Journal of Cellular Biochemistry 100:665–678 (2007)

Large-Scale Plant Protein Subcellular Location Prediction

Kuo-Chen Chou[1,2]* and Hong-Bin Shen[2]

[1]Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130
[2]Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, 1954 Hua-Shan Road, Shanghai 200030, China

Abstract Current plant genome sequencing projects have called for development of novel and powerful high throughput tools for timely annotating the subcellular location of uncharacterized plant proteins. In view of this, an ensemble classifier, Plant-PLoc, formed by fusing many basic individual classifiers, has been developed for large-scale subcellular location prediction for plant proteins. Each of the basic classifiers was engineered by the K-Nearest Neighbor (KNN) rule. Plant-PLoc discriminates plant proteins among the following 11 subcellular locations: (1) cell wall, (2) chloroplast, (3) cytoplasm, (4) endoplasmic reticulum, (5) extracell, (6) mitochondrion, (7) nucleus, (8) peroxisome, (9) plasma membrane, (10) plastid, and (11) vacuole. As a demonstration, predictions were performed on a stringent benchmark dataset in which none of the proteins included has $\geq$25% sequence identity to any other in a same subcellular location to avoid the homology bias. The overall success rate thus obtained was 32–51% higher than the rates obtained by the previous methods on the same benchmark dataset. The essence of Plant-PLoc in enhancing the prediction quality and its significance in biological applications are discussed. Plant-PLoc is accessible to public as a free web-server at http://202.120.37.186/bioinf/plant. Furthermore, for public convenience, results predicted by Plant-PLoc have been provided in a downloadable file at the same website for all plant protein entries in the Swiss-Prot database that do not have subcellular location annotations, or are annotated as being uncertain. The large-scale results will be updated twice a year to include new entries of plant proteins and reflect the continuous development of Plant-PLoc. J. Cell. Biochem. 100: 665–678, 2007.

Key words: plant protein; fusion classifier; gene ontology; GO discrete model; amphiphilic pseudo amino acid composition; KNN rule; plant-PLoc

Knowledge of the subcellular location of a protein is important because it can provide useful clues to reveal its function. Even if the function of a protein is known, it is equally important to find where and in what kind of environment the protein performs its function because one of the fundamental goals in cell biology and proteomics is to identify the functions of proteins in the context of compartments that organize them in the cellular environment. Although the knowledge of protein subcellular localization can be acquired by conducting various experiments, that is both expensive and time-consuming. Particularly, recent advances in large-scale genome sequencing have generated a huge number of protein sequences. For example, the Swiss-Prot [Bairoch and Apweiler, 2000] database contained only 3,939 protein sequence entries in 1986, but now the number has rapidly increased to 227,503 according to version 50.2 of the UniProtKB/Swiss-Prot Release as of June 27, 2006; that is, the number of protein sequences has increased by more than 57 times in about two decades.

The explosion of protein sequences has challenged us to develop an automated method for fast and reliably annotating the subcellular location of uncharacterized proteins. The knowledge thus obtained can help us timely utilize these newly found protein sequences for

This article contains supplementary material, which may be viewed at the Journal of Cellular Biochemistry website at http://www.interscience.wiley.com/jpages/0730-2312/suppmat/index.html.

*Correspondence to: Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130. E-mail: kchou@san.rr.com

Received 31 May 2006; Accepted 6 July 2006

DOI 10.1002/jcb.21096

both basic research and drug discovery (see, e.g., [Chou, 2004; Lubec et al., 2005]).

Many methods have been developed in this regard [Nakashima and Nishikawa, 1994; Cedano et al., 1997; Chou and Elrod, 1999; Nakai and Horton, 1999; Emanuelsson et al., 2000; Nakai, 2000; Feng, 2001; Chou and Cai, 2002; Feng, 2002; Pan et al., 2003; Zhou and Doctor, 2003; Garg et al., 2005; Matsuda et al., 2005; Shen and Chou, 2005a]. However, of these methods only the one by [Emanuelsson et al., 1999], called TargetP, was specialized for predicting the subcellular location of plant proteins. The predictor has been widely used since its inception, stimulating the studies of plant proteins and related areas. However, TargetP has the following problems that need to be further developed. (1) The prediction of TargetP actually only covers three locations if the uncertain location "other" as defined in TargetP is not counted. The three subcellular locations are: chloroplast, mitochondrion, and secretory pathway. Therefore, if a user wishes to use TargetP to predict a protein outside these three sites, such as endoplasmic reticulum, cell wall, and vacuole (Fig. 1), the predictor will fail to work, or the result thus obtained will be meaningless. (2) Protein sequences annotated as "POTENTIAL," "BY SIMILARITY," or "PROBABLE" were also included in deriving the prediction rule for TargetP, which might weaken the predictor due to lacking experimental evidences. (3) The benchmark dataset constructed for TargetP contains many homologous sequences. For example, after removing the 162 proteins labeled as "other" for the uncertain location, the benchmark dataset only contains 940–162 = 778 plant proteins, of which 141 belongs to chloroplast, 368 to mitochondrion, and 269 to secretary pathway. It has been found thru a sequence identity analysis [Wang and Dunbrack, 2003] that, among the 141 chloroplast proteins, there are 4 pairs, that is, (Q39734, Q42910), (P09195, P46275), (P26259, P24847), and (P15193, P12330), that have more than 80% sequence identity (Table I). Among the 368 mitochondrion proteins, 34 pairs have more than 80% sequence identity; and among the 269 secretary pathway, 12 pairs. If a cutoff is imposed to exclude those sequences which have $\geq$25% sequence identity to each other in a same subcellular location, the remaining protein sequences in the three locations would be reduced to 89, 182, and 110, respectively, which are only about 63%, 49%, and 41% of the proteins in the original dataset of TargetP.

To improve the quality of working dataset and avoid the redundancy and homology bias, a much more stringent dataset for plant proteins is needed. Also, to make the prediction practically more useful for plant cell, more subcellular locations need to be covered. Particularly, the plant genome

