Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 7, July 2018, pp. 1635-1640
DOI: 10.23940/ijpe.18.07.p28.16351640

Enhancing Subcellular Localization Prediction of Apoptosis Proteins by Ensemble SVMs with Random Under-Sampling

Xiao Wang*, Xiaohe Li, Hui Li, Hongwei Tao, Rong Wang, and Yinghui Meng

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

Abstract

The locations of apoptosis proteins in the cell determine their biological functions. So firstly, it is necessary to identify the subcellular locations of these proteins. In recent years, researchers have proposed a large number of prediction methods, specifically for apoptosis proteins. However, the vast majority of the methods have the following problems: (1) they utilize sequence-based methods rather than annotation-based methods for feature representation; (2) they ignore the negative impact of the imbalanced training dataset. In the work, a balanced predictor, GOIL-Apo, is proposed for dealing with the issues, which yields balanced solutions for predicting locations of apoptosis proteins. Firstly, by using gene ontology (GO) based methods, apoptosis proteins are represented as GO feature vectors. Subsequently, an ensemble classifier that fuses multiple SVMs with random under-sampling is proposed to deal with the data imbalance problem. Rigorous cross-validations show that the accuracy of GOIL-Apo is much better than the up-to-date predictors.

Keywords: apoptosis proteins; subcellular localization; imbalance learning; GO features

(Submitted on April 11, 2018; Revised on May 21, 2018; Accepted on June 16, 2018)

1. Introduction

Cell apoptosis, or programmed cell death, which was first proposed by Professor Kerr in 1972, is a normal and ubiquitous phenomenon in any multicellularity, and it is an autonomic ordered death process under certain physiological and pathological conditions [4]. This process occurs at all stages of individual growth and development. Apoptosis is not passive, but active, which involves a series of gene activation, expression and regulation, to pursue a better accommodation to the living environment [2]. Once the regulation of cell apoptosis deteriorates, diseases such as tumor, autoimmune diseases, and neurodegenerative diseases will emerge [5]. Therefore, it is necessary to understand the internal mechanism of apoptosis because it can help understand the pathogenesis of related diseases and provide a reliable basis for further treatment. Apoptosis proteins play an important role in programmed cell death. Accurately identifying subcellular locations of apoptosis proteins can help researchers explore functions of proteins and find drug targets. The conventional subcellular locations of proteins are gained via biochemical experiments. It is very time-consuming and expensive to do biological experiments, which cannot meet researchers' demands. So, researchers have developed a lot of automated computational methods.

For apoptosis proteins, a number of predictors have been developed. The feature extraction methods of these predictors included pseudo amino acid composition [1], distance frequency [8], and position-specific score matrices [3]. However, none of the existing predictors utilize the gene ontology (GO) based method to extract feature vectors. Our previous research [7] has proven that the GO-based method performs better than the sequence-based methods for predicting locations of apoptosis proteins. Consequently, we also use the GO-based method for feature representation in this paper.

The benchmark dataset CL317 that is cleared up by Chen and Li [1] is the most widely used dataset for predicting locations of apoptosis proteins. However, a typical imbalanced problem exists in this dataset, that is, the number of the proteins of different subcellular locations differs significantly. Unfortunately, the existing predictors have neglected the phenomenon of class imbalance. In this paper, by combining random under-sampling with the one-vs-rest SVMs, we develop a powerful ensemble classifier for dealing with the data imbalance problem.

* Corresponding author.
E-mail address: pandaxiaoxi@163.com

2. Materials and Methods

2.1. Dataset

In this paper, the benchmark dataset CL317 that contains 317 apoptosis proteins and covers six subcellular locations is utilized to test the prediction performance of our proposed predictor. These proteins are distributed as follows: 112 proteins belonging to cytoplasmic (Cy), 55 proteins to membrane (Me), 34 proteins to mitochondrial (Mi), 17 proteins to secreted (Se), 52 proteins to nuclear (Nu) and 47 proteins to endoplasmic reticulum (En), where cytoplasmic proteins are six times more than secreted proteins and three times more than mitochondrial proteins. An uneven sample distribution will exacerbate the prediction performance. Table 1. shows the protein distribution for each subcellular location in the dataset.

