Expert Reviews

Predicting multisite protein subcellular locations: progress and challenges

Expert Rev. Proteomics 10(3), 227–237 (2013)

亚细胞多位点预测问题

Pufeng Du[1] and Chao Xu*[2]

[1]*School of Computer Science and Technology, Tianjin University, Tianjin, 300072, China*
[2]*School of Computer Software, Tianjin University, Tianjin, 300072, China*
**Author for correspondence: cxu.cstju@gmail.com*

In the last two decades, predicting protein subcellular locations has become a hot topic in bioinformatics. A number of algorithms and online services have been developed to computationally assign a subcellular location to a given protein sequence. With the progress of many proteome projects, more and more proteins are annotated with more than one subcellular location. However, multisite prediction has only been considered in a handful of recent studies, in which there are several common challenges. In this special report, the authors discuss what these challenges are, why these challenges are important and how the existing studies gave their solutions. Finally, a vision of the future of predicting multisite protein subcellular locations is given.

Keywords: informative features • multisites • single site • subcellular locations

Predicting protein subcellular locations

In the last two decades, life science studies have gradually become complicated processes involving all sorts of modern technologies. Almost every experiment in today's biology laboratory requires sophisticated devices that can only be manufactured with highly developed modern industry, which is armed with automatic controlling systems, high-performance computers and precise manufacturing facilities. Bioinformatics and computational biology are now playing increasingly important roles in life sciences. We are in a new age in which the life sciences have become a precise and quantitative subject; theories and models can be applied to predict a number of experimental results and to guide laboratory practice.

As basic biology has pointed out long ago, a cell is deemed to be the most basic construction unit of almost every living creature on the planet. Every living cell is composed of even more basic components, which are known as the subcellular compartments or subcellular organelles. Most of the subcellular organelles can be roughly considered as isolated spaces surrounded by biomembranes. Some other subcellular structures, such as ribosomes, the cytoskeleton and the centriole, which are nonmembrane bound, may also be recognized as subcellular organelles. All these subcellular structures form a large dynamic system within the cell. The proteins and other macromolecules are synthesized, transferred and activated for their function inside this system. Because different compartments of the cells are usually involved in different biological processes, knowing where the protein is anchored or where the protein would perform its molecular function could be useful in understanding its role in the entire cellular system and would potentially inspire targeted drug discovery.

Although various experimental technologies have been developed for determining the protein subcellular locations, almost every available approach is costly and time consuming. With the development of proteome projects, large amounts of available protein sequences and functional annotations have enabled us to develop computational methods as alternative choices. The computational methods used provide results of low resolution. Only four subcellular compartments were considered when the artificial neuron network was first applied in predicting protein subcellular locations [1]. However, such methods were in so much demand that many efforts have been made in the last two decades to improve the prediction resolution as well as the prediction accuracy.

The basis for computationally predicting protein subcellular locations is the protein sorting theory. According to this theory, every protein

10.1586/EPR.13.16

ISSN 1478-9450

has a highly conserved region in its sequence, preferably close to the N-terminal or C-terminal, which can lead the protein to the correct place during or after the synthetic process. However, with the accumulation of data, this theory cannot explain more than half of the available data. Thus, another theory, especially for globular proteins, was proposed. This theory pointed out that the protein localization is actually a complicated interaction process between the molecular surface of a folded protein and the physicochemical microenvironment inside the subcellular organelles. Thus, the average physicochemical properties of the protein molecule surface must be adapted to the microenvironment to which it is localized [2,3].

According to the aforementioned theory basis, many computational approaches were developed. PSORT servers were the first and the most successful practice in applying the protein sorting theory [4]. By introducing the machine learning algorithms, such as artificial neuron network [1], support vector machine (SVM) [5], covariance separation [6], fuzzy-K nearest neighbor (Fuzzy-KNN) [7], optimized evidence-theoretical-K nearest neighbor (OET-KNN) [8] and multilabel K-nearest neighbor (ML-KNN) [9], a number of online services for identifying protein subcellular locations were established [10,11].

Usually, the subcellular location predictors would assign only one subcellular location to a given protein, assuming that the proteins with multiple localization sites are a minor part of all proteins in the cell. The authors term these predictors 'single-site predictors'. Although most of the relevant studies admitted that some proteins can localize to multiple subcellular organelles, they did not take those proteins into consideration. Unfortunately, this is actually a conceptual bias, which is the result of limited data resources in the early stage of computational protein science. According to the report of DBMLoc, more than 30% of proteins localized to more than one subcellular location [12]. Thus, having multiple subcellular localizations is a common phenomenon for proteins. The studies for predicting protein subcellular locations should try to provide more than one subcellular location in order to make the prediction more reliable. However, only a handful of recently developed predictors can provide more than one subcellular location for a given protein. These predictors are termed 'multisite predictors'.

In this special report, the authors focus on the problem of developing multisite predictors. The authors first briefly introduce the progress in this special topic, and then raise three main challenges that all multisite predictors face. Finally, state-of-the-art methods and the present list of available programs that are capable of predicting multisite subcellular locations, and the future of this topic, are discussed.

Early explorations for multisite prediction

As the authors have mentioned before, more than 30% of known proteins have been assigned more than one subcellular location. The nature of localizing a protein within the cell is not usually to a unique location, but commonly to multiple locations. Thus, the computational methods must consider this situation in their models to represent the biological facts.

An innovative study for predicting more than one subcellular location for proteins was carried out on the yeast proteome dataset in 2005 [13,14]. A modified version of GO-FunD-PseAAC predictor was applied for this purpose [15]. Twenty two different locations were considered in that study. Reviewing this study, considering current opinion of the problem, the authors believe that this study met the challenges they are going to discuss when planning to incorporate multisite prediction ability. Although there is no program released with this study, this work actually expanded the horizon of predicting protein subcellular locations to a different level.

