Analytical Biochemistry 394 (2009) 269–274

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0

Hong-Bin Shen *, Kuo-Chen Chou *

Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai 200240, China
Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

ARTICLE INFO

Article history:
Received 21 May 2009
Available online 3 August 2009

Keywords:
Multiplex protein
Homology search
Representative proteins
Gene ontology
Functional domain
Sequential evolution
Ensemble classifier
Fusion approach

ABSTRACT

Predicting subcellular localization of human proteins is a challenging problem, particularly when query proteins may have a multiplex character, i.e., simultaneously exist at, or move between, two or more different subcellular location sites. In a previous study, we developed a predictor called "Hum-mPLoc" to deal with the multiplex problem for the human protein system. However, Hum-mPLoc has the following shortcomings. (1) The input of accession number for a query protein is required in order to obtain a higher expected success rate by selecting to use the higher-level prediction pathway; but many proteins, such as synthetic and hypothetical proteins as well as those newly discovered proteins without being deposited into databanks yet, do not have accession numbers. (2) Neither functional domain nor sequential evolution information were taken into account in Hum-mPLoc, and hence its power may be reduced accordingly. In view of this, a top-down strategy to address these shortcomings has been implemented. The new predictor thus obtained is called Hum-mPLoc 2.0, where the accession number for input is no longer needed whatsoever. Moreover, both the functional domain information and the sequential evolution information have been fused into the predictor by an ensemble classifier. As a consequence, the prediction power has been significantly enhanced. The web server of Hum-mPLoc2.0 is freely accessible at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/.

Numerous efforts have been made to develop various methods for predicting protein subcellular localization based on the sequence information (see, e.g., [1–8] and a long list of references cited in two comprehensive review papers [9,10]). However, for practical applications in drug development, it is more important and urgent to timely determine the subcellular locations of human proteins. Unfortunately, relatively much fewer predictors were established that are specialized for predicting the subcellular localization of human proteins.

Although the HSLPred developed by Garg et al. [11] was specifically for human proteins, the predictor can only cover four subcellular location sites: cytoplasm, mitochondria, nucleus, and plasma membrane. If a user used HSLPred [11] to predict a query protein located outside the aforementioned four sites, such as lysosome and centriole, the predictor would fail to work, or the results thus obtained would not make any sense.

To improve the coverage limit, the predictor called Hum-PLoc [12] was developed to extend the coverage scope for human proteins from 4 to 12 location sites, i.e., the aforementioned 4 sites plus the following 8 sites: centriole, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, microsome, and peroxisome. However, Hum-Ploc [12] cannot be used to deal with multiplex proteins, which may simultaneously exist at, or move between, two or more different subcellular locations. Proteins with multiple locations or dynamic feature of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery [13,14]. According to a statistical analysis on the Swiss-Prot database (version 55.3), this kind of multiplex proteins might occupy about 20% of the human proteins.

To make Hum-Ploc be able to predict the multiplex protein locations as well, the predictor called Hum-mPLoc [15] was developed. Meanwhile, the subcellular location scope covered by Hum-mPLoc was further extended to the 14 sites; i.e., the aforementioned 12 location sites plus endosome and synapse. Even though, Hum-mPLoc could still yield about 70% jackknife cross-validation success rate when tested by a very stringent benchmark dataset in which none of the proteins included has $\geqslant$25% pairwise sequence identity to any other protein in the same subcellular location subset. The Hum-mPLoc predictor was established by hybridizing the "higher-level" GO (gene ontology [16]) approach and PseAAC (pseudo amino acid composition [17,18]) approach. Its power mainly came from the GO approach because proteins formulated in the GO database space would be clustered in a way

* Corresponding authors.
E-mail addresses: hbshen@sjtu.edu.cn (H.-B. Shen), kcchou@gordonlifescience.org (K.-C. Chou).

0003-2697/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2009.07.046

much better reflecting their subcellular locations, as elucidated in [19].