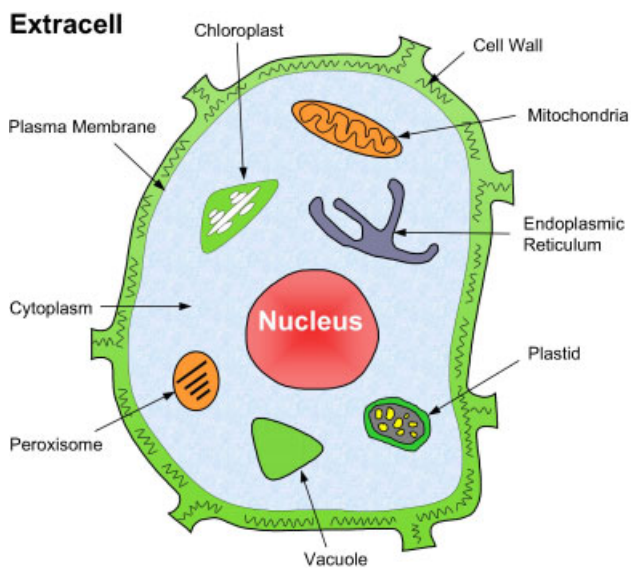

Fig. 1. Schematic illustration to show the eleven subcellular locations of plant proteins: (1) cell wall, (2) chloroplast, (3) cytoplasm, (4) endoplasmic reticulum, (5) extracell, (6) mitochondrion, (7) nucleus, (8) peroxisome, (9) plasma membrane, (10) plastid, and (11) vacuole. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TABLE I. List of Protein Pairs That Have More Than 80% Sequence Identity in the Benchmark Dataset Constructed for TargetP [Emanuelsson et al., 2000]

Subcellular location	Pair with more than 80% sequence identity	
Chloroplast	(Q39734, Q42910); 82%	(P26259, P24847); 81%
	(P09195, P46275); 81%	(P15193, P12330); 89%
Mitochondrion	(P11498, Q05920); 96%	(P15150, Q29552); 81%
	(P20004, Q99798); 96%	(Q61578, P08165); 86%
	(P40939, Q64428); 83%	(Q60587, P55084); 89%
	(P43304, Q64521); 92%	(Q60759, Q92947); 85%
	(Q08276, Q01899); 84%	(Q15118, Q63065); 92%
	(P29197, Q05045); 89%	(P00506, P00508); 83%
	(P29197, Q05046); 91%	(P16219, Q07417); 89%
	(P29197, Q43298); 86%	(P11066, Q00291); 83%
	(P38482, P19023); 83%	(P17783, Q43744); 82%
	(Q37683, Q95046); 86%	(Q16836, Q61425); 89%
	(Q03265, P19482); 98%	(Q14249, O08600); 86%
	(Q07536, Q02253); 94%	(Q07021, O35796); 82%
	(P06576, Q05825); 87%	(P51133, P51134); 91%
	(P54071, Q04467); 92%	(P46656, P08498); 87%
	(Q09128, Q07973); 82%	(Q39732, Q39733); 92%
	(Q09128, Q64441); 94%	(Q06056, Q06055); 87%
	(P14519, P34897); 95%	(Q06056, Q06646); 86%
Secretory pathway	(P14133, P24792); 81%	(P11955, P23951); 91%
	(P11515, P37891); 85%	(P06289, P06451); 83%
	(P22284, P22285); 91%	(P06289, P06452); 84%
	(P25778, Q10717); 81%	(P09762, P09761); 82%
	(P24101, P00433); 87%	(Q43194, Q43193); 82%
	(P23432, P52398); 84%	(P18263, P51317); 83%

Proteins are represented by their accession numbers.

sequencing projects [Jackson et al., 2006; Jorgensen, 2006] have called for development of novel and powerful high throughput tools to timely annotate the subcellular location of uncharacterized plant proteins. This kind of development may also stimulate the in-depth investigation of metabolic pathways, whose knowledge is indispensable for understanding a living system at the level of molecular networks [Chou et al., 2006].

The present study was initiated in an attempt to develop a new approach by which the identification can cover more subcellular locations of plant proteins and in the mean time bear less unwanted bias. To realize this, a new dataset was constructed that covers 11 subcellular locations, with a stringent criterion that none of proteins included has $\geq$25% sequence identity to any other in a same subcellular location.

As is well known, the more stringent criterion is imposed to exclude homologous proteins from a benchmark dataset, the harder it is to get a higher success rate. Also, the more the number of subcellular locations covered, the lower the odds are in getting a correct prediction. To overcome these extra difficulties, the technique by hybridizing and fusing different classifiers was introduced.

MATERIALS

Protein sequences were collected from the Swiss-Prot database [Bairoch and Apweiler, 2000] version 49.3 at http://www.ebi.ac.uk/swissprot/ released on March 21, 2006 according to the annotation information in the CC (comment or notes) and OC (organism classification) fields. In order to collect as much desired information as possible, but meanwhile ensure a high-quality for the working datasets, the data were screened strictly according to the following criteria. (1) Only those sequences annotated with "viridiplantae" in the OC field were collected because the current study was focused on plant proteins only. (2) Because a same subcellular location (SUBCELLULAR LOCATION) in the CC field might be annotated with different terms, several key words were used for a same subcellular location. For example, in search for cytoplasmic proteins, the key words "cytoplasm," and "cytoplasmic" were used; in search for extracell proteins, the key words "extracell," "extracellular," and "secreted" were used; in search for mitochondrial proteins, the key words "mitochondrion," "mitochondria," and "mitochondrial" were used; "in search for peroxisomal proteins," the key words "peroxisome," "peroxisomal," "microsome," "glyoxysomal," and "glycosomal" were used; in search for plasma membrane proteins, the key words "plasma membrane", "integral membrane," "multi-pass membrane," and "single-pass membrane" were used; in search for vacuolar proteins, the key words "vacuole"

and "vacuolar" were used; and so forth. (3) Sequences annotated with ambiguous or uncertain terms, such as "potential," "probable," "probably'," "maybe," or "by similarity," were excluded. (4) Sequences annotated by two or more locations were not included because of lack of the uniqueness. (5) Sequences annotated with "fragment" were excluded; also, sequences with less than 50 amino acid residues were removed because they might just be fragments. (6) To avoid any homology bias, a redundancy cutoff was operated by a culling program [Wang and Dunbrack, 2003] to winnow those sequences which have $\geq$25% sequence identity to any other in a same subcellular location. (7) Those subcellular locations (subsets) which contain less than ten protein sequences were left out because of lacking statistical significance.

After strictly following the above procedures, we finally obtained 671 protein sequences of which 12 belonged to cell wall, 204 to chloroplast, 101 to cytoplasm, 18 to endoplasmic reticulum, 46 to extracell, 96 to mitochondrion, 85 to nucleus, 16 to peroxisome, 40 to plasma membrane, 29 to plastid, and 24 to vacuole (Fig. 1). Thus, we have a dataset $\mathbb{S}^0$ which is a union of the following 11 subsets; that is,

$$\mathbb{S}^0 = S_1^0 \cup S_2^0 \cup S_3^0 \cup \cdots S_{11}^0 \qquad (1)$$

On the basis of dataset $\mathbb{S}^0$, two working datasets, that is, a learning (training) dataset $\mathbb{S}^{\mathrm{L}}$and an independent testing dataset $\mathbb{S}^{\mathrm{T}}$, were constructed. In order to fully use the data in $\mathbb{S}^0$ and meanwhile guarantee that $\mathbb{S}^{\mathrm{L}}$ and $\mathbb{S}^{\mathrm{T}}$ be completely independent of each other, the following condition was imposed:

