Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology

ARTICLE *in* CURRENT PROTEOMICS · DECEMBER 2009

Impact Factor: 0.44 · DOI: 10.2174/157016409789973707

1 AUTHOR:

Kuo-Chen Chou
Gordon Life Science Institute

509 PUBLICATIONS **30,060** CITATIONS

SEE PROFILE

Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology

Kuo-Chen Chou*

Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

Abstract: With the avalanche of protein sequences generated in the post-genomic age, it is highly desired to develop automated methods for efficiently identifying various attributes of uncharacterized proteins. This is one of the most important tasks facing us today in bioinformatics, and the information thus obtained will have important impacts on the development of proteomics and system biology. To realize that, one of the keys is to find an effective model to represent the sample of a protein. The most straightforward model in this regard is its entire amino acid sequence; however, the entire sequence model would fail to work when the query protein did not have significant homology to proteins of known characteristics. Thus, various non-sequential models or discrete models were proposed. The simplest discrete model is the amino acid (AA) composition. Using it to represent a protein, however, all the sequence-order information would be completely lost. To cope with such a dilemma, the concept of pseudo amino acid (PseAA) composition was introduced. Its essence is to keep using a discrete model to represent a protein yet without completely losing its sequence-order information. Therefore, in a broad sense, the PseAA composition of a protein is actually a set of discrete numbers that is derived from its amino acid sequence and that is different from the classical AA composition and able to harbour some sort of sequence order or pattern information. Ever since the first PseAA composition was formulated to predict protein subcellular localization and membrane protein types, it has stimulated many different modes of PseAA composition for studying various kinds of problems in proteins and proteins-related systems. In this review, we shall give a brief and systematic introduction of various modes of PseAA composition and their applications. Meanwhile, the challenges for finding the optimal PseAA composition are also briefly discussed**.**

Keywords: Protein attributes, sequential model, discrete model, PseAAC, functional domain, gene ontology, sequential evolution, optimal PseAAC

I. INTRODUCTION

The development of science is always driven or ***inextricably*** linked to a series of nagging questions, such as "Where are we?" "What's next?" and "How can be done?"

1). Where are we?

The accomplishment of human genome sequencing has indicated that we are in the Post Genomic Age. One of the most remarkable features in this age is the explosive increase in volume of ***protein sequence*** data. For instance, in 1986 the Swiss-Prot databank contained merely 3,939 protein sequence entries (Table **1**), but the number has since jumped to 428,650 according to version 57.0 of 24-Mar-2009 (www.ebi.ac.uk/swissprot), meaning that the number of protein sequence entries now is more than 108 times the number from about 23 years ago. The rapid increase in protein sequence entries is also shown by the (Fig. **1**), where a statistical illustration to show the growth of the UniProtKB/TrEMBL Protein Database (http://www.ebi.ac.uk/-uniprot/TrEMBLstats/) is given.

2). What's Next?

In the face of the avalanche of protein sequences, it is self-evident and clear what our next task is. In order to timely use these newly-found protein sequences for basic research and drug development, it is highly desired to determine their structures and functions in a large-scale manner. Thus, a new discipline, called "protein-based genomics" or "proteomics", is emerging. It is actually the science of the cellular protein universe that will hold a key position in the new biology and medicine reshaped by the monumental achievement of the sequencing of human genome. Meanwhile, in order to understand the protein-protein interactions and other complex interactions at a cellular level or in a living system, another discipline, called "system biology", is also emerging. As two rapidly developing frontiers in science, proteomics and system biology are often overlapped with each other. They are far more challenging than the classical molecular biology because they will involve not only very large structural components but also extremely complicated dynamic processes. Nevertheless, they are the area that has to be explored and the area that many believe will become the main stream in biological sciences in this century.

3). How can be Done?

Now, the question is how can the large-scale determination of protein structures and functions be realized? The X-

*Address correspondence to this author at the Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA;
E-mail: kcchou@gordonlifescience.org

1570-1646/09 $55.00+.00

Table 1. The Growth of Protein Sequences in SWISS-Protein Data Bank[a]

Release	Date	Number of Sequence Entries	Number of Amino Acids	Average Length Per Sequence[b]
2.0	09/86	3,939	900,163	229
5.0	09/87	5,205	1,327,683	236
9.0	11/88	8,702	2,498,140	287
12.0	10/89	12,305	3,797,482	309
16.0	11/90	18,364	5,986,949	326
20.0	11/91	22,654	7,500,130	331
24.0	12/92	28,154	9,545,427	339
27.0	10/93	33,329	11,484,420	345
30.0	10/94	40,292	14,147,368	351
32.0	11/95	49,340	17,385,503	352
34.0	10/96	59,021	21,210,389	359
35.0	11/97	69,113	25,083,768	363
37.0	12/98	77,977	28,268,293	363
38.0	07/99	80,000	29,085,965	364
39.0	05/00	86,593	31,411,114	363
40.0	10/01	101,602	37,315,215	367
42.0	10/03	135,850	50,046,799	368
45.0	10/04	163,235	59,631,787	365
48.0	09/05	194,317	70,391,852	362
51.0	10/06	241,242	88,541,632	367
56.0	07/08	392,667	141,217,034	360
57.0	03/09	428,650	154,416,236	360

[a]From http://www.ebi.ac.uk/swissprot/.
[b]The average length per sequence is defined as the total number of amino acids divided by the total number of sequences. The quotient is rounded to an integer.

ray crystallography is indeed a powerful tool in determining 3D (three-dimensional) structures of proteins; unfortunately, it is not feasible in dealing with those proteins difficult to crystallize. Encoded by 20-35% of genes [1], membrane proteins are extremely important to the life of cells, but they are difficult to crystallize and most of them will not dissolve in normal solvents. Therefore, so far very few membrane protein structures (less than 1% of known protein structures) have been determined. Although the recent reports [2-4] have indicated NMR is a very powerful tool in determine the 3D structures of membrane proteins, it is time-consuming and costly. To timely acquire the 3D structural information for these proteins, a feasible approach is by means of structural bioinformatics. Although the structures thus obtained would not be as accurate as those by X-ray and NMR techniques, they can timely provide us with many useful insights for both basic research and drug discovery (see, e.g., [5-18] as well as a review [19]).

Bioinformatics is generally defined as the use of mathematical and informational techniques to solve biological problems by creating or using computer programs, mathematical models or both. One of the remarkable advantages of bioinformatics is its ability in dealing with large-scale objects.

Therefore, bioinformatics and proteomics are often stimulating each other with some sort of overlapping in contents: many important topics in bioinformatics have come from proteomics, while the tools developed in bioinformatics have greatly helped the development of proteomics.

Bioinformatics can be roughly categorized into two branches: sequential bioinformatics and structural bioinformatics. The former is focused on sequences, with the main applications to the data mining, sequence alignment, metabolic networks, morphometrics, and virtual evolution; the latter focused on 3D structures and the related mechanisms. Since the 3D structure of a protein is determined by its sequence, sequence analysis and alignment are also heavily involved in the structural bioinformatics.

The most often used approach in structural bioinformatics is the homology modelling [19, 20] for deriving the 3D structures of proteins. Once the information of 3D structures is obtained, there are varieties of approaches for studying structure-related function and mechanisms from dynamic and other points of view, such as the quasi-continuum model for investigating the low-frequency (or Terahertz frequency) internal collective motion of biomacromolecules and its biological functions (see, e.g., [21-26] and the review articles [27,28]) as well as its medical applications [29,30], molecu-

