Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms

Kuo-Chen Chou[1] & Hong-Bin Shen[2]

[1]Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, USA. [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to K.-C.C. (kcchou@gordonlifescience.org) or H.-B.S. (hbshen@crystal.harvard.edu).

Published online 17 January 2008; doi:10.1038/nprot.2007.494

Information on subcellular localization of proteins is important to molecular cell biology, proteomics, system biology and drug discovery. To provide the vast majority of experimental scientists with a user-friendly tool in these areas, we present a package of Web servers developed recently by hybridizing the 'higher level' approach with the *ab initio* approach. The package is called Cell-PLoc and contains the following six predictors: Euk-mPLoc, Hum-mPLoc, Plant-PLoc, Gpos-PLoc, Gneg-PLoc and Virus-PLoc, specialized for eukaryotic, human, plant, Gram-positive bacterial, Gram-negative bacterial and viral proteins, respectively. Using these Web servers, one can easily get the desired prediction results with a high expected accuracy, as demonstrated by a series of cross-validation tests on the benchmark data sets that covered up to 22 subcellular location sites and in which none of the proteins included had $\geq 25\%$ sequence identity to any other protein in the same subcellular-location subset. Some of these Web servers can be particularly used to deal with multiplex proteins as well, which may simultaneously exist at, or move between, two or more different subcellular locations. Proteins with multiple locations or dynamic features of this kind are particularly interesting, because they may have some special biological functions intriguing to investigators in both basic research and drug discovery. This protocol is a step-by-step guide on how to use the Web-server predictors in the Cell-PLoc package. The computational time for each prediction is less than 5 s in most cases. The Cell-PLoc package is freely accessible at http://chou.med.harvard.edu/bioinf/Cell-PLoc.

INTRODUCTION

Knowledge of the subcellular localization of proteins is important because it can (i) provide useful insights about their functions, (ii) indicate how and in what kind of cellular environments they interact with each other and with other molecules and (iii) help in understanding the intricate pathways that regulate biological processes at the cellular level[1,2].

Although the subcellular localization of a protein can be determined by conducting various biochemical experiments, the approach by purely doing experiments is both time consuming and costly. With the avalanche of gene products in the post-genomic age, the gap between newly found protein sequences and the knowledge of their subcellular localization is becoming increasingly wide[3]. For instance, according to version 52.0 of the Swiss-Prot database released on March 6, 2007, at http://www.ebi.ac.uk/swissprot/, the number of total protein entries is 260,175. After excluding those annotated as 'fragment' or containing less than 50 amino-acid residues, the number is reduced to 247,262, out of which 133,652 are with subcellular-location annotations (item 1 of **Table 1**). However, out of the 133,652 proteins, only 49,367 are annotated with experimental observations (item 2 of **Table 1**) and 84,285 are annotated with uncertain labels such as 'probable', 'potential', 'perhaps' and 'by similarity' (item 3 of **Table 1**). A similar gap also exists in the gene ontology (GO) database[4], which was established based on the molecular function, biological process and cellular component. As shown in item 5 of **Table 1**, out of the 247,262 proteins, only 116,593 have GO annotations to indicate

TABLE 1 | Breakdown of the 247,262[a] protein entries from Swiss-Prot database (version 52.0, released on March 6, 2007) according to the nature of their subcellular location annotation and their expression in the GO database (released on March 6, 2007).

Item	Description	Number	Percentage
1	Proteins with subcellular location annotations in Swiss-Prot database	133, 652	$\frac{133,652}{247,262} = 54.1\%$
2	Proteins in Item 1 with experimentally observed subcellular locations	49, 367	$\frac{49,367}{247,262} = 20.0\%$
3	Proteins in Item 1 with uncertain terms, such as 'potential', 'probable' and 'by similarity'	84, 285	$\frac{84,285}{247,262} = 34.1\%$
4	Proteins that can be represented in the GO space	226, 596	$\frac{226,596}{247,262} = 91.6\%$
5	Proteins with subcellular component annotations in the GO database	116, 593	$\frac{116,593}{247,262} = 47.2\%$

[a]The original number of protein entries was 260,175, of which 12,913 were either annotated as 'fragment' or with less than 50 amino acid residues, and hence were removed for further consideration.

their subcellular components. In other words, the percentage (47.2%) of the protein entries with subcellular annotations in the GO database is even lower than that (54.1%) in the Swiss-Prot database. Moreover, it is instructive to point out that the GO database was derived from other more basic databases including the Swiss-Prot database. Therefore, the GO annotations might be contaminated by the uncertain information from the 84,285 entries as indicated in item 3 of **Table 1** (see refs. 4–6).

For the timely use of these newly found proteins for basic research and drug discovery[7,8], it is highly desirable to develop an automated method to bridge such a gap. During the past 15 years, many efforts have been devoted to deal with such a challenge, and significant progresses have been achieved in predicting the subcellular localization of proteins[6,9–45].

In spite of using many different advanced techniques, the aforementioned methods can be basically categorized into two strategies: the *ab initio* sequence-based approach and the 'higher level' sequence-based approach.

The so-called *ab initio* approach is that the prediction is made based on the sequence information alone, without using any information derived from the higher level databases such as GO. To develop a powerful method for predicting the subcellular localization of a protein, one of the most important things is how to represent the sample of a protein by a descriptor that not only contains as much information as possible but also can be handled by a powerful prediction engine. One of the typical *ab initio* approaches is the sequential model in which the sample of a protein is represented by its entire amino-acid sequence, and the sequence similarity search-based tools like BLAST[46] are used to conduct prediction. However, this model fails to work when the query protein does not have significant homology to proteins of known location[31,43,47]. To deal with this problem, various discrete models were developed in which the sample of a protein is represented by a set of discrete numbers. The simplest discrete model was based on the amino-acid (AA) composition or AAC[10,11,13,14,24]. In the AAC discrete model, all the sequence-order effects were missing. To avoid losing the sequence-order information completely, the concept of pseudo amino-acid (PseAA) composition or PseAAC was proposed[48] that can reflect the sequence-order information (at least partially) through a set of correlation factors, and the prediction quality has been remarkably improved[48,49]. The concept of PseAAC has also been used by many others in improving the prediction quality for subcellular localization of proteins and their other attributes[25,42,45,50–60]. Because PseAAC has been widely used, for the convenience of users, a free server called PseAAC (see ref. 61) has been established recently at the website http://chou.med.harvard.edu/bioinf/PseAAC/. By using this Web server, users can generate the PseAAC for any given protein sequence by selecting the mode they want. For a systematic introduction about the *ab initio* approach and its various models, the readers are referred to a recent paper by Emanuelsson *et al.*[43]. Although various *ab initio* models have their respective merits, all of them have a common limit: the success rate is very low when the query protein has less than 25% sequence identity to proteins of known location, particularly when the number of subcellular locations to be covered is greater than four or five[35,40,62].

