Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 347 (2006) 150–157
BBRC
www.elsevier.com/locate/ybbrc

Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization

Kuo-Chen Chou [a,b,*], Hong-Bin Shen [b]

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

Received 6 June 2006
Available online 21 June 2006

Abstract

Predicting subcellular localization of human proteins is a challenging problem, especially when unknown query proteins do not have significant homology to proteins of known subcellular locations and when more locations need to be covered. To tackle the challenge, protein samples are expressed by hybridizing the gene ontology (GO) database and amphiphilic pseudo amino acid composition (PseAA). Based on such a representation frame, a novel ensemble classifier, called "Hum-PLoc", was developed by fusing many basic individual classifiers through a voting system. The "engine" of these basic classifiers was operated by the KNN (*K*-nearest neighbor) rule. As a demonstration, tests were performed with the ensemble classifier for human proteins among the following 12 locations: (1) centriole; (2) cytoplasm; (3) cytoskeleton; (4) endoplasmic reticulum; (5) extracell; (6) Golgi apparatus; (7) lysosome; (8) microsome; (9) mitochondrion; (10) nucleus; (11) peroxisome; (12) plasma membrane. To get rid of redundancy and homology bias, none of the proteins investigated here had $\geqslant 25\%$ sequence identity to any other in a same subcellular location. The overall success rates thus obtained via the jackknife cross-validation test and independent dataset test were 81.1% and 85.0%, respectively, which are more than 50% higher than those obtained by the other existing methods on the same stringent datasets. Furthermore, an incisive and compelling analysis was given to elucidate that the overwhelmingly high success rate obtained by the new predictor is by no means due to a trivial utilization of the GO annotations. This is 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". The information and clues 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 with a different manner and "depth". To dig out the knowledge about their locations, a sophisticated operation engine is needed. And the current predictor is one of these kinds, and has proved to be a very powerful one. The Hum-PLoc classifier is available as a web-server at http://202.120.37.186/bioinf/hum.

Keywords: Cellular networking; Organelle; Gene ontology; Amphiphilic pseudo amino acid composition; KNN; Fusion; 25% Sequence identity cutoff

The function of a protein is closely linked to its cellular attributes, such as where it is located in a cell and how it is associated with the lipid bilayer [1,2]. 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. Determination of protein subcellular location purely using experimental approaches is both time-consuming and expensive. Particularly, the number of new protein sequences yielded by the high-throughput sequencing technology in the post-genomic era has increased explosively. For instance, in 1986 Swiss-Prot [3] contained only 3939 protein sequence entries, but now the number has jumped to 222,289 according to the version 50.0 of UniProtKB/Swiss-Prot Release as of 30-May-2006, meaning that the number of protein sequences now is more than 56 times of that in 1986. Facing such an avalanche of new protein sequences, it is both challenging and indispensable to develop an automated method for fast and accurately annotating the subcellular

* Corresponding author.
E-mail address: kchou@san.rr.com (K.-C. Chou).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.06.059

attributes of uncharacterized proteins. The knowledge thus obtained can help us timely utilize these newly found protein sequences for both basic research and drug discovery [4,5].

Although the similarity search-based tools such as BLAST can be used to help annotate the subcellular location of an uncharacterized protein, this approach fails when the query protein does not have significant homology to proteins of known localization. Thus, various different approaches for predicting protein subcellular location have been proposed [6–15]. However, all these prediction methods were established basically based on a single classifier derived from a single learning process regardless of whether the operation was engineered with the simple geometric rule, or neural network, or covariant discriminant algorithm, or SVM (support vector machine). Obviously, the prediction quality would be considerably limited by using only one single classifier to deal with piled-up complicated protein sequences with extreme variation in both sequence order and length.

Also, the datasets constructed to train the existing predictors cover very limited cellular locations. For instance, the datasets constructed by Nakashima and Nishikawa [8] only cover two locations, those by Garg et al. [15] four locations, and those by Cedano et al. [9] five locations. Although the datasets constructed by the authors in [14] extended to cover more locations, these datasets were constructed with a very tolerant criterion that allowed inclusion of those proteins with sequence identity up to 80% in a same subcellular location group. To avoid homology bias, a much stricter criterion should be adopted for constructing the working datasets.

