Predicting Golgi-resident proteins in plants by incorporating N-terminal transmembrane domain information in the general form of Chou's pseudo-amino acid compositions

Yasen Jiao, Pufeng Du
School of Computer Science and Technology
Tianjin University
Tianjin 300072, China
{yasenjiao, pufengdu}@gmail.com

Xiaoquan Su
Qingdao Institute of BioEnergy and Bioprocess Technology
Chinese Academy of Sciences
Qingdao 266101, China
suxq@qibebt.ac.cn

***Abstract*—Knowing the subcellular location of a protein is an important step in understanding its biological functions. In this paper, we developed a new method to identify whether a protein is a Golgi-resident protein or not in plant cells. We proposed to incorporate transmembrane domain information and six different kinds of physicochemical properties of amino acids in the general form of Chou's pseudo-amino acid compositions. By using SVM based classifiers, our method achieved over 90% prediction accuracy in a 5-fold cross validation, which is much better than the other state-of-the-art methods**

***Keywords*—Golgi apparatus; subcellular localization; SVM; transmembrane domain;**

I. Introduction

The Golgi apparatus, which can be found in most eukaryotic cells, usually serves as an intermediate station in the secretory pathway that transports proteins out of a cell [1]. It also plays an important role in the post-translational modification process [2].

Knowing whether a protein is a Golgi-resident protein is an important step in understanding its possible biological functions. However, it is a big challenge to identify Golgi-resident proteins only from primary sequences. Although there are many Golgi-resident proteins that have been sequenced, the computational methods for identifying Golgi-resident proteins are still lacking. Especially, as the plant Golgi-resident proteins seem to have no targeting signals, it is more difficult to identify Golgi-resident proteins in plants. As far as we know, GolgiP [3] is the only existing method that is specially designed to predict Golgi-resident proteins in plants. By mapping known functional domains on protein sequences, GolgiP has achieved a promising prediction performance. However, as two classes of proteins in the training and testing dataset of GolgiP are highly imbalanced, its prediction performance may be over-estimated.

Over the last few years, various features of sequence have been used to predict subcellular locations of a protein [4], such as amino acid compositions, dipeptide compositions, functional domain information, GO information and many more. Although the exact biological mechanism, which directs proteins to Golgi apparatus, is still unknown, the transmembrane domain properties of Golgi-resident proteins seem to be important.

In this paper, we integrated the transmembrane domain properties in the general form of pseudo-amino acid compositions [5] to predict the Golgi-resident proteins. Without any knowledge of functional domains, our method performed better than the GolgiP method.

II. Materials and Methods

A. Dataset curation

As we focused on predicting Golgi proteins in plants, we chose *Arabidopsis thaliana* as the model organism in this study. We extracted all known protein sequences of *Arabidopsis thaliana* from UniProt database release 2014_06 [6]. To establish a high quality working dataset, several screening procedures were strictly followed.

(1) The protein sequences were categorized into two classes, the Golgi-resident class and the non-Golgi class. If a protein was annotated with at least one Golgi associated subcellular location, it would be assigned to the Golgi-resident class. Otherwise, it would be assigned to the non-Golgi class.

(2) For the Golgi-resident class, the proteins with ambiguous annotations, such as 'PROBABLE', 'POTENTIAL', 'POSSIBLE', or 'BY SIMILARITY' were discarded. The proteins with subcellular location annotations other than "Golgi apparatus" or "Golgi apparatus membrane" were also discarded.

All correspondence should be addressed to Pufeng Du (Email: PufengDu@gmail.com; Tel/Fax: +86-22-23689450). This work was supported by National Science Foundation of China (NSFC 61005041 and 61303161), Tianjin Natural Science Foundation (No. 12JCQNJC02300), China Postdoctoral Science Foundation (2012T50240 and 2013M530114) and Hong Kong Scholar Program.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

(3) For both classes, the protein sequences, which contain ambiguous symbols, like "X", "B" or "Z", were discarded. The protein sequences, which are less than 15 in length, were also discarded, as they could be fragments of other proteins.

(4) CD-HIT program [7] was applied to reduce the homologues and redundancy sequences. According to GolgiP [3], the similarity cut-off value was set to 95%.

Eventually, we obtained a dataset containing 204 Golgi-resident proteins and 8795 non-Golgi proteins. As it is a highly imbalanced dataset, directly evaluating our method on this dataset is likely to result in an over-estimated performance. Therefore, we used a sub-sampling strategy to create 10 balanced datasets. In every balanced dataset, the Golgi-resident proteins were the same 204 Golgi-resident proteins. The non-Golgi proteins were 204 proteins, which were randomly selected from the 8795 non-Golgi proteins. These ten datasets were denoted as GDS95-01, GDS95-02, ..., GDS95-10.

