*Send Orders for Reprints to reprints@benthamscience.net*



# Recent Advances in Predicting Protein Classification and Their Applications to Drug Development

Xuan Xiao[1,2,4,*], Wei-Zhong Lin[1] and Kuo-Chen Chou[3,4]

[1]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China;* [2]*Information School, ZheJiang Textile & Fashion College, NingBo, 315211 China;* [3]*Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia;* [4]*Gordon Life Science Institute, 53 South Cottage Road, Belmont, MA 02478, USA*

**Abstract:** With the explosion of protein sequences generated in the postgenomic era, the gap between the number of attribute-known proteins and that of uncharacterized ones has become increasingly large. Knowing the key attributes of proteins is a shortcut for prioritizing drug targets and developing novel new drugs. Unfortunately, it is both time-consuming and costly to acquire these kinds of information by purely conducting biological experiments. Therefore, it is highly desired to develop various computational tools for fast and effectively classifying proteins according to their sequence information alone. The process of developing these high throughput tools is generally involved with the following procedures: (1) constructing benchmark datasets; (2) representing a protein sequence with a discrete numerical model; (3) developing or introducing a powerful algorithm or machine learning operator to conduct the prediction; (4) estimating the anticipated accuracy with a proper and objective test method; and (5) establishing a user-friendly web-server accessible to the public. This minireview is focused on the recent progresses in identifying the types of G-protein coupled receptors (GPCRs), subcellular localization of proteins, DNA-binding proteins and their binding sites. All these identification tools may provide very useful informations for in-depth study of drug metabolism.

**Keywords:** GPCR type, protein subcellular localization, DNA-binding protein, discrete model, PseAAC, protein attribute predictor, web-servers.

## 1. INTRODUCTION

Proteins perform various functions within living organisms, including catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another, etc. Proteins are also essential parts of organisms and participate in virtually every process within cells. In modern pharmaceutical therapies proteins drugs have become increasingly dependent on our knowledge on proteins. The functions of proteins are determined by their structures. Many studies have indicated that the informations derived from classifying various attributes of proteins, such as G-protein-coupled receptors (GPCRs) types, subcellular localization of proteins, and DNA-binding proteins, are very useful for rational drug design.

GPCRs, also called the 7-transmembrane receptors (Fig. **1**), are the largest family of cell surface receptors and are key mediators of the effects of numerous endogenous neurotransmitters, hormones, cytokines, therapeutic drugs, and drugs-of- abuse [1]. They mediate many important physiological functions and are considered as one of the most successful therapeutic targets for a broad spectrum of disease. The design and implementation of high-throughput GPCR assays that allow the cost-effective screening of large compound libraries to identify novel drug candidates are critical in early drug discovery [2].

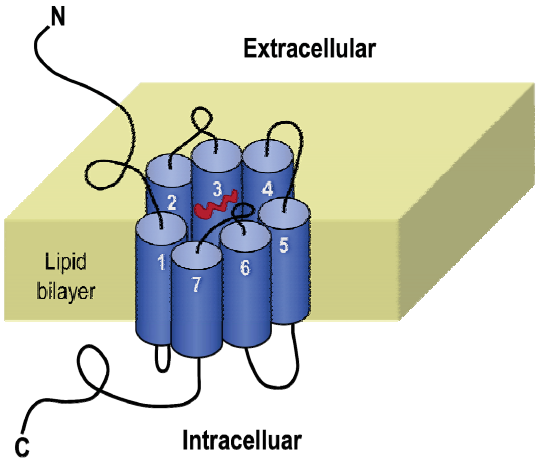


**Fig. (1).** Schematic representation of a GPCR with a trademark of seven-transmembrane helices, depicted as cylinders and connected by alternating cytoplasmic and extracellular hydrophilic loops. The 7-helix bundle thus formed has a central pore on its extracellular surface. The red entity located in the central pore represents a ligand messenger. Reproduced with permission from Chou [2, 165].

Knowledge of subcellular locations (Fig. **2**) of proteins can provide key hints and useful insights for revealing their functions [3-11], helping to understand the intricate path-

*Address correspondence to this author at the Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China; Tel/Fax: 086-0798-8499229; E-mail: xxiao@gordonlifescience.org

1873-5294/13 $58.00+.00 

ways that regulate biological processes at the cellular level [12-14]. It is also very useful for identifying and prioritizing drug targets during the process of drug development. For example, the functions of apoptosis proteins are closely related to their subcellular locations. These proteins are very important for understanding the mechanism of programmed cell death [10, 15, 16]. Also, knowledge of subcellular localization of viral proteins in a host cell or virus-infected cell is closely related to their destructive tendencies and consequence [17-19]. Subcellular localization of proteins is vital for the signaling, metabolic or structural properties of the cell. Proteins with incorrect subcellular locations can cause disorders that involve biogenesis, protein aggregation, cell metabolism or signaling [20]. Listed in (Table **1**) are some human diseases caused by proteins located to the wrong subcellular compartment.

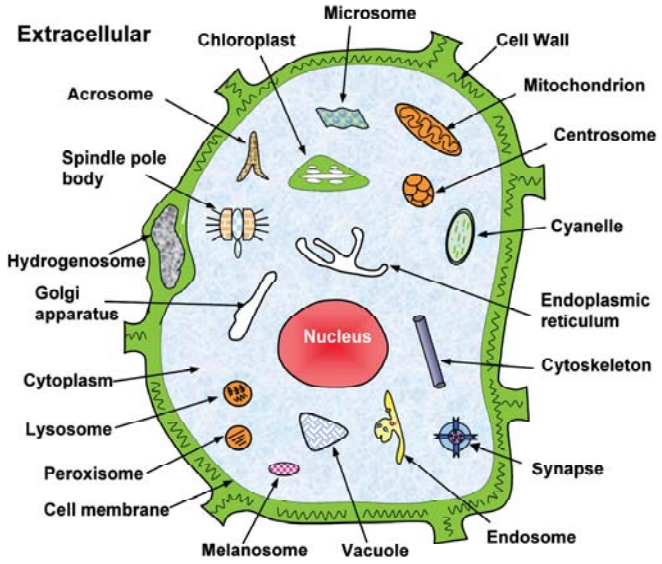


**Fig. (2).** Schematic illustration to show the 22 subcellular locations of eukaryotic proteins: (1) acrosome, (2) cell membrane, (3) cell wall, (4) centriole, (5) chloroplast, (6) cyanelle, (7) cytoplasm, (8) cytoskeleton, (9) endoplasmic reticulum, (10) endosome, (11) extracellular, (12) Golgi apparatus, (13) hydrogenosome, (14) lysosome, (15) melanosome, (16) microsome (17) mitochondria, (18) nucleus, (19) peroxisome, (20) spindle pole body, (21) synapse, and (22) vacuole. Reproduced from Chou [32] with permission.

DNA-binding proteins (Fig. **3**) play crucial roles in various biological processes in organisms [21, 22], such as recognition of specific nucleotide sequences, regulation of transcription, and regulation of gene expression. There are several different DNA binding domains in the promoter regions of transcription factors including zinc fingers, homeodomains, helix loop helices and leucine zippers. It is estimated that in the human genome the total number of transcription factors alone can be as high as 3000 or about 10% of all protein-coding genes.

With the explosion of protein sequences generated in the postgenomic era, the gap between the number of attribute-known proteins and that of uncharacterized ones has become increasingly large. Many efforts have been made in both academic institutions and pharmaceutical industries in order to determine the structures and functions of uncharacterized proteins by means of various techniques such as cryoelectron microscopy, crystallography, and NMR. However, some proteins, particularly membrane proteins important for drug development, are very difficult to crystallize and most of them will not dissolve in normal solvents. Although the recently developed state-of-the-art NMR technique is a very powerful tool in determining the 3D structures of membrane proteins [23-27], it is time-consuming and costly. Also, although some membrane protein structures can be derived by using homology approaches (see, e.g., [28-31]), unfortunately the number of templates for membrane proteins is quite limited. For example, more than thousand GPCR sequences are known, and much more are expected to come in the near future, yet the valid templates for them are only a dozen or so. In view of this, it would be highly desired to develop novel computational methods to predict various function-related attributes [32] of proteins based on their primary sequences alone.

During the last two decades or so, considerable efforts have been invested in this regard. For instance, Naveed *et al.* [33] proposed the GPCR-MPredictor, which can efficiently predict GPCRs at five levels. Xiao *et al.* [34, 35] developed GPCR-CA and GPCR-2L predictors by hybridizing the following three different modes of pseudo amino acid composition (PseAAC) [36]: the functional domain PseAAC, low-frequency Fourier spectrum PseAAC and protein cellular automata image PseAAC. Goldfeld *et al.* [37] presented loop structure prediction results of the intracellular and extracellular loops of four GPCRs; Wu and Xiao *et al.* [34, 38-42] used the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites; Wang *et al.* [43] constructed Virus-ECC-mPLoc predictor, which exploited correlations between sub-cellular locations and hybridized the gene ontology information with the dipeptide composition information; He *et al.* [44] discussed the imbalanced multi-modal multi-label learning for human proteins subcellular localization with both single and multiple sites; Chen *et al.* [45] predicted target DNA sequences of DNA-binding proteins based on unbound structures; van Dyke *et al.* [46] identified preferred drug-DNA binding sequences; Klein *et al.* [47] provided two powerful methods: the yeast one-hybrid system and the yeast two-hybrid system to identify DNA-binding proteins and protein-protein interactions; Lin *et al.* [48] identified DNA binding proteins using random forest. Besides, there are many studies for predicting protein subcellular localization (see, e.g, [5, 49-73] as well as a long list references cited in two review articles [9, 74].

