Vol.2, No.10, 1090-1103 (2010)
http://dx.doi.org/10.4236/ns.2010.210136

Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms

Kuo-Chen Chou[1,2], **Hong-Bin Shen**[1,2]

[1]Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA; kcchou@gordonlifescience.org
[2]Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai, 200240, China; hbshen@sjtu.edu.cn

Received 5 August 2010; revised 8 September 2010; accepted 12 September 2010.

ABSTRACT

Cell-PLoc 2.0 is a package of web-servers evolved from Cell-PLoc (Chou, K.C. & Shen, H.B., Nature Protocols, 2008, 2:153-162) by a top-down approach to improve the power for predicting subcellular localization of proteins in various organisms. It contains six predictors: Euk-mPLoc 2.0, Hum-mPLoc 2.0, Plant-mPLoc, Gpos-mPLoc, Gneg-mPLoc, and Virus-mPLoc, specialized for eukaryotic, human, plant, Gram-positive bacterial, Gram-negative bacterial, and virus proteins, respectively. Compared with Cell-PLoc, the predictors in the Cell-PLoc 2.0 have the following advantageous features: (1) they all have the capacity to deal with the multiplex proteins that can simultaneiously exist, or move between, two or more subcellular location sites; (2) no accession number is needed for the input of a query protein even if using the "high-level" GO (gene ontology) prediction engine; (3) the functional domain information and sequential evolution information are fused into the "*ab initio*" sequence-based prediction engine to enhance its accuracy. In this protocol, a step-to-step guide is provided for how to use the web server predictors in the Cell-PLoc 2.0 package, which is freely accessible to the public at http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/.

Keywords: Euk-mPLoc 2.0; Hum-mPLoc 2.0; Plant-mPLoc; Gpos-mPLoc; Gneg-mPLoc; Virus-mPLoc; Higher-level GO approach; Ab-initio approach; Functional domain; Sequential evolution; Multiplex proteins

1. INTRODUCTION

The localization of a protein in a cell is one of its most important attributes. It can provide useful insight about the function of the protein. It is also fundamental to system biology because knowledge of the subcellular locations of proteins is indispensable for in-depth understanding how the biological processes are regulated by the intricate pathways at the cellular level [1,2]. Particularly, the information of protein subcellular location is very useful for identifying and prioritizing drug targets [3] during the process of drug development.

Given an uncharacterized protein sequence, how can we identify which subcellular location site it resides at? Does the protein stay in a single subcellular location or can it simultaneously exist in, or move between, two and more subcellular location sites? Although the answers to these questions can be determined by means of various biochemical experiments, it is time-consuming and laborious to acquire the desired information with experimental methods alone. Particularly, in the post-genomic age, the number of newly found protein sequences has increased explosively. For instance, in 1986 the Swiss-Prot databank contained merely 3,939 protein sequence entries, but the number has since jumped to 519,348 according to the data released by the same databank on 10-Aug-2010 (www.expasy.org/sprot/relnotes/relstat.html), meaning that the number of protein sequence entries now is more than 131 times the number from about 24 years ago. Facing such an avalanche of protein sequences, it is highly desired to develop automated methods for timely identifying the subcellular locations of uncharacterized proteins based on their sequence information alone.

Actually, during the past 18 years or so, various computational methods were developed in this regard (see, e.g., [4-59].

All the aforementioned methods each have their own advantages and have indeed played a role in stimulating the development of this area. Meanwhile, they also each have their own limitations. For example, TargetP [15] is one of the popular methods in this area. Its remarkable merit is to make the prediction of the subcellular location of a protein related to its signal peptide and hence has a clearer biological meaning and basis. But TargetP [15] can only cover four subcellular location sites. For a query protein located outside its coverage scope, TargetP would either fail to predict or the predicted result thus obtained would not make any sense. The similar problem also exists for PSORTb [33], one of the other popular methods in this area.

The other problem for the existing methods listed above is that none of them can be used to deal with multiplex proteins that may simultaneously reside at, or move between, two or more different subcellular locations. Proteins with multiple location sites or dynamic feature of this kind are particularly interesting because they may have some unique biological functions worthy of our special notice [2,3]. Particularly, as pointed out by Millar *et al.* [60], recent evidence indicates that an increasing number of proteins have multiple locations in the cell.

About two years ago, a package of web-servers called **Cell-PLoc** was published [61] that can be used to predict subcellular localization of proteins in various organisms. It contained six web-server predictors: **Euk-mPLoc** [62], **Hum-mPLoc** [63], **Plant-PLoc** [64], **Gpos-PLoc** [65], **Gneg-PLoc** [66], and **Virus-PLoc** [67], specialized for eukaryotic, human, plant, Gram-positive bacterial, Gram-negative bacterial, and virus proteins, respectively. As elucidated in the protocol article [61], each of the six predictors in **Cell-PLoc** was established by hybridizing the "higher-level" GO (gene ontology) [68] approach and the "*ab initio*" PseAAC (pseudo amino acid composition) [16] approach, and hence could yield higher success rates as well as cover much wider scope. For example, the **Euk-mPLoc** predictor can cover up to 22 subcellular location sites. Moreover, of the six predictors in the **Cell-PLoc** package [61], **Euk-mPLoc** and **Hum- mPLoc** can be also used to deal with proteins with multiple-location sites. Therefore, ever since it was published, **Cell-PLoc** has been widely and increasingly used.

However, the existing version of **Cell-PLoc** [61] has the following shortcomings. **(1)** The accession number of a query protein is indispensable as an input in order to utilize the advantage of the "higher-level" GO approach. Many proteins, such as hypothetical and synthetic proteins as well as those newly-discovered proteins that have not been deposited into databanks yet, do not have accession numbers, and hence cannot be handled with the GO approach. **(2)** Even with their accession numbers available, many proteins cannot be meaningfully formulated in a GO space because the current GO database is far from complete yet. **(3)** Although the PseAAC approach was used as a complement in **Cell-PLoc** [61] that could take some partial sequence order effects into account, the original PseAAC [16,69] did not contain the sequential evolution and functional domain information, and hence would affect the prediction quality. **(4)** Except **Euk-mPLoc** (the predictor for eukaryotic proteins) and **Hum-mPLoc** (the predictor for human proteins), all the other predictors in **Cell-PLoc** package [61] cannot be used to deal with multiplex proteins.

To address the aforementioned four problems, a top-down approach to enhance the power of **Cell-PLoc** has been implemented. The new version thus obtained is denoted by **Cell-PLoc 2.0**. Compared with the old **Cell-PLoc** [61], **Cell-PLoc 2.0** has the following advantageous features.

Input Data. By means of the "homology-based GO extraction" strategy as developed recently (see, e.g., [70]), the requirement for the accession number of a query protein is no longer needed even if using the higher-level GO approach to perform the prediction. This is especially useful for predicting the subcellular location sites of hypothetical proteins or synthetic proteins, as well as those new protein sequences without being deposited into data banks and hence having no accession numbers assigned yet.

Sequence Information. For those proteins that have no useful GO information to carry out the higher-level prediction, a hybridization approach by fusing the functional domain information and sequential evolution information as illustrated in **Figure 1** is developed to replace the simple PseAAC approach [16] in the old **Cell-PLoc** [61]. As a consequence, the success rates have been remarkably increased for those proteins without useful GO numbers.

Multiplex Proteins. In the old **Cell-PLoc** package [61], only two predictors, i.e., the one specialized for eukaryotic proteins and the one specialized for human proteins, can be used to treat proteins with multiple location sites. In **Cell-PLoc 2.0**, all the six predictors, including those specialized for plant proteins, Gram- positive bacterial proteins, Gram-negative bacterial proteins, and virus proteins, can be used to deal with the multiplex proteins.