Table 1. Distribution for each subcellular location in the dataset

Subset	Subcellular location	Number of proteins
1	Cytoplasm	112
2	Membrane	55
3	Mitochondrion	34
4	Secreted	17
5	Nuclear	52
6	Endoplasmic reticulum	47
Total number	317	

2.2. Feature Representation

To develop an excellent predictor, extracting core and essential features of the proteins is a critical step, which accurately reflects the relationship between proteins and their subcellular locations. The GO-based feature extraction methods can take full advantage of the correlation between the annotation information of a protein and its subcellular location(s), so it is very effective for protein subcellular localization prediction. Our previous work [7] and other researches [6] have elaborated the legitimacy of using this method. Below, the detailed procedures of the GO-based method are given.

Protein P can be represented as:

$$P = [f_1, f_2, \cdots, f_\mu, \cdots, f_\omega]^T \tag{1}$$

where T is a transpose operator, f_μ is the μ-th feature, and ω is usually equal to the number of all GO terms in the GO database. However, with the rapid growth of GO items in the GO database, it may lead to high-dimension disaster. In this study, we use the GO subspace to avoid this problem. The details are as follows.

For each protein in the dataset, we searched the Swiss-Prot database (released on 24 July 2015) with the BLAST tool by setting the blast parameter expect value $E \le 0.001$ to search for homologous proteins. We used the accession numbers (ACs) of the homologous proteins as the keys to retrieve the relevant GO terms from the GOA database and put these GO terms into the set $\mathbb{S}$. After this process, suppose ω different GO terms are in the set $\mathbb{S}$, then these GO terms form a GO Euclidean space with ω dimensions.

For the protein P, the element f_μ of the feature vector can be denoted as:

$$f_\mu = \begin{cases} g_\mu & if\ GO\ hit \\ 0 & otherwise \end{cases} \tag{2}$$

where g_μ is the number of occurrences of the μ-th GO term in the protein P (term-frequency).

2.3. Under-sampling SVMs Ensemble Classifier

Due to the superior performance of SVM, SVM has become one of the most commonly used classification algorithms. Apoptosis protein subcellular localization prediction can be regarded as a multi-class classification problem. However, binary

SVM is only suitable for the binary classification. By applying the one-vs-rest strategy, SVM can be used to deal with the multi-class classification problem. The one-vs-rest strategy trains an independent SVM for each class. When constructing the SVM for each class, the samples that belong to the current class are labeled as positive samples, and the samples that do not belong to the class are negative ones. For example, given a multi-class benchmark dataset N that contains x protein samples and covers m subcellular locations, the dataset can be further classified into m subsets according to the different locations, that is, $N = N_1 \bigcup N_2 \bigcup \cdots \bigcup N_m$, where $N_i (i = 1, 2, \cdots, m)$ is the subset that contains the proteins that belong to the i-th subcellular location. For the i-th location, the training set T_i is the union of the positive set T_i^+ and the negative set T_i^-, where T_i^+ contains the proteins belonging to the i-th location, and T_i^- consists of the proteins that do not belong to the i-th location. They are constructed as following:

$$\begin{cases} T_i^+ = \{(p, +1) \mid p \in N_i\} \\ T_i^- = \{(q, -1) \mid q \notin N_i\} \end{cases} \tag{3}$$

where $(p, +1)$ represents protein p that belongs to the i-th location, and $(q, -1)$ represents protein q that does not belong to the i-th location. An independent SVM is trained from the training set for the the i-th subcellular location. For query protein y , its subcellular location is determined by:

$$C = \arg\max_{i=1}^{m} s_i(y) \tag{4}$$

where $s_i(y)$ is the prediction output of the i-th SVM to the protein y .

There is an obvious shortcoming in the one-vs-rest SVMs, that is, if uneven distributions exist in different categories, the sample number of the negative set will be significantly more than that of the positive set. It is clear that there is a typical imbalance in the CL317 dataset. A direct application will further exacerbate the imbalance. Therefore, we utilized the random under-sampling technique to solve the imbalance problem. The random under-sampling technique addresses the imbalance problem by decreasing the samples in majority classes, and the processes of training and predicting will be accelerated as a result of smaller training datasets. Note that missing important information caused by reducing the training sets potentially deteriorates the predictor performance. An effective way to circumvent this problem introduces ensemble learning to random under-sampling. For the i-th subcellular location, T_i^+ and T_i^- are the positive and negative training sets, respectively. The negative set T_i^- can be divided into n subsets by random under-sampling, i.e. $\{T_{(i,1)}^-, T_{(i,2)}^-, \cdots, T_{(i,k)}^-, \cdots, T_{(i,n)}^-\}$, where the size of each subset is equal to T_i^+. Then, n training sets can be obtained, where each training set is the union of one negative subset and T_i^+, i.e. $\{T_i^+ \bigcup T_{(i,1)}^-\}, \{T_i^+ \bigcup T_{(i,2)}^-\}, \cdots, \{T_i^+ \bigcup T_{(i,k)}^-\}, \cdots, \{T_i^+ \bigcup T_{(i,n)}^-\}$, and train one independent SVM from each training set. For query protein y , the prediction score of the i-th subcellular location is defined as:

$$S_i(y) = \frac{\sum_{k=1}^{n} r_k^i}{n} \tag{5}$$

where r_k^i denotes the prediction output of the k-th SVM. Finally, the predicted location of protein y is the location with the highest prediction score.