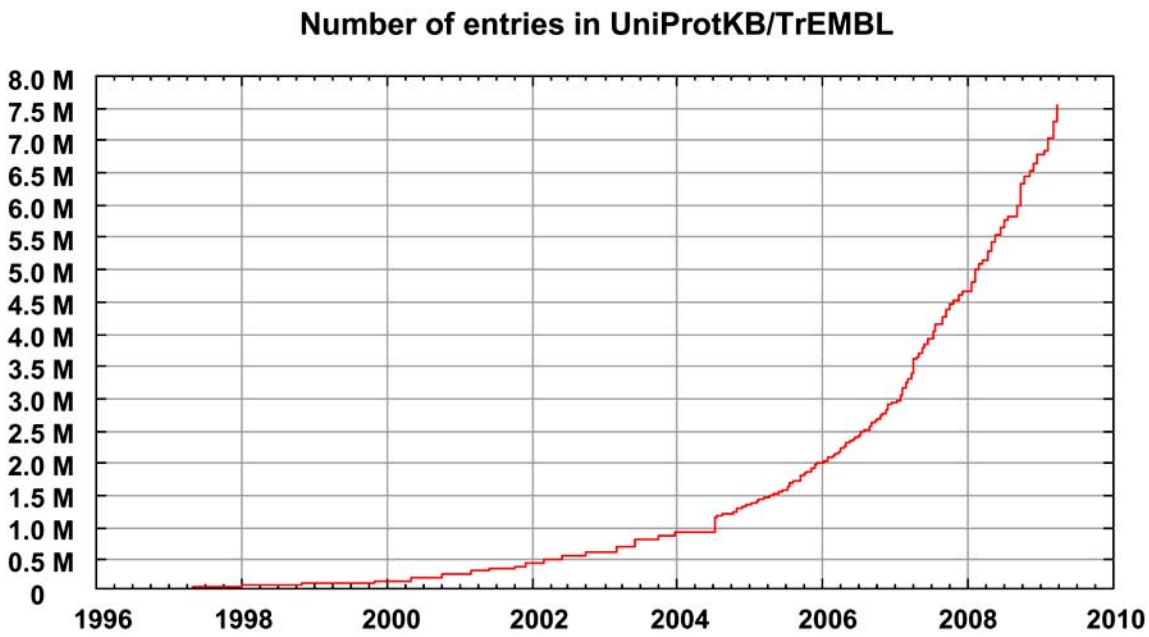

Fig. (1). A statistical illustration to show the growth of the UniProtKB/TrEMBL Protein Database (http://www.ebi.ac.uk/uniprot/-TrEMBLstats/).

lar dynamics for simulating the local motion of protein atoms [31,32], molecular docking for investigating the binding interactions of proteins and their ligands [33-35], QSAR for quantitatively studying structure-activity relationship [36-38], and pharmacophore for defining the essential features of biological activities of drugs [39, 40], among other approaches [41].

Many typical topics in sequential bioinformatics are relevant to prediction of protein attributes [42, 43]. For example, given an uncharacterized protein sequence, how can we identify which subcellular location site it resides [44-50]? Does the protein remain in a single subcellular location, or can it simultaneously exist in or move between two and more subcellular locations [50, 51]? Is the protein an enzyme or non-enzyme [52]? And if it is an enzyme, to which main functional class and sub-functional does it belong [53]? Is it a membrane protein or a non-membrane protein [54]? If it is the former, to which membrane protein type does it belong [54]? Which part of the protein serves as its signal sequence [55-58], and where might it be cleaved by proteases such as HIV protease [59-61] and SARS enzyme [35, 62]? What is its folding rate [63, 64]? Which structural class [65-68] and quaternary type does the protein belong to [69-71]? The list of questions is vast.

Although the answers to such questions can be determined by conducting a variety of biochemical experiments, the straightforward approach of performing experiments is not only time-consuming but also very costly. Consequently, the gap between the number of newly discovered protein sequences and the knowledge of their attributes continues to expand. In order to use these newly-found proteins for basic research and drug discovery in a timely manner [19, 72], it would be highly desirable if such a gap could be bridged by developing effective bioinformatics tools for predicting these various attributes of uncharacterized proteins, based on their sequence information alone.

II. SEQUENTIAL MODEL AND DISCRETE MODEL

To develop an effective method for predicting protein attributes, the following three things are indispensable: a valid benchmark dataset, a powerful prediction algorithm (or engine), and an effective mathematical expression for the samples that can truly reflect their intrinsic correlation with the object to be predicted. The current review is focused on the 3rd necessity.

There are many different ways to formulate protein sequence samples in this regard. However, they can be basically categorized into two different kinds of representations: the sequential representation and the discrete model.

1). The Sequential Model

The most straightforward sequential model for a protein sample is its entire amino acid sequence, which can contain the most complete information of the protein sequence. This is its obvious advantage. To get the desired results, the sequence-similarity-search-based tools, such as BLAST [73, 74], are usually utilized to conduct the prediction. However, this kind of approach failed to work when the query protein did not have significant homology to proteins with known characteristics (see, e.g., [75, 76]). Besides the entire protein sequences, the partial sequential model is sometimes also adopted. For example, in predicting the signal peptides for the secretory proteins, the first 70 amino acid sequential model [77] or the first 100 amino acid sequential model [57, 58] were used to represent the protein samples concerned. The rationale of such partial sequential model is based on the fact that the signal peptides of secretory proteins are located at their N-termini, and a statistical analysis [78] has shown

that the shortest one contains 8 amino acid residues and the longest one contains 90 residues, with the majority having a length within 18-25 residues.

2). The Discrete Model

Rather than a series of successive amino acid symbols according a certain order as in the sequential model, the discrete model, also termed as "non-sequential" model [79], is formulated by a set of discrete numbers.

The simplest discrete model used to represent a protein sample is its amino acid (AA) composition or AAC, as can be formulated as follows. Given a protein sequence **P** with L amino acid resides, it can be formulated as

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{1}$$

where R_1 represents the 1st residue, R_2 the 2nd residue, …, R_L the L-th residue, and they each belong to one of the 20 native amino acids. According to the AAC-discrete model, the protein **P** of **Eq. 1** can be expressed by

$$\mathbf{P} = \begin{bmatrix} f_1 & f_2 & \cdots & f_{20} \end{bmatrix}^{\mathrm{T}} \tag{2}$$

where f_u $(u = 1, 2, \cdots, 20)$ are the normalized occurrence frequencies of the 20 native amino acids [65, 80] in **P**, and **T** the transposing operator. Many methods for predicting various protein attributes were based on the AAC-discrete model (see, e.g., [45-47,67, 81-99]. However, as one can see from **Eq. 2**, all the sequence-order effects would be lost by using the AAC-discrete model, and hence the prediction quality thus obtained might be limited. This is the main shortcoming of the AAC discrete model. To avoid completely losing the sequence-order information, the concept of pseudo amino acid composition was proposed, as will be elaborated in the next section.

III. PSEUDO AMINO ACID COMPOSITION

The pseudo amino acid (PseAA) composition or PseAAC was first proposed [100] for improving the prediction quality of protein subcellular localization [47] and membrane protein types [101]. Its original formulation can be briefed as follows. Rather than **Eq.2**, the protein **P** of **Eq.1** in the PseAAC-discrete model should be formulated as

$$\mathbf{P} = \begin{bmatrix} p_1, p_2, \cdots, p_{20}, p_{20+1}, \cdots, p_{20+\lambda} \end{bmatrix}^{\mathrm{T}}, \quad (\lambda < L) \tag{3}$$

where the $20 + \lambda$ components are given by

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda}\tau_k}, & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda}\tau_k}, & (20+1 \le u \le 20+\lambda) \end{cases} \tag{4}$$

where w is the weight factor (which was set at 0.05 in [100]), and τ_k the k-th tier correlation factor that reflects the sequence order correlation between all the k-th most contiguous residues (Fig. **2**) as formulated by

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} \mathrm{J}_{i,i+k}, \quad (k < L) \tag{5}$$

