

# Plant-mPLOC: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization

Kuo-Chen Chou<sup>1,2\*</sup>, Hong-Bin Shen<sup>1,2</sup>

**1** Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, **2** Gordon Life Science Institute, San Diego, California, United States of America

## Abstract

One of the fundamental goals in proteomics and cell biology is to identify the functions of proteins in various cellular organelles and pathways. Information of subcellular locations of proteins can provide useful insights for revealing their functions and understanding how they interact with each other in cellular network systems. Most of the existing methods in predicting plant protein subcellular localization can only cover three or four location sites, and none of them can be used to deal with multiplex plant proteins that can simultaneously exist at two, or move between, two or more different location sites. Actually, such multiplex proteins might have special biological functions worthy of particular notice. The present study was devoted to improve the existing plant protein subcellular location predictors from the aforementioned two aspects. A new predictor called “Plant-mPLOC” is developed by integrating the gene ontology information, functional domain information, and sequential evolutionary information through three different modes of pseudo amino acid composition. It can be used to identify plant proteins among the following 12 location sites: (1) cell membrane, (2) cell wall, (3) chloroplast, (4) cytoplasm, (5) endoplasmic reticulum, (6) extracellular, (7) Golgi apparatus, (8) mitochondrion, (9) nucleus, (10) peroxisome, (11) plastid, and (12) vacuole. Compared with the existing methods for predicting plant protein subcellular localization, the new predictor is much more powerful and flexible. Particularly, it also has the capacity to deal with multiple-location proteins, which is beyond the reach of any existing predictors specialized for identifying plant protein subcellular localization. As a user-friendly web-server, Plant-mPLOC is freely accessible at <http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/>. Moreover, for the convenience of the vast majority of experimental scientists, a step-by-step guide is provided on how to use the web-server to get the desired results. It is anticipated that the Plant-mPLOC predictor as presented in this paper will become a very useful tool in plant science as well as all the relevant areas.

**Citation:** Chou K-C, Shen H-B (2010) Plant-mPLOC: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE* 5(6): e11335. doi:10.1371/journal.pone.0011335

**Editor:** Edward Newbigin, University of Melbourne, Australia

**Received:** April 19, 2010; **Accepted:** June 4, 2010; **Published:** June 28, 2010

**Copyright:** © 2010 Chou, Shen. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by the National Natural Science Foundation of China (Grant No. 60704047), Science and Technology Commission of Shanghai Municipality (Grant No. 08ZR1410600, 08JC1410600) and sponsored by the Shanghai Pujiang Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [kcchou@gordonlifescience.org](mailto:kcchou@gordonlifescience.org)

## Introduction

Information of the subcellular localization of proteins is important because it can (1) indicate how and under what kind of cellular environments they interact with each other and with other molecules, (2) provide useful clues for revealing their functions, and (3) help understand the intricate pathways that regulate biological processes at the cellular level [1,2]. Although this kind of information can be acquired by conducting various biochemical experiments, it is both time consuming and expensive to determine the subcellular localization of uncharacterized proteins one by one with experiments alone. With the avalanche of protein sequences generated in the Post-Genomic Age, it is highly desired to develop computational methods that can be used to identify the subcellular location site(s) of a newly found protein based on its sequence information alone.

During the past 17 years or so, numerous efforts have been made in this regard (see, e.g., [3,4,5,6,7,8,9,10] as well as a long list of references cited in two comprehensive review articles [11,12]).

However, relatively much fewer predictors were developed specialized for predicting the subcellular localization of plant proteins. To the best of our knowledge, of the aforementioned methods only the one called “TargetP” [6] and the one called “Predotar” [8] are specialized for plant proteins. Ever since the two predictors were proposed, they have been widely used for studying various plant protein systems and related areas. However, TargetP and Predotar can discriminate plant proteins among only three or four location sites. For instance, TargetP [6] only covers the following sites: (1) mitochondria, (2) chloroplast, (3) secretory pathway, and (4) other. And Predotar [8] only covers the following sites: (1) endoplasmic reticulum, (2) mitochondrion, (3) plastid, and (4) other. After removing the ambiguous location of “other”, TargetP or Predotar actually covers only three subcellular location sites. If a user tried to use TargetP and Predotar to predict a query protein located outside the aforementioned sites, such as cell wall, peroxisome, Golgi apparatus, or vacuole, the two predictors would either fail to work or generate meaningless outcomes.

To improve the situation, the predictor called “Plant-PLOC” [13] was developed to extend the coverage scope for plant proteins from the three locations covered by TargetP or Predotar to the following eleven: **(1)** cell wall, **(2)** chloroplast, **(3)** cytoplasm, **(4)** endoplasmic reticulum, **(5)** extracellular, **(6)** mitochondrion, **(7)** nucleus, **(8)** peroxisome, **(9)** plasma membrane, **(10)** plastid, and **(11)** vacuole. The Plant-PLOC predictor was established by integrating the “higher-level” GO (gene ontology) [14] approach and PseAAC (pseudo amino acid composition) [15] approach. GO is a controlled vocabulary used to describe the biology of a gene product in any organism [16,17]. The GO database was established based on the molecular function, biological process and cellular component [14], and hence proteins formulated in the GO database space would be clustered in a way much better reflecting their subcellular locations, as elucidated in [18]. For those proteins that cannot be meaningfully defined in the GO space, the PseAAC descriptor [15] would play a better complementary role than the classical AAC (amino acid composition) descriptor.

However, the existing Plant-PLOC [13] predictor has the following problems. **(1)** The accession number of a query protein is required as an input in order to utilize the advantage of GO approach. Many proteins, such as synthetic or hypothetical proteins, and newly discovered sequences without being deposited into databanks yet, do not have accession numbers, and hence cannot be treated with the GO approach. **(2)** Even with the accession numbers available, many proteins can still not be meaningfully formulated in a GO space because the current GO database is far from complete yet. **(3)** Although the PseAAC approach, a complementary approach to the GO approach in Plant-PLOC [13], can take into account some partial sequence order effects, the original PseAAC [15] did not contain the functional domain and sequential evolution informations, which have been proved to play an important role in enhancing the prediction quality of other protein attributes (see, e.g., [19,20]). **(4)** Plant-PLOC [13] cannot be used to deal with multiplex proteins that 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 [2,21]. Particularly, as pointed out by Millar et al. [22], recent evidence indicates that an increasing number of proteins have multiple locations in the cell.

The present study was initiated in an attempt to develop a new and more powerful predictor for predicting plant protein subcellular localization by addressing the above four problems.

## Materials and Methods

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 elaborated in [13]; the only differences are as follows. **(1)** To get the updated benchmark dataset, instead of version 49.3 of the Swiss-Prot database, the version 55.3 released on 29-Apr-2008 was adopted. **(2)** In order to make the new predictor also able to deal with proteins having two or more location sites, the multiplex proteins are no longer excluded in this study. Actually, according to a statistical analysis on the current database, about 8% of plant proteins were found located in more than one location.

After strictly following the aforementioned procedures, we finally obtained a benchmark dataset  $\mathbb{S}$  containing 978 different protein sequences, which are distributed among 12 subcellular

locations (**Fig. 1**); i.e.,

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \cup \mathbb{S}_6 \cup \dots \cup \mathbb{S}_{12} \quad (1)$$

where  $\mathbb{S}_1$  represents the subset for the subcellular location of cell membrane,  $\mathbb{S}_2$  for cell wall,  $\mathbb{S}_3$  for chloroplast, and so forth; while  $\cup$  represents the symbol for “union” in the set theory. A breakdown of the 978 plant proteins in the benchmark dataset  $\mathbb{S}$  according to their 12 location sites is given in **Table 1**. To avoid redundancy and homology bias, none of the proteins in  $\mathbb{S}$  has  $\geq 25\%$  pairwise sequence identity to any other in a same subset. The corresponding accession numbers and protein sequences are given in **Table S1**.