The papers published for protein classification prediction or computational proteomics are extremely large. The current review was focused on those methods with the following features: (i) the capacity to be able to deal with multi-label systems [32]; (ii) rigorous benchmark datasets established by imposing a stringent cutoff threshold to reduce the homology bias and redundancy; (iii) novel features to represent the protein samples; (iv) high application values by providing user-friendly web-servers.

**Table 1. Mislocalized Proteins that have been Associated with Human Disease [20].**

| Protein | Disease | Protein | Disease |
|---|---|---|---|
| SRY | Swyer syndrome | Rhodopsin | Retinitis pigmentosa |
| SHOX | Léri–Weill dyschondrosteosis | AVPR2 | Nephrogenic diabetes insipidus |
| TRPS1 | TRPS | ATP7B | Wilson disease |
| ARX | XLAG | ABCA1 | Tangier disease |
| FOXP2 | Speech–language disorder | Tau | Neurodegenerative diseases |
| AIRE | APECED | TARDBP | ALS and FTLD |
| RPS19 | Diamond–Blackfan anemia | FUS | FTLD |
| AGT | Primary hyperoxaluria type 1 | FOXO | Various types of cancer |
| hsMOK2 | Laminopathy | p53 | Various types of cancer |
| SHOC2 | Noonan-like syndrome | | |

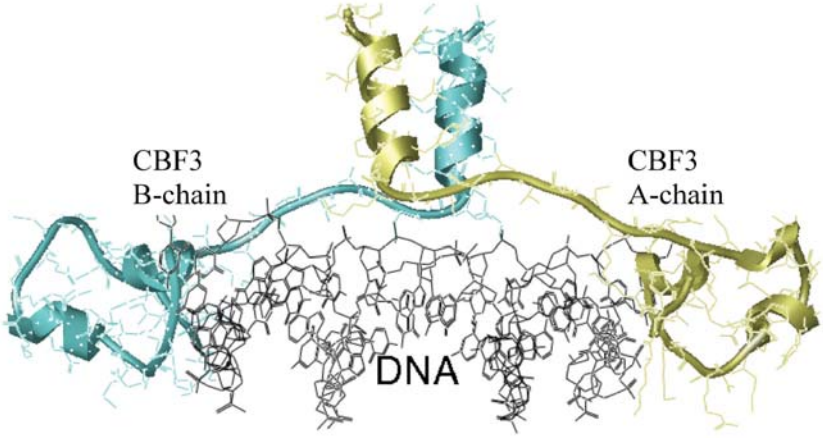


**Fig. (3).** Illustration to show the binding of a protein (CBF3) with DNA via its A-chain and B-chain. The protein is shown in the ribbon drawing, while DNA in dot-and-stick drawing. Reproduced from Chou [29] with permission.

## 2. BENCHMARK DATASET CONSTRUCTION

Constructing a high quality and updated benchmark dataset is crucially important for developing a protein attribute predictor. To realize this, a feasible way was to collect the data from some molecular biology databases, such as protein knowledgebase (UniProtKB: http://www.uniprot.org), protein data bank (PDB: http://www.rcsb.org/pdb/home/home.do), protein database (http://www.ncbi.nlm.nih.gov/protein ) provided by National Center for Biotechnology information (NCBI), etc.

Another way was to utilize the special databases. For example, GPCRDB [75, 76] is a molecular-class information system that collects, combines, validates and stores large amounts of heterogenous data on GPCRs. Using the GPCRDB, Worth *et al.* [77] presented a comprehensive database for GPCR template predictions and homology models, named GPCR-SSFE. Tanz *et al.* [78] constructed a subcellular location database for Arabidopsis proteins, named SUBA3 by combining the manual literature curation of largescal subcellular proteomics, fluorescent protein visualization and protein-protein interaction datasets with subcellular targeting calls from 22 prediction programs. Lum *et al.* [79] constructed a database called FunSecKB specially for the secreted fungal proteins.

The benchmark datasets constructed recently were generally according to the following the steps.

Step 1. Searching for protein samples from a database using some relevant key words, such as "DNA binding", "G protein couple receptors", and "subcellular location", etc.

Step 2. Sequences annotated with "fragment" were excluded; also, sequence with less than 50 amino acid (AA) residues were removed.

Step 3. Reducing redundancy and homology bias. Usually, the CD-HIT [80, 81] or similar programs was utilized to winnow those sequences according to some threshold. For example, setting the threshold at 40% [82] would get rid of all the proteins from the benchmark dataset that had $\geq$ 40% pairwise sequence identity to any other in a same subset; or,

to make the threshold even harder by setting it at 25% [74] to rid of all the proteins with $\geq 25\%$ pairwise sequence identity.

Finally, the benchmark dataset $S$ constructed via the above steps was further divided into a training dataset $S^{\text{Train}}$ and an independent testing dataset $S^{\text{Test}}$; i.e.,

$$\begin{cases} S^{\text{Train}} \bigcap S^{\text{Test}} = \varnothing \\ S^{\text{Train}} \bigcup S^{\text{Test}} = S \end{cases} \quad (1)$$

where $\bigcap$, $\bigcup$, and $\varnothing$ represent the symbols for "intersection", "union", and "empty set" in the set theory. If, however, a predictor was examined by subsampling test or jackknife test [83], no such a division was needed since the entire benchmark dataset could be used for the purposes of both training and testing the predictor without causing any memory bias problem, as elucidated in [84].

## 3. FORMULATING PROTEIN SAMPLES

According to [85] and demonstrated by a series of recent publication (see, e.g., [86-93]), a protein or peptide sequence in any discrete model can always be expressed by the general form of PseAAC [85]; i.e.,

$$\mathbf{P} = \left[ \begin{array}{cccccc} \psi_1 & \psi_2 & \cdots & \psi_u & \cdots & \psi_\Omega \end{array} \right]^{\mathbf{T}} \quad (2)$$

where $\mathbf{P}$ represent the protein sample or its feature vector, $\mathbf{T}$ the transpose operator, while the vector's dimension $\Omega$ and components $\psi_u$ $(u = 1,2, \cdots, \Omega)$ will depend on how to extract the desired information from the protein sequences. Described below are just for some of these formulations that were often used in protein attribute predictions.

### 3.1. GO (Gene Ontology) Formulation

The GO project (http://www.geneontology.org) is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. GO contains three domains: cellular component, molecular function, and biological process. Accordingly, protein samples defined in a GO database space would be clustered in a way better reflecting their structure and function. GO has become a general feature of proteomics that was commonly used [38, 39, 53, 94-104]. The detailed procedures for generating the GO formulation can be briefed as follows.

Step 1. Use BLAST [105] to search the homologous proteins of the query protein $\mathbf{P}$ from the Swiss-Prot database, with the expect value E$\leqslant$0.001 for the BLAST parameter.

Step 2. Those proteins which have $\geq 60\%$ pairwise sequence identity with the query protein $\mathbf{P}$ are collected into a set, $\mathbb{S}^{\mathbf{P-homo}}$, called the "homology set" of **P**. All the proteins in $\mathbb{S}^{\mathbf{P-homo}}$ can be deemed as the "representative proteins" of **P** and they have their accession numbers clearly defined in the Swiss-Prot database.

Step 3. Search each of these accession numbers collected in Step 2 against the GO database at http://www.ebi.ac.uk/GOA to find the corresponding GO numbers.

Step 4. The query protein **P** can be expressed via representative proteins in $\mathbb{S}^{\mathbf{P-homo}}$ via **Eq.2** with

$$\psi_u^{\text{GO}} = \begin{cases} 1, \text{ if a hit is found against the } u\text{-th GO number} \\ \quad \text{for any of the proteins in } \mathbb{S}^{\text{P-homo}} \\ 0, \text{ otherwise} \end{cases} \quad (u = 1,2,\cdots,\Omega) \quad (3)$$

and the value of $\Omega$ is the maximal number of terms contained in GO database.

Although the protein sample formulation derived via the above steps could incorporate some GO information, it has a shortcoming: only the integer number 0 and 1 were used to reflect the GO information. Such an over-simplified formulation might cause some important information loss and hence limit the prediction quality. Thus, Wu *et al.* [38-41, 99, 100] proposed the following procedures for improvement.

First, the GO database contains many GO terms' number. However, these numbers do not increase successively and orderly. For easier handling, a reorganization and compression procedure was used to renumber them as done in [41, 106, 107]. The GO database obtained through such a treatment is called GO_compress database. Using GO_compress database can reduce the number of elements in **Eq.2**.

Second, the elements in **Eq.3** were replaced by

$$\psi_u^{\text{GO}} = \frac{\sum_{k=1}^{\mathbb{N}_{\text{P}}^{\text{homo}}} g(u,k)}{\mathbb{N}_{\text{P}}^{\text{homo}}} \quad (4)$$

where $\mathbb{N}_{\text{P}}^{\text{homo}}$ is the number of representative proteins in $\mathbb{S}^{\mathbf{P-homo}}$, and

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

Although the GO formulation could yield better results in identifying protein subcellular locations, it might become a naught vector or meaningless when the protein **P** did not have significant homology to any protein in the Swiss-Prot database. When cases happened like that, the other kind of formulations such as the sequential evolutional formulation would be adopted as will be mentioned below.

