Virus-ECC-mPLoc: A Multi-Label Predictor for Predicting the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites Based on a General Form of Chou's Pseudo Amino Acid Composition

Xiao Wang, Guo-Zheng Li*

The Key Laboratory of Embedded System and Service Computing, Ministry of Education,
Department of Control Science and Engineering,
Tongji University,
Shanghai 201804, China

Abstract. Protein subcellular localization aims at predicting the location of a protein within a cell using computational methods. Knowledge of subcellular localization of viral proteins in a host cell or virus-infected cell is important because it is closely related to their destructive tendencies and consequences. Prediction of viral protein subcellular localization is an important but challenging problem, particularly when proteins may simultaneously exist at, or move between, two or more different subcellular location sites. Most of the existing protein subcellular localization methods specialized for viral proteins are only used to deal with the single-location proteins. To better reflect the characteristics of multiplex proteins, a new predictor, called Virus-ECC-mPLoc, has been developed that can be used to deal with the systems containing both singleplex and multiplex proteins by introducing a powerful multi-label learning approach which exploits correlations between subcellular locations and by hybridizing the gene ontology information with the dipeptide composition information. It can be utilized to identify viral proteins among the following six locations: (1) viral capsid, (2) host cell membrane, (3) host endoplasmic reticulum, (4) host cytoplasm, (5) host nucleus, and (6) secreted. Experimental results show that the overall success rates thus obtained by Virus-ECC-mPLoc are 86.9% for jackknife test and 87.2% for independent data set test, which are significantly higher than that by any of the existing predictors. As a user-friendly web-server, Virus-ECC-mPLoc is freely accessible to the public at the web-site http://levis.tongji.edu.cn:8080/bioinfo/Virus-ECC-mPLoc/.

* Address correspondence to this author at the The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Department of Control Science and Engineering. 4800 Cao An Road. Shang Hai, China, 201804. Tel: +86 021 6958 3706; Fax: +86 021 6958 3706; E-mail: gzli@tongji.edu.cn

Keywords: Protein subcellular localization; Multi-label learning; Classifier Chain; Multiplex proteins

1 Introduction

A virus is a small particle that infects cells in various organisms. As acellular organisms and obligate intracellular parasites, viruses can reproduce themselves only by invading and taking over other cells as they lack the cellular machinery for self-reproduction. Although viruses are acellular organisms, viral proteins are required to reside in various cellular compartments of the host cell or virus-infected cell to perform their functions. Therefore, knowledge of the subcellular localization of viral proteins within a host cell or virus-infected cell is very useful for studying the function of viral proteins and designing antiviral drugs because it is closely related to their destructive tendencies and consequences.

Although the subcellular localization of a protein can be determined by carrying out various biochemical experiments, the approach by purely doing experiments is both time consuming and high cost. In the post-genomic age, the gap between newly found protein sequences and the information of their subcellular localization is becoming increasingly wide. For example, according to the Swiss-Prot database,14 version 50.0 released on May 30, 2006, the number of viral proteins with reliable subcellular location annotations is about 12% of all the viral protein entries concerned [72]. To bridge such a gap, it is highly desirable to develop computational methods to predict protein subcellular localization automatically and accurately. During the past decade, many efforts have been devoted to deal with such a challenge, and a large number of computational methods have been developed in an attempt to predict the subcellular localization of proteins (see, e.g., [2,11–14,23,24,30,38,39,50,56–58,63,66,72,73,76,82]). In particular, machine-learning based approaches, such as Neural Networks, K Nearest Neighbor, Support Vector Machine, AdaBoost, are widely used to solve this problem. Among these approaches, Support Vector Machine is widely adopted in bioinformatics and is shown to achieve better performance as compared with others.

However, proteins may simultaneously reside at, or move between, two or more different subcellular locations. Unfortunately, the aforementioned methods didn't take multiple-location or multiplex proteins into account when predicting protein subcellular localization. In general, they were established under the assumption that a protein resides at one, and only one, subcellular location. 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 [33, 74]. In particular, recent evidences have indicated that an increasing number of proteins have multiple locations in the cell, as indicated by Millar et al. [51].

Recently, a powerful predictor, called iLoc-Virus [80] was developed that can be used to predict the subcellular localization of viral proteins among their 6 location sites in which some of the proteins may belong to two and more subcellular locations. To the best of our knowledge, iLoc-Virus is at present the best predictor able to deal with multiple-location or multiplex proteins when predicting viral protein subcellular localization. However, ML-KNN classifier used by iLoc-Virus is not optimal because it doesn't take correlations between subcellular locations into account.

In this paper, to better reflect the characteristics of multiplex proteins, a new predictor, called Virus-ECC-mPLoc, has been developed that can be used to deal with the systems containing both singleplex and multiplex proteins by introducing a powerful multi-label learning approach which exploits correlations between subcellular locations and by hybridizing the gene ontology information with the dipeptide composition information. Our experimental results on a benchmark dataset consisting of 207 viral protein sequences show that the overall success rates thus obtained by Virus-ECC-mPLoc are 86.9% for jackknife test and 87.2% for independent data set test, which are significantly higher than that by iLoc-Virus predictor.

According to a recent comprehensive review [9], to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target concerned; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us elaborate how to deal with these steps one-by-one.