Since some proteins in  $\mathbb{S}$  may occur in two or more locations, it is instructive to introduce the concept of “locative protein” [23], as briefed as follows. A protein coexisting at two different location sites will be counted as 2 locative proteins even though the two are with completely the same sequence; if coexisting at three sites, 3 locative proteins; and so forth. Thus, it follows

$$N(\text{loc}) = N(\text{seq}) + \sum_{m=1}^{\Omega} (m-1)N(m) \quad (2)$$

where  $N(\text{loc})$  is the number of total locative proteins,  $N(\text{seq})$  the number of total different protein sequences,  $N(1)$  the number of proteins with one location,  $N(2)$  the number of proteins with two locations, and so forth; while  $\Omega$  is the number of total subcellular location sites concerned (for the current case,  $\Omega = 12$  as shown in **Fig. 1**).

For the current 978 different protein sequences, 904 occur in one subcellular location, 71 in two locations, 3 in three locations, and none in four or more locations. Substituting these data into Eq.2, we have

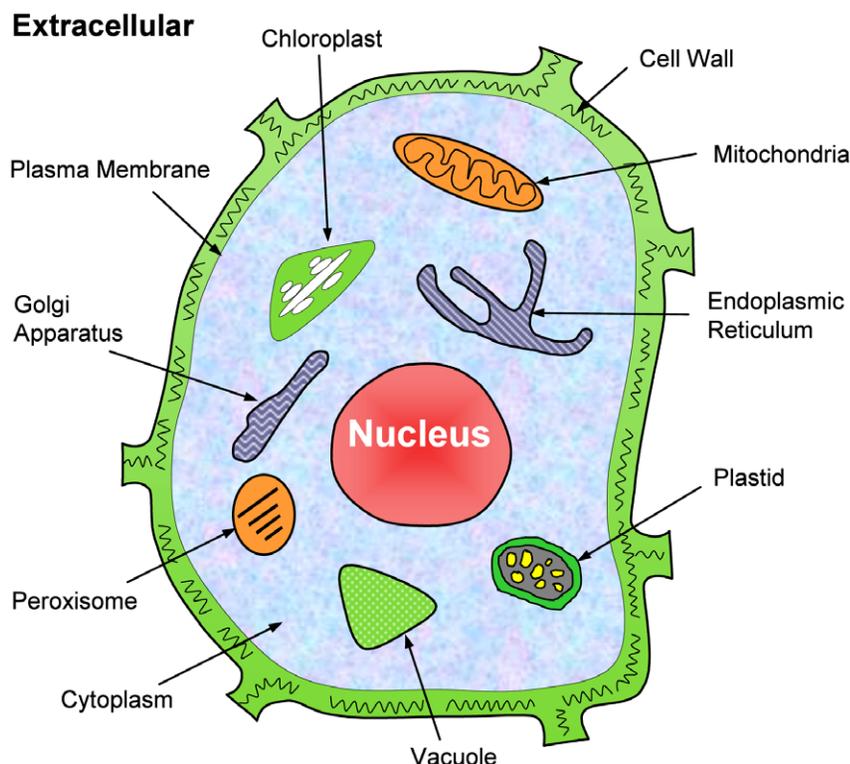
$$\begin{aligned} N(\text{loc}) &= N(\text{seq}) + (1-1) \times 904 + (2-1) \times 71 \\ &\quad + (3-1) \times 3 + \sum_{m=4}^{12} (m-1) \times 0 \\ &= 978 + 71 + 6 = 1,055 \end{aligned} \quad (3)$$

which is fully consistent with the figures in **Table 1** and the data in **Table S1**.

To develop a powerful method for predicting protein subcellular localization, it is very important to formulate the sample of a protein in terms of the core features that are intrinsically correlated with its localization in a cell. To realize this, the strategy by integrating the GO representation and PseAAC representation was adopted in the original Plant-PLOC [13]. In this study, the essence of such a strategy will be still kept. However, in order to overcome the four shortcomings as mentioned in Introduction for Plant-PLOC [13], a completely different combination approach has been developed, as described below.

## 1. Gene Ontology Descriptor

The gene ontology (GO) representation for a protein sample in the original Plant-PLOC [13] was derived through its accession number from the GO database [16]. Therefore, in using Plant-PLOC to conduct prediction, the accession number of a query protein would be indispensable as a part of input. To avoid such a requirement, the following different procedures are proposed to derive the GO representation.



**Figure 1. Schematic illustration to show the 12 subcellular locations of plant proteins.** The 12 location sites are: (1) cell membrane, (2) cell wall, (3) chloroplast, (4) cytoplasm, (5) endoplasmic reticulum, (6) extracellular, (7) Golgi apparatus, (8) mitochondrion, (9) nucleus, (10) peroxisome, (11) plastid, and (12) vacuole.

doi:10.1371/journal.pone.0011335.g001

**Step 1.** Use BLAST [24] to search the homologous proteins of the query protein  $\mathbf{P}$  from the Swiss-Prot database (version 55.3), with the BLAST parameter of expect value  $E \leq 0.001$ .

**Step 2.** Those proteins that have  $\geq 60\%$  pairwise sequence identity with the query protein  $\mathbf{P}$  are collected into a set,  $\mathbb{S}_{\mathbf{P}}^{\text{homo}}$ , called the “homology set” of  $\mathbf{P}$ . All the elements in  $\mathbb{S}_{\mathbf{P}}^{\text{homo}}$  can be deemed as the *representative proteins* of  $\mathbf{P}$ . Because these representative proteins were retrieved from the Swiss-Prot database, they must each have their own accession numbers.

**Step 3.** Search each of these accession numbers collected in Step 2 against the GO database at <http://www.ebi.ac.uk/GOA/> to find the corresponding GO numbers [16].

**Step 4.** The current GO database (version 70.0 released 10 March 2008) contains 60,020 GO numbers, thus the query protein  $\mathbf{P}$  can be formulated through its representative proteins in  $\mathbb{S}_{\mathbf{P}}^{\text{homo}}$  by the following equation

$$\mathbf{P}_{\text{GO}} = [\Delta_1^G \quad \Delta_2^G \quad \dots \quad \Delta_i^G \quad \dots \quad \Delta_{60020}^G]^T \quad (4)$$

where  $\mathbf{T}$  is the transposing operator, and

$$\Delta_i^G = \begin{cases} 1, & \text{if a hit is found against the } i\text{-th GO number} \\ & \text{for any of the proteins in } \mathbb{S}_{\mathbf{P}}^{\text{homo}} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Through the above steps, we can use Eq.4 derived from the representative proteins in  $\mathbb{S}_{\mathbf{P}}^{\text{homo}}$  to investigate the query protein  $\mathbf{P}$ . The rationale of such a practice is based on the fact that homology

proteins generally share similar attributes, such as folding patterns [25] and biological functions [26,27,28]. Thus, the accession number is no longer needed for the input of the query protein even when using the high-level GO approach to predict its subcellular localization as required in the old Plant-PLOC [13].

The above homology-based GO extraction method is particularly useful for studying those proteins which do not have UniProt accession numbers. However, it would still fail to work under any of the following situations: (1) the query protein does not have significant homology to any protein in the Swiss-Prot database, i.e.,  $\mathbb{S}_{\mathbf{P}}^{\text{homo}} = \emptyset$  meaning the homology set is an empty one; (2) its representative proteins do not contain any useful information for statistical prediction based on a given training dataset.

Therefore, it is necessary to consider the following representations for those proteins that fail to be meaningfully defined in the GO space.

## 2. Functional Domain Descriptor

The functional domain (FunD) is the core of a protein. Therefore, in determining the 3-D (dimensional) structure of a protein by experiments (see, e.g., [29,30]) or by computational modeling (see, e.g., [28,31]), the first priority was always focused on its FunD. Using FunD to formulate protein samples was originally proposed in [32,33] based on the 2005 FunDs in the SBASE-A database [34]. Since then, a series of new protein FunD databases were established, such as COG [35], KOG [35], SMART [36], Pfam [37], and CDD [38]. Of these databases, CDD contains the domains imported from COG, Pfam, and SMART, and hence is relatively much more complete [38] and will be adopted in this study. The version 2.11 of CDD contains

**Table 1.** Breakdown of the plant protein benchmark dataset  $\mathbb{S}$  derived from Swiss-Prot database (release 55.3) according to the procedures described in the Materials section.