$$\mathbb{S}^{\mathrm{L}} \cup \mathbb{S}^{\mathrm{T}} = \mathbb{S}^0 \text{ and } \mathbb{S}^{\mathrm{L}} \cap \mathbb{S}^{\mathrm{T}} = \emptyset \qquad (2)$$

where $\cup$, $\cap$, and $\emptyset$ represent the symbols for "union," "intersection," and "empty set" in the set theory, respectively. Protein samples in the corresponding subsets of $\mathbb{S}^{\mathrm{L}}$ and $\mathbb{S}^{\mathrm{T}}$ are randomly assigned according to the following "bracket percentage distribution" criterion:

$$n_i^{\mathrm{L}} = \begin{cases} 50 + \mathrm{INT}\{(n_i^0 - 50) \times 0.2\}, & \text{if } n_i^0 \geq 50 \\ \mathrm{INT}\{n_i^0 \times 0.8\}, & \text{if } 20 \leq n_i^0 < 50 \\ \mathrm{INT}\{n_i^0 \times 0.9\}, & \text{if } 10 \leq n_i^0 < 20 \end{cases} \qquad (3a)$$

with

$$n_i^{\mathrm{T}} = n_i^0 - n_i^{\mathrm{L}} \quad (i = 1, 2, \ldots, 11) \qquad (3b)$$

where n_i^0, n_i^{L}, n_i^{T} are the numbers of protein samples in the ith subset of the original dataset $\mathbb{S}^0$, learning dataset $\mathbb{S}^{\mathrm{L}}$, and testing dataset $\mathbb{S}^{\mathrm{T}}$, respectively, and INT is the integer-truncating operator meaning to take the integer part for the number in the brackets right after it. The numbers of proteins thus obtained for the 11 subcellular locations in the learning dataset $\mathbb{S}^{\mathrm{L}}$ and testing dataset $\mathbb{S}^{\mathrm{T}}$ are given in Table II. The accession numbers and sequences for the corresponding proteins in the learning and testing datasets are given in the Online Supplementary Materials A and B, respectively.

METHOD

Now the problem we are facing is how to use some known data to deduce some unknown information. For the current study, the known data are the sequences of proteins as well as the annotations of those proteins whose subcellular locations are known through experimental observations and clearly annotated; while the unknown data, or the desired results, are the subcellular locations of the remaining proteins. To deal with this kind of problem, the first important thing is how to effectively represent the sample of a protein. The most straightforward way in this regard is to use the sequential model, that is, represent a protein sample with its entire amino acid sequence, and then deduce the subcellular location of an uncharacterized protein according to the sequence similarity principle. However, this kind of straightforward sequence-based approach (such as BLAST [Altschul et al., 1997]) will fail to work when the query protein does not have significant

TABLE II. Number of Plant Proteins in Each of the 11 Subcellular Locations for the Learning and Testing Datasets, Respectively

Subcellular location	Learning dataset $\mathbb{S}^{\mathrm{L}}$	Testing dataset $\mathbb{S}^{\mathrm{T}}$
(1) Cell wall	10	2
(2) Chloroplast	80	124
(3) Cytoplasm	60	41
(4) Endoplasmic reticulum	16	2
(5) Extracell	36	10
(6) Mitochondrion	59	37
(7) Nucleus	57	28
(8) Peroxisome	14	2
(9) Plasma membrane	32	8
(10) Plastid	23	6
(11) Vacuole	19	5
Total	406	265

homology to proteins of known localization. For instance, for a protein of only 50 residues, the number of different sequence order combinations would be $20^{50} \simeq 1.1259 \times 10^{65}$. Actually, the average protein length is much longer than 50. According to Swiss-Prot data bank [Bairoch and Apweiler, 2000] the average length per sequence is getting longer each year: it was 229 in 1986, but it has been increased to 367 based on the release 49.7 of May 16, 2006. The number of different combinations for a protein with 367 residues will be $20^{367} = 10^{367\log 20} > 10^{477}$. For such an astronomical number, it is impractical to construct a training data set to statistically cover all the possible cluster patterns. Furthermore, protein sequence lengths vary widely. This has posed an additional difficulty for using the sequential model for protein subcellular location. To avoid this kind of difficulties caused by the sequential model, a feasible approach is to resort to the discrete model. The simplest discrete model for representing a protein sample is the amino acid composition (AA), which was widely used by many previous investigators to predict protein structural class [Klein, 1986; Klein and Delisi, 1986; Nakashima et al., 1986; Deleage and Roux, 1987; Metfessel et al., 1993; Chou and Zhang, 1994; Mao et al., 1994; Chandonia and Karplus, 1995; Chou, 1995; Bahar et al., 1997; Chou and Maggiora, 1998; Liu and Chou, 1998; Zhou, 1998; Zhou and Assa-Munt, 2001; Luo et al., 2002; Cao et al., 2006; Lee et al., 2006]. The AA discrete model consists of 20 numbers each representing the occurrence frequency of one of the 20 native amino acids in a protein. Its advantage is simple and easy to be formulated for various existing algorithms or predictors, such as the least Euclidean distance algorithm [Nakashima et al., 1986; Nakashima and Nishikawa, 1994], ProtLock predictor [Cedano et al., 1997], covariant discriminant algorithm [Chou and Elrod, 1999], neural network algorithm [Cai and Zhou, 2000], and support vector machines (SVM) [Vapnik, 1998]. However, the AA discrete model did not include any sequence-order information, and hence the success rates by the predictors based on it would be limited. To improve the situation, the pseudo amino acid composition (PseAA) was introduced [Chou, 2001]. The PseAA discrete model consists of $20 + \Lambda$ numbers, where the first 20 numbers are the same as those in the AA discrete model and the remaining numbers represent Λ sequence-correlation factors of different ranks. It is through the latter that a considerable amount of sequence-order information is incorporated [Chou, 2001], and the prediction quality has been remarkably improved [Feng, 2002; Pan et al., 2003; Shen and Chou, 2005b; Zhang et al., 2006; Xiao et al., 2006b]. Subsequently, the functional domain composition (FunD) was introduced [Chou and Cai, 2002]. The FunD discrete model was extremely successful for predicting protein structural class [Chou and Cai, 2004], implying that the structural class of a protein is closely correlated with the components of its FunD. In other words, the latter reflects the core feature of a protein in studying the structural classification. Actually, the AA, PseAA, and FunD discrete models are all reflecting some sort of core feature of a protein although from different angles or with different focuses. Now, the problem is how to find the optimal core feature for the focus of predicting protein subcellular location?

