Amino Acids (2007) 33: 57–67
DOI 10.1007/s00726-006-0478-8
Printed in The Netherlands

Amino Acids

Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction*

H.-B. Shen[1], **J. Yang**[1], and **K.-C. Chou**[1,2]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China
[2] Gordon Life Science Institute, San Diego, CA, U.S.A.

Received November 9, 2006
Accepted November 23, 2006
Published online January 19, 2007;

Summary. With the avalanche of newly-found protein sequences emerging in the post genomic era, it is highly desirable to develop an automated method for fast and reliably identifying their subcellular locations because knowledge thus obtained can provide key clues for revealing their functions and understanding how they interact with each other in cellular networking. However, predicting subcellular location of eukaryotic proteins is a challenging problem, particularly when unknown query proteins do not have significant homology to proteins of known subcellular locations and when more locations need to be covered. To cope with the challenge, protein samples are formulated by hybridizing the information derived from the gene ontology database and amphiphilic pseudo amino acid composition. Based on such a representation, a novel ensemble hybridization classifier was developed by fusing many basic individual classifiers through a voting system. Each of these basic classifiers was engineered by the KNN (K-Nearest Neighbor) principle. As a demonstration, a new benchmark dataset was constructed that covers the following 18 localizations: (1) cell wall, (2) centriole, (3) chloroplast, (4) cyanelle, (5) cytoplasm, (6) cytoskeleton, (7) endoplasmic reticulum, (8) extracell, (9) Golgi apparatus, (10) hydrogenosome, (11) lysosome, (12) mitochondria, (13) nucleus, (14) peroxisome, (15) plasma membrane, (16) plastid, (17) spindle pole body, and (18) vacuole. To avoid the homology bias, none of the proteins included has $\geq 25\%$ sequence identity to any other in a same subcellular location. The overall success rates thus obtained via the 5-fold and jackknife cross-validation tests were 81.6 and 80.3%, respectively, which were 40–50% higher than those performed by the other existing methods on the same strict dataset. The powerful predictor, named ''Euk-PLoc'', is available as a web-server at http://202.120.37.186/bioinf/euk. Furthermore, to support the need of people working in the relevant areas, a downloadable file will be provided at the same website to list the results predicted by Euk-PLoc for all eukaryotic protein entries (excluding fragments) in 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 the new entries of eukaryotic proteins and reflect the continuous development of Euk-PLoc.

Keywords: Cellular networking – Subcellular compartment – KNN classifier – Fusion – Voting – Gene ontology – Amphiphilic pseudo amino acid composition

1. Introduction

The cell is the basic structural and functional unit of all living organisms that is able to grow and reproduce independently. An adult human being is made up of approximately 100,000 billion (10^{14}) cells (Radford, 2003). Every cell contains approximately one billion (10^9) protein molecules that are located in many different compartments or organelles (Fig. 1), and perform a wide variety of activities in the cell. The organelles are specialized to carry out different tasks. For instance: the cell nucleus contains the genetic material (DNA) and thus governs all functions of the cell; cell membrane functions as a boundary layer to contain the cytoplasm, while cell wall provides protection from physical injury; the cytoplasm, a jelly-like material, fills the cell and serves as a ''molecular soup'' in which all of the cell's organelles are suspended; chloroplast is the site of photosynthesis; centriole forms spindle fibres to separate chromosomes during cell division; cytoskeleton is responsible for establishing cell shape, providing mechanical strength, locomotion, and intracellular transport of organelles; endoplasmic reticulum transports chemicals between cells and within cells; Golgi apparatus modifies chemicals to make them functional; lysosome breaks large molecules into small molecules by inserting a molecule of water into the chemical bond; mitochondrion is the ''power plant'' producing energy needed by the cell, and endoplasmic reticulum

* Electronic supplementary material: Supplementary material is available in the online version of this article at 10.1007/s00726-006-0478-8 and is accessible for authorised users.

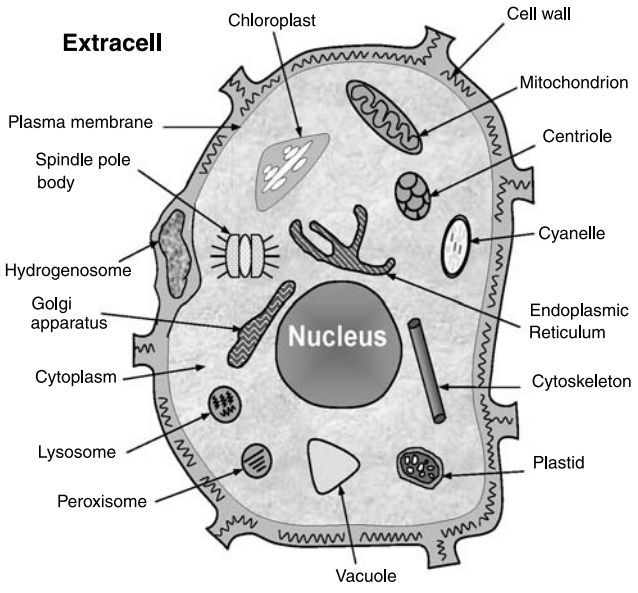

Fig. 1. Schematic illustration to show the 18 subcellular locations of eukaryotic proteins: (1) cell wall, (2) centriole, (3) chloroplast, (4) cyanelle, (5) cytoplasm, (6) cytoskeleton, (7) endoplasmic reticulum, (8) extracell, (9) Golgi apparatus, (10) hydrogenosome, (11) lysosome, (12) mitochondria, (13) nucleus, (14) peroxisome, (15) plasma membrane, (16) plastid, (17) spindle pole body, and (18) vacuole

is, together with the ribosome, responsible for synthesizing proteins; and peroxisome breaks down excess fatty acids and hydrogen peroxide (H_2O_2), a potentially dangerous product of fatty-acid oxidation. Most of these functions, which are critical to the cell's survival, are performed by the proteins in a cell. 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. Accordingly, the significance to identify the subcellular localization of an uncharacterized protein has become self-evident.

