Bioinformatics, 33(22), 2017, 3524–3531
doi: 10.1093/bioinformatics/btx476
Advance Access Publication Date: 24 July 2017
Original Paper

OXFORD

Sequence analysis

pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites

Xiang Cheng[1,2], Shu-Guang Zhao[1], Wei-Zhong Lin[2], Xuan Xiao[2,3,]* and Kuo-Chen Chou[3,4,5,]*

[1]College of Information Science and Technology, Donghua University, Shanghai, China, [2]Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, China, [3]The Gordon Life Science Institute, Boston, MA 02478, USA, and [4]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia and [5]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 18, 2017; revised on July 19, 2017; editorial decision on July 20, 2017; accepted on July 22, 2017

Abstract

Motivation: Cells are deemed the basic unit of life. However, many important functions of cells as well as their growth and reproduction are performed via the protein molecules located at their different organelles or locations. Facing explosive growth of protein sequences, we are challenged to develop fast and effective method to annotate their subcellular localization. However, this is by no means an easy task. Particularly, mounting evidences have indicated proteins have multi-label feature meaning that they may simultaneously exist at, or move between, two or more different subcellular location sites. Unfortunately, most of the existing computational methods can only be used to deal with the single-label proteins. Although the 'iLoc-Animal' predictor developed recently is quite powerful that can be used to deal with the animal proteins with multiple locations as well, its prediction quality needs to be improved, particularly in enhancing the absolute true rate and reducing the absolute false rate.

Results: Here we propose a new predictor called 'pLoc-mAnimal', which is superior to iLoc-Animal as shown by the compelling facts. When tested by the most rigorous cross-validation on the same high-quality benchmark dataset, the absolute true success rate achieved by the new predictor is 37% higher and the absolute false rate is four times lower in comparison with the state-of-the-art predictor.

Availability and implementation: To maximize the convenience of most experimental scientists, a user-friendly web-server for the new predictor has been established at http://www.jci-bioinfo.cn/pLoc-mAnimal/, by which users can easily get their desired results without the need to go through the complicated mathematics involved.

Contact: xxiao@gordonlifescience.org or kcchou@gordonlifescience.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Called by many as a 'building block of life', the cell contains many different protein molecules located at its different organelles or locations. It is through these proteins that the cell's growth and reproduction along with its many important functions are realized. Consequently, the importance of knowledge about their subcellular localization is self-evident. Unfortunately, there is a huge gap between the newly discovered protein sequences and their experiment-determined location sites. To reduce the gap and to timely use these new protein sequences for basic research and drug development, it is highly demanded to develop computation methods in this regard. In the last 25 years or so, many prediction methods were proposed to address this problem (see (Chou and Shen, 2007b; Nakai, 2000) as well as a long list of references cited in the two review articles).

But most of these prediction methods can only be used to deal with the so-called single-label system where each protein has one, and only one, subcellular location. With more experimental data available, however, it has been found that the distribution of proteins in a cell actually belongs to a multi-label system, in which some proteins may simultaneously occur in two or more different location sites. These proteins should not be ignored because they may have some exceptional biological functions (Glory and Murphy, 2007) worthy of our special notice.

About 10 years ago, considerable efforts have been made to study this kind of multi-label protein systems (Chou and Shen, 2007a, 2010a,b; Chou *et al.*, 2011, 2012; Dehzangi *et al.*, 2015; Huang and Yuan, 2013; Lin *et al.*, 2013; Mei, 2012; Pacharawongsakda and Theeramunkong, 2013; Shen and Chou, 2007a, 2009a,b, 2010a,b; Wang *et al.*, 2013; Wu *et al.*, 2011, 2012; Xiao *et al.*, 2011a). They can be basically categorized into two series (Chou, 2015): the 'PLoc' series (see, e.g. (Chou and Shen, 2007a, 2010b; Shen and Chou, 2009b, 2010b)) and 'iLoc' series (see, e.g. (Chou *et al.*, 2012; Lin *et al.*, 2013; Wu *et al.*, 2011; Xiao *et al.*, 2011a)).

pt?>Compared with the single-label systems, it is much more complicated and difficult to deal with the multi-label ones, particularly in achieving a descent 'absolute true' success rate (Chou, 2013). The score standard for the absolute true rate is very harsh. According to its definition, for a protein sample that actually simultaneously exists in the subcellular locations 'A, B and C'. If the predicted result is not exactly the three locations but 'A and B' or 'A, B, C and D', its score for the 'absolute true' rate is zero. In other words, when and only when the predicted result is perfectly identical to the actual situation, can its score be counted equal to 1. For instance, among the existing predictors, the iLoc-Animal (Lin *et al.*, 2013) is the most powerful one for predicting the subcellular location of animal proteins. But its reported absolute true success rate was only 45.62% (Lin *et al.*, 2013).

The present study was devoted to develop a new multi-label predictor that can remarkably improve the prediction quality for the subcellular localization of animal proteins, particularly in the absolute true success rate.

2 Materials and methods

According to the 5-step rule (Chou, 2011) and as done in a series of recent publications (Chen *et al.*, 2016b; Jia *et al.*, 2016b; Liu *et al.*, 2017a; Meher *et al.*, 2017; Qiu *et al.*, 2016), in reporting a new statistical prediction method, one should make the following five aspects very clear: (i) benchmark dataset, (ii) sample formulation, (iii) operation algorithm, (iv) cross-validation and (v) web-server establishment. The advantages of doing so are: (i) repeatability, i.e. easy for others to repeat the reported results; (ii) stimulativity, i.e. facilitating others to develop new prediction models in various relevant areas; and (iii) wide usage, i.e. being convenient for most experimental scientists to use the reported predictor. Below, we are to address the five steps one-by-one.

