Journal of Theoretical Biology 416 (2017) 81–87

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions

Ya-Sen Jiao, Pu-Feng Du*

School of Computer Science and Technology, Tianjin University, Tianjin 300350, China

ARTICLE INFO

Keywords:
SVM
Positional-Specific Physicochemical Properties
Multi-label
intermembrane space proteins

ABSTRACT

Predicting protein submitochondrial locations has been studied for about ten years. A dozen of methods were developed in this regard. Although a mitochondrion has four submitochondrial compartments, all existing studies considered only three of them. The mitochondrial intermembrane space proteins were always excluded in these studies. However, there are over 50 mitochondrial intermembrane space proteins in the recent release of UniProt database. We think it is time to incorporate these proteins in predicting protein submitochondrial locations. We proposed the functional domain enrichment score, which can be used as an enhancement to our positional-specific physicochemical properties method. We constructed a high-quality working dataset from the UniProt database. This dataset contains proteins from all four submitochondrial locations. Proteins with multiple submitochondrial locations are also included. Our method achieved over 70% prediction accuracy for proteins with single location on this dataset. On the M3-317 benchmarking dataset, our method achieved comparable prediction performance to other state-of-the-art methods. Our results indicate that the intermembrane space proteins can be incorporated in predicting protein submitochondrial locations. By evaluating our method with the proteins that have multiple submitochondrial locations, we conclude that our method is capable of predicting multiple submitochondrial locations. This is the first report of *ab initio* methods that can identify intermembrane space proteins. This is also the first attempt to incorporate proteins with multiple submitochondrial locations. The benchmarking dataset can be obtained by emails to the corresponding author.

1. Introduction

A eukaryotic cell usually contains a nucleus and some other organelles that are enclosed by membranes. A mitochondrion is one of these organelles. It is enclosed by two layers of membranes. A eukaryotic cell usually contains several mitochondria. The mitochondria are involved in many different cellular processes, such as energy metabolism, programmed cell death, and ionic homeostasis (Berardi et al., 2011; Oxenoid et al., 2016). Mitochondria are related to many human diseases and the aging process. Different number of proteins are found in mitochondria of different species.

Computational prediction of protein subcellular locations have been extensively studied in the last two decades. Many different computational methods and online services have been developed. These computational tools have been proved to be useful in helping life science studies. In recent years, the studies in predicting protein subcellular locations focused on four topics: (1) predicting protein sub-subcellular locations (Du et al., 2011), such as protein subnuclear locations (Wang and Liu, 2015), submitochondrial locations (Du and Li, 2006), subchloroplast locations (Wang et al., 2015), and sub-Golgi locations (Jiao and Du, 2016a); (2) predicting multi-label protein subcellular locations (Du and Xu, 2013); (3) predicting subcellular location for proteins with specific structural topology (Du et al., 2012); and (4) predicting alterations of protein subcellular locations under different conditions, such as disease, drug therapies, and environmental stress (Lee et al., 2013). In this paper, we will touch the first and the second topics.

A mitochondrion is enclosed by two layers of membranes, which are known as the outer membrane and the inner membrane. The outer membrane separate a mitochondrion from the cytosol, while the inner membrane separate the mitochondria internal space into two parts. The space within the inner membrane contains the mitochondrial matrix. The mitochondrial DNA, ribosomes, enzymes and many other kinds of molecules can be found in the matrix. The space between the inner membrane and the outer membrane is called the intermembrane space. In general, there are four different submitochondrial compart-

* Corresponding author.
E-mail address: PufengDu@gmail.com (P.-F. Du).

http://dx.doi.org/10.1016/j.jtbi.2016.12.026
Received 21 September 2016; Received in revised form 6 December 2016; Accepted 30 December 2016
Available online 08 January 2017

ments: the outer membrane, the intermembrane space, the inner membrane and the matrix.

Over the last few years, several studies focused on predicting protein submitochondrial locations. The first report appeared in the year 2006. Du and Li proposed the SubMito method, which can predict protein submitochondrial locations using sequence-based features (Du and Li, 2006). They also released the first benchmarking dataset in predicting protein submitochondrial locations. This dataset is currently known as the M3-317 dataset. Nanni and Lumini developed the GPLoc method based on genetic algorithm (Nanni and Lumini, 2008). Zeng et al. presented the Predict_subMITO method by using an augmented version of pseudo amino acid compositions (Zeng et al., 2009). Shi et al. introduced the wavelet-SVM based method to improve the prediction performance (Shi et al., 2011). Fan and Li further improved the prediction accuracy by hybridizing six different sequence descriptors (Fan and Li, 2012). Zakeri et al. employed another hybrid method to significantly improve the prediction performance (Zakeri et al., 2011). Lin et al. used the overrepresented tetra-peptides to predict the protein submitochondrial locations (Lin et al., 2013). Du and Yu developed the Submito-PSPCP with the concept of Positional Specific Physicochemical Properties (PSPCP) (Du and Yu, 2013). Ahmad et al. uses split amino acid compositions and feature selection methods to achieve better prediction performance (Ahmad et al., 2016). Li et al. significantly improved the prediction performance by integrating many different features (Li et al., 2014). All these methods reported the jackknife test results on the M3-317 dataset.