Here we are to use the gene ontology (GO) database [Ashburner et al., 2000; Harris et al., 2004] to formulate the core feature of a protein. The reason for us to do so was based on such an assumption that representation of protein samples in the GO database space would make them clustered in a way closely correlated with their subcellular locations because the GO database was established based on the following three species-independent principles: molecular function, biological process, and cellular component [Camon et al., 2004; Lee et al., 2005]. All these criteria are not only the attributes of genes, gene products or gene-product groups, but also closely correlated with the subcellular localization. However, how to establish a predictor based on the GO database to improve the prediction quality for protein subcellular location is by no means a trivial problem. The reasons are as follows. (1) For those proteins with "subcellular location unknown" annotation in Swiss-Prot database, most (more than 99%) of their corresponding GO numbers in GO database are also annotated with "cellular component unknown" (see, e.g., the proteins with accession numbers O75920, P07315, and Q92796 in Table III). (2) Even for some proteins whose subcellular locations are clearly annotated in Swiss-Prot database, their corresponding GO numbers in GO database are annotated with "cellular component unknown." For example, for the proteins with accession number

TABLE III. Examples to Show the Subcellular Location Annotations for Some Proteins in the Swiss-Prot Database and the Annotations for the Corresponding GO Numbers in the GO Database

Swiss-Prot database		GO database	
Accession number	Swiss-Prot annotation	GO number	GO annotation
O75920	No subcellular location annotated	GO:0005554	Molecular function unknown
		GO:0007399	Nervous system development
		GO:0008372	Cellular component unknown
P07315	No subcellular location annotated	GO:0000004	Biological process unknown
		GO:0005212	Structural constituent of eye lens
		GO:0008372	Cellular component unknown
Q92796	No subcellular location annotated	GO:0004385	Guanylate kinase activity
		GO:0005515	Protein binding
		GO:0008285	Negative regulation of cell proliferation
		GO:0008372	Cellular component unknown
O75897	Cytoplasm	GO:0000004	Biological process unknown
		GO:0008146	Sulfotransferase activity
		GO:0008372	Cellular component unknown
		GO:0016740	Transferase activity
P83683	Extracellular	GO:0005184	Neuropeptide hormone activity
		GO:0007218	Neuropeptide signaling pathway
		GO:0008372	Cellular component unknown
O43303	Centrosome	GO:0000004	Biological process unknown
		GO:0005554	Molecular function unknown
		GO:0008372	Cellular component unknown
P83168	Extracellular	GO:0004866	Endopeptidase inhibitor activity
		GO:0004867	Serine-type endopeptidase inhibitor activity
		GO:0008372	Cellular component unknown
		GO:0030162	Regulation of proteolysis

O75897, P83683, O43303, and P83168 in Table III, their subcellular locations are annotated with "cytoplasm," "extracellular," "centrosome," and "extracellular," respectively, in Swiss-Prot database, but these proteins are annotated with "cellular component unknown" in the GO database. (3) It was found through a statistical analysis that, of the 15,348 plant protein sequence entries in the Swiss-Prot database (version 50.0, released May 30, 2006), only 6,355 are annotated with experimentally observed subcellular locations, and 4,540 annotated with uncertain locations such as "potential," "maybe," and "probable." As mentioned in the Materials section, the uncertain annotations cannot be used as robust data for establishing a predictor. Actually, the plant proteins with uncertain annotations also belong to our target of prediction. The similar but even more complicated situation also exists in the GO database. Because the GO database was derived from the Swiss-Prot database, the subcellular component annotations in GO would unavoidably contain this kind of uncertain information so as to complicate the problem. Therefore, the subcellular component annotations in GO database cannot be used as the robust data for establishing a solid predictor either. Accordingly, the really useful data in this regard can only be taken from the 6,355 plant proteins with experimentally observed subcellular locations as clearly annotated in the Swiss-Prot database. In other words, to fill the gap, the number of plant proteins whose subcellular locations need to be predicted is $(15{,}348-6{,}355)=8{,}933$, which is more than 58% of all the plant proteins in Swiss-Prot database.

Accordingly, the information in GO database that may be useful for formulating the core feature of proteins is actually "buried" into many tedious GO numbers, just like it is "buried" into many amino acid components in the AA discrete model [Cedano et al., 1997], or pseudo amino acid components in the PseAA discrete model [Chou, 2001], or functional domain components in the FunD discrete model [Chou and Cai, 2002], or original amino acid sequences in the sequential model [Altschul et al., 1997]. To "dig out" the useful information, let us consider the GO discrete model, as formulated below.

Mapping UniProtKB/Swiss-Prot protein entries [Apweiler et al., 2004] to the GO database, one can get a list of data called "gene_association.goa_uniprot," where each UniProtKB/Swiss-Port protein entry corresponds to one or several GO numbers. In this

study, such a data file was directly downloaded from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/ (released on March 4, 2006). The relationships between the UniProtKB/Swiss-Port protein entries and the GO numbers may be one-to-many, "reflecting the biological reality that a particular protein may function in several processes, contain domains that carry out diverse molecular functions, and participate in multiple alternative interactions with other proteins, organelles or locations in the cell" [Ashburner et al., 2000], as exemplified in Table III. On the other hand, because the current GO database is not complete yet, some protein entries (such as "P27057," "Q8LGI2," and "P32034") have no corresponding GO numbers, that is, no mapping records at all in the GO database, and hence are not included in the data list of gene_association.goa_uniprot.

The GO numbers do not increase successively and orderly. For easier handling, some reorganization and compression procedure was taken to renumber them. For example, after such a procedure, the original GO numbers GO:0000001, GO:0000002, GO:0000003, GO:0000004, GO:0000006, ..., GO:0051912 would become GO_compress:0000001, GO_compress:0000002, GO_compress:0000003, GO_ compress :0000004, GO_compress:0000005, ..., and GO_ compress:0009918, respectively. The GO database thus obtained is called GO_compress database, whose dimensions were reduced from 51,912 in the original GO database to 9918. Each of the 9,918 entities in the GO_compress database served as a base to define a protein sample. Unfortunately, the current GO numbers failed to give a complete coverage in the sense that some proteins might not belong to any of the GO numbers as mentioned above. Although the problem will gradually become trivial or eventually be solved with the GO database developing, to tackle such a problem right now, a hybridization approach was introduced by fusing the GO approach and the amphiphilic pseudo amino acid composition (PseAA) approach [Chou, 2005], as described below.