Besides, for the practical application in drug discovery, it is more important and urgent to timely annotate the subcellular location of human proteins. However, very few of the aforementioned methods were established specialized for predicting the subcellular location of human proteins. Although the method developed recently by Garg et al. [15] was specifically for human proteins, its coverage was within only four subcellular locations, i.e., cytoplasm, mitochondria, nuclear, and plasma membrane. If a user wishes to use it to predict a protein located outside these four sites, such as endoplasmic reticulum and Golgi apparatus, the predictor will fail to work, or the results thus obtained will not make any sense. Moreover, the cutoff or threshold set by these authors to remove homologous sequences was 90%, meaning that the benchmark dataset thus constructed would contain those proteins with sequence identity up to 90%. Obviously, such a cutoff criterion is not stringent enough because proteins with 40–90% sequence identity are usually deemed quite homologous to each other [4,16].

To extend the scope of practical application and reduce the homology bias, new working datasets were constructed. The new datasets cover 12 subcellular locations with a 25% sequence identity cutoff; i.e., none of proteins has $\geqslant$25% sequence identity to any other within a same subcellular location group. As is well known, the more the subcellular locations covered, the lower the odds are in getting a correct prediction. Also, the more stringent the benchmark dataset in excluding homologous sequences, the harder it becomes to get a high success rate for cross-validation test.

To overcome the difficulties from these two aspects, the samples of proteins were formulated by hybridizing the information derived from the gene ontology [17] and amphiphilic pseudo amino acid composition [18]. Based on the hybridization representation, a novel ensemble classifier was formed by fusing many individual basic classifiers through a voting system. Such an approach has significantly empowered us in predicting human protein subcellular location.

Materials

Protein sequences were collected from the Swiss-Prot database [3] release 49.3 released on 21-March-2006 at http://www.ebi.ac.uk/swissprot/ according to the annotation information in the CC (comment or notes) and ID (identification) fields. In order to collect as much desired information as possible, but meanwhile ensuring a high-quality for the working datasets, the data were screened strictly according to the following criteria. (1) Only those sequences annotated with "human" in the ID field were collected because the current study was focused on human proteins only. (2) Because a same subcellular location (-!-SUBCELLULAR LOCATION) in the CC field might be annotated with different terms, the keywords listed in Table 1 were used to search against the categorization of subcellular locations. (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 [19] to winnow those sequences which have $\geqslant$25% sequence identity to any other in a same subcellular location. (7) Those subcellular locations (subsets) which contain less than eight protein sequences were left out because of lacking statistical significance.

After strictly following the above procedures, we finally obtained 2041 human protein sequences, which are distributed among the following 12 subcellular locations (Fig. 1): 25 proteins in centriole, 377 in cytoplasm, 14 in cytoskeleton, 35 in endoplasmic reticulum, 301 in extracell, 42 in Golgi apparatus, 40 in lysosome, 8 in microsome, 228 in mitochondrion, 580 in nucleus, 23 in peroxisome, and 368 in plasma membrane (Table 2). Thus, we have a dataset S^0 which is a union of the following 12 subsets; i.e.,

$$S^0 = S^0_1 \cup S^0_2 \cup S^0_3 \cup \cdots S^0_{12} \tag{1}$$

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

$$S^{\mathrm{L}} \cup S^{\mathrm{T}} = S^0 \text{ and } S^{\mathrm{L}} \cap S^{\mathrm{T}} = \emptyset \tag{2}$$

where $\cup$, $\cap$, and $\emptyset$ represent the symbols for "union", "intersection", and "empty set" in the set theory, respectively. To avoid the situation that the numbers of proteins in some subsets of the learning dataset S^{L} might overwhelm those of the others, the following "bracket percentage distribution" criterion was used to randomly assign the protein samples to the corresponding subsets of S^{L} and S^{T}

Table 1
Keywords used to search the Swiss-Prot database for known human proteins' subcellular locations

Subcellular location	Keywords
Centriole	Centriole; centrosome; centromer
Cytoplasm	Cytoplasm; cytoplasmic
Cytoskeleton	Cytoskeleton; cytoskeletal; filament; microtubule
Endoplasmic reticulum	Endoplasmic reticulum
Extracell	Extracell; extracellular; secreted
Golgi apparatus	Golgi
Lysosome	Lysosome; lysosomal
Microsome	Microsome; microsomal
Mitochondrion	Mitochondrion; mitochondria; mitochondrial
Nucleus	Nucleus; nuclear
Peroxisome	Peroxisome; peroxisomal; microsome; glyoxysomal; glycosomal
Plasma membrane	Plasma membrane; integral membrane; multi-pass membrane; single-pass membrane