Benchmark Datasets. With more experimental data available in Swiss-Prot database (www.ebi.ac.uk/swissprot), to update the data for training the predictors, instead of version 50.7 released on 9-Sept-2006 as used in the old **Cell-PLoc** [61], the benchmark datasets for training the

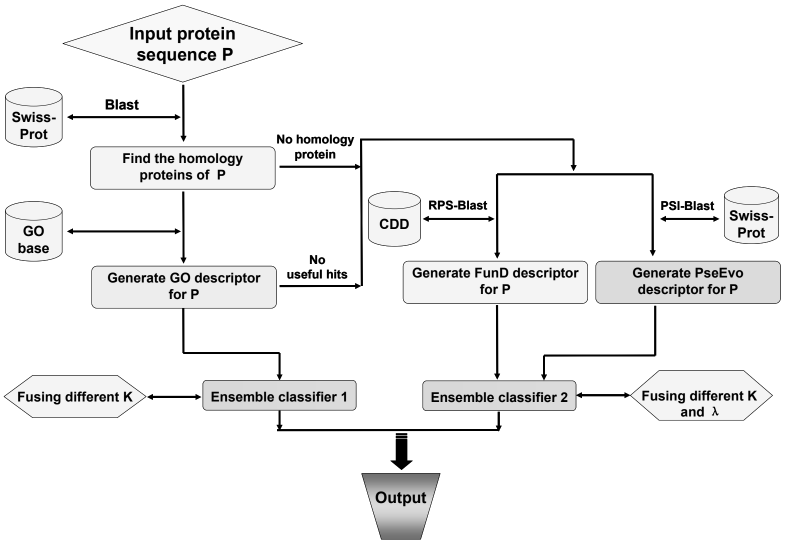


Figure 1. A flowchart to show the prediction process of the predictors in Cell-PLoc 2.0, where ensemble classifier 1 is for processing the GO descriptor samples, while ensemble classifier 2 is for the FunD (functional domain) and PseEvo (pseudo sequential evolution) descriptor samples. See [70,71] for further explanation.

predictors in **Cell-PLoc 2.0** were constructed based on version 55.3 released on 29-April-2008. Moreover, to make all the predictors in **Cell-PLoc 2.0** have the capacity to deal with the multiplex proteins as well, the sequences annotated with two or more subcellular location sites were no longer excluded even for plant proteins, Gram-positive bacterial proteins, Gram-negative bacterial proteins, and virus proteins as done previously in the old **Cell-PLoc** package [61].

Below, let us describe how to use the new **Cell-PLoc 2.0** package to get the desired results.

2. EQUIPMENT AND MATERIALS

Hardware. Same as in the old **Cell-PLoc** [61], i.e., you need a computer that is able to access to internet.

Data. Your input protein sequences should be in FASTA format. You can enter the sequence of a query protein by either typing or copying-and-pasting it into the input box. Spaces and line breaks will be ignored and will not affect the prediction result.

Programs. Cell-PLoc 2.0 contains the following programs: (1) **Euk-mPLoc 2.0** for predicting the subcellular localization of eukaryotic proteins; (2) **Hum-mPLoc 2.0** for human proteins; (3) **Plant-mPLoc** for plant proteins; (4) **Gpos-mPLoc** for Gram-positive bacterial proteins; (5) **Gneg-mPLoc** for Gram-negative bacterial proteins; (6) **Virus-mPLoc** for virus proteins. The six predictors were evolved from Euk-mPLoc [62], Hum-mPLoc [63], Plant- PLoc [64], Gpos-PLoc [65], Gneg-PLoc [66], and Virus- PLoc [67] in the original **Cell-PLoc** package [61] through a top-down approach to enhance their power, as elaborated in [70-75], respectively. Note that now all the six predictors in **Cell-PLoc 2.0** have the capacity to deal with multiplex proteins as well, as indicated by the character "m" in front of their partial name "PLoc" that stands for the first character of "multiple".

3. PROCEDURE

1) Go to the internet at http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/ and you will see the top page of the **Cell-PLoc 2.0** package on the screen of your computer, as shown in **Figure 2**.

2) You should use the relevant predictor to conduct the prediction: (1) if your query protein is an eukaryotic one, click the button Euk-mPLoc 2.0**;** (2) if it is a human protein, click Hum-mPLoc 2.0; (3) if it is a plant protein, click Plant-mPLoc; (4) if it is a Gram-positive bacterial protein, click Gpos-mPLoc; (5) if it is a Gram- negative bacterial protein, click Gneg-mPLoc; (6) if it is a viral protein, click Virus-mPLoc**.**

3) Without loss of generality, let us take **Hum-mPLoc 2.0** as an example. By clicking Hum-mPLoc 2.0, you will be prompted with the top page of the **Hum-mPLoc 2.0** web-server predictor (**Figure 3**). To find the coverage scope and caveat in using the predictor, click the Read Me button and you will see that the current **Hum-mPLoc 2.0** version can cover the following 14 human protein subcellular location sites: (1) centriole, (2) cytoplasm, (3) cytoskeleton, (4) endoplasmic reticulum, (5) endosome, (6) extracell, (7) Golgi apparatus, (8) ly-

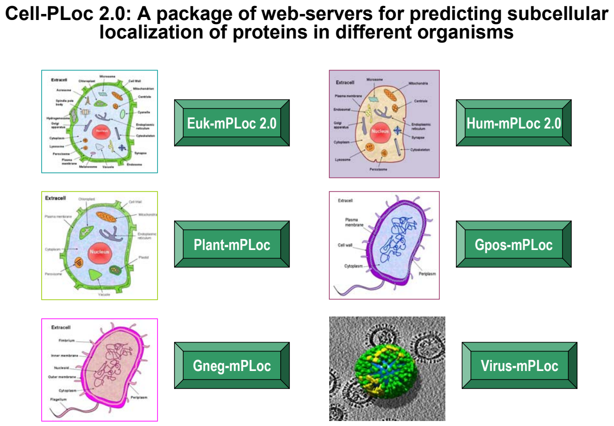
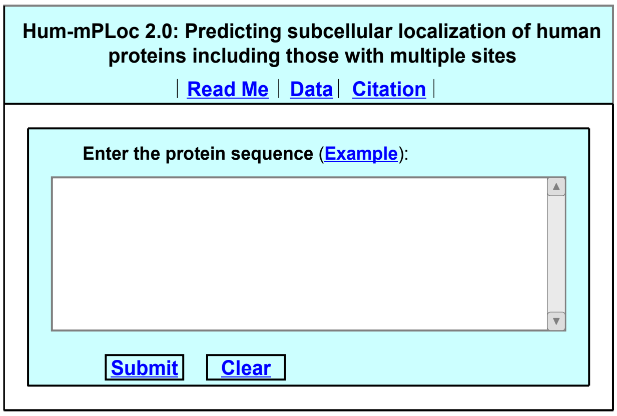

Cell-PLoc 2.0: A package of web-servers for predicting subcellular localization of proteins in different organisms
Euk-mPLoc 2.0
Hum-mPLoc 2.0
Plant-mPLoc
Gpos-mPLoc
Gneg-mPLoc
Virus-mPLoc

Figure 2. Illustration to show the **Cell-PLoc 2.0** web-page at http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/.


Hum-mPLoc 2.0: Predicting subcellular localization of human proteins including those with multiple sites

Enter the protein sequence (Example):

Figure 3. A semi-screenshot to show the top page of the web-server predictor **Hum-mPLoc 2.0** in the **Cell-PLoc 2.0** package.

sosome, (9) microsome, (10) mitochondrion, (11) nucleus, (12) peroxisome, (13) plasma membrane, and (14) synapse, as schematically shown in **Figure 4**. You will also see the caveat from the Read Me window how to avoid meaningless prediction. To continue the prediction, go back to the top page of the **Hum-mPLoc 2.0** web-server predictor by closing the Read Me window.

4) Enter your query protein sequence into the input box as shown at the centre of **Figure 3**. The input sequence should be in FASTA format. A sequence in FASTA format consists of a single-line description, followed by lines of sequence data. The first character of the description line is a greater-than symbol (">") in the first column. All lines should be shorter than 80 characters. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box. For more information about FASTA format, visit http://en.wikipedia.org/wiki/Fasta_format.

5) To get the predicted result, click the Submit button.