2 Materials and Methods

2.1 Datasets

In this study, we use the same dataset $\mathbb{X}$ in iLoc-Virus [80] as the benchmark dataset for the current study. Using the dataset $\mathbb{X}$ will make it easier to compare our new predictor with the existing one because the tested results by iLoc-Virus on $\mathbb{X}$ have been reported [80]. Furthermore, the dataset is constructed specialized for viral proteins, where none of proteins included in $\mathbb{X}$ has greater than or equal to 25% pairwise sequence identity to any other in a same subcellular location compared with most of the other benchmark datasets in this area.

The dataset $\mathbb{X}$ contains 207 viral protein sequences, of which 165 belong to one subcellular location, 39 to two locations, 3 to three locations,and none to four or more locations. The dataset covers 6 subcellular locations as shown in Fig.1, and hence can be represented as

$$\mathbb{X} = \mathbb{X}_1 \cup \mathbb{X}_2 \cup \mathbb{X}_3 \cup \mathbb{X}_4 \cup \mathbb{X}_5 \cup \mathbb{X}_6 \tag{1}$$

where $\mathbb{X}_1$ represents the subset for the subcellular location of "viral capsid", $\mathbb{X}_2$ for "host cell membrane", $\mathbb{X}_3$ for "host endoplasmic reticulum", and so forth.

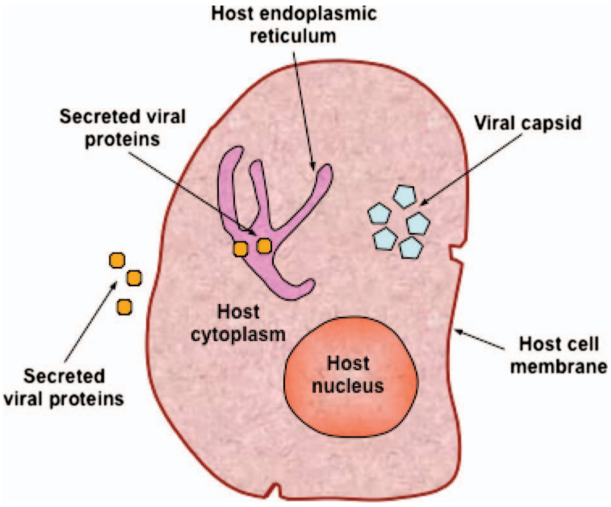

Fig. 1. Illustration to show the 6 subcellular locations of viral proteins. The 6 subcellular locations are: (1) viral capsid, (2) host cell membrane, (3) host endoplasmic reticulum, (4) host cytoplasm, (5) host nucleus, and (6) secreted.

A breakdown of the 207 viral proteins in the benchmark dataset $\mathbb{X}$ according to their six location sites is given in Table 1. To avoid redundancy and homology bias, none of the proteins in $\mathbb{X}$ has greater than or equal to 25% pairwise sequence identity to any other in a same subset. For convenience, hereafter let us just use the subscripts of Eq.(1) as the codes of the 6 location sites; i.e., "1" for "viralcapsid", "2" for "host cell membrane", "3" for "host endoplasmic reticulum", and so forth (Table 2).

Note that because some proteins may occur in two different locations, the 207 different proteins actually correspond to 252 "locative proteins" (Table 1). For the concept of locative proteins, readers are referred to [15,67,71] where the difference between "protein" and "locative protein" and their relationship are elaborated.

For readers' convenience, the corresponding accession numbers and protein sequences in $\mathbb{X}$ are given in Online Supporting Information A.

2.2 Feature Extraction

To develop a powerful method for statistically predicting protein subcellular localization, one of the most important things is to extract core and essential features of protein samples that are closely correlated with their subcellular locations. To avoid losing many important information hidden in protein sequences, the pseudo amino acid composition (PseAAC) was proposed [5,7] to replace the simple amino acid composition (AAC) for representing the sample of a protein. For a brief introduction about Chou's PseAAC, visit the Wikipedia web-page at

Table 1. Breakdown of the viral protein benchmark dataset $\mathbb{X}$ taken from [80]

Subset	Subcellular location	Number of proteins
$\mathbb{X}_1$	Viral capsid	8
$\mathbb{X}_2$	Host cell membrane	33
$\mathbb{X}_3$	Host endoplasmic reticulum	20
$\mathbb{X}_4$	Host cytoplasm	87
$\mathbb{X}_5$	Host nucleus	84
$\mathbb{X}_6$	Secreted	20
Total number of locative proteins N(loc)		252^a
Total number of different proteins N(seq)		207^b

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

[a] See Eqs.36-38 of [16] for the definition about the number of locative proteins, and its relation with the number of different proteins.