(1) Search a protein sample in the GO_compress database, if there is a hit corresponding to the ith GO_compress number, then the ith component of the protein in the 9918-D (dimensional) GO_compress space is assigned 1; otherwise, 0. Thus, the protein can be formulated as:

$$\mathbf{P} = [g_1\, g_2 \ldots g_{\mathrm{i}} \ldots g_{9918}]^{\mathbf{T}} \tag{4}$$

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

$$g_{\mathrm{i}} = \begin{cases} 1, & \text{hit found in GO\_compress} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

(2) If no hit (i.e., no record in the GO_compress database) is found at all, then the protein should be defined in the $(20 + 2\lambda)$-D amphiphilic PseAA space [Chou, 2005], as given below

$$\begin{aligned} \mathbf{P} &= [p_1 \ldots p_{20} \quad p_{20+1} \ldots p_{20+2\lambda}]^{\mathbf{T}} \\ &= [p_1 \ldots p_{20} \ldots p_{\Lambda}]^{\mathbf{T}}, \end{aligned} \tag{6}$$

where $p_1, \ldots p_{20}$ are associated with the amino acid composition reflecting the occurrence frequencies of the 20 native amino acids in the protein [Nakashima et al., 1986; Chou and Zhang, 1994], and $p_{20+1}, \ldots p_{20+2\lambda}$ are the 2λ correlation factors that reflect its sequence-order pattern thru the amphiphilic feature [Chou, 2005]. For simplifying the formulation later on, $\Lambda = 20 + 2\lambda$ is used for Equation 6. The protein representation thus defined is called the "amphiphilic pseudo amino acid composition" or PseAA, which has the same form as the conventional amino acid composition but contains more components and information. For reader's convenience, a brief introduction about the PseAA and the key equations for deriving its components are provided in Online Support Materials C.

Suppose there are N proteins ($\mathbf{P}_1, \mathbf{P}_2, \ldots \mathbf{P}_{\mathrm{N}}$) which have been classified into $\mathrm{M} = 11$ subsets (subcellular locations). Now, for a query protein $\mathbf{P}$, how can we identify which subset it belongs to? Below we shall use the K-Nearest Neighbor (KNN) rule [Cover and Hart, 1967; Keller et al., 1985; Denoeux, 1995] to deal with this problem. According to the KNN rule, the query protein should be assigned to the subset represented by the majority of its K nearest neighbors. 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. There are many different definitions to measure the "nearness" for the KNN classifier, such as Euclidean distance, Hamming distance [Mardia et al., 1979], and Mahalanobis distance [Mahalanobis, 1936; Pillai, 1985; Chou, 1995]. Here, we use the

following equation to measure the nearness between proteins $\mathbf{P}$ and $\mathbf{P}_i$

$$\delta(\mathbf{P}, \mathbf{P}_i) = 1 - \frac{\mathbf{P} \cdot \mathbf{P}_i}{||\mathbf{P}||||\mathbf{P}||} \quad (7)$$

where $\mathbf{P} \cdot \mathbf{P}_i$ is the dot product of the two vectors, and $||\mathbf{P}||$ and $||\mathbf{P}_i||$ their modulus, respectively. According to Equation 7, when $\mathbf{P} \equiv \mathbf{P}_i$ we have $\delta(\mathbf{P}, \mathbf{P}_i) = 0$, indicating the "distance" between the two proteins is zero and hence they are regarded as having perfect or 100% similarity.

In using the KNN rule, the predicted result will depend on the selection of the parameter K, the number of the nearest neighbors to the query protein $\mathbf{P}$. If $K = 1$, the protein $\mathbf{P}$ will be predicted belonging to the same subcellular location of the protein in the training dataset that has the shortest "distance" to $\mathbf{P}$ as defined by Equation 7. If there are two and more proteins in the training dataset ($\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N$) that have exactly the same shortest distance to $\mathbf{P}$, the query protein will be randomly assigned to one of their subcellular locations although this kind of tie case rarely happens. When $K > 1$, the subcellular location of the query protein $\mathbf{P}$ will be determined by the majority of its K nearest neighbors thru a vote. If there is a tie for the voting results, the query protein will be randomly assigned to one of the locations associated with the tie case. Generally speaking, the greater the K (the number of the nearest neighbors considered), the less likely the tie case occurs. In the current study, no tie case was observed when $K \geq 5$.

Because the predicted results by the KNN algorithm [Cover and Hart, 1967; Keller et al., 1985; Denoeux, 1995] depend on the selection of parameter K, hereafter we shall use NN(K) to represent the symbol of KNN, implying that the predicted result is the function of K, the number of the nearest neighbors concerned for the query protein $\mathbf{P}$.

During the course of prediction, the following self-consistency principle should be followed. If a query protein was defined in the 9918-D GO_compress space (Eq. 4), then the prediction should be carried out based on those proteins in the training dataset that could be defined in the same 9918-D space. If the query protein in the 9918-D GO_compress space was a naught vector and hence must be defined instead in the $(20 + 2\lambda)$-D or Λ-D PseAA space (Eq. 6), then the prediction should be conducted according to the principle that all the proteins in the training dataset be defined in the same Λ-D space as well. Accordingly, the current hybridization predictor actually consists of two sub-predictors: (1) the NN(K)-GO predictor that operates in the 9918-D GO_compress space, and (2) the NN(K,Λ)-PseAA predictor that operates in the Λ-D amphiphilic PseAA space. The former is the function of K, while the latter the function of both K and Λ. For a given learning dataset, selection of different K and Λ would result in different outcomes. To get the optimal success rate, one has to test the results by using different numbers of K and Λ one by one. However, it is both time-consuming and tedious to do so. To solve such a problem, the following two fusion processes are introduced for the NN(K) and NN(K, Λ) classifiers, respectively.

One Dimensional Fusion

It is for generating an ensemble classifier by fusing many 1-D individual basic NN(K) classifiers each having a different specified value of K, as formulated by

$$\mathbb{NN}^{GO} = \mathrm{NN}(1) \forall \mathrm{NN}(2) \forall \ldots \forall \mathrm{NN}(\Omega) \quad (8)$$

where the symbol $\forall$ denotes the fusing operator, and $\mathbb{NN}^{GO}$ the ensemble classifier formed by fusing NN(1), NN(2), ..., and NN(Ω) according to the flowchart of Figure 2. Here $\Omega = 10$ because preliminary tests indicated that the success rate obtained by the NN(K) classifier trained by the current learning dataset was lower when $K > 10$.