Subset	Subcellular location <sup>a</sup>	Number of proteins
$\mathbb{S}_1$	Cell membrane	56
$\mathbb{S}_2$	Cell wall	32
$\mathbb{S}_3$	Chloroplast	286
$\mathbb{S}_4$	Cytoplasm	182
$\mathbb{S}_5$	Endoplasmic reticulum	42
$\mathbb{S}_6$	Extracellular	22
$\mathbb{S}_7$	Golgi apparatus	21
$\mathbb{S}_8$	Mitochondrion	150
$\mathbb{S}_9$	Nucleus	152
$\mathbb{S}_{10}$	Peroxisome	21
$\mathbb{S}_{11}$	Plastid	39
$\mathbb{S}_{12}$	Vacuole	52
Total number of locative proteins $N(\text{loc})$		1,055 <sup>b</sup>
Total number of different proteins $N(\text{seq})$		978 <sup>c</sup>

None of proteins included here has  $\geq 25\%$  sequence identity to any other in a same subcellular location.

<sup>a</sup>The benchmark dataset  $\mathbb{S}$  here covers 12 plant subcellular locations and the "Golgi apparatus" is newly added in comparison with the dataset in [13] that covered 11 location sites.

<sup>b</sup>See Eqs.2–3 for the definition about the number of locative proteins, and its relation with the number of different proteins.

<sup>c</sup>Of the 978 different proteins, 904 have one subcellular location, 71 have two locations, 3 have three locations, and none have four or more locations.

doi:10.1371/journal.pone.0011335.t001

17,402 characteristic domains. Thus, using each of these domains as a base vector, a given protein sample can be defined as a vector in the 17402-D (dimensional) FunD space according to the following procedures:

**Step 1.** Use RPS-BLAST (Reverse PSI-BLAST) program [24] to conduct sequence alignment of the sequence of the query protein  $\mathbf{P}$  with each of the 17,402 domain sequences in the CDD database.

**Step 2.** If the significance threshold value (expect value) is  $\leq 0.001$  for the  $i$ -th domain meaning that a "hit" is found, then the  $i$ -th component of the protein  $\mathbf{P}$  in the 17402-D space is assigned 1; otherwise, 0.

**Step 3.** The protein sample  $\mathbf{P}$  in the FunD space can thus be formulated as

$$\mathbf{P}_{\text{FunD}} = [\Delta_1^D \quad \Delta_2^D \quad \cdots \quad \Delta_i^D \quad \cdots \quad \Delta_{17402}^D]^T \quad (6)$$

where  $\mathbf{T}$  has the same meaning as in Eq.4, and

$$\Delta_i^D = \begin{cases} 1, & \text{when a hit is found for } \mathbf{P} \text{ in CDD} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

### 3. SeqEvo (Sequential Evolution) Descriptor

Biology is a natural science with historic dimension. All biological species have developed continuously starting out from a very limited number of ancestral species. The evolution in protein sequences involves changes of single residues, insertions and deletions of several residues [39], gene doubling, and gene

fusion. In the course of time such changes accumulate, so that many similarities between initial and resultant amino acid sequences are eliminated, but the corresponding proteins may still share many common attributes, such as belonging to a same subcellular location and possessing basically the same function. To incorporate this kind of evolutionary effects, let us use the "Position-Specific Scoring Matrix" or "PSSM" [24] to express the protein sample  $\mathbf{P}$ , as formulated by

$$\mathbf{P}_{\text{Evo}} = \begin{bmatrix} V_{1 \rightarrow 1} & V_{1 \rightarrow 2} & \cdots & V_{1 \rightarrow 20} \\ V_{2 \rightarrow 1} & V_{2 \rightarrow 2} & \cdots & V_{2 \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ V_{i \rightarrow 1} & V_{i \rightarrow 2} & \cdots & V_{i \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ V_{L \rightarrow 1} & V_{L \rightarrow 2} & \cdots & V_{L \rightarrow 20} \end{bmatrix} \quad (8)$$

where  $V_{i \rightarrow j}$  represents the score of the amino acid residue in the  $i$ -th position of the protein sequence being changed to amino acid type  $j$  during the evolutionary process, and  $L$  the sequence length of protein  $\mathbf{P}$ . Here, the numerical codes 1, 2, ..., 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes. The  $L \times 20$  scores in Eq.8 were generated by using PSI-BLAST [24] to search the Swiss-Prot database (version 55.3 released on 29-Apr-2008) through three iterations with 0.001 as the  $E$ -value cutoff for multiple sequence alignment against the sequence of the protein  $\mathbf{P}$ , followed by a standard conversion given below:

$$V_{i \rightarrow j} = \frac{V_{i \rightarrow j}^0 - \langle V_i^0 \rangle}{\text{SD}(V_i^0)} \quad (i=1, 2, \dots, L; j=1, 2, \dots, 20) \quad (9)$$

where  $V_{i \rightarrow j}^0$  represent the original scores directly created by PSI-BLAST [24] that are generally shown as positive or negative integers (the positive score means that the corresponding mutation occurs more frequently than expected by chance, while the negative means just the opposite); the symbol  $\langle \rangle$  means taking the average of the quantity therein over 20 native amino acids, and  $\text{SD}$  means the corresponding standard deviation. The converted values obtained by Eq.9 will have a zero mean value over the 20 amino acids and will remain unchanged if going through the same conversion procedure again. However, according to the descriptor of Eq.8, proteins with different lengths will correspond to row-different matrices causing difficulty for developing a predictor able to uniformly cover proteins of any length. To make the descriptor become a size-uniform matrix, one possible avenue is to represent a protein sample  $\mathbf{P}$  by

$$\bar{\mathbf{P}}_{\text{Evo}} = [\bar{V}_1 \quad \bar{V}_2 \quad \cdots \quad \bar{V}_{20}]^T \quad (10)$$

where

$$\bar{V}_j = \frac{1}{L} \sum_{i=1}^L V_{i \rightarrow j} \quad (j=1, 2, \dots, 20) \quad (11)$$

where  $\bar{V}_j$  represents the average score of the amino acid residues in the protein  $\mathbf{P}$  being changed to amino acid type  $j$  during the evolutionary process. However, if  $\bar{\mathbf{P}}_{\text{Evo}}$  of Eq.10 was used to represent the protein  $\mathbf{P}$ , all the sequence-order information during the evolutionary process would be missed. To avoid complete loss

of the sequence-order information, the concept of the pseudo amino acid composition (PseAAC) as originally proposed in [15] was utilized; i.e., instead of Eq.10, let us use the pseudo position-specific scoring matrix as given by

$$\mathbf{P}_{\text{PseEvo}}^{\lambda} = [\bar{\mathbf{V}}_1 \quad \bar{\mathbf{V}}_2 \quad \cdots \quad \bar{\mathbf{V}}_{20} \quad \mathbf{V}_1^{\lambda} \quad \mathbf{V}_2^{\lambda} \quad \cdots \quad \mathbf{V}_{20}^{\lambda}]^T \quad (12)$$

to represent the protein  $\mathbf{P}$ , where

$$\mathbf{V}_j^{\lambda} = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} [\mathbf{V}_{i \rightarrow j}^{\lambda} - \mathbf{V}_{(i+\lambda) \rightarrow j}^{\lambda}]^2 \quad (j=1, 2, \dots, 20; \lambda < L) \quad (13)$$

meaning that  $\mathbf{V}_j^1$  is the correlation factor by coupling the most contiguous position-specific scoring matrix scores along the protein chain for the amino acid type  $j$ ;  $\mathbf{V}_j^2$  that by coupling the second-most contiguous position-specific scoring matrix scores; and so forth. Note that, as mentioned in the Material section of [13], the length of the shortest protein sequence in the benchmark dataset is  $L = 50$ , and hence the value allowed for  $\lambda$  in Eq.13 must be smaller than 50. When  $\lambda=0$ ,  $\mathbf{V}_j^{\lambda}$  becomes a naught element and Eq.12 is degenerated to Eq.10.

It is instructive to point out that the above three protein descriptors, i.e.,  $\mathbf{P}_{\text{GO}}$  of Eq.4,  $\mathbf{P}_{\text{FunD}}$  of Eq.6, and  $\mathbf{P}_{\text{PseEvo}}^{\lambda}$  of Eq.12, can be actually deemed as three different kinds of PseAAC as well [40]. This is because, according to its original definition, the PseAAC is actually a set of discrete numbers [15] as long as it is different from the classical amino acid composition and it is derived from a protein sequence that is able to harbor some sort of sequence order or pattern information. The concept of PseAAC has also been widely used to deal with many other protein-related problems and sequence-related systems (see, e.g., [41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56]).

