

Euk-mPLoc: A Fusion Classifier for Large-Scale Eukaryotic Protein Subcellular Location Prediction by Incorporating Multiple Sites

Kuo-Chen Chou^{*,†,‡} and Hong-Bin Shen^{‡,§}

Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 1954 Hua-Shan Road, Shanghai 200030, China, and School of Information Engineering, Southern Yangtze University, Wuxi, China

Received November 28, 2006

One of the critical challenges in predicting protein subcellular localization is how to deal with the case of multiple location sites. Unfortunately, so far, no efforts have been made in this regard except for the one focused on the proteins in budding yeast only. For most existing predictors, the multiple-site proteins are either excluded from consideration or assumed even not existing. Actually, proteins may simultaneously exist at, or move between, two or more different subcellular locations. For instance, according to the Swiss-Prot database (version 50.7, released 19-Sept-2006), among the 33 925 eukaryotic protein entries that have experimentally observed subcellular location annotations, 2715 have multiple location sites, meaning about 8% bearing the multiplex feature. 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. Meanwhile, according to the same Swiss-Prot database, the number of total eukaryotic protein entries (except those annotated with "fragment" or those with less than 50 amino acids) is 90 909, meaning a gap of $(90\ 909 - 33\ 925) = 56\ 984$ 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 blank, so far, all the existing methods for predicting eukaryotic protein subcellular localization are limited in the case of single location site only. To overcome such a barrier, a new ensemble classifier, named Euk-mPLoc, was developed that can be used to deal with the case of multiple location sites as well. Euk-mPLoc is freely accessible to the public as a Web server at <http://202.120.37.186/bioinf/euk-multi>. Meanwhile, to support the people working in the relevant areas, Euk-mPLoc has been used to identify all eukaryotic 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 at the same Web site via a downloadable file prepared with Microsoft Excel and named "Tab_Euk-mPLoc.xls". Furthermore, to include new entries of eukaryotic proteins and reflect the continuous development of Euk-mPLoc in both the coverage scope and prediction accuracy, we will timely update the downloadable file as well as the predictor, and keep users informed by publishing a short note in the Journal and making an announcement in the Web Page.

Keywords: Large-scale prediction • Eukaryotic protein • Multiple locations • Ensemble classifier • Fusion • Optimal threshold • Euk-mPLoc

1. Introduction

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 explosion of protein sequences generated in the postgenomic era, it is highly desired to develop a high-

throughput tool for rapidly 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, many efforts have been made in developing various methods for predicting protein subcellular location.^{3–15} Unfortunately, none of these methods can be used to deal with 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

* To whom correspondence should be addressed. E-mail: kchou@san.rr.com.

[†] Gordon Life Science Institute.

[‡] Shanghai Jiaotong University.

[§] Southern Yangtze University.

Table 1. Breakdown of the 90 909^a Eukaryotic 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	Eukaryotic proteins with subcellular location annotations in the Swiss-Prot database	63134	63134/90909 = 69.4%
2	Proteins in Item 1 with experimentally observed subcellular locations	33925	33925/90909 = 37.3%
3	Proteins in Item 1 with uncertain terms, such as “potential”, “probable”, and “by similarity”	29209	29209/90909 = 32.1%
4	Proteins in Item 2 with multiple subcellular locations	2715	2715/33925 = 8.0%
5	Proteins that have the corresponding GO numbers in the GO database	87029	87029/90909 = 95.7%
6	Proteins with subcellular component annotations in the GO database	59533	59533/90909 = 65.5%

^a The number of the original Eukaryotic protein entries was 99,777, of which 8,868 were either annotated as “fragment” or with less than 50 amino acid residues, and hence were removed for further consideration.

research and drug discovery. Although the multiple location problem has been addressed in two recent papers,^{16,17} the coverage is limited within the scope of the budding yeast proteins only. Besides, no statistical foundation was provided in these papers^{16,17} 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 establish a method for predicting the subcellular localization of all eukaryotic proteins including those with multiple locations as well.

