Progress in Machine Learning-based Predicting Subcellular Localizations of Proteins with Multiple Sites

Shanping Qiao
Shandong Provincial Key Laboratory of Network Based Intelligent Computing, School of Information Science and Engineering
University of Jinan
Jinan 250022, China
e-mail: qspzl@hotmail.com

***Abstract*—Prediction of protein subcellular localizations is a key step to determinate the functions of proteins. The experimental methods are both expensive and time-consuming. Therefore, many machine learning based computational approaches were proposed in the last two decades. Recently, it is proved that the number of proteins with multiple sites is rising. To determinate the subcellular localizations of this kind of proteins is a more difficult problem. Generally, dataset construction, feature representation, algorithm design and validation test are the four main aspects need to be considered in developing the predicting algorithms. This paper reviewed these four topics in detail. Although a great success has been got by many researchers, there are still a lot of problems need to study deeply.**

Keywords-protein subcellular localization; dataset; feature extraction; predicting algorithm; validation test

I. INTRODUCTION

According to cellular anatomy, a cell is constituted by many different components, such as cytoskeleton, mitochondrion, Golgi apparatus, etc. These parts are specialized to carry out different functions in a cell. Actually, most of these functions are performed by the proteins in it. However, to perform their functions properly, these proteins must be in the "designated regions of a cell," usually termed "subcellular locations." Otherwise, some kinds of diseases, such as Alzheimer and cancer, would occur [1]. So, knowledge of protein subcellular locations is very useful for both basic research and drug development.

The problem of predicting subcellular localizations of proteins needs to study deeply. First, the traditional methods to settle this problem, such as cell fractionation, X-ray crystallography, nuclear magnetic resonance, electron microscopy and fluorescence microscopy, are both expensive and time-consuming. Second, the number of the new-found proteins is growing exponentially (as shown in Fig. 1). Third, the gap between the sequence-known and the function-known proteins is becoming larger and larger (as depicted in Table I) [2]. Finally, existing of proteins with multiple sites has been a common phenomenon in fact [3].

Under this serious situation, it is high desirable to develop the computational methods to help to address this problem. In fact, a great progress has been made in this field in the last two decades [4,5,6,7]. In 2005, Chou [8] and Gardy [9] began to study the problem of predicting the subcellular localizations of proteins with multiple sites, respectively. Now, this investigation has become a hot and challenging topic.

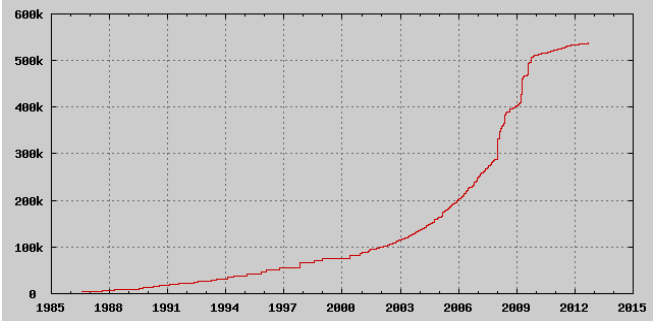

Figure 1. Number of entries in UniProtKB/Swiss-Prot.

TABLE I. NUMBER OF PROTEIN SEQUENCES IN THE UNIPROTKB/SWISS-PROT PROTEIN KNOWLEDGE BASE

Release Date	Database Version	Total	Experimental Annotations	Non-experimental Annotations
2008-7-22	14	390,787	64,733	167,972
2009-9-1	15.7	495,368	68,029	220,091
2010-7-13	2010_08	516,934	70,180	232,546
2011-7-27	2011_08	531,326	70,552	241,226
2012-5-16	2012_05	536,029	70,868	245,342

In this paper, the four aspects including dataset construction, feature representation, algorithm design, and validation test, are reviewed in details.