#### 4. Prediction Engine and Process

The prediction engine used in this study is the ensemble classifier  $\mathbb{C}^E$  [12] formed by fusing many basic individual classifiers operated according to the OET-KNN (Optimized Evidence-Theoretic K Nearest Neighbor) rule [57,58]. OET-KNN is a very powerful classifier as demonstrated in identifying membrane protein types [58]. For reader's convenience, a brief introduction about OET-KNN is given below.

Let us consider a problem of classifying  $N$  plant protein entities into 12 categories (subcellular location sites). The problem can be formulated as

$$\mathbb{L} = \{\mathbb{L}_1, \mathbb{L}_2, \mathbb{L}_3, \dots, \mathbb{L}_{11}, \mathbb{L}_{12}\} \quad (14)$$

The available information is assumed to consist in a training dataset

$$\mathbb{N} = \{(\mathbf{P}_1, u_1), \dots, (\mathbf{P}_N, u_N)\} \quad (15)$$

where the  $N$  plant proteins  $\mathbf{P}_i$  ( $i=1, 2, \dots, N$ ) and their corresponding location labels  $u_i$  ( $i=1, 2, \dots, N$ ) take the values in  $\mathbb{L}$  of Eq.14. According to the KNN ( $k$ -Nearest Neighbors) rule [59], an unclassified protein  $\mathbf{P}$  is assigned to the class (or location) represented by the majority of its  $k$  nearest neighbors of  $\mathbf{P}$ . Owing to its good performance and simple-to-use feature, the KNN rule, also named as "voting KNN rule", is quite popular in pattern recognition community.

The ET-KNN (Evidence Theoretic  $k$ -nearest Neighbors) rule is a pattern classification method based on the Dempster-Shafer theory of belief functions [57]. In the classification process, each neighbor of a protein to be classified is considered as a piece of evidence supporting certain hypotheses concerning the class (or location) membership of that protein. Based on this kind of evidence, the basic belief masses are assigned to each subset concerned. Such masses are obtained for each of the  $k$  nearest neighbors of the protein under consideration and aggregated using the Dempster's rule of combination [60]. A decision is made by assigning the query protein to the class (or location) with the maximum credibility.

Suppose  $\mathbf{P}$  is a query protein to be classified, and  $\mathcal{S}_K^{\mathbf{P}}$  is the set of its  $k$ -nearest neighbors in the training dataset  $\mathbb{N}$  of Eq.15. Thus, for any  $\mathbf{P}_i \in \mathcal{S}_K^{\mathbf{P}}$ , the knowledge that  $\mathbf{P}_i$  belongs to class (or location)  $\mathbb{L}_{\mu} \in \mathbb{L}$  can be considered as a piece of evidence that increases our belief that  $\mathbf{P}$  also belongs to  $\mathbb{L}_{\mu}$ . According to the basic belief assignment mapping theory [60], this piece of evidence can be formulated by

$$\mathbb{B}(\mathbf{P}_i, \mathbb{L}_{\mu}) = \alpha_0 \exp[-\gamma_{\mu}^2 D^2(\mathbf{P}_i, \mathbf{P})] \quad (16)$$

where  $\alpha_0$  is a fixed parameter,  $\gamma_{\mu}$  is a parameter associated with class (or location)  $\mathbb{L}_{\mu}$ , and  $D^2(\mathbf{P}_i, \mathbf{P})$  is the square distance between  $\mathbf{P}$  and  $\mathbf{P}_i$ . In this study, when the proteins are represented by the GO descriptor mode (cf. Eq.4) or the FunD mode (cf. Eq.6), then  $D(\mathbf{P}_i, \mathbf{P})$  is defined as  $1 - \cos(\mathbf{P}_i, \mathbf{P})$ , i.e.

$$D(\mathbf{P}_i, \mathbf{P}) = 1 - \frac{\mathbf{P}_i \cdot \mathbf{P}}{\|\mathbf{P}_i\| \|\mathbf{P}\|} \quad (17)$$

where  $\|\mathbf{P}_i\|$  and  $\|\mathbf{P}\|$  are the modulus of  $\mathbf{P}_i$  and  $\mathbf{P}$ , respectively. When the proteins are represented by the PseEvo model (cf. Eq.12), then  $D(\mathbf{P}_i, \mathbf{P})$  is defined as the Euclidean distance between  $\mathbf{P}$  and  $\mathbf{P}_i$ .

In the ET-KNN rule, it was not addressed how to optimally select the parameters. In 1998 an optimization procedure to determine the optimal or near-optimal parameter values of  $\alpha_0$  and  $\gamma_{\mu}$  was proposed from the data by minimizing an error function [61]. It was observed that the OET-KNN rule obtained thru such an optimization treatment would lead to a substantial improvement in classification accuracy.

The belief function of  $\mathbf{P}$  belonging to class (or location)  $\mathbb{L}_{\mu}$  is a combination of its  $k$ -Nearest Neighbors, and can be formulated as

$$\mathbb{B}(\mathbf{P}, \mathbb{L}_{\mu}, K) = (\cdots ((\mathbb{B}(\mathbf{P}_1, \mathbb{L}_{\mu}) \oplus \mathbb{B}(\mathbf{P}_2, \mathbb{L}_{\mu})) \oplus \mathbb{B}(\mathbf{P}_3, \mathbb{L}_{\mu})) \oplus \cdots) \oplus \mathbb{B}(\mathbf{P}_K, \mathbb{L}_{\mu}) \quad (18)$$

where  $\oplus$  is called the orthogonal sum, which is commutative and associative. According to Dempster's rule [60], the belief function of Eq.18 can be expressed as

$$\mathbb{B}(\mathbf{P}, \mathbb{L}_{\mu}, K) = \frac{\sum_{\mathcal{S}_{K,i}^{\mathbf{P}} \subseteq \mathcal{S}_K^{\mathbf{P}}, \mathcal{S}_{K,j}^{\mathbf{P}} \subseteq \mathcal{S}_K^{\mathbf{P}}, \mathcal{S}_{K,i}^{\mathbf{P}} \cap \mathcal{S}_{K,j}^{\mathbf{P}} = \mathbb{L}_{\mu}} \mathbb{B}(\mathbf{P}, \mathcal{S}_{K,i}^{\mathbf{P}}) \mathbb{B}(\mathbf{P}, \mathcal{S}_{K,j}^{\mathbf{P}})}{1 - \sum_{\mathcal{S}_{K,i}^{\mathbf{P}} \subseteq \mathcal{S}_K^{\mathbf{P}}, \mathcal{S}_{K,j}^{\mathbf{P}} \subseteq \mathcal{S}_K^{\mathbf{P}}, \mathcal{S}_{K,i}^{\mathbf{P}} \cap \mathcal{S}_{K,j}^{\mathbf{P}} = \emptyset} \mathbb{B}(\mathbf{P}, \mathcal{S}_{K,i}^{\mathbf{P}}) \mathbb{B}(\mathbf{P}, \mathcal{S}_{K,j}^{\mathbf{P}})} \quad (19)$$

where  $\mathcal{S}_{K,i}^{\mathbf{P}}$  is the  $i$ -th possible subset of  $\mathcal{S}_K^{\mathbf{P}}$ , and  $\subseteq$ ,  $\cap$ , and  $\emptyset$  are

the symbols in set theory, representing “contained in”, “intersection”, and the empty set, respectively.

A decision is made by assigning the query protein  $\mathbf{P}$  to the  $\mu$ -th ( $\mu=1,2,3, \dots, 11,12$ ) class (or location) with which the belief function of Eq.19 has the maximum value; i.e.,

$$\mu = \arg \max_v \{B(\mathbf{P}, \mathbb{L}_v, K)\}, \quad (v=1, 2, \dots, 11,12) \quad (20)$$

where  $\mu$  is the argument of  $v$  that maximizes the belief function  $B(\mathbf{P}, \mathbb{L}_v, K)$ . If there are two and more arguments leading to a same maximum value for  $B(\mathbf{P}, \mathbb{L}_v, K)$ , the query protein will be randomly assigned to one of the subcellular locations associated with these arguments although this kind of tie case rarely happens.

