Available online at www.sciencedirect.com

ScienceDirect

Biochemical and Biophysical Research Communications 355 (2007) 1006–1011

BBRC

www.elsevier.com/locate/ybbrc

Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites

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

[a] *Department of Biological Chemistry & Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA*
[b] *Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA*

Received 3 February 2007
Available online 23 February 2007

Abstract

Proteins may simultaneously exist at, or move between, two or more different subcellular locations. Proteins with multiple locations or dynamic feature of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery. For instance, among the 6408 human protein entries that have experimentally observed subcellular location annotations in the Swiss-Prot database (version 50.7, released 19-Sept-2006), 973 ($\approx$15%) have multiple location sites. The number of total human protein entries (except those annotated with "fragment" or those with less than 50 amino acids) in the same database is 14,370, meaning a gap of $(14{,}370 - 6408) = 7962$ entries for which no knowledge is available about their subcellular locations. Although one can use the computational approach to predict the desired information for the gap, so far all the existing methods for predicting human protein subcellular localization are limited in the case of single location site only. To overcome such a barrier, a new ensemble classifier, named Hum-mPLoc, was developed that can be used to deal with the case of multiple location sites as well. Hum-mPLoc is freely accessible to the public as a web server at http://202.120.37.186/bioinf/hum-multi. Meanwhile, for the convenience of people working in the relevant areas, Hum-mPLoc has been used to identify all human protein entries in the Swiss-Prot database that do not have subcellular location annotations or are annotated as being uncertain. The large-scale results thus obtained have been deposited in a downloadable file prepared with Microsoft Excel and named "Tab_Hum-mPLoc.xls". This file is available at the same website and will be updated twice a year to include new entries of human proteins and reflect the continuous development of Hum-mPLoc.

Keywords: Large-scale prediction; Human protein; Multiplex locations; Hum-mPLoc; Fusion; Optimal threshold

Being able to grow and reproduce independently, cells are deemed the most basic structural and functional units of all living creatures. Every cell contains numerous protein molecules located in different compartments or organelles, the so-called subcellular locations. One of the fundamental goals in molecular cell biology and proteomics is to identify their subcellular locations or environments because the function of a protein and its role in a cell are closely correlated with which compartment or organelle it resides in. With the avalanche of protein sequences generated in the post-genomic era, it is highly desired to develop an automated method for fast and reliably annotating the subcellular locations of uncharacterized proteins. The knowledge thus obtained can help us timely utilize these newly found protein sequences for both basic research and drug discovery [1,2]. Actually, various different approaches for predicting protein subcellular location have been proposed [3–18]. Unfortunately, none of these methods were established specialized for human proteins, while timely annotating their subcellular location is more impor-

[*] Corresponding author.
E-mail addresses: hbshen@crystal.harvard.edu (H.-B. Shen), kchou@san.rr.com (K.-C. Chou).

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.02.071

tant and urgent because this is directly related to the practical application in drug discovery for human beings. Recently, two methods were developed specifically for predicting the subcellular locations of human proteins: one is called HSLPred [19]; and the other, Hum-PLoc [20]. The former can cover four different subcellular location sites and the latter, 12 different sites. However, none of the two predictors can be used to deal with those human proteins which may simultaneously exist at, or move between, two or more different subcellular locations. Proteins with multiple locations or dynamic feature of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery. Although the multiple location problem has been addressed in two recent papers [21,22], the coverage is limited within the scope of the budding yeast proteins only. Besides no statistical foundation was provided in [21,22] for how to derive the optimal threshold in dealing with the case containing proteins with multiple locations.

The present study was initiated in an attempt to develop a method for predicting the subcellular localization of all human proteins including those with multiple locations as well.

Materials and methods

Materials

According to a statistical analysis on the Swiss-Prot database at www.ebi.ac.uk/swissprot/(version 50.7 released on 19-Sept-2006), the number of total human protein entries is 14,573. After excluding those annotated as "fragment" or containing less than 50 amino acid residues, the number is reduced to (14,573 – 203) = 14,370, of which 10,381 are with subcellular location annotations (Item 1 of Table 1). However, of the 10,381 proteins, only 6408 are annotated with experimental observations (Item 2 of Table 1) and 3973 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 [23], which was established according to molecular function, biological process, and cellular component. As shown in Item 5 of Table 1, of the 14,370 human protein entries, only 9649 have GO annotations to indicate their subcellular components. In other words, the percentage (67.2%) of the human protein entries with subcellular annotations in the GO database is even lower than that (72.2%) in the Swiss-Prot database. Moreover, it is instructive to point out that the GO database was derived from other more fundamental databases including the Swiss-Prot database. Therefore, the GO annotations might be contaminated by the uncertain information from the 3973 entries as indicated in Item 3 of Table 1.