Here, we will focus on the six Web servers in the Cell-PLoc package[6] that were developed recently based on the 'higher level' sequence-based approach, or strictly speaking, a hybridization of the 'higher level' and the *ab initio* sequence-based approaches. These Web servers distinguish themselves by having the following features. (i) User-friendly and quick to generate the result. By just typing or copying and pasting the query protein sequence into the input box, the user can generally get the desired result in less than 5 s. (ii) Wider coverage scope. In comparison with some popular predictors such as PSORT[12], TargetP[18] and PSORT-B[27] that cover five or less subcellular locations, some predictors in the Cell-PLoc package can cover up to 22 subcellular locations. (iii) Low pairwise sequence identity benchmark data set. Compared with the data sets constructed for many existing predictors that allow inclusion of protein samples with 80% (see ref. 26), 90% (see ref. 13,31) or even higher sequence identity, the benchmark data sets used to develop these Web-server predictors in the Cell-PLoc package were strictly complying with such a rule that none of the proteins included have $\geq$25% sequence identity to any other protein in the same subcellular location subset. This is important for avoiding homology and redundancy bias, and particularly useful for dealing with those proteins that do not have significant sequence homology to any of the proteins of known location[43,47]. (iv) Higher expected accuracy. Because these Web-server predictors were established by hybridizing the 'higher level' GO approach with the state-of-the-art *ab initio* sequence-based approach, the overall success rates of prediction are generally significantly higher than those by the best of the existing *ab initio* sequence-based approaches alone. This kind of enhancement in the success rate is particularly more remarkable when a query protein has less sequence similarity to proteins of known location and when the coverage of prediction is wider, that is, the number of subcellular locations to be covered is greater. For instance, for the benchmark data set of human proteins classified into 12 subcellular locations where none of the proteins has $\geq$25% sequence identity to any other protein within the same subset, the overall jackknife cross-validation success rate by the *ab initio* sequence-based approach based on the state-of-the-art technique such as support vector machine (SVM)[63] was lower than 40%, but that by the approach of hybridizing the 'higher level and *ab initio* approaches was about 81% (see ref. 35). For the benchmark data set of eukaryotic proteins classified into 16 subcellular locations with the same threshold to exclude homologous sequences, the overall jackknife success rate by the hybridization approach was about 82%, which is also more than 40% higher than those by various *ab initio* approaches[40]. (v) Ability to deal with multiplex proteins. Some proteins may simultaneously exist at, or move between, two or more subcellular locations. Proteins with multiple locations or dynamic features of this kind are particularly interesting because they may have some special biological functions intriguing to investigators in both basic research and drug discovery. Two of the Web servers in the current version of Cell-PLoc package, Hum-mPLoc and Euk-mPLoc, can be used to deal with biological systems containing both single-location and multiple-location proteins. To the best of our knowledge, so far no other Web server can do the same. (vi) Availability of large-scale predicted results. To maximize the convenience of people working in the relevant areas, we have used each of the six predictors to identify all the protein entries (except those annotated with 'fragment' or those with less than 50 amino acids) in the Swiss-Prot database for the corresponding organism that do not have subcellular location annotations or are annotated with uncertain terms such as 'probable', 'potential', 'likely' or 'by similarity'. We have

deposited the large-scale results thus obtained into the relevant downloadable files. To get these files, just follow the steps described in the Procedure section below. These large-scale results can serve the following two purposes: they can be directly used by those who need the information immediately and they can set a preceding mark to examine the accuracy of our predicted results by the future experimental results.

The Cell-PLoc package was developed for predicting the subcellular localization of proteins in various different organisms. If one wishes to predict the signal peptide of a query protein, we would recommend the use of SignalP[64], PrediSi[65] as well as Signal-CF at http://chou.med.harvard.edu/bioinf/Signal-CF/ (see ref. 66) and Signal-3L at http://chou.med.harvard.edu/bioinf/Signal-3L/ (see ref. 67) developed very recently.

The Web server predictors in the Cell-PLoc package have been well recognized[2,39,42,45,53,58–60,68–74]. For example, Plant-PLoc[62] was published in 2007 and has already been used by Ho and Ng[74] to predict the subcellular locations of chitinases from *Medicago sativa* and *Galege orientalis*.

What is the 'higher level' sequence-based approach?

A typical 'higher level' sequence-based approach is the one in which a protein sample is defined in the GO[4] space, as originally introduced in ref. 75. According to the GO database, a protein entry may correspond to several GO numbers. If each of the GO numbers in the GO database is used to serve as a vector basis, a protein entry can be defined as a high-dimensional vector in the GO space by searching the GO database for the protein entry; if a hit is found, then the corresponding vector component is assigned 1, otherwise 0. Thus, the dimensions of a protein vector will depend on the total GO numbers. For example, the GO database released on May 30, 2006, contains 10,173 GO numbers and, therefore, the protein entry will be defined in a 10,173-dimensional space.

The reason for representing protein samples as described above was based on the assumption that by this method proteins mapped into the GO database space would be clustered in a way better reflecting their subcellular locations. This enhances the success rate of prediction for those proteins that do not have significant sequence homology to proteins with known locations.

Besides the GO database, another higher level database is the FunD (functional domain) database, which has been derived from the integrated domain and motif database[76] or InterPro database. The FunD database consists of many sequences with well-known functional domain types. If each of these sequences in the FunD database is used to serve as a vector basis as originally proposed in ref. 23, a protein entry can be defined as a high-dimensional vector in the FunD space according to the following procedure: search the FunD database for the protein sequence; if a hit is found, then the corresponding vector component is assigned 1, otherwise 0. Thus, the dimensions of a protein vector will depend on the total FunD sequences. For example, the FunD database from the InterPro release 6.2 (April 24, 2003) contains 7,785 entries that are available from the website at http://www.ebi.ac.uk/interpro and, therefore, the protein will be defined in a 7,785-dimensional space. The FunD approach is very effective for predicting protein structural class, as demonstrated in ref. 77.

