Subchloroplast Location Prediction via Homolog Knowledge Transfer and Feature Selection

Xiaomei Li[1], Xindong Wu[1, 2], Gongqing Wu[1], Xuegang Hu[1]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, P.R. China

[2]Department of Computer Science, University of Vermont, Burlington, VT, USA

Abstract

The accuracy of subchloroplast location prediction algorithms often depends on predictive and succinct features derived from proteins. Thus, to improve the prediction accuracy, this paper proposes a novel SubChloroplast location prediction method, called SCHOTS, which integrates the HOmolog knowledge Transfer and feature Selection methods. SCHOTS contains two stages. First, discriminating features are generated by WS-LCHI, a Weighted Gene Ontology (GO) transfer model based on bit-Score of proteins and Logarithmic transformation of CHI-square. Second, the more informative GO terms are selected from the features. Extensive studies conducted on three real datasets demonstrate that SCHOTS outperforms three off-the-shelf subchloroplast prediction methods.

Introduction

Surrounded by two layers of membrane, chloroplasts can be divided into four main compartments: stroma, thylakoid lumen (ThyLum), thylakoid membrane (ThyMem) and envelope. Determining the subchloroplast locations of proteins is a vital step toward understanding the molecular mechanism that underlies the functions of cells. The performance of protein subcellular location (PSL) prediction often depends on protein features. Recently, there are four kinds of methods commonly used for the protein feature generation, such as sequence-based methods, evolutionary-information-based methods, annotation-based methods and hybrid methods. In this paper, we focus on the annotation-based approach, which extracts protein features from domain, protein networks or GO annotation information.

SCHOTS derives predictive and succinct features from amino acid sequences and GO annotation information based on the following considerations. From the viewpoints of evolutionary biology, a protein tends to have the same location with its homolog. The performance of several existing GO transfer models may degrade by incorporating evolutionarily divergent homolog. Motivated by this, we propose a novel weighted GO transfer model to reduce the impact from potentially noisy annotations in homolog transfer. From the viewpoints of computation, different GO terms may have different discriminative capability for PSL prediction. We hence use text-weighting methods to assign weights for relative GO terms (Chi and Nam 2012). Meanwhile, considering the redundancy and irrelevance of GO terms, we use the Fisher score (Duda, Hart, and Stork 2001) to filter redundant GO terms. From the above stages, we get a higher accuracy model SCHOTS of subchloroplast location prediction. Extensive experiments show that SCHOTS outperforms the state-of-the-art protein subchloroplast location prediction methods, including SubChlo (Du and Li 2009), subIdent (Shi et al. 2011) and BS-KNN (Hu and Yan 2012).

Our Method

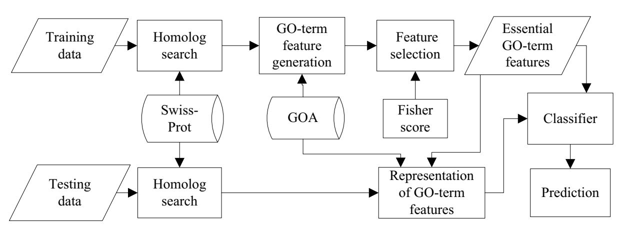

Figure 1. Processing flow of SCHOTS

Figure 1 shows the processing flow of SCHOTS. It consists of the following five components. (1) Homolog search: we find all homologous proteins of the target proteins from the UniProtKB/Swiss-Prot database (release 2012-01) using BLAST (Altschul et al. 1997) with a given E-value (0.001). (2) GO-term feature generation: we first get the GO annotations of the target protein and its homologs from the GOA database (Version 103). We

The research has been supported by the National 863 Program of China under grant 2012AA011005, the National 973 Program of China under grant 2013CB329604, the National Natural Science Foundation of China (NSFC) under grants 61229301, 61273297, 61273292.

transfer the GO terms of the protein regarding a high bit-score and a high identity (no less than 80%) calculated by BLAST, and then we assign the bit-score and the *LCHI* coefficient as weights to GO features. The bit-score depends on sizes of the query sequence and the database, so the coordinates of a protein feature maybe larger than 1. Thus, the protein features are subject to a standard conversion. (3) Feature selection: we use the Fisher score on training datasets to determine a set of features relevant to the predictor. The method chooses the top 50 GO-term features with the highest F-scores computed by Fisher score as the essential GO-term features. (4) Classifier: we select support vector machines SVM with default settings (γ=0, C=1 and RBF kernel) in Weka (Witten and Frank 2005) as the base classifier of our method. (5) Prediction: we use the Jackknife test and single independent dataset examination to evaluate the prediction ability.