Therefore, the number of human proteins that has reliable subcellular location annotations is 6408 (Item 2 of Table 1), which is about 45% of all the human protein entries concerned; i.e., there are 14,370 – 6408 = 7962 human protein entries for which the subcellular localization needs to be identified or further confirmed.

Also, as shown in Table 1, of the 6408 human proteins with experimentally annotated subcellular location annotation, 973 are with multiple location sites (Item 4 of Table 1); i.e., about 15% of the proteins may simultaneously exist at two or more subcellular locations. This kind of multiplex proteins were totally excluded during the process of data construction in the precious studies [19,20], but now they are to become an important constituent part of the training dataset, as will be illuminated below.

Protein sequences were collected from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/. The detailed procedures are basically the same as those described in [20]. The only differences are: (1) to get the updated data, instead of version 49.3, the version 50.7 released on 9-Sept-2006 is adopted. (2) To cover the proteins with multiple locations, those sequences which were excluded in the previous study [20] due to being annotated by two or more subcellular locations are included in the current study.

After strictly following the aforementioned procedures, we finally obtained a benchmark dataset $\mathbb{S}$ covering 14 subcellular locations (Fig. 1), as outlined in Table 2. The corresponding accession numbers and protein sequences are given in Online Supporting Information A.

Methods

Hum-PLoc [20] is a powerful ensemble classifier that can identify a query human protein among 12 possible subcellular locations with very high success rate even under the condition that it has less than 25% sequence identity to the proteins in the training dataset. However, Hum-PLoc was established on the assumption that each of the human proteins concerned dwells in one, and only one, subcellular location. To enable it to deal with proteins with multiple locations, let us consider the following procedures.

Training dataset. Because some proteins may simultaneously exist in two or more subcellular locations, it is instructive to introduce the concept of "locative protein" according to the following identity: given a same protein coexisting at two different subcellular locations, it will be counted as 2 locative proteins; if coexisting at three locations, 3 locative proteins; and so forth. Thus, the number of total locative proteins, $\tilde{N}$, can be expressed as

$$\tilde{N} = \tilde{n}_1 + \tilde{n}_2 + \cdots + \tilde{n}_m = \tilde{N}_1 + 2\tilde{N}_2 + \cdots + m\tilde{N}_m = \sum_{\tau=1}^{m} \tau\tilde{N}_\tau \quad (1)$$

Table 1
Breakdown of the 14,370[a] human protein entries from the Swiss-Prot database (version 50.7 released on 19-Sept-2006) according to the nature of their subcellular location annotation and their expression in the GO database (released on 12-Sept-2006)

Item	Description	Number	Percentage (%)
1	Human proteins with subcellular location annotations in the Swiss-Prot database	10,381	10,381/14,370 = 72.2
2	Proteins in Item 1 with experimentally observed subcellular locations	6408	6408/14,370 = 44.6
3	Proteins in Item 1 with uncertain terms, such as "potential", "probable", and "by similarity"	3973	3973/14,370 = 27.7
4	Proteins in Item 2 with multiple subcellular locations	973	973/6406 = 15.2
5	Proteins that have the corresponding GO numbers in the GO database	13,490	13,490/14,370 = 93.9
6	Proteins with subcellular component annotations in the GO database	9649	9649/14,370 = 67.2

[a] The number of the original human protein entries was 14,573, of which 203 were either annotated as "fragment" or with less than 50 amino acid residues, and hence were removed for further consideration.