2. 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 eukaryotic protein entries is 99 777. After excluding those annotated as “fragment” or containing less than 50 amino acid residues, the number is reduced to (99 777 – 8868) = 90 909, of which 63 134 are with subcellular location annotations (item 1 of Table 1). However, of the 63 134 proteins, only 33 925 are annotated with experimental observations (item 2 of Table 1) and 29 209 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,¹⁸ which was established according to molecular function, biological process, and cellular component. As shown in item 5 of Table 1, of the 90 909 eukaryotic protein entries, only 59 533 have GO annotations to indicate their subcellular components. In other words, the percentage (65.5%) of the eukaryotic protein entries with subcellular annotations in the GO database is even lower than that (69.4%) 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 29 209 entries as indicated in item 3 of Table 1.

Therefore, the number of eukaryotic proteins that have reliable subcellular location annotations is 33 925 (item 2 of Table 1), which is about 37% of all the eukaryotic protein entries concerned; that is, there are 90 909 – 33 925 = 56 984 eukaryotic protein entries for which the subcellular localization needs to be identified or further confirmed.

Also, as shown in Table 1, of the 33 925 eukaryotic proteins with experimentally annotated subcellular location annotation, 2715 are with multiple location sites (item 4 of Table 1); that

is, about 8% of these 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 study,¹³ but now they are to become an important constituent part of the training data set, 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 previously described.¹³ The only differences are the following: (1) To get the updated data, instead of version 49.3, the newer 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 study¹³ because they were annotated by two or more subcellular locations are included in the current study. (3) To avoid homology bias but meanwhile cover as many locations as possible, the 25% sequence identity cutoff procedure¹³ is imposed for all the subcellular locations except for the following three: acrosome, melanosome, and synapse; otherwise, the numbers of proteins remaining in the three subsets would be too few to have statistical significance.

After strictly following the aforementioned procedures, we finally obtained a benchmark data set S covering 22 subcellular locations (Figure 1), as outlined in Table 2. The corresponding accession numbers and protein sequences are given in Supporting Information A and the Web site at <http://202.120.37.186/bioinf/euk-multi>.

3. Method

Euk-OET-PLoc¹³ is a powerful ensemble classifier that can yield very high success rate on a very stringent benchmark data set which covers 16 possible eukaryotic protein subcellular locations and in which none of proteins have $\geq 25\%$ sequence identity to any other in a same subcellular location. However, Euk-OET-PLoc was established on the assumption that each of the eukaryotic proteins concerned locates at one, and only one, subcellular location. To enable it to deal with proteins with multiple locations, let us consider the following procedures.

3.1. Training Data Set. 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

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

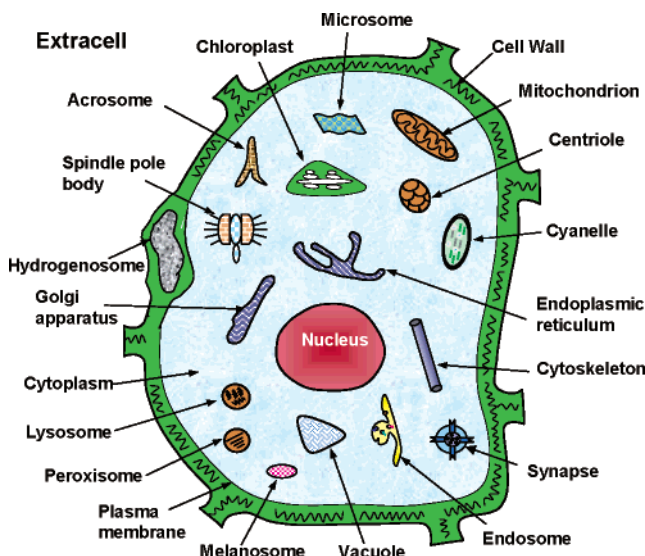


Figure 1. Schematic illustration to show the 22 subcellular locations of eukaryotic proteins: (1) acrosome, (2) cell wall, (3) centriole, (4) chloroplast, (5) cyanelle, (6) cytoplasm, (7) cytoskeleton, (8) endoplasmic reticulum, (9) endosome, (10) extracell, (11) Golgi apparatus, (12) hydrogenosome, (13) lysosome, (14) melanosome, (15) microsome (16) mitochondria, (17) nucleus, (18) peroxisome, (19) plasma membrane, (20) plastid, (21) spindle pole body, and (22) vacuole.