The power of the ensemble classifier  $\mathbb{C}^E$  is also reflected by the fact that a statistical predictor established by fusing many basic individual predictors will significantly improve its performance as demonstrated by the recent studies on protein folding rate predictions [62,63]. For the detailed procedures of how to fuse many individual OET-KNN classifiers to form the ensemble classifier  $\mathbb{C}^E$ , see Eqs.30–35 in [12]. For the procedures of how to make  $\mathbb{C}^E$  able to deal with both single-location and multiple-location proteins, see Eqs.36–48 of [12].

The prediction is processed according to the following order.

(1) If the query protein can be expressed as a meaningful or productive descriptor in the GO database via its representative proteins in  $\mathbb{S}_P^{\text{homo}}$ , then  $\mathbf{P}_{\text{GO}}$  of Eq.4 should be input into the prediction engine for identifying its subcellular location site(s). And the output will be determined by fusing many basic OET-KNN predictors [12] with different numbers of  $K$  (cf. Eq.18–20), the parameter of the nearest neighbor rule [57].

(2) If the query protein does not have significant homology to any protein in the Swiss-Prot database, i.e.,  $\mathbb{S}_P^{\text{homo}} = \emptyset$ , or its representative proteins in  $\mathbb{S}_P^{\text{homo}}$  do not contain any useful GO

information, then both the FunD representation  $\mathbf{P}_{\text{FunD}}$  of Eq.6 and the pseudo position-specific scoring matrix representation  $\mathbf{P}_{\text{PseEvo}}^{\lambda}$  of Eq.12 should be input into the prediction engine. The output will be determined by fusing many basic OET-KNN predictors [12] with different numbers of  $K$  (cf. Eq.20) and  $\lambda$  (cf. Eq.13).

The whole process can be formulated as

$$\mathbb{C}^E \triangleright \mathbf{P} = \begin{cases} \mathbb{C}^E \triangleright \mathbf{P}_{\text{GO}}, & \text{if Eq.4 contains useful information} \\ \mathbb{C}^E \triangleright \mathbf{P}_{\text{FunD}} \\ \mathbb{C}^E \triangleright \mathbf{P}_{\text{PseEvo}} \end{cases} \Rightarrow, \quad \text{otherwise} \quad (21)$$

where  $\triangleright$  represents the identification operator, and  $\} \Rightarrow$  means fusing the results generated from its left side.

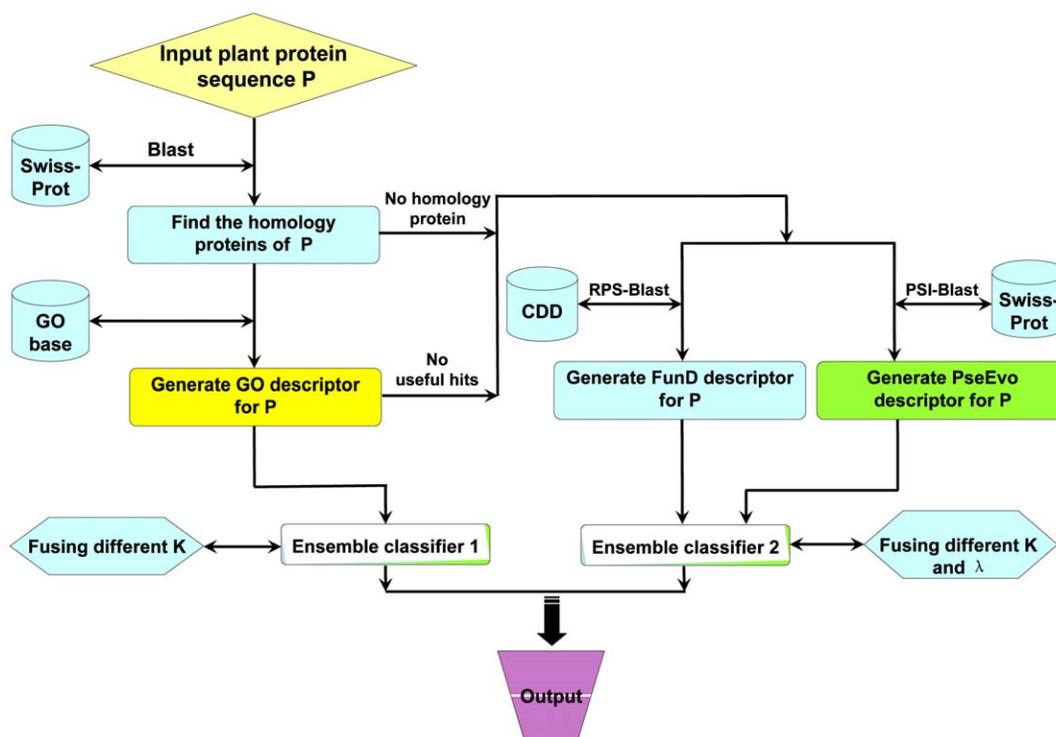
The entire ensemble classifier thus established is called “**Plant-mPLoc**”, where “m” stands for the first character of “multiple”, meaning that Plant-mPLoc is able to deal with proteins having both single and multiple subcellular locations. To provide an intuitive picture, a flowchart is given in **Fig. 2** to illustrate the prediction process of Plant-mPLoc.

### Protocol Guide

For the convenience of experimental scientists, a user-friendly web-server for Plant-mPLoc was established. Here let us provide a step-by-step guide on how to use the web-server to get the desired results.

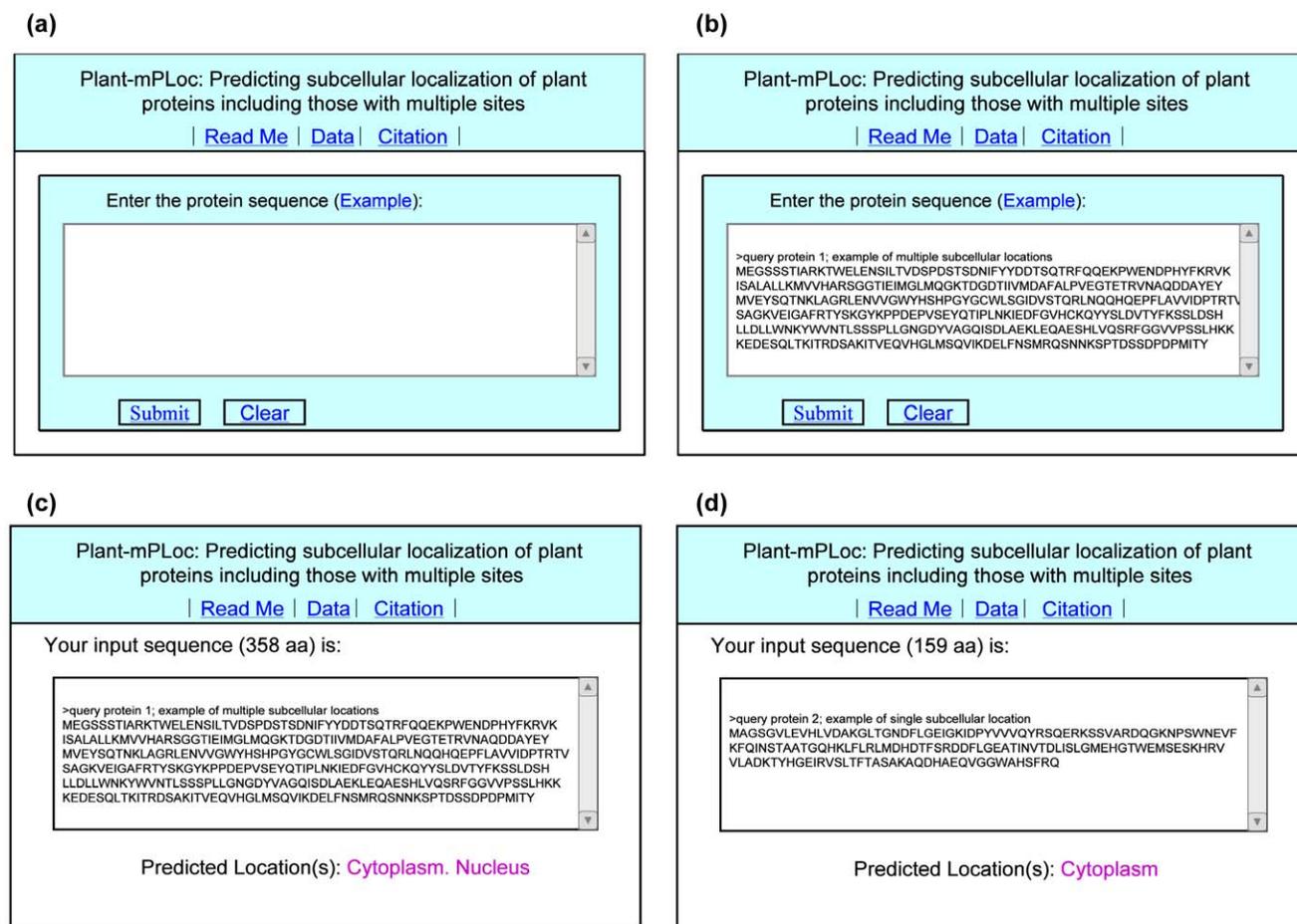
**Step 1.** Open the web server at <http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/> and you will see the top page of the predictor on your computer screen, as shown in **Fig. 3a**. Click on the **Read Me** button to see a brief introduction about Plant-mPLoc predictor and the caveat in using it.