Although the information about protein subcellular localization can be determined by conducting various experiments, that is both time-consuming and costly. Particularly, the number of newly-found protein sequences has increased explosively in the post genomic era. For instance, according to the Swiss-Prot database (Bairoch and Apweiler, 2000) version 50.0 released on 30-May-2006 the number of total eukaryotic protein entries is 95,502. After excluding those annotated as ''fragment'' or containing less than 50 amino acid residues, the number is reduced to 48,341, of which 21,275 are with subcellular location annotations (Item 1 of Table 1). However, of the 21,275 proteins, 6,704 are annotated with experimental observations (Item 2 of Table 1) and 14,571 annotated with uncertain labels such as ''probable'', ''potential'', ''perhaps'', and ''by similarity'' (Item 3 of Table 1). The uncertain annotations cannot be used as robust data for training a solid predictor. Actually, proteins with uncertain annotations also belong to the targets of identification either by newly developed predictors or by further experiments.

A similar gap also exists in the gene ontology (GO) database (Ashburner et al., 2000), which was established according to molecular function, biological process, and cellular component. As shown in Item 5 of Table 1, of the 48,341 eukaryotic proteins, only 26,000 have GO annotations to indicate their subcellular components. Moreover, it is instructive to point out that the GO database was derived from various other databases, including Swiss-Prot database. Therefore, the GO annotations might be contaminated by the uncertain information from the 14,571 entries as indicated in Item 3 of Table 1.

Therefore, the number of eukaryotic proteins that have reliable subcellular location annotations is 6,704 (Item 2 of Table 1), which is about 14% of all the eukaryotic

Table 1. Breakdown of the 48,341[a] eukaryotic protein entries from Swiss-Prot database (version 50.0, released 30-May-2006) according to the nature of their subcellular location annotation and their expression in GO database (released on 4-March-2006)

Item	Description	Number	Percentage
1	Proteins with subcellular location annotations in Swiss-Prot database	21,275	$\frac{21275}{48341} = 44.0\%$
2	Proteins in Item 1 with experimentally observed subcellular locations	6,704	$\frac{6704}{48341} = 13.9\%$
3	Proteins in Item 1 with uncertain terms, such as ''potential'', ''probable'', and ''by similarity''	14,571	$\frac{14571}{48341} = 30.1\%$
4	Proteins that can be represented in the GO space (cf. Eq. 3)	44,274	$\frac{44274}{48341} = 91.6\%$
5	Proteins with subcellular component annotations in GO database	26,000	$\frac{26000}{48341} = 53.8\%$

[a] The original eukaryotic protein entries was 95,502, of which 47,161 were either annotated as ''fragment'' or with less than 50 amino acid residues, and hence were removed for further consideration

protein entries concerned. In other words, there are (48,341–6,704) = 41,637 eukaryotic proteins for which the subcellular localization needs to be identified or further confirmed.

With the rapidly increase of gene products in the post-genomic era, it is expected that the gap between the newly-found protein sequences and the knowledge of their subcellular localization will be continuously enlarged. To timely use these new proteins for basic research and drug discovery (Chou, 2004; Lubec et al., 2005), it is highly desired to develop an effective method to bridge such gap, and the present study was initiated in an attempt to address the challenge with a focus on eukaryotic proteins.

Actually, many methods have been developed in this regard (Cedano et al., 1997; Chou, 2000a; Chou and Cai, 2002; Chou and Elrod, 1999; Chou and Shen, 2006; Feng, 2001, 2002; Gao et al., 2005a, b; Garg et al., 2005; Guo et al., 2006a; Matsuda et al., 2005; Nakai, 2000; Nakai and Horton, 1999; Nakashima and Nishikawa, 1994; Reinhardt and Hubbard, 1998; Xiao et al., 2005; Zhou and Doctor, 2003). However, the datasets used to train these predictors cover very limited subcellular locations. For instance, the datasets in Nakashima and Nishikawa (1994) only cover two locations; those in Cedano et al. (1997), 5 locations; those in Garg et al. (2005), 4 locations; and those in Matsuda et al. (2005), 4 locations. Although the datasets in Chou and Elrod (1999) and Park and Kanehisa (2003) expanded the coverage to 12 subcellular locations, they were constructed with a very tolerant criterion that allowed inclusion of those proteins with sequence identity up to 80% to each other. This will certainly result in overestimating a predictor due to the homology bias.