However, the existing version of Hum-mPLoc has the following problems. (1) In order to take advantage of the GO approach, the input for a query protein must include its accession number. Many proteins, such as synthetic and hypothetical proteins as well as those newly discovered proteins that have not been deposited into databanks yet, do not have accession numbers, and hence their subcellular locations cannot be predicted via the GO approach. (2) Since the current GO database is far from complete yet, many proteins cannot be meaningfully formulated in a GO space even if their accession numbers are available. (3) Although the PseAAC approach, a complement to the GO approach in Hum-mPLoc, can take into account some partial sequence order effects, the original PseAAC [17,20] missed the functional domain and sequential evolution information.

The present study was initiated in an attempt to develop a new and more powerful predictor, called Hum-mPLoc 2.0, for predicting human protein subcellular localization by addressing the above three problems.

Materials

Protein sequences were collected from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/. The detailed procedures are basically the same as those in [15]. The only difference is that, in order to obtain the updated data, instead of version 50.7 released on 9 September 2006, the version 55.3 released on 29 April 2008 is adopted. After strictly following the procedures as described in [15], we finally obtained a benchmark dataset of 3106 different protein sequences covering 14 subcellular locations (see Table 1), where 2580 proteins belong to one subcellular location, 480 to two locations, 43 to three locations, and 3 to four locations. The corresponding accession numbers and protein sequences are given in Online Supporting Information A. Note that because some proteins may occur in two or more locations, the 3106 different proteins actually correspond to 3681 locative proteins. The concept of "locative proteins" was introduced for studying proteins with multiple subcellular location sites, as illustrated in [10,15].

Table 1
Breakdown of the human protein benchmark dataset derived from Swiss-Prot database (release 55.3) according to the procedures described under Materials (none of proteins included here has ⩾25% pairwise sequence identity to any other in a same subcellular location).

Order	Subcellular location	Number of proteins
1	Centriole	77
2	Cytoplasm	817
3	Cytoskeleton	79
4	Endoplasmic reticulum	229
5	Endosome	24
6	Extracell	385
7	Golgi apparatus	161
8	Lysosome	77
9	Microsome	24
10	Mitochondrion	364
11	Nucleus	1021
12	Peroxisome	47
13	Plasma membrane	354
14	Synapse	22
Total number of locative proteins $\tilde{N}$		3681[a]
Total number of different proteins N		3106[b]

[a] See Eqs. (1)–(4) of [15] for the definition about the number of locative proteins, and its relation with the number of different proteins.
[b] Of the 3106 different proteins, 2580 belong to one subcellular location, 480 to two locations, 43 to three locations, and 3 to four location.

Methods

The key in developing a powerful method for predicting protein subcellular localization is to grasp the core features of proteins that are intrinsically correlated with their localization in a cell. In this regard, the strategy by hybridizing the GO representation and PseAAC representation was quite successful, as demonstrated in [12,15]. Therefore, we shall continue adopting the hybridization strategy in the current study. However, in order to solve the three problems raised in the Introduction, the detailed procedures to realize the hybridization approach will be completely different, as elaborated below.

GO representation

GO is a controlled vocabulary used to describe the biology of a gene product in any organism [21,22]. The GO representation for a protein sample in the original Hum-mPLoc [15] was derived by first searching for its accession number against all the UniProt accession numbers and their corresponding GO numbers in the GO database [21], followed by mapping the GO information thus obtained into the representation for the protein sample. Therefore, in using Hum-mPLoc for prediction, the accession number of a query protein would be indispensable. To avoid such a problem, here let us derive the GO representation according to the following procedures.

Step 1. Use BLAST [23] to search the homologous proteins of the query protein **P** from the Swiss-Prot database (version 55.3), with the BLAST parameter of expect value $E \leqslant 0.001$.

Step 2. Those proteins which have ⩾60% pairwise sequence identity with the query protein **P** are collected as its *representative proteins*; meanwhile, their corresponding accession numbers in the Swiss-Prot database are also obtained accordingly.

Step 3. Search each of these accession numbers collected in Step 2 against the GO database at http://www.ebi.ac.uk/GOA/ to retrieve the GO information [21].