II. DATASET CONSTRUCTION

Dataset construction is a basic and important problem. It is necessary to construct a dataset to train the learning

machine. Protein sequences in the dataset are mainly collected from UnProtKB/Swiss-Port and some other special databases, such as NPD and PPDB. There are two steps in constructing a dataset. The first step is to get the protein sequences using some searching tools, and the second one is to remove the sequences that do not meet the criteria using some screen programs. Most researchers use the existing datasets proposed in some papers for convenience and comparison. The quality of the dataset influences the performance and generalization ability of the predicting algorithm significantly. The two measurements of a dataset are Label Density and Label Diversity.

Though the number of protein sequences is very large in the post-genome era, there are even two difficult problems which need the further study. One is the number of some kinds of protein sequences is inadequate actually. Just because of this problem, the second one which named imbalance of classes occurred. Although some kinds of methods are proposed [2], such as active learning and semi-supervised learning, it still requires the better approaches to deal with this situation.

III. FEATURE REPRESENTATION

A protein is composed of some amino acids and its sequence can be formulated by the following representation.

$$P = R_1 R_2 \cdots R_n \tag{1}$$

where P represents the protein sequence, R_1 is the first reside, R_2 the second, and so forth.

Amino acids have many properties, such as physicochemical properties, sequence properties, annotation properties, etc. In order to extract the feature from the sequence, some informative properties must be used. The fact is that too many features would bring about the disaster of high dimension. On the contrary, too few features would make some important information lost. So, how to extract and represent the features is a vital problem. Some widely used feature descriptors are shown below.

A. Sorting Signal Features

According to molecular biology, in a protein, there is a sorting signal which is composed of about 15 to 70 amino acids located in N- or C-terminal along the sequence. This signal conducts the protein to transport to the correct locations. The subcellular locations can be predicted through this kind of signal [10].

B. Sequence Features

In 1986, Nakashima proposed amino acid composition (AAC) to represent protein features. This representation had been widely used in the past. However, the correlations between amino acids were lost completely. In order to keep these correlations, Chou proposed pseudo amino acid composition (PseAAC) [11] representation in 2001. This descriptor puts the correlations factors into the feature vector. Actually, PseAAC is a general form of feature representation and many transformations have been studied [12]. Several tools which can calculate PseAAC have been developed. The representative ones are PseAAC [13], PseAAC-Builder [14] and Propy [15].

C. Annotation Features

Chou and his colleague began to use functional domain (FunD) [16] to study predicting protein subcellular localizations. FunD utilizes the annotation information in feature representation. First, it needs to determinate the length of the feature vector according to the number of annotations in FunD database. And then a searching process for the homology sequences of the query protein will be executed. Finally, a feature vector will be built on the base of these homology sequences. Gene ontology (GO) [17] is another annotation which describes a protein in three parts, function, biology process and cell composition. The method of building the feature vector is similar to FunD. The difference lies in the database. In fact, FunD and GO may be not available when the homology sequences are not found in the corresponding databases. In this case, PseAAC or other forms would be used instead.

Except the above feature extracting methods, there are many other descriptors (e.g. position specific scoring matrix (PSSM), protein-protein interaction (PPI) and motif) were used in this field. Although a number of feature representations have been proposed, the more informative features are still need the deeper exploration. Simultaneously, the representing forms and measurements also need study.

IV. ALGORITHM DESIGN

The high complexity of protein data result in the difficulty of designing the predicting algorithms with good performance. Computational approaches have been put forward for many years. Most of them are based on machine learning. To deal with the proteins with multiple sites, multi-label learning and ensemble with other related intelligent computing algorithms have become a very popular way on this topic.

A. Basic Machine Learning Algorithms

The K nearest neighbor (KNN) algorithm is a simple, intuitive and effective classification algorithm. This algorithm finds out the k neighbors in the given dataset of the query protein based on the distance measurement firstly. On the base of these k neighbors, this algorithm counts the number of every class, and classifies the query protein to the majority class. OET-KNN and Fuzzy-KNN are the variants of KNN.

Artificial neural network (ANN) is a model which simulates the work mechanism of human brain and contains many neurons which are connected each other through

different weights. BP, RBF and some other kinds of ANN are widely used in present [18].