B. Sequence representations

The pseudo-amino acid compositions have been commonly used to represent protein sequences in predicting their subcellular locations [8]. We proposed a novel mode of the general form of Chou's pseudo-amino acid compositions [5], which incorporates the transmembrane domain information and various physicochemical properties of amino acids.

Let $r_1r_2...r_l$ be a protein sequence, where r_i (i=1, 2, ..., l) is the i-th amino acid on the sequence, l the length of the protein sequence. The representation contains four different parts. The first part is the amino acid compositions of the protein, which represents the occurrence frequencies of 20 different types of amino acids in the sequence. This part is defined as follows:

$$\boldsymbol{V}_1 = \frac{1}{l}[n_1, n_2, ..., n_{20}]^T, \tag{1}$$

where n_i (i = 1, 2, ..., 20) is the number of the i-th type of amino acid.

The second part is the di-peptide compositions of the protein, which represents the occurrence frequencies of 400 di-peptides. Because the sequence order information is lost in the amino acid compositions, this part provides some sequence order information. This part is defined as follows:

$$\boldsymbol{V}_2=\frac{1}{l\text{-}1}[d_1, d_2, ..., d_{400}]^T, \tag{2}$$

where d_i (i = 1, 2, ..., 400) is the number of the i-th dipeptide.

The third part describes the transmembrane domain information, as the transmembrane domain information has been proved to be useful in predicting Golgi-resident proteins [9].This part includes four parameters of the transmembrane domains: the average length of all transmembrane domains (l_{tmd}), the length of the first transmembrane domain (l_0), the number of transmembrane domains in the first 70 amino acids (m_{70}) and the probability that the N-terminal of the protein is on the cytoplasmic side of the membrane (p_{nc}). These parameters are computed for every protein sequence by using the TMHMM server [10]. This part can be denoted as follows:

$$\boldsymbol{V}_3=[l_{tmd}, l_0, m_{70}, p_{nc}]^T, \tag{3}$$

The last part is a serial of auto-correlation functions, which are calculated from the physicochemical properties of the amino acids. Let $H(k, j)$ be the k-th type of physicochemical properties of the j-th type of amino acid. The q-order auto-correlation function can be represented as follows:

$$f_{q,\,k} = \frac{1}{l\text{-}q}\sum_{i=1}^{l\text{-}q} h(k, r_i)h(k, r_{i+q}), \tag{4}$$

where $h(k, j)$ is the normalized value of $H(k, j)$, which can be defined as:

$$h(k,j) = \frac{H(k,j) - m_k}{s_k}, \tag{5}$$

where

$$m_k = \frac{1}{20}\sum_{j=1}^{20} H(k,j), \text{ and} \tag{6}$$

$$s_k = \sqrt{\frac{1}{20}\sum_{j=1}^{20}(H(k,j) - m_k)^2} \tag{7}$$

We applied six different physicochemical properties in this work, including hydrophobicity, hydrophilicity, side-chain mass, pK1 (α-$COOH$), pK2 (NH_3) and pI (at 25°C). The value of q varies from 1 to λ. Therefore, the forth part of the representation can be defined as the follows:

$$\boldsymbol{V}_4 = [f_{1,1}, f_{1,2}, ..., f_{1,\lambda}, f_{2,1}, f_{2,2}, ..., f_{2,\lambda}, ..., f_{6,1}, f_{6,2}, ..., f_{6,\lambda}]^T \tag{8}$$

By concatenating four parts, the whole representation is:

$$\boldsymbol{V} = [\boldsymbol{V}_1^T, \boldsymbol{V}_2^T, \boldsymbol{V}_3^T, \boldsymbol{V}_4^T]^T \tag{9}$$

To optimize the prediction performance, the parameter λ was set to 17.

C. Training and Testing method

The LIBSVM package was applied in this work [11]. We chose the Radial Basis Function (RBF) kernel and used a grid search method to optimize the parameters c and γ.

We carried out two different evaluation experiments, 5-fold cross validation and independent dataset test, to estimate the prediction performance of our method. In the independent

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

dataset test, 80% of the sequences were randomly selected to train the predictor and the remaining 20% were used to test the predictor.

We used four statistics to measure the prediction performance of our methods, including sensitivity (*Sen*), specificity (*Spe*), accuracy (*Acc*) and Matthew's Correlation Coefficients (*MCC*). These statistics are defined as the follows:

$$Sen = \frac{TP}{TP+FN}, \tag{10}$$

$$Spe = \frac{TN}{TN+FP}, \tag{11}$$

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, \text{ and} \tag{12}$$

$$MCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{13}$$

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, false positives and false negatives in the 5-fold cross validation or the independent dataset test.