with

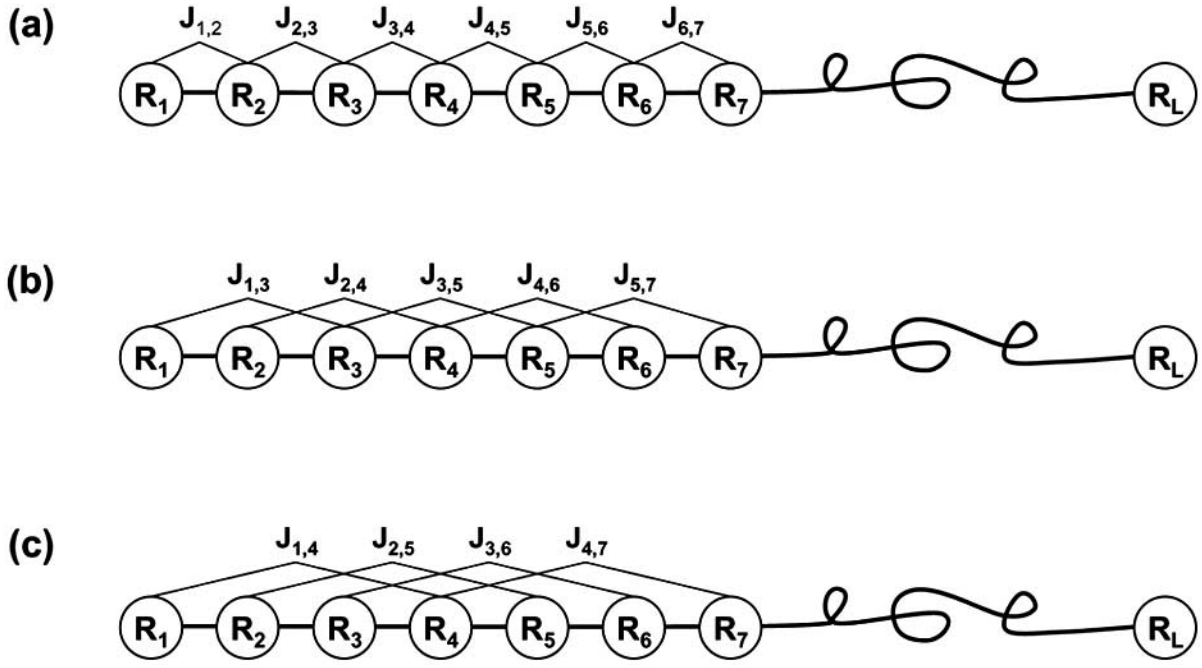

Fig. (2). A schematic drawing to show (**a**) the 1st-tier, (**b**) the 2nd-tier, and (3) the 3rd-tier sequence-order-correlation mode along a protein sequence, where R_1 represent the amino acid residue at the sequence position 1, R_2 at position 2, and so forth, and the coupling factors $\mathrm{J}_{i,j}$ are given by **Eq. 6**. Panel (a) reflects the correlation mode between all the most contiguous residues, panel (**b**) that between all the 2nd most contiguous residues, and panel (**c**) that between all the 3rd most contiguous residues. Adapted from [100] with permission.

$$J_{i,i+k} = \frac{1}{3}\left\{\begin{array}{l}\left[H_1(R_{i+k}) - H_1(R_i)\right]^2 + \left[H_2(R_{i+k}) - H_2(R_i)\right]^2 \\ + \left[M(R_{i+k}) - M(R_i)\right]^2\end{array}\right\} \quad (6)$$

where $H_1(R_i)$, $H_2(R_i)$, and $M(R_i)$ are respectively the hydrophobicity value, hydrophilicity value, and side chain mass for the amino acid R_i; while $H_1(R_{i+k})$, $H_2(R_{i+k})$, and $M(R_{i+k})$ are those for the amino acid R_{i+k}. Note that before substituting the values of hydrophobicity, hydrophilicity, and side-chain mass into Eq. **6**, they are all subjected to a *standard conversion* as described by the following equation:

$$\left\{\begin{array}{l} H_1(R_i) = \dfrac{H_1^0(R_i) - \left\langle H_1^0 \right\rangle}{SD(H_1^0)} \\ H_2(R_i) = \dfrac{H_2^0(R_i) - \left\langle H_2^0 \right\rangle}{SD(H_2^0)} \\ M(R_i) = \dfrac{M^0(R_i) - \left\langle M^0 \right\rangle}{SD(M^0)} \end{array}\right. \quad (7)$$

where the symbols $H_1^0(R_i)$ and $H_2^0(R_i)$ are the original hydrophobicity and hydrophilicity values for R_i that can be obtained from Tanford [102] and Hopp and Woods [103], respectively, and $M^0(R_i)$ the mass of the side chain for R_i that can be found from any biochemistry text book. In **Eq. 7** the symbol $\langle\ \rangle$ means taking the average of the quantity therein over 20 native amino acids, and SD means the corresponding standard deviation. The converted values obtained by **Eq. 7** will have a zero mean value over the 20 native amino acids, and will remain unchanged if going through the same conversion procedure again. As we can see, the first 20 components in **Eq. 3**, i.e., p_1, p_2, … , p_{20}, are associated with the conventional amino acid composition of P, while the remaining components p_{20+1}, p_{20+2},…, $p_{20+\lambda}$ are the λ correlation factors that reflect the 1st tier, 2nd tier, …, and the λ-th tier sequence order correlation patterns (Fig. **2**). It is these additional λ factors that approximately incorporate the sequence-order effects.

Note that using **Eq. 6** is just one of the modes for deriving the correlation factors or Pse-AAC components. The others, such as the physicochemical distance mode [104] and amphiphilic pattern mode [105], can also be used to derive different types of Pse-AAC.

Ever since the concept of PseAAC was introduced [100], it has been widely used to study various problems in proteins and protein-related systems, such as protein structural class [106-113], protein secondary structure content [114], protein quaternary structure [69, 71, 115], protein homo-oligomer types [116], classification of amino acids [117], protein subcellular localization [50,76, 118-133], protein subnuclear localization [134-136], G-protein-coupled receptor (GPCR) type classification [137,138], protein submitochondria localization [139-141], conotoxin superfamily classification [142,143], membrane protein type [54,144-148], transmembrane protein region [149], apoptosis protein subcellular localization [150-154], enzyme functional classification [52,53,105,155], cell wall lytic enzyme [156], protein fold pattern [157], cofactors of oxidoreductases [158], lipase type [159], protein-protein interactions [160], DNA-binding proteins [161], signal peptide [57, 58], and other protein-related systems [37,41].

Owing to the wide usage of PseAA composition, a web-server called "PseAAC" [162] was established at the website http://chou.med.harvard.edu/bioinf/PseAAC/, by which users can generate 63 different kinds of PseAA composition.

Actually, according to the original idea [100], the introduction of PseAAC was to keep representing a protein sequence with a discrete model yet without completely losing its sequence-order information. Therefore, although there are many other discrete models with a variety of different names, in essence they are actually different forms or modes of PseAAC, as summarized below.

1). Stochastic Signal Processing

Pan *et al.* [118] used the stochastic signal processing approach to derive the PseAAC components for predicting protein subcellular location.

2). Fourier Transform

Wang *et al*. used Fourier transform to formulate the PseAAC components for predicting membrane protein types [163].

3). Low-Frequency Fourier Spectrum Analysis

Liu *et al.* [144] pointed out that among the frequency spectrum derived from the Fourier transform, the high-frequency components are more noisy and hence only the low-frequency components are more important, just like the case of protein internal motions where the low-frequency (or Terahertz frequency) components are functionally more important (see, e.g., [22-27,164-166]). In view of this, they extracted the information from the low-frequency Fourier spectrum to formulate the PseAAC components for predicting membrane protein types [144].

4). Supervised Locally Linear Embedding (SLLE)

Wang *et al.* [167] introduced the SLLE technique to extract the essential features from the high-dimensional PseAAC space so as to enhance the cluster tolerant capacity [90] and increase the prediction quality for membrane protein types.

5). Complexity Measure Factor

Xiao *et al*. introduced the complexity measure factor as the PseAAC component for predicting protein subcellular localization [120] and protein structural classes [108].

6). Special Functions

Gao *et al*. [168] used Lyapunov index, Bessel function, and Chebyshev filter to formulate the PseAAC components for predicting protein subcellular location.

7). Cellular Automata

Xiao *et al*. introduced the cellular automaton model or image [169] to derive the PseAAC components [170] for predicting protein subcellular location [171]. Diao *et al*. [149] used cellular automata and Lempel-Ziv complexity to derive the PseAAC components for predicting transmembrane regions in proteins.

8). Sequence-Order Correlation Factors

Chen *et al*. used the sequence-order correlation factors as the PseAAC components for predicting protein structural class [106,107], and protein secondary structural contents [114].

9). Dipeptide Components

Lin and Li [109] used the dipeptide components [172,173] as the PseAAC components for predicting protein structural class.

10). Geometric Moments

Xiao *et al*. used the geometric moments of cellular automaton image (Fig. **3**) to formulate the PseAAC components for predicting protein structural classes [113].

11). Grey Dynamic Modeling

Xiao *et al*. [112] introduce the grey dynamic theory [174] to formulate the PseAAC components for predicting protein structural classes.

12). Gray Level Co-occurrence Matrix (GLCM)

Xiao *et al*. [137] used the GLCM factors extracted from the cellular automaton images to formulate the PseAAC components for predicting G-protein-coupled receptor functional classes.

13). Functional Domain (FunD)

The concept of PseAAC was expanded to incorporate the functional domain information for predicting protein subcellular localization [49,175], membrane protein types [176,177], enzyme functional classes [53], protein structural class [68], protease types [178, 179], and protein quaternary structural attributes [70,71]. The process of formulating the PseAAC in terms of the functional domain information can be briefed as follows.