2.1 Benchmark dataset

In the current study, the benchmark dataset was constructed based on the one reported in Lin *et al.* (2013), where a total of 5048 animal protein sequences were categorized into 20 subsets according to their different subcellular locations confirmed by experiments. To enhance its quality and to reduce the redundancy and homology bias, the CD-HIT (Fu *et al.*, 2012) was adopted to remove those sequences from the original benchmark dataset (Lin *et al.*, 2013) that have ≥40% pairwise sequence identity to any other in a same subset.

After such a cut-off procedure, the total number of protein sequences was reduced to 3919. Their protein codes and detailed sequences are given in Supplementary Material S1.

An overall distribution of these proteins in the 20 subcellular locations is given in Supplementary Material S2, from which we can see that, of the 3919 different proteins, 2113 occur in one location, 1293 in two locations, 286 in three locations, 173 in four locations, 43 in five locations, 5 in six locations, 3 in seven locations, 3 in eight locations and none in nine and more locations. When studying multi-label systems, it is instructive to introduce the concept of multiplicity degree (Chou, 2013) as defined by

$$\mathrm{MD} = \frac{N(\mathrm{vir})}{N} = \frac{\sum_{k=1}^{N} n^{L}(k)}{N} \tag{1}$$

where MD is the abbreviation of 'multiplicity degree', N the total number of protein samples with different amino acid sequences, and $N(\mathrm{vir})$ is the total number of virtual protein samples investigated. The number of virtual proteins is calculated as follows: if a protein sample with two different labels (or located in two different subcellular locations), it will be counted as two virtual or 'locative' protein samples; if with three different labels, counted as three virtual samples; and so forth. Thus, the total number of virtual protein samples can be derived from the numerator of Eq. 1, where $n^{L}(k)$ is the number of different labels (or subcellular locations) marked on the k-th protein sample. As we can see from Eq. 1, MD = 1 means the system containing no protein belonging to more than one location, while MD > 1 means some proteins occurring in more than one location. The higher the value of MD, the more protein samples that have multiple labels. For instance, the multiplicity degree is 1 for most protein subcellular prediction methods without covering the multiplex proteins; it is 1.146 for Euk-mPLoc (Chou and Shen, 2010a) and iLoc-Euk (Chou *et al.*, 2011), 1.185 for Hum-mPLoc (Shen and Chou, 2009b) and iLoc-Hum (Chou *et al.*, 2012), and 1.079 for Plant-mPLoc (Chou and Shen, 2010b) and iLoc-Plant (Wu *et al.*, 2011).

A breakdown of the 3919 proteins according to their occurrences in 20 different subcellular localizations is given in Table 1, from which we can also see that the multiplicity degree of the current benchmark dataset is (6539/3919) = 1.669.

The new benchmark dataset thus obtained is denoted by $\mathbb{S}$, which can be further formulated as

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \cdots \cup \mathbb{S}_u \cup \cdots \cup \mathbb{S}_{19} \cup \mathbb{S}_{20} \tag{2}$$

where $\mathbb{S}_1$ only contains the protein samples from the 'acrosome' location (cf. Table 1), $\mathbb{S}_2$ only those from the 'cell cortex' location, and so forth; ∪ denotes the symbol for 'union' in the set theory.

2.2 Proteins sample formulation

For a sequence-known protein **P**, its most general expression is

$$\mathrm{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{3}$$

where L denotes its length, R_1 is the 1st residue, R_2 the 2nd residue,

Table 1. Breakdown of the proteins in the benchmark dataset into 20 subsets according to their different subcellular localizations (cf. Supplementary Material S1 and Supplementary Material S2)

Subset	Subcellular location name	Number of proteins
$\mathbb{S}_1$	Acrosome	26
$\mathbb{S}_2$	Cell cortex	41
$\mathbb{S}_3$	Cell membrane	884
$\mathbb{S}_4$	Centriole	22
$\mathbb{S}_5$	Centrosome	86
$\mathbb{S}_6$	Cytoplasm	1283
$\mathbb{S}_7$	Cytoskeleton	310
$\mathbb{S}_8$	Endoplasmic reticulum	455
$\mathbb{S}_9$	Endosome	142
$\mathbb{S}_{10}$	Extracellular space	97
$\mathbb{S}_{11}$	Golgi apparatus	317
$\mathbb{S}_{12}$	Lysosome	114
$\mathbb{S}_{13}$	Melanosome	10
$\mathbb{S}_{14}$	Microsome	57
$\mathbb{S}_{15}$	Mitochondrion	514
$\mathbb{S}_{16}$	Nucleus	1064
$\mathbb{S}_{17}$	Peroxisome	64
$\mathbb{S}_{18}$	Plasma membrane	884
$\mathbb{S}_{19}$	Spindle	103
$\mathbb{S}_{20}$	Synapse	66
Total number of virtual proteins N(vir)[a]		6539
Total number of proteins with different sequences N		3919
The multiplicity degree MD[b]		1.669

[a] See the numerator of Eq. 1 and the relevant text for the definition of virtual proteins.