Out of the above two 'higher level' sequence-based approaches, the approach using the GO database has been extensively studied for predicting the subcellular localization of proteins. Therefore, in this protocol, we will focus on the GO approach.

Hybridization of the 'higher level' sequence-based approach with the *ab initio* sequence-based approach

A query protein entry may not have any corresponding GO number at all and hence its representation in the GO space will be a meaningless naught vector. This kind of situation may arise due to the following two reasons: the GO database is not complete yet or the query protein is a synthetic or a hypothetical one[8]. For the former, the problem will become trivial or eventually be solved with the continuous development of the GO database; but for the latter, the problem will always exist. A similar situation can also occur for the approach using FunD[23]. To cope with the naught vector problem, we adopted the strategy by hybridizing the GO approach with the *ab initio* approach, as described below.

Given a query protein entry, if any hit is found by searching the GO database for the protein entry, its subcellular location will be identified by the 'higher level' GO approach, otherwise, by the state-of-the-art *ab initio* sequence-based approach. The flowchart given in **Figure 1** illustrates the process of the hybridization approach.

Because more than 90% protein entries in the Swiss-Prot database have corresponding GO numbers, and the success rates obtained by using the 'higher level' GO approach are overwhelmingly higher than those by the *ab initio* approaches, and also because the *ab initio* predictor built in the hybridization approach is at least comparable to the best of the existing *ab initio* methods, it is conceivable that the overall success rate obtained by the hybridization approach should be significantly higher than those by the existing individual *ab initio* approaches.

The current Cell-PLoc package consists of six Web servers that were established recently based on the hybridization approaches, specialized on various organisms. The prediction engines in these Web servers are featured with a powerful ensemble classifier formed by fusing many individual basic classifiers, operated by either KNN[78] or OET-KNN[79,80] algorithm. Each of these Web-server predictors has been tested to yield a high overall jackknife success rate on a very stringent benchmark data set in which none of the proteins have $\geq$25% sequence identity with any other in the same subcellular location[35,40,62].

Below, we will describe the equipment and input data needed and give a step-by-step guide on how to use these Web servers.

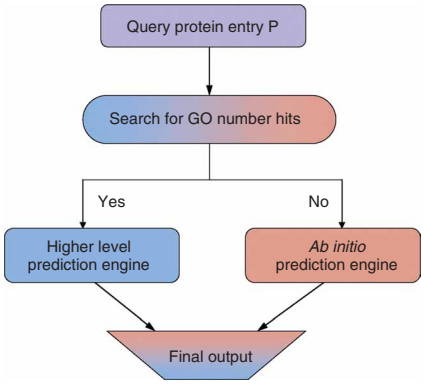

Figure 1 | A flowchart to show the process of hybridizing the higher level GO approach with the *ab initio* approach for predicting the subcellular localization of a query protein.

TABLE 2 | List of the Web servers and their Web addresses in the Cell-PLoc package that were developed recently for predicting subcellular locations of proteins in various organisms.

Predictor name	Website address	Organism	Number of subcellular locations to be covered
Euk-mPLoc[a]	http://chou.med.harvard.edu/bioinf/euk-multi/	Eukaryotic	22
Hum-mPLoc[b]	http://chou.med.harvard.edu/bioinf/hum-multi/	Human	14
Plant-PLoc	http://chou.med.harvard.edu/bioinf/plant/	Plant	11
Gpos-PLoc	http://chou.med.harvard.edu/bioinf/Gpos/	Gram-positive	5
Gneg-PLoc	http://chou.med.harvard.edu/bioinf/Gneg/	Gram-negative	8
Virus-PLoc	http://chou.med.harvard.edu/bioinf/virus/	Virus	7

[a]Evolved from Euk-PLoc[40]; Euk-mPLoc can be used to deal with proteins of multiple subcellular locations as well as single subcellular location. [b]Evolved from Hum-PLoc[35]; Hum-mPLoc can be used to deal with proteins of multiple subcellular locations as well as single subcellular location.

MATERIALS

EQUIPMENT SETUP

Hardware You need a computer with access to the Internet and a Web browser.

Data Your input protein sequences should be in FASTA format. You can either copy and paste or type the sequence of a query protein into the input box. Spaces and line breaks will be ignored and will not affect the predictions. The input sequence can be either with or without an exact accession number. If the query protein has an accession number, you should include its exact accession number as part of the input, because this will reduce the computation time and generally get a more accurate result; if no accession number is available, the prediction can still be performed by using a dummy accession number as described in Step 6 of the Procedure section below.

Programs The following predictors of the subcellular localization of proteins will be described in this protocol:

Euk-mPLoc For predicting the subcellular localization of eukaryotic proteins including those with multiple locations[81]; evolved from Euk-PLoc[40].

Hum-mPLoc For predicting the subcellular localization of human proteins including those with multiple locations[82]; evolved from Hum-PLoc[35].

Plant-Ploc For predicting the subcellular localization of plant proteins[62].

Gpos-Ploc For predicting the subcellular localization of Gram-positive bacterial proteins[83].

Gneg-PLoc For predicting the subcellular localization of Gram-negative bacterial proteins[84].

Virus-PLoc For predicting the subcellular localization of viral proteins within host and virus-infected cells[44]. See **Table 2** for a summarization of the above six Web servers and their websites.

PROCEDURE

1| Open the Web page http://chou.med.harvard.edu/bioinf/Cell-PLoc and you will see the top page of the Cell-PLoc package[6] on your computer screen, as shown in **Figure 2**.

2| If you have (i) eukaryotic protein sequences, click on the Euk-mPLoc button; (ii) human protein sequences, click Hum-mPLoc; (iii) plant sequences, click Plant-PLoc; (iv) Gram-positive bacterial protein sequences, click Gpos-PLoc; (v) Gram-negative bacterial protein sequences, click Gneg-PLoc and (vi) viral protein sequences, click Virus-PLoc.

3| For the convenience of description, let us take Euk-mPLoc as an example. After clicking Euk-mPLoc, you will see the top page of the Euk-mPLoc Web server (**Fig. 3**). To see the coverage scope, click the Read Me button and you will see the current Euk-mPLoc version can cover the following 22 subcellular location sites: (i) acrosome, (ii) cell wall, (iii) centriole, (iv) chloroplast, (v) cyanelle, (vi) cytoplasm, (vii) cytoskeleton, (viii) endoplasmic reticulum, (ix) endosome, (x) extracell, (xi) Golgi apparatus, (xii) hydrogenosome, (xiii) lysosome, (xiv) melanosome, (xv) microsome, (xvi) mitochondrion, (xvii) nucleus, (xviii) peroxisome, (xix) plasma membrane, (xx) spindle pole body, (xxi) synapse and (xxii) vacuole, as illustrated by the schematic drawing in **Figure 4**. If you already know that the query protein is not eukaryotic or is not localized in one of the above 22 locations, please stop prediction because the result obtained will not make any sense; otherwise, close the Read Me window and continue the prediction.

