest number of top features. As shown in Fig. 3, RNN yields a MCC of 0.830 when using the top 190 features and the performance stabilizes with increasing number of features. The overall accuracy is 0.858. The MCC and overall accuracy are slightly lower than the best values and are all higher than those of the two other classifiers. Considering the efficiency of classifiers, this RNN classifier is preferred for predicting protein subcellular location. In addition, we compare our method with another deep learningbased method DeepLoc. As shown in Table 2, our method is superior to DeepLoc. The potential reasons are belows: 1) The learned node embedding are different among different localizations since interacting proteins may share similar localizations; 2) Enrichment features derived from KEGG and GO information are also differently distributed among different locations. 3) SMOTE can reduce the impact of data imbalance since the number of proteins for different localization are quite different. 3.3. Results of IFS with DT Different from the above-used black-box classifiers, we use DT to clearly display the classification procedures and provide many information on the molecular mechanism of proteins with different subcellular locations. On the basis of the feature list yielded by mRMR, we perform step 10 to construct feature subsets. On each subset, a DT classifier is trained and assessed through 10-fold cross-validation. The performance of DT corresponding to the numbers of top features is listed in Supplementary Material S3. Similarly, an IFS curve is plotted in Fig. 5. DT yields the highest MCC value of 0.679 when using the top 3190 features, and the overall accuracy is 0.730. These values are lower than those of the optimum RNN, IBK, and RF classifiers. This result is reasonable because DT is a weak classifier. However, its classification procedures are completely open, indicating many biological insights. Accordingly, we use the 3190 features to construct the DT on all proteins and extract the corresponding decision rules, as listed in Supplementary Material S4. Analysis on these rules is helpful to reveal the molecular mechanism of proteins with different subcellular locations.

4. Discussion

With the rapid advances of proteomic research in recent years, numerous novel proteins that may play potential roles in diverse biological activities have been identified [45]. Understanding protein function is crucial to depict the hidden and complex biological processes during life. The cell is the basic functional unit of living organisms, and most proteins exert functional roles inside the cell, such as skeleton protein, or on the cell surface, such as membrane protein [46]. Subcellular localization is a decisive factor for the protein function that controls the target protein movement and determines its molecular interaction partners [47]. Knowledge on protein localization will remarkably contribute to characterizing the unknown functions of novel proteins and facilitate the progress of proteomic research. In this study, we apply several computational methods, including mRMR, IFS, RF, IBK, RNN, and DT, to construct an optimum RNN classifier using the validated protein localization data from Swiss-Prot, PPI data from STRING, and functional annotation content (GO and pathway). This classifier can classify each protein into one of the 16 categories of subcellular organelles with a high MCC value. Some decision rules are yielded in this study that can make the prediction procedures completely open and provide new insights into the study of different subcellular locations. The new findings reported in this study will be helpful in revealing the molecular function of given proteins in specific processes. The decision rules may provide a detailed and extended explanation about the mechanism of the protein in residing and operating in its environment. To verify the reliability of our findings, we search for existing evidence through literature review to determine the relevance between the protein subcellular location and the related GO terms or KEGG pathways. Here, we select the most key features and decision rules as examples to provide a detailed discussion, indicating that these GO terms or KEGG pathways may partially represent the molecular mechanism for the subcellular location of proteins.

