Journal of Theoretical Biology 398 (2016) 32–42

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional types of membrane proteins

Shibiao Wan [a,*], Man-Wai Mak [a,*], Sun-Yuan Kung [b]

[a] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China
[b] Department of Electrical Engineering, Princeton University, NJ, USA

HIGHLIGHTS

- Mem-ADSVM outperforms state-of-the-art multi-label membrane-protein predictors.
- Mem-ADSVM can predict both membrane proteins and their functional types.
- The proposed adaptive-decision scheme is conducive to performance improvement.
- Mem-ADSVM is accessible at http://bioinfo.eie.polyu.edu.hk/MemADSVMServer/.

ARTICLE INFO

Article history:
Received 20 November 2015
Received in revised form
7 March 2016
Accepted 7 March 2016
Available online 19 March 2016

Keywords:
Membrane protein type prediction
Multi-label classification
Adaptive-decision scheme
Gene ontology
Two-layer classification

ABSTRACT

Identifying membrane proteins and their multi-functional types is an indispensable yet challenging topic in proteomics and bioinformatics. However, most of the existing membrane-protein predictors have the following problems: (1) they do not predict whether a given protein is a membrane protein or not; (2) they are limited to predicting membrane proteins with single-label functional types but ignore those with multi-functional types; and (3) there is still much room for improvement for their performance. To address these problems, this paper proposes a two-layer multi-label predictor, namely Mem-ADSVM, which can identify membrane proteins (Layer I) and their multi-functional types (Layer II). Specifically, given a query protein, its associated gene ontology (GO) information is retrieved by searching a compact GO-term database with its homologous accession number. Subsequently, the GO information is classified by a binary support vector machine (SVM) classifier to determine whether it is a membrane protein or not. If yes, it will be further classified by a multi-label multi-class SVM classifier equipped with an adaptive-decision (AD) scheme to determine to which functional type(s) it belongs. Experimental results show that Mem-ADSVM significantly outperforms state-of-the-art predictors in terms of identifying both membrane proteins and their multi-functional types. This paper also suggests that the two-layer prediction architecture is better than the one-layer for prediction performance. For reader's convenience, the Mem-ADSVM server is available online at http://bioinfo.eie.polyu.edu.hk/MemADSVMServer/.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

According to the characteristic sequence and structural features, proteins are divided into four common types (Andreeva et al., 2014): membrane, soluble (or globular), fibrous and intrinsically disordered. Among them, membrane proteins are found to play essential roles in a variety of vital biological processes (Almén et al., 2009) by interacting with the membranes of a cell or an organelle. Membrane proteins are targets of almost half of all medicinal drugs (Bakheet and Doig, 2009; Overington et al., 2006), because they mediate many interactions between cells and extracellular surroundings as well as between the cytosol and membrane-bound organelles. Despite owning the same basic phospholipid bilayer structure (Lodish et al., 2000), membrane proteins perform various and diversified functions. Therefore, given a query protein, identifying whether it is a membrane protein or not is an indispensable yet challenging topic in proteomics and bioinformatics.

Membrane proteins can be further divided into different functional types. Conventionally, some studies (Lodish et al., 2000) broadly classified membrane proteins into two categories, namely integral (or intrinsic) and peripheral (or extrinsic), depending upon the interactions between membrane proteins and the membrane. Other studies (Gerald, 2013) grouped membrane proteins into three distinct classes:

* Corresponding authors.
E-mail addresses: shibiao.wan@connect.polyu.hk (S. Wan), enmwmak@polyu.edu.hk (M.-W. Mak), kung@princeton.edu (S.-Y. Kung).

http://dx.doi.org/10.1016/j.jtbi.2016.03.013
0022-5193/© 2016 Elsevier Ltd. All rights reserved.

Table 1
Summary of existing predictors for identifying membrane proteins (Layer I) and predicting membrane protein functional types. *Pse-PSSM*: pseudo position-specific score matrix (PSSM); *AA*: amino-acid composition; *PseAA*: pseudo amino acid composition; *improved PSSM*: PSSM with AA physical-chemical properties; *KNN*: K-nearest neighbor; *OET-KNN*: optimized evidence-theoretic KNN; *EN*: elastic net.

Stratification	Predictor	Features	Classifier	Multi-label	No. of classes
Layer I	MemType-2L (Chou and Shen, 2007)	Pse-PSSM	ensemble OET-KNN	No	2
	LeastEudist (Nakashima et al., 1986)	AA	least Euclidean distance	No	2
	ProtLoc (Cedano et al., 1997)	AA	least Mahalanobis distance	No	2
Layer II	Mem-PseAA (Huang and Yuan, 2013)	PseAA	multi-label KNN	Yes	8
	iMem-Seq (Xiao et al., 2015)	improved PSSM	multi-label KNN	Yes	8
	Mem-mEN (Wan et al., 2015)	GO terms	multi-label EN	Yes	8

integral, peripheral and lipid-anchored. With the exponentially growing number of protein sequences discovered in the post-genomic era, these three classes of membrane proteins are further divided into eight types (Chou and Shen, 2007): (1) single-pass type I; (2) single-pass type II; (3) single-pass type III; (4) single-pass type IV; (5) multi-pass; (6) lipid-anchor; (7) GPI-anchor and (8) peripheral. More information about the hierarchical relationships between these eight types and the former three classes can be found in Wan et al. (2015). Particularly, GPI-anchored proteins (Type 7) is a kind of special lipid-anchored proteins (Type 6).[1] Type 7 is singled out from Type 6 because GPI-anchored proteins ubiquitously exist in many species and have been intensively studied for their unique functions (Ikezawa, 2002). Details about these eight functional types are also elaborated in Wan et al. (2015).

