Chemometrics and Intelligent Laboratory Systems 124 (2013) 9–13

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Prediction of Golgi-resident protein types by using feature selection technique

Hui Ding [a,*], Shou-Hui Guo [a], En-Ze Deng [a], Lu-Feng Yuan [a], Feng-Biao Guo [a], Jian Huang [a], Nini Rao [a], Wei Chen [b,*], Hao Lin [a,*]

[a] Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China
[b] Department of Physics, Center for Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan 063000, China

ARTICLE INFO

Article history:
Received 10 April 2012
Received in revised form 6 March 2013
Accepted 10 March 2013
Available online 18 March 2013

Keywords:
cis-Golgi proteins
trans-Golgi proteins
Analysis of variance
Support vector machine
g-Gap dipeptide

ABSTRACT

The functions of Golgi apparatus are to store, package and distribute proteins. Knowing the type of a Golgi-resident protein will provide in-depth insight into its function. In this study, we developed a support vector machine-based method to identify the types of Golgi-resident proteins by using only amino acid sequence information. A strictly and objective dataset including 137 proteins with the sequence identity <25% was used for training and testing the support vector machine. The analysis of variance was proposed to find out the optimized feature set. In the leave-one-out cross-validation, the maximum overall accuracy of 85.4% was achieved with the area under the receiver operating characteristic curves of 0.878. The results demonstrate that the proposed method can be used to discriminate the types of Golgi-resident proteins. An on-line server subGolgi is freely available at http://lin.uestc.edu.cn/server/subGolgi2.

1. Introduction

The Golgi apparatus is an important eukaryotic organelle, which is consisted of a stack of membrane-bounded cisternae located between the endoplasmic reticulum and the cell surface. Many different enzymes and other proteins are retained in Golgi apparatus to perform their various synthetic activities [1]. The cis-Golgi and trans-Golgi apparatus are thought to be specialized cisternae leading proteins in and out of the Golgi apparatus [2]. The function of cis-Golgi proteins is to receive and process the biosynthetic output from endoplasmic reticulum. Then the proteins modified by cis-Golgi proteins are packaged and sent to the required destination by trans-Golgi proteins. Many studies have demonstrated that neurodegenerative diseases, such as Parkinson's disease [3] and Alzheimer's disease [4] are associated with the defects in the Golgi apparatus [5]. Therefore, correctly identifying the types of Golgi-resident proteins is very important for fully understanding Golgi-resident proteins' roles in the process of transport and modification of transited proteins.

Experimentally identifying the types of Golgi-resident proteins is a good starting point for in-depth study of their functions. However, it is time-consuming and costly for biochemical experiments to systematically investigate the types of Golgi-resident proteins because large amounts of potential proteins are required to interrogate [6]. Phylogenetic tree is an accepted method to identify the types of Golgi-resident proteins. Although this method is not particularly expensive, it is more time-consuming. Furthermore, phylogenetic tree can not provide any information about the proteins whose homologies can not be found in benchmark dataset.

Machine learning techniques are powerful tools for the annotation of protein functions [7]. Many methods for predicting protein localizations have provided predictive information about Golgi-resident proteins [8,9]. However, only a limited number of methods have been specifically designed for the study of Golgi-resident proteins. Yuan and Teasdale have predicted the Golgi type II membrane proteins based on their transmembrane domains [10]. However, only 25% of Golgi-resident proteins were estimated to contain transmembrane regions in *Arabidopsis thaliana* [11]. Chou et al. [6] have designed a server called GolgiP for the prediction of Golgi-resident proteins in plants. However, the sequences' similarity in training data is so high that the predictive performance of their proposed method might be overestimated. Recently, we have developed a tool subGolgi to discriminate between cis-Golgi and trans-Golgi proteins [12]. The overall accuracy was 74.7% which is less than adequate. Therefore, there is an urgent need to develop efficient computational tools for accurately identifying the types of Golgi-resident proteins.