**Step 2.** Either type or copy and paste the query protein sequence into the input box at the center of **Fig. 3a**. The input



**Figure 2. A flowchart to show the prediction process of Plant-mPLoc.**

doi:10.1371/journal.pone.0011335.g002



**Figure 3. Semi-screenshot to show the prediction steps.** (a) the top page of the Plant-mPLOC web server at <http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/>, (b) the input of a query protein in FASTA format, (c) the output predicted by Plant-mPLOC for the query protein 1 in the Example window, and (d) the output for the query protein 2 in the Example window.  
doi:10.1371/journal.pone.0011335.g003

sequence should be in the FASTA format. A sequence in FASTA format consists of a single-line description, followed by lines of sequence data. The first character of the description line is a greater-than symbol (“>”) in the first column. All lines should be shorter than 80 characters. Examples to show the input sequences format can be seen by clicking on the Example button right above the input box. For more information about FASTA format, visit [http://en.wikipedia.org/wiki/Fasta\\_format](http://en.wikipedia.org/wiki/Fasta_format).

**Step 3.** Click on the Submit button to see the predicted result. For example, if you use the sequence of query protein 1 in the Example window, the input screen should look like the illustration in **Fig. 3b**; after clicking the Submit button, you will see “Cytoplasm. Nucleus” shown on the predicted result window (**Fig. 3c**), meaning that the protein is a multiplex one, which can simultaneously occur in “cytoplasm” organelle and “nucleus” organelle, fully consistent with experimental observations. However, if using the sequence of query protein 2 in the Example window as an input, you will see “Cytoplasm” shown on the predicted result window (**Fig. 3d**), meaning that the protein is a single-location one occurring in “cytoplasm” compartment only, also consistent with experimental observations. It takes less than 15 seconds for a protein sequence of 300 amino acids before the predicted result appears on your computer screen. Generally speaking, the longer the sequence is, the more time it is needed.

**Step 4.** Click on the Citation button to find the relevant papers that document the detailed development and algorithm of Plant-mPLOC.

**Step 5.** Click on the Data button to download the benchmark datasets used to train and test the Plant-mPLOC predictor.

**Caveat.** To obtain the predicted result with the expected success rate, the entire sequence of the query protein rather than its fragment should be used as an input. A sequence with less than 50 amino acid residues is generally deemed as a fragment

## Results and Discussion

In statistical prediction, the following three methods are often used to examine the quality of a predictor: independent dataset test, subsampling test, and jackknife test [64]. Since independent dataset can be treated as a special case of sub-sampling test, one benchmark dataset is sufficient to serve all the three kinds of cross-validation. However, as elucidated in [18] and demonstrated by Eq.50 of [12], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset and hence has been increasingly and widely adopted to examine the power of various predictors (see, e.g., [42,46,51,53,55,65,66,67,68,69]). Particularly for a benchmark dataset in which none of proteins included has  $\geq 25\%$  pairwise sequence identity to any other in a same subset

(subcellular location), such as the one used in the current study (cf. **Table S1**), it would be highly unlikely to get an over-estimated success rate by the jackknife test. Quite the contrary, the success rate derived by the jackknife test on such kind of stringent dataset would actually be under-estimated in comparison with the success rates of practical applications in most cases, as will be seen later.

For the details of how to calculate the overall success rate for a statistical system with both single-location and multiple-location proteins, see Eqs.43–48 and Fig. 4 of [12], where the details of how to count the false positives (over-predictions) and false negatives (under-predictions) were also elaborated.

Let us first compare the current predictor Plant-mPLOC with the old Plant-PLOC [13]. Listed in **Table 2** are the results obtained with Plant-PLOC [13] and Plant-mPLOC, respectively, on the benchmark dataset (cf. **Table S1**) by the jackknife cross-validation test. During the testing process, only the sequences of proteins in **Table S1** but not their accession numbers were used as inputs in order to make the comparison between the two predictors under exactly the same condition. As we can see from **Table 2**, for such a stringent benchmark dataset, the overall success rate achieved by the new predictor is 63.7%, which is more than 25% higher than that by Plant-PLOC [13].

Now, let us compare the current predictor with TargetP [6] and Predotar [8], two popular predictors widely used for predicting the subcellular locations of plant proteins. As mentioned in Introduction, the two predictors only cover three or four location sites. Therefore, it can be easily conceived that they would yield even much lower success rates when tested by the current benchmark dataset that covers twelve location sites.

Actually, even if tested by a benchmark dataset within the scope that can be covered by TargetP [6] or Predotar [8], the success rate by the current Plant-mPLOC predictor is also much higher than those by the two predictors, as demonstrated below.

**Table 2.** A comparison of the jackknife success rates by Plant-PLOC [13] and the current Plant-mPLOC on the benchmark dataset (cf. **Table S1**) that covers 12 location sites of plant proteins in which none of the proteins included has  $\geq 25\%$  pairwise sequence identity to any other in a same location.

Subcellular location	Success rate <sup>a</sup>	
	Plant-PLOC	Plant-mPLOC
Cell membrane	15/56 = 26.8%	24/56 = 42.9%
Cell wall	7/32 = 21.9%	8/32 = 25.0%
Chloroplast	184/286 = 64.3%	248/286 = 86.7%
Cytoplasm	51/182 = 28.0%	72/182 = 39.6%
Endoplasmic reticulum	1/42 = 2.4%	17/42 = 40.5%
Extracellular	4/22 = 18.2%	3/22 = 13.6%
Golgi apparatus	6/21 = 28.6%	6/21 = 28.6%
Mitochondrion	26/150 = 17.3%	114/150 = 76.0%
Nucleus	92/152 = 60.5%	136/152 = 89.5%
Peroxisome	2/21 = 9.5%	14/21 = 66.7%
Plastid	9/39 = 23.1%	4/39 = 10.3%
Vacuole	4/52 = 7.7%	26/52 = 50.0%
Total	401/1055 = 38.0%	672/1055 = 63.7%

<sup>a</sup>Note that in order to make the comparison under exactly the same condition, only the sequences of proteins in the **Table S1** but not their accession numbers were used as inputs during the prediction.

doi:10.1371/journal.pone.0011335.t002

Let us compare Plant-mPLOC with TargetP [6] first. The TargetP predictor also has a web-server at <http://www.cbs.dtu.dk/services/TargetP/>, with a built-in training dataset covering the following four items: “mitochondria”, “chloroplast”, “secretory pathway”, and “other”. Since the “secretory pathway” is not a final destination of subcellular location as annotated in Swiss-Prot databank, and hence was removed from the comparison. Also, the location of “other” is not a clear site for comparison, and should be removed as well. Thus, in order to compare TargetP with the new predictor Plant-mPLOC, let us construct an independent testing dataset by randomly picking testing proteins according to the following criteria: (i) they must belong to plant proteins, as clearly annotated in Swiss-Prot database; (ii) they must neither occur in the training dataset of TargetP nor occur in the training dataset of Plant-mPLOC in order to avoid the memory bias; (iii) their experimentally observed subcellular locations are known as clearly annotated in Swiss-Prot database, and also these locations must be within the scope covered by TargetP as a compromise for rationally utilizing its web-server. By following the above procedures, we obtained a degenerate independent testing dataset consisting of 1,775 plant proteins, of which 1,500 belong to chloroplast and 275 belong to mitochondrion. The accession numbers and sequences of these 1,775 proteins are given in **Table S2**.