[b] See Eq. 1 for the definition of multiplicity degree.

R_3 the 3rd residue and so forth. Since all the existing machine-learning algorithms such as SVM (Chen *et al.*, 2016a) and Random Forest (Jia *et al.*, 2016a) can only handle vectors but not sequences (Chou, 2015), we have to convert Eq. 3 into a vector. Unfortunately, a vector defined in a discrete model might completely lose all the sequence-pattern information. To overcome this problem, the PseAAC (Pseudo Amino Acid Composition) (Chou, 2001) was proposed in 2001. Ever since then, the concept of PseAAC has been rapidly used in nearly all the areas of computational proteomics (Chou, 2009) and many fields of genome analysis (see, e.g. (Chen *et al.*, 2015) as well as a long list of references cited in (Chou, 2017; Liu *et al.*, 2017b)). According to the concept of general PseAAC (Chou, 2011), any protein sequence can be formulated as a PseAAC vector given by

$$\mathrm{P} = \left[\Psi_1\ \ \Psi_2\ \ \cdots\ \ \Psi_\mathrm{u}\ \ \cdots\ \ \Psi_\Omega\right]^\mathrm{T} \tag{4}$$

where $\mathbf{T}$ is a transpose operator, while the subscript Ω is an integer parameter and its value as well as the components Ψ_u $(u = 1,\ 2,\ \cdots,\ \Omega)$ will depend on how to extract the desired information from the amino acid sequence of $\mathbf{P}$, as elaborated below.

The information of gene ontology (GO) has been widely used to enhance the prediction quality of predicting protein subcellular localization (see, e.g. (Chou and Cai, 2003; Shen and Chou, 2007a; Wan *et al.*, 2013; Wu *et al.*, 2011; Xiao *et al.*, 2011a,b)). The advantage of using the GO approach is that proteins mapped into the GO space (instead of Euclidean space or any other simple geometric space) would be clustered in a manner much better for studying their subcellular locations, as elaborated in Chou and Shen (2008). For the rationale or justification of using the GO approach to predict the protein subcellular localization, see an incisive discussion or analysis given in Section VI of a comprehensive review paper (Chou, 2013).

However, the existing GO approaches (see, e.g. (Chou and Cai, 2003; Shen and Chou, 2007a; Wan *et al.*, 2013; Wu *et al.*, 2011; Xiao *et al.*, 2011a,b)) have the following shortcomings. (i) Only the digital numbers 0 and 1 (or their simple combination) were used to incorporate the GO information, and hence some very useful information was missed. (ii) The dimension of the protein vectors, namely Ω of Eq. 3, in the previous GO approaches was very high; e.g. it was 1930 in Chou and Cai (2003), 3043 in Lin *et al.* (2013) and 9567 in Chou and Shen (2006), and hence was prone to lead to the high-dimension disaster problem (Wang *et al.*, 2008).

Here, we are to propose a novel GO approach, by which not only the dimension of protein vectors can be significantly reduced, but the GO information incorporated can also be significantly optimized. The detailed procedures are as follows.

Step 1. Use BLAST to search all the proteins in the Swiss-Prot database for those proteins that have high homology (i.e. more than 60% pairwise sequence identity) with the protein **P** of Eq. 3. The proteins thus obtained are collected into a subset, $\mathbb{S}_\mathbf{P}^\mathrm{homo}$, called the homology set of **P**. Subsequently, retrieve the GO numbers of the protein in $\mathbb{S}_\mathbf{P}^\mathrm{homo}$ that has the highest homology with **P**. If it has no GO number at all, try the 2nd highest homologous protein in $\mathbb{S}_\mathbf{P}^\mathrm{homo}$; if it has no GO code either, try the 3rd highest homologous one; and so forth. The detailed description of this step as well as its rationale have been clearly elaborated in Chou *et al.*, (2011, 2012), and hence there is no need to repeat here. Eventually, suppose the homologous protein of **P** has a set of GO code given by

$$\left\{\mathrm{GO}_1^\mathrm{P}\ \ \mathrm{GO}_2^\mathrm{P}\ \ \cdots\ \ \mathrm{GO}_k^\mathrm{P}\ \ \cdots\ \ \mathrm{GO}_{n^\mathrm{g}}^\mathrm{P}\right\} \tag{5}$$

where $\mathrm{GO}_\mathrm{k}^\mathbf{P}$ $(k = 1,\ 2,\ \cdots,\ n^\mathrm{g})$ is the k-th GO code for the protein in $\mathbb{S}_\mathbf{P}^\mathrm{homo}$ that has been found with a set of GO codes according to the aforementioned order, and n^g is the total number of GO codes it has. Suppose the $N = 3,919$ sequence-different proteins in the benchmark dataset $\mathbb{S}$ are expressed as

$$\left\{\mathrm{P}^1,\ \mathrm{P}^2,\ \mathrm{P}^3,\ \ \cdots,\mathrm{P}^i,\ldots,\ \mathrm{P}^N\right\} \tag{6}$$

and the total number of proteins in the benchmark dataset $\mathbb{S}$ that have exactly the same GO code as $\mathrm{GO}_\mathrm{k}^\mathbf{P}$ is $N(k)$; i.e.

$$N(k) = \sum\nolimits_{i=1}^{N} \Delta(\mathrm{P}^i,\ \mathrm{GO}_k^\mathrm{P}) \tag{7}$$

where

$$\Delta\left(\mathrm{P}^i,\ \mathrm{GO}_k^\mathrm{P}\right) = \begin{cases} 1, & \text{if } \mathrm{GO}_k^\mathrm{P} \in\ \mathrm{P}^i \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Moreover, suppose among the $N(k)$ proteins, $n(k,\ u)$ belong(s) to the u-th subset $\mathbb{S}_u$; i.e.

$$N(k) = \sum\nolimits_{u=1}^{N^\mathrm{Loc}} n(k,\ u) \tag{9}$$

where $N^\mathrm{Loc} = 20$ is the total number of subcellular locations investigated (see Eq. 2 or Table 1).

