Indonesian Journal of Life Sciences Vol. 01 | Number 02 | September (2019)
http://journal.i3l.ac.id/ojs/index.php/IJLS

REVIEW ARTICLE

Prediction Methods of the Protein Subcellular Localization: A Systematic reviews

Gabriella Patricia[1], Nanda Rizqia Pradana Ratnasari[1], Arli Aditya Parikesit[1*]

[1]*Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, Indonesia*

**Corresponding author. Email: aril.parikesit@i3l.ac.id*

ABSTRACT (Calibri, 16)

The prediction of protein subcellular localization (SCL) has been a long-running challenge in bioinformatics. Protein SCL is crucial for a protein to exercise its functions properly. The reliance of protein localization on signaling peptides and the information available in gene ontology (GO) databases makes it possible to use computational approaches to predict protein SCL. SCL methods can be classified as either sequence-based or annotation-based. Machine learning algorithms and classifiers are used in protein SCL prediction tools. This review presents a list of protein SCL predictors published in the last 5 years.

Keywords: *Protein subcellular localization; bioinformatics; gene ontology; machine learning algorithm*

INTRODUCTION

Cells contain organelles, subcellular structures which perform independent biochemical processes and thus requiring different proteins. Various subcellular compartments assure function operation work well and normal in the entire cell (Kaushik, Chouhan & Dwivedi, 2017). Subcellular targeting, which is a process of directing a newly synthesized protein to its target organellle, is important for proteins to be able to perform their functions. Overall amino acid composition differs among proteins that are found in different organelles, suggesting that proteins may have evolved to function optimally in certain organelles (Dönnes & Höglund, 2004). As such, proper subcellular localization (SCL) is critical for optimal performance; thus it is unsurprising that protein SCL predictions are highly relevant in subjects such as genome annotation, protein function prediction, and drug vaccine target identification (Su et al., 2007).

Proteins possess targeting sequences with distinct physiochemical properties, which is used to determine its localization within the cell. These targeting sequences display high specificity and evolution conservation, specifically the conservation of the biochemical properties rather than the primary amino acid sequence. Computational approaches to prediction of SCL rely on these features.

Most SCL methods are sequence-based—relying on amino-acid compositions, sequence

homology, or sorting signals—or annotation-based, which uses information beyond provided in the protein sequences. One of the challenges that protein SCL prediction faces is multi-localization, in which a single protein can be localized to multiple locations. A multitude of bioinformatics algorithms have been developed for protein SCL prediction, many of them utilizing machine-learning algorithms such as support vector machines (SVM), artificial neural network (ANN), deep neural network (DNN), Hidden Markov model (HMM). This literature review expounds on the recently released tools for SCL prediction and the algorithms they use.

MATERIAL AND METHODS

In this literature review, the PubMed database was used for literature search using a combination of words that included “subcellular localization”, “protein”, “algorithm”, and “prediction”. The search term is as follows: *("subcellular localization"[Title]) AND protein AND algorithm AND prediction*. The paper title has to contain the term “subcellular localization”. Only papers published in 2014 or later are included in this study.

RESULT AND DISCUSSION

Table 1. List of protein subcellular localization predictors

Predictor	Category	Reference
DeepLoc	End-to-end sequence-based	Almagro et al., 2017
Hum-mPLoc 3.0	GO-based	Zhou et al., 2016
HybridGO-Loc	Multi-label GO-based	Wan et al., 2014
mPLR-Loc	Multi-label GO-based	Wan et al., 2015
MSLVP	Sequence-based	Thakur et al., 2016
pLoc-mHum	Multi-label GO-based	Cheng et al., 2017
pLoc-mPlant	Multi-label GO-based	Cheng & Chou, 2017
PMLPR	GO-based	Mirzaei Mehrabad et al., 2018

The above table displays the protein SCL tools published in the past 5 years. While some of them are for general use, several are targeted for specific categories of species such as humans, plants, or viruses, having optimized for the use of each of those datasets.