Knowing the functional types of membrane proteins can be helpful to elucidate the biological functions of membrane proteins. For example, phospholipases (Tappia and Dhalla, 2014), belonging to Type-8, are a group of water-soluble enzymes that are temporarily bound to the polar head groups of membrane phospholipids. Their major functions are lipid signaling, which can be achieved by hydrolyzing various bonds linking phospholipases with the lipid layer to which they are temporarily attached. Moreover, about 20–35% of genes contain the instructions for producing membrane proteins, whereas the structurally annotated membrane proteins only account for less than 1% of the proteins with known structures (Nanni and Lumini, 2008). Knowing the functional types of membrane proteins can accelerate the process of annotating their structures. Besides, because of their fluidity property, membrane proteins can freely move within the lipid bilayer to the location where their functions are performed. The knowledge of the functional type of a membrane protein can help reveal the mechanisms of this kind of biological activities. Therefore, it is highly necessary to develop computational approaches for timely and accurate prediction of the functional types of membrane protein. Ideally, the computational approaches should perform two-layer predictions.

1. Layer I: Given a query protein, the predictor determines whether the query protein is a membrane protein or not.
2. Layer II: If the answer in Layer I prediction is 'yes', the predictor determines the functional type(s) of the protein.

In recent years, impressive progress has been made in predicting the functional types of membrane proteins (Chou and Shen, 2007; Nanni and Lumini, 2008; Chou and Cai, 2005; Cai et al., 2003; Zou, 2014; Hayat et al., 2012; Ding et al., 2012; Wang et al., 2010) and in the prediction of membrane proteins in specific subcellular locations (Tripathi and Gupta, 2014; Yuan and Teasdale, 2002). While many advanced predictors have been developed, they still have several limitations, which are elaborated below.

1. These predictors assume that all query proteins are membrane proteins. If a query protein is not a membrane protein, these predictors will attempt to determine the most likely functional type of the protein. This is obviously undesirable because a non-membrane protein does not have a *membrane* functional type. Given that membrane proteins are just one of the four common types of proteins (Andreeva et al., 2014), it is important to ensure that the query protein is really a membrane protein prior to determining its functional type(s). As far as we know, only three predictors, namely LeastEudist Nakashima et al. (1986), ProtLoc Cedano et al. (1997) and MemType-2L Chou and Shen (2007),[2] are capable of predicting whether a query protein is a membrane protein or not. These predictors are summarized in Layer I of Table 1. As can be seen, these predictors use sequence-based features (i.e., amino-acid compositions and pseudo position-specific score matrices) to discriminate membrane proteins from non-membrane proteins.
2. They are limited to the prediction of membrane proteins with single-label functional types. However, many membrane proteins were found to simultaneously belong to multiple functional types. For example, the envelope glycoprotein p57 (Clemente and Juan, 2009; Vahlenkamp et al., 2002) was reported to belong to single-pass type I (Type 1) when locating in the host endoplasmic reticulum membrane, and simultaneously it belongs to peripheral (Type 8) when locating in the host cell membrane. Table 1 lists the existing multi-label predictors for membrane protein type prediction (Layer II). To the best of our knowledge, only three predictors, namely Mem-PseAA Huang and Yuan (2013), iMem-Seq Xiao et al. (2015) and Mem-mEN Wan et al. (2015),[3] are able to predict multi-label membrane proteins. In terms of feature extraction, iMem-Seq uses the information from position-specific score matrices and physical-chemical property matrices; Mem-PseAA uses feature information from pseudo-amino acid compositions; Mem-mEN uses the information from homologous gene ontology term frequencies. In terms of classification, both Mem-PseAA and iMem-Seq use multi-label kNN classifiers, whereas Mem-mEN uses a multi-label elastic net (EN) classifier. Particularly, Mem-mEN possesses the property of 'interpretability' (Wan et al., 2015), which can provide biological reasons on why a query protein belongs to the predicted type(s).
3. The performance of existing predictors are far from satisfactory. In particular, it has been shown (Wan et al., 2015) that Mem-mEN performs better than Mem-PseAA and iMem-Seq. However, the performance of all of these predictors still remains to be improved. Besides, the one-layer architecture of these predictors cannot model or capture the intrinsic inter-class correlations, such as the

[1] http://www.uniprot.org/locations/SL-9902.

[2] Note that LeastEudist and ProtLoc were implemented in Chou and Shen (2007). For ease of reference, we use the name LeastEudist to denote the method proposed in Nakashima et al. (1986).

[3] For ease of reference, we refer to the predictor proposed in Huang and Yuan (2013) as Mem-PseAA.

Download English Version:

https://daneshyari.com/en/article/4495814

Download Persian Version:

https://daneshyari.com/article/4495814

Daneshyari.com