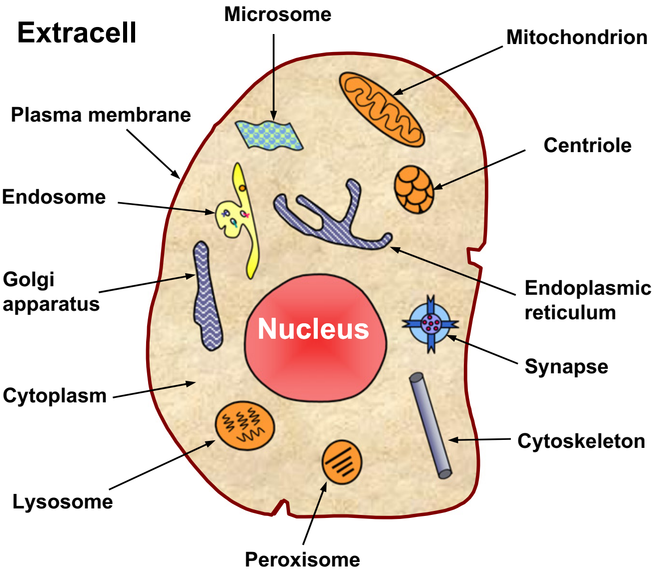

Fig. 1. Schematic illustration to show the fourteen subcellular locations of human proteins: (1) centriole, (2) cytoplasm, (3) cytoskeleton, (4) endoplasmic reticulum, (5) endosome, (6) extracell, (7) Golgi apparatus, (8) lysosome, (9) microsome, (10) mitochondrion, (11) nucleus, (12) peroxisome, (13) plasma membrane, and (14) synapse.

Table 2
Breakdown of the human protein benchmark dataset derived from Swiss-Prot database (release 50.7) according to the procedures described in Materials

Order	Subcellular location	Number of proteins
1	Centriole	39
2	Cytoplasm	633
3	Cytoskeleton	47
4	Endoplasmic reticulum	157
5	Endosome	48
6	Extracell	325
7	Golgi apparatus	112
8	Lysosome	63
9	Microsome	15
10	Mitochondrion	307
11	Nucleus	877
12	Peroxisome	46
13	Plasma membrane	455
14	Synapse	10
Total number of locative proteins $\tilde{N}$		3134[a]
Total number of different proteins N		2750[b]

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

[a] See Eqs. (1)–(4) for the definition about the number of locative proteins, and its relation with the number of different proteins.

[b] Of the 2750 different proteins, 2396 belong to one subcellular location, 325 to two locations, 28 to three locations, and 1 to four locations.

where m is the number of total subcellular locations (for the current study $m = 14$ as shown in Fig. 1); $\tilde{n}_1$ the number of locative proteins in the 1st subcellular location (Table 2), $\tilde{n}_2$ that in the 2nd subcellular location, and so forth; $\tilde{N}_1$ the number of proteins with one subcellular location, $\tilde{N}_2$ that with two locations, and so forth. Suppose N is the number of total different proteins; it can be expressed by

$$N = \tilde{N}_1 + \tilde{N}_2 + \cdots + \tilde{N}_m = \sum_{\tau=1}^{m} \tilde{N}_\tau \tag{2}$$

Subtracting Eq. 2 from Eq. 1, we obtain the relation between N and $\tilde{N}$ as given by

$$\tilde{N} = N + \tilde{N}_2 + \cdots + (m-1)\tilde{N}_m = N + \sum_{\tau=2}^{m} (\tau - 1)\tilde{N}_\tau \tag{3}$$

For example, for the dataset in Table 2, of the 3134 locative proteins, 2396 belong to one subcellular location, 325 to two locations, 28 to three locations, and 1 to four locations. Substituting these data into Eq. 3, we obtain the number of different proteins

$$N = 3134 - [(2-1) \times 325 + (3-1) \times 28 + (4-1) \times 1] = 2750 \tag{4}$$

which is fully consistent with the figures in Table 2. The number of locative proteins is greater than the number of different proteins, i.e., $\tilde{N} > N$; when, and only when, none of the proteins could exist in more than one subcellular location, should we have $\tilde{N} = N$.

Threshold value. For the single subcellular location predictor, the criterion in determining which subcellular location a query protein **P** should belong to is determined by $Y_\mu(\mathbf{P})$, a score function generated by an ensemble classifier formed by fusing many basic individual classifiers through a voting system (see Eqs. 10 and 15 of [20]), where $\mu = 1, 2, \ldots, 14$ refers to the μth subset (subcellular location). Note that the maximum value for μ is increased from 12 in [20] to 14 here because two more subcellular locations are covered in the current dataset derived from a newer version of Swiss-Prot database. The predicted result is made by assigning the query protein **P** to the λth subcellular location with which the score function has the maximum value; i.e.,

$$\lambda = \text{Arg-Max}_\mu\{Y_\mu(\mathbf{P})\}, \quad (\mu = 1, 2, \ldots, 14) \tag{5}$$