### 3.2. Sequential Evolution Information Formulation

Biology is a natural science with a historic dimension. All biological species have developed starting from a very limited number of ancestral species. Their evolution involves changes of single residues, insertions and deletions of several residues [108], gene doubling, and gene fusion. With these changes accumulated for a long period of time, many similarities between initial and resultant amino acid se-

quences are gradually eliminated, but the corresponding proteins may still share many common attributes [31], such as having basically the same biological function and residing at a same subcellular location. To extract the sequential evolution information and use it to define the components of **Eq.2**, the PSSM (Position Specific Scoring Matrix) was used [38, 40, 41, 97, 98, 100, 109-111].

The PSSM is a $L \times 20$ matrix which was generated by using PSI-BLAST [105] to search the Swiss-Prot database or other protein database.

$$\mathbf{P}_{\mathrm{PSSM}}^{(0)} = \begin{bmatrix} m_{1,1}^{(0)} & m_{1,2}^{(0)} & \cdots & m_{1,20}^{(0)} \\ m_{2,1}^{(0)} & m_{2,2}^{(0)} & \cdots & m_{2,20}^{(0)} \\ \vdots & \vdots & \vdots & \vdots \\ m_{L,1}^{(0)} & m_{L,2}^{(0)} & \cdots & m_{L,20}^{(0)} \end{bmatrix} \tag{6}$$

where $L$ is the length of **P**, $m_{i,j}^{(0)}\,(1 \le i \le L, 1 \le j \le 20)$ represents the score of the amino acid residue in the $i$-th position of the protein sequence being changed to amino acid type $j$ during the evolutionary process. Here, the numerical codes 1, 2, …, 20 are the alphabetical order of their single character codes. Because the value of $m_{i,j}^{(0)}$ may be less than zero, in order to make every element in **Eq.6** to be greater than zero, a conversion was performed through the standard sigmoid function to make Eq.6 become

$$\mathbf{P}_{\mathrm{PSSM}}^{(1)} = \begin{bmatrix} m_{1,1}^{(1)} & m_{1,2}^{(1)} & \cdots & m_{1,20}^{(1)} \\ m_{2,1}^{(1)} & m_{2,2}^{(1)} & \cdots & m_{2,20}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ m_{L,1}^{(1)} & m_{L,2}^{(1)} & \cdots & m_{L,20}^{(1)} \end{bmatrix} \tag{7}$$

where

$$m_{i,j}^{(1)} = \frac{1}{1+e^{-m_{i,j}^{(0)}}} \qquad \left(1 \le i \le L, 1 \le j \le 20\right) \tag{8}$$

In proteins, the number of amino acids ($L$) is not the same, hence PSSM could not be directly used in machine learning. In order to convert variable size $L \times 20$ dimension PSSM into fixed size dimension input vector used in automatic predictor, two simple strategies were usually adopted: (i) all rows of **Eq.7** were summed to form a 20-D vector and then divide by $L$; (ii) all rows of **Eq.7** belonging to the same amino acid were pooled together to form 20 matrices of size $N_{AA} \times 20$, where $N_{\mathrm{AA}}$ is the number of amino acid types. So we get $20 \times 20 = 400$ dimension vector to formulate **Eq.2** [112, 113].

Above two formulations include too simple statistics information. Wu *et al.* [38, 41, 42, 100] proposed "SeqEvo" formulation as descried below.

Step 1. Use the elements in **Eq.7** to define a new matrix **M** as formulated by

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,20} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ m_{L,1} & m_{L,2} & \cdots & m_{L,20} \end{bmatrix} \tag{9}$$

with

$$m_{i,j} = \frac{m_{i,j}^{(0)} - m_j^{(0)}}{\mathrm{SD}(m_j^{(0)})} \qquad (i = 1,2,...,L;\ j = 1,2,...,20) \tag{10}$$

where

$$m_j^{(0)} = \frac{1}{L}\sum_{i=1}^{L} m_{i,j}^{(0)} \qquad (j = 1,2,...,20) \tag{11}$$

is the mean for $m_{i,j}^{(0)}\ (i = 1,2,...,L)$ and

$$\mathrm{SD}\left(m_j^{(0)}\right) = \sqrt{\sum_{i=1}^{L}\left|m_{i,j}^{(0)} - m_j^{(0)}\right|^2 / L} \tag{12}$$

is the corresponding standard deviation.

Step 2. Introduce a new matrix generated by multiplying **M** with its own transpose matrix $\mathbf{M}^{\mathrm{T}}$; i.e.,

$$\mathbf{M}^{\mathrm{T}}\mathbf{M} = \begin{bmatrix} \sum_{i=1}^{L} m_{i,1}m_{i,1} & \sum_{i=1}^{L} m_{i,1}m_{i,2} & \cdots & \sum_{i=1}^{L} m_{i,1}m_{i,20} \\ \sum_{i=1}^{L} m_{i,2}m_{i,1} & \sum_{i=1}^{L} m_{i,2}m_{i,2} & \cdots & \sum_{i=1}^{L} m_{i,2}m_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{L} m_{i,20}m_{i,1} & \sum_{i=1}^{L} m_{i,20}m_{i,2} & \cdots & \sum_{i=1}^{L} m_{i,20}m_{i,20} \end{bmatrix} \tag{13}$$

which contains $20 \times 20 = 400$ elements. Since **Eq.13** is a symmetric matrix, we only need the information of its 210 elements, of which 20 are the diagonal elements and $(400-20)/2-190$ are the lower triangular elements, to formulate the protein **P**; i.e., now for **Eq.2**, we have $\Omega = 210$ and that the components $\psi_u\,(u = 1,2,...,210)$ are respectively taken from the 210 diagonal and lower triangular elements of **Eq.13** by following a given order, say from left to right and from the 1st row to the last as illustrated by the following equation

$$\begin{bmatrix} (1) & & & & \\ (2) & (3) & & & \\ (4) & (5) & (6) & & \\ \vdots & \vdots & \vdots & \ddots & \\ (191) & (192) & (193) & \cdots & (210) \end{bmatrix} \tag{14}$$

where the numbers in parentheses indicate the order of elements taken from **Eq.13**.

In the above PSSM models, however, only the statistical information of PSSM was utilized but the inner interactions among the constituent amino acid residues in a protein sample, or its sequence-order effects, were ignored. To avoid

completely lose the sequence-order information associated with PSSM, Lin *et al.* [114] introduced the Grey-PSSM by extracting the useful information from **Eq.7** to define the components of **Eq.2**. Using the grey system theory [115], the following information from the *j*-th column of Eq.10 can be extracted

$$\begin{bmatrix} a_1^j \\ a_2^j \\ b^j \end{bmatrix} = \left(\mathbf{B}_j^T \mathbf{B}_j\right)^{-1} \mathbf{B}_j^T \mathbf{U}_j \quad (j = 1,2,\cdots,20) \tag{15}$$

where

$$\mathbf{B}_j = \begin{bmatrix} -m_{2,\mathbf{j}}^{(1)} & -m_{1,\mathbf{j}}^{(1)} - 0.5m_{2,\mathbf{j}}^{(1)} & 1 \\ -m_{3,\mathbf{j}}^{(1)} & -\sum_{i=1}^{2} m_{\mathbf{i},\mathbf{j}}^{(1)} - 0.5m_{3,\mathbf{j}}^{(1)} & 1 \\ \vdots & \vdots & \vdots \\ -m_{k,\mathbf{j}}^{(1)} & -\sum_{i=1}^{k-1} m_{\mathbf{i},\mathbf{j}}^{(1)} - 0.5m_{\mathbf{k},\mathbf{j}}^{(1)} & 1 \\ \vdots & \vdots & \vdots \\ -m_{\mathbf{L},\mathbf{j}}^{(1)} & -\sum_{i=1}^{L-1} m_{\mathbf{i},\mathbf{j}}^{(1)} - 0.5m_{\mathbf{L},\mathbf{j}}^{(1)} & 1 \end{bmatrix} \tag{16}$$

and

$$\mathbf{U}_\mathbf{j} = \begin{bmatrix} m_{\mathbf{2,j}}^{(1)} - m_{\mathbf{1,j}}^{(1)} \\ m_{\mathbf{3,j}}^{(1)} - m_{\mathbf{2,j}}^{(1)} \\ \vdots \\ m_{\mathbf{k,j}}^{(1)} - m_{\mathbf{k\text{-}1,j}}^{(1)} \\ \vdots \\ m_{\mathbf{L,j}}^{(1)} - m_{\mathbf{L\text{-}1,j}}^{(1)} \end{bmatrix} \tag{17}$$

Therefore, when using the Grey-PSSM approach, the value of $\Omega$ in **Eq.2** would equal to 60 and the components $\psi_u\ (u = 1,2,...,60)$ were formulated by:

$$\begin{cases} \psi_{\mathrm{3j\text{-}2}}^{\mathrm{E}} = a_1^{\mathrm{j}} f_{\mathrm{j}} w_1 \\ \psi_{\mathrm{3j\text{-}1}}^{\mathrm{E}} = a_2^{\mathrm{j}} f_{\mathrm{j}} w_2 \quad (\mathrm{j} = 1,2,\cdots,20) \\ \psi_{\mathrm{3j}}^{\mathrm{E}} = b^{\mathrm{j}} f_{\mathrm{j}} w_3 \end{cases} \tag{18}$$

where $w_1, w_2$ and $w_3$ are the weight factor; $f_j\ (j = 1,2,\cdots,20)$ were the occurrence frequencies of the 20 amino acids.