Figure 1. illustrates the workflow of our proposed predictor GOIL-Apo. In the upper part of the figure, the rebalanced SVM ensemble classifier is constructed for the individual subcellular location as described above. The lower part of the figure describes the prediction process.

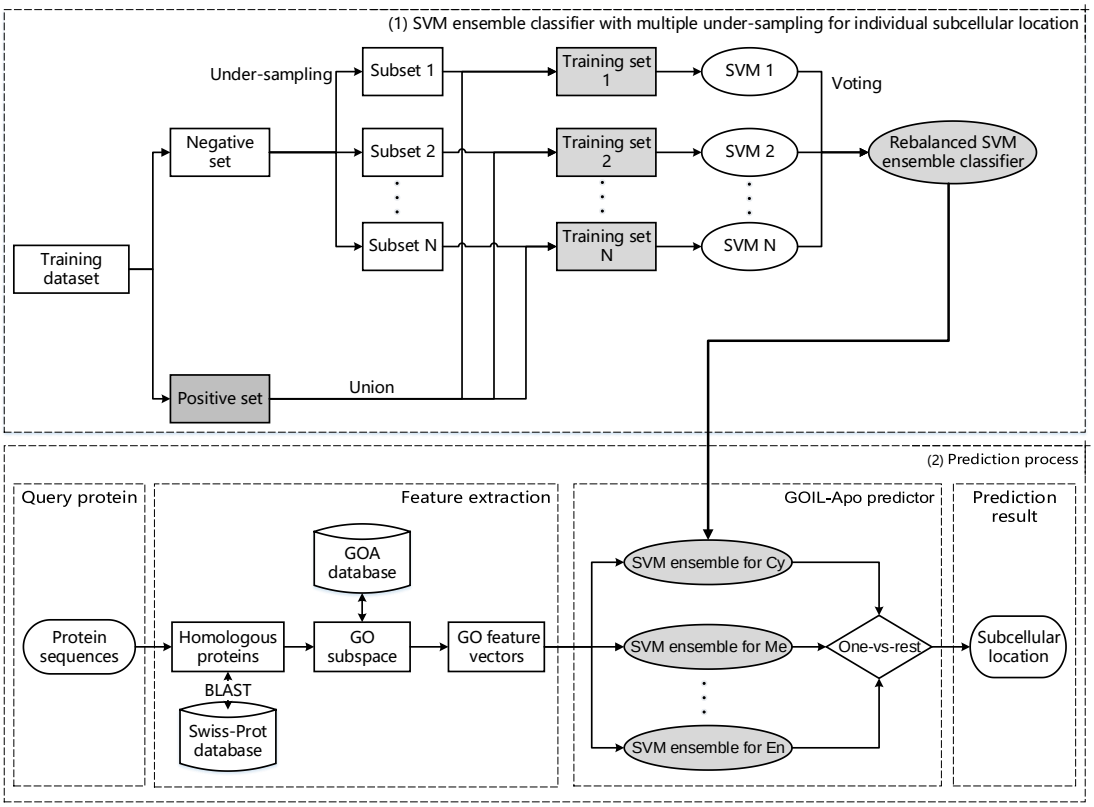

Figure 1. Workflow of the proposed method. The bold arrow from (1) pointing to (2) denotes that the independent rebalanced SVM ensemble classifier of (1) is trained for each subcellular location.

2.4. Performance Measures

In statistical prediction, independent inspection, k-fold cross-validation and Jackknife cross-validation are three common testing methods, where the Jackknife cross-validation is the most rigorous and objective testing method. The jackknife test will be used to test our proposed method.