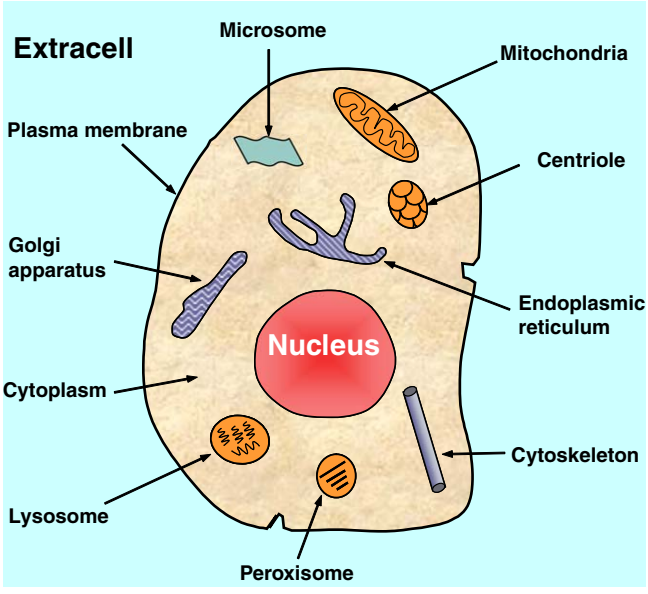

Fig. 1. Schematic illustration to show the 12 subcellular locations of human proteins: (1) centriole; (2) cytoplasm; (3) cytoskeleton; (4) endoplasmic reticulum; (5) extracell; (6) Golgi apparatus; (7) lysosome; (8) microsome; (9) mitochondrion; (10) nucleus; (11) peroxisome; (12) plasma membrane.

Table 2
Breakdown of the human protein sequences obtained by following steps (1)–(7) in Materials

Numerical code	Subcellular location	Number of protein sequences
1	Centriole	25
2	Cytoplasm	377
3	Cytoskeleton	14
4	Endoplasmic reticulum	35
5	Extracellular	301
6	Golgi apparatus	42
7	Lysosome	40
8	Microsome	8
9	Mitochondrion	228
10	Nucleus	580
11	Peroxisome	23
12	Plasma membrane	368
	Total	2041

Table 3
Number of proteins[a] in each of the 12 subcellular locations for the learning and testing datasets randomly generated according to Eq. (3)

Subcellular location	Learning dataset S^{L}	Testing dataset S^{T}
(1) Centriole	20	5
(2) Cytoplasm	155	222
(3) Cytoskeleton	12	2
(4) Endoplasmic reticulum	28	7
(5) Extracell	140	161
(6) Golgi apparatus	33	9
(7) Lysosome	32	8
(8) Microsome	7	1
(9) Mitochondrion	125	103
(10) Nucleus	196	384
(11) Peroxisome	18	5
(12) Plasma membrane	153	215
Total	919	1122

[a] The accession numbers and sequences for the proteins for each of the subsets in the learning and testing datasets are given in the Online Supplementary Materials A and B, respectively. None of the proteins has $\geqslant 25\%$ sequence identity to any other in the same subset (subcellular location) for either within the learning dataset and testing dataset, or between the two.

$$\begin{cases} n_i^{\mathrm{L}} = 100 + \mathrm{INT}\{(n_i^0 - 100) \times 0.2\}, & \text{if } n_i^0 \geqslant 100 \\ n_i^{\mathrm{L}} = \mathrm{INT}\{n_i^0 \times 0.8\}, & \text{if } 20 \leqslant n_i^0 < 100 \\ n_i^{\mathrm{L}} = \mathrm{INT}\{n_i^0 \times 0.9\}, & \text{if } 8 \leqslant n_i^0 < 20 \\ n_i^{\mathrm{T}} = n_i^0 - n_i^{\mathrm{L}} \end{cases} (i = 1, 2, \ldots, 12) \quad (3)$$

where n_i^0, n_i^{L}, andn_i^{T} are the numbers of protein samples in the ith subset of the original dataset S^0, learning dataset S^{L}, and testing dataset S^{T}, respectively, and the symbol INT is the "integer truncation operator" meaning to take the integer part for the number in the brackets right after it. The numbers of proteins thus obtained for the 12 subcellular locations in the learning dataset S^{L} and testing dataset S^{T} are given in Table 3. 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.