The rationale for introducing the grey model as done above is that we only know the score of each amino acid resides in protein sequence beings changed to other amino acid resides during evolutionary process, but we do not understand the intrinsic information of overall sequence evolutionary. So the protein sequence evolution information could be viewed as a "grey system". Grey model is particularly useful to deal with this kind of grey systems. Grey model is built on an Accumulated Generating Operation, which could reduce the stochastic noise of raw series. Hence, the aforementioned Grey-PSSM formulation could more effectively incorporate the protein sequence evolution information than the simple statistical approaches as reflected by remarkably enhancing the success rates thus achieved.

### 3.3. Cellular Automation (CA) Image

Protein sequences stored in databases are often strings of characters, and how to read or compare them is one of the basic problems we are often encountered with. It would act like a snail's pace for human beings to read these sequences with the naked eyes. Also, it is very hard to extract any key features by directly reading these sequences. Visualization may be a good choice. By encoding a protein sequence into digital format with genetic and physical chemistry information, followed by using cellular automaton to evolve a 2-dimensional image by taking into account the interaction between amino acids, many important features, which are originally hidden in the bimolecular sequence, can be clearly revealed thru its cellular automaton image as demonstrated Xiao *et al.* [35, 116]. According to Wolfram's theory, each protein sequence is corresponding to a cellular automaton image with its own textural feature. Accordingly, those proteins that belong to a same attribute must have some similar textures in their cellular automaton images [35]. Thus, the features extracted from their cellular automaton images can be used to cluster or distinguish various attributes of proteins.

It is instructive to point out that in most common visual methods, the point of the special curve corresponding to a certain amino acid was colligated only with the residues prior to it, while the effects of all the residues behind it were totally ignored. This is inconsistent with the real world that all the residues in a protein are coupled with each other as an entity in nature. In the aforementioned cellular automata image approach, however, the residues in a protein were coupled with each other as an entity. In the process of producing the protein image, the state of cell corresponding to a certain amino acid was colligated with resides both prior to and behind it. Accordingly, the cellular automata image approach could find some implicit sequence features, and these features were difficult to be found by other gene visualizations. The GLCM factors extracted from CA images of proteins could more effectively reflect their overall sequence patterns so as to enhance the power of the corresponding predictor.

### 3.4. Grey-PseAAC (Pseudo Amino Acid Composition)

Lin *et al.* [48] introduced a new and simple PseAAC model, the so-called Grey-PseAAC. In that model, they firstly converted a protein **P** to a series of real numbers ac-

cording to (Table **1**) of [48]: $\begin{pmatrix} x_1 & x_2 & \cdots & x_N \end{pmatrix}$ where $N$ is the length of **P.** Secondly, they extracted the grey system information $\begin{bmatrix} a_1 & a_2 & b \end{bmatrix}$ according to the following equation

$$\begin{bmatrix} a_1 \\ a_2 \\ b \end{bmatrix} = \left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{U} \tag{19}$$

where

$$\mathbf{B} = \begin{bmatrix} -x_2 & -x_1 - 0.5x_2 & 1 \\ -x_3 & -\sum_{i=1}^{2} x_i - 0.5x_3 & 1 \\ \vdots & \vdots & \vdots \\ -x_k & -\sum_{i=1}^{k-1} x_i - 0.5x_k & 1 \\ \vdots & \vdots & \vdots \\ -x_N & -\sum_{i=1}^{N-1} x_i - 0.5x_N & 1 \end{bmatrix} \tag{20}$$

and

$$\mathbf{U} = \begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_k - x_{k-1} \\ \vdots \\ x_N - x_{N-1} \end{bmatrix} \tag{21}$$

Thus, the protein **P** can be formulated by **Eq.2** with

$$\psi_i = \begin{cases} f_i & (1 \le i \le 20) \\ |a_1| & i = 21 \\ |a_2| & i = 22 \\ |b| & i = 23 \end{cases} \tag{22}$$

where $f_i$ (i=1,2, …, 20) are the normalized occurrence frequencies of the 20 native amino acids in the protein **P**.

### 3.5. Web-Servers for PseAAC Generators

Besides the above approaches, there are many other PseAAC modes to formulate the protein sample **P**. Among them are the Discrete Wavelet Transform (DWT) for PseAAC [118], Function Domain PseAAC, Dipeptide Composition PseAAC, Low-Frequency Fourier Spectrum Mode [119, 120], and protein-protein interaction (PPI) [48, 121, 122]. It is instructive to point out that, because PseAAC has been widely and increasingly used, recently two powerful soft-wares, called 'PseAAC-Builder' [123] and 'propy' [124], were established for generating various special PseAAC components, in addition to the web-server 'PseAAC' [125] built in 2008.

## 4. PREDICTION ALGORITHMS

Once the formulation for protein samples are defined, there are many well-known algorithms to operate the prediction, such as the Covariance Discriminant (CA) [10, 11, 126, 127], Nearest Neighbor (NN) [106, 128], Artificial Neural Network (ANN) [129, 130], support vector machine (SVM) [33, 60, 110, 131-139], K-Nearest neighbor [74, 140], GIA-Nearest neighbor [141], Adaptive K-nearest neighbor, and Fuzzy K-nearest neighbor. Below, let us focus on some other algorithms.

### 4.1. Random Forest Algorithms

The Random Forest (RF) algorithm [142-144] is a popular machine learning algorithm and recently it has been successfully employed in dealing with various biological prediction problems [48, 109, 133, 145-147]. RF builds many tree predictors on the values of a random vector sampled independently and they have the same distribution. Subsequently, RF integrates those tree predictors. It has been shown that combining multiple trees produced in randomly selected subspaces can significantly improve the prediction accuracy. RF performs a type of cross-validation by using out-of-bag samples. For more detailed information about the RF algorithm, refer to the web-page at http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, where the code of RF for FORTRAN 77 can be download. Besides, the RF software package for MATLAB is available at http://code.google.com/p/randomforest-matlab/. When using these RF tool, the user is not required to know much knowledge about RF but only the two important functions: one is "classRF_train" for training given data and returning the prediction model; the other is "classRF_predict" for predicting query input with the prediction model [117].

### 4.2. Multi-label Predicting Algorithms

Most of existing predictors were developed to deal with single-label systems. For example, in dealing with protein subcellular location, it was assumed that each protein in the system concerned had one and only one subcellular location and hence needed only a single label to annotate it. However, increasing evidences have indicated that many proteins in a cell are actually formed a multi-label system in which some of them may have two or more subcellular location sites or attributes, and hence were called "multiplex proteins" needing two or more labels to annotate them [32].

To deal with multi-label systems, various multi-label machine learning algorithms were introduced [38-41, 71, 99-101, 148]. Below, let us introduce the Multi-Label KNN classifier.

### 4.3. Maximum Relevance Minimum Redundancy Algorithms

In classification prediction, the feature selection is an important procedure for enhancing the performance of the

classifier. In this regard, the maximum relevance minimum redundancy (mRMR) method was used to select the optimal features in the protein feature space [133, 149-157] to enhance the prediction quality. For more detailed information about the mRMR approach, refer to the aforementioned papers and the references cited therein.

## 5. WEB SERVERS

A user-friendly and publicly accessible web-server represents the future direction for developing practically more useful models [158, 159]. In this section, let us summarize the relevant web-server available.

### 5.1. GPCR-MPredictor [33]

GPCR-MPredictor is freely available at http://111.68.99.218/gpcr-mpredictor/. It can efficiently predict GPCRs at five levels. The first level determines whether a protein sequence is a GPCR or a non-GPCR. If the predicted sequence is a GPCR, then it is further classified into family, subfamily, sub-subfamily, and subtype levels.

### 5.2. iLoc-Gpos [38]

iLoc-Gpos is freely available at http://www.jci-bioinfo.cn/iLoc-Gpos. It was developed for predicting the subcellular localization of Gram positive bacterial proteins with both single-location and multiple-location sites.

### 5.3. iLoc-Virus [39]

iLoc-Virus is freely accessible to the public at http://www.jci-bioinfo.cn/iLoc-Virus. It hybridized the gene ontology information with the sequential evolution information. It can be utilized to identify viral proteins among the following six locations: (1) viral capsid, (2) host cell membrane, (3) host endoplasmic reticulum, (4) host cytoplasm, (5) host nucleus, and (6) secreted. The iLoc-Virus predictor not only can more accurately predict the location sites of viral proteins in a host cell, but also have the capacity to deal with virus proteins having more than one location.

### 5.4. iLoc-Hum [100]

iLoc-Hum is freely accessible to the public at http://www.jci-bioinfo.cn/iLoc-Hum , It was developed for identifying the subcellular localization of human proteins with both single and multiple location sites. It covers the following 14 location sites: centrosome, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracellular, Golgi apparatus, lysosome, microsome, mitochondrion, nucleus, peroxisome, plasma membrane, and synapse, where some proteins belong to two, three or four locations.