To measure our achieved prediction results, four standard performance measures, sensitivity(SN), specificity (SP), Matthews correlation coefficient (MCC) and the overall accuracy(ACC) are used in this research. Sensitivity is regarded as the true positive rate, which is the metric of the predictor's ability to recognize where TP, FN, TN, and FP are abbreviations of true positive, false negative, true negative, and false positive, respectively.

$$ACC = \frac{\sum TP}{N}$$

$$SN = \frac{TP}{TP + FN} \tag{6}$$

$$SP = \frac{TN}{TN + FP}$$

$$MCC = \frac{TPTN - FPFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3. Results and Discussion

3.1. Effect of the Number of Homologous Proteins

Experimental results show that the overall prediction accuracy may be changed by increasing the number of the homologous proteins. We respectively selected 1 to 10 of the homologous proteins of one apoptosis protein and constructed 10 different

sizes of GO subspaces. The homology used first is the highest similarity. As concretely displayed in Figure 2, the horizontal coordinates represent the number of homologous proteins, and the longitudinal coordinates represent the overall prediction accuracy. As can be observed from the figure, the highest overall prediction accuracy in the figure is 97.8%, and the lowest rate is 97.1%. The prediction accuracy is essentially flat, especially from 3 to 7. The curve is very gentle and almost a straight line. This figure illustrates that with an increase in the number of homologous proteins, the more abundant GO annotation information provides little help in improving the prediction accuracy. The reason may be that different homologous proteins of an apoptosis protein are of high similarity. It is almost not useful to improve the performance for prediction proteins subcellular location.

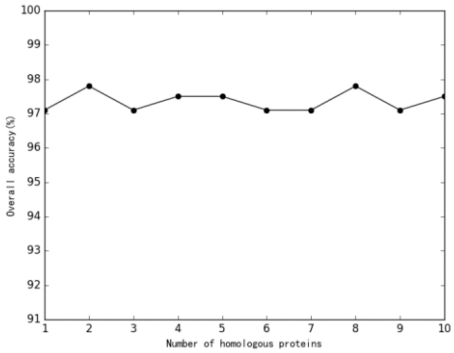

Figure 2. The graph shows how different number of homologous proteins affect the overall accuracy

3.2. Random Under-sampling Improving Performance

As we mentioned in the previous section, when the number of homologous proteins is 2, our predictor can achieve the best performance. In this case, we report SN, SP, MCC, and ACC to evaluate the prediction performance. Table 2 summarizes the performance comparisons between the under-sampling SVMs ensemble classifier and the one-vs-rest SVMs classifier based on an updated CL317 dataset by Jackknife cross-validation. In order to guarantee fairness, they both use the same features and the same test method on the same benchmark dataset. From Table 2, we can see that the overall accuracies reach 97.8% and 94.9%, respectively. The values of SN, SP, and MCC of the under-sampling SVMs ensemble classifier are consistently higher than or equal to that of the one-vs-rest SVM's. It demonstrates that the performance can be further improved by solving the imbalance problem. In detail, of the 17 apoptosis proteins secreted, 16 can be correctly predicted and all 34 apoptosis proteins in the mitochondria can be correctly predicted. The rates of SP can reach 98.9% and 100% on the mitochondrial and secreted, respectively. We have mentioned that MCC is an important measure for imbalanced learning. The MCC rates of our predictor are higher than that of the one-vs-rest SVMs, which fully illustrates solving the imbalance problem can be more efficient in enhancing the performances of the predictor. Furthermore, the predictor also performs satisfactorily on other subcellular locations, for example, the SN rates on nuclear and endoplasmic reticulum run up to 100%. In addition, the overall accuracy of the one-vs-rest SVMs can reach 94.9%, and the SN values are over 92%. The SP values are higher than 97%. It fully illustrates that the GO features are highly efficient for apoptosis protein subcellular localization prediction.

Table 2. Performance comparisons of our proposed method and one-vs-rest SVMs on CL317 dataset over Jackknife cross-validation

Location	Our proposed method			one-versus-rest SVMs		
	SN(%)	SP(%)	MCC	SN(%)	SP(%)	MCC
Cy	96.4	99.5	0.965	94.5	97.0	0.915
Me	96.4	99.2	0.956	94.5	99.2	0.944
Mi	100	98.9	0.953	97.1	98.5	0.921
Se	94.1	100	0.968	94.1	99.6	0.938
Nu	100	99.6	0.988	92.2	98.8	0.917
En	100	100	1.0	97.9	100	0.987
ACC	97.8			94.9		