To enlarge the coverage scope and avoid the homology bias, a much more extensive and stringent dataset is needed. To realize this, a new dataset was constructed that covers 18 subcellular locations, with a stringent criterion that none of proteins included has $\geq$25% sequence identity to any other within a same subcellular location. As is well known, the more stringent criterion is imposed to exclude homologous proteins from a benchmark dataset, the more difficult would be to get a higher success rate. Also, the more the number of subcellular locations covered, the lower the odds are in getting a successful prediction. To enhance the success rate under such two strict conditions, here a new approach, which is completely different from the aforementioned methods, is introduced. Below, let us first construct the new benchmark datasets.

2. Materials and methods

Protein sequences

Protein sequences were collected from the Swiss-Prot database release 50.0 (30-May-2006) at http://www.ebi.ac.uk/swissprot/ for eukary-

Table 2. Keywords used to search the Swiss-Prot database for known subcellular locations

Subcellular location	Keywords
cell wall	cell wall
centriole	centriole; centrosome; centromer
chloroplast	chloroplast
cyanelle	cyanelle
cytoplasm	cytoplasm; cytoplasmic
cytoskeleton	cytoskeleton; filament; microtubule
endoplasmic reticulum	endoplasmic reticulum
extracell	extracell; extracellular; secreted
Golgi apparatus	Golgi
hydrogenosome	hydrogenosome
lysosome	lysosome; lysosomal
mitochondrion	mitochondrion; mitochondria; mitochondrial
nucleus	nucleus; nuclear
peroxisome	peroxisome; peroxisomal; microsome; glyoxysomal; glycosomal
plasma membrane	plasma membrane; integral membrane
plastid	plastid
spindle pole body	spindle pole; spindle pole body
vacuole	vacuole; vacuolar

Table 3. Number of proteins in each of the 18 subcellular locations (Fig. 1) constructed in this paper

Subcellular location	Number of proteins
(1) cell wall	25
(2) centriole	21
(3) chloroplast	258
(4) cyanelle	97
(5) cytoplasm	718
(6) cytoskeleton	25
(7) endoplasmic reticulum	113
(8) extracell	806
(9) Golgi apparatus	85
(10) hydrogenosome	10
(11) lysosome	46
(12) mitochondrion	228
(13) nucleus	1169
(14) peroxisome	64
(15) plasma membrane	413
(16) plastid	38
(17) spindle pole body	15
(18) vacuole	44
Total	4175

None of the proteins included in the current benchmark dataset has $\geq$25% sequence identity to any of the proteins in the same subset. The corresponding accession numbers and sequences are given in the "Electronic Supplementary Material" of this paper (see http://dx.doi.org/10.1007/s00726-006-0478-8)

otic proteins according to their experimentally annotated subcellular locations. In order to obtain high-quality, well-defined working datasets, the data were collected strictly according to the following procedures: (a) To deal with the situation that a same location might be annotated with different terms, the keywords listed in Table 2 were used to search against the categorization of subcellular locations. (b) Sequences annotated with ambiguous or uncertain words, such as ''potential'', ''probable'', ''probably''', ''maybe'', or ''by similarity'', were excluded. (c) Sequences annotated by two or more locations were not included because of lack of the uniqueness; also, sequences annotated with ''prokaryotic'' were excluded because this study was focused on eukaryotic proteins only. (d) Proteins annotated with ''fragment'' were excluded; also, sequences with less than 50 amino acid residues were removed because they might just be fragments. (e) To avoid any homologous bias, a redundancy cutoff was operated by a culling program (Wang and Dunbrack Jr., 2003) to exclude those sequences which have $\geq$25% sequence identity to any other in a same subcellular location. (f) Those subcellular locations (subsets) which contain less than 10 protein sequences were left out because of lacking statistical significance.

After strictly following the above procedures, we finally obtained 4,175 protein sequences classified into the following 18 subcellular locations: cell wall, centriole, chloroplast, cyanelle, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, hydrogenosome, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, plastid, spindle pole body, and vacuole (Fig. 1). The number of proteins thus obtained for each of the 18 subcellular locations is given in Table 3. The accession numbers and sequences for the 4,175 proteins are given in the Online Supporting Information A.

Prediction method

The key to enhance the prediction quality for protein subcellular location is to grip the core features of a protein that are intimately related to its localization in a cell. Accordingly, the source of gene ontology (GO) consortium (Ashburner et al., 2000) can be used as a vehicle to formulate the prediction algorithm.

However, how to effectively use the GO database to improve the prediction quality for protein subcellular location is by no means a trivial problem (Camon et al., 2004; Lee et al., 2005). The reasons are as follows: (a) 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 protein with accession number O22892 and O00093 in Table 4). (b) Even for those proteins whose subcellular locations are clearly annotated in Swiss-Prot database, their corresponding GO numbers in GO database are not always directly indicating their corresponding subcellular locations; in some cases they are even annotated with ''cellular component unknown''. For example, for the protein with accession number O43303 in Table 4, its subcellular location is annotated with ''centriole'' in Swiss-Prot database, but none of its GO numbers indicates its subcellular location. Similar situations occur for the proteins with accession numbers P19877, Q29593, and Q9UIV8 as well (Table 4). (c) More important, it should be emphasized that during the cross-validation test for the current approach, only the GO numbers of a query protein but not its GO annotations are used, just like the case in testing all the previous predictors that only the sequence of a query protein but not its Swiss-Prot annotation is used; otherwise, the results obtained by the cross-validation test would not represent any prediction potential at all.