### 5.5. iLoc-Gneg [99]

iLoc-Gneg is freely accessible to the public at http://www.jci-bioinfo.cn/iLoc-Gneg. It was developed for predicting the subcellular localization of gram-positive bacterial proteins with both single-location and multiple-location sites. The dataset contains 1,392 gram-negative bacterial proteins classified into the following eight locations: (1) cytoplasm, (2) extracellular, (3) fimbrium, (4) flagellum, (5) inner membrane, (6) nucleoid, (7) outer membrane, and (8) periplasm. Of the 1,392 proteins, 1,328 are each with only one subcellular location and the other 64 are each with two subcellular locations.

### 5.6. iLoc-Euk [41]

iLoc-Euk is freely accessible to the public at the web-site http://www.jci-bioinfo.cn/iLoc-Euk. It works on a benchmark dataset of eukaryotic proteins classified into the following 22 location sites: (1) acrosome, (2) cell membrane, (3) cell wall, (4) centriole, (5) chloroplast, (6) cyanelle, (7) cytoplasm, (8) cytoskeleton, (9) endoplasmic reticulum, (10) endosome, (11) extracellular, (12) Golgi apparatus, (13) hydrogenosome, (14) lysosome, (15) melanosome, (16) microsome (17) mitochondrion, (18) nucleus, (19) peroxisome, (20) spindle pole body, (21) synapse, and (22) vacuole. It is significantly higher than that by any of the existing predictors that also have the capacity to deal with such a complicated and stringent system.

### 5.7. iLoc-Plant [40]

iLoc-Plant is freely accessible to the public at the web-site http://www.jci-bioinfo.cn/iLoc-Plant. It works on a benchmark dataset of plant proteins classified into the following 12 location sites: (1) cell membrane, (2) cell wall, (3) chloroplast, (4) cytoplasm, (5) endoplasmic reticulum, (6) extracellular, (7) Golgi apparatus, (8) mitochondrion, (9) nucleus, (10) peroxisome, (11) plastid, and (12) vacuole, where some proteins belong to two or three locations.

### 5.8. iLoc-Animal [160]

iLoc-Animal is freely accessible to the public at the web-site http://www.jci-bioinfo.cn/iLoc-Animal. It can be used to identified the subcellular locations of animal proteins among following 20 location sites: (1) acrosome, (2) cell membrane, (3) centriole, (4) centrosome, (5) cell cortex, (6) cytoplasm, (7) cytoskeleton, (8) endoplasmic reticulum, (9) endosome, (10) extracellular, (11) Golgi apparatus, (12) lysosome, (13) mitochondrion, (14) melanosome, (15) microsome, (16) nucleus, (17) peroxisome, (18) plasma membrane, (19) spindle, and (20) synapse for both single-label and multi-label cases.

### 5.9. ngLOC [161]

ngLOC web server is accessible at http://ngloc.unmc.edu. ngLOC is an n-gram-based Bayesian classifier that predicts subcellular localization of proteins both in prokaryotes and eukaryotes. This program can predict 11 distinct locations each in plant and animal species. ngLOC also predicts 4 and 5 distinct locations on gram-positive and gram-negative bacterial datasets, respectively.

### 5.10. iSMP-Grey [114]

iSMP-Grey is freely accessible to the public at http://www.jci-bioinfo.cn/iSMP-Grey. It can be used to identify the secretory proteins of malaria parasite based on the protein sequence information alone.

### 5.11. DR_bind [162]

DR_bind server is freely available at http://dnasite.limlab.ibms.sinica.edu.tw. It is a web server that automatically predicts DNA-binding residues, given the respective protein structure based on (i) electrostatics, (ii) evolution and (iii) geometry.

### 5.12. iDNA-Prot [48]

iDNA-Prot is freely accessible to the public at the website on http://www.jci-bioinfo.cn/iDNA-Prot. By incorporating the features into the general form of pseudo amino acid composition that were extracted from protein sequences via the "grey model" and by adopting the random forest operation engine, iDNA-Prot can identify uncharacterized proteins as DNA-binding proteins or non-DNA binding proteins based on their amino acid sequences information alone.

### 5.13. Swfoldrate [163]

The prediction server is freely available at http://www.jci-bioinfo.cn/swfrate/input.jsp. The predictor was achieved on the basis of multitudinous physicochemical features and statistical features from protein using nonlinear support vector machine (SVM) regression model, the method obtained an excellent agreement between predicted and experimentally observed folding rates of proteins.

### 5.14. iNR-PhysChem [139]

PhysChem is freely accessible to the public at either http://www.jci-bioinfo.cn/iNR-PhysChem. iNR-PhysChem introduced a novel mode of pseudo amino acid composition (PseAAC) whose components were derived from a physical-chemical matrix via a series of auto-covariance and cross-covariance transformations. It was observed that the overall success rate achieved by iNR-PhysChem was over 98% in identifying NRs or non-NRs, and over 92% in identifying NRs among the following seven subfamilies: NR1--thyroid hormone like, NR2--HNF4-like, NR3--estrogen like, NR4--nerve growth factor IB-like, NR5--fushi tarazu-F1 like, NR6--germ cell nuclear factor like, and NR0--knirps like.

### 5.15. NR-2L [164]

NR-2L is freely accessible at http://www.jci-bioinfo.cn/NR2L. It is a two-level predictor, which was developed that can be used to identify a query protein as a nuclear receptor or not based on its sequence information alone; if it is, the prediction will be automatically continued to further identify it among the following seven subfamilies: (1) thyroid hormone like (NR1), (2) HNF4-like (NR2), (3) estrogen like, (4) nerve growth factor IB-like (NR4), (5) fushi tarazu-F1 like (NR5), (6) germ cell nuclear factor like (NR6), and (7) knirps like (NR0). The identification was made by the Fuzzy K nearest neighbor (FK-NN) classifier based on the pseudo amino acid composition formed by incorporating various physicochemical and statistical features derived from the protein sequences, such as amino acid composition, dipeptide composition, complexity factor, and low-frequency Fourier spectrum components.

For reader's convenience, a brief description for each of the web servers described above are listed in (Table **2**).

**Table 2. List of the 15 Servers Introduced in this Paper As well As their Website Addresses and Targets.**

| No | Name | Website Address | Target |
|---|---|---|---|
| 1 | **GPCR-MPredictor** | http://111.68.99.218/gpcr-mpredictor/ | GPCRs family |
| 2 | **iLoc-Gpos** | http://www.jci-bioinfo.cn/iLoc-Gpos | Subcellular location |
| 3 | **iLoc-Virus** | http://www.jci-bioinfo.cn/iLoc-Virus | Subcellular location |
| 4 | **iLoc-Hum** | http://www.jci-bioinfo.cn/iLoc-Hum | Subcellular location |
| 5 | **iLoc-Gneg** | http://www.jci-bioinfo.cn/iLoc-Gneg | Subcellular location |
| 6 | **iLoc-Euk** | http://www.jci-bioinfo.cn/iLoc-Euk | Subcellular location |
| 7 | **iLoc-Plant** | http://www.jci-bioinfo.cn/iLoc-Plant | Subcellular location |
| 8 | **iLoc-Animal** | http://www.jci-bioinfo.cn/iLoc-Animal | Subcellular location |
| 9 | **ngLOC** | http://ngloc.unmc.edu | Subcellular location |
| 10 | **iSMP-Grey** | http://www.jci-bioinfo.cn/iSMP-Grey | Secretory proteins of malaria parasite |
| 11 | **DR_bind** | http://dnasite.limlab.ibms.sinica.edu.tw | DNA-binding residues |
| 12 | **iDNA-Prot** | http://www.jci-bioinfo.cn/iDNA-Prot | DNA-binding protein |
| 13 | **Swfoldrate** | http://www.jci-bioinfo.cn/swfrate/input.jsp | Protein folding |
| 14 | **iNR-PhysChem** | http://www.jci-bioinfo.cn/iNR-PhysChem | Nuclear receptors and their subfamilies |
| 15 | **NR-2L** | http://www.jci-bioinfo.cn/NR2L | Nuclear receptors and their subfamilies |

## 6. CONCLUSION AND PERSPECTIVES

As summarized in a review article [85], to develop a useful statistical predictor, one needs to consider the following steps: (i) construct or select a proper benchmark dataset to train and test the predictor; (ii) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm to operate the prediction; (iv) properly perform cross-validation tests to objectively measure the performance of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Measured against these five criteria, significant progresses have been achieved during the past decade in predicting protein classification, as reflected by the following facts.

i) The quality of benchmark datasets used for training and testing the predictors have been remarkably improved from the following three angles. The first is that their cut-off thresholds have become more stringent; the second is their coverage has become wider; and the third is more multi-label benchmark datasets have been constructed.

ii) More important and useful informations have been incorporated into PseAAC to formulate the protein samples via various effective approaches, such as gene ontology, functional domain, sequence evolution, and grey model.

iii) The algorithms for operating the prediction systems have become more powerful. Particularly, the algorithms for dealing with multi-label systems or multiplex proteins have been decently established that even did not exist 10 years ago.

iv) The metrics to measure the performance of predictors have been considerably developed by introducing the "absolute true" rate, which is a very intuitive and easy-to-understand measurement in studying multi-label systems.

v) Many web servers have been established to help experimental biologists easily to get their desired information without the need to follow the complicated mathematics.