Step 4. The current GO database (version 70.0 released March 10 2008) contains 60,020 GO numbers; thus the query protein **P** can be formulated through its representative proteins by the equation

$$\mathbf{P}_{\mathrm{GO}} = \left[\delta_1^{\mathrm{G}} \quad \delta_2^{\mathrm{G}} \quad \cdots \quad \delta_i^{\mathrm{G}} \quad \cdots \quad \delta_{60020}^{\mathrm{G}}\right]^{\mathrm{T}}, \tag{1}$$

where **T** is the transposing operator, and

$$\delta_i^{\mathrm{G}} = \begin{cases} 1, & \text{if a hit found against the } i\text{-th GO number} \\ & \text{for any of the representative proteins of } \mathbf{P} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Through the above steps, we can study the query protein **P** by means of the GO information derived from its representative proteins. The rationale to do so is based on the fact that homology proteins generally share similar attributes, such as biological functions and structural conformations [24,25]. The reason for using the value of 60% as a threshold here is due to the fact that for most cases proteins with 60% or higher sequence identity can be usually treated as homologous to each other [26]. Actually, our preliminary tests also indicated that such a threshold was a good choice.

Thus, the accession number is no longer required for the input of the query protein even when using the high-level GO approach to predict its subcellular localization as required in Hum-PLoc [12] and Hum-mPLoc [15].

The above homology-based GO extraction method is very useful for studying those proteins which do not have UniProt accession numbers. However, it would still fail to work under any one of the following two situations: (1) the query protein does not have

significant homology to any protein in the Swiss-Prot database; (2) its representative proteins do not contain any useful GO information for statistical prediction based on a given training dataset. Therefore, some complementary modes are needed, as introduced below.

FunD (functional domain) representation

Protein FunD databases, such as SMART [27], Pfam [28], COG [29], KOG [29], and CDD [30], were established according to the principle that proteins often contain several modules or domains, each with a distinct evolutionary origin and function. Of the aforementioned databases, CDD contains the domains imported from SMART, Pfam, and COG databases, and hence is relatively much more complete [30]. The version 2.11 of CDD contains 17,402 characteristic domains. Thus, a given protein sample can be defined as a vector in the 17402-D (dimensional) FunD space according to the following procedures [31]:

Step 1. Use RPS-BLAST (Reverse PSI-BLAST) program [23] to compare the protein sequence with each of the 17,402 domain sequences in the CDD database.

Step 2. If the significance threshold value (expect value) is $\leqslant$0.001 for the i-th profile meaning that a "hit" is found, then the i-th component of the protein in the 17402-D space is assigned 1; otherwise, 0.

Step 3. The protein sample **P** in the FunD space can thus be formulated as

$$\mathbf{P}_{\mathrm{FunD}} = \left[\delta_1^{\mathrm{D}} \quad \delta_2^{\mathrm{D}} \quad \cdots \quad \delta_i^{\mathrm{D}} \quad \cdots \quad \delta_{17402}^{\mathrm{D}}\right]^{\mathrm{T}}, \tag{3}$$

where **T** is the transpose operator, and

$$\delta_i^{\mathrm{D}} = \begin{cases} 1, & \text{when a hit found for } \mathbf{P} \text{ in CDD} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The reason for using the 0.001 as a threshold in searching FunD database (Step 2) was to optimize the prediction result, i.e., the overall success rate by the cross-validation tests. If allowing the threshold value to be lager than 0.001, more incorrect FunD domains or noisy information would be brought in so as to lower down the overall statistical prediction accuracy.

Evolutional expression

To incorporate the evolution information of proteins, the protein sample **P** should be expressed by a matrix, the so-called "Position-Specific Scoring Matrix" or "PSSM" [23]; i.e.,

$$\mathbf{P}_{\mathrm{Evo}} = \begin{bmatrix} \mathrm{V}_{1\to1} & \mathrm{V}_{1\to2} & \cdots & \mathrm{V}_{1\to20} \\ \mathrm{V}_{2\to1} & \mathrm{V}_{2\to2} & \cdots & \mathrm{V}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{V}_{i\to1} & \mathrm{V}_{i\to2} & \cdots & \mathrm{V}_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{V}_{L\to1} & \mathrm{V}_{L\to2} & \cdots & \mathrm{V}_{L\to20} \end{bmatrix} \tag{5}$$