4| Either type or copy and paste the query protein sequence into the input box (depicted by the box at the center of **Fig. 3**). The input sequence should be in FASTA format, as shown by clicking on the Example button right above the input box.

! CAUTION If the accession number entered does not represent the true input sequence, the result obtained may not make any sense. Note that an accession number is a unique identifier given to each protein sequence once it is submitted to UniProtKB, which is composed of Swiss-Prot and TrEMBL databases. The UniProtKB database (version 12.3 released on

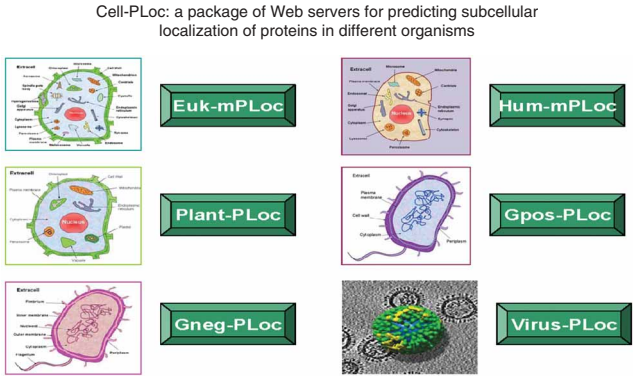

Figure 2 | Illustration to show the Cell-PLoc Web page at http://chou.med.harvard.edu/bioinf/Cell-PLoc/.

October 2, 2007) contains 5,217,756 protein sequences and hence the same number of accession numbers as well. Accession numbers are stable from release to release of the databases.

▲ CRITICAL STEP For speeding up the computation and getting a more accurate predicted result, that is, making the prediction by the 'higher level' GO approach, it is important to enter the exact accession number right above the protein sequence according to the FASTA format.

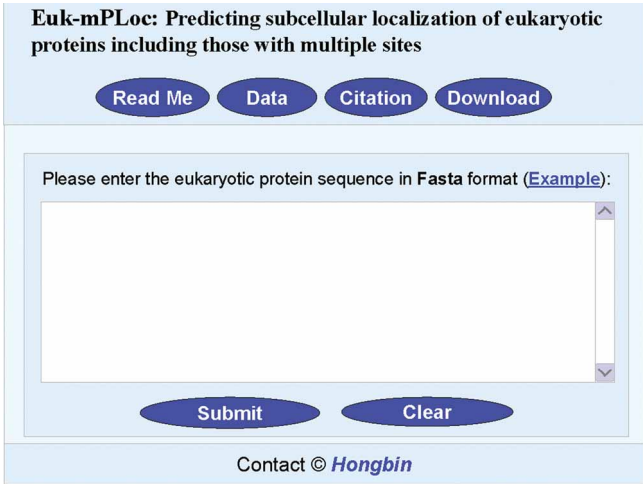

Figure 3 | An illustration to show the top page of the Web server Euk-mPLoc at http://chou.med.harvard.edu/bioinf/euk-multi/. See the text and **Table 2** for further explanation.

5| Click on the Submit button to see the subcellular location shown right under Predicted Result within a few seconds. For instance, if you use the first sequence in the Example window of Euk-mPLoc as an input, the input screen should look like the illustration in **Figure 5**; after clicking the Submit button, you will see both 'Mitochondrion' and 'Nucleus' shown on the output screen (**Fig. 6**), indicating that the query protein may exist in both the subcellular locations, fully in agreement with the experimental observations[85]. However, if you use the second sequence in the Example window as an input, you will see 'Lysosome' shown on the output screen, meaning that the query protein is localized only in the lysosome, also fully in agreement with the observation[86].

6| In the case where true accession numbers are not available for some proteins, such as synthetic and hypothetical proteins[8], Euk-mPLoc can still be used to predict their subcellular locations based on their sequences alone. Under such a circumstance, to make the prediction pass through the Web server, just add >?????? as the dummy accession number right above the query protein sequence, as shown in **Figure 7**. Thus, the prediction will be operated with the PseAAC approach, one of the state-of-the-art *ab initio* approaches, just like the case when the accession number does not have any corresponding GO number (see **Fig. 1**), and you will see the predicted result on the output screen as shown in **Figure 8**.

7| Click on the Citation button to find the relevant papers that document the detailed development and algorithm of Euk-mPLoc.

8| Click on the Data button to find all the benchmark data sets used to train and test the Euk-mPLoc predictor.

9| Click on the Download button to download the results predicted by Euk-mPLoc for all the eukaryotic protein entries (except those annotated with 'fragment' or those with less than 50 amino acids) in the Swiss-Prot database that do not have subcellular location annotations or are annotated with uncertain terms such as 'probable', 'potential', 'likely' or 'by similarity'. The large-scale predicted results have been deposited in a downloadable file prepared in 'Microsoft Excel' and 'PDF' formats. To download the former, click Tab Euk-mPLoc.xls; to download the latter, click Tab Euk-mPLoc.pdf. See **Table 3** for a few examples taken from the large-scale downloadable file. Note that the above large-scale results predicted by Euk-mPLoc will be updated periodically to include new entries of eukaryotic proteins and reflect the continuous development of Euk-mPLoc.

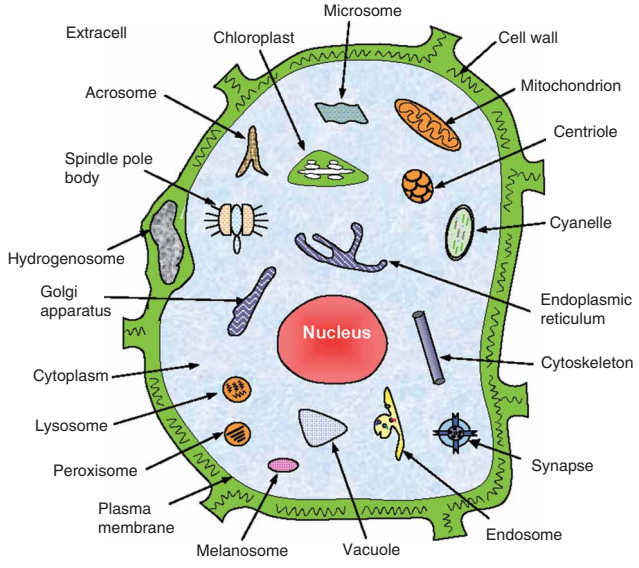

Figure 4 | A schematic illustration to show the various different components or organelles in a eukaryotic cell. Reprinted from ref. 81 with permission.