In this paper, we presented a discriminative computational framework to identify the cis-Golgi and trans-Golgi proteins according to primary sequence information. The analysis of variance (ANOVA)

* Corresponding authors. Tel.: +86 28 83208232; fax: +86 28 83208238.
E-mail addresses: hding@uestc.edu.cn (H. Ding), chenwei_imu@yahoo.com.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.03.005

was proposed to optimize the feature set. As a result, the predictive accuracy of 85.4% was achieved with an area under the receiver operating characteristic curves (auROC) of 0.878 in the leave-one-out cross-validation, suggesting that the proposed method can be efficiently used to annotate the types of Golgi-resident proteins.

2. Materials and methods

2.1. Dataset

Both amino acid sequences and annotation information of Golgi-resident proteins were extracted from Universal Protein Resource (Uniprot) [13]. The following criteria were performed to guarantee the quality of benchmark data: 1) Only those proteins annotated by one subGolgi location were selected, because the number of proteins with two subGolgi locations is too small to have statistical significance. 2) Proteins with ambiguous protein existence annotations, such as "uncertain," "predicted" and "inferred from homology" were excluded because they lack confidence. 3) Only those proteins with experimental confirmed subGolgi location were included because they can provide correct and validated information. 4) The sequences which are fragments of other proteins were excluded because their information is redundant and not integrity. 5) Sequences containing nonstandard letters, such as "B," "X" or "Z," were excluded because their meanings are ambiguous.

After strictly following the above procedures, we obtained 671 proteins including 162 cis-Golgi proteins and 509 trans-Golgi proteins. It is well-known that sequence similarity in dataset correlates with estimated accuracy. High similarity data would cause two problems: one is that the data would be lack of enough representativeness due to the high similarity of the sequences in cis-Golgi and trans-Golgi proteins; the other is that the results might be misleading because of using the biased data to train the proposed method. Therefore, it is necessary to get rid of redundancy and homology bias. In this study, we used the PISCES program [14] to remove the highly similar sequences. The following is a guide on how to use the program PISCES to remove the similar sequences. At first, open the web server (http://dunbrack.fccc.edu/Guoli/PISCES_InputD.php) and paste the protein sequences with FASTA format into the textbox or upload file with FASTA format. Secondly, set the cutoff of sequence identity (Maximum percentage identity), minimum chain length and maximum chain length. Thirdly, press the submit button at the bottom and the user need to fill user name, email address and institution in a new windows. Finally, click on the Submit button. Then the links of results will be sent to the provided email address. The present study sets 25% as the cutoff of sequence identity with the minimum chain length of 25aa and the maximum chain length of 10,000aa, and about 80% of proteins in raw data have been removed. As a result, a total of 42 cis-Golgi and 95 trans-Golgi proteins were obtained.

It is important to use the independent dataset to evaluate the performance of the method. Here, we collected 13 cis-Golgi and 51 trans-Golgi proteins from Uniprot, which is independent from training set. All training data and independent data can be found from http://lin.uestc.edu.cn/server/SubGolgi/data.

2.2. Features

It is one of the most important parts for pattern recognition to generate a set of informative parameters. In recent decades, various parameters such as PseAAC [15–17], physicochemical properties of amino acids [18–20] and GO information [21–23] have been successfully employed in many protein structure and function predictions. The proximate dipeptide compositions are also accepted parameters and have been widely applied in the realm of protein prediction [24,25]. Here, we extended proximate dipeptide compositions to *g*-gap dipeptide compositions [26]. A Golgi-resident protein with length of L can be characterized by a 400 dimensional feature vector and described as follows:

$$F_{400}^{g} = \left[f_1^g, f_2^g, \cdots, f_i^g, \cdots, f_{400}^g\right]^T \tag{1}$$

here symbol **T** denotes the transposition of the vector. f_i^g is the frequency of the i-th g-gap dipeptide and expressed as:

$$f_i^g = n_i^g / \sum_{i=1}^{400} n_i^g = n_i^g / (L-g-1) \tag{2}$$

here n_i^g denotes the number of the i-th g-gap dipeptide. 0-gap dipeptides denote proximate dipeptides.