where $\mathrm{V}_{i\to j}$ represents the score of the amino acid residue in the i-th position of the protein sequence being changed to amino acid type j during the evolution process. Here, the numerical codes 1, 2, ..., 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes. The $L \times 20$ scores in Eq. (5) were generated by using PSI-BLAST [23] to search the Swiss-Prot database (version 55.3 released on 29 April 2007) through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the sequence of the protein **P**, followed by a standard conversion given as

$$\mathrm{V}_{i\to j} = \frac{\mathrm{V}_{i\to j}^0 - \langle \mathrm{V}_i^0 \rangle}{\mathrm{SD}(\mathrm{V}_i^0)} \qquad (i = 1, 2, \cdots, L; j = 1, 2, \cdots, 20), \tag{6}$$

where $\mathrm{V}_{i\to j}^0$ represent the original scores directly created by PSI-BLAST [23] that are generally shown as positive or negative integers (the positive score means that the corresponding mutation occurs more frequently in the alignment than expected by chance, while the negative score means just the opposite); the symbol $\langle\rangle$ means taking the average of the quantity therein over 20 native amino acids, and SD means the corresponding standard deviation. The converted values obtained by Eq. (6) will have a zero mean value over the 20 amino acids and will remain unchanged if going through the same conversion procedure again. However, according to the descriptor of Eq. (5), proteins with different lengths will correspond to row-different matrices. To make the descriptor become a size-uniform matrix, one possible avenue is to represent a protein sample **P** by

$$\bar{\mathbf{P}}_{\mathrm{Evo}} = \left[\bar{\mathrm{V}}_1 \quad \bar{\mathrm{V}}_2 \quad \cdots \quad \bar{\mathrm{V}}_{20}\right]^{\mathrm{T}}, \tag{7}$$

where

$$\bar{\mathrm{V}}_j = \frac{1}{L}\sum_{i=1}^{L} \mathrm{V}_{i\to j} \qquad (j = 1, 2, \cdots, 20) \tag{8}$$

where $\bar{\mathrm{V}}_j$ represents the average score of the amino acid residues in the protein **P** being changed to amino acid type j during the evolution process. However, if $\bar{\mathbf{P}}_{\mathrm{Evo}}$ of Eq. (7) was used to represent the protein **P**, all the sequence-order information during the evolution process would be lost. To avoid complete loss of the sequence-order information, the concept of the pseudo amino acid composition (PseAAC) as originally proposed in [17] was adopted; i.e., instead of Eq. (7), let us use the pseudo position-specific scoring matrix as given by

$$\mathbf{P}_{\mathrm{PseEvo}}^{\lambda} = \left[\bar{\mathrm{V}}_1 \quad \bar{\mathrm{V}}_2 \quad \cdots \quad \bar{\mathrm{V}}_{20} \quad \mathrm{V}_1^{\lambda} \quad \mathrm{V}_2^{\lambda} \quad \cdots \quad \mathrm{V}_{20}^{\lambda}\right]^{\mathrm{T}} \tag{9}$$

to represent the protein **P**, where

$$\mathrm{V}_j^{\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\left[\mathrm{V}_{i\to j} - \mathrm{V}_{(i+\lambda)\to j}\right]^2 \qquad (j = 1, 2, \cdots, 20; \lambda < L), \tag{10}$$

meaning that V_j^1 is the correlation factor by coupling the most contiguous position-specific scoring matrix scores along the protein chain for the amino acid type j; V_j^2 that by coupling the second-most contiguous position-specific scoring matrix scores; and so forth. Note that, as mentioned under Materials, the length of the shortest protein sequence in the benchmark dataset is $L = 50$, and hence the value allowed for λ in Eq. (10) must be smaller than 50. When $\lambda = 0$, V_j^{λ} becomes a naught element and Eq. (9) is degenerated to Eq. (7).