When the protein subnuclear locations were predicted, there is another exploration study considering multisite prediction. The authors have discussed this study in their previous review [16]. Ninety two sequences with multisite subnuclear locations were applied as the testing dataset in predicting protein subnuclear locations [17]. Although this study did not make predictions for multisite subcellular locations, it is the first study that took the multisite subcellular locations into consideration.

The challenges for multisite prediction of protein subcellular locations

Challenge I: the informative representations of proteins

It is always a big challenge to represent the protein effectively in a discrete form, not only for the case of classic single-site predictions, but also for multisite cases. In the studies that considered only single-site prediction, the pseudo-amino acid composition is a commonly used method for protein representation. The pseudo-amino acid composition was proposed in 2001, when various protein attributes were predicted [18]. This method has been applied in almost every branch of computational protein science [19]. Recently, Chou proposed the concept of general form pseudo-amino acid compositions as a summarization of the relevant developments in the last decade [20].

The GO-based representation, which is a mode of general form pseudo-amino acid compositions, was widely applied in predicting multisite protein subcellular locations. Let n be the total number of GO numbers that are considered in a study. Every given protein is represented as follows:

$$P_{GO} = [\Delta_1, \Delta_2, ..., \Delta_n]^T \tag{1}$$

where Δ_i, i = 1, 2 ,...,n, can be defined in different ways.

For example, in Euk-mPLoc2 [21], every protein sequence was searched against the entire UniProtKB/SwissProt database using BLAST [22]. The sequences with similarity scoring higher than a threshold were collected to form a homology set. Let N be the total number of protein sequences in the homology set of a given protein, and δ (i, k) an indicator function as follows:

$$\delta(i,k) = \begin{cases} 1 & \text{if the k-th protein in the homology set hits the} \\ & \text{i-th GO number} \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

Δ_i was defined as:

第i个GO注释和N个蛋白

$$\Delta_i = \delta(i,1) \vee \delta(i,2) \vee \ldots \vee \delta(i,N) \qquad (3)$$

where $\vee$ is the disjunction operator in logical algebra. P_{GO} is a binary vector with the definition in Equation 3, of which the i-th dimension indicates whether a hit can be found against the i-th GO number for any of the protein in the homology set. However, the importance of i-th GO number cannot be represented using only a binary form. Therefore, in iLoc-Animal [23] and iLoc-Euk [24], an improved Δ_i was proposed as follows:

$$\Delta_i = \frac{1}{N}\sum_{k=1}^{N}\delta(i,k) \qquad (4)$$

This definition incorporates the enrichment information of a certain GO number in the homology set of a given protein, which can be used as the importance of GO number.

When the GO numbers are used in predicting the protein subcellular locations, there is always a concern that the application of these methods could be restricted by the availability of the GO annotations. The authors of iLoc-Animal carried out an analysis on how the availability of GO numbers would affect the prediction performance. According to their reports, the prediction accuracies on the testing datasets with and without GO annotations are almost the same in iLoc-Animal [23]. Based on their analysis, they emphasized two facts as remarks of their work. First, the GO-based predictors, once established, would not need any further input such as GO terms or exact ID numbers in the databases. The requirement of input is only the sequence, which is identical to those non-GO-predictors. Second, the superior performance of GO-based predictors was not an overestimated result. It is because the proteins would be clustered better with GO-based representations than others [23]. In our opinion, the concern of using the GO-based features can now be eliminated.

？？？
有问题

Another commonly used representation method relies on the sequential evolution information that resides in the positional specific scoring matrix (PSSM), which could be extracted by iteratively searching the given protein against the UniProtKB/Swiss-Prot database with PSI-BLAST [25]. A PSSM is a matrix like Equation 5.

第i个AA突变成第j个AA概率

$$\mathrm{PSSM}(P) = \begin{bmatrix} E_{1\to1} & E_{1\to2} & \ldots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \ldots & E_{2\to20} \\ \vdots & \vdots & \ddots & \vdots \\ E_{L\to1} & E_{L\to1} & \ldots & E_{L\to20} \end{bmatrix} \qquad (5)$$

In the above matrix, the element $E_{i\to j}$ represents the propensity of the i-th amino acid on the sequence P being mutated to the j-th type of amino acid in the evolutionary process. L is the length of the protein P. Because the score values of different rows in PSSM can vary in a large range, a row-wise normalization procedure would always be applied to the PSSM [19]. The normalized PSSM can be represented as:

标准化的矩阵

$$\mathrm{PS\tilde{S}M}(P) = \begin{bmatrix} \tilde{E}_{1\to1} & \tilde{E}_{1\to2} & \ldots & \tilde{E}_{1\to20} \\ \tilde{E}_{2\to1} & \tilde{E}_{2\to2} & \ldots & \tilde{E}_{2\to20} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{E}_{L\to1} & \tilde{E}_{L\to1} & \ldots & \tilde{E}_{L\to20} \end{bmatrix} \qquad (6)$$

where

$$\tilde{E}_{i\to j} = \frac{1}{D_i}\left(E_{i\to j} - \bar{E}_i\right) \qquad (7)$$

$$\bar{E}_i = \frac{1}{20}\sum_{j=1}^{20} E_{i\to j} \qquad (8)$$

$$D_i = \sqrt{\frac{1}{20}\sum_{j=1}^{20}\left(E_{i\to j} - \bar{E}_i\right)^2} \qquad (9)$$

应该说是PSSM的目标：同时表示序列信息和进化信息

With the normalized PSSM, there are several different methods for representing a protein. All these methods tried to represent both the sequence information and the evolutionary information at the same time. For example, Euk-mPLoc2 applied the Pseudo-PSSM (PsePSSM), which is a combination of pseudo-amino acid compositions and the PSSM. According to the authors, PsePSSM can represent not only the sequence information, but also the evolutionary history of the sequence [21,26]. iLoc-Euk introduced the self-correlation matrix of PSSM, which has similar implications to PsePSSM, but is much easier to compute [24]. iLoc-Animal took advantage of grey-system theory, where the evolution history of a protein is modeled as a dynamic system [23].