[b] Of the 207 different proteins, 165 have one subcellular location, 39 have two locations, and 3 have three locations. See Online Supporting Information A for the protein sequences.

http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. For a summary about its recent development and applications, see a comprehensive review [8]. Ever since the concept of PseAAC was proposed by Chou [5] in 2001, its has rapidly penetrated into almost all the fields of protein attribute prediction, such as identifying bacterial virulent proteins [55], predicting homo-oligomeric proteins [61], predicting protein secondary structure content [4], predicting supersecondary structure [89], predicting protein structural classes [45, 64], predicting protein quaternary structure [87], predicting enzyme family and sub-family classes [60,78,88], predicting protein subcellular location [44,87], predicting subcellular localization of apoptosis proteins [27, 40, 42, 47], predicting protein subnuclear location [41], predicting protein submitochondria locations [46, 54, 84], identifying cell wall lytic enzymes [26], identifying risk type of human papillomaviruses [28], identifying DNA-binding proteins [29], predicting G-Protein-Coupled Receptor Classes [34, 59], predicting protein folding rates [35], predicting outer membrane proteins [36], predicting cyclin proteins [52], predicting GABA(A) receptor proteins [53], identifying bacterial secreted proteins [83], identifying the cofactors of oxidoreductases [85], identifying lipase types [86], identifying protease family [37], predicting Golgi protein types [25], classifying amino acids [31], among many others. In the present study, we adopted *Gene Ontology* and *Dipeptide Composition* feature extraction methods to generate features of protein examples, which are widely used in many existing protein subcellular localization systems [15, 19–21, 43, 67–71, 80, 81]. For reader's convenience, a brief introduction on *Gene Ontology* and *Dipeptide Composition* is given below.

***GO* (Gene Ontology)** *GO* database [1] was established according to the molecular function, biological process, and cellular component. Accordingly, protein samples defined in a *GO* database space would be clustered in a way better reflecting their subcellular locations [16, 17]. So far, there are two main approaches to extract features from *GO* database space. However, in order to incorporate more information, instead of only using 0 and 1 elements as done in [19], here let us use another better approach [21] as described below.

Step 1 Compression and reorganization of the existing *GO* numbers. The *GO* database (version 94 released 08 April 2011) contains many *GO* numbers. However, these numbers do not increase successively and orderly. For easier handling, some reorganization and compression procedure was taken to renumber them. The *GO* database obtained through such a treatment is called GO_compress database, which contains 18,844 numbers increasing successively from 1 to the last one.

Step 2 According to Eq.6 of a recent comprehensive review [10], the general form of Chou's PseAAC can be formulated as

$$P = [\psi_1, \psi_2, \psi_3, \cdots, \psi_u, \cdots, \psi_\Omega]^T \tag{2}$$

where **T** is a transpose operator, while the subscript Ω reflects the dimension of the vector and its value as well as the components ψ_1, ψ_2, $\cdots$, ψ_Ω will be defined by the feature extractions as elaborated below. Using Eq.(2) with $\Omega = 18,844$, the protein P can be represented as

$$P_{GO} = \left[f_1^G, f_2^G, f_3^G, \cdots, f_u^G, \cdots, f_{18844}^G\right]^T \tag{3}$$

where $f_u^G (u = 1, 2, ..., 18, 844)$ are defined via the following steps.

Step 3 Use BLAST [65] to search the homologous proteins of the protein P from the Swiss-Prot database (version 55.3), with the expect value $E \leq 0.001$ for the BLAST parameter.

Step 4 Those proteins which have ≥60% pairwise sequence identity with the protein P are collected into a set, $\mathbb{X}^{P-homo}$, called the "homology set" of P. All the elements in $\mathbb{X}^{P-homo}$ can be deemed as the "representative proteins" of P, sharing some similar attributes such as structural conformations and biological functions [6,32,49]. Because they were retrieved from the Swiss-Prot database, these representative proteins must each have their own accession numbers.

Step 5 Search the *GO* database at http://www.ebi.ac.uk/GOA/ to find the corresponding *GO* number(s) [3] for each of the accession numbers collected in Step 4, followed by converting the *GO* numbers thus obtained to their GO_compress numbers as described in Step 1. (Note that the relationships between the UniProtKB/Swiss-Port protein entries and the *GO* numbers may be one-to-many, "reflecting the biological reality that a particular protein may function in several processes, contain domains that carry out diverse molecular functions, and participate in multiple alternative interactions with other proteins, organelles or locations in the cell" [1]. For example, the

Uni-ProtKB/ Swiss-Prot protein entry "P01040" corresponds to three *GO* numbers, i.e., "GO:0004866", "GO:0004869", and "GO:0005622").

Step 6 the elements in Eq.3 is given by

$$f_u^G = \frac{\sum_{k=1}^{N(rep)} g(u,k)}{N(rep)} (u = 1, 2, \cdots, 18844) \tag{4}$$

where $N(rep)$ is the number of representative proteins in $\mathbb{X}^{P-homo}$, and

$$g(u,k) = \begin{cases} 1, & \text{if the k-th representative protein hits the u-th} \\ & \text{GO\_compress number} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Note that the *GO* feature extraction method may become a naught vector or meaningless under any of the following situations: (1) the protein P does not have significant homology to any protein in the Swiss-Prot database, i.e., $\mathbb{X}^{P-homo} = \emptyset$ meaning the homology set $\mathbb{X}^{P-homo}$ is an empty one; (2) its representative proteins do not contain any useful *GO* information for statistical prediction based on a given training dataset.

Under such a situation, let us consider using the dipeptide composition method to extract features for the protein P, as described below.

