Bioinformatics Advance Access published August 23, 2010

GolgiP: prediction of Golgi resident proteins in plants

Wen-Chi Chou[1,2], Yanbin Yin[1,2], Ying Xu[1,2, 3]

[1]Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, USA

[2]BioEnergy Science Center (http://www.bioenergycenter.org), USA

[3]College of Computer Science and Technology, Jilin University, Changchun, Jilin, China

Associate Editor: Prof. Burkhard Rost

ABSTRACT

Summary: We present a novel Golgi-prediction server, ***GolgiP***, for computational prediction of both membrane-associated and non-membrane-associated Golgi resident proteins in plants. We have employed a support vector machine-based classification method for prediction of such Golgi proteins, based on three types of information, dipeptide composition, transmembrane domain(s), and functional domain(s) of a protein, where the functional domain information is generated through searching against the Conserved Domains Database (CDD), and the transmembrane domain (TMD) information includes the number of TMDs, the length of TMD, and the number of TMDs at the N-terminus of a protein. Using GolgiP, we have made genome-scale predictions of Golgi resident proteins in 18 plant genomes, and have made the preliminary analysis of the predicted data.

Availability: The GolgiP web service is publically available at http://csbl1.bmb.uga.edu/GolgiP/

Contact: xyn@csbl.bmb.uga.edu

Supplementary Information: Supplementary data are available at the database website and ***Bioinformatics*** online.

1 INTRODUCTION

The Golgi apparatus is an essential cellular organelle found in most, if not all, eukaryotic cells, serving as an intermediate station of the secretory pathway that transports proteins out of a cell. Besides, Golgi is also a major site for protein post-translational modifications (e.g. glycosylation (Nilsson, et al., 2009)) and synthesis of various polysaccharides. The plant cell walls are mainly comprised of lignins, glycosylated proteins, and polysaccharides, most of which are synthesized in Golgi (Lerouxel, et al., 2006).

Identification of the Golgi resident proteins represents a very challenging and a highly important problem for the understanding of the biological processes taking place in Golgi. While there are 1183 mouse and human Golgi resident proteins identified (Sprenger, et al., 2007), only a little over 400 plant Golgi proteins have been experimentally identified. A key challenging issue is that plant Golgi proteins do not seem to have obvious targeting signals as proteins targeted at other cellular compartments, like nucleus or extra-cellular space. Most of the existing computational tools for subcellular localization predictions are designed for the general subcellular localization prediction, and their predictions for Golgi resident proteins are less than adequate (Sprenger, et al., 2006). Only one program has been specifically designed for prediction of Golgi localized proteins but it focuses only on trans-membrane Golgi proteins (Yuan and Teasdale, 2002). The issue is that only 25% of Golgi proteins of *Arabidopsis thaliana* are estimated to contain transmembrane regions (Schwacke, et al., 2003), indicating the inadequacy of the current programs. Based on this consideration, we have designed a support vector machine (SVM) based classifier, called GolgiP, to predict both Golgi localized transmembrane proteins and non-transmembrane proteins. GolgiP currently provides multiple models for predicting plant Golgi proteins, based on the specific needs of a user.

2 METHODS AND DATASET

We have collected a large dataset comprising of 402 known Golgi proteins and 5,703 known non-Golgi proteins of *Arabidopsis thaliana* (91.2%), *Oryza sativa* (8.2%), and other plants (0.7%), from the SUBA (Heazlewood, et al., 2007) and the UniProt (Apweiler, et al., 2004) databases, as well as manually curated from the published literature. The non-Golgi proteins are proteins that have subcellular localization annotations, but not in Golgi according to the above databases. The redundant sequences in our dataset are removed by CD-hit using 95% sequence identity as the cut-off (Li and Godzik, 2006). Four fifth of the data was used to train the classifier and the remaining one fifth was used to test the trained classifier, where the dataset was randomly partitioned into training and test data sets.