Step 2. Based on Eqs. 7 and 9, the general PseAAC vector of Eq. 4 can be defined as

$$\Psi_u = \underset{1\le k\le n^\mathrm{g}}{\mathrm{Max}}\left[\frac{n(k,\ u)}{N(k)}\right]\ \left(u = 1,\ 2,\ \cdots,\ \Omega = N^\mathrm{Loc} = 20\right) \tag{10}$$

where the operator Max means taking the maximum value among those with different k values. It is through such a formulation to optimize the GO information and reduce the dimension of PseAAC vectors for predicting the subcellular localization of multi-label proteins.

Listed in Supplementary Material S3 are the PseAAC vectors formulated by Eq. 10 for the 3919 sequence-different proteins in the benchmark dataset. As we can see from there, the dimension of the current PseAAC vectors is reduced to 20, much lower than those in the previous GO approaches (Chou and Cai, 2003; Chou and Shen, 2006; Lin *et al.*, 2013).

2.3 Operation algorithm

In this study, the ML-GKR (multi-label Gaussian kernel regression) classifier (Cheng *et al.*, 2017) has been used to predict the protein subcellular localization, as described below. According to Eq. 10, the *i*-th protein P^i in the benchmark dataset can be formulated as

$$\mathrm{P}_{\mathrm{GO}}^{i} = \left[\Psi_1^i \quad \Psi_2^i \ \Psi_3^i \ \cdots \ \Psi_{20}^i\right]^{\mathrm{T}} \tag{11}$$

And its subcellular location(s) in the multi-label system can be formulated as a vector $\mathbf{L}^i$ given by

$$\mathrm{L}^i = \left[\ell_1^i \ \ell_2^i \ \ell_3^i \quad \cdots \quad \ell_{20}^i\right]^{\mathrm{T}} \tag{12}$$

where

$$\ell_u^i = \begin{cases} +1 \ \text{if} \ \mathrm{P}^i \in \mathbb{S}_u \\ -1 \ \text{otherwise} \end{cases} \quad (u = 1, \ 2, \ \cdots, \ 20) \tag{13}$$

Likewise, for a query protein $\mathbf{P}^{\mathrm{q}}$ we have

$$\mathrm{P}^{\mathrm{q}} = \left[\Psi_1^{\mathrm{q}} \quad \Psi_2^{\mathrm{q}} \ \Psi_3^{\mathrm{q}} \ \cdots \ \Psi_{20}^{\mathrm{q}}\right]^{\mathrm{T}} \tag{14}$$

Its subcellular location label(s) in the multi-label system should be given by

$$\mathrm{L}^{\mathrm{q}} = \left[\ell_1^{\mathrm{q}} \ \ell_2^{\mathrm{q}} \ \ell_3^{\mathrm{q}} \quad \cdots \quad \ell_{20}^{\mathrm{q}}\right]^{\mathrm{T}} \tag{15}$$

where

$$\ell_u^{\mathrm{q}} = \begin{cases} +1 \ \text{if} \ \Delta_{\mathrm{u}} \geq 0 \\ -1 \ \text{otherwise} \end{cases} \quad (u = 1, \ 2, \ \cdots, \ 20) \tag{16}$$

The Δ_u in Eq. 16 is given by

$$\Delta_u = \left[\sum\nolimits_{i=1}^{N} \ell_u^i \cdot \exp\left(-\frac{\|\mathrm{P}^{\mathrm{q}} - \mathrm{P}^i\|^2}{2\theta^2}\right)\right]\left[\sum\nolimits_{i=1}^{N} \exp\left(-\frac{\|\mathrm{P}^{\mathrm{q}} - \mathrm{P}^i\|^2}{2\theta^2}\right)\right]^{-1} \tag{17}$$

where θ is a parameter whose optimal value will be determined later, $\|\mathbf{P}^{\mathrm{q}} - \mathbf{P}^i\|$ is the Euclidean distance between the query protein (Eq. 14) and the *i*th protein (Eq. 11) in the benchmark dataset $\mathbb{S}$, as given by Chou and Zhang (1995); i.e.

$$\|\mathrm{P}_{\mathrm{GO}}^{\mathrm{q}} - \mathrm{P}_{\mathrm{GO}}^{i}\|^2 = \sum\nolimits_{u=1}^{20} \left(\Psi_u^{\mathrm{q}} - \Psi_u^i\right)^2 \tag{18}$$

Thus, the location label vector $\mathbf{L}^{\mathrm{q}}$ of Eq. 15 for the query protein $\mathbf{P}^{\mathrm{q}}$ is well defined, and hence its subcellular location or locations can be explicitly predicted as well. For example: if $\ell_1^{\mathrm{q}} = \ell_3^{\mathrm{q}} = \ell_{20}^{\mathrm{q}} = +1$ while all the other components in Eq. 15 are equal to -1, this means that the query protein $\mathbf{P}^{\mathrm{q}}$ is located in the 1st, 3rd and 20th subcellular locations; if $\ell_2^{\mathrm{q}} = +1$ while all the others are equal to -1, meaning that the query protein is located in the 2$^{\mathrm{nd}}$ subcellular location only; and so forth.