where m is the number of total subcellular locations (for the current study $m = 22$ as shown in Figure 1); \tilde{n}_1 is the number of locative proteins in the first subcellular location (Table 2), \tilde{n}_2 that in the second subcellular location, and so forth; and \tilde{N}_1 is 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

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

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

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

For example, for the data set in Table 2, of the 5618 different proteins, 5091 belong to one subcellular location, 495 to two locations, 28 to three locations, and 4 to four locations. Substituting these data into eq 3, we obtain the number of locative proteins

$$\tilde{N} = 5618 + [(2 - 1) \times 495 + (3 - 1) \times 28 + (4 - 1) \times 4] = 6181 \quad (4)$$

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

3.2. Threshold Value. For the single subcellular location predictor, the criterion in determining which subcellular location a query protein \mathbf{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

Table 2. Breakdown of the Eukaryotic Protein Benchmark Data Set Derived from Swiss-Prot Database (Release 50.7) According to the Procedures Described in the Materials Section^a

order	subcellular location	number of proteins
1	Acrosome	17
2	Cell wall	53
3	Centriole	64
4	Chloroplast	501
5	Cyanelle	85
6	Cytoplasm	1060
7	Cytoskeleton	74
8	Endoplasmic reticulum	364
9	Endosome	89
10	Extracell	640
11	Golgi apparatus	254
12	Hydrogenosome	13
13	Lysosome	80
14	Melanosome	13
15	Microsome	31
16	Mitochondrion	535
17	Nucleus	1333
18	Peroxisome	97
19	Plasma membrane	725
20	Spindle pole body	36
21	Synapse	15
22	Vacuole	102
Total number of locative proteins \tilde{N}		6181 ^b
Total number of different proteins N		5618 ^c

^a None of the proteins have $\geq 25\%$ sequence identity to any other in a same subset except for the following three locations: acrosome, melanosome, and synapse; otherwise, the numbers of proteins remaining in the three subsets would be too few to have statistical significance. ^b See eqs 1–4 for the definition about the number of locative proteins, and its relation with the number of different proteins. ^c Of the 5618 different proteins, 5091 belong to one subcellular location, 495 to two locations, 28 to three locations, and 4 to four locations.

voting system (see eqs 12 and 18 of ref 13), where $\mu = 1, 2, \dots, 16$ refers to the μ -th subset (subcellular location). The predicted result is made by assigning the query protein \mathbf{P} to the λ -th subcellular location with which the score function has the maximum value; that is,

$$\lambda = \arg \max_{\mu} \{Y_{\mu}(\mathbf{P})\}, \quad (\mu = 1, 2, \dots, 16) \quad (5)$$

where λ is the argument of μ that maximizes $Y_{\mu}(\mathbf{P})$. For the multiple subcellular locations predictor, eq 5 should be modified as

$$\{\lambda\} = \text{which}\{Y_{\mu}(\mathbf{P}) \geq \max[Y_{\mu}(\mathbf{P})] - \theta\}, \quad (\mu = 1, 2, \dots, 22) \quad (6)$$

where “which” is a function often used in the R language (<http://www.r-project.org/>) that returns the indices satisfying the condition stated in the braces. Also, the maximum value for μ here is increased from 16 to 22 because 6 more subcellular locations are covered in the current data set derived from a newer version of Swiss-Prot database, and 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 $Y_{\lambda 1}(\mathbf{P})$, that is, $Y_{\lambda 1}(\mathbf{P}) - Y_{\lambda 2}(\mathbf{P}) \leq \theta$, then the query protein \mathbf{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 indeces, corresponding to two or more subcellular locations predicted. The predictor obtained through the above modified procedures is called Euk-mPLoc (θ) that

will also cover the proteins with multiple subcellular locations; that is, the conversion can be formulated as

$$\text{Euk-OET-PLoc} \Rightarrow \text{Euk-mPLoc}(\theta) \quad (7)$$

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, this will lead to an over-prediction for multiplex proteins; if θ is too small, this will lead to 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,¹⁹ 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 fusion classifier for different μ are integers (see eqs 12 and 18 of ref 13), the θ can also be reduced to the scope of integers. Thus, we can assign $\theta = 0, 1, 2, 3, 4, \dots$ to eq 6, and find the optimal value for θ through the following procedure:

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

$$C(\theta) = \{C_1(\theta), C_2(\theta), \dots, C_{m(\theta)}(\theta)\} \quad (8)$$

while the real subcellular locations to which the protein \mathbf{P} belongs are

$$R = \{R_1, R_2, \dots, R_r\} \quad (9)$$

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

$$Q(\theta) = H(\theta) - ||S(\theta)|| \quad (10)$$

where $H(\theta)$ represents the number of successful hits as given by

$$H(\theta) = ||R \cap C(\theta)|| = \sum_{i=1}^{m(\theta)} \Delta_i(C_i(\theta), R) \quad (11)$$

where $||R \cap C(\theta)||$ represents the number of the elements in the intersection between the set R and the set $C(\theta)$ (see eqs 8 and 9 and Figure 2), and the delta function is given by

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

while $S(\theta)$ is the set formed by the over-hit and miss-hit elements during the prediction with Euk-mPLoc(θ) as given by (cf. Figure 2)

$$S(\theta) = [R \cup C(\theta)] - [R \cap C(\theta)] \quad (13)$$

where \cup and \cap represent the symbols of union and intersection, respectively, in the set theory. And $||S(\theta)||$ in eq 10 represents the number of the total elements in the set $S(\theta)$, that is, the sum of the over-hit and miss-hit events.

During the self-consistency test process²⁰ on the benchmark dataset S (Table 2), each of the proteins in testing with the Euk-mPLoc(θ) fusion classifier will yield a value of $Q(\theta)$. Suppose the sum for all these values are given by

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

where \in is a symbol in the set theory meaning “member of”,

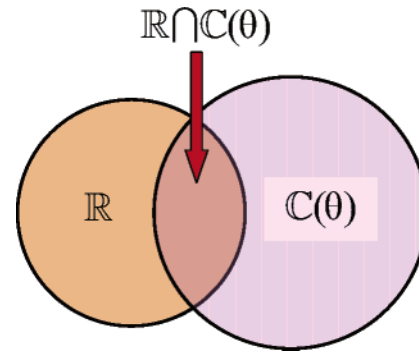


Figure 2. Schematic drawing to show (1) $R \cap C(\theta)$, the intersection region (pink) between the set R (see eq 9) and the set $C(\theta)$ (see eq 8); the number of the elements in such a region, that is, $||R \cap C(\theta)||$, represents the number of the successful hits obtained by the Euk-mPLoc(θ) fusion classifier. (2) $S(\theta) = [R \cup C(\theta)] - [R \cap C(\theta)]$, the remaining region (orange or purple), where the number of elements, that is, $||S(\theta)||$, represents the sum of the over-hits and miss-hits. See the relevant texts for further explanation.

and hence $\Omega(\theta)$ is a function of θ . The optimal value for θ is given by

$$\theta^* = \arg \max_{\theta} \{\Omega(\theta)\} \quad (15)$$

where “arg max” has the same meaning as that of eq 5, meaning taking the value of argument θ with which $\Omega(\theta)$ is the maximum. For the benchmark dataset S of 6181 proteins as listed in Table 2, we obtained $\theta^* = 2$, meaning that the optimal threshold value is 2 for the current benchmark data set. Therefore, the ensemble classifier of eq 7 can be further explicitly expressed as

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

3.3. Success Rate. For the current study, the proteins in the benchmark data set S consists of $\mu = 22$ subsets; that is,

$$S = S_1 \cup S_2 \cup S_3 \cup \dots \cup S_{21} \cup S_{22} \quad (17)$$

where each subset corresponds to one of the 22 subcellular locations according to the order of Table 2. For the single location case, suppose the result predicted by the ensemble classifier Euk-OET-PLoc¹³ on \mathbf{P}_k^μ , the k -th protein in the μ -th subset, is the site belonging to the u_k^μ -th subcellular location; that is,

$$\text{Euk-OET-PLoc}(\mathbf{P}_k^\mu) = u_k^\mu, \quad (\mu = 1, 2, \dots, 22; \quad u_k^\mu = 1, 2, \dots, 22) \quad (18)$$

then the overall success rate can be defined by

$$\frac{1}{N} \sum_{\mu=1}^{22} \sum_{k=1}^{n_\mu} \delta[u_k^\mu, \mu] \quad (19)$$

where n_μ is the number of proteins in the μ -th subcellular location of the benchmark data set, and the delta function