The development of systems biology enabled the application of network-based features. Lee *et al.* carried out a study in fusing network-based descriptors with the sequence information [27], in which the protein network topology was used to fuse different sequences. Their method has been proven to have much better performance than the sequence-based methods in three species: human, yeast and fly. The results of this study, as well as other similar works, have clearly demonstrated that the network information can improve the prediction performance significantly [27–32]. Several other representations have also been applied in predicting multisite subcellular locations, like protein domain compositions [33], subcellular location relationship [34] and many other forms [35–38].

模型性能评估

Challenge II: performance measures

The measure of the prediction performance is another challenge in developing multisite protein subcellular location predictors. In the studies for predicting single-site subcellular locations, several performance measures, including sensitivity, specificity, positive-predictive value (PPV) and Matthew's correlation coefficient have been commonly applied to indicate the prediction performance on a given benchmarking dataset. These performance measures can be defined as the Equations 10–13.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{11}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{TP}} \tag{12}$$

$$\text{MCC} = \frac{\text{TPTN - FPFN}}{\sqrt{\text{TP} + \text{FPTP} + \text{FNTN} + \text{FPTN} + \text{FN}}} \tag{13}$$

In the above definitions, TP, FP, TN and FN are numbers of true positives, false positives, true negatives and false negatives, respectively. The definitions of these four numbers can be found in Figure 1. In practice, these numbers can be obtained in three different ways: the leave-one-out crossvalidation, which is also known as the jackknife test; the n-fold crossvalidation; and the independent dataset test.

Particularly, the definition of specificity is not consistent in different literatures [39,40]. The PPV may be called specificity in some cases [40]. In other studies, the term precision and recall would be used. Actually, precision is identical to our PPV and recall is identical to our sensitivity [101]. In most cases of predicting protein subcellular locations, defining a global specificity would be difficult. Thus, the F1-score, which is the harmonic mean of precision and recall, has been widely applied [41], as it uses both the precision and recall and thus avoids calculation of the specificity.

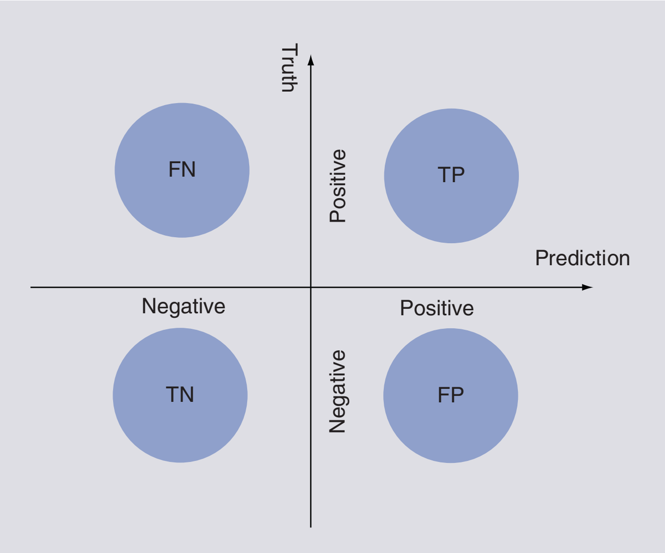

Figure 1. Definition of the basic statistics in performance measures. For a two-class classification problem, there is always a positive class and a negative class. When the prediction and the truth are both positive, the prediction result is a TP. When the prediction and the truth are both negative, the prediction result is a TN. When the prediction and the truth are not identical, the prediction result can be either FP or FN, depending on the whether the prediction result is positive or negative.
FN: False negative; FP: False positive; TN: True negative; TP: True positive.

When multisite predictions are first considered, an intuitive solution is to count the TP, FP, TN and FN numbers for the multisite prediction results. The performance measures for the single-site cases could be easily applied again. However, to count these numbers, a more basic problem must be solved in the first place. The problem is whether a prediction is correct or incorrect if only part of the predicted subcellular locations are correct. For example, given a protein P in the benchmarking dataset, its true subcellular location includes three locations. Although the prediction results also include three locations, only two of them are included in the three true locations. In this case, should this protein P be counted as correctly predicted ones or incorrectly predicted? In practice, there are two different methods to solve this problem. One method introduced the concept of locative proteins or virtual proteins. The other method introduced the measure of top-k accuracy rather than using single-site performance measures.

The concept of locative proteins or virtual proteins was proposed in several multisite predictors in the last few years. A locative protein is actually a two-tuple containing a protein and a location. When a protein in the benchmarking dataset has more than one subcellular location, it will be split into several locative proteins or virtual proteins, so that the methods for dealing with the single-site predictions will be easily transferred to the locative dataset [42]. For example, the protein P, which the authors have mentioned before, would be split into three locative proteins in the locative dataset. With this concept, the performance measure would be identical to those in the single-site cases. The performance comparisons between multisite predictors and the single-site predictors also become easy if the single-site predictors can be tested on the same locative dataset.

The top-k accuracy is a much lazier way to calculate the performance. It is defined as the fraction of correctly predicted samples, in which the prediction is considered correct if at least one of the known locations is included in the k predicted locations. This is equal to saying all partially correct predictions are counted as the correctly predicted one. This performance measure was applied in the earliest study concerning multisite locations [14,27,33] and the exploration to incorporate the protein interaction network information into subcellular location predictions [27]. Recently, this measure was also applied in predicting subcellular locations for topology-specific proteins [43].

Besides the above intuitive solutions, systematic performance measures for multisite prediction have been introduced from the machine learning field. In the machine learning field, the multisites prediction problem is under the topic of multilabel classifications. This topic has been studied in the machine learning field for over 10 years and is becoming more and more popular in recent years along with the relevant biological applications [44,45].