Support vector machine (SVM) is suitable to solve the small sample, high-dimension and non-linear problems. The statistical learning theory and VC dimension are the core concepts of SVM. A main part to design SVM is to find a proper kernel function. With the aid of kernel function, a non-linear classification problem in a low dimension space can be transformed to a linear one after mapped it to a high dimension space. Recently, algorithms based on SVM are emerged massively [19,20].

Additionally, decision tree (DT), Bayes method, Hidden Markov Model (HMM), Covariant Discrimination Algorithm (CDA), Gauss Process, Wavelet Transform, etc., were all used in the past.

B. Multi-Label Learning Algorithms

Multi-label learning is fit to settle the problems having the multiple sematic meanings. Proteins with multiple sites are categorized to this case. To cope with the multiple sites, the correlations among labels must be considered. There are about three types which are the first-order strategy, the second-order strategy and the high-order strategy, to settle this problem. These algorithms are classified into two categories. One is problem transformation methods, and the other is algorithm adaption methods. For the former, representative algorithms include first-order approaches Binary Relevance and Classifier Chains, second-order approach Calibrated Label Ranking, and high-order approach Random k-labelsets. For the latter, representative algorithms include first-order approaches ML-KNN [21] and ML-RBF [22], second-order approach Rank-SVM, and high-order approach LEAD [23]. In recent years, multi-label learning algorithms are widely used in predicting the subcellular localizations of proteins with multiple sites [24,25].

C. Ensemble Learning Algorithms

The ensemble learning uses several algorithms to solve one problem on the base of an effective integration of these algorithms. Generalization ability is the most important goal for ensemble learning. There are three main aspects which are generating individual classifiers with large differences, ensemble strategy of these classifiers and the synthesis of output from each classifier, in design an ensemble learning algorithm. Dietterich designed the Error Correcting Output Codes (ECOC) algorithm. Freund and Schapire proposed the famous AdaBoost algorithm, and Breiman proposed the Bagging algorithm, respectively. Despite the successes have been gained [26], the ensemble learning is still a very hot topic nowadays, and many problems are not solved successfully yet. There is a huge study space needs to research deeply. It is high desirable to develop the more effective algorithms to deal with the multiplex proteins.

V. VALIDATION TEST

After the dataset construction, feature representation and algorithm design are all finished, the following step is to test the predicting algorithm, and hence to validate the performance and the generalization ability of the predicting algorithm. This topic contains two parts, test method and evaluation metrics.

A. Test Methods

Self-consistency test and cross-validation test are the two basic test methods. The latter can be classified into independent dataset test, sub-sampling test and jackknife test [27]. Among the above test methods, jackknife is regarded as the most objective one. But it is more time-consuming than others. After the test result got, the following metrics can be used to measure the performance of the predicting algorithm.

B. Evaluation Metrics

Let N be the number of samples in the test dataset, M the number of all single labels, L_i and L_i^* the actual and predicted label set of protein i respectively, and $\|\bullet\|$ the module of a set. The five metrics are formulated as follows.

- Subset-accuracy

$$\text{Subset - Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\Delta\left(L_i, L_i^*\right) \tag{2}$$

- Hamming-loss

$$\text{Hamming - Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left\|L_i \cup L_i^*\right\| - \left\|L_i \cap L_i^*\right\|}{M}\right) \tag{3}$$

- Accuracy

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left\|L_i \cap L_i^*\right\|}{\left\|L_i \cup L_i^*\right\|}\right) \tag{4}$$

- Recall

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left\|L_i \cap L_i^*\right\|}{\left\|L_i\right\|}\right) \tag{5}$$

- Precision

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left\|L_i \cap L_i^*\right\|}{\left\|L_i^*\right\|}\right) \tag{6}$$

VI. CONCLUSION

Proteins with multiple sites are essentially the multiple semantic objects. The multi-label algorithms are suitable to solve this kind of problem. A high quality dataset is the base of this problem. An informative feature representation and an algorithm with good performance are the core parts of this problem. An appropriate test method can give an objective judgment to the predicting algorithm. The computational approach is a feasible way to solve this problem. Because the data of proteins are so massive, complex and imbalanced that it is very difficult to deal with. There are still so many topics to be studied deeply in this field. These issues must be studied from a "system" point of view. The research of protein subcellular localizations is benefit to the development of cellular biology, proteomics, system biology and bioinformatics. At the same time, this research would promote the improvement of machine learning.