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

However, for the multiple location case, the definition will be more complicated because the predicted result for a given

Table 3. Illustrations To Help Readers Understand the Supporting Information B Containing the Predicted Results by Euk-mPLoc for the 56 984 Eukaryotic Proteins That Either Have No Subcellular Location Annotations in Databanks or Are Annotated with Uncertain Terms Such as “Probable”, “Potential”, and “by Similarity”

A	B	C	D
accession number	Swiss-Prot code	annotation in Swiss-Prot database	identified location by Euk-mPLoc
Q8GY58	GUN23_ARATH		cell wall.
Q9BY78	RNF26_HUMAN		cytoplasm.
Q80U87	UBP8_MOUSE		cytoplasm. nucleus.
Q41853	RSH1_MAIZE	nucleus (probable).	nucleus.
Q19958	STO2_CAEEL		endoplasmic reticulum. Golgi.
Q9DCN1	NUD12_MOUSE	peroxisome (by similarity).	peroxisome.
O99795	CYB_VARVV		mitochondrion.
Q99PU7	BAP1_MOUSE	nucleus (by similarity).	cytoplasm. nucleus.
Q08326	MSS4_RAT		endosome.
O23735	CYSK2_BRAJU	cytoplasm (by similarity).	chloroplast.
Q9QZK8	DNS2A_RAT	lysosome (by similarity).	lysosome.
P08144	AMYA_DROME		lysosome. secreted protein.
Q5KF05	MVP1_CRYNE	cytoplasm (by similarity).	endosome.
P01671	KV3S_MOUSE		acrosome.
Q17029	VATF_ANOGA		chloroplast. mitochondrion.
Q8X1X3	G3P_PARBR	cytoplasm (by similarity).	peroxisome.
Q9WTI7	MYO1C_MOUSE		cytoskeleton.
Q9USS8	YNB2_SCHPO		centriole. cytoplasm.
Q9I968	PA22_TRIMU	secreted protein (by similarity).	secreted protein.
Q01771	STADS_BRANA	plastid; chloroplast (probable).	chloroplast.
Q01642	MS84A_DROME		cytoplasm. cytoskeleton.
P07597	NLTP1_HORVU		cell wall. cytoplasm.
P87027	SPG1_SCHPO		spindle pole body.
P59326	YTHD1_MOUSE		synapse.

protein may belong to one or more subcellular locations. Now, let us suppose the result operated by the multiple location predictor Euk-mPLoc on P_k^μ is U_k^μ ; that is,

$$\text{Euk-mPLoc}(P_k^\mu) = U_k^\mu \quad (21)$$

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

$$\frac{1}{\tilde{N}} \sum_{\mu=1}^{22} \sum_{k=1}^{\tilde{n}_\mu} \Delta[U_k^\mu, \mu] \quad (22)$$

where \tilde{n}_μ is the number of locative proteins in the μ -th subset of the benchmark data set (see eq 1), and the Δ function is defined by

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

4. Results and Discussion

In statistical prediction, the following three methods are often used for cross-validating the accuracy of a predictor: the single independent data set test, subsampling test, and jackknife test. Of these three, the jackknife test is deemed as the most rigorous and objective one, as illustrated by a comprehensive review.²¹ Therefore, jackknife test has been increasingly used in literatures^{9,11,22–36} for examining the quality of various prediction methods.

In jackknife test, each protein in the benchmark data set 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 data sets 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 $\leq 100\%$ 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 Euk-mPLoc (eq 16) on the data set of Supporting Information A. The overall success rate as defined by eq 22 was $4165/6181 = 67.4\%$. This is a very high success rate as can be seen from the following discussion.