III. Results and Discussions

A. Parameter Calibration

In order to optimize the prediction performance of our method, the value of parameter λ should be chosen carefully. We used GDS95-01 dataset to optimize this parameter. By generating sequence representations with λ from 1 to 25, we found that λ=17 could achieve the best prediction performance in 5-fold cross validation (Figure 1). Therefore, parameter λ was set to 17 in all following analysis. This may not be the optimized value for every dataset. However, optimizing parameters on every dataset is not a reasonable choice in practical applications and may result in over-estimated performance.

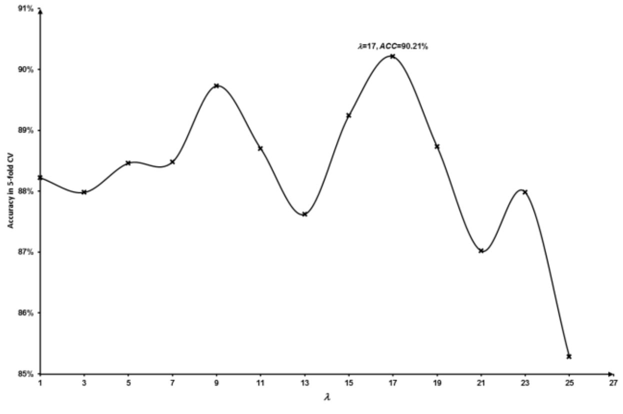

Fig. 1. Parameter calibration curve of pseudo-amino acid compositions.

TABLE I. Prediction performance

Dataset[a]	Sen[b]	Spe[c]	Acc[d]	MCC[e]
GDS95-01	86.3%	94.1%	90.2%	0.806
GDS95-02	83.8%	88.7%	86.3%	0.726
GDS95-03	82.8%	83.8%	83.3%	0.667
GDS95-04	80.4%	88.3%	84.4%	0.688
GDS95-05	81.8%	90.2%	86.0%	0.723
GDS95-06	81.4%	90.7%	86.0%	0.724
GDS95-07	82.3%	88.3%	85.3%	0.707
GDS95-08	80.9%	87.7%	84.3%	0.688
GDS95-09	81.8%	85.8%	83.8%	0.677
GDS95-10	86.8%	88.8%	87.8%	0.755

[a] Dataset names;

[b] Sensitivity, as defined in Eq. (10);

[c] Specificity, as defined in Eq. (11);

[d] Accuracy, as defined in Eq. (12);

[e] Matthew's correlation coefficient, as defined in Eq. (13).

B. Prediction performance in different datasets

As we have mentioned in the Dataset section, we generated 10 balanced datasets for evaluating the prediction performance. We carried out 5-fold cross validation on every dataset. The prediction performance can be found in Table 1. The prediction accuracy on different dataset varies in a range from 83% to 90%. The sensitivity varies from 81% to 87%. As the parameter of our algorithm was only optimized on the GDS95-01 dataset, these results indicated that our classifier was not over-optimized.

C. Performance comparison

We carried out an independent dataset test on GDS95-01 to compare the prediction performance of our method with GolgiP [3], which is the state-of-the-art method in predicting Golgi-resident proteins. Both predictors were tested by the same testing dataset. Our method was trained and optimized without any information from the testing dataset. Our method performs much better than the GolgiP predictor not only in prediction accuracy, but also in sensitivity (Table 2).

It should be noted that GolgiP reported much higher performances when evaluated on their own training and testing dataset. However, it should also be noted that the dataset for training and testing GolgiP is highly imbalanced. The non-Golgi proteins are much more than the Golgi-resident proteins in GolgiP dataset. This will result in an over-fitted predictor that will recognize the non-Golgi proteins much better than the Golgi-resident proteins, regardless to which kind of features it took into consideration. When GolgiP is tested by a balanced dataset, although its ability to recognize non-Golgi proteins is still promising, its ability in recognizing Golgi-resident

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

proteins becomes more visible. Therefore, we can observe that GolgiP performs better than our method in specificity, but worse in sensitivity, accuracy or MCC.

TABLE II. PERFORMANCE COMPARISON

Methods	Sen	Spe	Acc	MCC
Our method	90.0%	97.5%	93.8%	0.877
GolgiP-Comprehensive	57.5%	100%	78.8%	0.635

D. Further performance analysis

In our method, we used a number of different features to represent the protein sequences. However, it is interesting to investigate which feature is more useful in our method. Therefore, we used different sequence representations to test our method on GDS95-01 dataset. As shown in Table 3, the di-peptide composition is more useful to improve the sensitivity, while both the transmembrane domain information and the di-peptide compositions are useful to improve the specificity.