Proteins often contain several modules or domains, each with a distinct function. Based on such a fact, several FunD databases were developed, such as SMART [180], Pfam [181], COG [182], KOG [182], CDD [183]. As a paradigm, here let us consider version 2.11 of CDD database, which covers 17,402 common protein domains and families. With each of the 17,402 domain sequences as a gauge base, a given protein sample can be defined in a PseAAC space with 17,402 components according to the following procedures. **Step 1**: use RPS-BLAST (Reverse PSI-BLAST) program [184] to compare the protein sequence of **Eq.1** with each of the 17,402 domain sequences in the CDD database. **Step 2**: if the significance threshold value (expect value) is

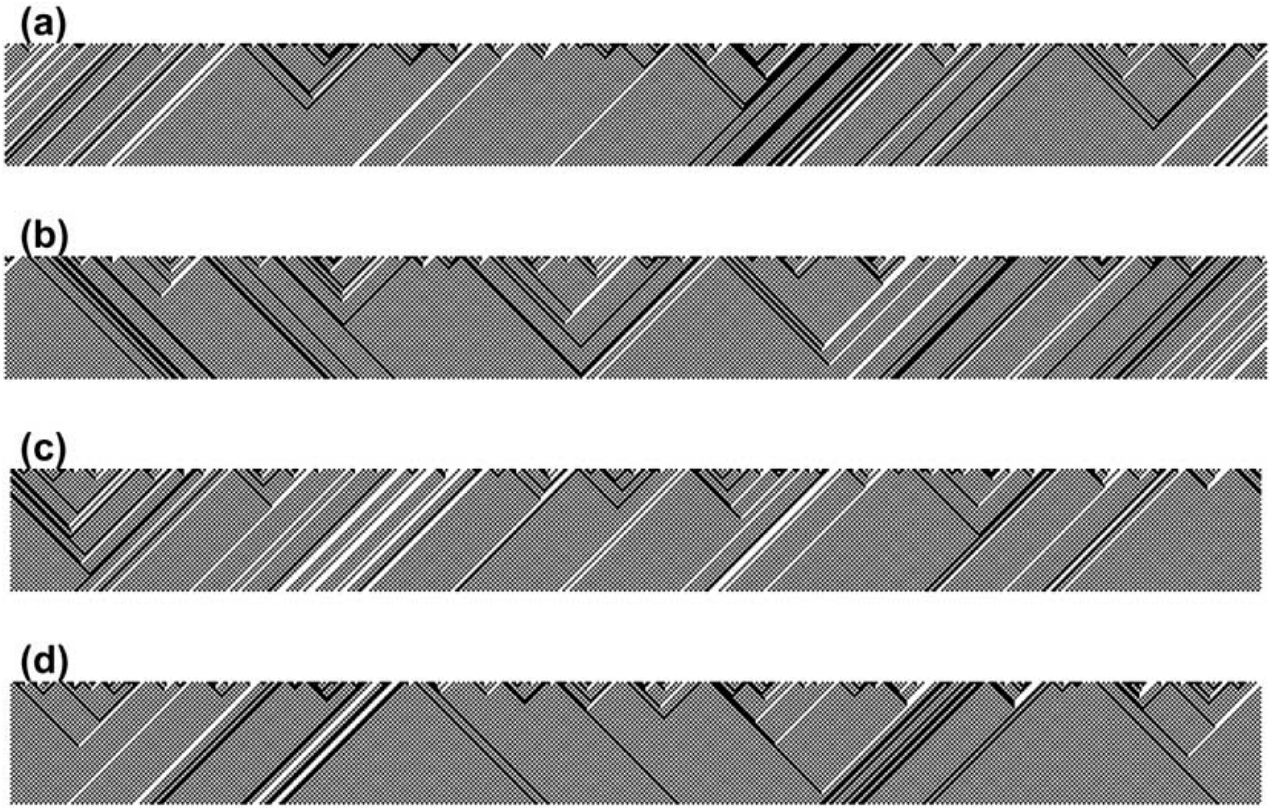

Fig. (3). The image generated by the cellular automaton evolving rule for a protein sequence in **(a)** the all-α structural class, **(b)** the all-β structural class, **(c)** the α/β structural class, and **(d)** the α+β structural class.

$E \leq 0.001$ for the i-th profile that means a "hit" is found, then the i-th component of the protein in the 17402-D PseAAC space is assigned 1; otherwise, 0. Accordingly, instead of **Eq. 3**, the PseAAC thus defined for the protein P should be formulated as

$$\mathbf{P} = \left[p_1 \quad p_2 \quad \cdots \quad p_i \quad \cdots \quad p_{17402} \right]^{\mathbf{T}} \tag{8}$$

where $\mathbf{T}$ is the transpose operator, and

$$p_i = \begin{cases} 1, & \text{when a hit is found for } \mathbf{P} \text{ in CDD} \\ 0, & \text{otherwise} \end{cases} \quad (i = 1, 2, \cdots, 17402) \tag{9}$$

14). Gene Ontology (GO)

The information extracted from the GO database [185-187] was used to formulate PseAAC for predicting protein subcellular localization [50,76,79,122,124-128,130,175,188-191], enzyme functional class [52,192], membrane protein types [193], and protein-protein interactions [160]. GO database [185] was established according to the molecular function, biological process, and cellular component. Accordingly, protein samples defined in a GO database space would be clustered in a way better reflecting their subcellular locations [79,122].

The procedures for formulating the PseAAC with the GO database can be briefed as follows. **Step 1**. Mapping UniProtKB/Swiss-Prot protein entries [194] to the GO database, one can get a list of data called "gene_association.goa_uniprot", where each UniProtKB/Swiss-Port protein entry (accession number) corresponds to one or several GO numbers. Actually, such a data file can be directly downloaded from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/. The relationships between the UniProtKB/Swiss-Port protein entries and the GO numbers may be one-to-many, "reflecting the biological reality that a particular protein may function in several processes, contain domains that carry out diverse molecular functions, and participate in multiple alternative interactions with other proteins, organelles or locations in the cell" [185], as illustrated in [76,125,128] through some examples for different organisms. **Step 2.** The GO numbers in the existing GO database do not increase successively and orderly. For easier handling, some reorganization and compression procedure was taken to renumber them; e.g., for the GO database released on 30-May-2006, after such a procedure, the original GO numbers GO:0000001, GO:0000002, GO:0000003, GO:0000004, GO:0000006, . . ., GO:0051990 would become GO_compress:0000001, GO_compress:-0000002, GO_compress:0000003, GO_compress:0000004, GO_compress:0000005, . . ., and GO_compress:0010173, respectively. The GO database thus obtained is called GO_compress database, whose maximum number was reduced from 51990 in the original GO database to 10173. Each of the 10173 entities in the GO_compress database serves as a base to define a protein sample. **Step 3.** Search the GO_compress database for the protein sequence of **Eq. 1**, if there is a hit corresponding to the i-th GO_compress number, then the i-th component of the protein in the 10173-D GO_compress space is assigned 1; otherwise, 0. Accordingly, instead of **Eq. 3**, the protein can now be formulated as:

$$\mathbf{P} = \left[p_1 \quad p_2 \quad \cdots \quad p_i \quad \cdots \quad p_{10173} \right]^{\mathbf{T}} \tag{10}$$

where

$$p_i = \begin{cases} 1, & \text{hit found in GO\_cpmpress} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

15). Sequential Evolution

The concept of PseAAC was also be extended to cover the sequential evolution information for predicting protein membrane protein type [54], enzyme functional classes [53], protease types [178,179], and protein quaternary structural attributes [70]. It was through the PSSM (**P**osition-**S**pecific **S**coring **M**atrix) [184] that the sequential evolution information was incorporated into PseAAC, as formulated below.