10| For predicting the subcellular localization of proteins of other organisms, and downloading their large-scale predicted results and other relevant data and information, click the corresponding Web server button (**Fig. 2**) as described in Step 2 and follow Steps 3–9.

11| To support the plant genome-sequencing projects[87,88], we have categorized the large-scale predicted results for plant proteins according to their species into the following 16 groups: (i) *Arabidopsis*, (ii) barley, (iii) *Chlamydomonas*, (iv) liverwort, (v) maize, (vi) mesostigma, (vii) pea, (viii) potato, (ix) rape, (x) rice, (xi) soybean, (xii) spinach,

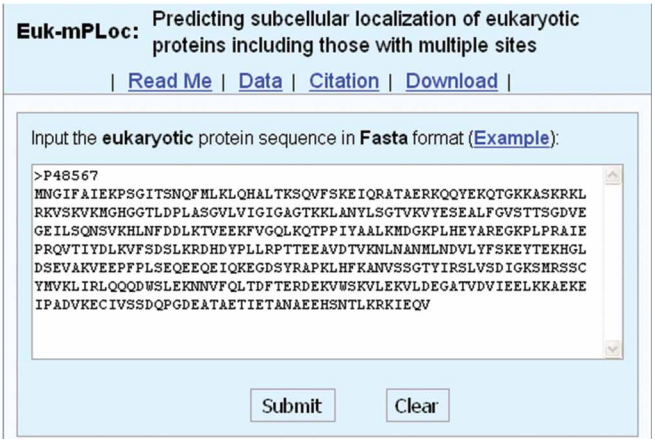

Figure 5 | A screenshot to show the input in FASTA format with the true accession number.

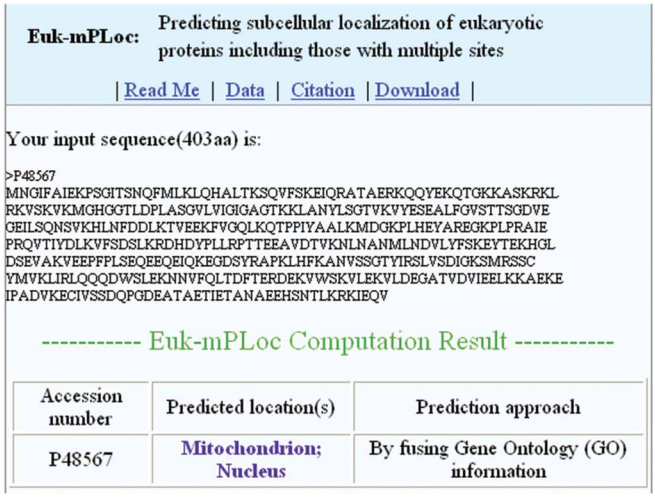

Figure 6 | A screenshot to show the output predicted by the higher level GO approach, where the predicted results are in purple.

(xiii) tobacco, (xiv) tomato, (xv) wheat and (xvi) others. To download the results thus categorized, click Tab Plant-PLoc category.xls in the Download window of Plant-PLoc.

! CAUTION For the six Web servers listed in **Table 2**, only Hum-mPLoc and Euk-mPLoc can be used to deal with both single-location and multiple-location proteins. This is because, of the human proteins with experimental location annotations, 15% are those with multiple locations[82]; of the eukaryotic proteins with experimental location annotations, 8% are those with multiple locations[81]. For the proteins in other organisms, so far such a percentage is still lower than 5%. However, as more experimental data for multiple-location proteins in these organisms become available in the future, the Cell-PLoc package will be periodically updated to deal with both single-location and multiple-location proteins in the other organisms as well.

● TIMING

The computational time for each prediction is within 5 s for most cases

? TROUBLESHOOTING

If the server does not accept the query protein input for computation, the trouble might be caused by one of the following reasons:

(1) Input is not in the FASTA format.

(2) Input sequence is less than 50 amino acids and hence it might represent a fragment rather than a real protein.

(3) The input sequence contains invalid characters; the valid single-letter characters for a protein sequence are ACDEFGHIKLMNPQRSTVWY.

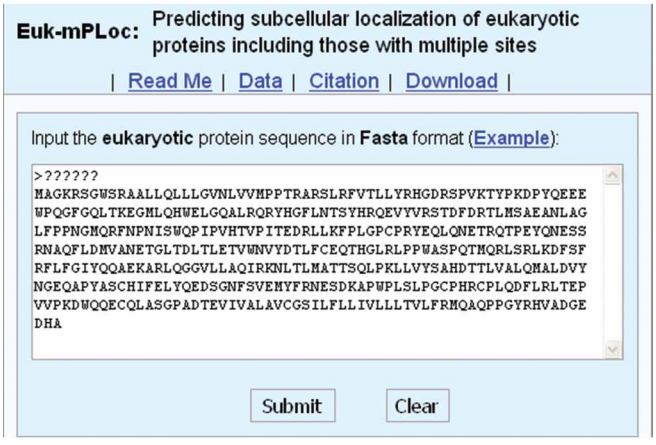

Figure 7 | A screenshot to show the input in FASTA format with the dummy accession number.

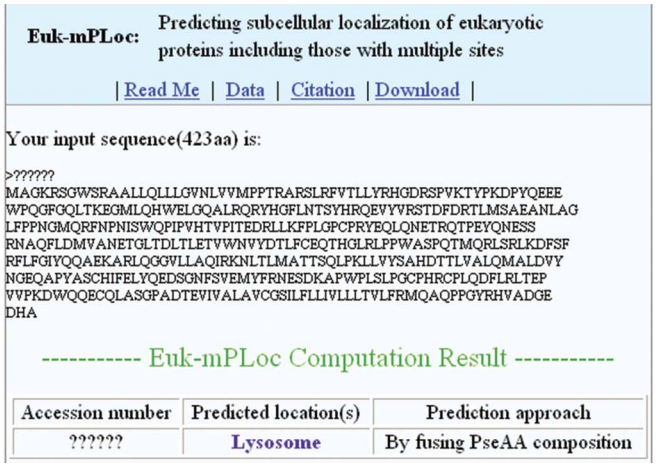

Figure 8 | A screenshot to show the output predicted by the *ab initio* PseAAC approach, where the predicted results are in purple.