Further efforts in this area should be focused on multiplex proteins because they may have some unique or special functions important for both basic research and drug development. Actually, as indicated by a recent review [32], many biomedical systems belong to the multi-label systems in which each of their constituent molecules possesses one or more than one function or feature, and hence needs one or more than one label to indicate its attribute(s).

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

## ACKNOWLEDGEMENTS

This work was supported by the grants from the National Natural Science Foundation of China (No. 31260273, 61261027), the Key Project of Chinese Ministry of Education (No. 210116), and the Department of Education of Jiang-Xi Province (No. GJJ11557, GJJ12490), and the Jiangxi Provincial Foundation for Leaders of Disciplines in Science (No. 20113BCB22008, No. 20114BAB211013, NO. 20122BAB201020), and the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No. 20121BDH80023). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## REFERENCES

[1] Teitler, M.; Klein, M. T. A new approach for studying GPCR dimers: drug-induced inactivation and reactivation to reveal GPCR dimer function in vitro, in primary culture, and in vivo. *Pharmacology & therapeutics*, **2012**, *133*, 205-217.

[2] Chou, K. C. Prediction of G-protein-coupled receptor classes. *Journal of Proteome Research*, **2005**, *4*, 1413-1418.

[3] Nakashima, H.; Nishikawa, K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol*, **1994**, *238*, 54-61.

[4] Cedano, J.; Aloy, P.; P'erez-Pons, J. A.; Querol, E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol*, **1997**, *266*, 594-600.

[5] Fan, G. L.; Li, Q. Z. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2012**, *304*, 88-95.

[6] Huang, C.; Yuan, J. 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*, **2013**, *113*, 50-57.

[7] Wan, S.; Mak, M. W.; Kung, S. Y. 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.*, **2013**, *323*, 40-48.

[8] Chou, K. C.; Elrod, D. W. Protein subcellular location prediction. *Protein Eng.*, **1999**, *12*, 107-118.

[9] Nakai, K. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.*, **2000**, *54*, 277-344.

[10] Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins: Struct., Funct., Genet.*, **2003**, *50*, 44-48.

[11] Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.*, **2003**, *22*, 395-402.

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

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

[14] Chou, K. C.; Shen, H. B. 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; doi:10.4236/ns.2010.210136). *Nature Protocols*, **2008**, *3*, 153-162.

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

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

[17] Cao, J. Z.; Liu, W. Q.; Gu, H. Predicting Viral Protein Subcellular Localization with Chou's Pseudo Amino Acid Composition and Imbalance-Weighted Multi-Label K-Nearest Neighbor Algorithm. *Protein and Peptide Letters*, **2012**, *19*, 1163-1169.

[18] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.*, **2010**, *263*, 203-209.

[19] Xiao, X.; Wu, Z. C.; Chou, K. C. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.*, **2011**, *284*, 42-51.

[20] Hung, M. C.; Link, W. Protein localization in disease and therapy. *J. Cell Sci.*, **2011**, *124*, 3381-3392.

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

[22] Zhao, X. W.; Li, X. T.; Ma, Z. Q.; Yin, M. H. Identify DNA-Binding Proteins with Optimal Chou's Amino Acid Composition. *Protein & Peptide Letters*, **2012**, *19*, 398-405.

[23] Schnell, J. R.; Chou, J. J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature*, **2008**, *451*, 591-595.

[24] OuYang, B.; Xie, S.; Berardi, M. J.; Zhao, X. M.; Dev, J.; Yu, W.; Sun, B.; Chou, J. J. Unusual architecture of the p7 channel fromhepatitis C virus *Nature* **2013** *498*, 521-525.

[25] Berardi, M. J.; Shih, W. M.; Harrison, S. C.; Chou, J. J. Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. *Nature*, **2011**, *476*, 109-113.

[26] Chou, J. J.; Li, S.; Klee, C. B.; Bax, A. Solution structure of Ca2+-calmodulin reveals flexible hand-like properties of its domains. *Nature Structural Biology*, **2001**, *8*, 990-997.

[27] Wang, J.; Pielak, R. M.; McClintock, M. A.; Chou, J. J. Solution structure and functional analysis of the influenza B proton channel. *Nature Structural and Molecular Biology*, **2009**, *16*, 1267-1271.

[28] Chou, K. C. Insights from modelling three-dimensional structures of the human potassium and sodium channels. *Journal of Proteome Research*, **2004**, *3*, 856-861.

[29] Chou, K. C. Insights from modeling the 3D structure of DNA-CBF3b complex. *Journal of Proteome Research*, **2005**, *4*, 1657-1660.

[30] Chou, K. C. Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. *Journal of Proteome Research*, **2005**, *4*, 1681-1686.

[31] Chou, K. C. Review: Structural bioinformatics and its impact to biomedical science. *Current Medicinal Chemistry*, **2004**, *11*, 2105-2134.

[32] Chou, K. C. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Molecular Biosystems*, **2013**, *9*, 1092-1100.

[33] Naveed, M.; Khan, A. GPCR-MPredictor: multi-level prediction of G protein-coupled receptors using genetic ensemble. *Amino Acids*, **2012**, *42*, 1809-1823.

[34] Xiao, X.; Wang, P.; Chou, K. C. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Molecular bioSystems*, **2011**, *7*, 911-919.

[35] Xiao, X.; Wang, P.; Chou, K. C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *Journal of computational chemistry*, **2009**, *30*, 1414-1423.

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

[37] Goldfeld, D. A.; Zhu, K.; Beuming, T.; Friesner, R. A. Loop prediction for a GPCR homology model: algorithms and results. *Proteins*, **2013**, *81*, 214-228.

[38] Wu, Z. C.; Xiao, X.; Chou, K. C. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins. *Protein Pept Lett*, **2012**, *19*, 4-14.

[39] Xiao, X.; Wu, Z. C.; Chou, K. C. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J Theor Biol*, **2011**, *284*, 42-51.

[40] Wu, Z. C.; Xiao, X.; Chou, K. C. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Molecular bioSystems*, **2011**, *7*, 3287-3297.

[41] Chou, K. C.; Wu, Z. C.; Xiao, X. iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. *PLoS One*, **2011**, *6*, e18258.

[42] Wu, Z. C.; Xiao, X.; Chou, K. C. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J Theor Biol*, **2010**, *267*, 29-34.

[43] Wang, X.; Li, G. Z.; Lu, W. C. 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*, **2013**, *20*, 309-317.

[44] 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*, **2012**, *7*, e37155.

[45] Chen, C. Y.; Chien, T. Y.; Lin, C. K.; Lin, C. W.; Weng, Y. Z.; Chang, D. T. Predicting target DNA sequences of DNA-binding proteins based on unbound structures. *PLoS ONE*, **2012**, *7*, e30446.

[46] van Dyke, M. W. REPSA: combinatorial approach for identifying preferred drug-DNA binding sequences. *Methods in molecular biology*, **2010**, *613*, 193-205.

[47] Klein, P.; Dietz, K. J. Identification of DNA-binding proteins and protein-protein interactions by yeast one-hybrid and yeast two-hybrid screen. *Methods in molecular biology*, **2010**, *639*, 171-192.

[48] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PLoS ONE*, **2011**, *6*, e24756.

[49] Scott, M. S.; Thomas, D. Y.; Hallett, M. T. Predicting subcellular localization via protein motif co-occurrence. *Genome Res*, **2004**, *14*, 1957-1966.

[50] Scott, M. S.; Calafell, S. J.; Thomas, D. Y.; Hallett, M. T. Refining protein subcellular localization. *Plos Comput Biol*, **2005**, *1*, 518-528.

[51] Mizuno, Y.; Kurochkin, I. V.; Herberth, M.; Okazaki, Y.; Schonbach, C. Predicted mouse peroxisome-targeted proteins and their actual subcellular locations. *Bmc Bioinformatics*, **2008**, *9*.

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

[53] Briesemeister, S.; Blum, T.; Brady, S.; Lam, Y.; Kohlbacher, O.; Shatkay, H. SherLoc2: A High-Accuracy Hybrid Method for Predicting Subcellular Localization of Proteins. *J Proteome Res*, **2009**, *8*, 5363-5366.

[54] Brustolini, O. J. B.; Fietto, L. G.; Cruz, C. D.; Passos, F. M. L. Computational analysis of the interaction between transcription factors and the predicted secreted proteome of the yeast Kluyveromyces lactis. *Bmc Bioinformatics*, **2009**, *10*.

[55] Mintz-Oron, S.; Aharoni, A.; Ruppin, E.; Shlomi, T. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics*, **2009**, *25*, I247-I252.

[56] Tung, T. Q.; Lee, D. A method to improve protein subcellular localization prediction by integrating various biological data sources. *Bmc Bioinformatics*, **2009**, *10*.

[57] Zhu, L.; Yang, J.; Shen, H. B. Multi Label Learning for Prediction of Human Protein Subcellular Localizations. *Protein J*, **2009**, *28*, 384-390.

[58] Briesemeister, S.; Rahnenfuhrer, J.; Kohlbacher, O. YLoc--an interpretable web server for predicting subcellular localization. *Nucleic acids research*, **2010**, *38*, W497-502.