Table 4. 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
O22892	No subcellular location annotated	GO:0000004	Biological process unknown
		GO:0005554	Molecular function unknown
		GO:0008372	Cellular component unknown
O00093	No subcellular location annotated	GO:0003993	Acid phosphatase activity
		GO:0016158	3-phytase activity
		GO:0016787	Hydrolase activity
O43303	Centriole	GO:0000004	Biological process unknown
		GO:0005554	Molecular function unknown
		GO:0008372	Cellular component unknown
P19877	Extracellular	GO:0006935	Chemotaxis
		GO:0008009	Chemokine activity
		GO:0008083	Growth factor activity
		GO:0008372	Cellular component unknown
Q29593	Cytoplasm	GO:0004801	Transaldolase activity
		GO:0005975	Carbohydrate metabolism
		GO:0006098	Pentose-phosphate shunt
		GO:0008372	Cellular component unknown
		GO:0016740	Transferase activity
Q9UIV8	Cytoplasm	GO:0004866	Endopeptidase inhibitor activity
		GO:0004867	Serine-type endopeptidase inhibitor activity
		GO:0008372	Cellular component unknown
		GO:0009411	Response to UV
		GO:0030162	Regulation of proteolysis

(d) As mentioned above, the GO database was derived from other databases including the Swiss-Prot database, and the annotations in GO might be contaminated by the uncertain annotations from Swiss-Prot (Item 3 of Table 1). Therefore, the subcellular component annotations from GO cannot be directly used to form a solid training dataset. It must be formed from the 6,704 eukaryotic protein sequences in Swiss-Prot database that are annotated with experimentally observed subcellular locations (see Item 2 of Table 1).

The information that may be useful for predicting subcellular locations of proteins are actually "buried" into a series of tedious GO numbers, just like they are "buried" into a pile of complicated amino acid sequences, although the manner and the "depth" they are "buried" are quite different. In view of this, the key problem is how to represent protein samples thru the tedious GO numbers, just like the efforts by many previous investigators in extracting various features from the complicated sequences to represent protein samples. The following approach was developed for such a purpose that is very important for effectively using complicated and tedious data to derive the desired results.

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 30-May-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 3. On the other hand, because the current GO database is not complete yet, some protein entries (such as "P54661", "Q9I969", and "Q09874") have no corresponding GO numbers, i.e., 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:0051990 would become GO_compress:0000001, GO_compress:0000002, GO_compress:0000003, GO_compress:0000004, GO_compress:0000005, ..., and GO_compress:0010173, respectively. The GO database thus obtained is called GO_compress database, whose dimensions were reduced from 51,990 in the original GO database to 10,173. Each of the 10,173 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 10173-D (dimensional) GO_compress space is assigned 1; otherwise, 0. Thus, the protein can be formulated as:

$$\mathbf{P} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_i \\ \vdots \\ g_{10173} \end{bmatrix} \tag{1}$$

where

$$g_i = \begin{cases} 1, & \text{hit found in GO\_compress} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

(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, as given below

$$\mathbf{P} = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+2\lambda} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{\Lambda} \end{bmatrix}, \quad (\Lambda = 20 + 2\lambda) \tag{3}$$

where $p_1, p_2, \ldots, p_{20}$ are associated with the amino acid composition reflecting the occurrence frequencies of the 20 native amino acids in the protein (Chou and Zhang, 1994; Nakashima et al., 1986), and $p_{20+1}, p_{20+2}, \ldots, p_{20+2\lambda}$ are the 2λ correlation factors that reflect its sequence-order pattern through the amphiphilic feature (Chou and Cai, 2005; Chou et al., 1997). The protein representation as defined by Eq. (3) is called the "amphiphilic pseudo amino acid composition", which is one kind of the pseudo amino acid composition (PseAA) originally introduced by Chou (Chou, 2001) and has the same form as the conventional amino acid composition but contains more components and sequence information. For a given protein sequence and the value of λ, the $20 + 2\lambda$ elements in Eq. (3) can be easily computed by following Eqs. (2–6) of Chou (2005).

Suppose there are N proteins $(\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N)$ which have been classified into 18 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; Denoeux, 1995; Keller et al., 1985) to deal with this problem. According to the KNN rule, the query protein should be assigned to the subset represented by a 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 (Chou, 1995; Mahalanobis, 1936; Pillai, 1985). Here, we use the following equation to measure the nearness between protein $\mathbf{P}$ and $\mathbf{P}_i$

$$D(\mathbf{P}, \mathbf{P}_i) = 1 - \frac{\mathbf{P} \cdot \mathbf{P}_i}{\|\mathbf{P}\| \|\mathbf{P}_i\|} \tag{4}$$

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 Eq. (4), when $\mathbf{P} \equiv \mathbf{P}_i$ we have $D(\mathbf{P}, \mathbf{P}_i) = 0$, indicating the "distance" between these two proteins is zero and hence they have 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 Eq. (4). 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 through 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 $K \geq 7$.

Because the predicted results by the KNN algorithm (Cover and Hart, 1967; Denoeux, 1995; Keller et al., 1985) 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 **P**.