According to the concept of PSSM [184], the protein $\mathbf{P}$ of **Eq. 1** can be represented by:

$$\mathbf{P} = \begin{bmatrix} \mathrm{A}_{1\to1} & \mathrm{A}_{1\to2} & \cdots & \mathrm{A}_{1\to20} \\ \mathrm{A}_{2\to1} & \mathrm{A}_{2\to2} & \cdots & \mathrm{A}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{A}_{i\to1} & \mathrm{A}_{i\to2} & \cdots & \mathrm{A}_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{A}_{L\to1} & \mathrm{A}_{L\to2} & \cdots & \mathrm{A}_{L\to20} \end{bmatrix} \tag{12}$$

where $\mathrm{A}_{i\to j}$ represents the score of the amino acid residue in the i-th position of the protein sequence being changed to amino acid type j during the evolution process. Here, the numerical codes 1, 2, …, 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes. The $L \times 20$ scores in **Eq. 12** were generated by using PSI-BLAST [184] to search the Swiss-Prot database (version 52.0 released on 6-March-2007) through three iterations with 0.001 as the *E*-value cutoff for multiple sequence alignment against the sequence of the protein $\mathbf{P}$, followed by a standardization procedure given by:

$$\mathrm{A}_{i\to j} = \frac{\mathrm{A}^0_{i\to j} - \bar{\mathrm{A}}^0_i}{\mathrm{SD}\left(\mathrm{A}^0_i\right)} \quad (i = 1, 2, \cdots, L;\ j = 1, 2, \cdots, 20) \tag{13}$$

where $\mathrm{A}^0_{i\to j}$ represent the original scores directly created by PSI-BLAST that are generally shown as positive or negative integers (the positive score means that the corresponding mutation occurs more frequently in the alignment than expected by chance, while the negative score means just the opposite); $\bar{\mathrm{A}}^0_i$ the mean of $\mathrm{A}^0_{i\to j}$ over 20 native amino acids; $\mathrm{SD}\left(\mathrm{A}^0_i\right)$ the standard deviation of $\mathrm{A}^0_{i\to j}$. The standardized scores will have a zero mean value over the 20 amino acids and will remain unchanged if going through the same conversion procedure again. However, according to the PSSM descriptor (**Eq. 12**), proteins with different lengths will correspond to row-different matrices. To make the PSSM de-

scriptor become a size-uniform matrix, one possible approach is to represent a protein sample **P** by

$$\mathbf{P} = \left[\bar{\mathrm{A}}_1 \quad \bar{\mathrm{A}}_2 \quad \cdots \quad \bar{\mathrm{A}}_{20} \right]^{\mathrm{T}} \tag{14}$$

where

$$\bar{\mathrm{A}}_j = \frac{1}{L} \sum_{i=1}^{L} \mathrm{A}_{i \to j} \qquad (j = 1,\ 2,\ \cdots,\ 20) \tag{15}$$

where $\bar{\mathrm{A}}_j$ represents the average score of the amino acid residues in the protein P being changed to amino acid type j during the evolution process. However, if the protein P of **Eq. 1** is represented by **Eq. 14**, all its sequence-order information during the evolution process would be lost. To avoid completely losing this kind of evolutional information, rather than **Eq. 14**, the PseAAC for the protein **P** is formulated by

$$\mathbf{P} = \left[\bar{\mathrm{A}}_1 \quad \bar{\mathrm{A}}_2 \quad \cdots \quad \bar{\mathrm{A}}_{20} \quad \Phi_1^{\xi} \quad \Phi_2^{\xi} \quad \cdots \quad \Phi_{20}^{\xi} \right]^{\mathrm{T}} \tag{16}$$

where

$$\Phi_j^{\xi} = \frac{1}{L-\xi} \sum_{i=1}^{L-\xi} \left[\mathrm{A}_{i \to j} - \mathrm{A}_{(i+\xi) \to j} \right]^2 \quad (j = 1,\ 2,\ \cdots,\ 20;\ \xi < L) \tag{17}$$

meaning that Φ_j^1 is the correlation factor by coupling the most contiguous PSSM scores along the protein chain for the amino acid type j; Φ_j^2 that by coupling the second-most contiguous PSSM scores; and so forth.

16). Local Linear Discriminant Analysis (LLDA)

Some of high-dimensional PseAAC representations might cause over-fitting or "dimension disaster" problems. To overcome these problems, Wang *et al.* [195] introduced the LLDA approach to extract the key features from the high dimensional PseAAC space. The dimension-reduced PseAAC descriptor was used to predict membrane protein types.

Besides the list discussed above, there still might be some other representations for the sample of a protein sequence. However, regardless of how much different these representations would be from one another, they could always be classified into the category of PseAAC as long as they are composed of a set of discrete numbers that is derived from its amino acid sequence (cf. **Eq. 1**) and that is different from the classical AAC of **Eq. 2** and able to reflect some sort of sequence order or pattern effects.

IV. CONCLUSION AND PERSPECTIVES

The huge amount of protein sequences generated in the postgenomic era has stimulated the development of bioinformatics, proteomics, and system biology. In order to timely use these protein sequences for basic research and drug development, it has been highly desired to develop computational methods by which one can quickly predict their various biological attributes based on the sequence information alone. One of the key and critical problems in this regard is how to formulate the samples of protein sequences that can truly reflect their intrinsic correlation with the object to be predicted.

The number of possible sequence order patterns in proteins is extremely large, which has posed a formidable barrier for using the sequential model to conduct effective statistical prediction. But using the simple AAC discrete model will miss all the sequence-order information and hence limit the prediction quality. To deal with such a dilemma, the concept of PseAAC was proposed. In PseAAC, the information of constituent amino acids and their partial sequence-order effects are reflected via a set of discrete numbers, the so-called PseAA components.

Ever since the concept of PseAAC was introduced, many different modes of PseAAC have been proposed for dealing with various kinds of problems in proteins and proteins-related systems. In principle, the more the number of the PseAA components is, the more sequence-order effects it contains. However, it might cause over-fitting or "dimension disaster" problem if the PseAAC contains too many components. Therefore, an optimal PseAAC should consist of as many key and as few trivial components as possible. Here, the key components mean that they can best reflect the intrinsic correlation with the object to be predicted.

Different protein attributes may correspond to different modes of optimal PseAA composition. For a given protein system, how to select and discern the key components from the trivial ones so as to find its optimal PseAAC, is an interesting and challenging topic. It is also an important direction for further improving the quality of protein attribute prediction.

The other strategy to cope with the aforementioned over-fitting or "dimension disaster" problem is to introduce the ensemble classifier by fusing many different dimensions of PseAAC through a voting system.

REFERENCES

[1] Douglas, S.M.; Chou, J.J. and Shih, W.M. DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. *Proc. Natl. Acad. Sci. USA*, **2007**, *104*, 6644-6648.

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

[3] Call, M.E.; Schnell, J.R.; Xu, C.; Lutz, R.A.; Chou, J.J. and Wucherpfennig, K.W. The structure of the zetazeta transmembrane dimer reveals features essential for its assembly with the T cell receptor. *Cell*, **2006**, *127*, 355-368.

[4] Oxenoid, K. and Chou, J.J. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 10870-10875.

[5] Chou, K.C.; Jones, D. and Heinrikson, R.L. Prediction of the tertiary structure and substrate binding site of caspase-8. *FEBS Lett.*, **1997**, *419*, 49-54.

[6] Chou, J.J.; Matsuo, H.; Duan, H. and Wagner, G. Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. *Cell*, **1998**, *94*, 171-180.

[7] Chou, K.C.; Watenpaugh, K.D. and Heinrikson, R.L. A model of the complex between cyclin-dependent kinase 5(Cdk5) and the activation domain of neuronal Cdk5 activator. *Biochem. Biophys. Res. Commun.*, **1999**, *259*, 420-428.

[8] Chou, K.C.; Tomasselli, A.G. and Heinrikson, R.L. Prediction of the tertiary structure of a caspase-9/inhibitor complex. *FEBS Lett.*, **2000**, *470*, 249-256.

[9] Chou, K.C. and Howe, W.J. Prediction of the tertiary structure of the beta-secretase zymogen. *Biochem. Biophys. Res. Commun.*, **2002**, *292*, 702-708.

[10] Zhang, J.; Luan, C.H.; Chou, K.C. and Johnson, G.V.W. Identification of the N-terminal functional domains of Cdk5 by molecular

truncation and computer modelling. *Proteins: Struct Funct. Genet.*, **2002**, *48*, 447-453.

[11] Chou, K.C. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5 *Biochem. Biophys. Res. Commun.*, **2004**, *316*, 636-642.

[12] Chou, K.C. Molecular therapeutic target for type-2 diabetes. *J. Proteome Res.*, **2004**, *3*, 1284-1288.

[13] Chou, K.C. Insights from modelling the tertiary structure of BACE2. *J. Proteome Res.*, **2004**, *3*, 1069-1072.

[14] Chou, K.C. Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. *Biochem. Biophys. Res. Commun.*, **2004**, *319*, 433-438.

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

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

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

[18] Chou, K.C. Modeling the tertiary structure of human cathepsin-E. *Biochem. Biophys. Res. Commun.*, **2005**, *331*, 56-60.

[19] Chou, K.C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **2004**, *11*, 2105-2134.

[20] Levitt, M. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **1992**, *226*, 507-533.

[21] Chou, K.C. Identification of low-frequency modes in protein molecules. *Biochem. J.*, **1983**, *215*, 465-469.

[22] Chou, K.C. The biological functions of low-frequency phonons: 3. Helical structures and microenvironment. *Biophys. J.*, **1984**, *45*, 881-890.

[23] Chou, K.C. Low-frequency vibration of DNA molecules. *Biochem. J.*, **1984**, *221*, 27-31.

[24] Chou, K.C. Low-frequency motions in protein molecules: beta-sheet and beta-barrel. *Biophys. J.*, **1985**, *48*, 289-297.

[25] Chou, K.C. Low-frequency resonance and cooperativity of haemoglobin. *Trends Biochem. Sci.*, **1989**, *14*, 212.

[26] Chou, K.C.; Maggiora, G.M. and Mao, B. Quasi-continuum models of twist-like and accordion-like low-frequency motions in DNA. *Biophys. J.*, **1989**, *56*, 295-305.

[27] Chou, K.C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys Chem.*, **1988**, *30*, 3-48.

[28] Martel, P. Biophysical aspects of neutron scattering from vibrational modes of proteins. *Prog. Biophys. Mol. Biol.*, **1992**, *57*, 129-179.

[29] Gordon, G. Designed electromagnetic pulsed therapy: clinical applications. *J. Cell Physiol.*, **2007**, *212*, 579-582.

[30] Gordon, G. Extrinsic electromagnetic fields, low frequency (phonon) vibrations, and control of cell function: a non-linear resonance system. *J. Biomed. Sci. Eng.*, **2008**, *1*, 152-156.

[31] Karplus, M., Shakhnovich, E. In *Protein Folding*; Creighton, T. E., Ed.; Freeman: New York, **1992**.

[32] MacKerell JrA.D.; Bashford, D.; Bellott, M.M.; Dunbrack Jr.R.L.; Evanseck, J.D.; Field, M.J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.T.K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.T.; Prodhom, B.; Reiher, I.W.E.; Roux, B.; Schlenkrich, M.; Smith, J.C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D. and Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem.*, **1998**, *102*, 3586-3616.

[33] Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K. and Olson, A.J. Automated docking using a Lamarckian Genetic Algorithm and empirical binding free energy function. *J. Comput. Chem.*, **1998**, *19*, 1639-1662.

[34] Rarey, M.; Kramer, B.; Lengauer, T. and Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **1996**, *261*, 470-489.

[35] Chou, K.C.; Wei, D.Q. and Zhong, W.Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: ibid., 2003, Vol.310, 675). *Biochem. Biophys. Res. Commun.*, **2003**, *308*, 148-151.

[36] Gonzalez-Diaz, H.; Sanchez-Gonzalez, A. and Gonzalez-Diaz, Y. 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. *J. Inorg. Biochem.*, **2006**, *100*, 1290-1297.

[37] Gonzalez-Diaz, H.; Vilar, S.; Santana, L. and Uriarte, E. Medicinal chemistry and bioinformatics - current trends in drugs discovery with networks topological indices. *Curr. Top. Med. Chem.*, **2007**, *10*, 1015-1029.

[38] Du, Q.S.; Huang, R.B. and Chou, K.C. Review: Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. *Curr. Protein Peptide Sci.*, **2008**, *9*, 248-259.

[39] Marchand-Geneste, N.; Watson, K.A.; Alsberg, B.K. and King, R.D. New approach to pharmacophore mapping and QSAR analysis using inductive logic programming. Application to thermolysin inhibitors and glycogen phosphorylase B inhibitors. *J. Med. Chem.*, **2002**, *45*, 399-409.

[40] Sirois, S.; Wei, D.Q.; Du, Q.S. and Chou, K.C. Virtual Screening for SARS-CoV protease based on KZ7088 pharmacophore points. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1111-1122.

[41] Gonzalez-Diaz, H.; Gonzalez-Diaz, Y.; Santana, L.; Ubeira, F.M. and Uriarte, E. Proteomics, networks, and connectivity indices. *Proteomics*, **2008**, *8*, 750–778.

[42] Chou, K.C. In *Gene Cloning & Expression Technologies*, Weinrer, P. W., Lu, Q., Eds.; Eaton Publishing: Westborough, MA, Chp. 4, pp. 57-70. **2002**.

[43] Chou, K.C. In *Automation in Proteomics and Genomics: An Engineering Case-Based Approach (Harvard-MIT interdisciplinary special studies courses)*, Alterovitz, G., Benson, R., and Ramoni, M.F., Eds. Wiley & Sons, Ltd.: West Sussex, UK, **2009**; Chap. 5, pp. 97-143.

[44] Nakai, K. and Kanehisa, M. Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins: Struct. Funct. Genet.*, **1991**, *11*, 95-110.

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

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

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

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

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

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

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

[52] Chou, K.C. and Cai, Y.D. Predicting enzyme family class in a hybridization space. *Protein Science*, **2004**, *13*, 2857-2863.

[53] Shen, H.B. and Chou, K.C. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.*, **2007**, *364*, 53-59.

[54] Chou, K.C. and Shen, H.B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.*, **2007**, *360*, 339-345.

[55] Nielsen, H.; Engelbrecht, J.; von Heijne, G. and Brunak, S. Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins*, **1996**, *24*, 165-177.

[56] Nielsen, H.; Engelbrecht, J.; Brunak, S. and von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **1997**, *10*, 1-6.

[57] Chou, K.C. and Shen, H.B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.*, **2007**, *357*, 633-640.

[58] Shen, H.B. and Chou, K.C. Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Commun.*, **2007**, *363*, 297-303.

[59] Poorman, R.A.; Tomasselli, A.G.; Heinrikson, R.L. and Kezdy, F.J. A cumulative specificity model for proteases from human immu-

nodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. *J. Biol. Chem.*, **1991**, *266*, 14554-14561.

[60] Chou, K.C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.*, **1993**, *268*, 16938-16948.

[61] Chou, K.C. Review: Prediction of HIV protease cleavage sites in proteins *Analytical. Biochemistry*, **1996**, *233*, 1-14.

[62] Anand, K.; Ziebuhr, J.; Wadhwani, P.; Mesters, J.R. and Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science*, **2003**, *300*, 1763-1767.

[63] Ouyang, Z. and Liang, J. Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci.*, **2008**, *17*, 1256-1263.

[64] Chou, K.C. and Shen, H.B. FoldRate: A web-server for predicting protein folding rates from primary sequence. *Open Bioinformatics J.*, **2009**, *3*, 31-50.

[65] Nakashima, H.; Nishikawa, K. and Ooi, T. The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **1986**, *99*, 152-162.

[66] Chou, K.C. and Zhang, C.T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.*, **1994**, *269*, 22014-22020.

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

[68] Chou, K.C. and Cai, Y.D. Predicting protein structural class by functional domain composition. *Biochem. Biophys. Res. Commun. (Corrigendum: ibid., 2005, Vol.329, 1362)*, **2004**, *321*, 1007-1009.

[69] Chou, K.C. and Cai, Y.D. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins: Structure, Function, and Genetics*, **2003**, *53*, 282-289.

[70] Shen, H.B. and Chou, K.C. QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *J. Proteome Res.*, **2009**, *8*, 1577–1584.

[71] Xiao, X.; Wang, P. and Chou, K.C. Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. *J. Applied Crystallography*, **2009**, *42*, 169-173.

[72] Chou, K.C. Structural bioinformatics and its impact to biomedical science and drug discovery, In *Frontiers in Medicinal Chemistry*; Atta-ur-Rahman and Reitz, A. B., Ed.; Bentham Science Publishers: The Netherlands, **2006**; Vol. *3*. pp. 455-502.

[73] Altschul, S.F. In *Theoretical and Computational Methods in Genome Research*; Suhai, S., Ed.; Plenum: New York, 1997;. pp. 455-502.

[74] Wootton, J.C. and Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases *Comput. Chem.*, **1993**, *17*, 149-163.

[75] Garg, A.; Bhasin, M. and Raghava, G.P. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.*, **2005**, *280*, 14427-14432.

[76] Chou, K.C. and Shen, H.B. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.*, **2006**, *347*, 150-157.

[77] Nielsen, H.; Brunak, S. and von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **1999**, *12*, 3-9.

[78] Chou, K.C. Review: Prediction of protein signal sequences. *Curr. Protein Peptide Science*, **2002**, *3*, 615-622.

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

[80] Chou, K.C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins: Struct. Funct. Genet.*, **1995**, *21*, 319-344.

[81] Klein, P. and Delisi, C. Prediction of protein structural class from amino acid sequence. *Biopolymers*, **1986**, *25*, 1659-1672.

[82] Klein, P. Prediction of protein structural class by discriminant analysis. *Biochim. Biophys. Acta*, **1986**, *874*, 205-215.

[83] Chou, P.Y. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York, **1989**.

[84] Chou, K.C. and Zhang, C.T. A correlation coefficient method to predicting protein structural classes from amino acid compositions. *Eur. J Biochem.*, **1992**, *207*, 429-433.

[85] Metfessel, B.A.; Saurugger, P.N.; Connelly, D.P. and Rich, S.T. Cross-validation of protein structural class prediction using statistical clustering and neural networks. *Protein Sci.*, **1993**, *2*, 1171-1182.

[86] Chou, K.C. and Maggiora, G.M. Domain structural class prediction. *Protein Eng.*, **1998**, *11*, 523-538.

[87] Chou, K.C.; Liu, W.; Maggiora, G.M. and Zhang, C.T. Prediction and classification of domain structural classes. *Proteins: Struct. Funct. Genet.*, **1998**, *31*, 97-103.

[88] Zhou, G.P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.*, **1998**, *17*, 729-738.

[89] Liu, W. and Chou, K.C. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. *J Protein Chem.*, **1998**, *17*, 209-217.

[90] Chou, K.C. A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.*, **1999**, *264*, 216-224.

[91] Chou, K.C. Review: Prediction of protein structural classes and subcellular locations. *Curr. Protein Pept. Sci.*, **2000**, *1*, 171-208.

[92] Zhou, G.P. and Assa-Munt, N. Some insights into protein structural class prediction *Proteins: Struct. Funct. Genet.*, **2001**, *44*, 57-59.

[93] Chou, K.C. and Elrod, D.W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.*, **2002**, *1*, 429-433.

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

[95] Chou, K.C. Prediction of G-protein-coupled receptor classes. *J Proteome Res.*, **2005**, *4*, 1413-1418.

[96] Feng, K.Y.; Cai, Y.D. and Chou, K.C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.*, **2005**, *334*, 213-217.

[97] Du, Q.S.; Jiang, Z.Q.; He, W.Z.; Li, D.P. and Chou, K.C. Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. *J. Biomol. Struct. Dyn.*, **2006**, *23*, 635-640.

[98] Niu, B.; Cai, Y.D.; Lu, W.C.; Zheng, G.Y. and Chou, K.C. Predicting protein structural class with AdaBoost learner. *Protein Pept. Lett.*, **2006**, *13*, 489-492.

[99] Jahandideh, S.; Abdolmaleki, P.; Jahandideh, M. and Asadabadi, E.B. Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. *Biophys. Chem.*, **2007**, *128*, 87-93.

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

[101] Chou, K.C. and Elrod, D.W. Prediction of membrane protein types and subcellular locations. *Proteins: Struct. Funct. Genet.*, **1999**, *34*, 137-153.

[102] Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.*, **1962**, *84*, 4240-4274.

[103] Hopp, T.P. and Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, **1981**, *78*, 3824-3828.

[104] Chou, K.C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **2000**, *278*, 477-483.

[105] Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **2005**, *21*, 10-19.

[106] Chen, C.; Zhou, X.; Tian, Y.; Zou, X. and Cai, P. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal. Biochem.*, **2006**, *357*, 116-121.

[107] Chen, C.; Tian, Y.X.; Zou, X.Y.; Cai, P.X. and Mo, J.Y. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J. Theor. Biol.*, **2006**, *243*, 444-448.

[108] Xiao, X.; Shao, S.H.; Huang, Z.D. and Chou, K.C. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J. Comput. Chem.*, **2006**, *27*, 478-482.

[109] Lin, H. and Li, Q.Z. Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components. *J. Comput. Chem.*, **2007**, *28*, 1463-1466.

[110] Ding, Y.S.; Zhang, T.L. and Chou, K.C. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein Pept. Lett.*, **2007**, *14*, 811-815.

[111] Zhang, T.L.; Ding, Y.S. and Chou, K.C. Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. *J. Theor. Biol.*, **2008**, *250*, 186-193.

[112] Xiao, X.; Lin, W.Z. and Chou, K.C. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *J. Comput. Chem.*, **2008**, *29*, 2018-2024.

[113] Xiao, X.; Wang, P. and Chou, K.C. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. *J. Theor. Biol.*, **2008**, *254*, 691-696.

[114] Chen, C.; Chen, L.; Zou, X. and Cai, P. Prediction of protein secondary structure content by using the concept of chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.*, **2009**, *16*, 27-31.

[115] Zhang, S.W.; Chen, W.; Yang, F. and Pan, Q. Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. *Amino Acids*, **2008**, *35*, 591-598.

[116] Zhang, S.W.; Pan, Q.; Zhang, H.C.; Shao, Z.C. and Shi, J.Y. Prediction protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and naive Bayes feature fusion. *Amino Acids*, **2006**, *30*, 461-468.

[117] Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J. and Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2009**, *257*, 17-26.

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

[119] Chou, K.C. and Cai, Y.D. Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. *J. Cell. Biochem.*, **2004**, *91*, 1197-1203.

[120] Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y. and Chou, K.C. Using complexity measure factor to predict protein subcellular location. *Amino Acids*, **2005**, *28*, 57-61.

[121] Chou, K.C. and Shen, H.B. Predicting protein subcellular location by fusing multiple classifiers. *J. Cell. Biochem.*, **2006**, *99*, 517-527.

[122] Chou, K.C. and Shen, H.B. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **2008**, *3*, 153-162.

[123] Li, F.M. and Li, Q.Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Pept. Lett.*, **2008**, *15*, 612-616.

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

[125] Chou, K.C. and Shen, H.B. Large-scale plant protein subcellular location prediction. *J. Cell. Biochem.*, **2007**, *100*, 665-678.

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

[127] Chou, K.C. and Shen, H.B. Large-scale predictions of gram-negative bacterial protein subcellular locations. *J. Proteome Res.*, **2006**, *5*, 3420-3428.

[128] Shen, H.B. and Chou, K.C. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng. Des. Select.*, **2007**, *20*, 39-46.

[129] Shi, J.Y.; Zhang, S.W.; Pan, Q.; Cheng, Y.-M. and Xie, J. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids*, **2007**, *33*, 69-74.

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

[131] Lin, H.; Ding, H.; Feng-Biao Guo, F.B.; Zhang, A.Y. and Huang, J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **2008**, *15*, 739-744.

[132] Zhang, S.W.; Zhang, Y.L.; Yang, H.F.; Zhao, C.H. and 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.

[133] Shi, J.Y.; Zhang, S.W.; Pan, Q. and Zhou, G.P. Using pseudo amino acid composition to predict protein subcellular location: Approached with amino acid composition distribution. *Amino Acids*, **2008**, *35*, 321-327.

[134] Shen, H.B. and Chou, K.C. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem. Biophys. Res. Commun.*, **2005**, *337*, 752-756.

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

[136] Jiang, X.; Wei, R.; Zhao, Y. and Zhang, T. Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear locationl. *Amino Acids*, **2008**, *34*, 669-675.

[137] Xiao, X.; Wang, P. and Chou, K.C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J. Comput. Chem.*, **2009**, *30*, 1414-1423.

[138] Qiu, J.D.; Huang, J.H.; Liang, R.P. and Lu, X.Q. Prediction of G-protein coupled receptors classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. *Anal. Biochem.*, **2009**, *390*, 68-73.

[139] Du, P. and Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence *BMC Bioinformatics*, **2006**, *7*, 518.

[140] Nanni, L. and Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **2008**, *34*, 653-660.

[141] Zeng, Y.H.; Guo, Y.Z.; Xiao, R.Q.; Yang, L.; Yu, L.Z. and Li, M.L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.*, **2009**, *259*, 366–372.

[142] Mondal, S.; Bhavna, R.; Mohan Babu, R. and Ramakumar, S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.*, **2006**, *243*, 252-260.

[143] Lin, H. and Li, Q.Z. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. *Biochem. Biophys. Res. Commun.*, **2007**, *354*, 548-551.

[144] Liu, H.; Wang, M. and Chou, K.C. Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem. Biophys. Res. Commun.*, **2005**, *336*, 737-739.

[145] Shen, H.B. and Chou, K.C. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.*, **2005**, *334*, 288-292.

[146] Wang, S.Q.; Yang, J. and Chou, K.C. Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. *J. Theor. Biol.*, **2006**, *242*, 941-946.

[147] Shen, H.B.; Yang, J. and Chou, K.C. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *J. Theor. Biol.*, **2006**, *240*, 9-13.