To train an SVM-based classifier for Golgi proteins, we have examined three different sets of features, all computed from protein sequences. The first set of features is related to the dipeptide composition (DiAA). For each protein in our training set, we calculated the composition of dipeptides. The second set of features is related to transmembrane domains (TMDs). We used TMHMM (Krogh, Larsson et al. 2001) and Phobius (Kall, et al., 2004) to predict the number of TMDs, the average length of TMDs, the number of TMDs within the N terminal region consisting of 70 amino acids, the length of the first TMD within the N terminal region, and the orientation of the N-terminal (i.e. in the cytosol side or in the Golgi lumen side). The third set of features is related to functional domains (FunD). We searched proteins in our datasets against the CDD database using RPS-BLAST (Marchler-Bauer, et al., 2009) with an e-value cutoff < 0.01. We did this because the Golgi apparatus is where proteins get post-translational modifications such as glycosylation (Nilsson, et al., 2009), and where the syntheses of most polysaccharides take place (Nilsson, et al., 2009). In

[*]To whom correspondence should be addressed.

addition, Komatsu et al. found that the distributions of functional categories of proteins are different in different membranes such as plasma membrane, vascular membrane, and Golgi membrane, respectively (Komatsu, et al., 2007). Hence it is expected that enzymes for the Golgi-related activities should be located in Golgi. The CDDs found for the Golgi proteins are then collected as the third set of features.

We applied the LIBSVM package (Fan, et al., 2005) to train the classifier. We used the Radial Basis Function kernel, and tuned the cost (c) and gamma (γ) parameters to optimize the classification performance on the training data set.

3 RESULTS AND DISCUSSION

We used the aforementioned three sets of sequence features, and trained three SVM classification models. Besides, we combined all three sets of features to train a comprehensive model. The training performances are shown in the Supplementary Material.

We have compared the models with the other Golgi protein prediction tools, including PSORT (Nakai and Horton, 1999), WoLF PSORT (Horton, et al., 2007), and Yuan's Golgi predictor (Yuan and Teasdale, 2002) by using the testing data set.

Table 1: Evaluation of Golgi protein prediction tools

Tools	Sensitivity	Specificity	Accuracy	MCC
Yuan	71.64%	23.18%	26.37%	-0.03
WoLF PSROT	15.92%	92.69%	87.63%	0.08
GolgiP-TMD	61.73%	67.75%	67.75%	0.15
PSORT	43.53%	83.10%	80.49%	0.17
GolgiP-DiAA	71.64%	80.76%	80.16%	0.31
GolgiP-Comprehensive	72.84%	98.42%	96.73%	0.73
GolgiP-FunD	57.50%	100.00%	97.21%	0.75

The performances are sorted by Matthews correlation coefficient (MCC).

As shown in Table 1, Yuan's Golgi predictor has the good sensitivity but the lowest specificity and the lowest accuracy. PSORT and WoLF PSORT are two general subcellular localization prediction tools, and have moderate level of classification performance, which may not be adequate to serve as a plant Golgi protein predictor based on our analysis. Our program, GolgiP, exhibits the better overall performances with a higher accuracy and MCC.

Table 2: Application of the GolgiP program on 18 plant genomes

Clade	Species	# predicted Golgi proteins/ # Total proteins	%
Red algae	*Cyanidioschyzon merolae 10D*	430/5014	8.58%
green algae	*Micromonas pusilla CCMP1545*	716/10475	6.84%
green algae	*Micromonas strain RCC299*	833/9815	8.49%
green algae	*Ostreococcus lucimarinus*	706/7651	9.23%
green algae	*Ostreococcus tauri*	656/7725	8.49%
green algae	*Chlamydomonas reinhardtii*	982/14598	6.73%
green algae	*Volvox carteri f. nagariensis*	1025/15544	6.59%
moss	*Physcomitrella patens ssp patens*	2344/35938	6.52%
spike moss	*Selaginella moellendorffii*	2912/34697	8.39%
monocot	*Oryza sativa*	4240/67393	6.29%
monocot	*Brachypodium distachyon*	2446/32255	7.58%
monocot	*Sorghum bicolor*	2197/35899	6.12%
monocot	*Zea mays*	4748/75387	6.30%
dicot	*Vitis vinifera*	2008/30434	6.60%
dicot	*Arabidopsis thaliana*	2727/33410	8.16%
dicot	*Medicago truncatula*	1856/30028	6.18%
dicot	*Glycine max*	5262/75778	6.94%