Methods

The key to formulate a powerful algorithm for predicting the protein subcellular location is to grasp the core features of proteins that are intrinsically related to their localization in a cell. But the problem is how do we find the core features from piled-up complicated protein sequences?

To realize this, the source of gene ontology consortium [17] may be a useful tool. The rationale is as follows. The gene ontology, or GO, is a controlled vocabulary used to describe the biology of a gene product in any organism. The GO database was established based on the following three species-independent principles: molecular function, biological process, and cellular component.

However, how to effectively use the GO database to improve the prediction quality for protein subcellular location is by no means a trivial problem because, for those proteins with "subcellular location unknown" annotation in Swiss-Prot database, most of their corresponding GO numbers in GO database are also annotated with "cellular component unknown", as will be further elaborated later. Actually, the information for predicting subcellular locations of proteins are "buried" into a series of tedious GO numbers, just like they are "buried" into a pile of complicated amino acid sequences, although the way and the "depth" they are "buried" are quite different. Therefore, the key is how to use these complicated and tedious data to derive the desired results. The following approach was developed for such a goal.

Mapping UniProtKB/Swiss-Prot protein entries [20] 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 04-March-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" [17]. For example, the UniProtKB/Swiss-Prot protein entry "P01040" corresponds to three GO numbers, i.e., "GO:0004866", "GO:0004869", and "GO:0005622". On the other hand, because the current GO database is not complete yet, some protein entries (such as "Q66GS9", "O94986", and "Q8WXD2") have no corresponding GO numbers, i.e., no mapping records at all in the GO database, and hence are not included in 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. The GO database obtained through such a treatment is called GO_compress database, whose dimensions were reduced to 9918 from 51,912 in the original GO database. Each of the 9918 entities in the GO_compress database served as a base to define a protein sample. Unfortunately, the current GO numbers failed to completely cover the proteins concerned, i.e., some proteins might not belong to any of the GO numbers. Although the problem would gradually become trivial or no longer exist with the continuous developing of GO database, to cope with such a problem right now, a hybridization approach was introduced by fusing the GO representation and the amphiphilic pseudo amino acid composition (PseAA) representation [18], as formulated 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:

$$\mathbf{P} = [\phi_1 \quad \phi_2 \quad \cdots \quad \phi_i \quad \cdots \quad \phi_{9918}]^{\mathbf{T}} \tag{4}$$

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

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

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 $(20 + 2\lambda) - \mathrm{D}$ amphiphilic PseAA space [18], as given below:

$$\mathbf{P} = \left[\psi_1 \quad \cdots \quad \psi_{20} \quad \psi_{20+1} \quad \cdots \quad \psi_{20+2\lambda}\right]^{\mathbf{T}}, \tag{6}$$

where $\psi_1, \psi_2, \ldots, \psi_{20}$ are associated with the amino acid composition reflecting the occurrence frequencies of the 20 native amino acids in the protein [21], and $\psi_{20+1}, \ldots, \psi_{20+2\lambda}$ are the 2λ correlation factors that reflect its sequence-order pattern through the amphiphilic feature. The protein representation as defined by Eq. (6) 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 Supplementary C.

Suppose there are N proteins $(\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N)$ which have been classified into 12 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 [22–24] 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 [8], Hamming distance [25,26], and Mahalanobis distance [27,28]. Here, we use the following equation to measure the nearness between protein $\mathbf{P}$ and $\mathbf{P}_i$

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

where $\mathbf{P}\cdot\mathbf{P}_i$ is the dot product of the two vectors, and $\|\mathbf{P}\|$ and $\|\mathbf{P}_i\|$ their moduli, respectively. According to Eq. (7), when $\mathbf{P} \equiv \mathbf{P}_i$ we have $\delta(\mathbf{P},\mathbf{P}_i) = 0$, indicating the "distance" between these two proteins is 0 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 learning dataset that has the shortest "distance" to $\mathbf{P}$ as defined by Eq. (7). If there are two and more proteins in the learning 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 $K \geqslant 5$.

Because the predicted results by the KNN algorithm [22–24] 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 counted in determining the query protein $\mathbf{P}$.

During the course of prediction, the following self-consistency principle should be strictly followed. If a query protein could be 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 $(20 + 2\lambda) - \mathrm{D}$ 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 $(20 + 2\lambda) - \mathrm{D}$ space as well. Accordingly, the current hybridization predictor actually consists of two subpredictors: (1) the NN(K) classifier that operates in the 9918-D GO_compress space, and (2) the NN(K,$20 + 2\lambda$) classifier that operates in the $(20 + 2\lambda) - \mathrm{D}$ amphiphilic PseAA space. The former is the function of K, 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,$20 + 2\lambda$) classifiers, respectively.