To further compare the performance of our method and the GolgiP, we carried out comparison on all 10 datasets. As shown in Table 4, our method performed better than GolgiP on every dataset in sensitivity. As GolgiP was trained with an imbalanced dataset, in which non-Golgi proteins are over ten times of Golgi-resident proteins. The almost 100% specificity of GolgiP is not indicating its superior performance, but a sign of potential over-fitting problems in its training. Although the non-Golgi proteins are much more than Golgi-resident proteins in nature, it is not reasonable to train a classifier with imbalanced dataset. This is simply because the imbalanced dataset will easily result in over-fitting problems [12, 13].

TABLE III. REPRESENTATION EFFECT ANALYSIS

Representations	Sen	Spe	Acc	MCC
PseACC	83.3%	85.8%	84.6%	0.691
PseACC+dip[f]	85.8%	90.2%	88.0%	0.761
PseAcc+dip+Trans[g]	86.3%	94.1%	90.2%	0.806

[f] dip: di-peptide composition

[g] Trans: Transmembrane domain information

TABLE IV. COMPREHENSIVE PERFORMANCE COMPARISON

S[j]	GolgiP-Comprehensive				Our Method			
	Sen	*Spe*	*Acc*	*MCC*	*Sen*	*Spe*	*Acc*	*MCC*
01	57.5%	100%	78.8%	0.635	90.0%	97.5%	93.8%	0.877
02	65.0%	97.5%	81.3%	0.661	85.0%	82.5%	83.8%	0.675
03	65.0%	100%	82.5%	0.694	82.5%	80.0%	81.3%	0.625
04	72.5%	100%	86.3%	0.754	90.0%	97.5%	93.8%	0.877
05	67.5%	100%	83.8%	0.714	75.0%	92.5%	83.8%	0.686
06	72.5%	100%	86.3%	0.754	77.5%	97.5%	87.5%	0.765
07	70.0%	100%	85.0%	0.734	75.0%	97.5%	86.3%	0.744
08	65.0%	100%	82.5%	0.694	82.5%	85.0%	83.8%	0.675
09	65.0%	100%	82.5%	0.694	72.5%	82.5%	77.5%	0.553
10	80.0%	97.5%	88.8%	0.787	92.5%	100%	96.3%	0.928

[h] Dataset id, i.e. 01 is GDS95-01

IV. CONCLUSIONS

In this paper, we present a new method for predicting Golgi-resident proteins in plants. Although we did not incorporate any functional domain related information, it still performs better than the other state-of-the-art methods. As there are no significant targeting signals for the Golgi-resident proteins in plants, we hope our method would be useful in identifying Golgi-proteins in plants.

REFERENCES

[1] Fabene PF, Bentivoglio M (1998) 1898-1998: Camillo Golgi and "the Golgi": one hundred years of terminological clones. Brain Res Bull 47: 195–198.

[2] Prydz K, Dalen KT (2000) Synthesis and sorting of proteoglycans. J Cell Sci 113 Pt 2: 193–205.

[3] Chou W-C, Yin Y, Xu Y (2010) GolgiP: prediction of Golgi-resident proteins in plants. Bioinformatics 26: 2464–2465. doi:10.1093/bioinformatics/btq446.

[4] Du P, Xu C (2013) Predicting multisite protein subcellular locations: progress and challenges. Expert Rev Proteomics 10: 227–237. doi:10.1586/epr.13.16.

[5] Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273: 236–247. doi:10.1016/j.jtbi.2010.12.024.

[6] The UniProt Consortium (2012) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Research 41: D43–D47. doi:10.1093/nar/gks1068.

[7] Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150–3152. doi:10.1093/bioinformatics/bts565.

[8] Chou K-C (2013) Some remarks on predicting multi-label attributes in molecular biosystems. Mol Biosyst 9: 1092–1100. doi:10.1039/c3mb25555g.

[9] Yuan Z, Teasdale RD (2002) Prediction of Golgi Type II membrane proteins based on their transmembrane domains. Bioinformatics 18: 1109–1115.

[10] Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580. doi:10.1006/jmbi.2000.4315.

[11] Chang C-C, Lin C-J (2011) LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2: 27:1–27:27. doi:10.1145/1961189.1961199.

[12] SUN Y, WONG AKC, KAMEL MS (2009) CLASSIFICATION OF IMBALANCED DATA: A REVIEW. Int J Patt Recogn Artif Intell 23: 687–719. doi:10.1142/S0218001409007326.

[13] He J, Gu H, Liu W (2012) Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. PLoS ONE 7: e37155. doi:10.1371/journal.pone.0037155.