Let us imaginethat if the protein samples are completely randomly distributed among the 22 possible locations, the overall success rate by random assignments would generally be $1/22 \approx 4.5\%$; if the random assignments are weighted according to the sizes of sub sets (Table 2), then the overall success rate would be⁵

$$\frac{1}{6181^2} (17^2 + 53^2 + 64^2 + 501^2 + 85^2 + 1060^2 + 74^2 + 364^2 + 89^2 + 640^2 + 254^2 + 13^2 + 80^2 + 13^2 + 31^2 + 535^2 + 1333^2 + 97^2 + 725^2 + 36^2 + 15^2 + 102^2) = 12.1\% \quad (24)$$

The rates thus obtained would be even lower for the statistical system where some proteins may have multiple locations as considered in the current study. Therefore, the overall success rate by the current multiple ensemble classifier Euk-mPLoc are overwhelmingly higher than the completely randomized rate and weighted randomized rate, implying that Euk-mPLoc is indeed very powerful in predicting subcellular localization of proteins including those with multiple location sites.

The results of the large-scale predictions performed by Euk-mPLoc for all eukaryotic protein entries in the Swiss-Prot

database that do not have subcellular location annotations or are annotated as being uncertain are given in Supporting Information B. Meanwhile, for the public convenience, the large-scale results have been deposited at <http://202.120.37.186/bioinf/euk-multi> via a downloadable file prepared with Microsoft Excel and named “Tab_Euk-mPLOC.xls”.

To help readers understand the entry data listed in Tab_Euk-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 Euk-mPLOC.

As we can see from the table, the protein with accession number “Q8GY58” and Swiss-Prot code “GUN23_ARATH” has no subcellular annotation available in Swiss-Prot, but was identified by Euk-mPLOC as belonging to the “cell wall”. Also, the protein with accession number “Q19958” and Swiss-Prot code “STO2_CAEEL” has no subcellular annotation in Swiss-Prot, but according to Euk-mPLOC, it was predicted belonging to both “endoplasmic reticulum” and “Golgi apparatus”. It is interesting to see that the majority of eukaryotic proteins identified by Euk-mPLOC belong to a single subcellular location as observed, and that in most cases the results identified by Euk-mPLOC are quite consistent with the uncertain annotations in Swiss-Prot database. However, some inconsistency does exist. For example, the protein with accession number “Q5KF05” and Swiss-Prot code “MVP1_CRYNE” has the uncertain annotation locating in “cytoplasm (by similarity)”, but it was identified by Euk-mPLOC belonging to “endosome”. Future experimental findings will tell which one of the two is correct.

5. Conclusion

Prediction of protein subcellular localization is an important problem. Although many different methods have been developed in this regard, we are challenged by the following problems: (1) So far, there is no method that can be generally used to deal with eukaryotic proteins with multiple subcellular location sites, but proteins of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery. (2) 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. (3) The benchmark data sets 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.

The new predictor Euk-mPLOC presented in this paper was devoted to deal with these problems, and quite encouraging results were obtained. It is anticipated that with more new entries of eukaryotic proteins into databanks, the predictor will be further developed in both the coverage scope and prediction quality. To keep the users timely informed of the development, a short note will be published in the Journal, and an announcement placed in the Web site.

Acknowledgment. We wish to express our gratitude to the two anonymous reviewers whose suggestions were very helpful in improving the presentation of this paper.