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

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

where the symbol $\forall$ denotes the fusing operator, and $\mathbb{NN}^{\mathrm{GO}}$ the ensemble classifier formed by fusing NN(1), NN(2), …, and NN(Ω) according to the flowchart of Fig. 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 $K > 10$.

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 Eq. (8) are $C_1, C_2, \ldots, C_{10}$, respectively; i.e.,

$$\{C_1, C_2, \ldots, C_{10}\} \in \{S_1, S_2, \ldots, S_{12}\} \tag{9}$$

where $\in$ is a symbol in the set theory meaning "member of", and $S_1, S_2, \ldots, S_{12}$ represent the 12 subsets defined by the 12 subcellular locations studied here (Fig. 1), and the voting score for the protein **P** belonging to the kth subset is defined by

$$Y_k^{\mathrm{GO}} = \sum_{i=1}^{10} w_i \Delta(C_i, S_k) \quad (k = 1, 2, \ldots, 12) \tag{10}$$

where w_i is the weight and was set at 1 for simplicity, and the delta function in Eq. (10) is given by

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

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

Two-dimensional fusion process. It is for generating an ensemble classifier by fusing many individual basic NN(K,$20 + 2\lambda$) 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 Eq. (8), let us consider $K = 1, 2, \ldots, 10$, and $\lambda = 0, 1, 2, \ldots, 18$; i.e.,

$$\{K\} = \{1, 2, \ldots, 10\}; \{20 + 2\lambda\} = \{20, 22, \ldots, 54, 56\} \tag{12}$$

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

$$\mathbb{NN}^{\mathrm{Pse}} = \mathrm{NN}(1, 20) \forall \mathrm{NN}(1, 22) \forall \cdots \forall \mathrm{NN}(10, 54) \forall \mathrm{NN}(10, 56) \tag{13}$$

where the fusion operator $\forall$ has the same meaning as that of Eq. (8), and the fusion flowchart can also be illustrated by Fig. 2 but with $\Omega = 10 \times 19 = 190$, meaning a process by fusing 190 basic individual classifiers now.

The detailed process of how the ensemble classifier $\mathrm{NN}^{\mathrm{Pse}}$ works is as follows. Suppose the predicted classification results for the query protein **P** by the 190 individual classifiers in Eq. (13) are

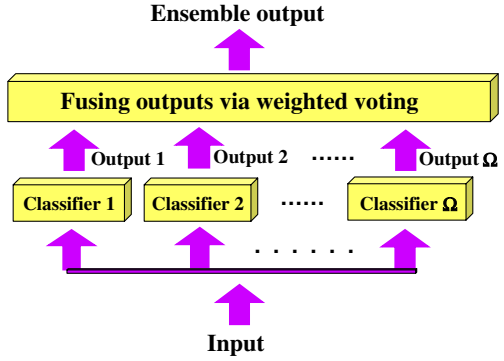

Fig. 2. Flowchart to show how the ensemble classifiers $\mathbb{NN}^{\mathrm{GO}}$ (Eq. (8)) and $\mathbb{NN}^{\mathrm{Pse}}$ (Eq. (13)) are formed by fusing Ω individual classifiers, where $\Omega = 10$ and 190 for the cases of $\mathbb{NN}^{\mathrm{GO}}$ and $\mathbb{NN}^{\mathrm{Pse}}$, respectively.

$$\begin{Bmatrix} C_{1,20} & C_{1,22} & \cdots & C_{1,56} \\ C_{2,20} & C_{2,22} & \cdots & C_{2,56} \\ \vdots & \vdots & \ddots & \vdots \\ C_{10,22} & C_{10,22} & \cdots & C_{10,56} \end{Bmatrix} \in \{S_1, S_2, \ldots, S_{12}\} \tag{14}$$

where $S_1, S_2, \ldots, S_{12}$ have the same meanings as in Eq. (9), i.e., represent the 12 subsets defined by the 12 subcellular locations studied here (Fig. 1), and the voting score for the protein **P** belonging to the kth subset is defined by

$$Y_k^{\mathrm{Pse}} = \sum_{i=1}^{10} \sum_{j=0}^{18} w_{i,20+2j} \Delta(C_{i,20+2j}, S_k) \quad (k = 1, 2, \ldots, 12) \tag{15}$$

where $w_{i,20+2j}$ is the weight and was set at 1 for simplicity, the delta function in Eq. (15) is given by

$$\Delta(C_{i,20+2j}, S_k) = \begin{cases} 1 & \text{if } C_{i,20+2j} \in S_k \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