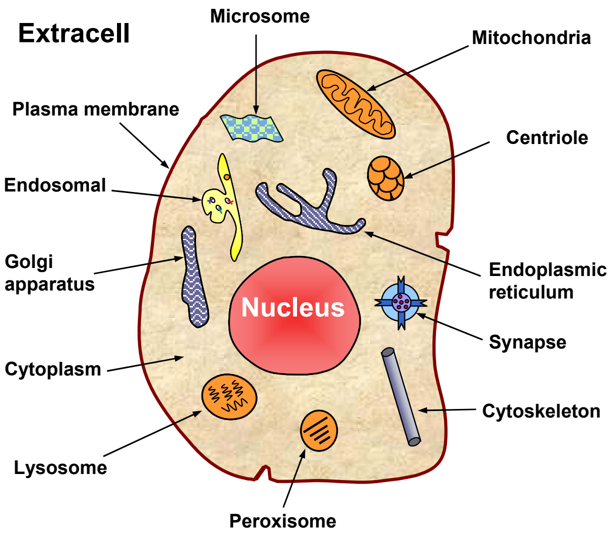


Figure 4. Schematic illustration to show the fourteen subcellular location sites of human proteins that are covered by the **Hum-mPLoc 2.0** predictor.

For example, if using the sequence of query protein 1 in the Example window as an input, you will see the input screen as shown in **Figure 5a**; after clicking the Submit button, you will see "**Cell membrane; Cytoplasm; Nucleus**" shown on the predicted location(s) window (**Figure 5b**), meaning that the query protein is a multiplex protein, which can simultaneously occur in "cell membrane", "cytoplasm" and "nucleus" sites, fully consistent with experimental observations. However, if using the sequence of query protein 2 in the Example window as an input, you will instead see the input screen as shown in **Figure 6a**; after clicking the Submit button, you will see "**Cytoplasm**" shown on the predicted location(s) window (**Figure 6b**), meaning that the query protein is a single-location protein residing in "cytoplasm" compartment only, also fully consistent with experimental observations.

6) By clicking the Citation button, you will find the relevant papers that document the detailed development and algorithm of **Hum-mPLoc 2.0**.

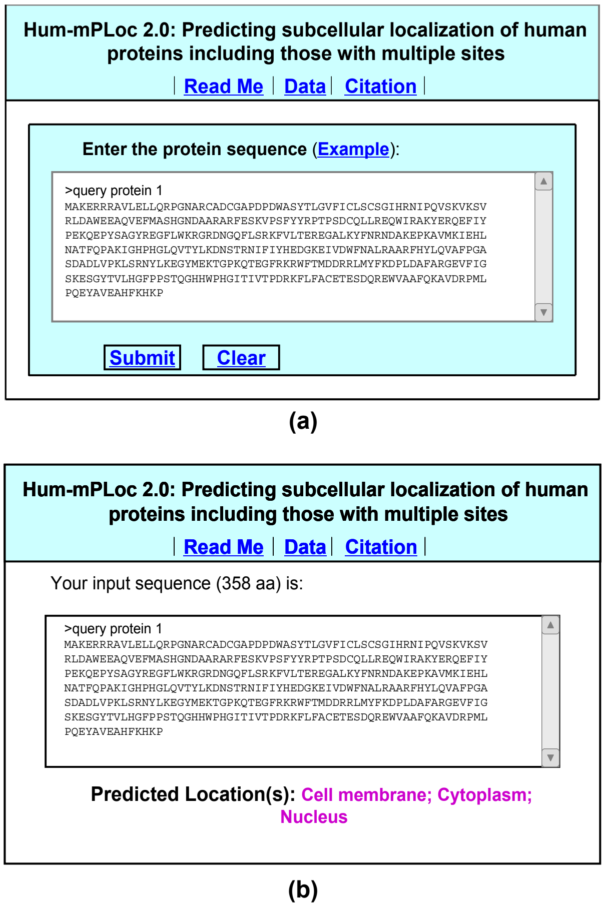
7) By clicking the Data button, you will find all the benchmark datasets used to train and test the **Hum-mPLoc 2.0** predictor.

8) If your query protein sequence is from other organism, click the relevant web-server button (**Figure 2**) as elaborated in **Step 2**, and repeat **Steps 3-6**.

TIMING The computational time for each prediction is within 15 seconds for most cases. The longer the query protein sequence is, the more time it is usually needed.

4. TROUBLESHOOTING

After you click the Submit button, if the server rejects



Hum-mPLoc 2.0: Predicting subcellular localization of human proteins including those with multiple sites

Enter the protein sequence (Example):

```
>query protein 1
MAKERRRAVLELLQRPGNARCADCGAPDPDWASYTLGVFICLSCSGIHRNIPQVSKVKSV
RLDAWEEAQVEFMASHGNDAARARFESKVPSFYYRPTPSDCQLLREQWIRAKYERQEFIY
PEKQEPYSAGYREGFLWKRGRDNGQFLSRKFVLTEREGALKYFNRNDAKEPKAVMKIEHL
NATFQPAKIGHPHGLQVTYLKDNSTRNIFIYHEDGKEIVDWFNALRAARFHYLQVAFPGA
SDADLVPKLSRNYLKEGYMEKTGPKQTEGFRKRWFTMDDRRLMYFKDPLDAFARGEVFIG
SKESGYTVLHGFPPSTQGHHWPHGITIVTPDRKFLFACETESDQREWVAAFQKAVDRPML
PQEYAVEAHFKHKP
```

Hum-mPLoc 2.0: Predicting subcellular localization of human proteins including those with multiple sites

Your input sequence (358 aa) is:

```
>query protein 1
MAKERRRAVLELLQRPGNARCADCGAPDPDWASYTLGVFICLSCSGIHRNIPQVSKVKSV
RLDAWEEAQVEFMASHGNDAARARFESKVPSFYYRPTPSDCQLLREQWIRAKYERQEFIY
PEKQEPYSAGYREGFLWKRGRDNGQFLSRKFVLTEREGALKYFNRNDAKEPKAVMKIEHL
NATFQPAKIGHPHGLQVTYLKDNSTRNIFIYHEDGKEIVDWFNALRAARFHYLQVAFPGA
SDADLVPKLSRNYLKEGYMEKTGPKQTEGFRKRWFTMDDRRLMYFKDPLDAFARGEVFIG
SKESGYTVLHGFPPSTQGHHWPHGITIVTPDRKFLFACETESDQREWVAAFQKAVDRPML
PQEYAVEAHFKHKP
```

Predicted Location(s): Cell membrane; Cytoplasm; Nucleus

Figure 5. A semi-screenshot to show the input in the FASTA format for **(a)** the query protein 1 taken from the Example window, and **(b)** the output predicted by **Hum-mPLoc 2.0** for the query protein sequence in panel **(a)**.

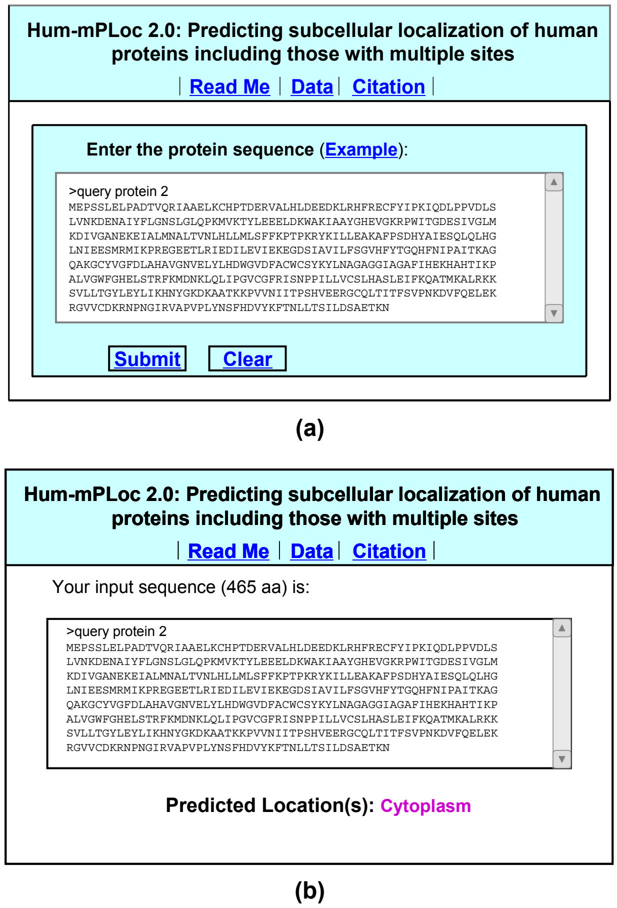
your submission for prediction, consider the following points for troubleshooting.