During the course of prediction, the following self-consistency principle should be followed. If a query protein could be defined in the 10173-D GO_compress space (Eq. 1), then the prediction should be carried out based on those proteins in the training dataset that could be defined in the same 10173-D space. If the query protein in the 10173-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. 3), 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 subpredictors: (a) the NN(K)-GO predictor that operates in the 10173-D GO_compress space, and (b) the NN(K,Λ)-PseAA predictor that operates in the Λ-D amphiphilic PseAA space. The predicted results by the latter also depend on Λ, the dimension of the PseAA (see Eqs. 3–4). Therefore, even for exactly the same training dataset and same fixed value of K, using different values of Λ will yield different results. Generally speaking, the more components the PseAA contains, the more information it carries. However, it will reduce the cluster-tolerant capacity if the PseAA contains too many components, so as to lower down the success rate of cross validation. To get the optimal result, two different ensemble classifiers were introduced for NN(K)-GO predictor and NN(K,Λ)-PseAA predictor, respectively, as formulated below.

For the NN(K)-GO predictor, the ensemble classifier was formed by fusing many single classifiers each having a different specified value for K, as described below.

Preliminary tests indicated that the success rates obtained by the NN(K)-GO predictor were lower when $K=1$, 2, or >10, and hence these numbers can be ignored. Let us suppose

$$\{\mathrm{K}\} = \{3, 4, \ldots, 10\} \tag{5}$$

represent a set of possible numbers for K, then we have a set of corresponding classifiers as formulated by

$$\{\mathrm{NN(K)}\} = \{\mathrm{NN}(3), \mathrm{NN}(4), \ldots, \mathrm{NN}(10)\} \tag{6}$$

where NN(3) is the NN classifier trained with 3 nearest neighbors in the 10173-D GO_compress space, NN(4) is the one trained with 4 nearest neighbors, and so forth. The ensemble classifier formed by fusing such a set of 8 individual classifiers is formulated by

$$\mathbb{C}^{\mathrm{GO}} = \mathrm{NN}(3) \oplus \mathrm{NN}(4) \oplus \cdots \oplus \mathrm{NN}(10) \tag{7}$$

where the symbol $\oplus$ denotes the fusion operator, and $\mathbb{C}^{\mathrm{GO}}$ the ensemble classifier formed by fusing NN(3), NN(4), …, and NN(10).

The process of how the ensemble classifier $\mathbb{C}^{\mathrm{GO}}$ works is as follows. Suppose the predicted classification results for the query protein **P** by the 8 individual classifiers in Eq. (7) are $\mathrm{Q}_3^{\mathrm{GO}}$, $\mathrm{Q}_4^{\mathrm{GO}}, \ldots, \mathrm{Q}_{10}^{\mathrm{GO}}$, respectively; i.e.,

$$\{\mathrm{Q}_3^{\mathrm{GO}}, \mathrm{Q}_4^{\mathrm{GO}}, \ldots, \mathrm{Q}_{10}^{\mathrm{GO}}\} \in \{S_1, S_2, \ldots, S_{18}\} \tag{8}$$

where $\in$ is a symbol in the set theory meaning "member of", $S_1, S_2, \cdots, S_{18}$ represent the 18 subsets defined by the 18 subcellular locations studied here (Fig. 1), and the voting score for the protein **P** belonging to the u-th subset is defined by

$$Y_u^{\mathrm{GO}} = \sum_{i=3}^{10} \delta(\mathrm{Q}_i^{\mathrm{GO}}, S_u), \quad (u = 1, 2, \ldots, 18) \tag{9}$$

where the delta function in Eq. (9) is given by

$$\delta(\mathrm{Q}_i^{\mathrm{GO}}, S_u) = \begin{cases} 1, & \text{if } \mathrm{Q}_i^{\mathrm{GO}} \in S_u \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

thus the query protein **P** is predicted belonging to the subset (subcellular location) with which its score of Eq. (9) is the highest.

The above fusion process can be straightforwardly extended to the case of the NN(K,Λ)-PseAA predictor. However, since it has two parameters, the ensemble classifier for NN(K,Λ)-PseAA should be formed by fusing many single classifiers with different K or Λ, respectively; i.e., the fusion process should involve a two dimensional process, as formulated below.

For the similar reason as mentioned above regarding Eq. (5), we can suppose

$$\{\mathrm{K}\} = \{3, 4, \ldots, 10\}; \qquad \{\Lambda\} = \{22, 24, \ldots, 58, 60\} \tag{11}$$

represent two sets of possible numbers for K and Λ, then we have a set of $8 \times 20 = 160$ individual classifiers as formulated by

$$\{\mathrm{NN(K,}\Lambda)\} = \begin{Bmatrix} \mathrm{NN}(3,22) & \mathrm{NN}(3,24) & \cdots & \mathrm{NN}(3,60) \\ \mathrm{NN}(4,22) & \mathrm{NN}(4,24) & \cdots & \mathrm{NN}(4,60) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{NN}(10,22) & \mathrm{NN}(10,24) & \cdots & \mathrm{NN}(10,60) \end{Bmatrix} \tag{12}$$

where NN(3,22) is the NN classifier trained with 3 nearest neighbors in the 22-D PseAA space, NN(4,24) is the one trained with 4 nearest neighbors in the 24-D PseAA space, and so forth. The ensemble classifier formed by fusing such 160 individual classifiers is formulated by

$$\mathbb{C}^{\mathrm{Pse}} = \mathrm{NN}(3,22) \oplus \mathrm{NN}(3,24) \oplus \cdots \oplus \mathrm{NN}(10,58) \oplus \mathrm{NN}(10,60) \tag{13}$$

where the fusion operator $\oplus$ has the same meaning as that of Eq. (7).

