Analytical Biochemistry 394 (2009) 269–274

Contents lists available at ScienceDirect

Analytical Biochemistry

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

A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0

Hong-Bin Shen *, Kuo-Chen Chou *

Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai 200240, China
Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

ARTICLE INFO

Article history:
Received 21 May 2009
Available online 3 August 2009

Keywords:
Multiplex protein
Homology search
Representative proteins
Gene ontology
Functional domain
Sequential evolution
Ensemble classifier
Fusion approach

ABSTRACT

Predicting subcellular localization of human proteins is a challenging problem, particularly when query proteins may have a multiplex character, i.e., simultaneously exist at, or move between, two or more different subcellular location sites. In a previous study, we developed a predictor called "Hum-mPLoc" to deal with the multiplex problem for the human protein system. However, Hum-mPLoc has the following shortcomings. (1) The input of accession number for a query protein is required in order to obtain a higher expected success rate by selecting to use the higher-level prediction pathway; but many proteins, such as synthetic and hypothetical proteins as well as those newly discovered proteins without being deposited into databanks yet, do not have accession numbers. (2) Neither functional domain nor sequential evolution information were taken into account in Hum-mPLoc, and hence its power may be reduced accordingly. In view of this, a top-down strategy to address these shortcomings has been implemented. The new predictor thus obtained is called Hum-mPLoc 2.0, where the accession number for input is no longer needed whatsoever. Moreover, both the functional domain information and the sequential evolution information have been fused into the predictor by an ensemble classifier. As a consequence, the prediction power has been significantly enhanced. The web server of Hum-mPLoc2.0 is freely accessible at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/.

Numerous efforts have been made to develop various methods for predicting protein subcellular localization based on the sequence information (see, e.g., [1–8] and a long list of references cited in two comprehensive review papers [9,10]). However, for practical applications in drug development, it is more important and urgent to timely determine the subcellular locations of human proteins. Unfortunately, relatively much fewer predictors were established that are specialized for predicting the subcellular localization of human proteins.

Although the HSLPred developed by Garg et al. [11] was specifically for human proteins, the predictor can only cover four subcellular location sites: cytoplasm, mitochondria, nucleus, and plasma membrane. If a user used HSLPred [11] to predict a query protein located outside the aforementioned four sites, such as lysosome and centriole, the predictor would fail to work, or the results thus obtained would not make any sense.

To improve the coverage limit, the predictor called Hum-PLoc [12] was developed to extend the coverage scope for human proteins from 4 to 12 location sites, i.e., the aforementioned 4 sites plus the following 8 sites: centriole, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, microsome, and peroxisome. However, Hum-Ploc [12] cannot be used to deal with multiplex proteins, which may simultaneously exist at, or move between, two or more different subcellular locations. Proteins with multiple locations or dynamic feature of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery [13,14]. According to a statistical analysis on the Swiss-Prot database (version 55.3), this kind of multiplex proteins might occupy about 20% of the human proteins.

To make Hum-Ploc be able to predict the multiplex protein locations as well, the predictor called Hum-mPLoc [15] was developed. Meanwhile, the subcellular location scope covered by Hum-mPLoc was further extended to the 14 sites; i.e., the aforementioned 12 location sites plus endosome and synapse. Even though, Hum-mPLoc could still yield about 70% jackknife cross-validation success rate when tested by a very stringent benchmark dataset in which none of the proteins included has $\geqslant$25% pairwise sequence identity to any other protein in the same subcellular location subset. The Hum-mPLoc predictor was established by hybridizing the "higher-level" GO (gene ontology [16]) approach and PseAAC (pseudo amino acid composition [17,18]) approach. Its power mainly came from the GO approach because proteins formulated in the GO database space would be clustered in a way

* Corresponding authors.
E-mail addresses: hbshen@sjtu.edu.cn (H.-B. Shen), kcchou@gordonlifescience.org (K.-C. Chou).

0003-2697/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2009.07.046

much better reflecting their subcellular locations, as elucidated in [19].

However, the existing version of Hum-mPLoc has the following problems. (1) In order to take advantage of the GO approach, the input for a query protein must include its accession number. Many proteins, such as synthetic and hypothetical proteins as well as those newly discovered proteins that have not been deposited into databanks yet, do not have accession numbers, and hence their subcellular locations cannot be predicted via the GO approach. (2) Since the current GO database is far from complete yet, many proteins cannot be meaningfully formulated in a GO space even if their accession numbers are available. (3) Although the PseAAC approach, a complement to the GO approach in Hum-mPLoc, can take into account some partial sequence order effects, the original PseAAC [17,20] missed the functional domain and sequential evolution information.

The present study was initiated in an attempt to develop a new and more powerful predictor, called Hum-mPLoc 2.0, for predicting human protein subcellular localization by addressing the above three problems.