ACKNOWLEDGMENT

This research work was supported by National Natural Science Foundation of China under Grant No. 61302128 and 61070130.

REFERENCES

[1] L. L. Hu, K. Y. Feng, Y. D. Cai and K. C. Chou, "Using Protein-protein Interaction Network Information to Predict the Subcellular Locations of Proteins in Budding Yeast," Protein & Peptide Letters, Jun. 2012, vol. 19, pp. 644-651, doi:10.2174/092986612800494066.

[2] J. Z. Cao, W. Q. Liu and J. J. He, "Mining Proteins with Non-Experimental Annotations Based on an Active Sample Selection Strategy for Predicting Protein Subcellular Localization," PLoS One, Jun. 2013, vol. 8, pp. e67343, doi:10.1371/journal.pone.0067343.

[3] S. Zhang, X. Xia, J. Shen, Y. Zhou and Z. Sun, "DBMLoc: A Database of Proteins with Multiple Subcellular Localizations," BMC Bioinformatics, Feb. 2008, vol. 28, pp. 127, doi:10.1186/1471-2105-9-127.

[4] K. C. Chou and H. B. Shen, "Recent Progress in Protein Subcellular Location Prediction," Anal Biochem, Nov. 2007, vol. 370, pp. 1-16, doi:10.1016/j.ab.2007.07.006.

[5] K. Imai and K. Nakai, "Prediction of Subcellular Locations of Proteins: Where to Proceed? ," Proteomics, Nov. 2010, vol. 10, pp. 3970-3983, doi:10.1002/pmic.201000274.

[6] K. C. Chou, "Some Remarks on Predicting Multi-label Attributes in Molecular Biosystems," Mol Biosyst, Jun. 2013, vol. 9, pp. 1092-1100, doi:10.1039/c3mb25555g.

[7] P. F. Du and C. Xu, "Predicting Multisite Protein Subcellular Locations: Progress and Challenges," Expert Rev Proteomics, Jun. 2013, vol. 10, pp. 227-237, doi:10.1586/epr.13.16.

[8] K. C. Chou and Y. D. Cai, "Predicting Protein Localization in Budding Yeast," Bioinformatics, Apr. 2005, vol. 21, pp. 944-950, doi:10.1093/bioinformatics/bti104.

[9] J. L. Gardy, M. R. Laird and F. Chen, "PSORTb v.2.0: Expanded Prediction of Bacterial Protein Subcellular Localization and Insights Gained from Comparative Proteome Analysis," Bioinformatics, Mar. 2005, vol. 21, pp. 617-623, doi:10.1093/bioinformatics/bti057.

[10] N. Y. Yu, J. R. Wagner and M. R. Laird, "PSORTb 3.0: Improved Protein Subcellular Localization Prediction with Refined Localization Subcategories and Predictive Capabilities for All Prokaryotes," Bioinformatics, Jul. 2010, vol. 26, pp. 1608-1615, doi:10.1093/bioinformatics/btq249.

[11] K. C. Chou, "Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition," Proteins: Struct, Funct, Genet, May. 2001, vol. 43, pp. 246-255, doi:10.1002/prot.1035.

[12] K. C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," Curr Proteomics, Sep. 2009, vol. 6, pp. 262-274, doi:10.2174/157016409789973707.

[13] H. B. Shen and K. C. Chou, "PseAAC: A Flexible Web-server for Generating Various Kinds of Protein Pseudo Amino Acid Composition," Anal Biochem, Feb. 2008, vol. 373, pp. 386-388, doi:10.1016/j.ab.2007.10.012.