Prediction process

The prediction engine used in this study is the ensemble classifier $\mathbb{C}^E$ formed by fusing many basic individual classifiers operated according to the OET-KNN (Optimized Evidence-Theoretic K Nearest Neighbor) rule [32]. For a detailed mathematical formulation of $\mathbb{C}^E$, refer to [10]. $\mathbb{C}^E$ is a very powerful classification engine as demonstrated by many previous studies (see, e.g., [33,34] as well as a list of relevant papers cited in [10]). Note that in order to make the prediction engine able to deal with the system containing both single and multiple location proteins, the ensemble classifier $\mathbb{C}^E$ needs to be modified as formulated by $\mathbb{C}^E(\theta)$, where θ is a threshold parameter for controlling the count of multiple locations and optimizing the predicted results, as elaborated in [10].

The prediction process is governed by the following criterion.

(1) If the query protein can be expressed as a meaningful or productive descriptor in the GO database via its representative proteins, then $\mathbf{P}_{GO}$ of Eq. (1) should be input into the prediction engine for identifying its subcellular location site(s). And the output will be determined by fusing many basic OET-KNN predictors with different parameters of K [10].
(2) If the query protein does not have significant homology to any protein in the Swiss-Prot database, or its representative proteins do not contain any useful GO information, then both the FunD representation $\mathbf{P}_{FunD}$ of Eq. (3) and the pseudo position-specific scoring matrix representation $\mathbf{P}^{\lambda}_{PseEvo}$ of Eq. (9) should be input into the prediction engine, as described in [35]. The output will be determined by fusing many basic OET-KNN predictors with different parameters of K and λ [10].

The entire ensemble classifier thus established is called Hum-mPLoc 2.0, where "m" right before "PLoc" stands for "multiple" meaning it can be used to deal with proteins with both single and multiple subcellular locations, and "2.0" refers to an updated version evolved from Hum-mPLoc [15]. To provide an intuitive picture, a flowchart is given in Fig. 1 to illustrate the prediction process of Hum-mPLoc 2.0.

It is instructive to point out that during the prediction process, rather than equally treating the GO, FunD, and evolutional representations, we adopted the hierarchy as described above and Fig. 1. The reasons doing so are as follows. (1) GO representation can generally yield overwhelmingly higher success rates than the other representations in predicting protein subcellular localization (see, e.g., a recent review [10] as well as the references cited therein). (2) The FunD database is still quite limited yet, and hence many protein sequences cannot be meaningfully defined in the FunD space (see, e.g., [31,35–37]). (3) Many preliminary tests have indicated that the method with the flowchart as described in Fig. 1 can lead to the highest overall cross-validation success rate. However, it will certainly be worthy of reconsidering the hierarchy with more complete FunD databases available in future.

Also, the current prediction method, just like many others in this area, is developed for predicting the subcellular localization of proteins rather than peptides. As is well known, in most cases a protein molecule usually contains more than 50 amino acids. Again, we'll consider the short peptide location problem in future when the statistically sufficient "short peptide subcellular location" data are available.

Results and discussion

In statistical prediction, the following three methods are often used to examine the quality of a predictor: independent dataset test, subsampling (such as dividing a benchmark dataset into 5 or 10 subgroups) test, and jackknife test [38]. Of these three cross-validations, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset [10], and hence has been increasingly and widely used to test the power of various statistical predictors (see, e.g., [39–54]). Note that it is a little more complicated to count the success prediction rate for a system containing both single-location proteins and multiple-location proteins. For example, the number of the total prediction targets should be the total number of different locative proteins rather than the total number of different proteins, as elaborated in [15,55].

The overall jackknife success rate by Hum-mPLoc 2.0 on the current benchmark dataset was 62.7%. This indicates that, compared with Hum-mPLoc [15], a significant advance has been made in enhancing the power of Hum-mPLoc 2.0, as can be viewed from the following aspects.

(1) The current benchmark dataset is extremely stringent that covers 14 subcellular location sites where none of proteins included has $\geqslant$25% pairwise sequence identity to any other in a same subset (subcellular location). As is well known, the more stringent of a benchmark dataset in excluding homolo-