***DC* (Dipeptide Composition)** Dipeptide composition represents the occurrence frequency of each two adjacent amino acid residues. It is used to describe the global information about each protein sequence in the form of 420- dimensional (420-D) feature vector. An advantage of *DC* over amino acid composition is that it uses some sequence-order information. Dipeptide composition will generate 420 components for each protein sequence, the first 20 components are the conventional amino acid composition(AAC); the following 400 components are the fractions of 400 dipeptides, i.e. AA, AC, AD, ... , YV, YW, YY; the 400 components are calculated using the following equation

$$\text{fraction of dip(i)} = \frac{\text{total number of dip(i)}}{\text{total number of all possible dipeptides}} \tag{6}$$

where dip(i) is the i-th dipeptide of the 400 dipeptides, i=1, 2 ,..., 400.

2.3 Prediction Algorithm: Ensemble of Classifier Chains

To enhance the success rate, the powerful ECC (Ensemble of Classifier Chains) classifier [62] was adopted to perform prediction. Below, let us briefly introduce the Ensemble of Classifier Chains classifier.

Consider the problem of classifying N proteins into M subcellular locations, which can be formulated as

$$\mathbb{Y} = \{\lambda_1, \lambda_2, ...\lambda_\mu, ..., \lambda_M\} \tag{7}$$

The available information is assumed to consist in a training dataset

$$T = \{(P_1, Y_1), (P_2, Y_2), \cdots, (P_N, Y_N)\} \tag{8}$$

where the N proteins P_i $(i = 1, 2, ..., N)$ and their corresponding class labels Y_i $(i = 1, 2, ..., N)$ is a subset of $\mathbb{Y}$ of Eq.7. This indicates that each protein may belong to two or more subcellular locations.

Binary relevance (BR) [75] is a popular approach to convert a multi-label learning problem into a number of independent binary classification ones. Specifically, a separate classifier h_μ is learned for each class label λ_μ in $\mathbb{Y}$, where each protein associated with class label set Y will be regarded as positive example when class label $\lambda_\mu \in Y$ while regarded as negative example when class label $\lambda_\mu \notin Y$. For the classification of a new protein, BR outputs the union of the class labels that are predicted by the M classifiers. BR is conceptually simple and easy to implement, whereas may be less effective since it don't take label correlations into account.

However, in contrast to BR, individual classifiers of CC (Classifier Chain) have to be trained sequentially. Furthermore, classifiers are linked along a chain where each classifier is responsible for prediction of presence or absence of class label $\lambda_\mu \in \mathbb{Y}$. The feature space of each classifier in the chain is extended with the 0/1 label associations of all previous classifiers. In other words, a chain $C_1, \cdots, C_M$ of binary classifiers is constructed. Each classifier C_μ in the chain is responsible for predicting the binary association of class label λ_μ given the feature space, incremented by all prior binary relevance associations in the chain $\lambda_1, \cdots, \lambda_{\mu-1}$.

The chaining method passes label information between classifiers, allowing CC to take into account label correlations and thus overcoming the label independence problem of BR method. However, the order of the chain itself clearly has an effect on accuracy. In [62], the issue is solved by using an ensemble framework with a different random chain ordering for each iteration.

In contrast to the traditional single-label ensemble learning, ECC is an ensemble of multiple multi-label methods, i.e. the CC method. Following the typical strategy of ensemble learning, ECC also has two steps, in which the first is to train M CC classifiers $C_1, C_2, \cdots, C_M$ and the second is to combine their predictions. In the first step, each C_k is trained with both a random chain ordering and a random subset of original training data set. In the second step, multi-label predictions of each C_k model are summed by label so that each label gets some votes, and then, we use a threshold to select the most possible labels which form the final multi-label prediction. Specifically, each C_k model predicts a vector $y_k = (l_1^k, \cdots, l_M^k) \in \{0, 1\}^M$. The sums are stored in a vector $W = (l_1, \cdots, l_M) \in R^M$ such that $l_j = \sum_{k=1}^{M} l_j^k$. Hence each $l_j \in W$ represents the sum of the votes for the jth label. We then normalize W to W^{norm}, which represents a distribution of scores for each label in [0, 1]. A threshold is used to choose the final multi-label set Y such that $\lambda_j \in Y$ where $l_j \geq t$ for threshold t. Here we simply set the threshold to be 0.5. Hence the relevant labels in Y represent the final multi-label prediction.

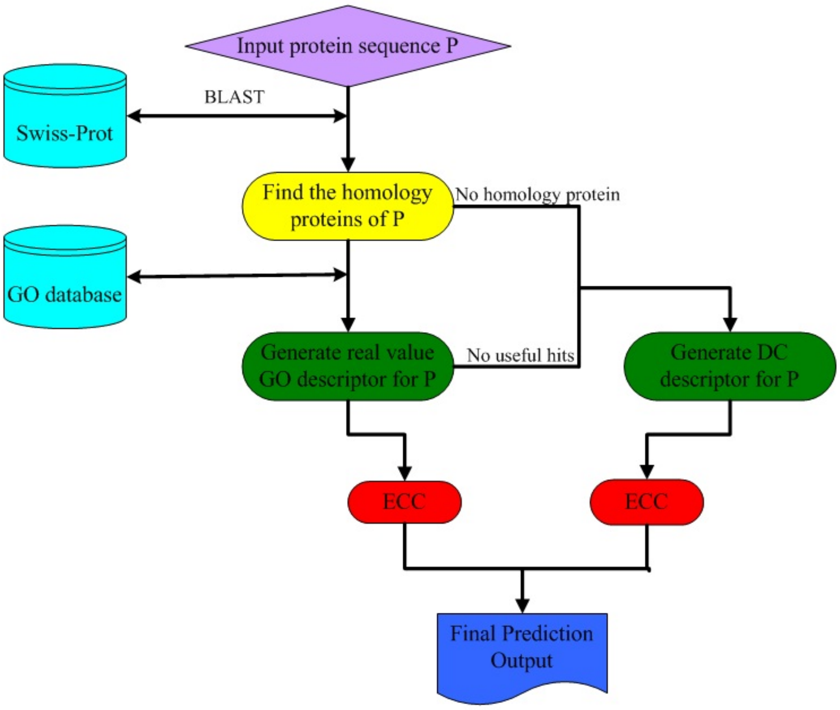

Fig. 2. A flowchart to show the prediction process of Virus-ECC-mPLoc

The entire predictor thus established is called Virus-ECC-mPLoc, which can be used to predict the subcellular localization of both singleplex and multiplex viral proteins. To provide an intuitive picture, a flowchart is provided in Fig. 2 to illustrate the prediction process of Virus-ECC-mPLoc.

3 Results and Discussions

As mentioned in the above section, the benchmark dataset used in this study is $\mathbb{X}$ (cf. Online Supporting Information A), which is the same benchmark dataset constructed in [80] for iLoc-Virus.

To evaluate the proposed new prediction algorithm of this study, we compare it with iLoc-Virus [80]. Actually, for such a dataset containing both single-location and multiple-location viral proteins distributed among 6 subcellular location sites, so far only two existing predictor, i.e., Virus-mPLoc [71] and iLoc-Virus [80], had the capacity to deal with it. Furthermore, iLoc-Virus [80] outperforms Virus-mPLoc [71]. Therefore, to demonstrate the power of the current predictor, it would suffice to just compare Virus-ECC-mPLoc with iLoc-Virus [80].

Table 2. A comparison of the independent data set test success rates by iLoc-Virus [80] and the current Virus-ECC-mPLoc on the benchmark dataset $\mathbb{X}$ (cf. Online Supporting Information A) that covers 6 location sites of viral proteins in which none of the proteins included has ≥25% pairwise sequence identity to any other in a same location.

Code	Subcellular location	Success rate by independent data set test	
		iLoc-Virus^a	Virus-ECC-mPLoc^b
1	Viral capsid	100.0%	100%
2	Host cell membrane	82.3%	91.3%
3	Host endoplasmic reticulum	76.8%	81.33%
4	Host cytoplasm	78.3%	85.2%
5	Host nucleus	88.1%	89.7%
6	Secreted	71.5%	84.5%
Overall		81.1%	87.2%

[a] The predictor from [80].
[b] The predictor proposed in this paper.

Table 3. A comparison of the jackknife success rates by iLoc-Virus [80] and the current Virus-ECC-mPLoc on the benchmark dataset $\mathbb{X}$ (cf. Online Supporting Information A) that covers 6 location sites of viral proteins in which none of the proteins included has ≥25% pairwise sequence identity to any other in a same location.

Code	Subcellular location	Success rate by jackknife test	
		iLoc-Virus^a	Virus-ECC-mPLoc^b
1	Viral capsid	100.0%	100%
2	Host cell membrane	78.8%	90.9%
3	Host endoplasmic reticulum	75.0%	70.0%
4	Host cytoplasm	79.3%	86.2%
5	Host nucleus	88.1%	91.7%
6	Secreted	75.0%	80.0%
Overall		82.1%	86.9%

[a] The predictor from [80].
[b] The predictor proposed in this paper.

In statistical prediction, there are three commonly used methods, that is, the independent data set test, subsampling (e.g., K-fold cross validation) test, and jackknife test, which are often used for testing the accuracy of a statistical prediction method. Among the three methods, the jackknife test is deemed the most objective because it can always yield a unique result for a given benchmark data set, as elucidated in two comprehensive reviews [16,22]. Therefore, the jackknife test has been increasingly and widely adopted by investigators to examine the power of various prediction methods (see, e.g., [13–15,19–21,23,24,48,66–73,77,79–81]). In the present study, we use both independent data set test and jackknife test to evaluate the power of Virus-ECC-mPLoc.

Table 2 and 3 report the detailed results on the 6 viral subcellular locations obtained with iLoc-Virus [80] and Virus-ECC-mPLoc on the aforementioned benchmark dataset $\mathbb{X}$ by the independent data set test and the jackknife test. For a fair algorithmic comparison between Virus-ECC-mPLoc and iLoc-Virus, we use the same GOA database that is described in this study to extract *GO* features. As we can see from Table 2 and 3, for such a stringent dataset, the overall success rate achieved by Virus-ECC-mPLoc is 86.9% for jackknife test which is about 5% higher than that by iLoc-Virus [80], while the overall success rate achieved by Virus-ECC-mPLoc is 87.2% for independent data set test which is about 6% higher than that by iLoc-Virus [80].

Note that during the process of the independent data set test and the jackknife test by iLoc-Virus and Virus-ECC-mPLoc, the false positives (over-predictions) and false negatives (under-predictions) were also taken into account to reduce the scores in calculating the overall success rate. As for the detailed process of how to count the over-predictions and under-predictions for a system containing both single-location and multiple-location proteins, see Eqs.43-48 and Fig. 4 in a comprehensive review [16].