We have applied GolgiP with the functional domain model to predict Golgi proteins on 18 selected fully sequenced plant genomes using the same cutoff. The reason we chose the functional domain model is that the model performs the best specificity, and therefore tends to avoid false positive results in this genome-wide prediction. The numbers and percentages of the predicted Golgi proteins in these organisms are shown in Table 2. Across algae, moss, monocot and dicot plants, the average percentages of predicted Golgi proteins is 7.25% among all the encoded proteins by these genomes. The stability in the percentage of the predicted Golgi proteins across different genomes indirectly suggests the reliability of our predictions. The trend of distribution of Golgi proteins from lower to higher plant species shows the similar percentage of Golgi proteins. This may suggest that the functionality of the Golgi apparatus has evolved and matured fairly early in the plant evolution.

In conclusion, we developed a Golgi protein prediction tool, GolgiP, and demonstrated its superior performance in predicting plant Golgi proteins over existing prediction servers. In addition, our predictions across multiple plant genomes give an estimation of the percentage of plant Golgi proteins across different plant organisms, which is in general agreement with the previous estimations.

ACKNOWLEDGEMENTS

We would like to thank Dr. Maor Bar-Peled for the helpful discussions.

Funding: This work is supported in part by the BioEnergy Science Center (BESC) grant from the Office of Biological and Environmental Research in the DOE Office of Science and National Science Foundation (DOE 4000063512).

Conflict of Interest: none declared.

REFERENCES

Apweiler, R., *et al.* (2004) UniProt: the Universal Protein knowledgebase, *Nucleic Acids Res*, **32**, D115-119.

Fan, R.-E., Chen, P.-H. and Lin, C.-J. (2005) Working set selection using second order information for training SVM, *Journal of Machine Learning Research*, **6**.

Heazlewood, J.L., *et al.* (2007) SUBA: the Arabidopsis Subcellular Database, *Nucleic Acids Res*, **35**, D213-218.

Horton, P., *et al.* (2007) WoLF PSORT: protein localization predictor, *Nucleic Acids Res*, **35**, W585-587.

Kall, L., Krogh, A. and Sonnhammer, E.L. (2004) A combined transmembrane topology and signal peptide prediction method, *J Mol Biol*, **338**, 1027-1036.

Komatsu, S., Konishi, H. and Hashimoto, M. (2007) The proteomics of plant cell membranes, *J Exp Bot*, **58**, 103-112.

Lerouxel, O., *et al.* (2006) Biosynthesis of plant cell wall polysaccharides - a complex process, *Curr Opin Plant Biol*, **9**, 621-630.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics (Oxford, England)*, **22**, 1658-1659.

Marchler-Bauer, A., *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database, *Nucleic Acids Res*, **37**, D205-210.

Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends Biochem Sci*, **24**, 34-36.

Nilsson, T., Au, C.E. and Bergeron, J.J. (2009) Sorting out glycosylation enzymes in the Golgi apparatus, *FEBS Lett*, **583**, 3764-3769.

Schwacke, R., *et al.* (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins, *Plant Physiol*, **131**, 16-26.

Sprenger, J., *et al.* (2007) LOCATE: a mammalian protein subcellular localization database, *Nucleic Acids Research*.

Sprenger, J., Fink, J.L. and Teasdale, R.D. (2006) Evaluation and comparison of mammalian subcellular localization prediction methods, *BMC Bioinformatics*, **7 Suppl 5**, S3.

Yuan, Z. and Teasdale, R.D. (2002) Prediction of Golgi Type II membrane proteins based on their transmembrane domains, *Bioinformatics (Oxford, England)*, **18**, 1109-1115.