[148] Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2008**, *252*, 350-356.

[149] Diao, Y.; Ma, D.; Wen, Z.; Yin, J.; Xiang, J. and Li, M. Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids*, **2008**, *34*, 111-117.

[150] Jiang, X.; Wei, R.; Zhang, T.L. and Gu, Q. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein Pept. Lett.*, **2008**, *15*, 392-396.

[151] Chen, Y.L. and Li, Q.Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *J. Theor. Biol.*, **2007**, *248*, 377–381.
[152] Chen, Y.L. and Li, Q.Z. Prediction of the subcellular location of apoptosis proteins. *J. Theor. Biol.*, **2007**, *245*, 775-783.
[153] Ding, Y.S. and 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 Recognit. Lett.*, **2008**, *29*, 1887-1892.
[154] Lin, H.; Wang, H.; Ding, H.; Chen, Y.L. and Li, Q.Z. Prediction of subcellular localization of apoptosis protein using chou's pseudo amino acid composition. *Acta Biotheor.*, **2009**, *57*, 321-330.
[155] Zhou, X.B.; Chen, C.; Li, Z.C. and Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.*, **2007**, *248*, 546–551.
[156] Ding, H.; Luo, L. and Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Pept. Lett.*, **2009**, *16*, 351-355.
[157] Shen, H.B. and Chou, K.C. Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **2006**, *22*, 1717-1722.
[158] Zhang, G.Y. and Fang, B.S. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. *J. Theor. Biol.*, **2008**, *253*, 310-315.
[159] Zhang, G.Y.; Li, H.C. and Fang, B.S. Predicting lipase types by improved Chou's pseudo-amino acid composition. *Protein Pept. Lett.*, **2008**, *15*, 1132-1137.
[160] Chou, K.C. and Cai, Y.D. Predicting protein-protein interactions from sequences in a hybridization space. *J. Proteome Res.*, **2006**, *5*, 316-322.
[161] Fang, Y.; Guo, Y.; Feng, Y. and 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.
[162] Shen, H.B. and Chou, K.C. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **2008**, *373*, 386-388.
[163] Wang, M.; Yang, J.; Liu, G.P.; Xu, Z.J. and Chou, K.C. Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. *Protein Eng. Des. Select.*, **2004**, *17*, 509-516.
[164] Chou, K.C. and Chen, N.Y. The biological functions of low-frequency phonons. *Sci. Sin.*, **1977**, *20*, 447-457.
[165] Chou, K.C. Low-frequency vibrations of helical structures in protein molecules. *Biochem. J.*, **1983**, *209*, 573-580.
[166] Vermont, C. In *http://www.sover.net/~bell/newFrontierpics.htm* Bellows Falls, Vermont USA [January 15, 2009].
[167] Wang, M.; Yang, J.; Xu, Z.J. and Chou, K.C. SLLE for predicting membrane protein types. *J. Theor. Biol.*, **2005**, *232*, 7-15.
[168] Gao, Y.; Shao, S.H.; Xiao, X.; Ding, Y.S.; Huang, Y.S.; Huang, Z.D. and Chou, K.C. Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. *Amino Acids*, **2005**, *28*, 373-376.
[169] Wolfram, S. Cellular automation as models of complexity. *Nature*, **1984**, *311*, 419-424.
[170] Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X. and Chou, K.C. Using cellular automata to generate Image representation for biological sequences. *Amino Acids*, **2005**, *28*, 29-35.
[171] Xiao, X.; Shao, S.H.; Ding, Y.S.; Huang, Z.D. and Chou, K.C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids*, **2006**, *30*, 49-54.
[172] Liu, W. and Chou, K.C. Protein secondary structural content prediction. *Protein Eng.*, **1999**, *12*, 1041-1050.
[173] Bhasin, M. and Raghava, G.P. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **2004**, *32*, W414-419.
[174] Deng, J.L. Grey system control. *Syst. Control Lett.*, **1985**, *1*, 288-294.
[175] Chou, K.C. and Cai, Y.D. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem. Biophys. Res. Commun.*, **2004**, *320*, 1236-1239.
[176] Cai, Y.D.; Zhou, G.P. and Chou, K.C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **2003**, *84*, 3257-3263.
[177] Cai, Y.D. and Chou, K.C. Predicting membrane protein type by functional domain composition and pseudo amino acid composition. *J. Theor. Biol.*, **2006**, *238*, 395-400.
[178] Chou, K.C. and Shen, H.B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun.*, **2008**, *376*, 321-325.
[179] Shen, H.B. and Chou, K.C. Identification of proteases and their types. *Anal. Biochem.*, **2009**, *385*, 153-160.
[180] Letunic, I.; Copley, R.R.; Pils, B.; Pinkert, S.; Schultz, J. and Bork, P. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **2006**, *34*, D257-D260.
[181] Finn, R.D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S.R.; Sonnhammer, E.L. and Bateman, A. Pfam: clans, web tools and services. *Nucleic Acids Res.*, **2006**, *34*, D247-D251.
[182] Tatusov, R.L.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Kiryutin, B.; Koonin, E.V.; Krylov, D.M.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; Rao, B.S.; Smirnov, S.; Sverdlov, A.V.; Vasudevan, S.; Wolf, Y.I.; Yin, J.J. and Natale, D.A. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **2003**, *4*, 41.
[183] Marchler-Bauer, A.; Anderson, J.B.; Derbyshire, M.K.; DeWeese-Scott, C.; Gonzales, N.R.; Gwadz, M.; Hao, L.; He, S.; Hurwitz, D.I.; Jackson, J.D.; Ke, Z.; Krylov, D.; Lanczycki, C.J.; Liebert, C.A.; Liu, C.; Lu, F.; Lu, S.; Marchler, G.H.; Mullokandov, M.; Song, J.S.; Thanki, N.; Yamashita, R.A.; Yin, J.J.; Zhang, D. and Bryant, S.H. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*, **2007**, *35*, D237-D240.
[184] Schaffer, A.A.; Aravind, L.; Madden, T.L.; Shavirin, S.; Spouge, J.L.; Wolf, Y.I.; Koonin, E.V. and Altschul, S.F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **2001**, *29*, 2994-3005.
[185] Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A.; Hill, D.P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.C.; Richardson, J.E.; Ringwald, M.; Rubin, G.M. and Sherlock, G. Gene ontology: tool for the unification of biology. *Nat. Genet.*, **2000**, *25*, 25-29.
[186] Camon, E.; Magrane, M.; Barrell, D.; Lee, V.; Dimmer, E.; Maslen, J.; Binns, D.; Harte, N.; Lopez, R. and Apweiler, R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **2004**, *32*, D262-D266.
[187] Harris, M.A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G.M.; Blake, J.A.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J.T.; Hill, D.P.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J.M.; Christie, K.R.; Costanzo, M.C.; Dwight, S.S.; Engel, S.; Fisk, D.G.; Hirschman, J.E.; Hong, E.L.; Nash, R.S.; Sethuraman, A.; Theesfeld, C.L.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S.Y.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E.M.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.; Tonellato, P.; Jaiswal, P.; Seigfried, T. and White, R. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **2004**, *32*, D258-D261.
[188] Chou, K.C. and 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.
[189] Cai, Y.D. and Chou, K.C. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochem. Biophys. Res. Commun.*, **2003**, *305*, 407-411.

[190] Lee, V.; Camon, E.; Dimmer, E.; Barrell, D. and Apweiler, R. Who tangos with GOA?-Use of Gene Ontology Annotation (GOA) for biological interpretation of '-omics' data and for validation of automatic annotation tools. *In Silico Biol.*, **2005**, *5*, 5-8.

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

[192] Chou, K.C. and Cai, Y.D. Using GO-PseAA predictor to predict enzyme sub-class. *Biochem. Biophys. Res. Commun.*, **2004**, *325*, 506-509.

[193] Chou, K.C. and Cai, Y.D. Using GO-PseAA predictor to identify membrane proteins and their types. *Biochem. Biophys. Res. Commun.*, **2005**, *327*, 845-847.

[194] Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M.J.; Natale, D.A.; O'Donovan, C.; Redaschi, N. and Yeh, L.S. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **2004**, *32*, D115-D119.

[195] Wang, T.; Yang, J.; Shen, H.B. and Chou, K.C. Predicting membrane protein types by the LLDA algorithm. *Protein Pept. Lett.*, **2008**, *15*, 915-921.

Received: May 02, 2009 Revised: September 01, 2009 Accepted: September 17, 2009