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

Results and discussion

The prediction process was operated according to the following procedures: if a query protein was defined in the 9918-D GO_compress space, then the ensemble classifier $\mathbb{NN}^{\mathrm{GO}}$ (Eq. (8)) was used to predict its subcellular location; otherwise, the ensemble classifier $\mathbb{NN}^{\mathrm{Pse}}$ (Eq. (13)) was used to predict its subcellular location. The prediction quality was examined by two standard test methods in statistics: the jackknife test and the independent dataset test.

Jackknife test

In the jackknife test, each protein in the learning 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 learning and testing datasets were actually open, and a protein was in turn moving from one to the other. The jackknife test result on the dataset of Online Supplementary Materials A obtained by hybridizing the ensemble classifiers $\mathbb{NN}^{\mathrm{GO}}$ (Eq. (8)) and $\mathbb{NN}^{\mathrm{Pse}}$ (Eq. (13)) is given in Table 4, where, for facilitating comparison, the corresponding results by various other methods are also listed. It can be seen from Table 4 that the overall jackknife success rate by the current ensemble hybridization classifier is more than 50% higher than those by the other existing approaches.

Independent dataset test

In the independent dataset test, the rule parameters were derived from the proteins only in the learning dataset S^{L} (Online Supplementary Materials A), and the prediction was made for proteins in an independent testing dataset S^{T} (Online Supplementary Materials B). The predicted results thus obtained are also given in Table 4, from which

Table 4
Overall success rates for the 12 subcellular locations (Fig. 1) of human proteins by different classifiers and test methods

Classifier	Input form	Test method	
		Jackknife[a]	Independent dataset[b]
Least Euclidean distance [8]	Amino acid composition	$\frac{277}{919} = 30.1\%$	$\frac{331}{1122} = 29.5\%$
ProtLock [9]	Amino acid composition	$\frac{273}{919} = 29.7\%$	$\frac{296}{1122} = 26.4\%$
SVM [14]	Amino acid composition and amino acid pairs [14]	$\frac{280}{919} = 30.5\%$	$\frac{385}{1122} = 34.3\%$
SVM [15]	Amino acid composition and dipeptide composition [15]	$\frac{282}{919} = 30.7\%$	$\frac{374}{1122} = 33.3\%$
Hybridization of ensemble classifiers $\mathbb{NN}^{GO}$ (Eq. (8)) and $\mathbb{NN}^{Pse}$ (Eq. (13))	Hybridization of GO (Eq. (4)) and amphiphilic PseAA (Eq. (6))	$\frac{745}{919} = 81.1\%$	$\frac{954}{1122} = 85.0\%$

[a] Jackknife cross-validation test was performed for the 919 proteins in the Online Supplementary Materials A, where none of the proteins has $\geqslant 25\%$ sequence identity to any other in the same subcellular location.
[b] Prediction was performed for the 1122 independent proteins in the Online Supplementary Materials B; none of proteins in the Online Supplementary Materials A and B has $\geqslant 25\%$ sequence identity to any others in the same subcellular location.

we can see that the current ensemble hybridization classifier outperformed the other methods by 51–58%.

It should be pointed out that the independent dataset test performed here was just for a demonstration of practical application. Because the selection of independent dataset often bears some sort of arbitrariness [29], the jackknife test is deemed more objective than the independent dataset test. In statistical prediction, three cross-validation approaches have been often used in literatures, i.e., the independent dataset test, sub-sampling test, and jackknife test. Of these three, the jackknife test is deemed the most rigorous and objective (see, e.g., a monograph [25] for the mathematical principle and a review [29] for an comprehensive discussion about this). Actually, the jackknife cross-validation has been recently used by more and more investigators [11–13,30–33] in examining the power of various prediction methods. Therefore, the power of a predictor should be measured by the success rate of jackknife test.