In practice, several different performance measures for multilabel classifications have been introduced, including the Hamming loss, multisite accuracy, multisite sensitivity, multisite PPV and absolute true rate. The definitions of these measures can be found in Equations 14–18.

$$H = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{M}\left[\,|L_p(j)\cup L_T(j)| - |L_p(j)\cap L_T(j)|\,\right] \qquad (14)$$

$$\mathrm{Acc} = \frac{1}{N}\sum_{j=1}^{N}\frac{|L_p(j)\cap L_T(j)|}{|L_p(j)\cup L_T(j)|} \qquad (15)$$

$$\mathrm{Sensitivity} = \frac{1}{N}\sum_{j=1}^{N}\frac{|L_p(j)\cap L_T(j)|}{|L_T(j)|} \qquad (16)$$

$$\mathrm{PPV} = \frac{1}{N}\sum_{j=1}^{N}\frac{|L_p(j)\cap L_T(j)|}{|L_p(j)|} \qquad (17)$$

$$\mathrm{ATR} = \frac{1}{N}\sum_{j=1}^{N}\delta(L_p(j), L_T(j)) \qquad (18)$$

In the above equations, H is the Hamming loss. It describes the average error rate of every prediction. The error here includes both the over predicted and the under predicted locations. M is the total number of subcellular locations that were considered in the study. N is the total number of proteins in the benchmarking dataset. $L_p(j)$ is a set that contains the predicted subcellular locations of the j-th protein in the benchmarking dataset, while L_T (j) is the set that contains the true subcellular locations of the j-th protein. The operator |.|, which can be applied on a set, is to calculate the total number of elements in that set. The function δ ($L_p(j)$, $L_T(j)$) is the identical indicator function of two sets, which is defined as Equation 19.

$$\delta(L_p(j), L_T(j)) = \begin{cases} 1 & L_p(j) = L_T(j) \\ 0 & L_p(j) \neq L_T(j) \end{cases} \qquad (19)$$

All the above performance measures should not be applied to a particular subcellular location. As some studies have pointed out, it is meaningless and misleading to measure the prediction performance for a particular subcellular location with these measures [45,46]. In practice, it would be very difficult to define $L_p(j)$ and $L_T(j)$, if only one particular subcellular location is considered in the multisite cases. Among the above performance measures, absolute true rate is the most stringent measure. It has been applied in measuring the performance of several recently developed multisite predictors [23,24]. 绝对真实率是最真实的度量

Actually, the mathematical essence of the above measures is just to expand the single-site measures by converting the concept of predicting a single location to a set of locations. When a protein P is given, we should think its subcellular location is a set of locations rather than a single location. The task of a multisite predictor is to estimate the content of this set. The TP, TN, FP and FN can be naturally defined for this single protein. As shown in Figure 2, the set of predicted locations $L_p(j)$ was represented as the red circle, while the set of true locations $L_T(j)$ was denoted as the blue circle. The yellow box contains all possible M locations. The number of elements could be used as TP in the A area, as FP in the B area, as FN in the C area and as TN in the D area. With these definitions, it is easy to see that the above definitions of multisite performance measures (Acc, PPV and sensitivity) are identical to the single-site performance measures.

Challenge III: determining the number of subcellular locations

预测到底有多少个位点的问题

Compared to the single-site predictions, a new problem that needs to be solved is how many subcellular locations should be reported for a given protein sequence. In the existing studies, there are several different ways to obtain this number. Most existing algorithms would give scores that can represent the membership degree of every given protein to every subcellular location. For example, the score for the protein P_i in subcellular location L_k can be denoted as $S(P_i,L_k)$. When we have N proteins and M subcellular locations, this will create an M×N matrix. The number of the subcellular locations can be obtained by analyzing this matrix. Here, the authors briefly introduce the basic idea of two methods for determining the number of subcellular locations.

One of these methods is called the top-k method. This method makes the number of subcellular locations as a parameter of the algorithm. The users of the algorithm are required to input the number of subcellular locations, which they expected, along with their protein sequences. The predictor will then select the subcellular locations according to this number and the values in the score matrix. For example, given a protein P, if the number k was entered as the expected number of its subcellular locations, the algorithm would choose the max k scores from the scores $S(P, L_1)$, $S(P, L_2)$, ..., $S(P, L_M)$ and output the k corresponding locations. The top-k method actually leaves the risk of errors partially to the users. For a newly sequenced protein, the number of all possible subcellular locations is usually something that is even more difficult to obtain than a subcellular location itself. Nevertheless, this method can be the easiest and simplest one that could easily extend the single-site algorithms to multisite ones [42,47]. As the authors have just described, there is an obvious issue with the top-k method. When the parameter k was entered, this method would predict every protein to have the same number of subcellular locations, which is unlikely to happen for all proteins in a test set unless these proteins are selected for this purpose.

The other more commonly used method is called the cutoff value-based method. When the M×N matrix of scores $S(P_i, L_k)$ is available, a threshold parameter C is given. For every protein P_i, the set of locations $L(P) = \{L_j|S(P_i,L_j)>C\}$ will be predicted.