where the operator Arg-Max$_\mu$ means taking the value of subscript μ with which $Y_\mu(\mathbf{P})$ is the maximum. Now for the multiple subcellular locations predictor, Eq. 5 should be modified as

$$\{\lambda\} = \text{Arg-Max}_\mu\{Y_\mu(\mathbf{P});\ \text{threshold} \leqslant \theta\}, \quad (\mu = 1, 2, \ldots, 14) \tag{6}$$

where θ is the threshold value for the allowable deviation in determining the optimal score for $Y_\mu(\mathbf{P})$, meaning that any subset, say $\lambda 2$, whose score $Y_{\lambda 2}(\mathbf{P})$ is within a deviation of θ from the highest score, say $Y_{\lambda 1}(\mathbf{P})$, i.e., $Y_{\lambda 1}(\mathbf{P}) - Y_{\lambda 2}(\mathbf{P}) \leqslant \theta$, then the query protein **P** will be assigned to the subcellular location $\lambda 2$ as well. Accordingly, in addition to a single index, $\{\lambda\}$ in Eq. 6 may also represent two or more indexes, corresponding to two or more subcellular locations predicted. The predictor obtained through the above modified procedures is called Hum-mPLoc (θ) that will also cover the proteins with multiple subcellular locations; i.e., the conversion can be formulate as

$$\text{Hum-PLoc} \Rightarrow \text{Hum-mPLoc}(\theta) \tag{7}$$

Thus, the number of proteins predicted with multiple locations will depend on the value of θ (see Eq. 6): the larger the value of θ, the more the proteins will be predicted having multiple locations. In other words, if θ is too large, which will lead to an over-prediction; if θ is too small, under-prediction. In view of this, the key is how to find the optimal value for θ.

Similar to the procedure in determining the threshold value for predicting HIV protease cleavage sites in proteins [24], the optimal value of θ can be determined by an optimizing process as illustrated below. Because the score functions, $Y_\mu(\mathbf{P})$, generated by the ensemble classifier for different μ are integers (see Eqs. 10 and 15 of [20]), the θ can also be reduced to the scope of integers. Thus, we can assign $\theta = 0, 1, 2, 3, 4, \ldots$ to Eq. 6, and find the optimal value for θ through the following procedure:

Suppose the predicted subcellular locations for a query protein by Hum-mPLoc (θ) ensemble classifier for a given value of θ is

$$\mathbb{C}(\theta) = \{C_1(\theta), C_2(\theta), \ldots, C_{m(\theta)}(\theta)\} \tag{8}$$

while the real subcellular locations to which the protein **P** belongs are

$$\mathbb{R} = \{R_1, R_2, \ldots, R_r\} \tag{9}$$

Define a quality control function for the threshold θ as given by

$$Q(\theta) = H(\theta) - \mathbb{N}\{\mathbb{S}(\theta)\} \tag{10}$$

where $H(\theta)$ is a hit function given by

$$H(\theta) = \sum_{i=1}^{m(\theta)} \Delta_i(C_i(\theta), \mathbb{R}) \tag{11}$$

where

$$\Delta_i(C_i(\theta), \mathbb{R}) = \begin{cases} 1, & \text{if } C_i(\theta) \in \mathbb{R} \\ 0, & \text{if } C_i(\theta) \notin \mathbb{R} \end{cases} \tag{12}$$

and $\mathbb{N}\{\mathbb{S}(\theta)\}$ represents the number of elements in the set $\mathbb{S}(\theta)$ formed by

$$\mathbb{S}(\theta) = [\mathbb{R} \cup \mathbb{C}(\theta)] - [\mathbb{R} \cap \mathbb{C}(\theta)] \tag{13}$$

where $\cup$ and $\cap$ represent the symbols of union and intersection, respectively, in the set theory. During the self-consistency test process [25] on the benchmark dataset $\mathbb{S}$ (Table 2), each of the proteins singled out for test will yield a value of $Q(\theta)$. Suppose the sum for all these values are given by

$$\Omega(\theta) = \sum_{\mathbf{P} \in \mathbb{S}} Q(\theta) \tag{14}$$

where $\in$ is a symbol in the set theory meaning "member of", and hence $\Omega(\theta)$ is a function of θ. The optimal value for θ is given by

$$\theta^* = \text{Arg-Max}_\theta\{\Omega(\theta)\} \tag{15}$$

where Arg-Max_θ means taking the value of argument θ with which $\Omega(\theta)$ is the maximum. For the benchmark dataset $\mathbb{S}$ of 3134 proteins as listed in Table 2, we obtained $\theta^* = 2$, meaning that the optimal threshold value is 2 for the current benchmark dataset. Therefore, the ensemble classifier of Eq. 7 can be further explicitly expressed as

$$\text{Hum-mPLoc} = \text{Hum-mPLoc}(\theta) \quad \text{(with the optimal threshold value of } \theta = 2) \tag{16}$$