The predictor developed via the aforementioned procedures is called pLoc-mAnimal, where 'pLoc' stands for 'predict subcellular localization', and 'mAnimal' for 'multi-label animal proteins'. A flowchart to show how the pLoc-mAnimal predictor works is given in Figure 1.

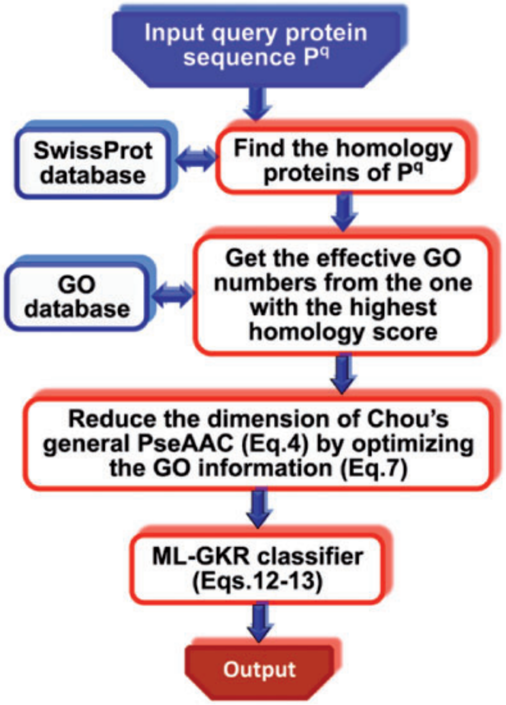
Input query protein sequence P^q
SwissProt database
Find the homology proteins of P^q
GO database
Get the effective GO numbers from the one with the highest homology score
Reduce the dimension of Chou's general PseAAC (Eq.4) by optimizing the GO information (Eq.7)
ML-GKR classifier (Eqs.12-13)
Output

Fig. 1. A flowchart to show how the pLoc-mAnimal predictor works

3 Results and discussion

As mentioned in the Chou's 5-step rule (Chou, 2011), one of the important procedures in developing a new predictor is how to objectively evaluate its anticipated accuracy. To address this, two issues need to be considered. (i) What metrics should be used to quantitatively reflect the predictor's quality? (ii) What test approach should be adopted to score the metrics?

3.1 A set of five metrics for multi-label systems

Different from the metrics used to measure the prediction quality of single-label systems, the metrics for the multi-label systems are much more complicated. To make them more intuitive and easier to understand for most experimental scientists, here we use the following five metrics proposed by Chou (2013) that have recently been widely used for studying various multi-label systems (see, e.g. (Cheng *et al.*, 2017; Lin and Xu, 2016; Qiu *et al.*, 2016))

$$\begin{cases} \text{Aiming} \uparrow = \dfrac{1}{N^{\text{test}}} \sum\nolimits_{k=1}^{N^{\text{test}}} \left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k^*\|}\right), \ [0, \ 1] \\[2ex] \text{Coverage} \uparrow = \dfrac{1}{N^{\text{test}}} \sum\nolimits_{k=1}^{N^{\text{test}}} \left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k\|}\right), \ [0, \ 1] \\[2ex] \text{Accuracy} \uparrow = \dfrac{1}{N^{\text{test}}} \sum\nolimits_{k=1}^{N^{\text{test}}} \left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k \cup \mathbb{L}_k^*\|}\right), \ [0, \ 1] \\[2ex] \text{Absolute true} \uparrow = \dfrac{1}{N^{\text{test}}} \sum\nolimits_{k=1}^{N^{\text{test}}} \Delta\left(\mathbb{L}_k, \ \mathbb{L}_k^*\right), \ [0, \ 1] \\[2ex] \text{Absolute false} \downarrow = \dfrac{1}{N^{\text{test}}} \sum\nolimits_{k=1}^{N^{\text{test}}} \left(\dfrac{\|\mathbb{L}_k \cup \mathbb{L}_k^*\| - \|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{M}\right), \quad [1, \ 0] \end{cases} \tag{19}$$

where N^{test} is the total number of the tested samples, M is the total number of labels for the investigated system, means the operator acting on the set therein to count the number of its elements, $\cup$ means the symbol for the 'union' in the set theory, $\cap$ denotes the symbol for the ‘intersection’, $\mathbb{L}_k$ denotes the subset that contains all the labels observed by experiments forthe k-th sample, $\mathbb{L}_k^*$ represents the subset that contains all the labels predicted for the k-th sample, and

$$\sum_{k=1}^{N^{\text{test}}} \Delta\left(\mathbb{L}_k,\ \mathbb{L}_k^*\right) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k^* \text{ are identical to those in } \mathbb{L}_k \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

In Eq. 19, the first four metrics with an upper arrow ↑ are called positive metrics, meaning that the larger the rate is, the better the prediction quality will be; the 5th metrics with a down arrow ↓ is called negative metrics, implying just the opposite meaning.