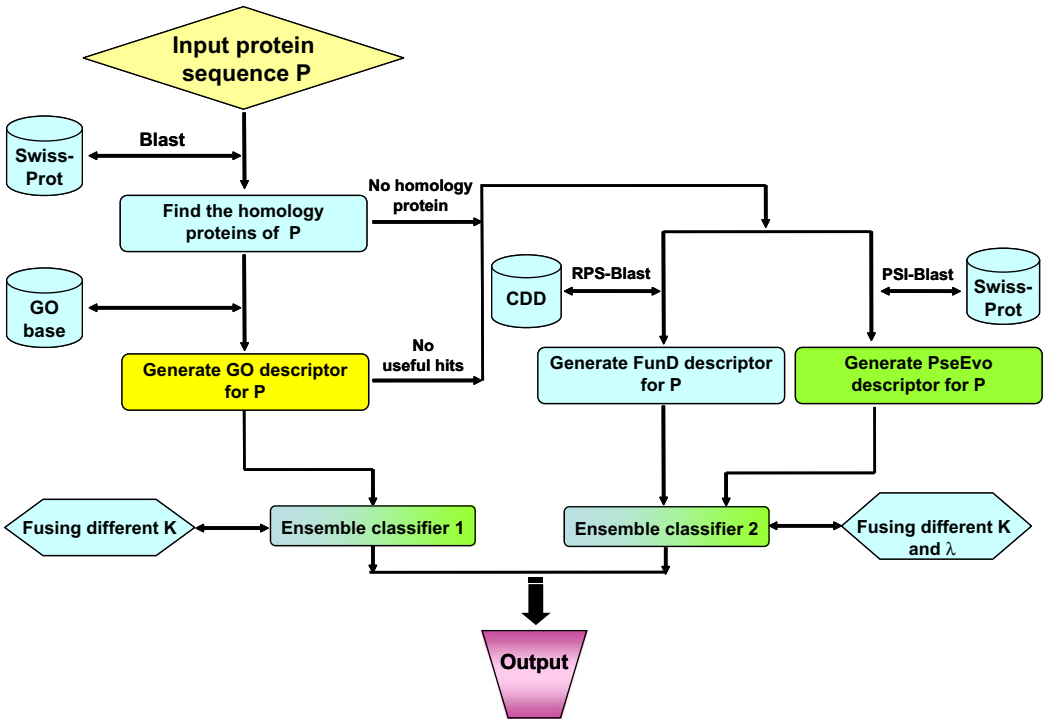

Fig. 1. A flowchart to show the prediction process of Hum-mPLoc 2.0, where ensemble classifier 1 is for processing the GO descriptor samples, while ensemble classifier 2 is for the FunD and PseEvo descriptor samples. See the text for further explanation.

gous sequences, or the more subcellular locations under its coverage, the more difficult it will be to enhance the overall success rate.

(2) Inclusion of proteins with multiple location sites will further complicate the difficulty of prediction.

(3) Besides Hum-mPLoc [15], so far no other predictor is available for dealing with a system of human proteins with both single and multiple subcellular locations.

(4) Although Hum-mPLoc could yield about 70% overall success rate as reported in [15], the accession number information of query proteins was required for utilizing the higher-level GO approach to perform the prediction. If no such information was given for the input data, then all the predictions by Hum-mPLoc would be performed without using any GO information, and the overall jackknife success rate by Hum-mPLoc [15] on the current stringent dataset would drop to 38.1%, which is about 25% lower than that by the current Hum-mPLoc 2.0.

(5) For those proteins without useful GO information, the simple PseAAC approach [17] was used in Hum-mPLoc, but in the current Hum-mPLoc 2.0, a much more sophisticated PseAAC approach is implemented by fusing the functional domain information and the sequential evolution information. This will enhance the success rate by 10% in comparison with the simple PseAAC approach.

Conclusions

Hum-mPLoc 2.0 is an updated predictor evolved from Hum-mPLoc [15] for dealing with the system of human proteins with both single subcellular location and multiple locations. In the updated predictor, a top-down approach to enhance the prediction power has been implemented via the following aspects. (1) The input of accession number for using the higher-level GO approach [19] to perform the prediction is no longer needed; this is particularly useful when dealing with synthetic proteins or hypothetical proteins, as well as those newly-discovered proteins without accession numbers assigned yet. (2) For those proteins without useful GO information to conduct the higher-level prediction, a more advanced PseAAC approach by fusing the FunD information and sequential evolution information is implemented to replace the simple PseAAC approach [17]. Hum-mPLoc 2.0 is available as a web server at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/, by which one can get the desired result for a query protein sequence in about 15 s.