- Check the format of your input data to make sure it complies with the FASTA format as elaborated in Step 4 of the PROCEDURE.
- Check the length of your input sequence to make sure it is at least 50 amino acids long; otherwise, it might not be a real protein but its fragment.
- Check the amino acid codes of your input sequence to make sure it does not contain any invalid characters.

You might also get meaningless result if the query protein is not among the subcellular location sites covered by the web-server predictor.

5. ANTICIPATED RESULTS

In statistical prediction of subcellular localization of proteins or their any other attributes, it would be meaningless to simply say the success rate of a predictor without specifying what method and benchmark dataset were used to test its accuracy.



Hum-mPLoc 2.0: Predicting subcellular localization of human proteins including those with multiple sites

Enter the protein sequence (Example):

```
>query protein 2
MEPSSLELPADTVQRIAAELKCHPTDERVALHLDEEDKLRHFRECFYIPKIQDLPPVDLS
LVNKDENAIYFLGNSLGLQPKMVKTYLEEELDKWAKIAAYGHEVGKRPWITGDESIVGLM
KDIVGANEKEIALMNALTVNLHLLMLSFFKPTPKRYKILLEAKAFPSDHYAIESQLQLHG
LNIEESMRMIKPREGEETLRIEDILEVIEKEGDSIAVILFSGVHFYTGQHFNIPAITKAG
QAKGCYVGFDLAHAVGNVELYLHDWGVDFACWCSYKYLNAGAGGIAGAFIHEKHAHTIKP
ALVGWFGHELSTRFKMDNKLQLIPGVCGFRISNPPILLVCSLHASLEIFKQATMKALRKK
SVLLTGYLEYLIKHNYGKDKAATKKPVVNIITPSHVEERGCQLTITFSVPNKDVFQELEK
RGVVCDKRNPNGIRVAPVPLYNSFHDVYKFTNLLTSILDSAETKN
```

Hum-mPLoc 2.0: Predicting subcellular localization of human proteins including those with multiple sites

Your input sequence (465 aa) is:

```
>query protein 2
MEPSSLELPADTVQRIAAELKCHPTDERVALHLDEEDKLRHFRECFYIPKIQDLPPVDLS
LVNKDENAIYFLGNSLGLQPKMVKTYLEEELDKWAKIAAYGHEVGKRPWITGDESIVGLM
KDIVGANEKEIALMNALTVNLHLLMLSFFKPTPKRYKILLEAKAFPSDHYAIESQLQLHG
LNIEESMRMIKPREGEETLRIEDILEVIEKEGDSIAVILFSGVHFYTGQHFNIPAITKAG
QAKGCYVGFDLAHAVGNVELYLHDWGVDFACWCSYKYLNAGAGGIAGAFIHEKHAHTIKP
ALVGWFGHELSTRFKMDNKLQLIPGVCGFRISNPPILLVCSLHASLEIFKQATMKALRKK
SVLLTGYLEYLIKHNYGKDKAATKKPVVNIITPSHVEERGCQLTITFSVPNKDVFQELEK
RGVVCDKRNPNGIRVAPVPLYNSFHDVYKFTNLLTSILDSAETKN
```

Predicted Location(s): Cytoplasm

Figure 6. A semi-screenshot to show the input in the FASTA format for **(a)** the query protein 2 taken from the Example window, and **(b)** the output predicted by **Hum-mPLoc 2.0** for the query protein sequence in panel **(a)**.

The following three cross-validation methods are generally used for examining the effectiveness of a statistical prediction method: **(1)** the independent dataset test, **(2)** the sub-sampling (K-fold cross-validation) test, and **(3)** the jackknife test [76].

For the independent dataset test, although all the proteins to be tested are outside the training dataset used to train the predictor and hence can avoid the "memory" effect or bias, the way of how to select the independent proteins for testing could be quite arbitrary unless the number of independent proteins is sufficiently large. This kind of arbitrariness might lead to completely different conclusions. For instance, a predictor achieving a higher success rate than the other predictor for a given independent testing dataset might fail to keep so when tested by another independent testing dataset [76].

For the subsampling test, the concrete procedure usually used in literatures is the 5-fold, 7-fold or 10-fold cross-validation. The problem with the K-fold cross-validation test as such is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset. For example, let us consider a highly simplified dataset that consists of 300 proteins classified into five subsets, in which 60 proteins belong to subcellular location #1, 55 to location #2, 70 to location #3, 65 to location #4, and 50 to location #5. For such a simple dataset, the number of possible combinations of taking one-fifth proteins from each of the five subsets will be

$$\begin{aligned}\Omega &= \Omega_1 \cdot \Omega_2 \cdot \Omega_3 \cdot \Omega_4 \cdot \Omega_5 \\ &= \frac{60!}{(60-12)!12!} \cdot \frac{55!}{(55-11)!11!} \cdot \frac{70!}{(70-14)!14!} \\ &\quad \cdot \frac{65!}{(65-13)!13!} \cdot \frac{50!}{(50-10)!10!} > 5.45 \times 10^{60}\end{aligned} \tag{1}$$

where Ω_1 is the number of possible different ways of taking $60/5 = 12$ proteins from subset #1, Ω_2 that of taking $55/5 = 11$ proteins from subset #2, Ω_3 that of taking $70/5 = 14$ proteins from subset #3, Ω_4 that of taking $65/5 = 13$ proteins from subset #4, and Ω_5 that of taking $50/5 = 10$ proteins from site-site-5. As we can see from **Eq.1**, even for such a simple and small dataset the number of possible ways in selecting the testing dataset for the 5-fold cross-validation would be greater than 5.45×10^{60}. It can be easily conceived that for a benchmark dataset containing over a thousand proteins that are classified into more than five subcellular location sites, the number of the possible selections for subsampling test will be even much greater. Accordingly, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark dataset and a same predictor, the subsampling test (such as 5-fold cross-validation) cannot avoid the arbitrariness either. A test method unable to yield a unique outcome cannot be deemed as an ideal one.

In the jackknife test, all the proteins in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining protein samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each protein sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the arbitrariness problem as mentioned above for the independent dataset test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. As for the possible overestimation in success rate by jackknife test because of only one sample being singled out at a time for testing, the answer is that as long as the jackknife test is performed on a stringent benchmark dataset in which none of proteins has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location such as those benchmark datasets specially constructed for the six predictors in **Cell-PLoc 2.0**, it is highly unlikely to yield an overestimated rate compared with the actual success rate in practical applications, as demonstrated in [72,74] and will be further discussed later. Besides, when the jackknife test was used to compare two predictors, even if there was some overestimate due to using a less stringent benchmark dataset for one predictor, the same overestimate would exist for the other as long as they were both tested by a same dataset.

Accordingly, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors (see, e.g., [47,51,55,58,59,77- 107]).

However, even if using the jackknife approach for cross-validation, a same predictor may still generate obviously different success rates when tested by different benchmark datasets. This is because the more stringent of a benchmark dataset in excluding homologous and high similarity sequences, or the more number of subcellular location sites it covers, the more difficult for a predictor to achieve a high overall success rate, as will be shown later.

The predictors in the old **Cell-PLoc** package [61] were established by hybridizing the "higher-level" GO approach with the "*ab initio*" sequence-correlated PseAAC [16] approach. Accordingly, their overall success prediction rates are generally higher than those by the best of the existing "*ab initio*" sequence-based approaches without combining with any higher level approach, as elucidated in [61] and demonstrated in a series of previous publications [62-67,108,109], and hence there is no need to repeat here.