No existing study considered all four compartments of a mitochondrion. The intermembrane space is always missing. When the SubMito was published in the year 2006, the number of known protein sequences in the intermembrane space is too limited to train a machine learning-based predictor. Although ignoring the intermembrane space proteins was a feasible solution in those days to construct a working model with limited data resources, the proteins in the intermembrane space are always wrongly predicted. Therefore, the problem of predicting protein submitochondrial locations is not fully solved. In the UniProt release 2016_04, there are over 50 reviewed protein sequences that are experimentally annotated with “mitochondrion intermembrane space”. This number is still not large, but it is possible to train a machine learning-based predictor that can take all four submitochondrial locations into consideration. Therefore, in the current paper, by incorporating the intermembrane space as the fourth location, our work actually goes a big step further in predicting protein submitochondrial locations.

Moreover, the proteins with multiple submitochondrial locations actually exist. But no existing method can predict multi-label protein submitochondrial locations (Chou, 2013). Therefore, we tried to predict the multi-label protein submitochondrial locations. In this work, we integrated the functional domain information to further improve the performance of SubMito-PSPCP.

2. Materials and methods

2.1. Datasets

We used two datasets in this work. One is the M3-317 dataset. As all state-of-the-art methods reported the prediction performance on the M3-317 dataset, it is a common ground for performance comparison. The other dataset was extracted from the UniProt database. We considered four submitochondrial locations in this new dataset. We describe the procedures to create this dataset as follows:

Firstly, all reviewed sequences, which have been experimentally annotated with subcellular location “Mitochondrion”, were extracted from the UniProt database (Release 2016_04) (UniProt Consortium, 2015) using the online query system.

Secondly, the sequences without experimentally annotated submitochondrial locations were discarded. In this work, we took only four

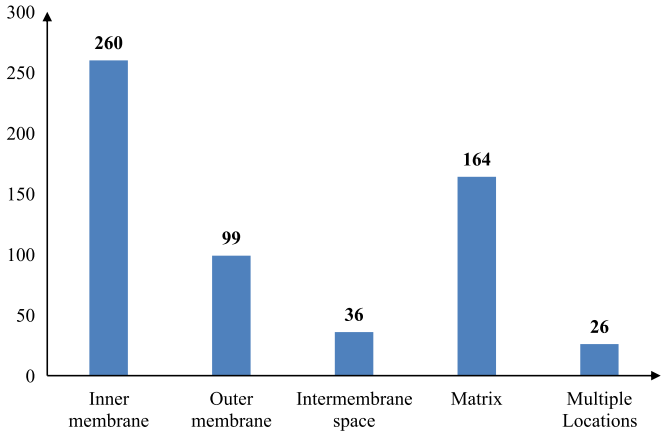

Fig. 1. Number of proteins in different submitochondrial locations in M4-585 dataset. There are 260 proteins in inner membrane, 99 in outer membrane, 36 in intermembrane space and 164 in matrix. These 559 proteins are single-location proteins. The remaining 26 proteins are multiple-locations proteins.

different terms in the controlled vocabulary of UniProt as effective submitochondrial location annotations. They are the “Mitochondrion inner membrane”, “Mitochondrion intermembrane space”, “Mitochondrion matrix”, and “Mitochondrion outer membrane”. The accession numbers of these four terms in the UniProt controlled vocabulary are “SL-0168”, “SL-0169”, “SL-0170”, and “SL-0172”, respectively.

Thirdly, all sequences that contain ambiguous amino acid notations, such as “X”, “B”, and “Z”, were discarded.

Finally, the sequences were processed by the CD-HIT program (Fu et al., 2012) to remove the highly homologous sequences. The sequence similarity cutoff was set to 40%. The remaining 585 sequences formed our working dataset. We call this dataset the M4-585 dataset.

Among the 585 protein sequences, there are 559 proteins with a unique submitochondrial location. 23 of the 585 protein sequences have two submitochondrial locations. 3 of the 585 protein sequences have three submitochondrial locations. The breakdown of the dataset can be found in Fig. 1.

Due to the limited number of proteins with multiple locations, it is impossible to train a real multi-label predictor. We used only proteins with single location to train our predictor. However, by our design, our predictor can rank all possible locations. If the number of desired locations is given, our predictor is able to output multiple locations as results. As the proteins with multiple locations are totally blinded to the training process, we used them as an independent testing dataset.

2.2. Positional-Specific Physicochemical Properties

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elucidated in a recent review (Chou, 2015). However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition or PseAAC was proposed. Ever since the concept of pseudo amino acid composition or Chou's PseAAC was proposed, it has penetrated into many biomedicine and drug development areas (Zhong and Zhou, 2014) and nearly all the areas of computational proteomics (Chou, 2009). Because it has been widely and increasingly used, several open source programs were established, such as the famous propy (Cao et al., 2013) and PseAAC-General (Du