Acknowledgments

The authors thank the two anonymous reviewers whose constructive comments are very helpful for strengthening the presentation of this paper. This work was supported by the National Natural Science Foundation of China (Grant 60704047), Science and Technology Commission of Shanghai Municipality (Grants 08ZR1410600 and 08JC1410600), and sponsored by Shanghai Pujiang Program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2009.07.046.

References

[1] K. Nakai, P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem. Sci. 24 (1999) 34–36.
[2] K.C. Chou, D.W. Elrod, Protein subcellular location prediction, Protein Eng. 12 (1999) 107–118.
[3] R.F. Murphy, M.V. Boland, M. Velliste, 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 (2000) 251–259.
[4] K.J. Park, M. Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs, Bioinformatics 19 (2003) 1656–1663.
[5] S. Matsuda, J.P. Vert, H. Saigo, N. Ueda, H. Toh, T. Akutsu, A novel representation of protein sequences for prediction of subcellular location using support vector machines, Protein Sci. 14 (2005) 2804–2813.
[6] K. Lee, D.W. Kim, D. Na, K.H. Lee, D. Lee, PLPD: reliable protein localization prediction from imbalanced and overlapped datasets, Nucleic Acids Res. 34 (2006) 4655–4666.
[7] S.W. Zhang, Y.L. Zhang, H.F. Yang, C.H. Zhao, Q. Pan, 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 (2008) 565–572.
[8] J.Y. Shi, S.W. Zhang, Q. Pan, G.P. Zhou, Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution, Amino Acids 35 (2008) 321–327.
[9] K. Nakai, Protein sorting signals and prediction of subcellular localization, Adv. Protein Chem. 54 (2000) 277–344.
[10] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[11] A. Garg, M. Bhasin, G.P. Raghava, Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search, J. Biol. Chem. 280 (2005) 14427–14432.
[12] K.C. Chou, H.B. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, Biochem. Biophys. Res. Commun. 347 (2006) 150–157.
[13] C. Smith, Subcellular targeting of proteins and drugs, 2008. Available from: <http://www.biocompare.com/Articles/TechnologySpotlight/976/Subcellular-Targeting-Of-Proteins-And-Drugs.html>.
[14] E. Glory, R.F. Murphy, Automated subcellular location determination and high-throughput microscopy, Dev. Cell 12 (2007) 7–16.
[15] H.B. Shen, K.C. Chou, Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, Biochem. Biophys. Res. Commun. 355 (2007) 1006–1011.
[16] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[17] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. 43 (2001) 246–255. Erratum, Proteins: Struct. Funct. Genet. 44 (2001) 60.
[18] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[19] K.C. Chou, H.B. Shen, Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2008) 153–162.
[20] H.B. Shen, K.C. Chou, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.
[21] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, R. Apweiler, The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro, Genome Res. 13 (2003) 662–672.
[22] D. Barrell, E. Dimmer, R.P. Huntley, D. Binns, C. O'Donovan, R. Apweiler, The GOA database in 2009—an integrated Gene Ontology Annotation resource, Nucleic Acids Res. 37 (2009) D396–D403.
[23] A.A. Schaffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, Nucleic Acids Res. 29 (2001) 2994–3005.
[24] Y. Loewenstein, D. Raimondo, O.C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, A. Tramontano, Protein function annotation by homology-based inference, Genome Biol. 10 (2009) 207.
[25] M. Gerstein, J.M. Thornton, Sequences and topology, Curr. Opin. Struct. Biol. 13 (2003) 341–343.
[26] K.C. Chou, Review: structural bioinformatics and its impact to biomedical science, Curr. Med.Chem. 11 (2004) 2105–2134.
[27] I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz, P. Bork, SMART 5: domains in the context of genomes and networks, Nucleic Acids Res. 34 (2006) D257–D260.
[28] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucleic Acids Res. 34 (2006) D247–D251.
[29] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale, The COG database: an updated version includes eukaryotes, BMC Bioinf. 4 (2003) 41.