From Eq. 19 we can see the following: (i) the ‘Aiming’ defied by the 1st sub-equation is for checking the rate or percentage of the correctly predicted labels over the practically predicted labels; (ii) the ‘Coverage’ defined in the 2nd sub-equation is for checking the rate of the correctly predicted labels over the actual labels in the system concerned; (iii) the ‘Accuracy’ in the 3rd sub-equation is for checking the average ratio of correctly predicted labels over the total labels including correctly and incorrectly predicted labels as well as those real labels but are missed in the prediction; (iv) the ‘Absolute true’ in the 4th sub-equation is for checking the ratio of the perfectly or completely correct prediction events over the total prediction events; (v) the ‘Absolute false’ in the 5th sub-equation is for checking the ratio of the completely wrong prediction over the total prediction events.

3.2 Jackknife test

Three cross-validation methods are often used in statistical prediction. They are: (i) independent dataset test, (ii) subsampling (or K-fold cross-validation) test and (iii) jackknife test (Chou and Zhang, 1995). Of these three, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in Chou (2011). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g. (Ahmad *et al.*, 2016; Chou and Cai, 2005; Dehzangi *et al.*, 2015; Khan *et al.*, 2017; Nanni *et al.*, 2014; Shen and Chou, 2007b; Zhou and Doctor, 2003)). Accordingly, the jackknife test was also used in this study.

3.3 Parameter determination

Since Eq. 17 contains a parameter θ, the predicted results obtained by pLoc-mAnimal will depend on the parameter’s value. In this study, the optimal value for θ was determined by maximizing the absolute true rate (see the 4th sub-equation in Eq. 19) by the jackknife validation on the benchmark dataset. It was observed that when θ = 1/6, the absolute true rate reached its highest score. And such a value would be used for further study.

3.4 Comparison with the state-of-the-art predictor

Listed in Table 2 are the rates obtained by the current pLoc-mAnimal predictor via the jackknife test on the benchmark dataset (Supplementary Material S1). For facilitating comparison, listed in that table are also the corresponding results obtained by the iLoc-Animal, the existing most powerful predictor for identifying the subcellular localization of animal proteins with both single and multiple sites.

Table 2. Comparison with the state-of-the-art method in predicting animal protein subcellular localization[a]

Predictor	Aiming (↑)	Coverage (↑)	Accuracy (↑)	Absolute true (↑)	Absolute false (↓)
pLoc-mAnimal[b]	87.96%	85.33%	84.64%	73.11%	1.650%
iLoc-Animal[c]	72.45%	34.18%	42.76%	35.93%	6.330%

[a] The rates listed below were derived by the jackknife test on the benchmark dataset $\mathbb{SS}$(Supplementary Material S1).

[b] The predictor proposed in this paper.

[c] The predictor proposed in Lin et al. (2013). Note that the rates in this table are somewhat different with the values originally reported in Lin et al. (2013). This is because the original values were derived based on a benchmark dataset that contained some proteins with ≥40% pairwise sequence identity to each other. See Section 2.1 for more explanation.

As shown in Table 2, the rates for the four positive metrics by pLoc-mAnimal are significantly higher than those by iLoc-Animal (Lin *et al.*, 2013), while the opposite is true for the negative metrics. As pointed out in a comprehensive review (Chou, 2013), among the aforementioned five metrics, the most important are ‘absolute true’ and ‘absolute false’. It is extremely difficult to increase the absolute true rate and reduce the absolute false rate for a multi-label predictor. Therefore, in reporting the results of their various prediction methods for multi-label systems, many investigators (see, e.g. (Chen *et al.*, 2012; Chou and Shen, 2007a, 2008, 2010a,b; Shen and Chou, 2007a, 2009a,b, 2010a,b) even did not mention the ‘absolute true’ and ‘absolute false’ rates. It has been demonstrated in this study, however, that the absolute true rate obtained by the new predictor is over 37% higher than that by iLoc-Animal (Lin *et al.*, 2013), while the absolute false rate by the new predictor is almost four times lower. It is indeed a compelling fact to show the superior of the new predictor over the existing state-of-the-art one in predicting the subcellular locations of multi-label animal proteins (Lin *et al.*, 2013).

Besides, no prediction quality was reported for each of 20 subcellular locations in the iLoc-Animal paper (Lin *et al.*, 2013). To in-depth analyze the corresponding prediction quality by the proposed predictor pLoc-mAnimal for the samples in each of the 20 subsets in the Supplementary Material S1, let us introduce the following set of metrics:

$$\begin{cases} \mathrm{Sn}(i) = 1 - \dfrac{N_{-}^{+}(i)}{N^{+}(i)} & 0 \ \le \mathrm{Sn} \le 1 \\[2ex] \mathrm{Sp}(i) = 1 - \dfrac{N_{+}^{-}(i)}{N^{-}(i)} & 0 \le \mathrm{Sp} \le 1 \\[2ex] \mathrm{Acc}(i) = 1 - \dfrac{N_{-}^{+}(i) + N_{+}^{-}(i)}{N^{+}(i) + N^{-}(i)} & 0 \le \mathrm{Acc} \le 1 \\[2ex] \mathrm{MCC}(i) = \dfrac{1 - \left(\dfrac{N_{-}^{+}(i)}{N^{+}(i)} + \dfrac{N_{+}^{-}(i)}{N^{-}(i)}\right)}{\sqrt{\left(1 + \dfrac{N_{+}^{-}(i) - N_{-}^{+}(i)}{N^{+}(i)}\right)\left(1 + \dfrac{N_{-}^{+}(i) - N_{+}^{-}(i)}{\mathrm{N}^{-}(i)}\right)}} \\ \qquad\qquad -1 \le \mathrm{MCC} \le 1 \\[2ex] (i = 1,\ 2,\ \cdots,\ 20) \end{cases} \tag{21}$$

where Sn, Sp, Acc and MCC represent the sensitivity, specificity, accuracy and Mathew’s correlation coefficient, respectively (Chen