Since the cutoff value can be determined when the predictor is developed, the users would not be required to enter the cutoff

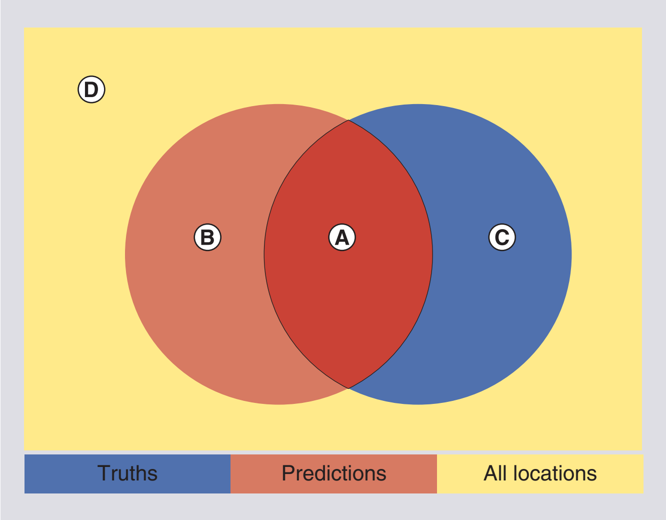

Figure 2. The interpretation of multisite performance measures. The yellow box contains all locations in a study. The red circle contains the predicted locations of a protein. The blue circle contains the true locations of a protein. **(A)**, **(B)**, **(C)** and **(D)** can be treated as true positives, false positives, false negatives and true negatives, respectively. Under these interpretations, the performance measures of the multisite predictors are the same as those of single-site predictors.

value when they were using the predictor. The user experience of this method is much better than the top-k method, because they can simply input the sequences and get different numbers of predicted subcellular locations for every protein. For the users, it seems that the predictor ‘knows’ how many subcellular locations should be predicted. However, the cutoff method still has two problems. First, in some cases, the scores for different proteins would vary in a large range, so that the cutoff value will generate a null result for some proteins. For example, given proteins P_1 and P_2, the max value of $S(P_1,L_k)$ ($k = 1,2,\dots,M$) may be 120, while the max value of $S(P_2,L_k)$ could be only 1.2 or even smaller. When the cutoff value was set to 100, the prediction results for protein P_2 would be an empty set, as no score is larger than the cutoff value. Intuitively, this could be solved by applying a normalization procedure to the scores of each protein. In practice, a smarter way was actually taken. In several studies, the subcellular locations of a given protein P were determined by Equation 20:

$$L(P)=\{L_k \mid S(P, L_k) > S_{max} - C(S_{max} - S_{min})\} \tag{20}$$

where

$$S_{max}(P) = \max\{S(P, L_k), k = 1,2,\dots,M\} \tag{21}$$

$$S_{min}(P)=\min\{S(P,L_k), k = 1,2,\dots,M\} \tag{22}$$

and C is a real number in the range (0,1).

In practice, C can be chosen directly, for example 0.2 or 0.4. These values can make the predictor work, but are usually unable to optimize the prediction accuracy. To maximize the prediction accuracy, an optimization procedure should be carried out to find a reasonable and better C. The optimized cutoff value C was usually achieved by scanning the cutoff value in a range with small steps to minimize the Hamming loss in the jackknife test. For example, we let C be selected in a set like [0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8]. For every value in this set, the Hamming loss of the jackknife test was calculated. The value that can generate the minimal Hamming loss would be used as the cutoff value in the predictor.

The cutoff value method is a very promising choice for developing the multisite predictors as long as the M×N score matrix is available. The cutoff method is useful not only because it can extend the single-site algorithms easily to multisite versions, but also because the user interface of the single-site predictors would not be changed after they were upgraded to multisite versions. This method has been commonly applied in the most recent development of multisite subcellular location predictors [9,23,24,48–51].

There are two other simple methods for determining the number of subcellular locations in multisite predictors. In the mGOASVM [52], the prediction was the union of M SVMs. For every subcellular location, an SVM was trained against the rest all locations. When a protein is given, the locations where the SVM reports positive are collected as the final results. In iLoc-Hum [51], iLoc-Gneg [50], iLoc-Gpos [49], iLoc-Plant [48] and iLoc-Virus [9], with the application of the MultiLabel K-nearest neighbor (ML-KNN) algorithm, the number of subcellular locations could be directly transferred from the nearest protein in the training dataset to the given testing protein.

Expert commentary

Currently, there are several existing methods that can predict multisite protein subcellular locations. The iLoc serials predictor, which includes iLoc-Hum, iLoc-Euk, iLoc-Animal, iLoc-Gneg, iLoc-Gpos, iLoc-Plant and iLoc-Virus, can predict multisite subcellular locations using ML-KNN algorithm and general form pseudo-amino acid compositions. The Cell-PLoc package [53], which includes Hum-mPLoc2 [54], Euk-mPLoc2 [21], Gneg-mPLoc [55], Gpos-mPLoc [56], Plant-mPLoc [57] and Virus-mPLoc [58], was established based on the ensemble classifiers and general form pseudo-amino acid compositions. YLoc [59], which provides the function of interpreting prediction results, was established based on Bayes inferences. As a summary of state-of-the-art methods, the authors present a list of recently developed multisite predictors in Table 1. This list contains 21 predictors that could tackle the multisite subcellular location prediction problem under different conditions. They are designed for different organisms, different subcellular compartments and different software platforms.

These methods have been designed, implemented and calibrated using different benchmarking datasets. Establishing a high-quality benchmarking dataset in itself is of fundamental importance in predicting multisite protein subcellular locations.

Table 1. A list of available multisite protein subcellular location predictors.