Materials

Protein sequences were collected from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/. The detailed procedures are basically the same as those in [15]. The only difference is that, in order to obtain the updated data, instead of version 50.7 released on 9 September 2006, the version 55.3 released on 29 April 2008 is adopted. After strictly following the procedures as described in [15], we finally obtained a benchmark dataset of 3106 different protein sequences covering 14 subcellular locations (see Table 1), where 2580 proteins belong to one subcellular location, 480 to two locations, 43 to three locations, and 3 to four locations. The corresponding accession numbers and protein sequences are given in Online Supporting Information A. Note that because some proteins may occur in two or more locations, the 3106 different proteins actually correspond to 3681 locative proteins. The concept of "locative proteins" was introduced for studying proteins with multiple subcellular location sites, as illustrated in [10,15].

Table 1
Breakdown of the human protein benchmark dataset derived from Swiss-Prot database (release 55.3) according to the procedures described under Materials (none of proteins included here has ⩾25% pairwise sequence identity to any other in a same subcellular location).

Order	Subcellular location	Number of proteins
1	Centriole	77
2	Cytoplasm	817
3	Cytoskeleton	79
4	Endoplasmic reticulum	229
5	Endosome	24
6	Extracell	385
7	Golgi apparatus	161
8	Lysosome	77
9	Microsome	24
10	Mitochondrion	364
11	Nucleus	1021
12	Peroxisome	47
13	Plasma membrane	354
14	Synapse	22
Total number of locative proteins $\tilde{N}$		3681[a]
Total number of different proteins N		3106[b]

[a] See Eqs. (1)–(4) of [15] for the definition about the number of locative proteins, and its relation with the number of different proteins.
[b] Of the 3106 different proteins, 2580 belong to one subcellular location, 480 to two locations, 43 to three locations, and 3 to four location.

Methods

The key in developing a powerful method for predicting protein subcellular localization is to grasp the core features of proteins that are intrinsically correlated with their localization in a cell. In this regard, the strategy by hybridizing the GO representation and PseAAC representation was quite successful, as demonstrated in [12,15]. Therefore, we shall continue adopting the hybridization strategy in the current study. However, in order to solve the three problems raised in the Introduction, the detailed procedures to realize the hybridization approach will be completely different, as elaborated below.

GO representation

GO is a controlled vocabulary used to describe the biology of a gene product in any organism [21,22]. The GO representation for a protein sample in the original Hum-mPLoc [15] was derived by first searching for its accession number against all the UniProt accession numbers and their corresponding GO numbers in the GO database [21], followed by mapping the GO information thus obtained into the representation for the protein sample. Therefore, in using Hum-mPLoc for prediction, the accession number of a query protein would be indispensable. To avoid such a problem, here let us derive the GO representation according to the following procedures.

Step 1. Use BLAST [23] to search the homologous proteins of the query protein **P** from the Swiss-Prot database (version 55.3), with the BLAST parameter of expect value $E \leqslant 0.001$.

Step 2. Those proteins which have ⩾60% pairwise sequence identity with the query protein **P** are collected as its *representative proteins*; meanwhile, their corresponding accession numbers in the Swiss-Prot database are also obtained accordingly.

Step 3. Search each of these accession numbers collected in Step 2 against the GO database at http://www.ebi.ac.uk/GOA/ to retrieve the GO information [21].

Step 4. The current GO database (version 70.0 released March 10 2008) contains 60,020 GO numbers; thus the query protein **P** can be formulated through its representative proteins by the equation

$$\mathbf{P}_{\mathrm{GO}} = \left[\delta_1^{\mathrm{G}} \quad \delta_2^{\mathrm{G}} \quad \cdots \quad \delta_i^{\mathrm{G}} \quad \cdots \quad \delta_{60020}^{\mathrm{G}}\right]^{\mathrm{T}}, \tag{1}$$

where **T** is the transposing operator, and

$$\delta_i^{\mathrm{G}} = \begin{cases} 1, & \text{if a hit found against the } i\text{-th GO number} \\ & \text{for any of the representative proteins of } \mathbf{P} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Through the above steps, we can study the query protein **P** by means of the GO information derived from its representative proteins. The rationale to do so is based on the fact that homology proteins generally share similar attributes, such as biological functions and structural conformations [24,25]. The reason for using the value of 60% as a threshold here is due to the fact that for most cases proteins with 60% or higher sequence identity can be usually treated as homologous to each other [26]. Actually, our preliminary tests also indicated that such a threshold was a good choice.

Thus, the accession number is no longer required for the input of the query protein even when using the high-level GO approach to predict its subcellular localization as required in Hum-PLoc [12] and Hum-mPLoc [15].

The above homology-based GO extraction method is very useful for studying those proteins which do not have UniProt accession numbers. However, it would still fail to work under any one of the following two situations: (1) the query protein does not have