[59] Briesemeister, S.; Rahnenfuhrer, J.; Kohlbacher, O. Going from where to why-interpretable prediction of protein subcellular localization. *Bioinformatics*, **2010**, *26*, 1232-1238.

[60] Chou, K. C.; Cai, Y. D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **2002**, *277*, 45765-45769.

[61] Gu, Q.; Ding, Y. S.; Jiang, X. Y.; Zhang, T. L. Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection. *Amino Acids*, **2010**, *38*, 975-983.

[62] Imai, K.; Nakai, K. Prediction of subcellular locations of proteins: Where to proceed? *Proteomics*, **2010**, *10*, 3970-3983.

[63] Kandaswamy, K. K.; Pugalenthi, G.; Moller, S.; Hartmann, E.; Kalies, K. U.; Suganthan, P. N.; Martinetz, T. Prediction of Apoptosis Protein Locations with Genetic Algorithms and Support Vector Machines Through a New Mode of Pseudo Amino Acid Composition. *Protein Peptide Lett*, **2010**, *17*, 1473-1479.

[64] Khan, A.; Majid, A.; Choi, T. S. Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers. *Amino Acids*, **2010**, *38*, 347-350.

[65] Qiu, J. D.; Luo, S. H.; Huang, J. H.; Sun, X. Y.; Liang, R. P. Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine. *Amino Acids*, **2010**, *38*, 1201-1208.

[66] Sun, C. L.; Zhao, X. M.; Tang, W. H.; Chen, L. N. FGsub: Fusarium graminearum protein subcellular localizations predicted from primary structures. *Bmc Syst Biol*, **2010**, *4*.

[67] Du, P. F.; Li, T. T.; Wang, X. Recent progress in predicting protein sub-subcellular locations. *Expert Rev Proteomic*, **2011**, *8*, 391-404.

[68] Khan, A.; Majid, A.; Hayat, M. CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. *Comput Biol Chem*, **2011**, *35*, 218-229.

[69] Shi, S. P.; Qiu, J. D.; Sun, X. Y.; Huang, J. H.; Huang, S. Y.; Suo, S. B.; Liang, R. P.; Zhang, L. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: Approach from the strategy of discrete wavelet transform feature extraction. *Bba-Mol Cell Res*, **2011**, *1813*, 424-430.

[70] Li, L. Q.; Zhang, Y.; Zou, L. Y.; Zhou, Y.; Zheng, X. Q. Prediction of Protein Subcellular Multi-Localization Based on the General form of Chou's Pseudo Amino Acid Composition. *Protein Peptide Lett*, **2012**, *19*, 375-387.

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

[72] Tejedor-Estrada, R.; Nonell, S.; Teixido, J.; Sagrista, M. L.; Mora, M.; Villanueva, A.; Canete, M.; Stockert, J. C. An Artificial Neural Network Model for Predicting the Subcellular Localization of Photosensitisers for Photodynamic Therapy of Solid Tumours. *Curr Med Chem*, **2012**, *19*, 2472-2482.

[73] Yoon, Y.; Lee, G. G. Subcellular Localization Prediction through Boosting Association Rules. *Ieee Acm T Comput Bi*, **2012**, *9*, 609-618.

[74] Chou, K. C.; Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[75] Horn, F.; Bettler, E.; Oliveira, L.; Campagne, F.; Cohen, F. E.; Vriend, G. GPCRDB information system for G protein-coupled receptors. *Nucleic acids research*, **2003**, *31*, 294-297.

[76] Vroling, B.; Sanders, M.; Baakman, C.; Borrmann, A.; Verhoeven, S.; Klomp, J.; Oliveira, L.; de Vlieg, J.; Vriend, G. GPCRDB: information system for G protein-coupled receptors. *Nucleic acids research*, **2011**, *39*, D309-319.

[77] Worth, C. L.; Kreuchwig, A.; Kleinau, G.; Krause, G. GPCR-SSFE: a comprehensive database of G-protein-coupled receptor template predictions and homology models. *Bmc Bioinformatics*, **2011**, *12*, 185.

[78] Tanz, S. K.; Castleden, I.; Hooper, C. M.; Vacher, M.; Small, I.; Millar, H. A. SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic acids research*, **2013**, *41*, D1185-1191.

[79] Lum, G.; Min, X. J. FunSecKB: the Fungal Secretome KnowledgeBase. *Database-Oxford*, **2011**.

[80] Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **2010**, *26*, 680-682.

[81] Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **2012**, *28*, 3150-3152.

[82] Xiao, X.; Wang, P.; Chou, K. C. GPCR-2L: Predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Molecular Biosystems*, **2011**, *7*, 911-919.

[83] Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*, 275-349.

[84] Chou, K. C.; Shen, H. B. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms (doi:10.4236/ns.2010.210136). *Natural Science*, **2010**, *2*, 1090-1103 (openly accessible at http://www.scirp.org/journal/NS/).

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

[86] Chen, C.; Shen, Z. B.; Zou, X. Y. Dual-Layer Wavelet SVM for Predicting Protein Structural Class Via the General Form of Chou's Pseudo Amino Acid Composition. *Protein & Peptide Letters*, **2012**, *19*, 422-429.

[87] Hayat, M.; Khan, A. Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. *Protein & Peptide Letters*, **2012**, *19*, 411-421.

[88] Xu, Y.; Ding, J.; Wu, L. Y.; Chou, K. C. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition *PLoS ONE*, **2013**, *8*, e55844.

[89] Liu, L.; Hu, X. Z.; Liu, X. X.; Wang, Y.; Li, S. B. Predicting Protein Fold Types by the General Form of Chou's Pseudo Amino Acid Composition: Approached from Optimal Feature Extractions. *Protein & Peptide Letters*, **2012**, *19*, 439-449.

[90] Sun, X. Y.; Shi, S. P.; Qiu, J. D.; Suo, S. B.; Huang, S. Y.; Liang, R. P. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Molecular BioSystems*, **2012**, *8*, 3178-3184.

[91] Zhao, X. W.; Ma, Z. Q.; Yin, M. H. Predicting protein-protein interactions by combing various sequence- derived features into the general form of Chou's Pseudo amino acid composition. *Protein & Peptide Letters*, **2012**, *19*, 492-500.

[92] Chen, Y. K.; Li, K. B. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2013**, *318*, 1-12.

[93] Xiao, X.; Wang, P.; Chou, K. C. Recent Progresses in Identifying Nuclear Receptors and Their Families. *Curr Top Med Chem*, **2013**, *13*, 1192-1200.

[94] Shen, H. B.; Yang, J.; Chou, K. C. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, **2007**, *33*, 57-67.

[95] Huang, W. L.; Tung, C. W.; Huang, H. L.; Ho, S. Y. Predicting protein subnuclear localization using GO-amino-acid composition features. *Biosystems*, **2009**, *98*, 73-79.

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

[97] 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*, **2010**, *5*, e9931.

[98] Chou, K. C.; Shen, H. B. Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE*, **2010**, *5*, e11335.

[99] 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*, **2011**, *6*, e20592.

[100] Chou, K. C.; Wu, Z. C.; Xiao, X. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems*, **2012**, *8*, 629-641.

[101] Wang, X.; Li, G. Z. A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins. *PLoS ONE*, **2012**, *7*, e36317.

[102] Mei, S.; Fei, W.; Zhou, S. Gene ontology based transfer learning for protein subcellular localization. *Bmc Bioinformatics*, **2011**, *12*, 44.

[103] Wan, S.; Mak, M. W.; Kung, S. Y. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. *Bmc Bioinformatics*, **2012**, *13*, 290.

[104] Wan, S.; Mak, M. W.; Kung, S. Y. 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*, **2013**, *323C*, 40-48.

[105] Schaffer, A. A.; Aravind, L.; Madden, T. L.; Shavirin, S.; Spouge, J. L.; Wolf, Y. I.; Koonin, E. V.; Altschul, S. F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research*, **2001**, *29*, 2994-3005.

[106] Cai, Y. D.; Chou, K. C. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochem Biophys Res Comm*, **2003**, *305*, 407-411.

[107] Chou, K. C.; Cai, Y. D. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun.*, **2003**, *311*, 743-747.

[108] Chou, K. C. The convergence-divergence duality in lectin domains of the selectin family and its implications. *FEBS Lett.*, **1995**, *363*, 123-126.

[109] Kumar, K. K.; Pugalenthi, G.; Suganthan, P. N. DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. *J Biomol Struct Dyn*, **2009**, *26*, 679-686.

[110] Huang, H. L.; Lin, I. C.; Liou, Y. F.; Tsai, C. T.; Hsu, K. T.; Huang, W. L.; Ho, S. J.; Ho, S. Y. Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. *Bmc Bioinformatics*, **2011**, *12 Suppl 1*, S47.

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

[112] Kumar, M.; Gromiha, M. M.; Raghava, G. P. S. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit*, **2011**, *24*, 303-313.

[113] Mundra, P.; Kumar, M.; Kumar, K. K.; Jayaraman, V. K.; Kulkarni, B. D. Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recogn Lett*, **2007**, *28*, 1610-1615.

[114] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model. *PLoS ONE*, **2012**, *7*, e49040.

[115] Deng, J. L. Introduction to Grey System Theory. *The Journal of Grey System*, **1989**, 1-24.