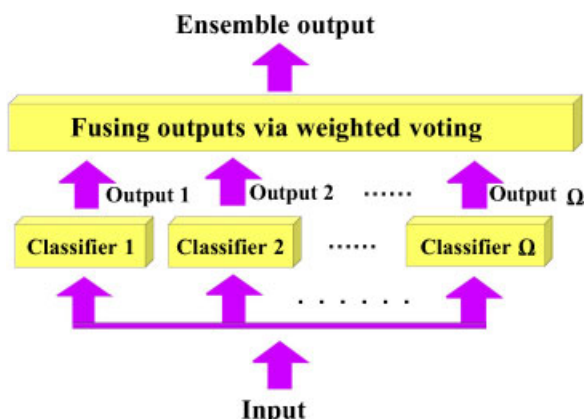

Fig. 2. Flowchart to show how the ensemble classifiers $\mathbb{NN}^{GO}$ (Eq. 8) and $\mathbb{NN}^{Pse}$ (Eq. 13) are formed by fusing Ω individual classifiers, where $\Omega = 10$ and 210 for the cases of $\mathbb{NN}^{GO}$ and $\mathbb{NN}^{Pse}$, respectively. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The process of how the ensemble classifier $\mathbb{NN}^{\mathrm{GO}}$ works is as follows. Suppose the predicted classification results for the query protein **P** by the 10 individual classifiers in Equation 8 are C_1,C_2,...,C_{10}, respectively; that is,

$$C_i \in S_\mu (i = 1, 2, \ldots, 10; \quad \mu = 1, 2, \ldots, 11) \quad (9)$$

where $\in$ is a symbol in the set theory meaning "member of", and S_1, S_2, S_3,...,S_{11} represent the 11 subsets defined by the 11 subcellular locations studied here (Fig. 1), and the voting score for the protein **P** belonging to the μth subset is defined by

$$Y_\mu^{\mathrm{GO}} = \sum_{i=1}^{10} w_i \Delta(C_i, S_\mu), \quad (\mu = 1, 2, \ldots, 11) \quad (10)$$

where w_i is the weight and was set at 1 for simplicity, and the delta function in Equation 10 is given by

$$\Delta(C_i, S_\mu) = \begin{cases} 1, & \text{if } C_i \in S_\mu \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

thus the query protein **P** is predicted belonging to the subset (subcellular location) with which its score of Equation 10 is the highest.

Two Dimensional Fusion

It is for generating an ensemble classifier by fusing many 2-D individual basic NN(K,Λ) classifiers each having different specified values of K and Λ. Owing to the similar reason as mentioned above in setting the value of Ω for Equation 8, let us consider K=1, 2,...,10, and $\Lambda = 20, 22, \ldots, 60$; that is,

$$\begin{aligned} \{\mathrm{K}\} &= \{1, 2, \ldots, 10\}; \\ \{\Lambda\} &= \{20, 22, \ldots, 58, 60\} \end{aligned} \quad (12)$$

Thus, the ensemble classifier obtained by the two-dimensional fusion process can be formulated as

$$\begin{aligned} \mathbb{NN}^{\mathrm{Pse}} = \mathrm{NN}(1,20) \forall \mathrm{NN}(1,22) \forall \ldots \\ \forall \mathrm{NN}(10,58) \forall \mathrm{NN}(10,60) \end{aligned} \quad (13)$$

where the fusion operator $\forall$ has the same meaning as that of Equation 8, and the fusion flowchart can also be illustrated by Figure 2 but with $\Omega = 10 \times 21 = 210$, meaning a process by fusing 210 basic individual classifiers now.

The detailed process of how the ensemble classifier $\mathbb{NN}^{\mathrm{Pse}}$ works is as follows. Suppose the predicted classification results for the query protein **P** by the 210 individual classifiers in Equation 13 are

$$\begin{aligned} C_{i,2j} \in S_\mu \quad (i = 1, 2, \ldots, 10; j \\ = 10, 11, \ldots, 30; \quad \mu = 1, 2, \ldots, 11) \end{aligned} \quad (14)$$

where S_1, S_2,...,S_{11} have the same meanings as in Equation 9, that is, represent the 11 subsets defined by the 11 subcellular locations studied here (Fig. 1), and the voting score for the protein **P** belonging to the μth subset is defined by

$$Y_\mu^{\mathrm{Pse}} = \sum_{i=1}^{10} \sum_{j=10}^{30} w_{i,2j} \Delta(C_{i,2j}, S_\mu), \ (\mu = 1, 2, \ldots, 11) \quad (15)$$

where $w_{i,2j}$ is the weight and was set at 1 for simplicity, the delta function in Equation 15 is given by

$$\Delta(C_{i,2j}, S_\mu) = \begin{cases} 1, & \text{if } C_{i,2j} \in S_\mu \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

thus the query protein **P** is predicted belonging to the subset (subcellular location) with which its score of Equation 15 is the highest.

The predictor thus established is named Plant-PLoc.

RESULTS AND DISCUSSION

For the proteins listed in the Online Supplementary Materials A and B we obtained the following results according to Steps 1–2 of Methods: (1) of the 406 proteins in the training dataset, 402 got hits in the GO_compress database, and hence were defined in the 9918-D GO_compress space (Eqs. 4 and 5), and the remainder defined in the Λ-D PseAA space (Eq. 6); (2) of the 265 proteins in the testing dataset, all got hits and were defined in the 9918-D GO_compress space, and none defined in the Λ-D PseAA space. However, this does not mean that there is no need to include the $\mathbb{NN}^{\mathrm{Pse}}$ predictor because in the practical application, cases do exist where the query proteins cannot be defined in GO system. It can be seen from a statistical analysis that currently there still are more than 3% plant proteins that have no any corresponding GO numbers. Although such a problem will be eventually solved with the continuous development of the GO database, it would be harmless and make the predictor more complete to keep the $\mathbb{NN}^{\mathrm{Pse}}$ classifier in the system since the prediction process is logically

operated according to the following priority: if a query protein can be defined in the 9918-D GO_compress space, then the classifier $\mathbb{NN}^{GO}$ is used to predict its subcellular location; otherwise, the classifier $\mathbb{NN}^{Pse}$ is used to predict its subcellular location.

The prediction quality was examined by two test methods: the jackknife test and the independent dataset test. In the jackknife test, each protein in the training dataset was singled out in turn as a "test protein" and all the rule parameters were calculated from the remaining N-1 proteins. In other words, the subcellular location of each protein was predicted by the rules derived using all the other proteins except the one that was being predicted. During the jackknifing process, both the training and testing dataset were actually open, and a protein was in turn moving from one to the other. In the independent dataset test, the rule parameters were derived from the proteins only in the training dataset, and the prediction was made for proteins in an independent dataset. Because the selection of independent dataset often bears some sort of arbitrariness, the jackknife test is deemed more objective than the independent dataset test. Actually, jackknife tests are thought one of the most rigorous and objective methods for cross-validation in statistics (see [Chou and Zhang, 1995] for a comprehensive review and [Mardia et al., 1979] for the mathematical principles), and have been increasingly used by investigators [Zhou, 1998; Feng, 2001; Zhou and Assa-Munt, 2001; Feng, 2002; Luo et al., 2002; Liu et al., 2005; Wang et al., 2005; Guo et al., 2006; Zhou and Cai, 2006; Xiao et al., 2006a] in examining the power of various prediction methods. Therefore, the power of a predictor should be measured by the success rate of jackknife test. The independent dataset test performed here was just for a demonstration of practical application.