Supporting Information Available: Data set containing the 6181 locative protein sequences classified into 22 eukaryotic subcellular locations, and table listing the predicted subcellular locations for the 56 984 eukaryotic proteins. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Chou, K. C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **2004**, *11*, 2105–2134.
- (2) Lubec, G.; Afjehi-Sadat, L.; Yang, J. W.; John, J. P. Searching for hypothetical proteins: theory and practice based upon original data and literature. *Prog. Neurobiol.* **2005**, *77*, 90–127.
- (3) Nakashima, H.; Nishikawa, K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* **1994**, *238*, 54–61.
- (4) Cedano, J.; Aloy, P.; P'erez-Pons, J. A.; Querol, E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **1997**, *266*, 594–600.
- (5) Chou, K. C.; Elrod, D. W. Protein subcellular location prediction. *Protein Eng.* **1999**, *12*, 107–118.
- (6) Nakai, K.; Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **1999**, *24*, 34–36.
- (7) Nakai, K. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* **2000**, *54*, 277–344.
- (8) Chou, K. C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct., Funct., Genet.* (Erratum: **2001**, *44*, 60) **2001**, *43*, 246–255.
- (9) Feng, Z. P. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers* **2001**, *58*, 491–499.
- (10) Feng, Z. P. An overview on predicting the subcellular location of a protein. *In Silico Biol.* **2002**, *2*, 291–303.
- (11) Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 44–48.
- (12) Park, K. J.; Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. *Bioinformatics* **2003**, *19*, 1656–1663.
- (13) Chou, K. C.; Shen, H. B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res.* **2006**, *5*, 1888–1897.
- (14) Shen, H. B.; Chou, K. C. Virus-PLOC: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers* **2007**, *85*, 233–240.
- (15) Shen, H. B.; Chou, K. C. Gpos-PLOC: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng., Des. Sel.* **2007**, *20*, 39–46.
- (16) Chou, K. C.; Cai, Y. D. Predicting protein localization in budding yeast. *Bioinformatics* **2005**, *21*, 944–950.
- (17) Lee, K.; Kim, D. W.; Na, D.; Lee, K. H.; Lee, D. PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic Acids Res.* **2006**, *34*, 4655–4666.
- (18) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29.
- (19) Chou, K. C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.* **1993**, *268*, 16938–16948.
- (20) Chou, K. C. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 319–344.
- (21) Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.
- (22) Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729–738.
- (23) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y.; Chou, K. C. Using complexity measure factor to predict protein subcellular location. *Amino Acids* **2005**, *28*, 57–61.

- (24) Gao, Q. B.; Wang, Z. Z.; Yan, C.; Du, Y. H. Prediction of protein subcellular location using a combined feature of sequence. *FEBS Lett.* **2005**, *579*, 3444–3448.
- (25) Guo, Y. Z.; Li, M.; Lu, M.; Wen, Z.; Wang, K.; Li, G.; Wu, J. Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. *Amino Acids* **2006**, *30*, 397–402.
- (26) Sun, X. D.; Huang, R. B. Prediction of protein structural classes using support vector machines. *Amino Acids* **2006**, *30*, 469–475.
- (27) Wen, Z.; Li, M.; Li, Y.; Guo, Y.; Wang, K. Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. *Amino Acids* **2006**, *32*, 277–283.
- (28) Mondal, S.; Bhavna, R.; Mohan Babu, R.; Ramakumar, S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.* **2006**, *243*, 252–260.
- (29) Zhang, S. W.; Pan, Q.; Zhang, H. C.; Shao, Z. C.; Shi, J. Y. Prediction protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and naive Bayes feature fusion. *Amino Acids* **2006**, *30*, 461–468.
- (30) Chen, C.; Zhou, X.; Tian, Y.; Zou, X.; Cai, P. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal. Biochem.* **2006**, *357*, 116–121.
- (31) Cao, Y.; Liu, S.; Zhang, L.; Qin, J.; Wang, J.; Tang, K. Prediction of protein structural class with Rough Sets. *BMC Bioinf.* **2006**, *7*, 20.
- (32) Gao, Q. B.; Wang, Z. Z. Classification of G-protein coupled receptors at four levels. *Protein Eng., Des. Sel.* **2006**, *19*, 511–516.
- (33) Guo, J.; Lin, Y.; Liu, X. GNBSL: A new integrative system to predict the subcellular location for Gram-negative bacteria proteins. *Proteomics* **2006**, *6*, 5099–5105.
- (34) Du, P.; Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinf.* **2006**, *7*, 518.
- (35) Zhang, Z. H.; Wang, Z. H.; Zhang, Z. R.; Wang, Y. X. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett.* **2006**, *580*, 6169–6174.
- (36) Chen, C.; Tian, Y. X.; Zou, X. Y.; Cai, P. X.; Mo, J. Y. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J. Theor. Biol.* **2006**, *243*, 444–448.

PR060635I