The process of how the ensemble classifier $\mathbb{C}^{\mathrm{Pse}}$ works is as follows. Suppose the predicted classification results for the query protein **P** by the 160 individual classifiers in Eq. (13) are

$$\begin{Bmatrix} \mathrm{Q}_{3,22}^{\mathrm{Pse}} & \mathrm{Q}_{3,22}^{\mathrm{Pse}} & \cdots & \mathrm{Q}_{3,22}^{\mathrm{Pse}} \\ \mathrm{Q}_{4,24}^{\mathrm{Pse}} & \mathrm{Q}_{4,24}^{\mathrm{Pse}} & \cdots & \mathrm{Q}_{4,24}^{\mathrm{Pse}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Q}_{10,60}^{\mathrm{Pse}} & \mathrm{Q}_{10,60}^{\mathrm{Pse}} & \cdots & \mathrm{Q}_{10,60}^{\mathrm{Pse}} \end{Bmatrix} \in \{S_1, S_2, \ldots, S_{18}\} \tag{14}$$

where $S_1, S_2, \ldots, S_{18}$ have the same meaning as in Eq. (8), i.e., represent the 18 subsets defined by the 18 subcellular locations studied here (Fig. 1), and the voting score for the protein **P** belonging to the u-th subset is defined by

$$Y_u^{\mathrm{Pse}} = \sum_{i=3}^{10} \sum_{j=11}^{30} \delta(\mathrm{Q}_{i,2j}^{\mathrm{Pse}}, S_u), \quad (u = 1, 2, \ldots, 18) \tag{15}$$

where the delta function in Eq. (15) is given by

$$\delta(\mathrm{Q}_{i,2j}^{\mathrm{Pse}}, S_u) = \begin{cases} 1, & \text{if } \mathrm{Q}_{i,2j}^{\mathrm{Pse}} \in S_u \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

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

The predictor thus found by hybridizing the fusion classifier $\mathbb{C}^{\mathrm{GO}}$ Eq. (7) and fusion classifier $\mathbb{C}^{\mathrm{Pse}}$ Eq. (13) is called "Euk-PLoc".

Finally, it should be pointed out that, although using GO database to predict protein subcellular location has been explored by previous investigators (Chou and Cai, 2003, 2004), the predictors formulated there has much less power than the current predictor owing to the following reasons. (a) The GO approach in (Chou and Cai, 2003, 2004) was operated by the nearest neighbor rule with $K=1$ only, which is corresponding to NN(1) according to the symbol used in this paper. As mentioned above, the success rate obtained by NN(1) is much lower than the rate by NN(3), NN(4), …, or NN(10), needless to say the rate by the ensemble classifier $\mathbb{C}^{\mathrm{GO}}$ formed by fusing these classifiers as formulated in Eq. (7). A similar difference also exists for the PseAA predictor part, in which the current approach is even much more sophisticated because it involves a 2-dimensional fusion problem as formulated by Eq. (13). (b) The dimension of

the GO database space in Chou and Cai (2003, 2004) is 1930, but the dimension of GO database space here is 10173, indicating the need to catch up with the rapid development in GO. (c) The benchmark dataset used in Chou and Cai (2003) only covers 4 subcellular locations with a 90% sequence identity cutoff, the dataset used in Chou and Cai (2004) covers 12 subcellular locations with an 80% sequence identity cutoff, while the dataset used here covers 18 subcellular locations with a 25% sequence identity cutoff. Besides, it is through Tables 1 and 4 presented here that the relationship between GO and Swiss-Prot is more clearly elucidated than in the previous papers (Chou and Cai, 2003, 2004).

3. Results and discussion

For the proteins listed in the ''Electronic Supplementary Material'' of this paper, we obtained the following results according to Steps 1–2 of Materials and methods. Of the 4175 proteins in the current benchmark dataset, 4102 got hits in the GO_compress database, and hence were defined in the 10173-D GO_compress space (Eqs. 1–2), and the remainder defined in the Λ-D PseAA space (Eq. 3). Therefore, if the protein samples were represented only based on the GO_compress database, $4175–4102 = 73$ proteins would have no definition, leading to a failure of utilizing their information. Although most of proteins studied can be defined in the 10173-D GO_compress space, it is better to hybridize with the PseAA approach, by which not only a protein can always be defined but also a considerable amount of sequence-order information can be incorporated. Thus, the prediction process was operated according to the following procedures: if a query protein was defined in the 10173-D GO_compress space, then the ensemble classifier $\mathbb{C}^{\mathrm{GO}}$ was used to predict its subcellular location; otherwise, the ensemble classifier $\mathbb{C}^{\mathrm{Pse}}$ was used to predict its subcellular location.