A question might be raised as asking why the SVM methods originally reported in [14] and [15] could yield an overall success rate higher than 70%, but here only slightly higher than 30%? The reasons are as follows: (1) The benchmark datasets originally used by these authors contained many homologous sequences in a same subcellular location. For example, the dataset used in [15] contained proteins with up to 90% sequence identity; and the dataset in [14] up to 80% sequence identity. When predictions were made by their methods on the current stringent dataset in which none of protein has $\geqslant 25\%$ sequence identity to any other in a same subcellular location, the success rates would of course decrease significantly. (2) Not only the original high success rate reported in [15] was derived from a high homology dataset, but also the identification was made among merely four subcellular locations, in contrast to 12 locations in the current stringent dataset. (3) The success rates by these methods as reported in [14,15] were obtained by the sub-sampling cross-validation test. When tested by the jackknife cross-validation, their success rates would naturally further diminish because, as mentioned above, the jackknife cross-validation is more rigorous and stringent.

Another question prone to ask is: Was the high success rate obtained here due to a trivial utilization of the subcellular component annotations in the GO database? The answer is absolutely no. The reasons are as follows: (1) Although GO database is constructed based on protein function and cellular component, 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 Nos. O22892 and O00093 in Table 5). (2) 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 actually annotated with "cellular component unknown". For example, for the protein with Accession No. O43303 in Table 5, its subcellular location is annotated with "centrosome" in Swiss-Prot database, but none of its GO numbers indicates its subcellular location. Similar situations do occur for P19877, Q29593, and Q9UIV8 as well. (3) 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 were 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 was used; otherwise, the results obtained by the cross-validation test would not represent any prediction power at all. (4) Finally, as shown by a compelling statistical analysis given in Table 6, the percentage (45.02%) of proteins with GO annotations to indicate their subcellular components is even less than the percentage (51.76%) of proteins with known subcellular location annotation in the Swiss-Prot database. Accordingly, the high success rate obtained by the current predictor was by no means due to a trivial prediction of subcellular location annotation for one database (Swiss-Prot) from another (GO), as often misinterpreted by some people. Also, it can be seen from Table 6 that there are a huge number of proteins with given accession numbers and the corresponding GO numbers, but their

Table 5
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 No.	Swiss-Prot annotation	GO No.	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

Table 6
Breakdown of the 212,425 protein sequence entries from Swiss-Prot database (version 49.3, released 21-March-2006) according to the nature of their subcellular location annotation and their expression in GO

Item	Description	No.	Percentage (%)
(1)	Proteins with known subcellular locations annotated in Swiss-Prot database	109,944	$\frac{109944}{212425} = 51.76$
(2)	Proteins with uncertain subcellular locations annotated in Swiss-Prot database, such as "potential" and "probable"	62,669	$\frac{62669}{212425} = 29.50$
(3)	Proteins that can be represented in the GO space (cf. Eq. (4))	199,389	$\frac{199389}{212425} = 93.86$
(4)	Proteins whose GO numbers are annotated with known subcellular components in the GO database	95,624	$\frac{95624}{212425} = 45.02$

subcellular locations are still unknown. In view of this, the significance of the novel and powerful ensemble classifier as presented in this paper is self-evident.

Conclusion

With the avalanche of new protein sequences generated in the postgenomic era, it is highly desirable to develop a computational method for fast and reliably annotating their subcellular locations because knowledge thus obtained can provide useful clues for revealing their functions and understanding how they interact with each other in cellular networking, one of the fundamental goals in cell biology and proteomics. However, predicting protein subcellular location is a very challenging and complicated problem, particularly for the cases where predictions are made among more subcellular locations and unknown query proteins do not have significant homology to proteins of known subcellular locations. That is why the human protein subcellular location predictor ("hslpred") developed recently [15] could only cover four locations (cytoplasm, mitochondria, nuclear, and plasma membrane), and the benchmark dataset contained proteins with sequence identity up to 90% in a same subset. The former would limit the prediction power while the latter could not avoid homology bias and overestimate the true success rate.

In this study, we constructed a more extensive and meanwhile much more stringent dataset, which covers 12 subcellular locations and in which none of proteins has $\geqslant 25\%$ sequence identity to any others in a same subcellular location. To improve the prediction quality, we adopted the strategy of (1) representing protein samples by hybridizing GO (Eq. (4)) and PseAA (Eq. (6)), and (2) introducing the ensemble classifier that was formed by fusing many basic individual classifiers operated by the engine of the KNN rule. Using GO to represent the sample of a protein

could effectively grasp the core features that might be closely correlated with the subcellular location, and hence enhanced the prediction success rate. However, because the GO database is not complete yet, some 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.