Now, in the new version of **Cell-PLoc 2.0**, the same high success rates will still be achieved by the "higher-level" GO prediction engine but no requirement for the accession number is needed for the input. And for those proteins without useful GO numbers, the corresponding success prediction rates will be further enhanced due to fusing the functional domain information and sequential evolution information into the "*ab initio*" prediction engine in the **Cell-PLoc 2.0** package as illustrated in **Figure 1**. Accordingly, the overall success rates by the predictors in **Cell-PLoc 2.0** are not only higher than those by the other predictors but also those by the predictors in the old **Cell-PLoc** package [61], as can be seen from the following comparisons.

Table 1. Comparison between each of the six predictors in Cell-PLoc [61] and that in Cell-PLoc 2.0 by jackknife test.

Organism	Number of subcellular locations covered	Cell-PLoc		Cell-PLoc 2.0	
		Predictor	Overall success rate[g]	Predictor	Overall success rate
Eukaryotic	22[a]	Euk-mPLoc	39.3%	Euk-mPLoc 2.0	64.2%
Human	14[b]	Hum-mPLoc	38.1%	Hum-mPLoc 2.0	62.7%
Plant	12[c]	Plant-PLoc	38.0%	Plant-mPLoc	63.7%
Gram-positive	4[d]	Gpos-PLoc	72.5%	Gpos-mPLoc	82.2%
Gram-negative	8[e]	Gneg-PLoc	71.5%	Gneg-mPLoc	85.7%
Virus	6[f]	Virus-PLoc	43.7%	Virus-mPLoc	60.3%

[a]The corresponding benchmark dataset was taken from the Supporting Information S1 of [70], in which none of protein included has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; [b]The corresponding benchmark dataset was taken from the Online Supporting Information A of [71], in which none of protein included has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; [c]The corresponding benchmark dataset was taken from Table S1 of [72], in which none of protein included has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; [d]The corresponding benchmark dataset was taken from the Online Supporting Information A of [73], in which none of protein included has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; [e]The corresponding benchmark dataset was taken from the Online Supporting Information A of [74], in which none of protein included has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; [f]The corresponding benchmark dataset was taken from the Online Supporting Information A of [75], in which none of protein included has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; [g]Note that in order to make the comparison under exactly the same condition, only the sequences of proteins but not their accession numbers were used as inputs during the prediction.

1) Comparison with the six predictors in **Cell-PLoc** [61]. Listed in **Table 1** are the overall success rates by **Cell-PLoc** [61] and **Cell-PLoc 2.0** using jackknife tests on six stringent benchmark datasets for eukaryotic, human, plant, Gram-positive bacterial, Gram-negative bacterial, and virus proteins, respectively. For the case of eukaryotic proteins, the comparison was made between the predictor **Euk-mPLoc** of **Cell-PLoc** [61] and the predictor **Euk-mPLoc 2.0** of **Cell-PLoc 2.0** using the benchmark dataset classified into 22 subcellular locations as given in the Supporting Information S1 of [70]. For human proteins, the comparison was made between the predictor **Hum-mPLoc** of **Cell-PLoc** [61] and the predictor **Hum-mPLoc 2.0** of **Cell-PLoc 2.0** using the benchmark dataset classified into 14 subcellular locations as given in the Online Supporting Information A of [71]. And so forth. To avoid homology bias and redundancy, none of the proteins included in the six datasets has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location. Also, to make the comparison between the two counterparts under exactly the same condition, only the sequences of proteins but not their accession numbers were used as inputs during the prediction. Meanwhile, the false positives (over-predictions) and false negatives (under-predictions) were also taken into account to reduce the scores for calculating the overall success rate. It is instructive to point out that it is much more complicated to count the over-predictions and under-predictions for a system containing both single-location and multiple-location proteins. For the detailed calculation formulation, see Eqs.43-48 as well as **Figure 4** in a comprehensive review [110]. It can be seen from **Table 1** that the overall success rates obtained by the predictors in **Cell-PLoc 2.0** are about 10-25% higher than those by their counterparts in **Cell-PLoc** [61].

2) Comparison with **PSORTb v.2.0** [33]. The predictor is widely used by biologists for predicting the subcellular locations of Gram-negative bacterial proteins. It is with a built-in training dataset covering the following five subcellular location sites: (1) cytoplasm, (2) extracellular, (3) inner membrane, (4) outer membrane, and (5) periplasm. The corresponding predictor in **Cell-PLoc 2.0** is **Gneg-mPLoc** that can cover eight subcellular locations of Gram-negative proteins; i.e., in addition to the above five locations, it also covers “fimbrium”, “flagellum”, and “nucleoid”. In order to make the two predictors with different coverage scopes comparable, a degenerate testing dataset was generated by randomly picking testing proteins according to the following criteria: (1) the testing samples must be Gram-negative bacterial proteins; (2) to avoid the unfair “memory” effect, the testing samples must be not in the training dataset of **PSORTb v.2.0**, nor in the training dataset of **Gneg-mPLoc**; (3) the experimentally observed subcellular locations of the testing proteins are known as clearly annotated in Swiss-Prot database; (4) their location sites must be within the scope covered by **PSORTb v.2.0** for properly using it (for the proteins with multiple location sites, at least one of them should be within the scope covered by **PSORTb**

v.2.0). For the detailed information about the testing dataset thus generated, see the Online Supporting Information B of [74] that contains 759 Gram-negative proteins, of which 116 are of cytoplasm, 62 of extracellular, 397 of inner membrane, 89 of outer membrane, and 95 of periplasm. As shown in **Table 2**, the overall success rates by **Gneg-mPLoc** and **PSORTb v.2.0** [33] in identifying the subcellular locations of proteins in such a testing dataset were 98.0% and 79.3%, respectively, indicating the success rate by **Gneg-mPLoc** of **Cell-PLOc 2.0** was 19% higher than that by **PSORTb v.2.0** [33]. Furthermore, some examples are given in **Table 3** to show how the results mispredicted by **PSORTb v.2.0** were successfully corrected by **Greg-mPLoc**. It is interesting to see from the table that the first protein with accession number P62532 was predicted by **Gneg-mPLoc** belonging to two subcellular location sites, "extracellular" and "fimbrium", fully consistent with experimental observation as annotated in Swiss-Prot database (version 55.3 released on 29- April-2008).

3) Comparison with **TargetP** [15]. The predictor is widely used by biologists for predicting the subcellular locations of plant proteins. It has a web-server at http://www.cbs.dtu.dk/services/TargetP/, with a built-in training dataset covering the following four items: "mitochondria", "chloroplast", "secretory pathway", and "other". Since the "secretory pathway" is not a final destination of subcellular location as annotated in Swiss-Prot databank, and should be removed from the comparison. Also, the location of "other" is not a clear site for comparison, and should be removed too. The corresponding predictor in **Cell-PLoc 2.0** is **Plant-mPLoc** that can cover 12 subcellular locations of plant proteins; i.e., in addition to "mitochondria" and "chloroplast", it also covers "cell membrane", "cell wall", "cytoplasm", "endoplasmic reticulum", "extracellular", "Golgi apparatus", "nucleus", "peroxisome", "plastid", and "vacuole". Thus, to make the two predictors with different coverage scopes comparable, a degenerate testing dataset was generated according to the similar procedures as described in section **5.2**. For the detailed information about the testing dataset thus generated, see Table S2 of [72] that contains 1,775 plant proteins of which 1,500 are of chloroplast and 275 of mitochondrion. As reported in [72], the overall success rates by **Plant-mPLoc** on such a testing dataset was 86%, which is more than 40% higher than that by **TargetP** [15] on the same testing dataset.