Experiments

In this section, we compare SCHOTS with three baseline methods of SubChlo, SubIdent and BS_KNN on five evaluation measures, such as sensitivity (*SE*), specificity (*SP*), the Matthews' correlation coefficient (*MCC*), average sensitivity (*AVG*) and overall accuracy (*ACC*).

Table 1 shows the performance of SCHOTS and three baseline methods [1] for the subchloroplast location prediction on the S60 dataset. From the experimental results, the overall accuracy of SCHOTS is up to 98.47%, which is improved by 31.29%, 9.16% and 22.57% in comparison with SubChlo, SubIdent and BS_KNN respectively.

Table 1. Performance comparison on the S60 dataset

Location	SubChlo	SubIdent	BS_KNN	SCHOTS
	***SE*(%)**	***SE*(%)**	***SE*(%)**	***SE*(%)**
Envelope	40.0	85.7	47.5	95
ThyLum	43.2	64.4	77.5	100
ThyMem	83.7	98.2	85.0	100
Stroma	67.3	80.0	73.9	98.4
***AVG*/*ACC*(%)**	58.5/67.2	82.1/89.3	70.9/75.9	98.4/98.5

Table 2. Performance comparison on the independent test dataset

Location	SubIdent	SCHOTS		
	***SE*(%)**	***SE*(%)**	***SP*(%)**	***MCC***
Envelope	76.2	95.2	80.0	0.82
ThyLum	66.7	100	100	1
ThyMem	96.9	100	92.3	0.95
Stroma	83.3	84.4	100	0.87
***AVG*/*ACC*(%)**	80.8/84.4	94.9/92.2		

We use the S60 dataset as the training dataset. None of the proteins in the independent test dataset is included in the S60 dataset. Table 2 shows the predictive ability for new proteins using SCHOTS and the baseline method SubIdent. We do not compare SCHOTS with SubChlo and BS_KNN, because they have no reports on this dataset. From this table, we can see that the prediction accuracy of SCHOTS is 7.79% which is higher than that of SubIdent. This is because WS-LCHI generates effective protein features from GO annotation information.

Conclusion

We proposed a novel subchloroplast location prediction method (SCHOTS) based on the weighted GO transfer model WS-LCHI and Fisher score in this paper. Experimental studies have shown the effectiveness of SCHOTS. However, in our study, we assumed that the sequence identity was no less than a cutoff threshold (e.g., 80%) as homolog. Similarity (identity) is not equal to homology. Thus, in our future work, we will (1) improve our weighted GO transfer model by more accurate homolog models, and (2) integrate information from different sources such as motifs and PSSM into protein features.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25(17), 3389-3402.

Chi, S. M., and Nam, D. G. 2012. WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms. *Bioinformatics* 28(7), 1028-1030.

Du, P. F., Cao, S. J., and Li, Y. D. 2009. SubChlo: Predicting Protein Subchloroplast Locations with Pseudo-amino Acid Composition and the Evidence-theoretic K-nearest Neighbor (ET-KNN) Algorithm. *J. Theor. Biol.* 261(2), 330-335.

Hu, J., and Yan, X. H. 2012. BS-KNN: An Effective Algorithm for Predicting Protein Subchloroplast Localization. *Evol. Bioinform.* 8, 79-87.

Duda, R. O., Hart, P. E., and Stork, D. G. 2001. *Pattern Classication*. Wiley-Interscience , New York.

Shi, S. P., Qiu, J. D., Sun, X. Y., Huang, J. H., Huang, S. Y., Suo, S. B., Liang, R. P., and Zhang, L. 2011. Identify Submitochondria and Subchloroplast Locations with Pseudo Amino Acid Composition: Approach from the Strategy of Discrete Wavelet Transform Feature Extraction. *BBA-Mol. Cell Res.* 1813, 424-430.

Witten, I. H., Frank, E. 2005. *Data Mining: Pracitcal Machine Learning Tools and Techniques with JAVA Implementations (Second Edition).* Morgan Kaufmann, San Francisco.

[1] We use the prediction results of SubChlo, SubIdent and BS-KNN directly from their original papers.