Finally, the significance of the novel predictor can be briefly and vividly expressed as follows. The information and clues of the subcellular locations of proteins are buried into a series of tedious GO numbers, just like they are buried into a pile of complicated amino acid sequences. To dig out the knowledge about their locations, an operation engine is needed. And the current predictor is one of these kinds and has proved to be a very powerful one.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.06.059.

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, Molecular Biology of the Cell, Garland Publishing, New York/London, 1994 (Chapter 1).
[2] H. Lodish, D. Baltimore, A. Berk, S.L. Zipursky, P. Matsudaira, J. Darnell, Molecular Cell Biology, Scientific American Books, New York, 1995 (Chapter 3).
[3] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL, Nucleic Acids Res. 25 (2000) 31–36.
[4] K.C. Chou, Review: Structural bioinformatics and its impact to biomedical science, Curr. Med. Chem. 11 (2004) 2105–2134.
[5] G. Lubec, L. Afjehi-Sadat, J.W. Yang, J.P. John, Searching for hypothetical proteins: theory and practice based upon original data and literature, Prog. Neurobiol. 77 (2005) 90–127.
[6] K. Nakai, P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem. Sci. 24 (1999) 34–36.
[7] K. Nakai, Protein sorting signals and prediction of subcellular localization, Adv. Protein Chem. 54 (2000) 277–344.
[8] H. Nakashima, K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, J. Mol. Biol. 238 (1994) 54–61.
[9] J. Cedano, P. Aloy, J.A. P'erez-Pons, E. Querol, Relation between amino acid composition and cellular location of proteins, J. Mol. Biol. 266 (1997) 594–600.
[10] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, PROTEINS: Struct. Funct. Genet. 43 (2001) 246–255 (Erratum: ibid., 2001, vol. 44, 60).
[11] Z.P. Feng, Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition, Biopolymers 58 (2001) 491–499.
[12] Z.P. Feng, An overview on predicting the subcellular location of a protein, In Silico. Biol. 2 (2002) 291–303.
[13] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, PROTEINS: Struct. Funct. Genet. 50 (2003) 44–48.
[14] K.J. Park, M. Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs, Bioinformatics 19 (2003) 1656–1663.
[15] A. Garg, M. Bhasin, G.P. Raghava, Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search, J. Biol. Chem. 280 (2005) 14427–14432.
[16] L. Holm, C. Sander, Protein folds and families: sequence and structure alignments, Nucleic Acids Res. 27 (1999) 244–247.
[17] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[18] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[19] G.L. Wang, R.L. Dunbrack Jr., PISCES: a protein sequence culling server, Bioinformatics 19 (2003) 1589–1591.
[20] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh, UniProt: the Universal Protein knowledgebase, Nucleic Acids Res. 32 (2004) D115–D119.
[21] K.C. Chou, C.T. Zhang, Predicting protein folding types by distance functions that make allowances for amino acid interactions, J. Biol. Chem. 269 (1994) 22014–22020.
[22] T.M. Cover, P.E. Hart, Nearest neighbour pattern classification, IEEE Trans. Inf. Theory IT-13 (1967) 21–27.
[23] T. Denoeux, A *k*-nearest neighbor classification rule based on Dempster–Shafer theory, IEEE Trans. Syst. Man Cybern. 25 (1995) 804–813.
[24] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy *k*-nearest neighbours algorithm, IEEE Trans. Syst. Man Cybern. 15 (1985) 580–585.
[25] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis: Chapter 11 Discriminant Analysis; Chapter 12 Multivariate Analysis of Variance; Chapter 13 Cluster Analysis, Academic Press, London, 1979, pp. 322–381.
[26] P.Y. Chou, in: G.D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 549–586.
[27] K.C.S. Pillai, in: S. Kotz, N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, John Wiley, New York, 1985, pp. 176–181, This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics, New York.
[28] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins: Struct. Funct. Genet. 21 (1995) 319–344.
[29] K.C. Chou, C.T. Zhang, Review: Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[30] G.P. Zhou, An intriguing controversy over protein structural class prediction, J. Protein Chem. 17 (1998) 729–738.
[31] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, PROTEINS: Struct. Funct. Genet. 44 (2001) 57–59.
[32] R.Y. Luo, Z.P. Feng, J.K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, Eur. J. Biochem. 269 (2002) 4219–4225.
[33] G.P. Zhou, Y.D. Cai, Predicting protease types by hybridizing gene ontology and pseudo amino acid composition, PROTEINS: Struct. Funct. Bioinformatics 63 (2006) 681–684.