Success rate. For the current study, the proteins in the benchmark dataset $\mathbb{S}$ consists of $\mu = 14$ subsets; i.e.,

$$\mathbb{S} = S_1 \cup S_2 \cup S_3 \cup \cdots S_{13} \cup S_{14} \tag{17}$$

where each subset corresponds to one of the 14 subcellular locations according to the order of Table 2. For the single location case, suppose the result predicted by the ensemble classifier Hum-PLoc [20] on $\mathbf{P}_k^\mu$, the kth protein in the μth subset, is the site belonging to the u_k^μth subcellular location; i.e.,

$$\text{Hum-PLoc}\langle \mathbf{P}_k^\mu \rangle = u_k^\mu, \quad (\mu = 1, 2, \ldots, 14; u_k^\mu = 1, 2, \ldots, 14), \tag{18}$$

then the overall success rate can be defined by

$$\frac{1}{N} \sum_{\mu=1}^{14} \sum_{k=1}^{n_\mu} \delta[u_k^\mu, \mu], \tag{19}$$

where n_μ is the number of proteins in the μth subcellular location of the benchmark dataset, and the delta function

$$\delta[u_k^\mu, \mu] = \begin{cases} 1, & \text{if } u_k^\mu = \mu \\ 0, & \text{if } u_k^\mu \neq \mu \end{cases} \tag{20}$$

However, for the multiple location case, the definition will be more complicated because the predicted result for a given protein may belong to one or more subcellular locations. Now, let us suppose the result operated by the multiple location predictor Hum-mPLoc on $\mathbf{P}_k^\mu$ is U_k^μ; i.e.,

$$\text{Hum-mPLoc}\langle \mathbf{P}_k^\mu \rangle = U_k^\mu, \tag{21}$$

where U_k^μ is not a number but a set that contains one or more subscript numbers in Eq. 17. Thus, the overall success rate is defined by

$$\frac{1}{\tilde{N}} \sum_{\mu=1}^{14} \sum_{k=1}^{\tilde{n}_\mu} \Delta[U_k^\mu, \mu]. \tag{22}$$

where $\tilde{n}_\mu$ is the number of locative proteins in the μth subset of the benchmark dataset (see Eq. 1), and the Δ function is defined by

$$\Delta[U_k^\mu, \mu] = \begin{cases} 1, & \text{if } \mu \in U_k^\mu \\ 0, & \text{if } \mu \notin U_k^\mu \end{cases} \tag{23}$$

Results and discussion

In statistical prediction the single independent dataset test, sub-sampling test and jackknife test are the three methods often used for cross-validation. Of these three, the jackknife test is deemed as the most rigorous and objective one, as illustrated by a comprehensive review [26]. Therefore, jackknife test has been increasingly used in literatures [8,11,27–39] for examining the accuracy of various prediction methods.

In jackknife test, each protein in the benchmark dataset was singled out in turn as a "test protein" and all the rule parameters were calculated from the remaining 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. For the case that includes proteins with multiple subcellular locations, it is instructive to note that during the process of jackknife test each of the N different proteins was singled out only once for testing although it may coexist at more than one location corresponding to several locative proteins (Table 2). However, to keep the maximum success rate $\leqslant$100% in accordance with the conventional definition, the denominator of Eq. 22 should be $\tilde{N}$ rather than N (see Eq. 3).

The jackknife test was performed with Hum-mPLoc (Eq. 16) on the dataset of Online Supporting Information A. The overall success rate as defined by Eq. 22 was $2218/3134 = 70.8\%$. This is a very high success rate as can be seen from the following discussion.