TABLE 3 | Some examples from the large-scale downloadable file for the predicted results by the Euk-mPLoc for those eukaryotic proteins that either have no subcellular location annotations in databanks or are annotated with uncertain terms such as 'probable','potential' and 'by similarity' (reprinted from ref. 81 with permission).

Accession number	Swiss-Prot code	Annotation in Swiss-Prot database	Identified location by Euk-mPLoc
Q8GY58	GUN23_ARATH		Cell wall
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
Q9QZK8	DNS2A_RAT	Lysosome (by similarity)	Lysosome
P08144	AMYA_DROME		Lysosome; secreted protein
Q17029	VATF_ANOGA		Chloroplast; mitochondrion
Q8X1X3	G3P_PARBR	Cytoplasm (by similarity)	Peroxisome
Q9WTI7	MYO1C_MOUSE		Cytoskeleton
Q9USS8	YNB2_SCHPO		Centriole; cytoplasm
Q01771	STADS_BRANA	Plastid; chloroplast (probable)	Chloroplast
P07597	NLTP1_HORVU		Cell wall; cytoplasm

(4) The accession number is invalid; the accession number of the query protein has to be checked carefully as defined in Step 4 of the Procedure section. If the query protein does not have a true accession number (e.g., for the case of synthetic or hypothetical protein), just use the dummy accession number '>??????' as shown in **Figure 7**. However, as mentioned above, prediction by using a dummy accession number might yield less accurate result and take longer computation time.

Odd results/anomalies might also occur if the query protein is outside the subcellular locations covered by the Web-server predictor.

ANTICIPATED RESULTS

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent data set test, subsampling test and jackknife test. In the independent data set test, although none of the proteins to be tested occurs in the training data set used to train the predictor, the selection of proteins for the testing data set could be quite arbitrary unless it is sufficiently large. This kind of arbitrariness may directly affect the conclusion. For instance, a predictor yielding higher success rate than the others for a testing data set might fail to yield so when applied to another testing data set[89]. For the subsampling test, the practical procedure often used in literatures is the fivefold, sevenfold or tenfold cross-validation. The problem with the subsampling examination as such is that the number of possible selections in dividing a benchmark data set is an astronomical figure even for a very simple data set[6]. Therefore, any practical result by the subsampling test alone represents one of many possible results, and hence cannot avoid the arbitrariness either. In the jackknife cross-validation, each of the protein samples in the benchmark data set is in turn singled out as a tested protein and the predictor is trained by the remaining proteins. During the jackknifing process, both training and testing data sets are actually open, and a protein will in turn move from one to the other. The jackknife cross-validation can exclude the memory effects during the entire testing process and, also, the result thus obtained is always unique for a given benchmark data set. Therefore, of the above three examination methods, the jackknife test is deemed the most objective[89], and has been increasingly used by investigators to examine the accuracy of various predictors[24,25,28,33,36,39,42,45,50–56,59,70,90–96].

However, even being tested by the jackknife cross-validation, the same predictor can still yield different success rates for different benchmark data sets. Generally speaking, more stringent the threshold to exclude homologous sequences from a benchmark data set, or larger the number of the subcellular locations it covers, lower the corresponding overall success rate yielded. For instance, some *ab initio* predictors based on state-of-the-art techniques such as SVM could yield an overall success rate of higher than 80% for a high homologous benchmark data set that contains proteins with pairwise sequence identity up to 90% and covers only four subcellular location sites. However, when tested by the low homologous benchmark data set that only contains proteins with pairwise sequence identity lower than 25% and covers 16 subcellular location sites, these same predictors could only yield an overall success rate of lower than 35% (see ref. 40).

On the other hand, for the predictors formed by hybridizing the 'higher level' and the *ab initio* approaches such as Euk-PLoc[40], Hum-PLoc[35], Plant-PLoc[62], Gpos-PLoc[83], Gneg-PLoc[84] and Virus-PLoc[44], even when tested by the very stringent benchmark data sets that only contain protein samples with pairwise sequence identity lower than 25% and cover up to 16 subcellular location sites, the overall jackknife success rates can still reach 71% to ~87% (**Table 4**). Particularly, for the multiple-location

TABLE 4 | The overall success rates by the jackknife cross-validation tests conducted on a series of stringent benchmark datasets.

Predictor	Organism	Can it deal with the systems also including multiplex proteins?[a]	Number of location sites covered	Pairwise sequence identity percentage allowed in the benchmark dataset (%)	Jackknife success rate (%)	Reference
Euk-PLoc	Eukaryotic	No	16	<25[b]	81.6	40
Euk-mPLoc	Eukaryotic	Yes	22	<25	67.4	81
Hum-PLoc	Human	No	12	<25	81.1	35
Hum-mPLoc	Human	Yes	16	<25	70.8	82
Plant-PLoc	Plant	No	11	<25	71.4	62
Gpos-PLoc	Gram-positive	No	5	<25	82.7	83
Gneg-PLoc	Gram-negative	No	8	<25	87.3	84
Virus-PLoc	Virus	No	4	<25	80.0	44
Virus-PLoc	Virus	No	7	<80[c]	89.2	44

[a]Here the so-called multiplex protein means that it can simultaneously exist at, or move between, two or more different subcellular locations. [b]Meaning that none of the proteins included in the benchmark dataset have ≥25% sequence identity to any other protein in the same subcellular location. [c]Meaning that none of the proteins included in the benchmark dataset have ≥80% sequence identity to any other protein in the same subcellular location.

predictors, such as Euk-mPLoc[81] and Hum-mPLoc[82], when tested by the benchmark data sets with the same strict threshold to exclude homologous proteins and covering up to 22 subcellular location sites, the overall jackknife success rates can still reach 67% to ~70% (**Table 4**). As can be conceived, it is much more difficult to get a decent overall success rate for a benchmark data set that also contains multiple-location proteins[6]. Moreover, besides Euk-mPLoc and Hum-PLoc, so far no other Web server can be used to deal with both single-location and multiple-location proteins[43].