To provide a more intuitive and easier-to-understand measurement, let us introduce a new measure, the so-called "exact match" success rate, to reflect the accuracy of a predictor, as defined by

$$\Lambda = \frac{\sum_{i=1}^{N} \triangle(i)}{N} \tag{9}$$

where Λ represents the exact match rate, N the number of total proteins investigated, and

$$\triangle(i) = \begin{cases} 1, & \text{if all the subcellular locations of the ith protein are} \\ & \text{correctly predicted without any overprediction} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

According to the above definition, for a protein belonging to, say, three subcellular locations, if only two of the three are correctly predicted, or the predicted result contains a location not belonging to the three, the prediction score will be counted as 0. In other words, when and only when all the subcellular locations of a query protein are exactly predicted without any underprediction or overprediction, can the prediction be scored with 1. Therefore, the exact match measure is much more strict and harsh than the measure used previously [16,80] in measuring the success rate. However, even if using such a stringent criterion on the same benchmark dataset, the overall exact match success rates achieved by Virus-ECC-mPLoc are 82.5% for jackknife test and 81.0% for independent data set test, which are about 9% and 6% higher than that by iLoc-Virus [80].

4 Conclusion

Prediction of protein subcellular localization is a challenging problem, particularly when the system concerned contains both singleplex and multiplex proteins. In this paper, we have proposed a novel multi-label predictor, called Virus-ECC-mPLoc, for predicting viral protein subcellular locations based on the powerful ECC algorithm and a hybrid of GO and DC feature extraction methods, which has been demonstrated very powerful for handling the multiplex proteins. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [18], here we have provided a web-server for the method presented in this paper at http://levis.tongji.edu.cn:8080/bioinfo/Virus-ECC-mPLoc/. The current approach represents a new strategy to deal with the multi-label biological problems, and hence may become a useful vehicle in the area of bioinformatics and proteomics.

Acknowledgment

This work was supported by the Natural Science Foundation of China under the Grant No. 60873129 and 61005006, and the Open Projects Program of National Laboratory of Pattern Recognition in China.

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature genetics 25(1), 25 (2000)
2. Bhasin, M., Raghava, G.P.S.: ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Research 32(Web Server), W414–W419 (Jul 2004)
3. Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., Apweiler, R.: The gene ontology annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Research 13(4), 662 –672 (Apr 2003)
4. Chen, C., Chen, L., Zou, X., Cai, P.: Prediction of protein secondary structure content by using the concept of chous pseudo amino acid composition and support vector machine. Protein and Peptide Letters 16(1), 2731 (2009)
5. Chou, K.: Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics 43(3), 246–255 (2001)
6. Chou, K.: Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry 11(16), 21052134 (2004)
7. Chou, K.: Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21(1), 10–19 (2005)
8. Chou, K.: Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6(4), 262274 (2009)
9. Chou, K.: Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273(1), 236–247 (2011)
10. Chou, K.: Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273(1), 236–247 (Mar 2011), http://www.sciencedirect.com/science/article/pii/S002251931000679X

11. Chou, K., Cai, Y.: Using functional domain composition and support vector machines for prediction of protein subcellular location. Journal of Biological Chemistry 277(48), 45765 –45769 (Nov 2002)
12. Chou, K., Elrod, D.W.: Protein subcellular location prediction. Protein Engineering 12(2), 107 –118 (Feb 1999), http://peds.oxfordjournals.org/content/12/2/107.abstract
13. Chou, K., Shen, H.: Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochemical and Biophysical Research Communications 347(1), 150–157 (Aug 2006)
14. Chou, K., Shen, H.: Predicting eukaryotic protein subcellular location by fusing optimized Evidence-Theoretic K-Nearest neighbor classifiers. Journal of Proteome Research 5(8), 1888–1897 (2006)
15. Chou, K., Shen, H.: Euk-mPLoc: a fusion classifier for Large-Scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research 6(5), 1728–1734 (May 2007)
16. Chou, K., Shen, H.: Recent progress in protein subcellular location prediction. Analytical Biochemistry 370(1), 1–16 (Nov 2007)
17. Chou, K., Shen, H.: Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3(2), 153–162 (Jan 2008)
18. Chou, K., Shen, H.: REVIEW: recent advances in developing web-servers for predicting protein attributes. Natural Science 1(2), 63–92 (2009)
19. Chou, K., Shen, H.: A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE 5(4), e9931 (Apr 2010)
20. Chou, K., Shen, H.: Plant-mPLoc: a Top-Down strategy to augment the power for predicting plant protein subcellular localization. PLoS ONE 5(6), e11335 (Jun 2010)
21. Chou, K., Wu, Z., Xiao, X.: iLoc-Euk: a Multi-Label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS ONE 6(3), e18258 (Mar 2011)
22. Chou, K., Zhang, C.: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30(4), 275349 (1995)
23. Chou, K., Shen, H.: Predicting protein subcellular location by fusing multiple classifiers. Journal of Cellular Biochemistry 99(2), 517–527 (Oct 2006)
24. Chou, K., Shen, H.: Largescale plant protein subcellular location prediction. Journal of Cellular Biochemistry 100(3), 665–678 (Feb 2007)
25. Ding, H., Liu, L., Guo, F., Huang, J., Lin, H.: Identify golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. Protein and Peptide Letters 18(1), 5863 (2011)
26. Ding, H., Luo, L., Lin, H.: Prediction of cell wall lytic enzymes using chou's amphiphilic pseudo amino acid composition. Protein and Peptide Letters 16(4), 351355 (2009)
27. Ding, Y., Zhang, T.: 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(13), 1887–1892 (Oct 2008)
28. Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S.: Using the concept of chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263(2), 203–209 (Mar 2010)