The predicted results obtained by Plant-PLoc are given in Table IV, where, for facilitating comparison, the corresponding rates obtained by various other predictors are also listed. As we can see from Table IV, the overall success rates obtained by the current Plant-PLoc in both jackknife cross-validation test and independent dataset test were 32–51% higher than those by the other predictors, indicating that Plant-PLoc is indeed very powerful. Also, it can be seen from Table IV that the overall success rate by the current approach for the independent dataset test is 7% higher than that for the jackknife test. This is because in the benchmark dataset the 25% sequence identity cutoff was imposed only for the proteins in a same subcellular location; no such a cutoff was imposed for the proteins with different subcellular locations in order for reflecting the reality. The latter will make it even harder to enhance the jackknife success rate. Similar phenomenon can also be seen from Table IV for the results obtained by the SVM approach.

Why the SVM methods and other predictors reported in the previous studies could yield much higher success rates than those listed in Table IV? The reasons are as follows. (1) The benchmark datasets originally used in those predictors contained many homologous sequences in a same subcellular location. Some of the benchmark datasets used there contained proteins with up to 90% sequence identity. When predictions were made by them on the current stringent dataset in which none of protein has $\geq$25% sequence identity to any other in a same

TABLE IV. Overall Success Rates for the 11 Subcellular Locations (Fig. 1) of Plant Proteins by Different Classifiers and Test Methods

		Test method	
Classifier	Input form	Jackknife[a]	Independent dataset[b]
Least Euclidean distance [Nakashima and Nishikawa, 1994]	Amino acid composition	141/406 = 34.7%	82/265 = 30.9%
ProtLock [Cedano et al., 1997]	Amino acid composition	141/406 = 34.7%	88/265 = 33.2%
SVM [Vapnik, 1998]	Amino acid composition and amino acid pairs	80/406 = 19.7%	124/265 = 46.8%
Hybridization of ensemble classifiers ^{GO}(Eq. 5) and ^{Pse}(Eq. 10)	Hybridization of GO (Eq. 1) and amphiphilic PseAA (Eq. 3a)	290/406 = 71.4%	209/265 = 78.9%

[a]Jackknife cross-validation test was performed for the 406 proteins in the Online Supplementary Material A, where none of the proteins has $\geq$25% sequence identity to any other in the same subcellular location.
[b]Prediction was performed for the 265 independent proteins in the Online Supplementary Materials B; none of proteins in the Online Supplementary Materials A and B has $\geq$25% sequence identity to any others in the same subcellular location.

subcellular location, the success rates would of course decrease dramatically. (2) Most of the success rates reported by the previous investigators were derived from the benchmark datasets covering only 3–5 subcellular locations; when prediction was made on the current benchmark dataset that covers 11 subcellular locations, the odds in getting a correct prediction would of course become lower. (3) Most of the previously reported success rates were obtained by the sub-sampling cross-validation. When tested by the jackknife cross-validation, the corresponding rates would be further diminished because, as mentioned above, the jackknife cross-validation is much more stringent for conducting an objective test and tougher for getting a high success rate.

Since TargetP [Emanuelsson et al., 2000] is a predictor with a built-in training dataset covering only three subcellular location sites, to compare it with the current predictor Plant-PLoc, let us randomly pick 30 protein samples from Swiss-Prot databank according to the following criteria: (1) they must belong to plant proteins, as annotated with "viridiplantae" in the OC field; (2) they must neither occur in the training dataset of TargetP nor occur in the training dataset of Plant-PLoc in order for avoiding the unfair memory effect; (3) their experimentally observed subcellular locations are known as clearly annotated in the CC field, and also these locations must be within the scope covered by TargetP as a compromise for rationally using TargetP. The predicted results for the 30 plant proteins by TargetP and Plant-PLoc are given in Table V, from which we can see how the results miss-predicted by TargetP were successfully corrected by Plant-PLoc.

CONCLUSION

Prediction of plant protein subcellular location is an important problem but meanwhile a very difficult one. The more the number of subcellular locations is considered, or the more stringent condition is imposed to exclude the sequence redundancy and homology bias, the more difficult will be to get a higher success prediction rate. That is why for the benchmark dataset investigated here, which involves 11 subcellular locations and in which none of protein has $\geq$25% sequence identity to any

TABLE V. Examples to Show How the Results Miss-Predicted by TargetP Were Corrected by Plant-PLoc

	Subcellular localization		
Protein (accession number)	Annotation in Swiss-Prot	Predicted by TargetP	Predicted by Plant-PLoc
P12853	Chloroplast	Mitochondrion	Chloroplast
P14226	Chloroplast	Mitochondrion	Chloroplast
Q9SBN6	Chloroplast	Mitochondrion	Chloroplast
Q41643	Chloroplast	Mitochondrion	Chloroplast
P28260	Chloroplast	Any other location	Chloroplast
P48706	Chloroplast	Any other location	Chloroplast
P27065	Chloroplast	Any other location	Chloroplast
P25832	Chloroplast	Any other location	Chloroplast
Q6EW14	Chloroplast	Mitochondrion	Chloroplast
Q3BAJ9	Chloroplast	Mitochondrion	Chloroplast
O98456	Chloroplast	Mitochondrion	Chloroplast
P06510	Chloroplast	Mitochondrion	Chloroplast
Q42690	Chloroplast	Secretory pathway	Chloroplast
Q85FH6	Chloroplast	Secretory pathway	Chloroplast
P12127	Chloroplast	Secretory pathway	Chloroplast
P29685	Mitochondrion	Chloroplast	Mitochondrion
P17614	Mitochondrion	Chloroplast	Mitochondrion
Q9FT52	Mitochondrion	Any other location	Mitochondrion
P29380	Mitochondrion	Any other location	Mitochondrion
P62773	Mitochondrion	Any other location	Mitochondrion
P00075	Mitochondrion	Chloroplast	Mitochondrion
Q8LFT2	Mitochondrion	Chloroplast	Mitochondrion
Q9ZT91	Mitochondrion	Chloroplast	Mitochondrion
P52901	Mitochondrion	Chloroplast	Mitochondrion
P26871	Mitochondrion	Chloroplast	Mitochondrion
Q36665	Mitochondrion	Chloroplast	Mitochondrion
P46742	Mitochondrion	Secretory pathway	Mitochondrion
Q95747	Mitochondrion	Secretory pathway	Mitochondrion
P60099	Mitochondrion	Any other location	Mitochondrion
P42056	Mitochondrion	Any other location	Mitochondrion

others in a same subcellular location, the success rates obtained by various powerful existing methods were only within the range of 20–46%, which are 32–51% lower than the rates obtained by Plant-PLoc, a new predictor developed in this paper.

The overwhelmingly high success rates obtained by Plant-PLoc indicate that proteins, if represented through the GO discrete model, can be more distinctly clustered according to their different subcellular locations, and that the ensemble classifier presented here is indeed a powerful operation engine in distinguishing these clusters.