Name	URL	Availability	Dataset†	Organisms	Features	Learning method	Number determination	Performance measure	Ref.
MLMK-TLM	http://soft.synu.edu.cn/upload/msy.rar	RAR package	3106, 14, 25%	Human	GO-based features	MK-SVM	Cutoff value	Locative measures	[35]
mGOASVM	http://bioinfo.eie.polyu.edu.hk/mGoaSvmServer/mGOASVM.html	Web access	207, 6, 25% (virus); 978, 12, 25% (plants)	Virus or plants	GO-based features	ML-SVM	Union of SVM results	Locative measures	[52]
iLoc-Animal	www.jci-bioinfo.cn/iLoc-Animal	Web access	5048, 20, 25%	Metazoan except human	General form PseAAC	ML-KNN	Cutoff value	Multisite measures	[23]
iLoc-Euk	www.jci-bioinfo.cn/iLoc-Euk	Web access	7776, 22, 25%	Eukaryotes	General form PseAAC	ML-KNN	Cutoff value	Multisite measures	[24]
iLoc-Gpos	www.jci-bioinfo.cn/iLoc-Gpos	Web access	519, 4, 25%	Gram-positive bacteria	General form PseAAC	ML-KNN	Direct transfer	Locative measures; ATR	[49]
iLoc-Gneg	www.jci-bioinfo.cn/iLoc-Gneg	Web access	1392, 8, 25%	Gram-negative bacteria	General form PseAAC	ML-KNN	Direct transfer	ATR	[50]
iLoc-Hum	www.jci-bioinfo.cn/iLoc-Hum	Web access	3106, 14, 25%	Human	General form PseAAC	ML-KNN	Direct transfer	ATR	[51]
iLoc-Plant	www.jci-bioinfo.cn/iLoc-Plant	Web access	978, 12, 25%	Plants	General form PseAAC	ML-KNN	Direct transfer	ATR	[48]
iLoc-Virus	www.jci-bioinfo.cn/iLoc-Virus	Web access	207, 6, 25%	Virus	General form PseAAC	ML-KNN	Direct transfer	ATR	[9]
Euk-mPLoc2	www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/	Web access	7766, 22, 25%	Eukaryotes	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[21]
Euk-mPLoc	www.csbio.sjtu.edu.cn/bioinf/euk-multi/	Web access	5618, 22, 25%	Eukaryotes	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[42]
Hum-mPLoc	www.csbio.sjtu.edu.cn/bioinf/hum-multi/	Web access	2750, 14, 25%	Human	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[47]
Hum-mPLoc2	www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/	Web access	3106, 14, 25%	Human	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[54]
Plant-mPLoc	www.csbio.sjtu.edu.cn/bioinf/plant-multi/	Web access	978, 12, 25%	Plants	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[57]

†The dataset description usually contain three numbers: X, Y, Z%. X is the number of all proteins in the dataset. Y is the number of all subcellular locations. Z% is the similarity threshold of dataset. NA means the relevant data are not available in the original literature. The description in the brackets indicates a benchmarking dataset provided by other subcellular predictors, which may be for single-site predictions.

ATR: Absolute true rate; GO: Gene ontology; MK-SVM: Multikernel support vector machine; ML-KNN: MultiLabel K-nearest neighbour; OET-KNN: Optimized-evidence-theoretical k-nearest neighbour; PseAAC: Pseudo-amino acid composition; SVM: Support vector machine.

Table 1. A list of available multisite protein subcellular location predictors (cont.).

Name	URL	Availability	Dataset[†]	Organisms	Features	Learning method	Number determination	Performance measure	Ref.
Gneg-mPLoc	www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/	Web access	1392, 8, 25%	Gram-negative bacteria	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[55]
Gpos-mPLoc	www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/	Web access	519, 4, 25%	Gram-positive bacteria	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[56]
Virus-mPLoc	www.csbio.sjtu.edu.cn/bioinf/virus-multi/	Web access	207, 6, 25%	Virus	General form PseAAC	OET-KNN; ensemble classifier	Cutoff value	Locative measures; ATR	[58]
ML-PLoc	www.csbio.sjtu.edu.cn/bioinf/ML-PLoc/	ZIP package	5909, 14, 80%	Human	General form PseAAC	SVM; ensemble classifier	Cutoff value	Locative measures; ATR	[67]
PSLT	www.mcb.mcgill.ca/~hera/PSLT/	Web access	2216, NA (human); 1612, NA (yeast); 2095, 9, NA (mouse)	Human, yeast, mouse	Protein motif frequency	Bayes inference	Top-k	Top-k	[33]
YLoc	http://abi.inf.uni-tuebingen.de/Services/YLoc	Web access	(Bacello); (Höglund); (DBMLoc)	Animal, fungal, plant (Bacello); eukaryotes (Höglund); not specific (DBMLoc)	Hybrid features	Bayes inference	Cutoff value	Locative measures; multisite measures; F1-score	[59]
KnowPred	http://bio-cluster.iis.sinica.edu.tw/kbloc/introduce.html	Web access	28056, 10, NA (ngLOC)	Eukaryotes, no plants	Peptide segments	Knowledge-based inference	Top-k	Top-k	[62]

[†] The dataset description usually contain three numbers: X, Y, Z%. X is the number of all proteins in the dataset. Y is the number of all subcellular locations. Z% is the similarity threshold of dataset. NA means the relevant data are not available in the original literature. The description in the brackets indicates a benchmarking dataset provided by other subcellular predictors, which may be for single-site predictions.

ATR: Absolute true rate; GO: Gene ontology; MK-SVM: Multikernel support vector machine; ML-KNN: MultiLabel K-nearest neighbour; OET-KNN: Optimized-evidence-theoretical k-nearest neighbour; PseAAC: Pseudo-amino acid composition; SVM: Support vector machine.

Here, 'high quality' means the dataset is both diverse enough and sufficient in size. The diversity of a dataset is measured by the maximal sequence similarity in the dataset. In practice, most of the studies used a culling program, such as CD-HIT [60] and PISCES [61], to control the sequence similarity level to be less than 25%, which can result in a very limited number of sequences for some organisms. For example, the dataset for establishing iLoc-Virus and Virus-mPLoc contains only 207 proteins. In machine learning, a dataset that is sufficient in size is important to represent the actual data, and hence important to train an effective model. Since most of the existing methods establish their dataset from the UniProtKB/SwissProt database, it may be very difficult to obtain more proteins with all the above conditions. However, if the original data sources are not limited to the UniProtKB/SwissProt database, there could be more sequences that can be applied. The KnowPred method has tried to apply the sequences from nonredundant dataset from NCBI [62], which may be a new and better data source for future studies.