Concluding remarks

With the explosion of newly found protein sequences entering into protein databanks in the post-genomic age, it is highly desired to develop an automated method by which one can get a fast and often a reliable suggestion for the localization of an uncharacterized protein. In comparison with many of the existing predictors that are based on the *ab initio* approach alone, the Web servers can, by hybridizing the 'high level' and *ab initio* approaches as described in this article, yield much higher success rates and cover much wider scope and hence are more powerful in practical applications. However, there is much room for further improvement.

Although Plant-PLoc[62] in the current version of Cell-PLoc package can cover 11 subcellular location sites of plant proteins, which is much more than three or four location sites covered by TargetP[18], if a query protein is outside of the 11 location sites, the predicted result would still be meaningless. A similar limitation in the coverage scope also exists for the other Web-server predictors in Cell-PLoc. In view of this, as more experimental subcellular location data become available, we will periodically expand the coverage scope for the Web servers in the future version of Cell-PLoc.

Owing to the reasons described in Step 11 of the Procedure section, for the current version of Cell-PLoc, only two predictors (Hum-mPLoc and Euk-mPLoc) can be used to deal with a biological system that contains both single-location and multiple-location proteins. As more experimental multiple-location data become available for plant, Gram-positive bacterial, Gram-negative bacterial and viral proteins, we will periodically convert Plant-PLoc, Gpos-PLoc, Gneg-PLoc and Virus-PLoc to Plant-mPLoc, Gpos-mPLoc, Gneg-mPLoc and Virus-mPLoc, respectively, so as to enable these Web-server predictors in the future version of Cell-PLoc to deal with multiplex proteins as well.

To enhance the prediction quality for those proteins that have no corresponding GO numbers or accession numbers (such as synthetic and hypothetical proteins), we will continue to improve the prediction engine and protein descriptor, compare the results obtained by different powerful *ab initio* predictors on various stringent benchmark data sets and periodically use the new state-of-the-art *ab initio* approach to replace the current one in the future version of Cell-PLoc.

Once a future version of Cell-PLoc is established, we will make an announcement through a Web page or a publication.

ACKNOWLEDGMENTS We express our gratitude to the editor and the anonymous reviewers for their valuable suggestions that were very helpful for strengthening the presentation of this article.

Published online at http://www.natureprotocols.com
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Ehrlich, J.S., Hansen, M.D. & Nelson, W.J. Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell–cell adhesion. *Dev. Cell* **3**, 259–270 (2002).
2. Glory, E. & Murphy, R.F. Automated subcellular location determination and high-throughput microscopy. *Dev. Cell* **12**, 7–16 (2007).
3. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* **25**, 31–36 (2000).
4. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
5. Hill, D.P., Blake, J.A., Richardson, J.E. & Ringwald, M. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res.* **12**, 1982–1991 (2002).
6. Chou, K.C. & Shen, H.B. Review: recent progresses in protein subcellular location prediction. *Anal. Biochem.* **370**, 1–16 (2007).

7. Chou, K.C. Review: structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **11**, 2105–2134 (2004).
8. 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.* **77**, 90–127 (2005).
9. Nakai, K. & Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14**, 897–911 (1992).
10. Nakashima, H. & Nishikawa, K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* **238**, 54–61 (1994).
11. Cedano, J., Aloy, P., P'erez-Pons, J.A. & Querol, E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **266**, 594–600 (1997).
12. Nakai, K. & Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34–36 (1999).
13. Reinhardt, A. & Hubbard, T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* **26**, 2230–2236 (1998).
14. Chou, K.C. & Elrod, D.W. Protein subcellular location prediction. *Protein Eng.* **12**, 107–118 (1999).
15. Yuan, Z. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* **451**, 23–26 (1999).
16. Nakai, K. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* **54**, 277–344 (2000).
17. Murphy, R.F., Boland, M.V. & Velliste, M. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 251–259 (2000).
18. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their *N*-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
19. Feng, Z.P. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers* **58**, 491–499 (2001).
20. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17**, 721–728 (2001).
21. Feng, Z.P. & Zhang, C.T. Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *Int. J. Biol. Macromol.* **28**, 255–261 (2001).
22. Feng, Z.P. An overview on predicting the subcellular location of a protein. *In Silico Biol.* **2**, 291–303 (2002).
23. Chou, K.C. & Cai, Y.D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **277**, 45765–45769 (2002).
24. Zhou, G.P. & Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins* **50**, 44–48 (2003).
25. Pan, Y.X. *et al.* Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.* **22**, 395–402 (2003).
26. Park, K.J. & Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. *Bioinformatics* **19**, 1656–1663 (2003).
27. Gardy, J.L. *et al.* PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* **31**, 3613–3617 (2003).
28. Huang, Y. & Li, Y. Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* **20**, 21–28 (2004).
29. Xiao, X. *et al.* Using complexity measure factor to predict protein subcellular location. *Amino Acids* **28**, 57–61 (2005).
30. Lei, Z. & Dai, Y. An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics* **6**, 291 (2005).
31. Garg, A., Bhasin, M. & Raghava, G.P. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.* **280**, 14427–14432 (2005).
32. Matsuda, S. *et al.* A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci.* **14**, 2804–2813 (2005).
33. Gao, Q.B., Wang, Z.Z., Yan, C. & Du, Y.H. Prediction of protein subcellular location using a combined feature of sequence. *FEBS Lett.* **579**, 3444–3448 (2005).
34. Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D. & Chou, K.C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* **30**, 49–54 (2006).
35. Chou, K.C. & Shen, H.B. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* **347**, 150–157 (2006).
36. Guo, J., Lin, Y. & Liu, X. GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. *Proteomics* **6**, 5099–5105 (2006).
37. Hoglund, A., Donnes, P., Blum, T., Adolph, H.W. & Kohlbacher, O. MultiLoc: prediction of protein subcellular localization using *N*-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* **22**, 1158–1165 (2006).
38. 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.* **34**, 4655–4666 (2006).
39. 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.* **580**, 6169–6174 (2006).
40. Chou, K.C. & Shen, H.B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res.* **5**, 1888–1897 (2006).
41. Pierleoni, A., Martelli, P.L., Fariselli, P. & Casadio, R. BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* **22**, e408–e416 (2006).
42. Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.-M. & Xie, J. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids* **33**, 69–74 (2007).
43. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971 (2007).
44. 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* **85**, 233–240 (2007).
45. Chen, Y.L. & Li, Q.Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *J. Theor. Biol.* **248**, 377–381 (2007).
46. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
47. Nair, R. & Rost, B. Sequence conserved for subcellular localization. *Protein Sci.* **11**, 2836–2847 (2002).
48. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins,* (Erratum: *ibid.*, 2001, Vol. 44, 60) **43**, 246–255 (2001).
49. Chou, K.C. & Shen, H.B. Predicting protein subcellular location by fusing multiple classifiers. *J. Cell. Biochem.* **99**, 517–527 (2006).
50. 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.* **357**, 116–121 (2006).
51. 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.* **243**, 444–448 (2006).
52. 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* **30**, 461–468 (2006).
53. Du, P. & Li, Y. Prediction of protein submitochondrial locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformatics* **7**, 518 (2006).
54. 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.* **243**, 252–260 (2006).
55. Lin, H. & Li, Q.Z. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. *Biochem. Biophys. Res. Commun.* **354**, 548–551 (2007).
56. Lin, H. & Li, Q.Z. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. *J. Comput. Chem.* **28**, 1463–1466 (2007).
57. Pu, X., Guo, J., Leung, H. & Lin, Y. Prediction of membrane protein types from sequences and position-specific scoring matrices. *J. Theor. Biol.* **247**, 259–265 (2007).
58. Kurgan, L.A., Stach, W. & Ruan, J. Novel scales based on hydrophobicity indices for secondary protein structure. *J. Theor. Biol.* **248**, 354–366 (2007).
59. Zhou, X.B., Chen, C., Li, Z.C. & Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.* **248**, 546–551 (2007).
60. Mundra, P., Kumar, M., Kumar, K.K., Jayaraman, V.K. & Kulkarni, B.D. Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. *Pattern Recogn. Lett.* **28**, 1610–1615 (2007).
61. Shen, H.B. & Chou, K.C. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* doi: 10.1016/j.ab.2007.10.012 (2007).
62. Chou, K.C. & Shen, H.B. Large-scale plant protein subcellular location prediction. *J. Cell. Biochem.* **100**, 665–678 (2007).