Since many plant proteins in Swiss-Prot and GO databases have no annotations to indicate their subcellular locations, a downloadable file listing the predicted results by Plant-PLoc for all these proteins has been provided at http://202.120.37.186/bioinf/plant. The file will be updated twice a year to support the new entries of plant proteins and reflect the continuous development of Plant-PLoc.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to the two anonymous reviewers whose valuable suggestions are very helpful for strengthening the presentation of this paper.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs (Review). Nucleic Acids Res 25:3389–3402.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. 2004. UniProt: The universal protein knowledgebase. Nucleic Acids Res 32:D115–D119.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: Tool for the unification of biology. Nat Genet 25:25–29.

Bahar I, Atilgan AR, Jernigan RL, Erman B. 1997. Understanding the recognition of protein structural classes by amino acid composition. Proteins Struct Funct Genet 29:172–185.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 25:31–36.

Cai YD, Zhou GP. 2000. Prediction of protein structural classes by neural network. Biochimie 82:783–785.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. 2004. The gene ontology annotation (GOA) database: Sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32: D262–D266.

Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K. 2006. Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7:20.

Cedano J, Aloy P, P'erez-Pons JA, Querol E. 1997. Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600.

Chandonia JM, Karplus M. 1995. Neural networks for secondary structure and structural class prediction. Protein Sci 4:275–285.

Chou KC. 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins Struct Funct Genet 21:319–344.

Chou KC. 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet (Erratum: ibid., 2001, Vol. 44, 60) 43:246–255.

Chou KC. 2004. Review: Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134.

Chou KC. 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19.

Chou KC, Cai YD. 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277:45765–45769.

Chou KC, Cai YD. 2004. Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun (Corrigendum: ibid., 2005, Vol. 329, 1362) 321:1007–1009.

Chou KC, Elrod DW. 1999. Protein subcellular location prediction. Protein Eng 12:107–118.

Chou KC, Maggiora GM. 1998. Domain structural class prediction. Protein Eng 11:523–538.

Chou KC, Zhang CT. 1994. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269:22014–22020.

Chou KC, Zhang CT. 1995. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349.

Chou KC, Cai YD, Zhong WZ. 2006. Predicting networking couples for metabolic pathways of Arabidopsis. EXCLI J 5:55–65.

Cover TM, Hart PE. 1967. Nearest neighbour pattern classification. IEEE Trans Inf Theory IT 13:21–27.

Deleage G, Roux B. 1987. An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1:289–294.

Denoeux T. 1995. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans Syst Man Cybern 25:804–813.

Emanuelsson O, Nielsen H, von Heijne G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 8:978–984.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016.

Feng ZP. 2001. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58:491–499.

Feng ZP. 2002. An overview on predicting the subcellular location of a protein. In Silico Biol 2:291–303.

Garg A, Bhasin M, Raghava GP. 2005. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 280:14427–14432.

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J. 2006. Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30:397–402.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. 2004. The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.

Jackson S, Rounsley S, Purugganan M. 2006. Comparative sequencing of plant genomes: Choices to make. Plant Cell 18:1100–1104.

Jorgensen R. 2006. Plant genomes. Plant Cell 18:1099.

Keller JM, Gray MR, Givens JA. 1985. A fuzzy k-nearest neighbours algorithm. IEEE Trans Syst Man Cybern 15:580–585.

Klein P. 1986. Prediction of protein structural class by discriminant analysis. Biochim Biophys Acta 874:205–215.

Klein P, Delisi C. 1986. Prediction of protein structural class from amino acid sequence. Biopolymers 25:1659–1672.

Lee V, Camon E, Dimmer E, Barrell D, Apweiler R. 2005. Who tangos with GOA?-Use of Gene Ontology Annotation (GOA) for biological interpretation of '-omics' data and for validation of automatic annotation tools. In Silico Biol 5:5–8.

Lee S, Lee BC, Kim D. 2006. Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins 62:1107–1114.

Liu W, Chou KC. 1998. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J Protein Chem 17:209–217.

Liu H, Wang M, Chou KC. 2005. Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336:737–739.

Lubec G, Afjehi-Sadat L, Yang JW, John JP. 2005. Searching for hypothetical proteins: Theory and practice based upon original data and literature. Prog Neurobiol 77:90–127.

Luo RY, Feng ZP, Liu JK. 2002. Prediction of protein strctural class by amino acid and polypeptide composition. Eur J Biochem 269:4219–4225.

Mahalanobis PC. 1936. On the generalized distance in statistics. Proc Natl Inst Sci India 2:49–55.

Mao B, Chou KC, Zhang CT. 1994. Protein folding classes: A geometric interpretation of the amino acid composition of globular proteins. Protein Eng 7:319–330.

Mardia KV, Kent JT, Bibby JM. 1979. Multivariate analysis: Chapter 11 discriminant analysis; chapter 12 multivariate analysis of variance; chapter 13 cluster analysis. London: Academic Press. pp 322–381.

Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, Akutsu T. 2005. A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci 14:2804–2813.

Metfessel BA, Saurugger PN, Connelly DP, Rich ST. 1993. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci 2:1171–1182.

Nakai K. 2000. Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 54:277–344.

Nakai K, Horton P. 1999. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–36.

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.

Nakashima H, Nishikawa K, Ooi T. 1986. The folding type of a protein is relevant to the amino acid composition. J Biochem 99:152–162.

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L. 2003. Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach. J Protein Chem 22:395–402.

Pillai KCS. 1985. Mahalanobis D2. In: Kotz S, Johnson NL, editors. Encyclopedia of statistical sciences. New York: John Wiley & Sons. This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics, pp 176–181.

Shen HB, Chou KC. 2005a. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Comm 337:752–756.

Shen HB, Chou KC. 2005b. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334:288–292.

Vapnik V. 1998. Statistical learning theory. New York: Wiley-Interscience.

Wang GL, Dunbrack RL, Jr. 2003. PISCES: A protein sequence culling server. Bioinformatics 19:1589–1591.

Wang M, Yang J, Xu ZJ, Chou KC. 2005. SLLE for predicting membrane protein types. J Theor Biol 232:7–15.

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC. 2006a. Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30:49–54.

Xiao X, Shao SH, Huang ZD, Chou KC. 2006b. Using pseudo amino acid composition to predict protein structural classes: Approached with complexity measure factor. J Comput Chem 27:478–482.

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY. 2006. Prediction protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature

extraction and naive Bayes feature fusion. Amino Acids 30:461–468.
Zhou GP. 1998. An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738.
Zhou GP, Assa-Munt N. 2001. Some insights into protein structural class prediction. Proteins Struct Funct Genet 44:57–59.
Zhou GP, Cai YD. 2006. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins Struct Funct Genet 63:681–684.
Zhou GP, Doctor K. 2003. Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Genet 50: 44–48.