While all the predictors can be broadly distinguished as either sequence-based or GO-based, each tool has its own unique algorithm and approach to protein SCL. The tools incorporated a variety of algorithms and classifiers, including: SVM, recurrent neural networks (RNNs), convolutional neural networks (CNNs), HMM, mRMR (minimum redundancy maximum relevance), PseAAC (pseudo-amino acid composition), ML-GKR (multi-label Gaussian kernel recognition), and NBI (network-based inference).

DeepLoc is a sequence-based SCL prediction algorithm which utilizes deep learning in its algorithm, particularly attention models, RNNs with long short-term memory cells (LSTMs) and CNNs. Attention models and LSTMs were used to detect sorting signals in any position in a protein, while CNNs were used to train short motifs detection filters. Comparison of generalization performances show that the DeepLoc dataset has higher accuracy compared to other datasets. The prediction model used in the subcellular localization utilizes a reccurent neural network and an attention mechanism that can process the entire protein and identify the protein regions.

Hum-mPLoc 3.0 is a gene ontology (GO) based predictor for SCL. The authors proposed a novel feature presentation protocol called HCM (Hidden Correlational Modeling) for reflecting the structural hierarchy of the domain knowledge bases. The compact and discrimative feature vectors that HCM produce takes into account the residue statistics and biological background of the proteins using the data provided by GO terms. Using the correlation matrices of GO terms, Zhou et al. (2016) generated a feature vector of 84-D. SVM was used as a classifier, with 12 class labels corresponding to 12 subcellular locations. The disadvantage of this tool is that it relies on the GO database, which is still incomplete as of the present, with over 25% proteins associated with less than 5 terms. Human datasets were used in performance comparison tests; however, the authors have stated that this tool can also be used on datasets from other species.

HybridGO-Loc is multi-label GO-based SCL predictor. Instead of just focusing on the frequency of GO terms, HybridGO-Loc highlights the relationships between the GO terms by also taking into account semantic similarities between GO terms. Each of these two factors forms frequency vectors and similarity vectors, which are then combined to output a fusion vector. For multi-label classification, SVM was used due to the lack of suitability of using HMM and ANNs for GO-based predictors. GO vectors have high dimensionality, which SVMs are well equipped to handle. The experimental shows that the proposed hybrid-feature predictor outperforms predictors based on individual GO features and other state-of-the-art predictors.

mPLR-Loc is another multi-label GO-based SCL predictor from the same authors as HybridGO-Loc (Wan et al., 2014; Wan et al., 2015). There are two notable differences: only frequency vectors were used, and that an adaptive decision-based penalized logistic regression classifier was used instead of SVM. Additionally, mPLR-Loc is specifically designed for viral and plant protein predictions. However, mPLR-Loc can also predict protein SCLs of other species as it uses GO terms, which also possess cross-species properties. Beside predicting subcellular localization, mPLR-Loc is also able to provide probabilitic confidence score for prediction decision.

MSLVP is a two-tier prediction algorithm for multiple SCL predictions of viral proteins. The predictor uses SVM for classification and regression analysis. It takes into account sequence features such as amino acid composition, dipeptide composition, and physicochemical properties for multiclass classification. Feature selection was based on the mRMR method, which minimizes redundancy and maximizes the relevance of selected features.

pLoc-mHum is a multi-label GO-based predictor which utilizes an algorithm called PseAAC, which generates the pseudo-amino acid composition of a protein to represent protein sequences. This algorithm converts a protein sequence into a digital vector that can then be processed by pattern-recognition algorithms (Du et al., 2012). A problem with combining GO and PseAAC in past approaches is that the dimensions of the protein vectors became very high, as high as 9567 in one case (Chou & Shen, 2006). Cheng et al.'s (2017) approach with pLoc-mHum is that trivial GO information is discarded in order to minimize the dimensions of the PseAAC vector. pLoc-mHum utilizes the ML-GKR classifier. This predictor was intended for human datasets; its counterpart for plant datasets is pLoc-mPlant (Cheng & Chou, 2017). The only major difference between the two predictors is that

the pLoc-mPlant was trained using plant datasets, as opposed to pLoc-mHum.