2.3. Feature selection technique

The original feature set generally contains redundant information or noise which will reduce the predictive accuracy. Thus, it is necessary to pick out informative parameters. Some techniques such as principal component analysis (PCA) [27] and minimal-redundancy-maximal-relevance (mRMR) [19] have been presented for feature selection. In this study, we proposed the ANOVA to perform feature selection. The ANOVA method can rank the features by measuring the ratio between their variances between groups and within groups [24]. The ratio reveals how strong the λ-th feature is related to the group variables. The ratio F value ($F(\lambda)$) of λ-th g-gap dipeptide in two benchmark datasets is defined as the following equation:

$$F(\lambda) = \frac{s_B^2(\lambda)}{s_W^2(\lambda)} \tag{3}$$

here $s_B^2(\lambda)$ and $s_W^2(\lambda)$ are the sample variance between groups (also called Means Square Between, MSB) and sample variance within groups (also called Mean Square Within, MSW) and can be given by:

$$s_B^2(\lambda) = \sum_{i=1}^{K} n_i \left(\frac{\sum_{j=1}^{n_i} f_{ij}(\lambda)}{n_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} f_{ij}(\lambda)}{\sum_{i=1}^{K} n_i} \right)^2 / df_B \tag{4}$$

$$s_W^2(\lambda) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(f_{ij}(\lambda) - \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} f_{ij}(\lambda)}{\sum_{i=1}^{K} n_i} \right)^2 / df_W \tag{5}$$

here $df_B = K - 1$ and $df_W = N - K$ are degrees of freedom for MSB and MSW, respectively. K and N represent the number of groups and total number of samples, respectively. $f_{ij}(\lambda)$ denotes the frequency of the λ-th feature of the j-th sample in the i-th group. n_i denotes the number of sample in the i-th group.

The $F(\lambda)$ reveals how strong the λ-th feature is related to the group variables. If there are m g-gap dipeptides whose $F(\lambda)$ is larger than a given cutoff F_{cutoff}, the frequencies of these g-gap dipeptides are selected as optimized feature set and expressed as:

$$F_m = [f_1, f_2, \cdots, f_i, \cdots, f_m]^T \tag{6}$$

If F_{cutoff} is set to zero, 400 g-gap dipeptides are all selected. The larger the F_{cutoff} is, the less the number of features is. By setting an appropriate F_{cutoff}, high-dimensional data can be projected into a low-dimensional space and the best accuracy can be achieved. The parameter m or F_{cutoff} was chosen by using cross-validation.

2.4. Support vector machine

Support vector machine (SVM) is a powerful machine learning method and has been successfully applied in protein structure and function prediction [28–31]. The SVM can find a decision boundary that separates two training data. The decision boundary is a hyperplane which maximizes the margin between the two sets in the

feature vector space. In this study, our implementation utilized the software LibSVM [32]. The RBF is used to perform the prediction. The grid search program was applied to tune the regularization parameter C and kernel width parameter γ by using cross-validation.

2.5. Performance evaluation

As previously described [24,25], three standard measurements: sensitivity (Sn), overall accuracy (OA) and Matthews correlation coefficient (MCC) were used to evaluate the performance of the proposed method, and can be defined by the following formulas:

$$\mathrm{Sn} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{7}$$

$$\mathrm{OA} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}) \tag{8}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})}} \tag{9}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively.

To estimate the robustness of the prediction system, the leave-one-out cross-validation was carried out. In the leave-one-out cross-validation, each sequence in the training dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified. As a unique outcome can be yielded, leave-one-out cross-validation was employed to evaluate the performance of all models. To describe the performance of models across the entire range of SVM decision values, receiver operating characteristic (ROC) curves were performed.

3. Results and discussion

3.1. Prediction accuracies for Golgi protein types

The g-gap dipeptides reflect correlations between two amino acids with a distance of g amino acids. Let us use as an example the 2-gap dipeptides, whose frequencies can be achieved by Eq. (2). We rank the importance of the 400 2-gap dipeptides according to their F values as defined by Eq. (3). A larger F value of a 2-gap dipeptide suggests that it is more important for the prediction of the types of Golgi proteins. Furthermore, by adding the ranked 2-gap dipeptides one by one, we can build 400 individual predictors for the 400 sub-feature sets. If large F values are used, the selected features will give more reliable information for classification. However, the number of the selected features is too small to afford enough information, which deduces the poor predictive accuracy. For example, when setting F values larger than 7, the optimized 2-gap dipeptides set contains 10 features. The overall accuracy is only 77.4% in leave-one-out cross-validation. On the contrary, if the F values are set small, the number of the selected features is so large that the cluster-tolerant capacity is reduced which also deduces a bad prediction in cross-validations. An example is that 400 2-gap dipeptides can only produce the overall accuracy of 70.8% in leave-one-out cross-validation (Table 1). Therefore, using appropriate 2-gap dipeptides can yield a prediction with higher accuracy.