[30] A. Marchler-Bauer, J.B. Anderson, M.K. Derbyshire, C. DeWeese-Scott, N.R. Gonzales, M. Gwadz, L. Hao, S. He, D.I. Hurwitz, J.D. Jackson, Z. Ke, D. Krylov, C.J. Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, G.H. Marchler, M. Mullokandov, J.S. Song, N. Thanki, R.A. Yamashita, J.J. Yin, D. Zhang, S.H. Bryant, CDD: a conserved domain database for interactive domain family analysis, Nucleic Acids Res. 35 (2007) D237–D240.
[31] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[32] T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Trans. Sys. Man Cybernetics 25 (1995) 804–813.
[33] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers, J. Proteome Res. 5 (2006) 1888–1897.
[34] S.W. Zhang, Q. Pan, H.C. Zhang, Z.C. Shao, J.Y. Shi, Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion, Amino Acids 30 (2006) 461–468.
[35] H.B. Shen, K.C. Chou, QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information, J. Proteome Res. 8 (2009) 1577–1584.
[36] Y.D. Cai, G.P. Zhou, K.C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (2003) 3257–3263.
[37] X. Xiao, P. Wang, K.C. Chou, Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition, J. Appl. Crystallogr. 42 (2009) 169–173.
[38] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[39] G.F. Zhou, X. Xu, C.T. Zhang, A weighting method for predicting protein structural class from amino acid composition, Eur. J. Biochem. 210 (1992) 747–749.
[40] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins: Struct. Funct. Genet. 50 (2003) 44–48.
[41] S.W. Zhang, Q. Pan, H.C. Zhang, Y.L. Zhang, H.Y. Wang, Classification of protein quaternary structure with support vector machine, Bioinformatics 19 (2003) 2390–2396.
[42] G.Y. Zhang, H.C. Li, B.S. Fang, Predicting lipase types by improved Chou's pseudo-amino acid composition, Protein Pept. Lett. 15 (2008) 1132–1137.
[43] S. Kannan, A.M. Hauth, G. Burger, Function prediction of hypothetical proteins without sequence similarity to proteins of known function, Protein Pept. Lett. 15 (2008) 1107–1116.
[44] K.C. Chou, H.B. Shen, ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information, Biochem. Biophys. Res. Commun. 376 (2008) 321–325.
[45] S.W. Zhang, W. Chen, F. Yang, Q. Pan, Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach, Amino Acids 35 (2008) 591–598.
[46] M.A. Rezaei, P. Abdolmaleki, Z. Karami, E.B. Asadabadi, M.A. Sherafat, H. Abrishami-Moghaddam, M. Fadaie, M. Forouzanfar, Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks, J. Theor. Biol. 254 (2008) 817–820.
[47] L. Nanni, A. Lumini, A further step toward an optimal ensemble of classifiers for peptide classification, a case study: HIV protease, Protein Pept. Lett. 16 (2009) 163–167.
[48] J.Y. Yang, Z.L. Peng, Z.G. Yu, R.J. Zhang, V. Anh, D. Wang, Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation, J. Theor. Biol. 257 (2009) 618–626.
[49] K.C. Chou, H.B. Shen, FoldRate: a web-server for predicting protein folding rates from primary sequence, Open Bioinf. Res. 3 (2009) 31–50. Available from: <http://www.bentham.org/open/tobioij/>.
[50] H. Ding, L. Luo, H. Lin, Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition, Protein Pept. Lett. 16 (2009) 351–355.
[51] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, J. Theor. Biol. 257 (2009) 17–26.
[52] H.B. Shen, J.N. Song, K.C. Chou, Prediction of protein folding rates from primary sequence by fusing multiple sequential features, J. Biomed.Sci. Eng. (JBiSE) 2 (2009) 136–143. Available from: <http://www.srpublishing.org/journal/jbise/>.
[53] C. Chen, L. Chen, X. Zou, P. Cai, Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine, Protein Pept. Lett. 16 (2009) 27–31.
[54] Y.H. Zeng, Y.Z. Guo, R.Q. Xiao, L. Yang, L.Z. Yu, M.L. Li, Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, J. Theor. Biol. 259 (2009) 366–372.
[55] K.C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, J. Proteome Res. 6 (2007) 1728–1734.