63. Vapnik, V. *Statistical Learning Theory* (Wiley-Interscience, New York, 1998).
64. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
65. Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32**, W375–W379 (2004).
66. Chou, K.C. & Shen, H.B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **357**, 633–640 (2007).
67. Shen, H.B. & Chou, K.C. Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Commun.* **363**, 297–303 (2007).
68. Regev-Rudzki, N. & Pines, O. Eclipsed distribution: a phenomenon of dual targeting of protein and its significance. *Bioessays* **29**, 772–782 (2007).
69. Lubec, G. & Afjehi-Sadat, L. Limitations and pitfalls in protein identification by mass spectrometry. *Chem. Rev.* **107**, 3568–3584 (2007).
70. Jahandideh, S., Abdolmaleki, P., Jahandideh, M. & Asadabadi, E.B. Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. *Biophys. Chem.* **128**, 87–93 (2007).
71. Afjehi-Sadat, L. *et al.* Structural and functional analysis of hypothetical proteins in mouse hippocampus from two-dimensional gel electrophoresis. *J. Proteome Res.* **6**, 711–723 (2007).
72. Diao, Y. *et al.* Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel–Ziv complexity. *Amino Acids,* doi: 10.1007/s00726-007-0550-z (2007).
73. Chen, Y.L. & Li, Q.Z. Prediction of the subcellular location of apoptosis proteins. *J. Theor. Biol.* **245**, 775–783 (2007).
74. Ho, V.S.M. & Ng, T.Z. Chitinase-like proteins with antifungal activity from emperor banana fruits. *Protein Pept. Lett.* **14**, 828–831 (2007).
75. Chou, K.C. & Cai, Y.D. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem. Biophys. Res. Commun.* **320**, 1236–1239 (2004).
76. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40 (2001).
77. Chou, K.C. & Cai, Y.D. Predicting protein structural class by functional domain composition. *Biochemical and Biophysical Research Communications,* Corrigendum: *ibid.*, 2005, Vol. 329, 1362 **321**, 1007–1009 (2004).
78. Cover, T.M. & Hart, P.E. Nearest neighbour pattern classification. *IEEE Trans. Inf. Theor. IT* **13**, 21–27 (1967).
79. Denoeux, T. A k-nearest neighbor classification rule based on Dempster–Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**, 804–813 (1995).
80. Zouhal, L.M. & Denoeux, T. An evidence-theoretic k-NN rule with parameter optimization. *IEEE Trans. Syst. Man Cybern.* **28**, 263–271 (1998).
81. Chou, K.C. & Shen, H.B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **6**, 1728–1734 (2007).
82. Shen, H.B. & Chou, K.C. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun.* **355**, 1006–1011 (2007).
83. Shen, H.B. & Chou, K.C. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng. Des. Sel.* **20**, 39–46 (2007).
84. Chou, K.C. & Shen, H.B. Large-scale predictions of Gram-negative bacterial protein subcellular locations. *J. Proteome Res.* **5**, 3420–3428 (2006).
85. Becker, H.F., Motorin, Y., Planta, R.J. & Grosjean, H. The yeast gene YNL292w encodes a pseudouridine synthase (Pus4) catalyzing the formation of psi55 in both mitochondrial and cytoplasmic tRNAs. *Nucleic Acids Res.* **25**, 4493–4499 (1997).
86. Geier, C., von Figura, K. & Pohlmann, R. Structure of the human lysosomal acid phosphatase gene. *Eur. J. Biochem.* **183**, 611–616 (1989).
87. Jorgensen, R. Plant genomes. *Plant Cell* **18**, 1099 (2006).
88. Jackson, S., Rounsley, S. & Purugganan, M. Comparative sequencing of plant genomes: choices to make. *Plant Cell* **18**, 1100–1104 (2006).
89. Chou, K.C. & Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30**, 275–349 (1995).
90. Zhou, G.P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **17**, 729–738 (1998).
91. Cao, Y. *et al.* Prediction of protein structural class with rough sets. *BMC Bioinformatics* **7**, 20 (2006).
92. Gao, Q.B. & Wang, Z.Z. Classification of G-protein coupled receptors at four levels. *Protein Eng. Des. Sel.* **19**, 511–516 (2006).
93. Kedarisetti, K.D., Kurgan, L.A. & Dick, S. Classifier ensembles for protein structural class prediction with varying homology. *Biochem. Biophys. Res. Commun.* **348**, 981–988 (2006).
94. Zhou, G.P. & Cai, Y.D. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. *Proteins* **63**, 681–684 (2006).
95. Chou, K.C. & Shen, H.B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **360**, 339–345 (2007).
96. Shen, H.B. & Chou, K.C. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.*, **364**, 53–59 (2007).