Prediction of subcellular localization of multi-location proteins in *Arabidopsis thaliana* using a probabilistic model integrating family-driven and protein-protein features

S.H. Razavi[a,*], S.A. Katanforoush[a]
[a] Department of Computer and Data Sciences, Shahid Beheshti University, Tehran, Iran.
hourirazavi@gmail.com

Abstract: The eukaryotic cells are organized into membrane-covered compartments that are characterized by specific sets of proteins and biochemically distinct cellular processes. Identifying the functions of proteins in various cellular organelles and pathways is one of the fundamental goals in proteomics, cell biology, and drug design research. Predicting appropriate protein subcellular localization can provide useful insights for revealing their functions and dysfunctions. The two principal obstacles of this problem are the prediction of multi-location proteins and deficiency of suitable knowledge of proper data modeling. Most of the existing methods were designed to treat the proteins as single-location. Whereas, recent experimental data report that location of the protein in a cell is actually a multi-label system, where some proteins may simultaneously occur in two or more different location regions [1, 2]. In this paper, we presented a robust classifier to predict the multiple-location of proteins. Our method is based on a graphical probabilistic model that combining two sources of information, the features derived from protein families, and protein-protein interactions. It seems that proteins in a subcellular location have interaction with each other, hence, the PPI network would be a beneficial resource for prediction of proteins locations. We benchmark our method using dataset taken from *SUBA4* [3], that is a comprehensive data center for *Arabidopsis thaliana* subcellular proteins. The dataset contains 5331 proteins in 11 subcellular locations. We obtain protein sequences and PPI from [3]. Our algorithm includes the following three steps: (1) each protein sequence was aligned to each HMM profile of the top 20 largest families of Pfam [4], that results in a vector represented similarity indices to major Pfam Protein families, (2) a probabilistic model merged this family-derived features and protein-protein interaction network that relates probability of co-location of a pair of proteins to the features, (3) via maximum likelihood, the method predicts a set of overlapping cluster which is assigned to subcellular locations for each protein. Finally, we compare our algorithm with several recent location predictors [1,5] and results proved the superiority of algorithm in comparison with others.

Keywords: Protein subcellular localization; multi-location proteins; PPI networks, profile-HMM

References

[1] X. Cheng, X. Xiao, K. C Chou, “pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC”, J. Mol. biol., 13 (2017) 1722-1727.

[2] E. Glory, RF. Murphy, “Automated subcellular location determination and high-throughput microscopy,” J. Dev. Cell., 12(2007) 7-16.

[3] C.M. Hooper, I.R. Castleden, S.K. Tanz, N. Aryamanesh, “SUBA4: the interactive data analysis center for Arabidopsis subcellular protein locations”, Nucleic Acids Res., 45 (2017) 1064-1074.

[4] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman, The Pfam protein families database: towards a more sustainable future”, Nucleic Acids Res., 44 (2016) 279–285.

[5] T. Blum, S. Briesemeister, O. Kohlbacher, “MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction”, BMC Bioinformatics, 10 (2009), 274.