Among the features that can indicate the category of “biological membrane,” GO: 0044743 is an important feature. It refers to the biological process of protein transmembrane import into intracellular organelle. Protein transport is a necessary function to living organisms that can move large molecules across the membrane of a cell or organelle. Several important processes, such as signal transduction, cell–cell interaction, and enzyme secretion, are dependent on protein transport [48]. Proteins attached to the surface or embedded in the membrane of a cell or organelle are termed as membrane proteins that are responsible for protein transport. Therefore, the enrichment of GO: 0044743 indicates a high probability for the given protein to be a membrane protein, suggesting the subcellular localization of the biological membrane. This result confirms the reliability of some decision rules and depicts an important relationship between protein localization and its biological function. The GO term GO: 0044425, which represents the cellular component of the membrane part, shows a direct indication for the classification of proteins located in the “biological membrane.” Logically, the high enrichment score of this GO term implies that the given protein is implicated in the formation of lipid bilayer along with the protein complexes embedded, directly indicates the object belonging to membrane proteins, and is located in the “biological membrane.” For example, peripheral myelin protein 22, which is strongly related to GO:0044425, is an integral membrane protein that is a major component of myelin in the peripheral nervous system [49]. This membrane protein plays a crucial role in pathological processes in the nervous system and functions in cell growth, differentiation, and apoptosis [50]. The term of GO: 0044425 has strong discriminative ability to identify the proteins located in the membrane. A relatively high enrichment score of GO: 0005654 is related to the rules in predicting the subcellular location of “nucleus” for given proteins. GO: 0005654 refers to the cellular component of nucleoplasm, indicating that the proteins related to this functional term are part of the nuclear content. As a GO term of cellular component, GO: 0005654 provides a direct indication for the proteins located in the nucleus, similar to GO: 0044425. Several experimental evidence have confirmed this decision rule. Nucleus accumbens-associated protein 1, which is encoded by the NACC1 gene and is associated to GO:0005654, is involved in tumor progression and tumor cell proliferation [51]. As demonstrated by immunohistochemical staining, NACC1 is locally expressed in the nucleus of carcinoma cells and shows potential correlation with tumor progression [52]. The POU4F3 gene with high association to GO:0005654 is exclusively located in the nucleus, as revealed by transient transfection studies; mutations in this gene can cause part of the protein product to be present in the cytoplasm and is associated with impaired hearing [53]. These findings are consistent with some rules indicating that proteins related to GO: 0005654 have the subcellular localization of the “nucleus.” The following GO term GO: 0005667 shows a powerful indicatory role for the identification of proteins located in the “nucleus.” GO: 0005667 refers to the cellular component of transcription factor complex, that is, a protein complex associated with DNA binding during transcription. Transcription is an elaborate process in which a particular segment of DNA is copied into the RNA with the function of several transcription factors, and this process usually occurs within the nucleus where the DNA is packaged into nucleosomes and chromatin structures [54]. Therefore, we infer that proteins involved in transcription are naturally located in the nucleus. For example, the TAF7 gene functions as a DNA-binding general transcription factor and plays a central role in regulating promoter responses to various activators and repressors in chromatin [55]. It effectively identifies the proteins with the subcellular localization of nucleus on the basis of the enrichment of GO: 0005667. GO: 0005615 refers to the cellular component of extracellular space and exhibits strong relevance to the identification of subcellular localization in the “extracellular space or cell surface” in some decision rules. The part outside the cells refers to the extracellular matrix, and the gene products are secreted from cells into the interstitial fluid or blood. Products, such as circulating immunoglobulin complex, secreted cytokine, and extracellular exosome, usually belong to the extracellular components and are annotated in GO:0005615. For instance, the GO: 0005615-related TNC gene encodes an extracellular matrix protein with a spatially and temporally restricted tissue distribution. It guides migrating neurons during development, synaptic plasticity, and neuronal regeneration [56]. TNC participates in the functional pathways of cell adhesion and may support the growth of epithelial tumors [57]. This result inspires us that the subcellular localization of proteins can partially reveal their functions and contribute to the understanding of their molecular mechanism. We then identify GO: 0005740 as an influential feature that can recognize the proteins located in the “mitochondrion.” GO: 0005740 represents the cellular component of mitochondrial envelope, which refers to the double lipid bilayer enclosing the mitochondrion and separating its contents from the cell cytoplasm. As an early term of mitochondrion, the enrichment of GO: 0005740 reasonably indicates the component located in the mitochondrion, which is consistent with some decision rules involving this GO term. We found a representative gene called UQCC2 that is annotated to GO: 0005740, providing strong evidence for the correlation between GO: 0005740 and subcellular localization of the “mitochondrion.” UQCC2 encodes a nucleoid protein localized to the mitochondria inner membrane and affects mitochondrial adenosine triphosphate production via its modulation of the respiratory chain activity [58]. Another publication reported that UQCC2 is expressed in mitochondrial nucleoids as shown by carboxy-terminally tagged signals and exhibits close relation to the amount of mitochondrial DNA, suggesting that this gene may regulate the organization and metabolism of mitochondrial DNA [59].