29. Fang, Y., Guo, Y., Feng, Y., Li, M.: Predicting DNA-binding proteins: approached from chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34(1), 103–109 (2008)
30. Garg, A., Bhasin, M., Raghava, G.P.S.: 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(15), 14427 –14432 (Apr 2005)
31. Georgiou, D., Karakasidis, T., Nieto, J., Torres, A.: Use of fuzzy clustering technique and matrices to classify amino acids and its impact to chou's pseudo amino acid composition. Journal of Theoretical Biology 257(1), 17–26 (Mar 2009)
32. Gerstein, M., Honig, B.: Sequences and topology. Current Opinion in Structural Biology 11(3), 327329 (2001)
33. Glory, E., Murphy, R.F.: Automated subcellular location determination and High-Throughput microscopy. Developmental Cell 12(1), 7–16 (Jan 2007)
34. Gu, Q., Ding, Y., Zhang, T.: Prediction of G-Protein-Coupled receptor classes in low homology using chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein and Peptide Letters 17(5), 559567 (2010)
35. Guo, J., Rao, N., Liu, G., Yang, Y., Wang, G.: Predicting protein folding rates using the concept of chou's pseudo amino acid composition. Journal of Computational Chemistry 32(8), 1612–1617 (Jun 2011)
36. Hao, L.: The modified mahalanobis discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition. Journal of Theoretical Biology 252(2), 350–356 (May 2008)
37. Hu, L., Zheng, L., Wang, Z., Li, B., Liu, L.: Using pseudo amino acid composition to predict protease families by incorporating a series of protein biological features. Protein and Peptide Letters 18(6), 552558 (2011)
38. Hua, S., Sun, Z.: Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17(8), 721 –728 (2001)
39. Huang, Y., Li, Y.: Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20(1), 21 –28 (Jan 2004)
40. Jian, X., Wei, R., Zhan, T., Gu, Q.: Using the concept of chous pseudo amino acid composition to predict apoptosis proteins subcellular location: An approach by approximate entropy. Protein and peptide letters 15(4), 392396 (2008)
41. Jiang, X., Wei, R., Zhao, Y., Zhang, T.: Using chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34(4), 669–675 (Feb 2008)
42. Kandaswamy, K.K., Pugalenthi, G., Moller, S., Hartmann, E., Kalies, K.U., Suganthan, P.N., Martinetz, T.: Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein and Peptide Letters 17(12), 14731479 (2010)
43. Khan, A., Majid, A., Hayat, M.: CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. Computational Biology and Chemistry 35(4), 218–229 (Aug 2011)
44. Li, F., Li, Q.: Predicting protein subcellular location using chous pseudo amino acid composition and improved hybrid approach. Protein and Peptide Letters 15(6), 612616 (2008)
45. Li, Z., Zhou, X., Dai, Z., Zou, X.: Prediction of protein structural classes by chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37(2), 415–425 (Aug 2009)

46. Lin, H., Ding, H., Guo, F., Zhang, A., Huang, J.: Predicting subcellular localization of mycobacterial proteins by using chou's pseudo amino acid composition. Protein and Peptide Letters 15(7), 739744 (2008)
47. Lin, H., Wang, H., Ding, H., Chen, Y., Li, Q.: Prediction of subcellular localization of apoptosis protein using chous pseudo amino acid composition. Acta biotheoretica 57(3), 321330 (2009)
48. Lin, W., Fang, J., Xiao, X., Chou, K.: iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS ONE 6(9), e24756 (2011)
49. Loewenstein, Y., Raimondo, D., Redfern, O.C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., Tramontano, A.: Protein function annotation by homology-based inference. Genome Biology 10(2), 207 (2009)
50. Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D., Poulin, B., Anvik, J., Macdonell, C., Eisner, R.: Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20(4), 547 –556 (Mar 2004)
51. Millar, A.H., Carrie, C., Pogson, B., Whelan, J.: Exploring the Function-Location nexus: Using multiple lines of evidence in defining the subcellular location of plant proteins. The Plant Cell Online 21(6), 1625 –1631 (Jun 2009)
52. Mohabatkar, H.: Prediction of cyclin proteins using chou's pseudo amino acid composition. Protein and Peptide Letters 17(10), 12071214 (2010)
53. Mohabatkar, H., Beigi, M.M., Esmaeili, A.: Prediction of GABAA receptor proteins using the concept of chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology 281(1), 18–23 (Jul 2011)
54. Nanni, L., Lumini, A.: Genetic programming for creating chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34(4), 653–660 (Jan 2008)
55. Nanni, L., Lumini, A., Gupta, D., Garg, A.: Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Transactions on Computational Biology and Bioinformatics (Aug 2011)
56. Niu, B., Jin, Y., Feng, K., Lu, W., Cai, Y., Li, G.: Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Molecular Diversity 12(1), 41–45 (May 2008)
57. Park, K., Kanehisa, M.: Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19(13), 1656 –1663 (2003)
58. Pierleoni, A., Martelli, P.L., Fariselli, P., Casadio, R.: BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22(14), e408 –e416 (Jul 2006)
59. Qiu, J., Huang, J., Liang, R., Lu, X.: Prediction of g-protein-coupled receptor classes based on the concept of chou's pseudo amino acid composition: An approach from discrete wavelet transform. Analytical Biochemistry 390(1), 68–73 (Jul 2009)
60. Qiu, J., Huang, J., Shi, S., Liang, R.: Using the concept of chous pseudo amino acid composition to predict enzyme family classes: An approach with support vector machine based on discrete wavelet transform. Protein and Peptide Letters 17(6), 715722 (2010)
61. Qiu, J., Suo, S., Sun, X., Shi, S., Liang, R.: OligoPred: a web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into chou's pseudo amino acid composition. Journal of Molecular Graphics and Modelling 30, 129–134 (Sep 2011)
62. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Proceedings of European Conference on Machine Learning and