By testing the predictive performances of each of 400 predictors, we plotted a 3-dimension graph in Fig. 1 for F value, feature dimension and overall accuracy. As shown in Fig. 1, the overall accuracy reaches its maximum when 83 features ($F_{\mathrm{cutoff}} = 2.17$) are used. Such 83 features were regarded as the optimal sub-feature set of our classifier. Based on these features, we drew the ROC curves of leave-one-out cross-validation in Fig. 2 to investigate the change of the sensitivity of cis-Golgi proteins *vs.* false cis-Golgi proteins rate (trans-Golgi proteins are predicted as cis-Golgi proteins) by varying decision values of SVM. It shows that the area under ROC curves (auROC) achieves 0.878 with the MCC of 0.652 (Table 1).

Table 1
The comparison of proposed method with other methods.

	Sn (%)		OA(%)	auROC	MCC
	cis-Golgi	trans-Golgi			
SVM (83 dipeptides)	73.8	90.5	85.4	0.878	0.652
SVM (400 dipeptides)	21.4	92.6	70.8	0.569	0.202
SVM (PseAAC)	47.6	87.4	75.2	0.659	0.381
SVM (AAC)	50.0	85.3	74.5	0.683	0.373
PLS(83 dipeptides)	71.4	91.6	85.4	0.851	0.649
Naïve Bayes (63 dipeptides)	50.0	92.6	79.6	0.814	0.487
RBF Network (83 dipeptides)	54.8	89.5	78.8	0.763	0.477

It is necessary to investigate whether other g-gap sub-feature sets can obtain higher accuracies or not. We varied the g from 0 to 6 and repeated the feature selection process for finding the maximum accuracy of each g-gap dipeptides. Results in Fig. 3 show that the feature set of optimized 2-gap dipeptides is the best one among the 7 optimized feature sets.

For the display of relationship between sequence identity and predicted accuracy, we further evaluated the predictive accuracies of proposed model on the datasets with different sequence identity using leave-one-out cross-validation. Results in Fig. 4 show that, with the sequence identity varying from 100% to 25%, the overall accuracies decrease and stabilize at about 85%, demonstrating that our proposed method is robust.

3.2. Comparison accuracies

It is natural to ask whether the proposed method has a better performance than other methods or not. We carried out two comparisons: one was to compare the predictive capability of the optimized 2-gap dipeptides with that of PseAAC and amino acid composition by using SVM algorithm; the other was to compare the predictive capability of SVM with that of Naïve Bayes, RBF Network and partial least squares regression (PLS regression) by using the optimized 2-gap dipeptides.

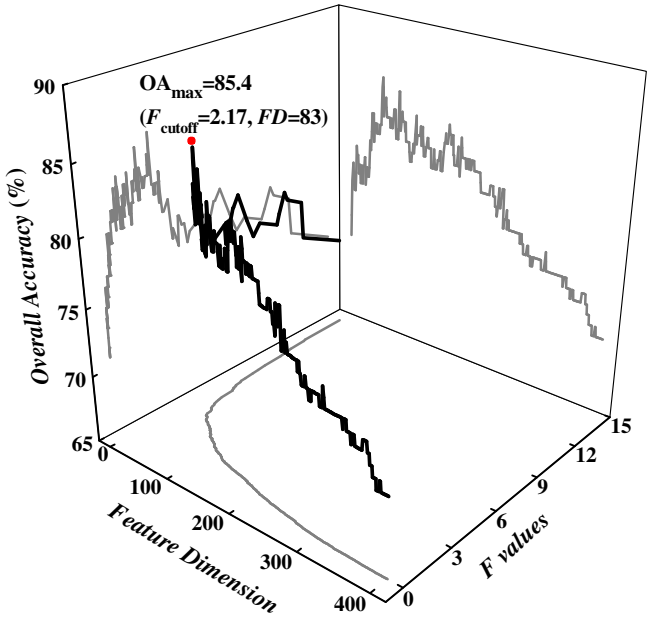

Fig. 1. The graph for predicting the types of Golgi-resident proteins. Dark line denotes 3-D curve. Three gray lines are projections on three planes (overall accuracy/feature dimension plane, overall accuracy/F value, F value/feature dimension plane).