The predicted results by TargetP [6] and the current Plant-mPLOC for each of the 1,775 independent testing proteins are listed in **Table S3**, where for facilitating comparison, the corresponding experimental results are also given. By examining **Table S3**, we can see the following. (1) Many proteins whose subcellular locations were misidentified by TargetP have been corrected by Plant-mPLOC. (2) Many proteins, which were identified by TargetP as belonging to the location of “other”, have been identified as “chloroplast” or “mitochondrion”, fully consistent with experimental observations. (3) There are quite a few proteins whose subcellular locations were incorrectly predicted by Plant-mPLOC, or the results yielded by Plant-mPLOC contain some false positives. Even though, the overall success rate by Plant-mPLOC on the 1,755 independent proteins is over 86%, which is at least more than 40% higher than that by TargetP [6].

Now, let us compare Plant-mPLOC with Predotar [8]. The web-server of Predotar is at: <http://urgi.versailles.inra.fr/predotar/predotar.html>, with a built-in training dataset covering the following four items: “endoplasmic reticulum”, “mitochondrion”, “plastid”, and “other”. Since the term “other” is not a clear description for subcellular location, and was removed from comparison. Thus, by following the aforementioned similar criteria as in constructing the independent dataset for comparing TargetP with Plant-mPLOC, we also constructed a degenerate independent dataset to compare Predotar [8] with Plant-mPLOC. The dataset consists of 381 plant proteins, of which 48 belong to endoplasmic reticulum, 253 belong to mitochondrion, and 70 belong to plastid. The accession numbers and sequences of these 381 proteins are given in **Table S4**. The predicted results by Predotar [8] and the current Plant-mPLOC for the 381 independent testing proteins and their corresponding experimental results are listed in **Table S5**, from which we can see the following. (1) Many proteins whose subcellular locations were correctly identified by Plant-mPLOC were unable to identify by Predotar [8] although all these location sites are within its coverage scope. (2) Many proteins whose subcellular locations were misidentified by Predotar [8] have been corrected by Plant-mPLOC. (3) Although Plant-mPLOC also had quite a few incorrect and false positive predicted results, its overall success rate for the 381 independent proteins could still be over 70%, which is at least more than 30% higher than that by Predotar [8].

Furthermore, it is interesting to see from **Table S3** and **Table S5** that some proteins with multiple locations have been correctly predicted by Plant-mPLoc. For example, according to the annotations of Swiss-Prot databank, the proteins with codes Q5YLB5, Q9FV51, and Q9LJL3 can coexist in both “chloroplast” and “mitochondrion” while the protein with code Q42560 can coexist in both “cytoplasm” and “mitochondrion”, and the predicted results by Plant-mPLoc are exactly so. This is beyond the reach of TargetP [6] and Predotar [8].

From the above three comparisons, we can now make the following points more clear.

The more stringent a benchmark dataset is in excluding homologous and high similarity sequences, or the more subcellular location sites it covers, the more difficult for a predictor to achieve a high overall success rate, as can be easily understood by considering the following cases. For a benchmark dataset only covering three subcellular locations each containing same number of proteins, the overall success rate by random assignments would generally be  $1/3 \approx 33.3\%$ ; while for a benchmark dataset covering 12 subcellular locations, the overall success rate by random assignments would be only  $1/12 \approx 8.3\%$ . This means that the former is more than four times the latter.

Also, a predictor tested by jackknife cross-validation is very difficult to yield a high success rate when performed on a stringent benchmark dataset in which none of proteins included has  $\geq 25\%$  pairwise sequence identity to any other in a same subset (subcellular location). That is why the overall success rate achieved by Plant-mPLoc was only 63.7% when tested by the jackknife cross-validation on the benchmark dataset of **Table S1** but was over 86% and 70% when tested by the independent datasets of **Table S2** and **Table S4**, respectively. However, regardless of using what test methods or test datasets, one thing is crystal clear, i.e., the overall success rates achieved by the current Plant-mPLoc are significantly higher than those by its counterparts.

Meanwhile, it has also become understandable why the success rates as originally reported for TargetP [6] and Predotar [8] were over-estimated. This is because the benchmark datasets adopted by the two predictors only cover less than one-third of the location sites that are covered by the current Plant-mPLoc. Besides, the benchmark datasets used by TargetP and Predotar to estimate their success rates contain many homologous sequences. For the benchmark dataset used by Predotar [8], the cutoff threshold was set at 80%, meaning that only those sequences which have  $\geq 80\%$  pairwise sequence identity to any other in a same subset were excluded [8]; while for the benchmark dataset used in TargetP [6], even no such a cutoff percentage was indicated. Compared with the current benchmark dataset (cf. **Table S1**) in which none of proteins included has  $\geq 25\%$  pairwise sequence identity to any other in a same subset, the benchmark datasets adopted in Predotar and TargetP are much less stringent and hence cannot avoid homologous bias and over estimation.

Plant-mPLoc was evolved from Plant-PLoc [13] through a top-down approach improvement. The new predictor distinguishes itself from the old one by the following remarkable features. **(1)** The ability of prediction is extended to cover both single-location and multiple-location proteins. **(2)** The input of accession number for using the higher-level GO approach [18] to perform the prediction is no longer needed; this is particularly useful when dealing with protein sequences without accession numbers available. **(3)** For those plant proteins without useful GO information to conduct the higher-level prediction, a sophisticated combination approach by fusing the FunD information and SeqEvo information is developed to replace the simple PseAAC approach [15].

It is instructive to point out that in a broader sense the protein descriptors,  $\mathbf{P}_{GO}$ ,  $\mathbf{P}_{FunD}$ , and  $\mathbf{P}_{PseEvo}^2$  as introduced in the current study, are actually three different forms of PseAAC [40]. Accordingly, it is essentially through the concept of PseAAC [15] that the GO information, FunD information, and SeqEvo information have been effectively incorporated into the predictor Plant-mPLoc. Plant-mPLoc is available as a web-server at <http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/>.

Finally, let us consider the following hypothetical case: a single amino acid mutation in the signal part of a protein sequence might lead it to a completely different subcellular location site. Can Plant-mPLoc be used to deal with such a subtle case? Like all existing predictors in this area, Plant-mPLoc is a statistical predictor. As a statistical predictor, it would generally not be so sensitive to reflect the change of only one amino acid. Nevertheless, since Plant-mPLoc is an ensemble classifier formed by fusing many basic individual classifiers as well as by incorporating functional domain and evolution informations, it would be relatively more competent in dealing with the cases of mutated sequences than those predictors based on single classifier alone. Of course, it remains a challenging problem how to incorporate into a statistical predictor with the subtle effect of a single amino acid mutation at the signal peptide of a protein.

## Supporting Information

**Table S1** This benchmark dataset S for Plant-mPLoc includes 1,055 plant protein sequences (978 different proteins), classified into 12 plant subcellular locations. Among the 978 different proteins, 904 belong to one subcellular location, 71 to two locations, and 3 to three locations. Both the accession numbers and sequences are given. None of the proteins has  $\geq 25\%$  sequence identity to any other in the same subset (subcellular location). See the text of the paper for further explanation.

Found at: doi:10.1371/journal.pone.0011335.s001 (0.78 MB PDF)

**Table S2** The degenerate testing dataset used for comparing the performance between TargetP (Emanuelsson, et al. *J. Mol. Biol.* 2000, 300: 1005–1016) and Plant-mPLoc of this paper. The dataset contains 1,775 plant proteins classified into 2 subcellular locations: (1) chloroplast, and (2) mitochondrion. To avoid bias, none of the proteins included here occurs in the training dataset of TargetP, nor in the training dataset of Plant-mPLoc. See the text of the paper for further explanation.

