Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: xx-xx-2016
Application Note

OXFORD

Sequence analysis

FUEL-mLoc: Feature-Unified Prediction and Explanation of Multi-Localization of Cellular Proteins in Multiple Organisms

Shibiao Wan [1,*], Man-Wai Mak [2,*] and Sun-Yuan Kung [1]

[1] Department of Electrical Engineering, Princeton University, New Jersey, USA
[2] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Although many web-servers for predicting protein subcellular localization have been developed, they often have the following drawbacks: (1) lack of interpretability or interpreting results with heterogenous information which may confuse users; (2) ignoring multi-location proteins; and (3) only focusing on specific organism. To tackle these problems, we present an interpretable and efficient web-server, namely FUEL-mLoc, using **F**eature-**U**nified prediction and **E**xplanation of m̲ulti-**L̲o̲c̲**alization of cellular proteins in multiple organisms. Compared to conventional localization predictors, FUEL-mLoc has the following advantages: (1) using unified features (i.e., essential GO terms) to interpret why a prediction is made; (2) being capable of predicting both single- and multi-location proteins; and (3) being able to handle proteins of multiple organisms, including *Eukaryota, Homo sapiens, Viridiplantae, Gram-positive Bacteria, Gram-negative Bacteria* and *Virus*. Experimental results demonstrate that FUEL-mLoc outperforms state-of-the-art subcellular-localization predictors.

Availability and implementation: `http://bioinfo.eie.polyu.edu.hk/FUEL-mLoc/`

Contact: shibiao.wan@princeton.edu (Shibiao Wan) or enmwmak@polyu.edu.hk (Man-Wai Mak)

Supplementary information: Supplementary materials are available at *Bioinformatics* online and `http://bioinfo.eie.polyu.edu.hk/FUEL-mLoc/suppl.html`.

1 Introduction

With a huge number of novel proteins found in the post-genomic era, determining where a protein locates within a cell is an essential yet challenging step to unravel its functions, especially realizing the facts that many proteins are co-located or move between, several subcellular compartments (Murphy, 2010). Recent decades have witnessed numerous computational methods proposed to assist laborious and time-consuming wet-lab experimental methods. Existing computational methods use different kinds of feature information (Emanuelsson *et al.*, 2007; Mott *et al.*, 2002; Chou and Shen, 2010). Gene Ontology (GO) based methods are found to be superior to other methods in terms of performance (Chi and Nam, 2012).

However, existing methods have the following drawbacks. First, most of the existing methods lack interpretability in that they do not provide the necessary information to decipher why the prediction results are obtained. Recently, an interpretable predictor named YLoc (Briesemeister *et al.*, 2010) was proposed to tackle this problem. However, YLoc requires heterogeneous biological features such as sorting signals, PROSITE[1] patterns and GO terms, which are not always available for every protein. Besides, using heterogeneous information to interpret the results may also confuse the users. Second, many methods can only predict single-location proteins while multi-location proteins are often ignored. Actually, multi-label proteins are found to participate in various metabolic activities in multiple cellular compartments (Murphy, 2010). Third, most of the existing methods focus on proteins from one or two organisms. It is highly required to develop a predictor that can handle proteins from most of the common organisms.

To tackle the aforementioned problems, we present an interpretable and efficient web-server, namely FUEL-mLoc, which uses **F**eature-**U**nified

[1] http://prosite.expasy.org/

prediction and **E**xplanation of multi-**Loc**alization of proteins from multiple organisms. Unlike YLoc which uses heterogenous information for prediction interpretation, FUEL-mLoc interprets the prediction decisions by unified features (i.e., GO terms), which are more clear, self-evident and structured. Also, FUEL-mLoc can handle both single- and multi-location proteins from organisms of eukaryota, human, plant, Gram-positive bacteria, Gram-negative bacteria and virus.

2 Web-Server Description

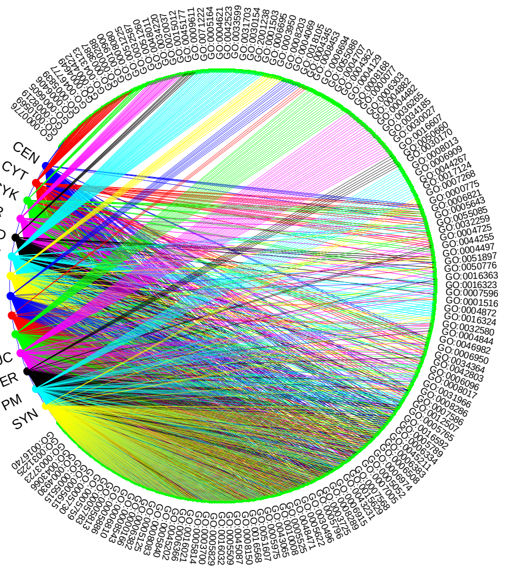

Figure 1. A network showing the relationships between the essential GO terms and each subcellular location for the human organism. Small green dots represent the GO terms and the large dots in different colors represent the 14 subcellular locations. A line connecting an essential GO term and a subcellular location denotes that the GO term contributes to the prediction of the subcellular location. On the contrary, if there is no line connecting an essential GO term with a particular subcellular location, then this GO term does not provide any information about the presence or absence of a protein in this particular subcellular location. See the supplementary materials for the acronyms of subcellular locations.