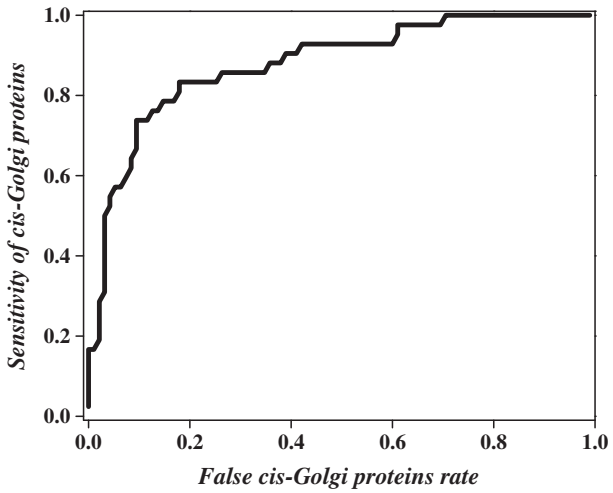

Fig. 2. The ROC curve of leave-one-out cross-validation of the proposed method.

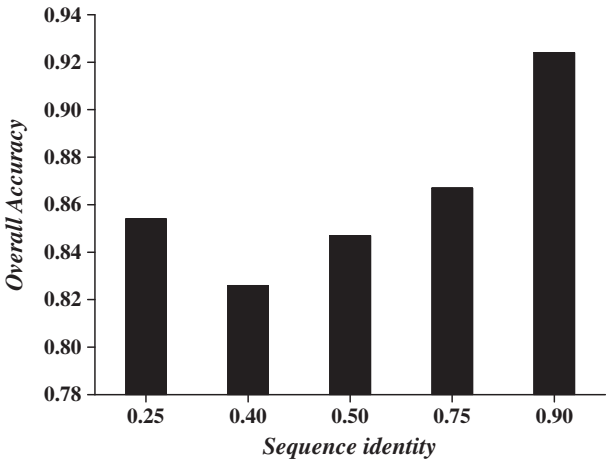

Fig. 4. The predictive accuracies for the datasets with different sequence identities.

For the PseAAC, we selected hydrophobicity, hydrophilicity and mass as amino acid characters. The correlation parameter lambda varied from 0 to 7 with the step of 1. The weight factor varied from 0.05 to 0.7 with the step of 0.05. By examining 98 sets of PseAAC parameters, the highest accuracy was obtained when $\lambda = 1$ with $w = 0.05$ and recorded in Table 1. It shows that optimal dipeptides can achieve the highest accuracy among other parameters (PseAAC, AAC, 400 dipeptides). Furthermore, we compare the performance of SVM with that of Naïve Bayes, RBF Network and PLS regression by using optimal dipeptides. The results were also listed in Table 1. It shows that the accuracy of PLS regression is higher than that of Naïve Bayes and RBF Network. Although the OA of PLS regression is as high as the accuracy obtained by SVM, the auROC and MCC of SVM are better. Thus we recommend using the SVM with optimal dipeptides for prediction.

For further evaluating the proposed method, we examined the performance of our method on the independent dataset. Results show that 69.2% cis-Golgi and 90.2% trans-Golgi proteins can be correctly predicted. In our recent work [12], a total of 95 Golgi-resident proteins with the sequence identity less than 40% were used to train and test the modified Mahalanobis Discriminant. The overall accuracy of the leave-one-out cross-validation was only 74.7% with the auROC of 0.772, which is lower than the results of the current study. Based on above comparisons and analysis, we concluded that the proposed method is a powerful tool for the annotation of the types of Golgi-resident proteins.