Table 3. Performance of pLoc-mAnimal for each of the 20 subcellular locations

i	Location[a]location	Sn(i)[b]	Sp(i)[b]	Acc(i)[b]	MCC(i)[b]
1	Acrosome	0.8077	0.9997	0.9985	0.8773
2	Cell cortex	0.8846	0.9641	0.9462	0.8463
3	Cell membrane	0.7273	0.9979	0.9964	0.6945
4	Centriole	0.6860	0.9927	0.9860	0.6749
5	Centrosome	0.5366	0.9964	0.9916	0.5684
6	Cytoplasm	0.8176	0.9203	0.8867	0.7416
7	Cytoskeleton	0.8032	0.9787	0.9648	0.7642
8	Endoplasmic reticulum	0.8527	0.9835	0.9684	0.8444
9	Endosome	0.7394	0.9886	0.9796	0.7137
10	Extracellular space	0.7938	0.9971	0.9920	0.8294
11	Golgi apparatus	0.8391	0.9853	0.9735	0.8220
12	Lysosome	0.7807	0.9924	0.9862	0.7603
13	Melanosome	0.9086	0.9850	0.9750	0.8907
14	Microsome	0.7000	1.0000	0.9992	0.8363
15	Mitochondrion	0.8947	0.9987	0.9972	0.9013
16	Nucleus	0.8336	0.9366	0.9087	0.7693
17	Peroxisome	0.9219	0.9987	0.9974	0.9206
18	Plasma membrane	0.8846	0.9641	0.9462	0.8463
19	Spindle	0.7379	0.9934	0.9867	0.7383
20	Synapse	0.8182	0.9961	0.9931	0.7967

[a]See Table 1 and the relevant context for further explanation.
[b]See Eq. 21 for the metrics definition.

et al., 2007), and i denotes the i-subcellular location in the benchmark dataset. $N^{+}(i)$ is the total number of the samples investigated in the i-th subset, whereas $N^{+}_{-}(i)$ is the number of the samples in $N^{+}(i)$ that are incorrectly predicted to be of other locations; $N^{-}(i)$ is the total number of samples in any location but not the i-th location, whereas $N^{-}_{+}(i)$ is the number of the samples in $N^{-}(i)$ that are incorrectly predicted to be of the i-th location. The metrics of Eq. 21 have been widely used to examine the quality of predictors in genome/proteome analysis (see, e.g. (Chen *et al.*, 2013; Lin *et al.*, 2014, 2017a; Xu *et al.*, 2014)) and computational biomedicine (see, e.g. (Liu *et al.*, 2017c,d; Qiu *et al.*, 2017; Xu *et al.*, 2017)).

Listed in Table 3 are the corresponding results obtained by pLoc-mAnimal for each of the 20 subcellular locations. As we can see from the table, the scores for each of the 20 subcellular locations are also very high, fully consistent with its overall performance as reported in Table 2.

3.5 Web server and user guide

For the convenience of most experimental scientists, the web-server of pLoc-mAnimal predictor has been established. Moreover, to maximize their convenience, a step-by-step guide is given below in Supplementary Material S4.

4 Conclusion

Protein subcellular location prediction is a challenging problem, particularly when the query proteins have multi-label features meaning that they may occur at two or more different location sites. Here, we have developed a new predictor called pLoc-mAnimal. Compared with iLoc-Animal (Lin *et al.*, 2013), the existing most powerful predictor also having the capacity to deal with the multiple locations of animal proteins, the scores achieved by the new predictor are remarkably better in all the five metrics widely used to check the quality of a multi-label predictor.

Why could the new predictor be so powerful? The key is that the feature vectors used in the new predictor have been optimized via a special general PsaAAC approach to substantially reduce their dimension but significantly optimize their cluster features as shown by Eq. 10.

Since the publically accessible web-server represents the future direction for developing practically more useful prediction method (Chou and Shen, 2009), the web-server for pLoc-Animal has been established and its user guide is given in Supplementary Material S4. It is anticipated that pLoc-Animal will become a very useful high throughput tool for annotating the subcellular location(s) of animal proteins.

Acknowledgements

The authors wish to thank the four anonymous reviewers whose constructive comments were very helpful to strengthen the presentation of this study.

Funding

The authors wish to thank the four anonymous reviewers whose constructive comments were very helpful to strengthen the presentation of this study. This work was supported by the grants from the National Natural Science Foundation of China (No. 31560316, 61262038, 61202313, 61271114 and 31260273), the Province National Natural Science Foundation of JiangXi (No. 20171ACB20023 and 20132BAB201053), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No.20120BDH80023), Innovation Program of Shanghai Municipal Education Commission (No. 14ZZ068).