Found at: doi:10.1371/journal.pone.0011335.s002 (0.91 MB PDF)

**Table S3** List of the results predicted by TargetP (Emanuelsson et al. *J. Mol. Biol.* 2000, 300: 1005–1016) and Plant-mPLoc on the 1,775 independent proteins in the Table S2, and their experimental subcellular locations as annotated in Swiss-Prot databank (version 55.3 released on 29-Apr-2008). Note for TargetP outputs, “C” means “Chloroplast”, “M” means “Mitochondrion”, “S” means “Secretory pathway”, and “\_” means “Any other location”.

Found at: doi:10.1371/journal.pone.0011335.s003 (0.41 MB PDF)

**Table S4** The degenerate testing dataset used for comparing the performance between Predotar (Small et al., *Proteomics* 2004, 4: 1581–1590) and Plant-mPLoc of this paper. The dataset contains 381 plant proteins classified into 3 subcellular locations: (1) endoplasmic reticulum, (2) mitochondrion, and (3) plastid. To avoid bias, none of the proteins included here occurs in the

training dataset of TargetP, nor in the training dataset of Plant-mPLOC. See the text of the paper for further explanation.

Found at: doi:10.1371/journal.pone.0011335.s004 (0.25 MB PDF)

**Table S5** List of the results predicted by Predotar (Small et al., *Proteomics* 2004, 4:1581–90) and Plant-mPLOC on the 381 independent proteins in the Table S4, and their experimental subcellular locations as annotated in Swiss-Prot databank (version 55.3 released on 29-Apr-2008). Note for the Predotar output, “ER” means “Endoplasmic reticulum”.

Found at: doi:10.1371/journal.pone.0011335.s005 (0.16 MB PDF)

## References

- Ehrlich JS, Hansen MD, Nelson WJ (2002) Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell adhesion. *Dev Cell* 3: 259–270.
- Glory E, Murphy RF (2007) Automated subcellular location determination and high-throughput microscopy. *Dev Cell* 12: 7–16.
- Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol* 238: 54–61.
- Cedano J, Aloy P, Perez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. *J Mol Biol* 266: 594–600.
- Chou KC, Elrod DW (1999) Protein subcellular location prediction. *Protein Engineering* 12: 107–118.
- Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* 300: 1005–1016.
- Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. *PROTEINS: Structure, Function, and Genetics* 50: 44–48.
- Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4: 1581–1590.
- Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, et al. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci* 14: 2804–2813.
- Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* 22: e408–416.
- Nakai K (2000) Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry* 54: 277–344.
- Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry* 370: 1–16.
- Chou KC, Shen HB (2007) Large-scale plant protein subcellular location prediction. *Journal of Cellular Biochemistry* 100: 665–678.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. *Nature Genetics* 25: 25–29.
- Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics* (Erratum: ibid, 2001, Vol44, 60) 43: 246–255.
- Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, et al. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* 13: 662–672.
- Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, et al. (2009) The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 37: D396–403.
- Chou KC, Shen HB (2008) Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols* 3: 153–162.
- Chou KC, Shen HB (2007) MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Comm* 360: 339–345.
- Chou KC, Shen HB (2008) ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem Biophys Res Comm* 376: 321–325.
- Smith C (2008) Subcellular targeting of proteins and drugs. <http://www.biocompare.com/Articles/TechnologySpotlight/976/Subcellular-Targeting-Of-Proteins-And-Drugs.html>.
- Millar AH, Carrie C, Pogson B, Whelan J (2009) Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. *Plant Cell* 21: 1625–1631.
- Chou KC, Shen HB (2007) Euk-mPLOC: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *Journal of Proteome Research* 6: 1728–1734.
- Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 29: 2994–3005.
- Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. *Journal of Biological Chemistry* 269: 22014–22020.
- Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, et al. (2009) Protein function annotation by homology-based inference. *Genome Biol* 10: 207.
- Gerstein M, Thornton JM (2003) Sequences and topology. *Curr Opin Struct Biol* 13: 341–343.
- Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. *Current Medicinal Chemistry* 11: 2105–2134.
- Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* 451: 591–595.
- Wang J, Pielak RM, McClintock MA, Chou JJ (2009) Solution structure and functional analysis of the influenza B proton channel. *Nat Struct Mol Biol* 16: 1267–1271.
- Chou KC (2004) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochemical and Biophysical Research Communications* 316: 636–642.
- Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry* 277: 45765–45769.
- Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical Journal* 84: 3257–3263.
- Murvai J, Vlahovicek K, Barta E, Pongor S (2001) The SBASE protein domain library, release 8.0: a collection of annotated protein sequence segments. *Nucleic Acids Research* 29: 58–60.
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.
- Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, et al. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34: D257–260.
- Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34: D247–251.
- Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, et al. (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 35: D237–240.
- Chou KC (1995) The convergence-divergence duality in lectin domains of the selectin family and its implications. *FEBS Letters* 363: 123–126.
- Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics* 6: 262–274.
- Esmacili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *Journal of Theoretical Biology* 263: 203–209.
- Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. *Journal of Theoretical Biology* 253: 310–315.
- Lin H, Wang H, Ding H, Chen YL, Li QZ (2009) Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. *Acta Biotheoretica* 57: 321–330.
- Ding YS, Zhang TL (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognition Letters* 29: 1887–1892.
- Lin H, Ding H, Feng-Biao Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein & Peptide Letters* 15: 739–744.
- Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *Journal of Theoretical Biology* 252: 350–356.
- Qiu JD, Huang JH, Liang RP, Lu XQ (2009) Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition:

## Acknowledgments

The authors wish to thank the reviewers for the valuable suggestions and comments, which are very helpful for strengthening the presentation of this paper.

## Author Contributions

Conceived and designed the experiments: KCC HBS. Performed the experiments: KCC HBS. Analyzed the data: KCC HBS. Wrote the paper: KCC.

- an approach from discrete wavelet transform. *Analytical Biochemistry* 390: 68–73.
48. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *Journal of Theoretical Biology* 257: 17–26.
  49. Gu Q, Ding YS, Zhang TL (2010) Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. *Protein Pept Lett* 17: 559–567.
  50. Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, et al. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *Journal of Theoretical Biology* 259: 366–372.
  51. Jiang X, Wei R, Zhang TL, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein & Peptide Letters* 15: 392–396.
  52. Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein & Peptide Letters* 15: 612–616.
  53. Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein & Peptide Letters* 16: 351–355.
  54. Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *Journal of Theoretical Biology* 248: 546–551.
  55. Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein & Peptide Letters* 16: 27–31.
  56. Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks, and connectivity indices. *Proteomics* 8: 750–778.
  57. Denoeux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics* 25: 804–813.
  58. Shen HB, Chou KC (2005) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochemical & Biophysical Research Communications* 334: 288–292.
  59. Cover TM, Hart PE (1967) Nearest neighbour pattern classification. *IEEE Transaction on Information Theory IT-13*: 21–27.
  60. Shafer G (1976) *A mathematical theory of evidence*. Princeton N.J.: Princeton University Press.
  61. Zouhal LM, Denoeux T (1998) An evidence-theoretic K-NN rule with parameter optimization. *IEEE Transactions on Systems, Man and Cybernetics* 28: 263–271.
  62. Shen HB, Song JN, Chou KC (2009) Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *Journal of Biomedical Science and Engineering (JBISE)* 2: 136–143. (openly accessible at <http://www.srpublishing.org/journal/jbise/>).
  63. Chou KC, Shen HB (2009) FoldRate: A web-server for predicting protein folding rates from primary sequence. *The Open Bioinformatics Journal* 3: 31–50. (openly accessible at <http://www.bentham.org/open/tobioij/>).
  64. Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 30: 275–349.
  65. Zhou GP (1998) An intriguing controversy over protein structural class prediction. *Journal of Protein Chemistry* 17: 729–738.
  66. Chen K, Kurgan LA, Ruan J (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J Comput Chem* 29: 1596–1604.
  67. Jiang Y, Iglinski P, Kurgan L (2008) Prediction of protein folding rates from primary sequences using hybrid sequence representation. *J Comput Chem*.
  68. Yang JY, Peng ZL, Yu ZG, Zhang RJ, Anh V, et al. (2009) Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *Journal of Theoretical Biology* 257: 618–626.
  69. He ZS, Zhang J, Shi XH, Hu LL, Kong XG, et al. (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE* 5: e9603.