PMLPR is a unique predictor in that its method uses recommendation systems to predict protein SCL. The recommendation system the authors chose was the NBI, a network-based algorithm. This algorithm constructs a bipartite network of users and objects; in this case, information was retrieved from SWISS-PROT and the cellular component in GO to construct the network. PMLPR predicts a list of locations for each protein based on recommender systems and it can properly overcome the multiple location prediction problems. PMLPR has a better efficiency compared to other methods.

CONCLUSION

Protein subcellular localization plays a critical role in determining the performance of a protein's function. Knowing where a protein is localized is important for gaining a better understanding on how each protein functions and the organization of the cell as a whole. A number of different tools and algorithms are available for protein subcellular localization prediction. When choosing a predictor, it is important to understand how each tool works and which algorithm will be most relevant to solve the problem. GO-based predictors tend to be more relevant to whole-scale or proteome analyses that place importance on relationships between proteins instead of individual proteins, whereas sequence-based predictors would be more efficient in predicting the SCL of individual proteins.

ACKNOWLEDGEMENTS

The authors would like to thanks Institute for Research and Corporate Social Responsibility of Indonesia International Institute for Life Sciences for supporting the development of this manuscript.

REFERENCES

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics, 33(21), 3387–3395. doi:10.1093/bioinformatics/btx431

Cheng, X., Xiao, X., & Chou, K.-C. (2017). pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics, 34(9), 1448–1456. doi:10.1093/bioinformatics/btx711

Cheng, X., Xiao, X., & Chou, K.-C. (2017). pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. Molecular BioSystems, 13(9), 1722–1727. doi:10.1039/c7mb00267j

Chou, K.C. and Shen, H.B. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteome Res., 5, 1888–1897.

Dönnes, P., & Höglund, A. (2004). Predicting protein subcellular localization: past, present, and future. Genomics, proteomics & bioinformatics, 2(4), 209–215. doi:10.1016/s1672-0229(04)02027-3

Du, P., Wang, X., Xu, C., & Gao, Y. (2012). PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Analytical Biochemistry,

425(2), 117–119. doi:10.1016/j.ab.2012.03.015

Kaushik, S., Chouhan, U., & Dwivedi, A. (2017). Study of Protein Subcellular Localization Prediction: A Review. International Journal of Life Science & Pharma Research Vol:7/Issue:3 ISSN 2250-0480. https://pdfs.semanticscholar.org/5217/579f89dab1832617dbc11b75246bcfbe4b00.pdf

Mirzaei Mehrabad, E., Hassanzadeh, R., & Eslahchi, C. (2018). PMLPR: A novel method for predicting subcellular localization based on recommender systems. Scientific reports, 8(1), 12006. doi:10.1038/s41598-018-30394-w

Qi, T., Qiu, T., Zhang, Q., Tang, K., Fan, Y., Qiu, J., … Cao, Z. (2014). SEPPA 2.0--more refined server to predict spatial epitope considering species of immune host and subcellular localization of protein antigen. Nucleic acids research, 42(Web Server issue), W59–W63. doi:10.1093/nar/gku395

Su, E., Chiu, H.-S., Lo, A., Hwang, J.-K., Sung, T.-Y., & Hsu, W.-L. (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics, 8(1), 330. doi:10.1186/1471-2105-8-330

Thakur, A., Rajput, A., & Kumar, M. (2016). MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. Molecular BioSystems, 12(8), 2572–2586. doi:10.1039/c6mb00241b

Wan, S., Mak, M. W., & Kung, S. Y. (2014). HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins. PloS one, 9(3), e89545. doi:10.1371/journal.pone.0089545

Wan, S., Mak, M.-W., & Kung, S.-Y. (2015). mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. Analytical Biochemistry, 473, 14–27. doi:10.1016/j.ab.2014.10.014

Zhou, H., Yang, Y., & Shen, H.-B. (2016). Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. Bioinformatics, btw723. doi:10.1093/bioinformatics/btw723