3.3. Performance Comparison with the Existing Methods

In order to further test the prediction accuracy of our proposed predictor, we compare our method with other methods using a CL317 dataset by jackknife test. The SN of each location and the classification accuracy as performance measures are selected and Table 3 lists the prediction results. As can be seen from Table 3, the prediction accuracy of our method is much better than the up-to-date predictors. For some subcellular locations, such as mitochondrial(Mi) and secreted(Se), the SN rates are more than 5% higher than tri-gram encoding and APSLAP. It illustrates that the performance can be further improved by

alleviating the imbalance problem. For nuclear(Nu) and endoplasmic reticulum(En), the SN rates are about 4% higher than tri-gram encoding. However, we also realize that the SN rates of our predictor on some subcellular locations are lower than that of some predictors. Among these predictors, EN FKNN and APSLAP utilize ensemble classifiers to make the final prediction and actually obtain very high accuracies, which indicates that ensemble learning is very promising in improving the prediction performance. In conclusion, the GO-based feature extraction method is very effective for predicting locations of apoptosis proteins, and dealing with the imbalance problem can further enhance the prediction performance.

Table 3. Comparison of different methods on CL317 dataset

Methods	**SN（%）**						**ACC（%）**
	Cy	Me	Mi	Se	Nu	En	
ID	81.3	81.8	85.3	88.2	82.7	83.0	82.7
ID_SVM	91.1	89.1	79.4	58.8	73.1	87.2	84.2
DF_SVM	92.9	85.5	76.5	76.5	93.6	86.5	88.0
Auto_Cova	86.4	90.7	93.8	85.7	92.1	93.8	90.0
FKNN	93.8	92.7	82.4	76.5	90.4	93.6	90.9
PseAAC_SVM	93.8	90.9	85.3	76.5	90.4	95.7	91.1
EN_FKNN	98.2	83.6	79.4	82.4	90.4	97.9	91.5
PSSM-AC	93.8	90.9	91.2	82.4	86.5	95.7	91.5
APSLAP	99.1	89.1	85.3	88.2	84.3	95.8	92.4
TGE	98.2	96.4	94.1	82.4	96.2	95.7	95.9
GO DWKNN	98.2	98.2	97.1	94.1	90.2	100	96.8
Our Method	96.4	96.4	100	94.1	100	100	97.8

4. Conclusions

In this study, we developed a novel predictor called GOIL-Apo for apoptosis protein subcellular localization by combining GO features of homologous proteins with an under-sampling SVMs ensemble classifier. We construct the GO subspace to avoid the curse of dimensionality by selecting relevant GO terms instead of using all GO terms in the GO database. Under-sampling SVMs ensemble classifier was proposed to solve the class imbalance problem in the data set, which balanced the data set by reducing the number of large class samples and integrating the effect of multiple reductions. Results on the CL317 dataset show that the prediction accuracy can be improved by alleviating the imbalance problem.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61402422, 61501405), Key Project of Science and Technology Research of the Education Department of Henan Province (14A520063), and Doctoral Research Fund of Zhengzhou University of Light Industry (2013BSJJ082).

References

1. Ying-Li Chen and Qian-Zhong Li, “Prediction of the subcellular location of apoptosis proteins,” *Journal of Theoretical Biology*, vol. 245, no. 4, pp. 775-783, April 2007.
2. M Kettunen and K Brindle, “Apoptosis detection using magnetic resonance imaging and spectroscopy,” *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 47, no. 3-4, pp. 175-185, December 2005.
3. Taigang Liu, Xiaoqi Zheng, Chunhua Wang, and Jun Wang, “Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation,” *Protein and Peptide Letters*, vol. 17, no. 10, pp. 1264-1269, October 2010.
4. Michael D. Jacobson, Miguel Weil, and Martin C. Raff, “Programmed cell death in animal development,” *Cell*, vol. 88, no. 3, pp. 347-354, 1997.
5. R. Sgonc and J. Gruber, “Apoptosis detection: an overview,” *Experimental gerontology*, vol. 33, no. 6, pp. 525-533, 1998.
6. Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung, “GOA-SVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of chou’s pseudoamino acid composition,” *Journal of Theoretical Biology*, vol. 323, pp. 40-48, April 2013.
7. Xiao Wang, Hui Li, Qiuwen Zhang, and Rong Wang, “Predicting subcellular localization of apoptosis proteins combining go features of homologous proteins and distance weighted knn classifier,” *BioMed Research International*, vol. 2016, pp. 1-8, April 2016.
8. Li Zhang, Bo Liao, Dachao Li, and Wen Zhu, “A novel representation for apoptosis protein subcellular localization prediction using support vector machine,” *Journal of Theoretical Biology*, vol. 259, no. 2, pp. 361-365, July 2009.