[14] P. F. Du, X. Wang and C. Xu, "PseAAC-Builder: A Cross-Platform Stand-Alone Program for Generating Various Special Chou's Pseudo-Amino Acid Compositions," Anal Biochem, Jun. 2012, vol. 425, pp. 117-119, doi:10.1016/j.ab.2012.03.015.

[15] D. S. Cao, Q. S. Xu and Y. Z. Liang, "Propy: A Tool to Generate Various Modes of Chou's PseAAC," Bioinformatics, Apr. 2013, vol. 29, pp. 960-962, doi:10.1093/bioinformatics/btt072.

[16] K. C. Chou and Y. D. Cai, "Using Functional Domain Composition and Support Vector Machines for Prediction of Protein Subcellular Location," J Biol Chem, Nov. 2002, vol. 277, pp. 45765-45769, doi:10.1074/jbc.M204161200.

[17] K. C. Chou and Y. D. Cai, "A New Hybrid Approach to Predict Subcellular Localization of Proteins by Incorporating Gene Ontology," Biochem Biophys Res Commun, Nov. 2003, vol. 311, pp. 743-747, doi:10.1016/j.bbrc.2003.10.062.

[18] C. Huang and J. Yuan, "Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites," Biosystems, Jul. 2013, vol. 113, pp. 50-57, doi:10.1016/j.biosystems.2013.04.005.

[19] S. B. Wan, M. W. Mak and S. Y. Kung, "mGOASVM: Multi-label Protein Subcellular Localization Based on Gene Ontology and Support Vector Machines," BMC Bioinformatics, 2012, vol. 13, pp. 290, doi:10.1186/1471-2105-13-290.

[20] T. H. Chang, L. C. Wu, T. Y. Lee, "EuLoc: A Web-server for Accurately Predict Protein Subcellular Localization in Eukaryotes by Incorporating Various Features of Sequence Segments into the General form of Chou's PseAAC," J Comput Aided Mol Des, Jan. 2013, vol. 27, pp. 91-103, doi:10.1007/s10822-012-9628-0.

[21] M. L. Zhang and Z. H. Zhou, "ML-KNN: A Lazy Learning Approach to Multi-Label Learning," Pattern Recognition. Jul. 2007, vol. 40, pp. 2038-2048, doi:10.1016/j.patcog.2006.12.019.

[22] M. L. Zhang, "ML-RBF: RBF neural network for multi-label learning.," Neural Process Lett, Apr. 2009, vol. 29, pp. 61-74, doi: 10.1007/s11063-009-9095-3.

[23] M. L. Zhang and Z. H. Zhou, "A Review on Multi-label Learning Algorithms," IEEE Transactions on Knowledge and Data Engineering, in press.

[24] S. Y. Mei, "Predicting Plant Protein Subcellular Multi-localization by Chou's PseAAC Formulation Based Multi-label Homolog Knowledge Transfer Learning," J Theor Biol, Oct. 2012, vol. 310, pp. 80-87, doi:10.1016/j.jtbi.2012.06.028.

[25] X. Wang, G. Z. Li and W. C. Lu, "Virus-ECC-mPLoc: A Multi-label Predictor for Predicting the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites Based on a General Form of Chou's Pseudo Amino Acid Composition," Protein Pept Lett, Mar. 2013, vol. 20, pp. 309-317, doi:10.2174/092986611320030009.

[26] G. H. Han, Z. G. Yu and V. Anh, "An Ensemble Method for Predicting Subnuclear Localizations from Primary Protein Structures," PLoS One, Feb. 2013, vol. 8, pp. e57225, doi:10.1371/journal.pone.0057225.

[27] K. C. Chou, "Prediction of Protein Structural Classes and Subcellular Locations," Curr. Protein Pept. Sci. Sep. 2000, vol. 1, pp. 171-208, doi:10.2174/1389203003381379.