The prediction quality was examined by the 5-fold cross-validation test and jackknife test, respectively. 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. As we can see from the above description, the jackknife test will take more computational time, especially for the large-scale dataset. Therefore, many chose to use the 5-fold cross-validation, in which the benchmark set is randomly divided into 5 subsets, each containing nearly equal number of proteins. These subsets are then grouped into a training set and a testing set. The training set consists of 4 of the 5 subsets and the testing set consists of the remaining one. This procedure is repeated 5 times and each of the subsets is used once for testing. The final prediction results are the average of the five testing results. However, the results obtained by the 5-fold cross-validation test have some sort of arbitrariness, as discussed in Chou and Zhang (1995). In statistical prediction, the independent dataset test, 5-fold test, and jackknife test are the three methods often used to cross-validate the power of a predictor. Among these three, the jackknife test is deemed the most rigorous and objective one [see (Chou and Zhang, 1995) for a comprehensive review about this], and hence have been increasingly used by investigators in examining the power of various prediction methods (Cao et al., 2006; Chen et al., 2006; Chou, 2000b; Du et al., 2006; Feng, 2001, 2002; Gao et al., 2005a, b; Guo et al., 2006a, b; Liu et al., 2005a, b; Luo et al., 2002; Shen and Chou, 2005; Shen et al., 2005, 2006; Sun and Huang, 2006; Wang et al., 2004, 2005a, b, 2006; Wen et al., 2006; Xiao et al., 2005, 2006a, b; Zhang et al., 2006; Zhou, 1998; Zhou and Assa-Munt, 2001).

The predicted results obtained by Euk-PLoc are given in Table 5, from which we can see that the overall success rate by 5-fold cross-validation test is 81.6%, and that by the jackknife test 80.3%. For such a stringent and wide-coverage benchmark dataset, if using the other methods, such as the HSLPred (Garg et al., 2005) and the SVM-based method (Matsuda et al., 2005), the corresponding overall success rates were in the range of 30–40%, much lower than the rates obtained by Euk-PLoc. From Table 5 we can also see that the overall success rate obtained by the jackknife test is 1.3% lower than that by the 5-fold cross-validation test, which is expected because the jackknife test is more strict than the 5-fold cross-validation test as mentioned above.

Why the methods reported in Garg et al. (2005) and Matsuda et al. (2005) could yield an overall success rate

Table 5. Overall success prediction rates by Euk-PLoc on the 4175 eukaryotic proteins[a] among the 18 subcellular locations (Fig. 1) by 5-fold cross-validation and jackknife test

Test method	Overall success rate
5-fold cross validation	$\frac{3406}{4175} = 81.6\%$
Jackknife	$\frac{3352}{4175} = 80.3\%$

[a] The accession numbers and sequences are given in the ''Electronic Supplementary Material'' of this paper available under http://dx.doi.org/10.007/s00726-006-0478-8, where none of proteins included has $\geq$25% sequence identity to any other in a same subset

Table 6. Comparison between Euk-PLoc and MultiLoc in performing the 5-fold cross-validation test on the dataset constructed by Hoglund et al. (2006)

Subcellular location	MCC (Matthews, 1975)		Overall success rate	
	Euk-PLoc[a]	MultiLoc[b]	Euk-PLoc[a]	MultiLoc[b]
(1) chloroplast	0.91	0.85	90.8%	75%
(2) cytoplasm	0.84	0.69		
(3) endoplasmic reticulum	0.78	0.61		
(4) extracell	0.88	0.75		
(5) Golgi apparatus	0.93	0.56		
(6) lysosome	0.76	0.48		
(7) mitochondrion	0.91	0.83		
(8) nucleus	0.95	0.73		
(9) peroxisome	0.85	0.43		
(10) plasma membrane	0.95	0.77		
(11) vacuole	0.78	0.42		

[a] Predictor proposed in this paper
[b] Predictor proposed in Hoglund et al. (2006)

higher than 80%, but here these methods only yielded a success rate within the range of 30–40%? The reasons are as follows: (a) The benchmark datasets originally used by these authors contained many homologous sequences in a same subcellular location. For example, the dataset used in Garg et al. (2005) and Matsuda et al. (2005) contained proteins with up to 90% sequence identity. When predictions were made by their methods on the current stringent dataset in which none of protein has $\geq$25% sequence identity to any other in a same subcellular location, the success rates would of course decrease significantly. (b) The success rate originally reported in Garg et al. (2005) was derived from the prediction among cytoplasm, mitochondria, nuclear, and plasma membrane, while that in Matsuda et al. (2005) derived among cytoplasm, extracellular, mitochondria, and nuclear. The benchmark datasets used in both cases cover only 4 subcellular locations, in contrast to 18 as in the current stringent dataset. Let us imagine: if the protein samples are completely randomly distributed among 4 possible locations, the overall success rate by random assignments would generally be $1/4 = 25\%$; however, the corresponding rate would be reduced to $1/18 \simeq 5.6\%$ if the protein samples are distributed among 18 possible locations. That is why the more the number of subcellular locations covered, the lower the odds are in getting a higher success rate. (c) As a further demonstration, let us compare Euk-Ploc with another SVM-based MultiLoc predictor (Hoglund et al., 2006). MultiLoc used a benchmark dataset including 5,959 proteins and covering 11 eukaryotic subcellular locations, i.e. chloroplast, cytoplasmic, endoplasmic reticulum, extracellular, Golgi, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane, and vacuolar. Their dataset allows inclusion of protein sequences with up to 80% sequence identity to one another. The dataset can be downloaded from http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc/information. As reported in Hoglund et al. (2006), the overall success rate obtained by MultiLoc using 5-fold cross-validation test was about 75%. To make the comparison under exactly the same condition, let us also use the same dataset and the same 5-fold cross-validation procedure to test Euk-PLoc. As shown in Table 6, not only the overall success rate obtained by Euk-PLoc is significantly higher than that by MultiLoc, but the Matthews correlation coefficient (MCC) obtained by Euk-PLoc for each of the eleven subcellular locations is also remarkably higher than that by MultiLoc, indicating that Euk-PLoc is much more reliable and stable.