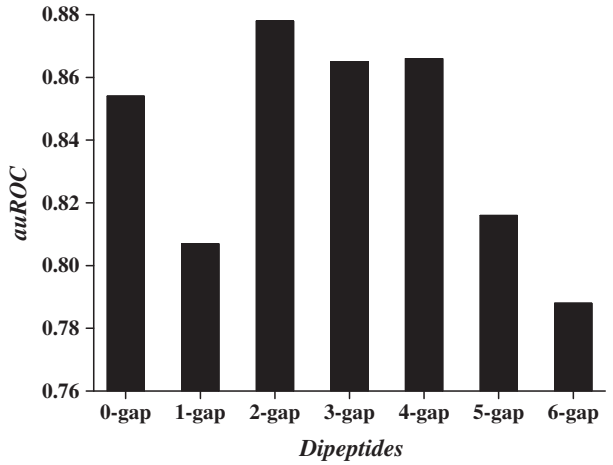

Fig. 3. The maximum auROCs of 7 optimized dipeptide sets by ANOVA.

3.3. Prediction of Golgi proteins

Generally speaking, when one obtains a protein sequence, it must judge whether it is a Golgi-resident protein before predicting the Golgi protein type of this protein. Thus, a non-Golgi protein dataset must be constructed. However, after filtering the uncertainty proteins in Uniprot according to the step in dataset section, we still achieved a large numbers of non-Golgi proteins (over 20,000 items). This can result in extremely imbalance of the size between positive data and negative data, which further cause learning bias in class identification techniques. Therefore, for balancing the size of two classes of data, we randomly selected 137 non-Golgi proteins who share the sequence identities less than 25%. We then repeated the feature selection procession for finding the best feature set which can achieve the highest accuracy. As a result, the maximum accuracy of 79.9% with MCC of 0.599 were achieved when 66 2-gap dipeptides ($F_{\text{cutoff}} = 5.2$) are used. 77.4% Golgi proteins and 82.5% non-Golgi proteins can be correctly predicted. These results suggest that our method is efficient.

4. Conclusion

In this study, we developed a powerful method for the prediction of types of Golgi-resident proteins by using ANOVA to filter *g*-gap dipeptide. Feature selection has widely used in pattern identification; however, the ANOVA was rarely used in protein bioinformatics, especially in Golgi-resident protein prediction. By using a series of experiments, we demonstrated the power of the method. Based on this method, we built an on-line server, namely, subGolgi v2.0 which can be freely available from http://lin.uestc.edu.cn/server/subGolgi2. We hope that the server will be useful to discriminate between cis-Golgi and trans-Golgi proteins in the absence of experimental data and elucidate the biological function of newly discovered Golgi-resident proteins.

Conflict of interest

The author has no conflict of interests concerning this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61202256, 61100092) and the Fundamental Research Funds for the Central Universities (ZYGX2012J113).