[116] Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. Using cellular automata to generate image representation for biological sequences. *Amino Acids*, **2005**, *28*, 29-35.

[117] Lin, T. H.; Bar-Joseph, Z.; Murphy, R. F. Learning Cellular Sorting Pathways Using Protein Interactions and Sequence Motifs. *J Comput Biol*, **2011**, *18*, 1709-1722.

[118] Liang, R. P.; Huang, S. Y.; Shi, S. P.; Sun, X. Y.; Suo, S. B.; Qiu, J. D. A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization. *Comput Biol Med*, **2012**, *42*, 180-187.

[119] Xiao, X.; Lin, W. Z.; Chou, K. C. Recent Advances in Predicting G-Protein Coupled Classification. *Current Bioinformatics*, **2012**, *7*, 132-142.

[120] Liu, H.; Yang, J.; Wang, M.; Xue, L.; Chou, K. C. Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. *The Protein Journal*, **2005**, *24*, 385-389.

[121] Shin, C. J.; Wong, S.; Davis, M. J.; Ragan, M. A. Protein-protein interaction as a predictor of subcellular location. *Bmc Syst Biol*, **2009**, *3*.

[122] Kumar, G.; Ranganathan, S. Network analysis of human protein location. *Bmc Bioinformatics*, **2010**, *11*.

[123] 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.*, **2012**, *425*, 117-119.

[124] Cao, D. S.; Xu, Q. S.; Liang, Y. Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **2013**, *29*, 960-962.

[125] Shen, H. B.; Chou, K. C. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **2008**, *373*, 386-388.

[126] Chou, K. C.; Elrod, D. W. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem. Biophys. Res. Commun.*, **1998**, *252*, 63-68.

[127] Chen, W.; Lin, H.; Feng, P. M.; Ding, C.; Zuo, Y. C.; Chou, K. C. iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PLoS ONE*, **2012**, *7*, e47843.

[128] Cai, Y. D.; Chou, K. C. Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics*, **2004**, *20*, 1151-1156.

[129] Cai, Y. D.; Chou, K. C. Artificial neural network for predicting alpha-turn types. *Anal. Biochem.*, **1999**, *268*, 407-409.

[130] Cai, Y. D.; Li, Y. X.; Chou, K. C. Using neural networks for prediction of domain structural classes. *BBA*, **2000**, *1476*, 1-2.

[131] Cai, Y. D.; Zhou, G. P.; Chou, K. C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **2003**, *84*, 3257-3263.

[132] Kaundal, R.; Raghava, G. P. RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. *Proteomics*, **2009**, *9*, 2324-2342.

[133] Li, B. Q.; Hu, L. L.; Chen, L.; Feng, K. Y.; Cai, Y. D.; Chou, K. C. Prediction of Protein Domain with mRMR Feature Selection and Analysis. *PLoS One*, **2012**, *7*, e39308.

[134] Li, L.; Wei, D. Q.; Wang, J. F.; Chou, K. C. SCYPPred: a web-based predictor of SNPs for human cytochrome P450. *Protein Pept Lett*, **2012**, *19*, 57-61.

[135] Lin, K.; Qian, Z. L.; Lu, L.; Lu, L. Y.; Lai, L. H.; Gu, J. Y.; Zeng, Z. B.; Li, H. P.; Cai, Y. D. Predicting miRNA's target from primary structure by the nearest neighbor algorithm. *Mol Divers*, **2010**, *14*, 719-729.

[136] Wang, W.; Geng, X.; Dou, Y.; Liu, T.; Zheng, X. Predicting protein subcellular localization by pseudo amino acid composition with a segment-weighted and features-combined approach. *Protein Pept Lett*, **2011**, *18*, 480-487.

[137] Xiong, Y.; Liu, J.; Wei, D. Q. An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins*, **2011**, *79*, 509-517.

[138] Xu, Q.; Pan, S. J.; Xue, H. H.; Yang, Q. Multitask learning for protein subcellular location prediction. *IEEE/ACM Trans Comput Biol Bioinform*, **2011**, *8*, 748-759.

[139] Xiao, X.; Wang, P.; Chou, K. C. iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. *PLoS ONE*, **2012**, *7*, e30869.

[140] Garg, P.; Sharma, V.; Chaudhari, P.; Roy, N. SubCellProt: predicting protein subcellular localization using machine learning approaches. *In silico biology*, **2009**, *9*, 35-44.

[141] Lin, W. Z.; Xiao, X.; Chou, K. C. GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. *Protein Eng Des Sel*, **2009**, *22*, 699-705.

[142] Rogers, J.; Gunn, S. Identifying feature relevance using a random forest. *Lect Notes Comput Sc*, **2006**, *3940*, 173-184.

[143] Breiman, L. Random forests. *Mach Learn*, **2001**, *45*, 5-32.

[144] Breiman, L. Randomizing outputs to increase prediction accuracy. *Mach Learn*, **2000**, *40*, 229-242.

[145] Nimrod, G.; Schushan, M.; Szilagyi, A.; Leslie, C.; Ben-Tal, N. iDBPs: a web server for the identification of DNA binding proteins. *Bioinformatics*, **2010**, *26*, 692-693.

[146] Nimrod, G.; Szilagyi, A.; Leslie, C.; Ben-Tal, N. Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. *J Mol Biol*, **2009**, *387*, 1040-1053.

[147] Pugalenthi, G.; Kandaswamy, K. K.; Chou, K. C.; Vivekanandan, S.; Kolatkar, P. RSARF: prediction of residue solvent accessibility from protein sequence using random forest method. *Protein Pept Lett*, **2012**, *19*, 50-56.

[148] Cao, J. Z.; Liu, W. Q.; Gu, H. Predicting Viral Protein Subcellular Localization with Chou's Pseudo Amino Acid Composition and Imbalance-Weighted Multi-Label K-Nearest Neighbor Algorithm. *Protein Pept Lett*, **2012**, *19*, 1163-1169.

[149] He, Z.; Zhang, J.; Shi, X. H.; Hu, L. L.; Kong, X.; Cai, Y. D.; Chou, K. C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, **2010**, *5*, e9603.

[150] Huang, T.; Chen, L.; Cai, Y. D.; Chou, K. C. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. *PLoS ONE*, **2011**, *6*, e25297.

[151] Huang, T.; He, Z. S.; Cui, W. R.; Cai, Y. D.; Shi, X. H.; Hu, L. L.; Chou, K. C. A Sequence-based Approach for Predicting Protein Disordered Regions. *Protein and Peptide Letters*, **2013**, *20*, 243-248.

[152] Huang, T.; Shi, X. H.; Wang, P.; He, Z.; Feng, K. Y.; Hu, L.; Kong, X.; Li, Y. X.; Cai, Y. D.; Chou, K. C. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks *PLoS ONE*, **2010**, *5*, e10972.

[153] Li, B. Q.; Hu, L. L.; Niu, S.; Cai, Y. D.; Chou, K. C. Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. *Journal of Proteomics*, **2012**, *75*, 1654-1665.

[154] Li, B. Q.; Huang, T.; Liu, L.; Cai, Y. D.; Chou, K. C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS ONE*, **2012**, *7*, e33393.

[155] Hu, L. L.; Chen, C.; Huang, T.; Cai, Y. D.; Chou, K. C. Predicting biological functions of compounds based on chemical-chemical interactions. *PLoS ONE*, **2011**, *6*, e29491.

[156] Hu, L. L.; Huang, T.; Cai, Y. D.; Chou, K. C. Prediction of Body Fluids where Proteins are Secreted into Based on Protein Interaction Network. *PLoS One*, **2011**, *6*, e22989.

[157] Huang, T.; Wang, J.; Cai, Y. D.; Yu, H.; Chou, K. C. Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. *PLoS ONE*, **2012**, *7*, e34460.

[158] Chou, K. C.; Shen, H. B. Review: recent advances in developing web-servers for predicting protein attributes (doi: 10.4236/ns.2009.12011). *Natural Science*, **2009**, *2*, 63-92 (openly accessible at http://www.scirp.org/journal/NS/)

[159] Lin, S. X.; Lapointe, J. Theoretical and experimental biology in one —A symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. *J. Biomedical Science and Engineering (JBiSE)*, **2013**, *6*, 435-442 (open accessible at http://www.scirp.org/journal/jbise/, doi:410.4236/jbise.2013.64054).

[160] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins *Molecular BioSystems*, **2013**, *9*, 634-644.

[161] King, B. R.; Vural, S.; Pandey, S.; Barteau, A.; Guda, C. ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes. *BMC research notes*, **2012**, *5*, 351.

[162] Chen, Y. C.; Wright, J. D.; Lim, C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. *Nucleic acids research*, **2012**, *40*, W249-256.

[163] Cheng, X.; Xiao, X.; Wu, Z. C.; Wang, P.; Lin, W. Z. Swfoldrate: Predicting protein folding rates from amino acid sequence with sliding window method. *Proteins*, **2013**, *81*, 140-148.

[164] Wang, P.; Xiao, X.; Chou, K. C. NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PLoS ONE*, **2011**, *6*, e23505.

[165] Chou, K. C.; Elrod, D. W. Bioinformatical analysis of G-protein-coupled receptors. *Journal of Proteome Research*, **2002**, *1*, 429-433.

**Received: June 25, 2013** **Revised: June 30, 2013** **Accepted: June 30, 2013**