References

Ahmad,K. *et al.* (2016) Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition. *J. Membr. Biol.*, **249**, 293–304.

Chen,J. *et al.* (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, **33**, 423–428.

Chen,J. *et al.* (2016a) dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. *Scientific Rep.*, **6**, 32333.

Chen,L. *et al.* (2012) Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS ONE*, **7**, e35254.

Chen,W. *et al.* (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **41**, e68.

Chen,W. *et al.* (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol. BioSyst.*, **11**, 2620–2634.

Chen,W. *et al.* (2016b) iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids*, **5**, e332.

Cheng,X. *et al.* (2017) iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics*, **33**, 341–346.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins Struct. Funct. Genet.* **43**, 246–255. (Erratum: ibid., 2001, Vol.44, 60).

Chou,K.C. (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics*, **6**, 262–274.

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.*, **273**, 236–247.

Chou,K.C. (2013) Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.*, **9**, 1092–1100.

Chou,K.C. (2015) Impacts of bioinformatics to medicinal chemistry. *Med. Chem.*, **11**, 218–234.

Chou,K.C. (2017) An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem.*, **17** 2337–2358.

Chou,K.C. and Cai,Y.D. (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun. (BBRC)*, **311**, 743–747.

Chou,K.C. and Cai,Y.D. (2005) Prediction of membrane protein types by incorporating amphipathic effects. *J. Chem. Inf. Model.*, **45**, 407–413.

Chou,K.C. and Shen,H.B. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res.*, **5**, 1888–1897.

Chou,K.C. and Shen,H.B. (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.*, **6**, 1728–1734.

Chou,K.C. and Shen,H.B. (2007b) Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **370**, 1–16.

Chou,K.C. and Shen,H.B. (2008) Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms, Natural Science, 2010, 2, 1090-1103). *Nat. Protoc.*, **3**, 153–162.

Chou,K.C. and Shen,H.B. (2009) Review: recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.*, **1**, 63–92.

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

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

Chou,K.C. *et al.* (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One*, **6**, e18258.

Chou,K.C. *et al.* (2012) iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.*, **8**, 629–641.

Chou,K.C. and Zhang,C.T. (1995) Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Dehzangi,A. *et al.* (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *J. Theor. Biol.*, **364**, 284–294.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

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

Huang,C. and Yuan,J. (2013) Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. *Biosystems*, **113**, 50–57.

Jia,J. *et al.* (2016a) iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem.*, **497**, 48–56.

Jia,J. *et al.* (2016b) pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.*, **394**, 223–230.

Khan,M. *et al.* (2017) Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. *J. Theor. Biol.*, **415**, 13–19.

Lin,H. *et al.* (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.*, **42**, 12961–12972.

Lin,W. and Xu,D. (2016) Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. *Bioinformatics*, **32**, 3745–3752.

Lin,W.Z. *et al.* (2013) iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. BioSyst.*, **9**, 634–644.

Liu,B. *et al.* (2017a) iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*, **33**, 35–41.

Liu,B. *et al.* (2017b) Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein Sequences. *Nat. Sci.*, **9**, 67–91.

Liu,B. *et al.* (2017c) 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol. Ther. Nucleic Acids*, **7**, 267–277.

Liu,L.M. *et al.* (2017d) iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem.*, **13**, 552–559.

Meher,P.K. *et al.* (2017) Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.*, **7**, 42362.

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

Nakai,K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.*, **54**, 277–344.

Nanni,L. *et al.* (2014) Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *J. Theor. Biol.*, **360**, 109–116.

Pacharawongsakda,E. and Theeramunkong,T. (2013) Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC. *IEEE Trans. Nanobiosci.*, **12**, 311–320.

Qiu,W.R. *et al.* (2017) iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. *Med Chem.*, doi:10.2174/1573406413666170623082245.

Qiu,W.R. *et al.* (2016) iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics*, **32**, 3116–3123.

Shen,H.B. and Chou,K.C. (2007a) Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun. (BBRC)*, **355**, 1006–1011.

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

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

Shen,H.B. and Chou,K.C. (2009b) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.*, **394**, 269–274.

Shen,H.B. and Chou,K.C. (2010a) Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J. Theor. Biol.*, **264**, 326–333.

Shen,H.B. and Chou,K.C. (2010b) Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J. Biomol. Struct. Dyn. (JBSD)*, **28**, 175–186.

Wan,S. *et al.* (2013) GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.*, **323**, 40–48.

Wang,T. *et al.* (2008) Predicting membrane protein types by the LLDA algorithm. *Protein Pept. Lett.*, **15**, 915–921.

Wang,X. *et al.* (2013) 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. *Protein Pept. Lett.*, **20**, 309–317.

Wu,Z.C. *et al.* (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. BioSyst.*, **7**, 3287–3297.

Wu,Z.C. *et al*. (2012) iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein Pept. Lett*., **19**, 4–14.

Xiao,X. *et al*. (2011a) 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**, 42–51.

Xiao,X. *et al*. (2011b) A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PLoS ONE*, **6**, e20592.

Xu,Y. *et al*. (2017) iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med Chem*., **13**, 544–551.

Xu,Y. *et al*. (2014) iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. J. Mol. Sci*., **15**, 7594–7610.

Zhou,G.P. and Doctor,K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins Struct. Funct. Genet*., **50**, 44–48.