4. Conclusion

Prediction of protein subcellular location is an important but meanwhile very difficult problem. 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 for us to get a higher success prediction rate. That is why for the benchmark dataset investigated here, which involves 18 subcellular locations and in which none of protein has $\geq$25% sequence identity to any others in a same subcellular location, the success rates obtained by various

powerful existing methods were only within the range of 30–40%.

To improve the prediction quality, we adopted the strategy of (a) representing protein samples by hybridizing GO (Eq. 1) and PseAA (Eq. 3), and (b) introducing the ensemble classifier that were formed by fusing many basic individual classifiers operated according to the nearest neighbor rule. Using GO to represent the sample of a protein could effectively grasp its core features. However, it was by no means a trivial use of the GO annotations to assign the subcellular locations of query proteins because 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''. As a matter of fact, the information useful for predicting subcellular locations of proteins are actually ''buried'' into a series of GO numbers, just like they are ''buried'' into a pile of complicated amino acid sequences. Also, because the GO database is not complete yet, many proteins might not be meaningfully represented in the GO system. For these proteins, the PseAA representation was used because it could incorporate a considerable amount of sequence order effects and yield better predicted results than the conventional amino acid composition representation.

In using the nearest neighbor rule, selecting different number (K) of the nearest neighbors counted for prediction, or using different dimension (Λ) of PseAA to represent protein samples, might lead to a different result. It is both tedious and time-consuming to find the optimal result by testing different values of K and Λ one by one. The ensemble classifier $\mathbb{C}^{\mathrm{GO}}$ formed by fusing a set of basic classifiers with different K and the ensemble classifier Λ^{Pse} formed by fusing a set of basic classifiers with different K and Λ can automatically solve the problem. Actually, it was observed that the result predicted by the ensemble classifier was better than the best of those by the individual classifiers. That is why the approach by hybridizing the ensemble classifiers $\mathbb{C}^{\mathrm{GO}}$and $\mathbb{C}^{\mathrm{Pse}}$ is so powerful, yielding over 80% success rates even for the current stringent datasets in which none of proteins has $\geq$25% sequence identity to any other in a same subcellular location. These rates are 40–50% higher than those by the other existing approaches. It is anticipated that the powerful approach may become a useful high throughput tool for many other relevant area in bioinformatics, proteomics, and molecular biology.

A web-server has been designed for the powerful predictor Euk-PLoc, and it is freely available at http://202.120.37.186/bioinf/euk to the public. Moreover, for the convenience of people who are working in the relevant fields, a downloadable file will be provided at the same website to list the results predicted by Euk-PLoc for all eukaryotic protein entries in Swiss-Prot database that are not fragments and that have no subcellular location annotations or are annotated with uncertain terms such as ''probable'', ''likely'', or ''by similarity''.

References

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. Nature Genet 25: 25–29

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

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, doi: 10.1186/1471-2105-7-20

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

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

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 (2000a) Review: prediction of protein structural classes and subcellular locations. Curr Protein Peptide Sci 1: 171–208

Chou KC (2000b) Review: prediction of tight turns and their types in proteins. Anal Biochem 286: 1–16

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

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 (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun 311: 743–747

Chou KC, Cai YD (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Biophys Res Commun 320: 1236–1239

Chou KC, Cai YD (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inform Model 45: 407–413

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

Chou KC, Shen HB (2006) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

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, Zhang CT, Maggiora GM (1997) Disposition of amphiphilic helices in heteropolar environments. Proteins Struct Funct Genet 28: 99–108

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

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

Du QS, Jiang ZQ, He WZ, Li DP, Chou KC (2006) Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. J Biomol Struct Dyn 23: 635–640

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

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005b) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

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 J, Lin Y, Liu X (2006a) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6: 5099–5105

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

Hoglund A, Donnes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22: 1158–1165

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

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

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

Liu H, Yang J, Ling JG, Chou KC (2005b) Prediction of protein signal sequences and their cleavage sites by statistical rulers. Biochem Biophys Res Commun 338: 1005–1011

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

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis chapter 11: Discriminant analysis; chapter 12: Multivariate analysis of variance; chapter 13: Cluster analysis. Academic Press, London 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

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451

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

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. Bioinformatics 19: 1656–1663

[#] Pillai KCS (1985) Mahalanobis D2. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, Vol 5. Wiley, New York, pp 176–181

Radford T (2003) Metaphors and dreams. The Scientist 17: 24–26

Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 26: 2230–2236

Shen HB, Chou KC (2005) 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

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240: 9–13

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334: 577–581

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

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

Wang M, Yang J, Chou KC (2005a) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28: 395-402 (Erratum, ibid. 2005, 29: 301)

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Des Select 17: 509–516

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

Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J Theor Biol 242: 941–946

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for

[#] This reference also presents a brief biography of Mahalanobis, who was a man of great originality and who made considerable contributions to statistics.

G-protein coupled receptors classification and fast structure recognition. Amino Acids (in press) (DOI: 10.1007/s00726-006-0341-y)
Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61
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 Prot 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, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Genet 50: 44–48

Authors' address: Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, U.S.A.,
Fax: +1-858-484-1018, E-mail: kchow@san.rr.com