Principles and Practice of Knowledge Discovery in Databases. pp. 254–269. Bled, Slovenia (2009)

63. Reinhardt, A., Hubbard, T.: Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Research 26(9), 2230–2236 (May 1998)
64. Sahu, S.S., Panda, G.: A novel feature representation method based on chou's pseudo amino acid composition for protein structural class prediction. Computational Biology and Chemistry 34(5-6), 320–327 (Oct 2010)
65. Schffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F.: Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research 29(14), 2994 –3005 (Jul 2001)
66. Shen, H., Chou, K.: Gpos-PLoc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins. Protein Engineering Design and Selection 20(1), 39 –46 (Jan 2007)
67. Shen, H., Chou, K.: 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(4), 1006–1011 (Apr 2007)
68. Shen, H., Chou, K.: Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of gram-positive bacterial proteins. Protein and Peptide Letters 16(12), 1478–1484 (2009)
69. Shen, H., Chou, K.: A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Analytical Biochemistry 394(2), 269–274 (Nov 2009)
70. Shen, H., Chou, K.: Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of gram-negative bacterial proteins. Journal of Theoretical Biology 264(2), 326–333 (May 2010)
71. Shen, H., Chou, K.: Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. Journal of Biomolecular Structure & Dynamics 28(2), 175–186 (Oct 2010)
72. Shen, H., Chou, K.: Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virusinfected cells. Biopolymers 85(3), 233–240 (Feb 2007)
73. Shen, H., Yang, J., Chou, K.: Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33(1), 57–67 (Jan 2007)
74. Smith, C.: Subcellular targeting of proteins and drugs (2008), http://www.biocompare.com/Articles/TechnologySpotlight/976/Subcellular-Target-ing-Of-Proteins-And-Drugs.html
75. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer US, Boston, MA (2010)
76. Wang, J., Sung, W., Krishnan, A., Li, K.: Protein subcellular localization prediction for gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. BMC Bioinformatics 6(1), 174 (2005)
77. Wang, P., Xiao, X., Chou, K.: NR-2L: a Two-Level predictor for identifying nuclear receptor subfamilies based on Sequence-Derived features. PLoS ONE 6(8), e23505 (2011)
78. Wang, Y., Wang, X., Yang, Z., Deng, N.: Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. Protein and Peptide Letters 17(11), 14411449 (2010)

79. Xiao, X., Wang, P., Chou, K.: GPCR-2L: predicting g protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Molecular BioSystems 7(3), 911–919 (2011)
80. Xiao, X., Wu, Z., Chou, K.: iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. Journal of Theoretical Biology 284(1), 42–51 (Sep 2011)
81. Xiao, X., Wu, Z., Chou, K.: A Multi-Label classifier for predicting the subcellular localization of Gram-Negative bacterial proteins with both single and multiple sites. PLoS ONE 6(6), e20592 (Jun 2011)
82. Yu, C., Lin, C., Hwang, J.: Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on npeptide compositions. Protein Science 13(5), 1402–1406 (May 2004)
83. Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., Xiong, W., Qin, W.: SecretP: identifying bacterial secreted proteins by fusing new features into chou's pseudo-amino acid composition. Journal of Theoretical Biology 267(1), 1–6 (Nov 2010)
84. Zeng, Y.h., Guo, Y.z., Xiao, R.q., Yang, L., Yu, L.z., Li, M.l.: Using the augmented chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. Journal of Theoretical Biology 259(2), 366–372 (Jul 2009)
85. Zhang, G., Fang, B.: Predicting the cofactors of oxidoreductases based on amino acid composition distribution and chou's amphiphilic pseudo-amino acid composition. Journal of Theoretical Biology 253(2), 310–315 (Jul 2008)
86. Zhang, G., Li, H., Gao, J., Fang, B.: Predicting lipase types by improved chou's Pseudo-Amino acid composition. Protein and Peptide Letters 15(10), 11321137 (2008)
87. Zhang, S., Chen, W., Yang, F., Pan, Q.: Using chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. Amino Acids 35(3), 591–598 (Apr 2008)
88. Zhou, X., Chen, C., Li, Z., Zou, X.: Using chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology 248(3), 546–551 (Oct 2007)
89. Zou, D., He, Z., He, J., Xia, Y.: Supersecondary structure prediction using chou's pseudo amino acid composition. Journal of Computational Chemistry 32(2), 271–278 (Jan 2011)