Another fact we can see from Table 1 is that most of the existing predictors are designed for a limited range of organisms. iLoc-Euk and Euk-mPLoc2 covers the eukaryotes branch, which should be the largest range that is covered currently. For a given organism, there may be more than one predictor that can be applied. For example, if we have a protein from some kind of plant, iLoc-Plant, iLoc-Euk, Euk-mPLoc2 and Plant-mPLoc can be applied at the same time. If the prediction results from these four predictors are not identical, there will be a need to combine these results. Therefore, a new type of method, which is known as the meta-predictors, was recently developed [63]. By combining the results of existing predictors, the meta-methods can provide more accurate and more comprehensive predictions.

元预测器：多个预测器的结合使用

Since there are more and more proteins that are discovered to have more than one subcellular location, it is natural to ask the question: can the proteins localize to every combination of possible subcellular locations? If so, can the predictors output all types of the combinations? By using the cutoff value-based method, the existing methods can actually assign every possible combination to a given protein, even if that combination has never been observed in the training dataset. This is just why the computational methods are thought to be useful in determining protein subcellular locations.

As a reminder, the users should choose the predictor according to their own requirements and the source of their data. We strongly recommend the users read the original literature before they apply the predictors in their studies. After all, applying a computational method without understanding its mechanism is no better than mixing reagents in a tube without knowing what they are.

Five-year view

The future development for this topic is still promising. There are several problems that should be solved in the next 5 years. First, the prediction of subcellular locations for topology-specific proteins should be considered. With the initial study of MemLoci [64], computationally determining the subcellular location of membrane proteins has become a hot topic in the last year [43]. The topology factor should also be considered in the multisite predictors. Second, although there are many methods that could predict the protein subcellular location with single-site or multisite results, why the protein should be predicted to the location is still a problem for most of the methods. In other words, an explanation of the prediction results would be more important. Several initial attempts to do this have been carried out in recent years, such as MemBrain [65] and YLoc [41,59]. Third, it is also necessary to interpret the mechanism and functional importance of multisite subcellular locations of a given protein. In computational terms, this may be achievable by incorporating the protein interaction networks. A protein may interact with different groups of partners when it is localized to different subcellular compartments. Therefore, the multisite subcellular locations may imply that a protein can be involved in different biological processes. Furthermore, since the protein structures are related to their functions, the mechanism of localizing a protein to multisite subcellular locations may be buried in the structural level, especially for different folding patterns or different isoforms of a protein. As the computational prediction of protein structures is never an easy problem, may be the multisite subcellular locations could be helpful in determining protein structures. Fourth, the efficiency of the prediction is still a problem. For several existing predictors, the efficiency for dealing with a large dataset should be improved. This is largely the problem of the implementation technology. Recently, the PseAAC-Builder, which is a stand-alone program, has significantly improved the computational efficiency of the pseudo-amino acid composition [66]. The technology of implementing the PseAAC-Builder could be transferred to other programs. Finally, choosing a proper program in a practical study is usually a big problem. Is it possible to develop an integrated and unified system that could help the investigators to choose the program automatically? With the current development of cloud computation and its application in bioinformatics and computational biology, we can expect a unified analysis platform that could save the users' time in choosing the right program for their study.

Financial & competing interests disclosure

The authors were supported by National Science Foundation of China (NSFC 61005041), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20100032120039), Tianjin Natural Science Foundation (No. 12JCQNJC02300), China Postdoctoral Science Foundation (2012T50240 and 2013M530114) and the Seed Foundation of Tianjin University (No. 60302006 and 60302024). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Key issues

- Most of the known proteins should have multiple subcellular locations. However, most of the existing predictors can only assign a single location to a given protein.
- There are several existing predictors concerning the multisite protein subcellular location.
- There are three main challenges that the multisite protein subcellular location predictors must face: the informative feature extraction, the performance measures and how the number of subcellular locations should be determined.
- Several solutions have been proposed to solve these problems. However, there is still room for further improvement.

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1 Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* 26(9), 2230–2236 (1998).

2 Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.* 276(2), 517–525 (1998).

3 Cedano J, Aloy P, Pérez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* 266(3), 594–600 (1997).

4 Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24(1), 34–36 (1999).

5 Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17(8), 721–728 (2001).

6 Chou KC, Elrod DW. Protein subcellular location prediction. *Protein Eng.* 12(2), 107–118 (1999).

7 Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* 20(1), 21–28 (2004).

8 Shen HB, Chou KC. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem. Biophys. Res. Commun.* 337(3), 752–756 (2005).

9 Xiao X, Wu ZC, Chou KC. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.* 284(1), 42–51 (2011).

10 Chou KC, Shen HB. Recent progress in protein subcellular location prediction. *Anal. Biochem.* 370(1), 1–16 (2007).

11 Shen HB, Yang J, Chou KC. Methodology development for predicting subcellular localization and other attributes of proteins. *Expert Rev. Proteomics* 4(4), 453–463 (2007).

12 Zhang S, Xia X, Shen J, Zhou Y, Sun Z. DBMLoc: a database of proteins with multiple subcellular localizations. *BMC Bioinformatics* 9(1), 127 (2008).

13 Cai YD, Chou KC. Predicting 22 protein localizations in budding yeast. *Biochem. Biophys. Res. Commun.* 323(2), 425–428 (2004).

14 Chou KC, Cai YD. Predicting protein localization in budding yeast. *Bioinformatics* 21(7), 944–950 (2005).

•• **The first study for predicting multisite subcellular locations.**

15 Chou KC, Cai YD. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem. Biophys. Res. Commun.* 320(4), 1236–1239 (2004).

16 Du P, Li T, Wang X. Recent progress in predicting protein sub-subcellular locations. *Expert Rev. Proteomics* 8(3), 391–404 (2011).

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

18 Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43(3), 246–255 (2001).

19 Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics* 6(4), 262–274 (2009).