4) Comparison with Predotar [111]. This is another popular predictor used by biologists for predicting the subcellular locations of plant proteins. Its web-server is at http://urgi.versailles.inra.fr/predotar/predotar.html, with a built-in training dataset covering the following four items: "endoplasmic reticulum", "mitochondrion", "plastid", and "other". Since the term "other" is not a clear description for subcellular location, and was removed from comparison. The corresponding predictor in **Cell-PLoc 2.0** is **Plant-mPLoc** that can cover 12 subcellular locations of plant proteins; *i.e.*, in addition to "endoplasmic reticulum", "mitochondria" and "plastid", it also covers "cell membrane", "cell wall", "chloroplast", "cytoplasm", "extracellular", "Golgi apparatus", "nucleus", "peroxisome", and "vacuole". Again, to make the two predictors with different coverage scopes comparable, a degenerate testing dataset was generated by following the similar procedures as described in section **5.2**. For the detailed information about the testing dataset thus generated, see Table S4 of [72], where it was also reported that the overall success rates by **Plant-mPLoc** on such a testing dataset was 70%, which is more than 30% higher than that by **Predotar** [111] on the same testing dataset.

Moreover, it was also shown in [72,74] that some proteins coexisting in two or more subcellular location sites were successfully identified by **Gneg-mPLoc** [74] and **Plant-mPLoc** [72]; cases like that are beyond the reach of **PSORTb v.2.0** [33], **TargetP** [15], or **Predotar** [111].

Table 2. A comparison of the predicted results by **Gneg-mPLoc** and **PSORTb v.2.0** [33] on the testing dataset of Online Supporting Information B of [74].

Subcellular location	Success rate	
	PSORTb v.2.0	Gneg-mPLoc
Cytoplasm	99/116=85.3%	115/116=99.1%
Extracellular	20/62=32.3%	52/62=83.9%
Inner membrane	329/397=82.9%	397/397=100%
Outer membrane	75/89=84.3%	87/89=97.8%
Periplasm	79/95=83.2%	93/95=97.9%
Total	602/759=79.3%	744/759=98.0%

Table 3. Some examples to show how the subcellular location sites mispredicted by **PSORTb v.2.0** were corrected by **Gneg-mPLoc.**

Protein accession number[a]	Experimental result annotated in Swiss-Prot database	Predicted result by PSORTb v.2.0	Predicted result by Gneg-mPLoc
P62532	Extracellular; Fimbrium	Unknown	Extracellular; Fimbrium
Q8X9H8	Cytoplasm	Unknown	Cytoplasm
P00962	Cytoplasm	Unknown	Cytoplasm
Q83LY4	Cytoplasm	Unknown	Cytoplasm
Q8DFR1	Cytoplasm	Unknown	Cytoplasm
Q84H44	Cytoplasm	Unknown	Cytoplasm
P27475	Extracellular	Unknown	Extracellular
O50319	Extracellular	Unknown	Extracellular
P31518	Extracellular	Unknown	Extracellular
Q89AD4	Cytoplasm	Unknown	Cytoplasm
Q56027	Extracellular	Unknown	Extracellular
O52623	Extracellular	Unknown	Extracellular
P26219	Cell inner membrane	Unknown	Cell inner membrane
P77293	Cell inner membrane	Unknown	Cell inner membrane
P95655	Cell inner membrane	Unknown	Cell inner membrane.
P04123	Cell inner membrane	Periplasm	Cell inner membrane
Q47879	Cell outer membrane	Unknown	Cell outer membrane
P0A935	Cell outer membrane	Unknown	Cell outer membrane
P00211	Periplasm	Cytoplasm	Periplasm
P0A182	Periplasm	Unknown	Periplasm
Q9Z4N3	Periplasm	Unknown	Periplasm
P31330	Periplasm	Cytoplasm	Periplasm

[a] Only the sequences but not the accession numbers were used as inputs during the prediction by Gneg-mPLoc. The accession numbers here are just for the usage of identification.

From the above four comparisons, we can now make the following points very clear.

- The more stringent a benchmark dataset is in excluding homologous and high similarity sequences, or the more subcellular location sites it covers, the more difficult for a predictor to achieve a high overall success rate. The impact of the coverage scope on the success rate can be easily understood by just considering the following cases. For a benchmark dataset only covering four subcellular locations each containing same number of proteins, the overall success rate by random assignments would generally be $1/4=25\%$; while for a benchmark dataset covering 22 subcellular locations, the overall success rate by random assignments would be only $1/25\approx4.5\%$. This means that the former is more than five times the latter.
- Also, a predictor examined by jackknife test is very difficult to yield a high success rate when performed on a stringent benchmark dataset in which none of proteins included has $\geq 25\%$ pairwise sequence identity to any other in a same subset (subcellular location). That is why the overall success rate achieved by **Gneg-mPLoc** was 85.7% when examined by the jackknife test on the benchmark dataset of the Online Supporting Information A of [74] but was 98.0% when examined by the independent dataset test for the proteins in the Online Supporting Information B of [74]. That is also why the overall success rate achieved by **Plant-mPLoc** was only 63.7% when examined by the jackknife test on the benchmark dataset of Table S1 of [72] but was over 86% and 70% when tested by the independent proteins of Table S2 and

Table S4 of [72], respectively. However, regardless of using what test methods or test datasets, one thing is crystal clear, i.e., the overall success rates achieved by the six predictors in **Cell-PLoc 2.0** are significantly higher than those by its counterparts.

- Meanwhile, it has also become understandable why the success rates as originally reported by **PSORTb v.2.0** [33], **TargetP** [15] and **Predotar** [111] were over-estimated. This is because none of the success rates reported for these predictors was derived by the jackknife test. Also, the benchmark datasets used to test these predictors covered much less subcellular location sites than those used in their counterparts in **Cell-PLoc 2.0**. Particularly, the benchmark datasets used by **PSORTb v.2.0** , **TargetP** and **Predotar** to estimate their success rates contained many homologous sequences. For instance, the cutoff threshold to reduce the homology bias for the benchmark dataset used in **Predotar** [111] was set at 80%, meaning that only those sequences which have ≥ 80% pairwise sequence identity to any other in a same subset were excluded [111]; while for the benchmark dataset used in **TargetP** [15] and **PSORTb v.2.0** [33], even no cutoff threshold was indicated to remove homologous sequences. Compared with the benchmark datasets used in [70-75] where none of proteins included has ≥ 25% pairwise sequence identity to any other in a same subset, the benchmark datasets adopted by **PSORTb v.2.0**, **TargetP**, and **Predotar** are much less stringent and hence cannot avoid homology bias and overestimation.

6. CONCLUDING REMARKS

Evolved from the old **Cell-PLoc** package [61], **Cell-PLoc 2.0** is much more flexible and powerful than the former. In addition to yielding higher success rates than the existing prediction method, all the predictors in **Cell-PLoc 2.0** have the capacity to deal with proteins with two or more subcellular location sites. Besides, the predictors in **Cell-PLoc 2.0** cover much wider scopes than most of the existing predictors in this area. For instance, **Hum-mPLoc 2.0** and **Euk-mPLoc 2.0** can cove up to 14 sites of human proteins and 22 sites of eukaryotic, respectively, which are about two to five times the number of subcellular location sites covered by most of the existing predictors.

However, **Cell-PLoc 2.0** also has the following limitations and further improvements will be needed with more experimental data available in future. **(1)** Although **Euk-mPLoc 2.0** in the **Cell-PLoc 2.0** package can cover 22 sites of eukaryotic proteins, if a query protein is outside of the 22 location sites, it would still generate meaningless result. Therefore, we shall continuously extend the coverage scope for each of the predictors in the **Cell-PLoc** series in a timely manner once more statistically significant experimental data will be available in future. **(2)** For some subcellular locations with very small numbers of proteins, the prediction success rates are still quite low. This is because there are not sufficient location-known proteins in these sites to effectively train the prediction engine. It is anticipated that with more experimental data available for these sites in the future, this kind of situation will be improved. **(3)** Since the power of **Cell-PLoc 2.0** is closely associated with the GO database [68,112,113] and functional domain database [114], with the continuous development of the GO database and functional domain database, more useful GO numbers and functional domain information will be incorporated into the prediction engine, further strengthening its prediction power.

Once further improvements are implemented, the future version of **Cell-PLoc** series will be announced via a publication or a webpage.

REFERENCES

[1] Ehrlich, J.S., Hansen, M.D., Nelson, W.J. (2002) Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell adhesion. *Dev Cell*, **3**, 259-270.

[2] Glory, E., Murphy, R.F. (2007) Automated subcellular location determination and high-throughput microscopy. *Dev Cell*, **12**, 7-16.

[3] Smith, C. (2008) Subcellular targeting of proteins and drugs. http://www.biocompare.com/Articles/TechnologySpotlight/976/Subcellular-Targeting-Of-Proteins-And-Drugs.html

[4] Nakai, K., Kanehisa, M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*: *Structure, Function and Genetics*, **11**, 95-110.

[5] Nakashima, H., Nishikawa, K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of Molecular Biology*, **238**, 54-61.

[6] Cedano, J., Aloy, P., P'erez-Pons, J.A., Querol, E. (1997) Relation between amino acid composition and cellular location of proteins. *Journal of Molecular Biology*, **266**, 594-600.

[7] Nakai, K., Horton, P. (1999) PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Science*, **24**, 34-36.

[8] Chou, K.C., Elrod, D.W. (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochemical and Biophysical Research Communications*, **252**, 63-68.

[9] Reinhardt, A., Hubbard, T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nu-*

cleic Acids Research, **26**, 2230-2236.

[10] Chou, K.C., Elrod, D.W. (1999) Protein subcellular location prediction. *Protein Engineering*, **12**, 107-118.

[11] Yuan, Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Letters*, **451**, 23-26.

[12] Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry*, **54**, 277-344.

[13] Murphy, R.F., Boland, M.V., Velliste, M. (2000) 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.

[14] Chou, K.C. (2000) Review: Prediction of protein structural classes and subcellular locations. *Current Protein and Peptide Science*, **1**, 171-208.

[15] Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300**, 1005-1016.

[16] Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics* (*Erratum*: *ibid*., 2001, Vol.44, 60), **43**, 246-255.

[17] Feng, Z.P. (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, **58**, 491-499.

[18] Hua, S., Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721-728.

[19] Feng, Z.P., Zhang, C.T. (2001) Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *Int. J. Biol. Macromol.*, **28**, 255-261.

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

[21] Chou, K.C., Cai, Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry*, **277**, 45765-45769.

[22] Zhou, G.P., Doctor, K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins*: *Structure, Function, and Genetics*, **50**, 44-48.

[23] Pan, Y.X., Zhang, Z.Z., Guo, Z.M., Feng, G.Y., Huang, Z.D., He, L. (2003) Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach. *Journal of Protein Chemistry*, **22**, 395-402.

[24] Park, K.J., Kanehisa, M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. *Bioinformatics*, **19**, 1656-1663.

[25] Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., Brinkman, F.S. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research*, **31**, 3613-3617.

[26] Huang, Y., Li, Y. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, **20**, 21-28.

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

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

[29] Lei, Z., Dai, Y. (2005) An SVM-based system for predicting protein subnuclear localizations BMC. *Bioinformatics*, **6**, 291.

[30] Shen, H.B., Chou, K.C. (2005) Predicting protein subnuclear location with optimized evidence-theoretic K-earest classifier and pseudo amino acid composition. *Biochem. Biophys. Res. Comm.*, **337**, 752-756.

[31] Garg, A., Bhasin, M., Raghava, G.P. (2005) Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *Journal of Biological Chemistry*, **280**, 14427-14432.

[32] Matsuda, S., Vert, J.P., Saigo, H., Ueda, N., Toh, H., Akutsu, T. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci.*, **14**, 2804-2813.

[33] Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., Brinkman, F.S. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617-623.

[34] Gao, Q.B., Wang, Z.Z., Yan, C., Du, Y.H. (2005) Prediction of protein subcellular location using a combined feature of sequence. *FEBS Letters*, **579**, 3444-3448.

[35] Chou, K.C., Shen, H.B. (2006) Predicting protein subcellular location by fusing multiple classifiers. *Journal of Cellular Biochemistry*, **99**, 517-527.

[36] Guo, J., Lin, Y., Liu, X. (2006) GNBSL: A new integrative system to predict the subcellular location for Gramegative bacteria proteins. *Proteomics*, **6**, 5099-5105.

[37] Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C. (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids*, **30**, 49-54.

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

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

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

[41] Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.-M., Xie, J. (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo mino acid composition. *Amino Acids*, **33**, 69-74.

[42] Chen, Y.L., Li, Q.Z. (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *Journal of Theo-*

retical Biology, **248**, 377–381.

[43] Chen, Y.L., Li, Q.Z. (2007) Prediction of the subcellular location of apoptosis proteins. *Journal of Theoretical Biology*, **245**, 775-783.

[44] Mundra, P., Kumar, M., Kumar, K.K., Jayaraman, V.K., Kulkarni, B.D. (2007) Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recognition Letters*, **28**, 1610-1615.

[45] Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols*, **2**, 953-971.

[46] Lin, H., Ding, H., Feng-Biao Guo, F.B., Zhang, A.Y., Huang, J. (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein & Peptide Letters*, **15**, 739-744.

[47] Shi, J.Y., Zhang, S.W., Pan, Q., Zhou, G.P. (2008) Using Pseudo Amino Acid Composition to Predict Protein Subcellular Location: Approached with Amino Acid Composition Distribution. *Amino Acids*, **35**, 321-327.

[48] Li, F.M., Li, Q.Z. (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein & Peptide Letters*, **15**, 612-616.

[49] Tantoso, E., Li, X.B. (2008) AAIndexLoc: Predicting Subcellular Localization of Proteins Based on a New Representation of Sequences Using Amino Acid Indices. *Amino Acids*, **35**, 345-353.

[50] Jiang, X., Wei, R., Zhang, T.L., Gu, Q. (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein & Peptide Letters*, **15**, 392-396.

[51] Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y. (2008) Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. *Amino Acids*, **35**, 383-388.

[52] Ding, Y.S., Zhang, T.L. (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognition Letters*, **29**, 1887-1892.

[53] Zhang, S.W., Zhang, Y.L., Yang, H.F., Zhao, C.H., Pan, Q. (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids*, **34**, 565-572.

[54] Jin, Y., Niu, B., Feng, K.Y., Lu, W.C., Cai, Y.D., Li, G.Z. (2008) Predicting subcellular localization with AdaBoost learner. *Protein & Peptide Letters*, **15**, 286-289.

[55] Lin, H., Wang, H., Ding, H., Chen, Y.L., Li, Q.Z. (2009) Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. *Acta Biotheoretica*, **57**, 321-330.

[56] Zhang, L., Liao, B., Li, D., Zhu, W. (2009) A novel representation for apoptosis protein subcellular localization prediction using support vector machine. *Journal of Theoretical Biology*, **259**, 361-365.

[57] Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., Li, M.L. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *Journal of Theoretical Biology*, **259**, 366-72.

[58] Du, P., Cao, S., Li, Y. (2009) SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. *Journal of Theoretical Biology*, **261**, 330-335.

[59] Cai, Y.D., He, J., Li, X., Feng, K., Lu, L., Kong, X., Lu, W. (2010) Predicting protein subcellular locations with feature selection and analysis. *Protein Pept. Lett.*, **17**, 464-472.

[60] Millar, A.H., Carrie, C., Pogson, B., Whelan, J. (2009) Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. *Plant Cell*, **21**, 1625-1631.

[61] Chou, K.C., Shen, H.B. (2008) Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols*, **3**, 153-162.

[62] Chou, K.C., Shen, H.B. (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *Journal of Proteome Research*, **6**, 1728-1734.

[63] Shen, H.B., Chou, K.C. (2007) Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochemical and Biophysical Research Communications*, **355**, 1006-1011.

[64] Chou, K.C., Shen, H.B. (2007) Large-scale plant protein subcellular location prediction. *Journal of Cellular Biochemistry*, **100**, 665-678.

[65] Shen, H.B., Chou, K.C. (2007) Gpos-PLoc: An ensemble classifier for predicting subcellular localization of Gram-ositive bacterial proteins. *Protein Engineering, Design, and Selection*, **20**, 39-46.