Let us imagine: if the protein samples are completely randomly distributed among the 14 possible locations, the overall success rate by random assignments would generally be $1/14 \simeq 7.1\%$; if the random assignments are weighted according to the sizes of subsets (Table 2), then the overall success rate would be

$$\begin{aligned} &\frac{1}{3134^2}(39^2 + 633^2 + 47^2 + 157^2 + 48^2 + 325^2 + 112^2 \\ &\quad + 63^2 + 15^2 + 307^2 + 877^2 + 46^2 + 455^{22} + 10^2) \\ &\quad \simeq 16.6\% \end{aligned} \tag{24}$$

Therefore, the overall success rate by the current multiple ensemble classifier Hum-mPLoc are overwhelmingly higher than the completely randomized rate and weighted randomized rate, implying that Hum-mPLoc is indeed very powerful in predicting subcellular localization of proteins including those with multiple location sites.

The results of the large-scale identifications performed by Hum-mPLoc for all human protein entries in the Swiss-Prot database that do not have subcellular location annotations or are annotated as being uncertain are given in Online Supporting Information B. Meanwhile, for the public convenience, the large-scale results have been deposited in a downloadable file prepared with Microsoft Excel and named “Tab_Hum-mPLoc.xls”. This file is available at http://202.120.37.186/bioinf/hum-multi and will be updated twice a year to include new entries of human proteins and reflect the continuous development of Hum-mPLoc.

To help readers understand the entry data listed in Tab_Hum-mPLoc.xls, some examples are illustrated through Table 3. It can be seen from the table that the Microsoft Excel data file consists of the following 4 columns:

- Column A is for the protein accession numbers.
- Column B is for the Swiss-Prot codes.
- Column C is for the annotations from Swiss-Prot database: the component in this column is either empty meaning no subcellular annotation available from Swiss-Prot database for the corresponding protein entry, or with uncertain terms such as “probable”, “potential”, and “by similarity”.
- Column D is for the subcellular locations identified by Hum-mPLoc.

As we can see from the table, the protein with Accession No. “Q15417” and Swiss-Prot code “CNN3_HUMAN” has no subcellular annotation available in Swiss-Prot, but was identified by Hum-mPLoc as belonging to “cytoskeleton”. Also, the protein with accession number “O00295” and Swiss-Prot code “TULP2_HUMAN” has no subcellular annotation in Swiss-Prot, but according to Hum-mPLoc, it belongs to both “centriole” and cytoplasm”. It is interesting to see that the majority of human proteins identified by Hum-mPLoc belong to a single subcellular location as observed, and that in most cases the results identified by Hum-mPLoc are quite consistent with the uncertain annotations in Swiss-Prot database. However, inconsistency does exist. For example, the protein with Accession No. P46926 and Swiss-Prot code “GNPI_HUMAN” has the uncertain annotation locating in “cytoplasm (by similarity)”, but it was identified by Hum-mPLoc belonging to “lysosome”. Future experimental findings will tell which one of the two is correct.

Conclusion

Although many different methods have been developed for predicting protein subcellular location, further development is needed due to the following reasons. (1) Most of the existing methods only cover a limited number of subcellular locations and will fail to work if a query protein is outside their coverage. (2) The benchmark datasets used in most existing methods contain proteins with high sequence identity with those in a same subcellular location, and hence will lead to undesired bias or fail to work if a query protein has no significant sequence similarity to proteins of known subcellular location. (3) So far there is no method whatsoever that can be used to deal with human proteins with multiple subcellular location sites, but human proteins of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in drug discovery.

The new predictor Hum-mPLoc presented in this paper was devoted to deal with these problems, and quite encouraging results were obtained.

Table 3
Illustrations to help readers understand the Online Supporting Information B containing the predicted results by Hum-mPLoc for the 7962 human proteins without subcellular location annotations available from databanks or annotated with uncertain terms such as “probable”, “potential”, and“by similarity”

A Accession No.	B Swiss-Prot code	C Annotation in Swiss-Prot database	D Identified location by Hum-mPLoc
Q15417	CNN3_HUMAN		Cytoskeleton
Q8NDT2	RB15B_HUMAN	Nucleus (probable)	Nucleus
O00295	TULP2_HUMAN		Centriole; cytoplasm
P83111	LACTB_HUMAN		Endosome
O95340	PAPS2_HUMAN		Cytoplasm; mitochondrion
Q92783	STAM1_HUMAN	Cytoplasm (probable)	Cytoplasm; Golgi
P46926	GNPI_HUMAN	Cytoplasm (by similarity)	Lysosome
O94812	BAIP3_HUMAN		Synapse
Q9NUI1	DECR2_HUMAN	Peroxisome (by similarity)	Peroxisome
Q9UJQ7	CT079_HUMAN		Mitochondrion; peroxisome
Q8WWZ3	EDAD_HUMAN	Cytoplasm (probable)	Cytoplasm
Q96A47	ISL2_HUMAN	Nucleus (by similarity)	Nucleus
Q8IVH4	MMAA_HUMAN	Mitochondrion (probable)	Cytoplasm; mitochondrion
O15520	FGF10_HUMAN	Secreted protein (potential)	Secreted protein
Q9NYA3	GOGA6_HUMAN		Golgi
Q66LE6	2ABD_HUMAN	Cytoplasm (by similarity)	Cytoplasm
O14607	UTY_HUMAN	Nucleus (potential)	Centriole; nucleus
Q96P15	SPB11_HUMAN	Cytoplasm (by similarity)	Secreted protein
Q8NCQ2	CS034_HUMAN		Nucleus
P54802	ANAG_HUMAN		Lysosome