Table 1. Comparing FUEL-mLoc with state-of-the-art multi-label predictors based on leave-one-out cross validation (LOOCV). $m(n)$ means the overall locative accuracy (overall actual accuracy) (Wan et al., 2016). ^{a} and ^{b} are from (Chou and Shen, 2008) and (Chou and Shen, 2010), respectively. ^{c} is available at http://www.jci-bioinfo.cn/iLoc-Cell. Gram-Pos and Gram-Neg represent the Gram-positive bacteria and Gram-negative bacteria, respectively.

Dataset	Cell-PLoc^a	Cell-PLoc 2.0^b	iLoc-Cell^c	FUEL-mLoc
Eukaryote	0.393 (–)	0.642 (–)	0.791 (0.713)	**0.884 (0.793)**
Human	0.381 (–)	0.627 (–)	0.763 (0.682)	**0.836 (0.743)**
Plant	0.380 (–)	0.637 (–)	0.717 (0.681)	**0.916 (0.883)**
Gram-Pos	0.725 (–)	0.822 (–)	0.931 (0.929)	**0.968 (0.963)**
Gram-Neg	0.715 (–)	0.857 (–)	0.914 (0.899)	**0.953 (0.945)**
Virus	0.437 (–)	0.603 (–)	0.782 (0.748)	**0.925 (0.899)**

From the algorithmic perspective,[2] FUEL-mLoc has the following advantages. First, FUEL-mLoc adopts two newly created databases, namely ProSeq and ProSeq-GO (Wan *et al.*, 2014) instead of the conventional Swiss-Prot and GOA databases, to overcome the problem caused by the unavailability of GO information in novel proteins. With these two databases, FUEL-mLoc not only guarantees that every query protein can be associated with at least one GO term, but also significantly reduces the memory consumption. Details about these two databases can be found in Supplementary Information S1.

Second, FUEL-mLoc uses an elastic-net (EN) based multi-label classifier to predict protein subcellular localization and simultaneously selects essential GO terms to interpret the results. Fig. 1 shows the relationship between the essential GO terms and the subcellular locations (GO-SCL relationships) for human proteins. The GO-SCL relationships for the other five organisms can be found in Supplementary Information S2.

Third, with sparse yet interpretable features, FUEL-mLoc performs better than existing state-of-the-art predictors for six organisms, as demonstrated in Table 1. More performance comparisons between FUEL- mLoc and other predictors can be found in Supplementary Information S3.

Fourth, existing interpretable predictors, such as YLoc, use heterogeneous features such as sorting signals and PROSITE patterns. For some novel proteins, information on these features may not be available yet. Moreover, using diverse features makes interpretation of results difficult because users have to combine different interpretations by themselves. The details about how FUEL-mPLoc and YLoc allow users to interpret the prediction results can be found in Supplementary Information S4.

Fifth, compared to other predictors, FUEL-mLoc has an extra function that allows users to use the online GO-term database from the QuickGO server, which can provide the most up-to-date GO information for predicting protein subcellular localization. Details can be found in Supplementary Information S6.

3 Conclusion

We developed a web-server, namely FUEL-mLoc, which not only predicts single- and multi-label proteins from almost all of the common organisms (such as eukaryote, human, plant, Gram-positive, Gram-negative and virus), but also provides interpretable information to explain why the prediction decisions are made. Compared to existing interpretable web-servers (e.g., YLoc), FUEL-mLoc uses unified feature information for more clear, self-evident and structured interpretation. Besides, FUEL-mLoc performs better than other subcellular-localization predictors.

Acknowledgement

This work was in part supported by The Hong Kong Research Grant Council Grant Nos. PolyU152068/15E and PolyU152518/16E.

References

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). YLoc—an interpretable web server for predicting subcellular localization. *Nucleic Acids Research*, **38**(Suppl 2), W497–W502.

Chi, S.-M. and Nam, D. (2012). Wegoloc: accurate prediction of protein subcellular localization using weighted gene ontology terms. *Bioinformatics*, **28**(7), 1028–1030.

Chou, K. C. and Shen, H. B. (2008). Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols*, **3**, 153–162.

Chou, K. C. and Shen, H. B. (2010). Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Sci.*, **2**, 1090–1103.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols*, **2**(4), 953–971.

Mott, R., Schultz, J., Bork, P., and Ponting, C. (2002). Predicting protein cellular localization using a domain projection method. *Genome research*, **12**(8), 1168–1174.

Murphy, R. F. (2010). communicating subcellular distributions. *Cytometry*, **77**(7), 686–92.

Wan, S., Mak, M. W., and Kung, S. Y. (2014). R3P-Loc: A compact multi-label predictor using ridge regression and random projection for protein subcellular localization. *Journal of Theoretical Biology*, **360**, 34–45.

Wan, S., Mak, M. W., and Kung, S. Y. (2016). Sparse regressions for predicting and interpreting subcellular localization of multi-label proteins. *BMC Bioinformatics*, **17**(97).

[2] See (Wan *et al.*, 2016) for details