[66] Chou, K.C., Shen, H.B. (2006) Large-scale predictions of Gram-negative bacterial protein subcellular locations. *Journal of Proteome Research*, **5**, 3420-3428.

[67] Shen, H.B., Chou, K.C. (2007) Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers*, **85**, 233-240.

[68] 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-arver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000) Gene ontology: Tool for the unification of biology. *Nature Genetics*, **25**, 25-29.

[69] Chou, K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10-19.

[70] Chou, K.C., Shen, H.B. (2010) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE*, **5**, e9931.

[71] Shen, H.B., Chou, K.C. (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Analytical Biochemistry*, **394**, 269-74.

[72] Chou, K.C., Shen, H.B. (2010) Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS ONE*, **5**, e11335.

[73] Shen, H.B., Chou, K.C. (2009) Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. *Protein & Peptide Letters*, **16**, 1478-1484.

[74] Shen, H.B., Chou, K.C. (2010) Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *Journal of Theoretical Biology*, **264**, 326-333.

[75] Shen, H.B., Chou, K.C. (2010) Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *Journal of Biomolecular Structure & Dynamics*, **28**, 175-186.

[76] Chou, K.C., Zhang, C.T. (1995) Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, **30**, 275-349.

[77] Fang, Y., Guo, Y., Feng, Y., Li, M. (2008) Predicting DNA-binding proteins: Approached from Chou's pseudo amino acid composition and other specific sequence features. *Amino Acids*, **34**, 103-109.

[78] Feng, Y.E., Luo, L.F. (2008) Use of tetrapeptide signals for protein secondary-structure prediction. *Amino Acids*, **35**, 607-614.

[79] Li, Z.C., Zhou, X.B., Dai, Z., Zou, X.Y. (2009) Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. *Amino Acids*, **37**, 415-425.

[80] Nanni, L., Lumini, A. (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **34**, 653-660.

[81] Wang, Y., Xue, Z., Shen, G., Xu, J. (2008) PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*, **35**, 295-302.

[82] Zhao, X.M., Chen, L., Aihara, K. (2008) Protein function prediction with high-throughput data. *Amino Acids*, **35**, 517-530.

[83] Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Asadabadi, E.B. (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. *Biophys. Chem.*, 128, 87-93.

[84] Chen, K., Kurgan, L.A., Ruan, J. (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *Journal of Computational Chemistry*, **29**, 1596-1604.

[85] Jahandideh, S., Sarvestani, A.S., Abdolmaleki, P., Jahandideh, M., Barfeie, M. (2007) Gamma-Turn types prediction in proteins using the support vector machines. *Journal of Theoretical Biology*, **249**, 785-790.

[86] Shao, X., Tian, Y., Wu, L., Wang, Y., L., J., Deng, N. (2009) Predicting DNA- and RNA-binding proteins from sequences with kernel methods. *Journal of Theoretical Biology*, **258**, 289-293.

[87] Yang, J.Y., Peng, Z.L., Yu, Z.G., Zhang, R.J., Anh, V., Wang, D. (2009) Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *Journal of Theoretical Biology*, **257**, 618-626.

[88] Anand, A., Suganthan, P.N. (2009) Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates. *Journal of Theoretical Biology*, **259**, 533-540.

[89] Chen, C., Chen, L.X., Zou, X.Y., Cai, P.X. (2008) Predicting protein structural class based on multi-features fusion. *Journal of Theoretical Biology*, **253**, 388-392.

[90] Du, P., Li, Y. (2008) Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. *Journal of Theoretical Biology*, **253**, 579-589.

[91] Jahandideh, S., Hoseini, S., Jahandideh, M., Hoseini, A., Disfani, F.M. (2009) Gamma-turn types prediction in proteins using the two-stage hybrid neural discriminant model. *Journal of Theoretical Biology*, **259**, 517-522.

[92] Lin, H. (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *Journal of Theoretical Biology*, **252**, 350-356.

[93] Munteanu, C.B., Gonzalez-Diaz, H., Magalhaes, A.L. (2008) Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. *Journal of Theoretical Biology*, **254**, 476-482.

[94] Rezaei, M.A., Abdolmaleki, P., Karami, Z., Asadabadi, E.B., Sherafat, M.A., Abrishami-Moghaddam, H., Fadaie, M., Forouzanfar, M. (2008) Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks. *Journal of Theoretical Biology*, **254**, 817-820.

[95] Vilar, S., Gonzalez-Diaz, H., Santana, L., Uriarte, E. (2009) A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *Journal of Theoretical Biology*, **261**, 449-458.

[96] Wang, T., Xia, T., Hu, X.M. (2010) Geometry preserving projections algorithm for predicting membrane protein types. *Journal of Theoretical Biology*, **262**, 208-213.

[97] Chen, Y., Han, K. (2009) BSFINDER: Finding Binding Sites of HCV Proteins Using a Support Vector Machine. *Protein & Peptide Letters*, **16**, 373-382.

[98] Kannan, S., Hauth, A.M., Burger, G. (2008) Function prediction of hypothetical proteins without sequence similarity to proteins of known function. *Protein & Peptide Letters*, **15**, 1107-1116.

[99] Nanni, L., Lumini, A. (2009) A Further Step Toward an Optimal Ensemble of Classifiers for Peptide Classification, a Case Study: HIV Protease. *Protein & Peptide Letters*, **16**, 163-167.

[100] Gu, F., Chen, H. (2009) Evaluating Long-term Relationship of Protein Sequence by Use of d-Interval Conditional Probability and its Impact on Protein Structural Class Prediction. *Protein Pept. Lett.*, **16**, 1267-1276.

[101] Ji, G., Wu, X., Shen, Y., Huang, J., Quinn Li, Q. (2010) A classification-based prediction model of messenger RNA polyadenylation sites. *Journal of Theoretical Biology*, **265**, 287-296.

[102] Yang, X.Y., Shi, X.H., Meng, X., Li, X.L., Lin, K., Qian, Z.L., Feng, K.Y., Kong, X.Y., Cai, Y.D. (2010) Classification of transcription factors using protein primary structure. *Protein & Peptide Letters*, **17**, 899-908.

[103] Gu, Q., Ding, Y.S., Zhang, T.L. (2010) Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. *Protein Pept. Lett.*, **17**, 559-567.

[104] Liu, L., He, D., Yang, S., Xu, Y. (2010) Applying chemometrics approaches to model and predict the binding

affinities between the human amphiphysin SH3 domain and its peptide ligands. *Protein Pept. Lett.*, **17**, 246- 253.

[105] Shi, R., Hu, X. (2010) Predicting enzyme subclasses by using support vector machine with composite vectors. *Protein Pept. Lett.*, **17**, 599-604.

[106] Wang, T., Yang, J. (2010) Predicting subcellular localization of gram-negative bacterial proteins by linear dimensionality reduction method. *Protein Pept. Lett.*, **17**, 32-37.

[107] Yang, J., Jiang, X.F. (2010) A novel approach to predict protein-protein interactions related to Alzheimer's disease based on complex network. *Protein Pept. Lett.*, **17**, 356-366.

[108] Chou, K.C., Shen, H.B. (2006) Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochemical and Biophysical Research Communications*, **347**, 150-157.

[109] Chou, K.C., Shen, H.B. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *Journal of Proteome Research*, **5**, 1888-1897.

[110] Chou, K.C., Shen, H.B. (2007) Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry*, **370**, 1-16.

[111] Small, I., Peeters, N., Legeai, F., Lurin, C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581-1590.

[112] Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., Apweiler, R. (2003) The gene ontology annotation (GOA) project: Implementation of GO in SWISS- PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 662-672.

[113] Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., Apweiler, R. (2009) The GOA database in 2009-an integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, **37**, D396-403.

[114] Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Krylov, D., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Thanki, N., Yamashita, R.A., Yin, J.J., Zhang, D., Bryant, S.H. (2007) CDD: A conserved domain database for interactive domain family analysis. *Nucleic Acids Research*, **35**, D237-D240.