An extremely high enrichment score of GO: 0044432 is required to indicate proteins with subcellular localization in the “endoplasmic reticulum.” This GO only represents the cellular component of the endoplasmic reticulum part, which consists of the irregular network of unit membranes with occasional ribosomes adhered on the outer surface. The endoplasmic reticulum is a vital organelle that serves multiple functions, being important particularly in the synthesis, folding, modification, and transport of proteins. PLD4, which is annotated by GO:0044432, plays a role in many intracellular signal transduction events [60]. As demonstrated by immunocytochemistry experiment in splenic cells, PLD4 is locally expressed in the endoplasmic reticulum and Golgi apparatus. The spatially restricted expression patterns implies that this gene may have specific function in splenic marginal zone cells [61]. Another GO:0044432-related gene Zfand2b is observed to have an endoplasmic reticulum expression pattern, as shown by immunostaining [62]. Zfand2b is involved in the regulation of signalmediated translocation of proteins into the endoplasmic reticulum [63]. These evidence validate the indicatory ability of GO: 0044432 to predict such subcellular localization of proteins. Among the criteria that can indicate the proteins located in the “cytoplasm,” we observe that two GO terms play important roles in some decision rules. These terms are GO: 0005829 and GO: 0008047. GO: 0005829 refers to the cellular component of cytosol, indicating the part of the cytoplasm that does not contain organelles but contain other particulate matter, such as protein complexes. Various kinase complexes, oxidoreductase, and cytosolic ribosome belong to this term, and most chemical reactions of metabolism occur in the cytosol. A GO:0005829 related gene, RRAGB, is localized in the cytoplasm and participates in the relocalization of mTORC1 to the lysosomes [64,65]. GO: 0008047 refers to the molecular function of enzyme activator activity. Given that most enzymes and enzyme reaction occur in the cytoplasm, proteins related to this GO term should be localized in the “cytoplasm.” An example is that the protein Rab GTPase-activating protein 1, which is encoded by RABGAP1, is mainly located in the cytosol through immunofluorescence and plays a role in microtubule nucleation by centrosome [66]. These findings provide direct confirmation for the indicatory ability of the two GO terms. The “Golgi apparatus” is an essential organelle in eukaryotic organisms that can move molecules from the endoplasmic reticulum to their destination. Its major functions are the modification, sorting, and packaging of proteins for secretion [67]. GO: 0032580, which is the cellular component of Golgi cisterna membrane and refers to the lipid bilayer surrounding the sacs or folds of the Golgi apparatus, is strongly related to the classification of proteins located in the “Golgi apparatus” of some decision rules. The product encoded by APH1A is an endoprotease complex that can catalyze the intramembrane cleavage of integral membrane proteins, such as notch receptors [68]. APH1A is highly related to GO:0032580 and is predominantly localizes in the endoplasmic reticulum and Golgi apparatus [69]. This observation confirms the rules where the high enrichment score of GO: 0032580 can indicate the subcellular localization of the Golgi apparatus. Considering the length limitation, the last GO term with detailed discussion is GO: 0043202 that refers to the cellular component of lysosomal lumen. In our decision rules, the high enrichment score of these GO terms indicate the subcellular localization in the “vacuole.” Vacuoles are essentially enclosed compartments filled with water and contain inorganic and organic molecules, including enzymes. In animal cells, vacuoles play roles in many processes, such as exocytosis and endocytosis. Vacuoles are also important in autophagy, which aids in destroying and recycling broken proteins in cells [70]. Lysosomal lumen, which is also called as vacuolar lumen, is the storage of digestive enzymes. Therefore, the genes related to GO: 0043202 may be the coding genes of the corresponding protease. GLB1, which is related to GO: 0043202, has multiple transcript variants depending on alternative splicing. One of the isoforms can generate mature lysosomal enzyme that can catalyze the hydrolysis of a terminal beta-linked galactose residue from ganglioside substrates and other glycoconjugates [71]. The protein products of GLB1 are found in the lysosomes through immunoelectron microscopy [72]. This result provides evidence for the above-mentioned inference and some decision rules, indicating that GO: 0043202 can serve as an indicator for the classification of subcellular location in the “vacuole. In this section, an extended description is discussed about the mostly related features extracted from decision rules. Several GO terms have indicatory roles for the subcellular localization prediction, and these evidence have validated the reliability and efficacy of the proposed rules. 5. Conclusions This study investigates protein subcellular location with novel combination of features. These features indicate the two important essential properties of proteins. The first property is the relationship to other proteins, and the second property represents the functional properties of proteins. The final RNN classifier with proper features provides good performance. We construct several decision rules with open classification procedures. The analysis on the obtained rules confirms their reasonability. These rules are valuable in revealing the molecular mechanism of proteins with different subcellular locations. Supplementary data to this article can be found online at https:// doi.org/10.1016/j.bbapap.2020.140477. Funding This study was supported by the Shanghai Municipal Science and Technology Major Project [2017SHZDZX01], National Key R&D Program of China [2018YFC0910403] and National Natural Science Foundation of China (No. 61903248). Declaration of Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.