References

[1] S.W. Rhee, T. Starr, K. Forsten-Williams, et al., The steady state distribution of glycosyltransferases between the Golgi apparatus and the endoplasmic reticulum is approximately 90:10, Traffic 6 (2005) 978–990.
[2] S.R. Pfeffer, Constructing a Golgi complex, The Journal of Cell Biology 155 (2001) 873–875.
[3] Y. Fujita, E. Ohama, M. Takatama, et al., Fragmentation of Golgi apparatus of nigral neurons with alphasynuclein-positive inclusions in patients with Parkinson's disease, Acta Neuropathologica 112 (2006) 261–265.
[4] N.K. Gonatas, J.O. Gonatas, A. Stieber, The involvement of the Golgi apparatus in the pathogenesis of amyotrophic lateral sclerosis, Alzheimer's disease, and ricin intoxication, Histochemistry and Cell Biology 109 (1998) 591–600.
[5] Z. Hu, L. Zeng, L. Xie, et al., Morphological alteration of Golgi apparatus and subcellular compartmentalization of TGF-beta1 in Golgi apparatus in gerbils following transient forebrain ischemia, Neurochemical Research 32 (2007) 1927–1931.
[6] W.C. Chou, Y. Yin, Y. Xu, GolgiP: prediction of Golgi-resident proteins in plants, Bioinformatics 26 (2010) 2464–2465.
[7] L. Hu, T. Huang, Y.D. Cai, et al., Prediction of body fluids where proteins are secreted into based on protein interaction network, PloS One 6 (2011) e22989.
[8] K.C. Chou, H.B. Shen, Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms, Nature Protocols 3 (2008) 153–162.
[9] K.C. Chou, H.B. Shen, A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0, PloS One 5 (2010) e9931.
[10] Z. Yuan, R.D. Teasdale, Prediction of Golgi Type II membrane proteins based on their transmembrane domains, Bioinformatics 18 (2002) 1109–1115.
[11] R. Schwacke, A. Schneider, E. van der Graaff, et al., ARAMEMNON, a novel database for Arabidopsis integral membrane proteins, Plant Physiology 131 (2003) 16–26.
[12] H. Ding, L. Liu, F.B. Guo, et al., Identify Golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition, Protein and Peptide Letters 18 (2011) 58–63.
[13] E.C. Dimmer, R.P. Huntley, Y. Alam-Faruque, et al., The UniProt-GO Annotation database in 2011, Nucleic Acids Research 40 (2012) D565–D570.
[14] G. Wang, R.L. Dunbrack Jr., PISCES: recent improvements to a PDB sequence culling server, Nucleic Acids Research 33 (2005) W94–W98.
[15] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, Journal of Theoretical Biology 273 (2011) 236–247.
[16] H.B. Shen, K.C. Chou, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition, Analytical Biochemistry 373 (2008) 386–388.
[17] L. Nanni, S. Brahnam, A. Lumini, High performance set of PseAAC and sequence based descriptors for protein classification, Journal of Theoretical Biology 266 (2010) 1–10.
[18] Y.X. Fan, J. Song, X. Kong, et al., PredCSF: an integrated feature-based approach for predicting conotoxin superfamily, Protein and Peptide Letters 18 (2011) 261–267.
[19] B.Q. Li, L.L. Hu, S. Niu, et al., Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches, Journal of Proteomics 75 (2012) 1654–1665.
[20] I.E. Kaya, T. Ibrikci, O.K. Ersoy, Prediction of disorder with new computational tool: BVDEA, Expert Systems with Applications 38 (2011) 14451–14459.
[21] W. Chen, H. Lin, Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information, Biochemical and Biophysical Research Communications 401 (2010) 382–384.
[22] W.L. Huang, C.W. Tung, S.W. Ho, et al., ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization, BMC Bioinformatics 9 (2008) 80.
[23] L. Li, Y. Zhang, L. Zou, et al., An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino Acid hydrophobicity, PloS One 7 (2012) e31057.
[24] H. Lin, H. Ding, Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition, Journal of Theoretical Biology 269 (2011) 64–69.
[25] W. Chen, H. Lin, Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine, Computers in Biology and Medicine 42 (2012) 504–507.
[26] H. Lin, The modified Mahalanobis Discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition, Journal of Theoretical Biology 252 (2008) 350–356.
[27] J. Tian, H. Gu, W. Liu, et al., Robust prediction of protein subcellular localization combining PCA and WSVMs, Computers in Biology and Medicine 41 (2011) 648–652.
[28] Y.S. Ding, T.L. Zhang, K.C. Chou, Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network, Protein and Peptide Letters 14 (2007) 811–815.
[29] W. Zhou, H. Yan, Prediction of DNA-binding protein based on statistical and geometric features and support vector machines, Proteome Science 9 (2011) S1.
[30] R.P. Liang, S.Y. Huang, S.P. Shi, et al., A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization, Computers in Biology and Medicine 42 (2012) 180–187.
[31] J. Luo, L. Yu, Y. Guo, M. Li, Functional classification of secreted proteins by position specific scoring matrix and auto covariance, Chemometrics and Intelligent Laboratory Systems 110 (2012) 163–167.
[32] R.E. Fan, P.H. Chen, C.J. Lin, Working set selection using the second order information for training SVM, Journal of Machine Learning Research 6 (2005) 1889–1918.