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.02.071.

References

[1] K.C. Chou, Review: structural bioinformatics and its impact to biomedical science, Curr. Med. Chem. 11 (2004) 2105–2134.

[2] 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.

[3] 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.

[4] K. Nakai, Protein sorting signals and prediction of subcellular localization, Adv. Protein Chem. 54 (2000) 277–344.

[5] 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.

[6] 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.

[7] K.C. Chou, D.W. Elrod, Protein subcellular location prediction, Protein Eng. 12 (1999) 107–118.

[8] 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.

[9] 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).

[10] Z.P. Feng, An overview on predicting the subcellular location of a protein, In Silico Biol. 2 (2002) 291–303.

[11] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins Struct. Funct. Genet. 50 (2003) 44–48.

[12] 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.

[13] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers, J. Proteome Res. 5 (2006) 1888–1897.

[14] H.B. Shen, K.C. Chou, Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells, Biopolymers 85 (2006) 233–240.

[15] K.C. Chou, H.B. Shen, Large-scale predictions of Gram-negative bacterial protein subcellular locations, J. Proteome Res. 5 (2006) 3420–3428.

[16] H.B. Shen, K.C. Chou, Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins, Protein Eng. Des. Sel. 20 (2007) 39–46.

[17] K.C. Chou, H.B. Shen, Large-scale plant protein subcellular location prediction, J. Cell. Biochem. 100 (2007) 665–678.

[18] H.B. Shen, J. Yang, K.C. Chou, Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction, Amino Acids (2007), doi:10.1007/s00726-00006-00478-00728.

[19] 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.

[20] K.C. Chou, H.B. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, Biochem. Biophys. Res. Commun. 347 (2006) 150–157.

[21] K.C. Chou, Y.D. Cai, Predicting protein localization in budding yeast, Bioinformatics 21 (2005) 944–950.

[22] K. Lee, D.W. Kim, D. Na, K.H. Lee, D. Lee, PLPD: reliable protein localization prediction from imbalanced and overlapped datasets, Nucleic Acids Res. 34 (2006) 4655–4666.

[23] 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.

[24] K.C. Chou, A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins, J. Biol. Chem. 268 (1993) 16938–16948.

[25] 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.

[26] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[27] G.P. Zhou, An intriguing controversy over protein structural class prediction, J. Protein Chem. 17 (1998) 729–738.

[28] X. Xiao, S. Shao, Y. Ding, Z. Huang, Y. Huang, K.C. Chou, Using complexity measure factor to predict protein subcellular location, Amino Acids 28 (2005) 57–61.

[29] Y. Gao, S.H. Shao, X. Xiao, Y.S. Ding, Y.S. Huang, Z.D. Huang, K.C. Chou, Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter, Amino Acids 28 (2005) 373–376.

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

[31] X.D. Sun, R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids 30 (2006) 469–475.

[32] Z. Wen, M. Li, Y. Li, Y. Guo, K. Wang, Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition, Amino Acids 32 (2006) 277–283.

[33] S.W. Zhang, Q. Pan, H.C. Zhang, Z.C. Shao, J.Y. Shi, Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion, Amino Acids 30 (2006) 461–468.

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

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

[36] Q.B. Gao, Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Eng. Des. Sel. 19 (2006) 511–516.

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

[38] P. Du, Y. Li, Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence, BMC Bioinformatics 7 (2006) 518.

[39] Z.H. Zhang, Z.H. Wang, Z.R. Zhang, Y.X. Wang, A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine, FEBS Lett. 580 (2006) 6169–6174.