20 Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273(1), 236–247 (2011).

21 Chou K-C, Shen H-B. 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 (2010).

22 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 215(3), 403–410 (1990).

23 Lin WZ, Fang JA, Xiao X, Chou KC. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.* 9(4), 634–644 (2013).

• **State-of-the-art method for predicting multisite protein subcellular locations.**

24 Chou KC, Wu ZC, Xiao X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS ONE* 6(3), e18258 (2011).

25 Altschul SF, Madden TL, Schäffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17), 3389–3402 (1997).

26 Shen HB, Chou KC. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng. Des. Sel.* 20(11), 561–567 (2007).

27 Lee K, Chuang HY, Beyer A *et al.* Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res.* 36(20), e136 (2008).

•• **A representative study that uses protein interaction networks as well as the protein sequence information.**

28 Kumar G, Ranganathan S. Network analysis of human protein location. *BMC Bioinformatics* 11(Suppl. 7), S9 (2010).

29 Mintz-Oron S, Aharoni A, Ruppin E, Shlomi T. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics* 25(12), i247–i252 (2009).

30 Shin CJ, Wong S, Davis MJ, Ragan MA. Protein-protein interaction as a predictor of subcellular location. *BMC Syst. Biol.* 3, 28 (2009).

31 Hu LL, Feng KY, Cai YD, Chou KC. Using protein–protein interaction network information to predict the subcellular locations of proteins in budding yeast. *Protein Pept. Lett.* 19(6), 644–651 (2012).

32 Scott MS, Calafell SJ, Thomas DY, Hallett MT. Refining protein subcellular localization. *PLoS Comput. Biol.* 1(6), e66 (2005).

33 Scott MS, Thomas DY, Hallett MT. Predicting subcellular localization via protein motif co-occurrence. *Genome Res.* 14(10A), 1957–1966 (2004).

34 He J, Gu H, Liu W. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *PLoS ONE*. 7(6), e37155 (2012).

35 Mei S. Multi-label multi-kernel transfer learning for human protein subcellular localization. *PLoS ONE* 7(6), e37716 (2012).

36 Mei S. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.* 310, 80–87 (2012).

37 Li GZ, Wang X, Hu X, Liu JM, Zhao RW. Multilabel learning for protein subcellular location prediction. *IEEE Trans. Nanobioscience* 11(3), 237–243 (2012).

38 Xu Q, Pan SJ, Xue HH, Yang Q. Multitask learning for protein subcellular location prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8(3), 748–759 (2011).

39 Burset M, Guigó R. Evaluation of gene structure prediction programs. *Genomics* 34(3), 353–367 (1996).

40 Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16(5), 412–424 (2000).

41 Briesemeister S, Rahnenführer J, Kohlbacher O. Going from where to why – interpretable prediction of protein subcellular localization. *Bioinformatics* 26(9), 1232–1238 (2010).

•• **The first study that tries to interpret the prediction result.**

42 Chou KC, Shen HB. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* 6(5), 1728–1734 (2007).

43 Du P, Tian Y, Yan Y. Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. *J. Theor. Biol.* 313, 61–67 (2012).

44 Zhang M-L, Zhou Z-H. ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognition* 40(7), 2038–2048 (2007).

45 Tsoumakas G, Katakis I. Multi-label classification. *Int. J. Data Warehousing Mining.* 3(3), 1–13 (2007).

46 Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. In: *Data Mining and Knowledge Discovery Handbook.* Springer, NY, USA, 667–685 (2010).

47 Shen HB, Chou KC. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun.* 355(4), 1006–1011 (2007).

48 Wu ZC, Xiao X, Chou KC. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. Biosyst.* 7(12), 3287–3297 (2011).

49 Wu ZC, Xiao X, Chou KC. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins. *Protein Pept. Lett.* 19(1), 4–14 (2012).

50 Xiao X, Wu Z-C, Chou K-C. 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 (2011).

51 Chou KC, Wu ZC, Xiao X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* 8(2), 629–641 (2012).

52 Wan S, Mak MW, Kung SY. mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics* 13, 290 (2012).

53 Chou KC, Shen HB. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3(2), 153–162 (2008).

• **A collection of OET-KNN-based multisite protein subcellular locations.**

54 Shen HB, Chou KC. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.* 394(2), 269–274 (2009).

55 Shen HB, Chou KC. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J. Theor. Biol.* 264(2), 326–333 (2010).

56 Shen HB, Chou KC. Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. *Protein Pept. Lett.* 16(12), 1478–1484 (2009).

57 Chou KC, Shen HB. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS ONE* 5(6), e11335 (2010).

58 Shen HB, Chou KC. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J. Biomol. Struct. Dyn.* 28(2), 175–186 (2010).

59 Briesemeister S, Rahnenführer J, Kohlbacher O. YLoc – an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.* 38(Web Server issue), W497–W502 (2010).

60 Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23), 3150–3152 (2012).

61 Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics* 19(12), 1589–1591 (2003).

62 Lin HN, Chen CT, Sung TY, Ho SY, Hsu WL. Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics* 10(Suppl. 15), S8 (2009).

63 Magnus M, Pawlowski M, Bujnicki JM. MetaLocGramN: a meta-predictor of protein subcellular localization for Gram-negative bacteria. *Biochim. Biophys. Acta* 1824(12), 1425–1433 (2012).

64 Pierleoni A, Martelli PL, Casadio R. MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics* 27(9), 1224–1230 (2011).

65 Shen H, Chou JJ. MemBrain: improving the accuracy of predicting transmembrane helices. *PLoS ONE* 3(6), e2399 (2008).

66 Du P, Wang X, Xu C, Gao Y. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* 425(2), 117–119 (2012).

67 Zhu L, Yang J, Shen H-B. Multi label learning for prediction of human protein subcellular localizations. *Protein J.* 28(9–10), 384–390 (2009).

Website

101 Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf