Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Journal of Theoretical Biology 273 (2011) 236–247

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

50th Anniversary Year Review

Some remarks on protein attribute prediction and pseudo amino acid composition

Kuo-Chen Chou

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

ARTICLE INFO

Available online 17 December 2010

Keywords:
PseAAC
Functional domain mode
Gene ontology mode
Sequential evolution mode
Cross-validation

ABSTRACT

With the accomplishment of human genome sequencing, the number of sequence-known proteins has increased explosively. In contrast, the pace is much slower in determining their biological attributes. As a consequence, the gap between sequence-known proteins and attribute-known proteins has become increasingly large. The unbalanced situation, which has critically limited our ability to timely utilize the newly discovered proteins for basic research and drug development, has called for developing computational methods or high-throughput automated tools for fast and reliably identifying various attributes of uncharacterized proteins based on their sequence information alone. Actually, during the last two decades or so, many methods in this regard have been established in hope to bridge such a gap. In the course of developing these methods, the following things were often needed to consider: (1) benchmark dataset construction, (2) protein sample formulation, (3) operating algorithm (or engine), (4) anticipated accuracy, and (5) web-server establishment. In this review, we are to discuss each of the five procedures, with a special focus on the introduction of pseudo amino acid composition (PseAAC), its different modes and applications as well as its recent development, particularly in how to use the general formulation of PseAAC to reflect the core and essential features that are deeply hidden in complicated protein sequences.

© 2010 Elsevier Ltd. All rights reserved.

Contents

1. Introduction . . . 236
2. Benchmark dataset . . . 237
3. Protein sample representation . . . 238
   3.1. Functional domain mode . . . 240
   3.2. Gene ontology mode . . . 240
   3.3. Sequential evolution mode . . . 240
4. Prediction algorithm (operating engine) . . . 241
   4.1. Nearest neighbor classifier . . . 241
   4.2. KNN classifier . . . 241
   4.3. One-dimensional fusion . . . 242
   4.4. Two-dimensional fusion . . . 242
5. Cross-validation test . . . 242
6. Web-server . . . 244
7. Conclusion and perspectives . . . 244
   Acknowledgements . . . 244
   References . . . 244

1. Introduction

With the explosive growth of protein sequences generated in the postgenomic age, scientists are anxious to know their attributes because they are closely correlated with the structures and functions of the proteins as well as their roles in biological processes, and hence

E-mail address: kcchou@gordonlifescience.org

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.12.024

Author's personal copy

are very important to both basic research and drug target development. For instance, given an uncharacterized protein sequence, what is its folding rate? Which structural class and quaternary structural attribute does it belong to? Which subcellular location site does it resides? Can it simultaneously exist in or move between two and more subcellular locations? How can we identify it as an enzyme or non-enzyme? If it is an enzyme, to which enzyme functional class does it belong? Is it a membrane protein or non-membrane protein? If the former, to which membrane protein type does it belong? Is it a protease? If it is, to which protease type does it belong? Is it a G protein-coupled receptor (GPCR)? If it is, to which GPCR type does it belong? Which part of the protein serves as its signal sequence? Where are its cleavage sites by proteases such as HIV (human immunodeficiency virus) protease and SARS (severe acute respiratory syndrome) enzyme? And so forth. Although the answers to these questions can be determined by conducting various biochemical experiments, it is both time-consuming and costly by relying on experimental approaches alone. As a consequence, the gap between the number of newly discovered protein sequences and the knowledge of their attributes is continuing to expand. To bridge such a gap and acquire these kinds of information in a timely manner, scientists are challenged to develop computational methods for predicting various attributes of proteins based on their sequence information alone.

To establish a really useful predictor in this regard, one usually needs to accomplish the following procedures: (1) construct a valid benchmark dataset to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the accuracy of the predictor; and (5) establish a user-friendly web-server for the predictor that is accessible to the public.

This review will discuss each of the above five procedures, with a special focus on procedure 2, particularly on how to use various different modes of pseudo amino acid composition to represent protein samples by incorporating their core and essential features.

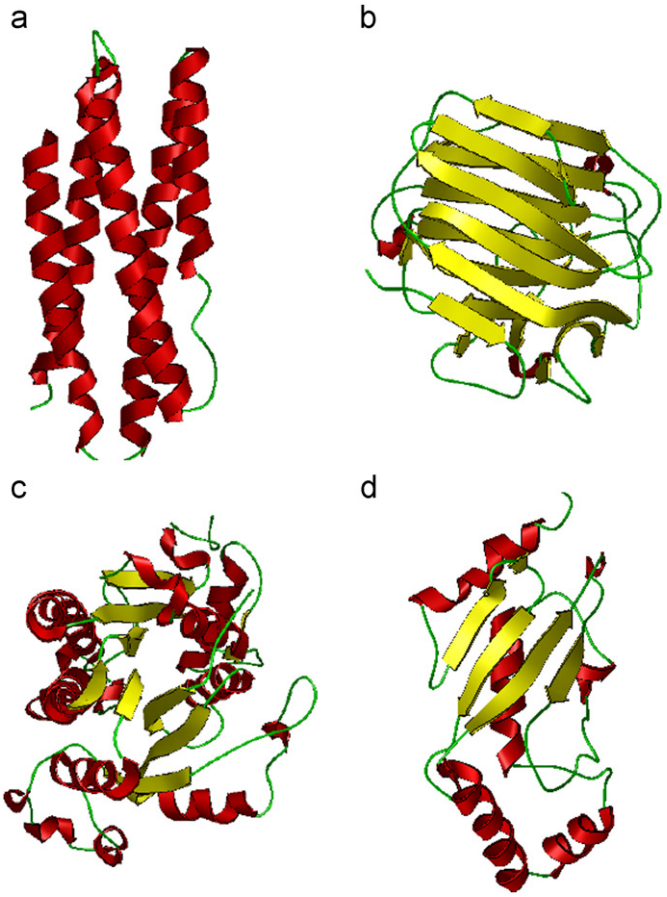


Fig. 1. Illustration to show the four categories of protein structural class: (a) all-α, (b) all-β, (c) α/β, and (d) α+β, where the α-helix is colored in red, β-strand in yellow, and the other in green. The PDB codes used to draw the representatives of the four structural classes are 1aep, 1gbg, 1enp, and 1aak, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Benchmark dataset

To develop a statistical prediction method for a given attribute, the first important thing is to construct a benchmark dataset $\mathbb{S}$ according to its possible classification, i.e.

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \cdots \cup \mathbb{S}_m \cup \cdots \cup \mathbb{S}_M \tag{1}$$

where $\mathbb{S}_1$ represents the subset for category 1 of the attribute, $\mathbb{S}_2$ for category 2, and so forth; while $\cup$ represents the symbol for "union" in the set theory, and M the number of different categories for the attribute concerned. For example, when the attribute concerned was about the protein structural classification as investigated in Chou (1995a), Chou and Zhang (1994), Chou (1989), Levitt and Chothia (1976), Nakashima et al. (1986) and Zhou (1998), M would be four as illustrated in Fig. 1; when the structural classification was defined according to the SCOP database (Murzin et al., 1995) or investigated in Chou and Cai (2004b), M would be seven as shown in Fig. 2; when the attribute was about the membrane protein type as investigated in Chou and Shen (2007d), M would be eight (Chou and Shen, 2007d) as illustrated in Fig. 3; when the attribute was about the subcellular localization of eukaryotic proteins as investigated in Chou and Shen (2010a), M would be 22 as illustrated in Fig. 4.

To avoid homology bias and redundancy, it is important to introduce a cutoff threshold when constructing a benchmark dataset. Different cutoff threshold values were used, such as 90% (Reinhardt and Hubbard, 1998), 80% (Small et al., 2004), 40% (Shen and Chou, 2007a), and 25% (Chou and Shen, 2010a; Chou and Shen, 2010c). When a benchmark dataset was constructed with the cutoff threshold of 25%, none of the proteins included would have $\geq 25\%$ pairwise sequence identity to any other in the same subset (category). Accordingly, the smaller the cutoff threshold is, the more stringent the benchmark dataset will be in excluding the homology bias.

The benchmark datasets constructed in the earlier stage (see, e.g., Cedano et al., 1997; Chou, 1989; Nakashima et al., 1986) usually consisted of a learning (or training) dataset and an independent testing dataset, as can be formulated as

$$\begin{cases} \mathbb{S} = \mathbb{S}^{\mathrm{L}} \cup \mathbb{S}^{\mathrm{T}} \\ \varnothing = \mathbb{S}^{\mathrm{L}} \cap \mathbb{S}^{\mathrm{T}} \end{cases} \tag{2}$$

where $\mathbb{S}^{\mathrm{L}}$ is the learning dataset, $\mathbb{S}^{\mathrm{T}}$ the training dataset, $\varnothing$ the empty set, and $\cap$ the symbol for "intersection" in the set theory. The learning dataset is used for training the predictor's "engine", while the testing dataset used for evaluating the predictor's accuracy via a cross-validation. As we can see from Eq. (2), none of the proteins in the testing dataset $\mathbb{S}^{\mathrm{T}}$ should occur in the learning dataset $\mathbb{S}^{\mathrm{L}}$. Therefore, $\mathbb{S}^{\mathrm{T}}$ is also called an independent dataset for performing cross-validation. However, as will be shown later, there is no need to artificially separate the benchmark dataset into a learning dataset and a testing dataset when the cross-validation is performed by the jackknife test, in which case one benchmark dataset can serve both the training and testing purposes.

Author's personal copy

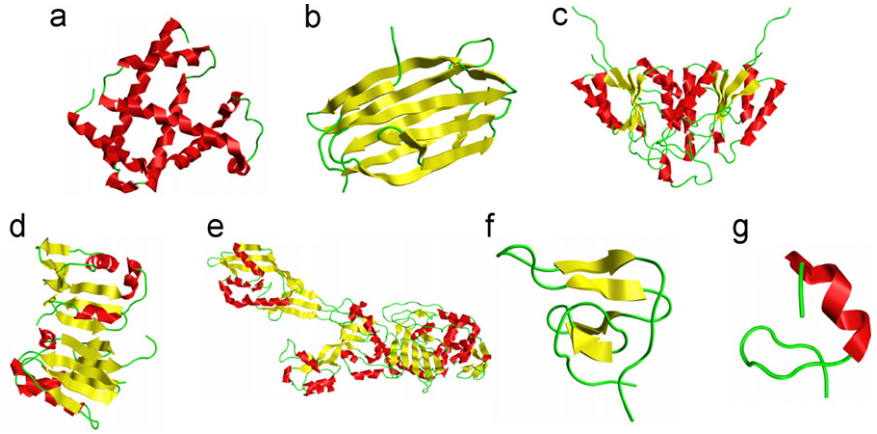


Fig. 2. Illustration to show the seven categories of protein structural class: (a) all-α, (b) all-β, (c) α/β, (d) α+β, (e) μ (multi-domain), (f) σ (small protein), and (g) ρ (peptide), where the α-helix is colored in red, β-strand in yellow, and the other in green. The PDB codes used to draw the representatives of the seven structural classes are 1a6m, 1uzv, 2f62, 2bf5, 1vqq, 4hir, and 1ter, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

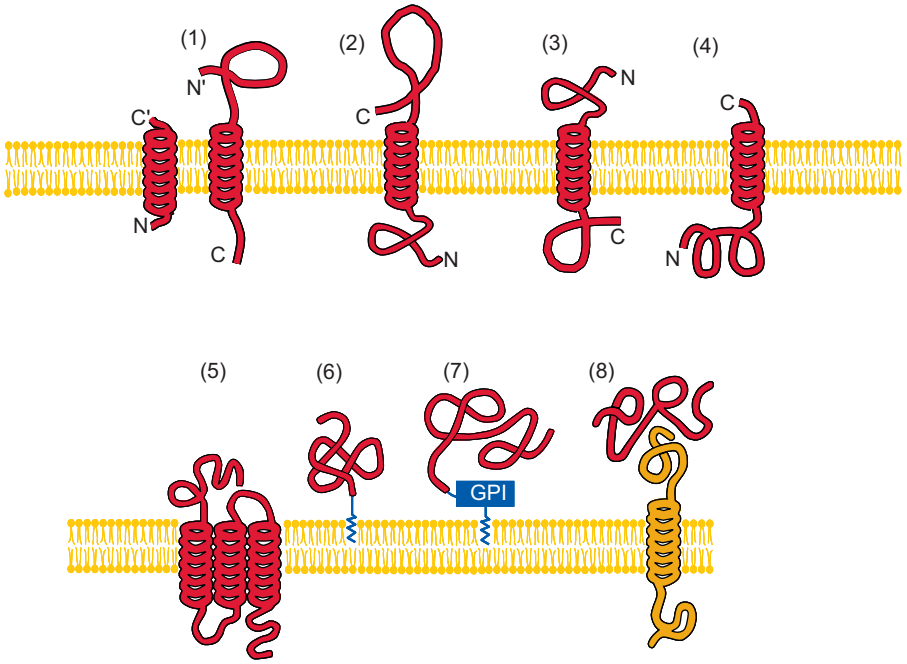


Fig. 3. Schematic drawings to show the eight categories of membrane protein types: (1) type I transmembrane, (2) type II, (3) type III, (4) type IV, (5) multipass transmembrane, (6) lipid-chain-anchored membrane, (7) GPI-anchored membrane, and (8) peripheral membrane. As shown in the figure, types I, II, III, and IV are all of single-pass transmembrane proteins; see Spiess (1995) for a detailed description about their difference.

3. Protein sample representation

Two kinds of models were usually used to represent protein samples. One is the sequential model, and the other the discrete model. The most straightforward sequential model for a protein sample is its entire amino acid sequence, as expressed by

$$\mathbf{P} = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \cdots R_L \tag{3}$$

where R_1 represents the 1st residue of the protein **P**, R_2 the 2nd residue,…, R_L the L-th residue, and they each belong to one of the 20 native amino acid types. To get the desired results, the sequence-similarity-search-based tools, such as BLAST (Altschul, 1997; Wootton and Federhen, 1993), are usually utilized to conduct the prediction. However, this kind of approach failed to work when the query protein did not have significant sequence similarity to any attribute-known proteins. Thus, various non-sequential models, or discrete models, were proposed, as illustrated below.

The simplest discrete model used to represent a protein sample is its amino acid (AA) composition or AAC (Nakashima et al., 1986). According to the AAC-discrete model, the protein **P** of Eq. (3) can be expressed by (Chou, 1995a)

$$\mathbf{P} = \begin{bmatrix} f_1 & f_2 & \cdots & f_{20} \end{bmatrix}^{\mathbf{T}} \tag{4}$$

where f_i (i=1, 2 ,…,20) are the normalized occurrence frequencies of the 20 native amino acids in **P**, and **T** the transposing operator. Many methods for predicting various protein attributes were based on the AAC-discrete model (see, e.g., Cedano et al., 1997; Chou, 1999, 2000, 2005b; Chou and Zhang, 1992, 1995; Chou and Maggiora, 1998; Chou and Elrod, 1999, 2002; Chou et al., 1998; Chou, 1989; Du et al., 2006; Feng et al., 2005; Jahandideh et al., 2007a; Klein, 1986; Klein and Delisi, 1986; Liu and Chou, 1998; Metfessel et al., 1993; Nakashima and Nishikawa, 1994; Niu et al., 2006; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). However, as one can see from Eq. (4), all the sequence-order effects would be missing using the AAC-discrete

Author's personal copy

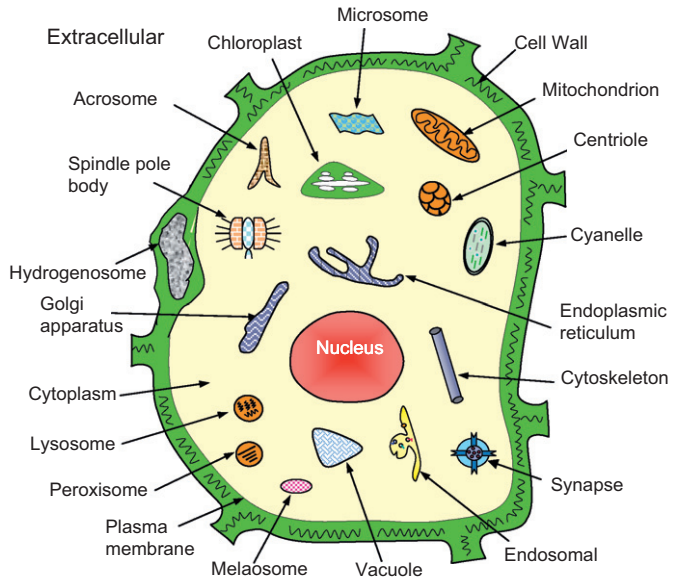


Fig. 4. Schematic illustration to show the 22 subcellular locations of eukaryotic proteins: (1) acrosome, (2) cell wall, (3) centriole, (4) chloroplast, (5) cyanelle, (6) cytoplasm, (7) cytoskeleton, (8) endoplasmic reticulum, (9) endosome, (10) extracellular, (11) Golgi apparatus, (12) hydrogenosome,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

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, a completely different discrete model, or the so-called "pseudo amino acid composition" (PseAAC) model (Chou, 2001), was proposed to represent the sample of a protein, as formulated by

$$\mathbf{P} = \left[p_1 \quad p_2 \quad \cdots \quad p_{20} \quad p_{20+1} \quad \cdots \quad p_{20+\Lambda} \right]^{\mathbf{T}} \tag{5}$$

where the first 20 elements are associated with the 20 elements in Eq. (4) or the 20 amino acid components of the protein, while the additional Λ factors are used to incorporate some sequence-order information via various modes. Typically, these additional factors are a series of rank-different correlation factors along a protein chain, but they can also be any combinations of other factors so long as they can reflect some sorts of sequence-order effects in one way or the other. For the convenience of users, a web-server called "PseAAC" (Shen and Chou, 2008) was established at http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/, by which some commonly used PseAAC forms can be automatically generated.

The concept of PseAAC has been widely used to study various problems in proteins and protein-related systems, such as predicting enzymes and their family/sub-family classification (Cai and Chou, 2005; Cai et al., 2005; Qiu et al., 2010; Wang et al., 2010b; Zhou et al., 2007), protein subcellular location prediction (Cai and Chou, 2003; Chou and Cai, 2003c, 2004e; Gao et al., 2005; Li and Li, 2008b; Pan et al., 2003; Shi et al., 2007, 2008; Xiao et al., 2006b; Zhang et al., 2008c), apoptosis protein subcellular location prediction (Chen and Li, 2007; Jiang et al., 2008b; Kandaswamy et al., 2010; Lin et al., 2009a; Liu et al., 2010b), mycobacterial protein subcellular location prediction (Lin et al., 2008), predicting protein subnuclear localization (Jiang et al., 2008a; Li and Li, 2008a; Shen and Chou, 2005b), predicting protein subchloroplast locations (Du et al., 2009), predicting protein submitochondria locations (Du and Li, 2006; Nanni and Lumini, 2008; Zeng et al., 2009), predicting membrane proteins and their types (Cai and Chou, 2006; Chou and Shen, 2007d; Liu et al., 2005; Shen and Chou, 2005a; Shen et al., 2006; Wang et al., 2004; Wang et al., 2006), discrimination of outer membrane proteins (Gao et al., 2010; Lin, 2008), identifying transmembrane regions in proteins (Diao et al., 2008), identifying proteases and their types (Chou and Shen, 2008a; Zhou and Cai, 2006), predicting protein solubility (Xiaohui et al., 2010), identifying GPCRs and their classes (Gu et al., 2010a, 2010b; Lin et al., 2009b; Qiu et al., 2009; Xiao et al., 2009b, 2010b), prediction of nuclear receptors (Gao et al., 2009), prediction of cyclin proteins (Mohabatkar, 2010), identifying bacterial secreted proteins (Yu et al., 2010), identifying risk type of human papillomaviruses (Esmaeili et al., 2010), prediction of cell wall lytic enzymes (Ding et al., 2009), prediction of lipases types (Zhang et al., 2008a), predicting conotoxin superfamily and family (Lin and Li, 2007a; Mondal et al., 2006), predicting the cofactors of oxidoreductases (Zhang and Fang, 2008), predicting DNA-binding proteins (Fang et al., 2008), predict protein structural classes (Chen et al., 2006a; Chen et al., 2006b; Ding et al., 2007; Li et al., 2009; Lin and Li, 2007b; Wu et al., 2010; Xiao et al., 2008a; Xiao et al., 2008b; Xiao et al., 2006a; Zhang and Ding, 2007; Zhang et al., 2008d), supersecondary structure prediction (Zou et al., 2011), protein secondary structure content prediction (Chen et al., 2009), predicting protein quaternary structural attributes (Chou and Cai, 2003a; Shen and Chou, 2009b; Xiao et al., 2009a; Xiao et al., 2010a; Zhang et al., 2008b; Zhang et al., 2006), fold pattern prediction (Shen and Chou, 2006; Shen and Chou, 2009a), and others (e.g., Georgiou et al., 2009).

Meanwhile, various modes of PseAAC by extracting different features from protein sequences were proposed, including stochastic signal processing mode (Pan et al., 2003), Fourier spectrum analysis mode (Liu et al., 2005), special functions mode (Gao et al., 2005), complexity measure factor mode (Xiao et al., 2005, 2006a), cellular automaton mode (Xiao et al., 2006b, 2008b, 2009b), geometric moments mode (Xiao et al., 2008b), gray dynamic mode (Xiao et al., 2008a), approximate entropy mode (Jiang et al., 2008a), continuous wavelet transform mode (Li et al., 2009), discrete wavelet transform mode (Qiu et al., 2009, 2010), sequence-segmented mode (Zhang et al., 2008b), evolutionary information and von Neumann entropy mode (Zhang et al., 2008c), and so forth.

However, according to its original concept, the essence of PseAAC 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 PseAAC 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 AAC and able to harbor some sort of sequence order or pattern information. Therefore, the PseAAC for a protein **P** should be generally formulated as

$$\mathbf{P} = \left[\psi_1 \quad \psi_2 \quad \cdots \quad \psi_u \quad \cdots \quad \psi_\Omega \right]^{\mathbf{T}} \tag{6}$$

where the subscript Ω is an integer, and its value and the components $\psi_1, \psi_2, \ldots$ will depend on how to extract the desired information from the amino acid sequence of **P** (cf. Eq. (3)). The form of Eq. (6) can cover all the aforementioned modes of PseAAC. For example, when

$$\psi_u = \begin{cases} f_u \Big/ \left(\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j \right), & (1 \le u \le 20) \\ w\theta_{u-20} \Big/ \left(\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j \right), & (20+1 \le u \le 20+\lambda = \Omega; \quad \lambda < L) \end{cases} \tag{7}$$

we immediately obtain the formulation of PseAAC originally introduced in Chou (2001), where the meanings for w, θ_j, and λ were clearly elaborated and hence there is no need to repeat here. When

$$\psi_u = \begin{cases} f_u \Big/ \left(\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j \right), & (1 \le u \le 20) \\ w\tau_{u-20} \Big/ \left(\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j \right) & (20+1 \le u \le 20+2\lambda = \Omega; \quad \lambda < L) \end{cases} \tag{8}$$

Author's personal copy

we obtain the formulation for the amphiphilic PseAAC (Chou, 2005a), where the meanings of w, τ_j, and λ were also clearly given.

It is instructive to point out that, with the general formulation of Eq. (6), the PseAAC can be used to reflect much more essential core features deeply hidden in complicated protein sequences through the following modes.

3.1. Functional domain mode

The functional domain (FunD) is the core of a protein. Therefore, in determining the 3-D (dimensional) structure of a protein by experiments (see, e.g., Call et al., 2010; Pielak and Chou, 2010; Schnell and Chou, 2008; Wang et al., 2009) or by computational modeling (see, e.g., Chou, 2004a; Chou, 2004b), the first priority was always focused on its FunD.

Using the FunD information to formulate protein samples was originally proposed in Cai et al. (2003) and Chou and Cai (2002) based on the 2005 FunDs in the SBASE-A database (Murvai et al., 2001). Since then, a series of new protein FunD databases were established, such as COG (Tatusov et al., 2003), KOG (Tatusov et al., 2003), SMART (Letunic et al., 2006), Pfam (Finn et al., 2006), and CDD (Marchler-Bauer et al., 2007). Of these databases, CDD contains the domains imported from COG, Pfam, and SMART, and hence is relatively much more complete (Marchler-Bauer et al., 2007) and was adopted in most of the recent publications (see, e.g., Chou and Shen, 2010a, 2010c; Shen and Chou, 2009d). The version 2.11 of CDD contains 17,402 characteristic domains. Thus, when using the general formulation of PseAAC (Eq. (6)) to incorporate the FunD information, we have $\Omega=17{,}402$, i.e.

$$\mathbf{P}_{\mathrm{FunD}}=\left[\psi_1^{\mathrm{D}}\quad\psi_2^{\mathrm{D}}\quad\cdots\quad\psi_u^{\mathrm{D}}\quad\cdots\quad\psi_{17402}^{\mathrm{D}}\right]^{\mathbf{T}} \tag{9}$$

where $\mathbf{T}$ has the same meaning as in Eq. (4), and

$$\psi_u^{\mathrm{D}}=\begin{cases}1, & \text{when a hit is found for } \mathbf{P} \text{ in CDD}\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

For the detailed procedure of how to find the hit for $\mathbf{P}$ in CDD, refer to Chou and Shen (2010a).

Similar approaches of representing protein samples with the FunD mode were also used for predicting protein subcellular localization (Chou and Cai, 2002; Chou and Cai, 2004d), membrane protein types (Cai and Chou, 2006; Cai et al., 2003), enzyme functional classes (Shen and Chou, 2007a), protease types (Chou and Shen, 2008a; Shen and Chou, 2009c), GPCRs types (Xiao et al., 2010b), protein structural class (Chou and Cai, 2004b), protein fold pattern (Shen and Chou, 2009a), and protein quaternary structural attributes (Shen and Chou, 2009b; Xiao et al., 2009a, 2010a).

3.2. Gene ontology mode

Gene ontology (GO) database (Ashburner et al., 2000) 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 some of their important attributes, such as subcellular localization and biological function (Chou and Shen, 2007c, 2008b).

The GO database (version 70.0 released 10 March 2008) contains 60,020 GO numbers. Thus, when using the general formulation of PseAAC to incorporate the GO information, we have $\Omega=60{,}020$, i.e.

$$\mathbf{P}_{\mathrm{GO}}=\left[\psi_1^{\mathrm{G}}\quad\psi_2^{\mathrm{G}}\quad\cdots\quad\psi_u^{\mathrm{G}}\quad\cdots\quad\psi_{60020}^{\mathrm{G}}\right]^{\mathbf{T}} \tag{11}$$

where

$$\psi_u^{\mathrm{G}}=\begin{cases}1, & \text{if a hit is found against the } u\text{-th GO number for protein } \mathbf{P}\\ 0, & \text{otherwise}\end{cases} \tag{12}$$

For the detailed procedure of how to find the hit for $\mathbf{P}$ in the GO database, refer to Chou and Shen (2010a).

The information extracted from the GO database (Ashburner et al., 2000; Camon et al., 2004; Harris et al., 2004) was used to formulate PseAAC for predicting protein subcellular localization (Cai and Chou, 2003; Chou and Cai, 2003b; Chou and Cai, 2004d; Chou and Shen, 2006a; Chou and Shen, 2006b; Chou and Shen, 2006c; Chou and Shen, 2007a; Chou and Shen, 2007b; Chou and Shen, 2007c; Chou and Shen, 2008b; Lee et al., 2005; Shen and Chou, 2007b; Shen and Chou, 2007c; Shen and Chou, 2007d; Shen et al., 2007), enzyme functional class (Chou and Cai, 2004a; Chou and Cai, 2004c), membrane protein types (Chou and Cai, 2005), protease types (Zhou and Cai, 2006), and protein–protein interactions (Chou and Cai, 2006).

3.3. Sequential evolution mode

Biology is a natural science with historic dimension. All biological species have developed continuously starting out from a very limited number of ancestral species. It is true for protein sequence as well (Chou, 2004b). Their evolution involves changes of single residues, insertions, and deletions of several residues (Chou, 1995b), gene doubling, and gene fusion. With these changes accumulated for a long period of time, many similarities between initial and resultant amino acid sequences are gradually eliminated, but the corresponding proteins may still share many common attributes, such as having basically the same biological function and residing in the same subcellular location.

The general formulation of PseAAC can be used to incorporate this kind of information via its sequential evolution mode, i.e.

$$\mathbf{P}_{\mathrm{Evo}}^{\lambda}=\left[\psi_1^{\lambda}\quad\psi_2^{\lambda}\quad\cdots\quad\psi_u^{\lambda}\quad\cdots\quad\psi_{\Omega}^{\lambda}\right]^{\mathbf{T}} \tag{13}$$

where

$$\psi_u^{\lambda}=\begin{cases}\dfrac{1}{L}\sum_{i=1}^{L}E_{i\to u}, & (u=1,2,\ldots,20)\\[2ex] \dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}[E_{i\to(u-20)}-E_{(i+\lambda)\to(u-20)}]^2, & (u=21,22,\ldots,40=\Omega;\quad \lambda<L)\end{cases} \tag{14}$$

where λ is an uncertain number that will be further discussed later, L is the length of $\mathbf{P}$ (counted in the total number of its constituent amino acids), and $E_{i\to j}$ 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 evolutionary process (Schaffer et al., 2001), which can be derived by using PSI-BLAST (Schaffer et al., 2001) to search the Swiss-Prot database as described in Chou and Shen (2010c). 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 above equations were used to identify membrane proteins and their types (Chou and Shen, 2007d), enzymes and their functional classes (Shen and Chou, 2007a), proteases and their types (Chou and Shen, 2008a), protein quaternary structural attributes (Shen and Chou, 2009b), as well as protein subcellular localization (Chou and Shen, 2010a; Chou and Shen, 2010b).

Besides the aforementioned PseAAC modes, there may be some other feature extraction methods to represent protein samples, but they can always be formulated with the form of Eq. (6), the general formulation of PseAAC.

It is instructive to point out that, regardless of which kind of PseAAC mode is adopted for protein samples, the query proteins and the

Author's personal copy

proteins used to train the prediction engine must be defined in the same infrastructural frame with exactly the same dimension. For instance, if a query protein is defined in the 17402-D FunD space (see Eq. (9)), then the prediction should be carried out based on those proteins in the training set that can be defined in the exactly same 17402-D FunD space as well. If a query protein is defined in the 60020-D GO space (see Eq. (11)), then the prediction should be carried out based on those proteins in the training set that can be defined in the exactly same 60020-D GO space as well. If the query protein in both the 17402-D FunD space and 60020-D GO space is a naught vector and hence must be defined instead in the sequential evolution space (see Eq. (13)), then all the proteins used to train the prediction engine must also be formulated in the same sequential evolution space. It is particularly important to follow such a self-consistency principle when hybridizing different PseAAC modes or building an ensemble classifier by fusing many individual classifiers (Chou and Shen, 2006d).

4. Prediction algorithm (operating engine)

The problem of predicting protein attributes can be generally described as follows. Suppose a system containing N proteins ($\mathbf{P}_1,\mathbf{P}_2,\ldots,\mathbf{P}_N$), which have been classified into M subsets (categories) as formulated by Eq. (1), where each subset S_m ($m=1,2,\ldots,M$) is composed of proteins with the same attribute category and its size (the number of proteins therein) is N_m. Obviously, we have $N=N_1+N_2+\cdots+N_M$. According to Eq. (6), we can suppose without losing generality that the k-th protein in the subset S_m (see Eq. (1)) is expressed by

$$\mathbf{P}_m^k=\left[\begin{array}{ccccccc}\psi_{m,1}^k & \psi_{m,2}^k & \cdots & \psi_{m,j}^k & \cdots & \psi_{m,\Omega}^k\end{array}\right]^{\mathbf{T}} \tag{15}$$

where $\psi_{m,j}^k\,(j=1,2,\ldots,\Omega)$ is the j-th component of the k-th protein in $\mathbb{S}_m$. Now, for a query protein $\mathbf{P}$ as defined by Eq. (6), how can we identify which subset it belongs to?

Many different prediction algorithms have been introduced to address this problem, such as discriminant algorithm (Chou and Maggiora, 1998; Chou and Elrod, 1999), neural network algorithm (Cai et al., 2000; Cai et al., 2001), support vector machine (SVM) (Cai et al., 2003; Cai et al., 2004; Chou and Cai, 2002), and K-nearest Neighbor algorithm (Cai and Chou, 2003; Chou and Shen, 2006b). In this paper we shall focus on the K-nearest neighbor algorithm (Denoeux, 1995) and show how to generate a powerful ensemble classifier by fusing many individual basic classifiers characterized with different control parameters.

The K-nearest neighbor (KNN) classifier is quite popular in pattern recognition community owing to its good performance and simple-to-use feature. According to the KNN rule (Denoeux, 1995; Keller et al., 1985), named also as the "voting KNN rule", the query protein should be assigned to the subset represented by a majority of its K nearest neighbors, as illustrated in Fig. 5

There are many different definitions to measure the "nearness" for the KNN classifier, such as Euclidean distance, Hamming distance (Mardia et al., 1979), and Mahalanobis distance (Chou, 1995a; Mahalanobis, 1936; Pillai, 1985). Usually, the following equation was adopted to measure the nearness between proteins $\mathbf{P}$ and $\mathbf{P}_m^k$ (cf. Eqs. (6) and (15)):

$$D(\mathbf{P},\mathbf{P}_m^k)=1-\frac{\mathbf{P}\cdot\mathbf{P}_m^k}{\|\mathbf{P}\|\,\|\mathbf{P}_m^k\|} \tag{16}$$

where $\mathbf{P}\cdot\mathbf{P}_m^k$ is the dot product of the two vectors, and $\|\mathbf{P}\|$ and $\|\mathbf{P}_m^k\|$ their modulus, respectively. According to Eq. (16), when $\mathbf{P}\equiv\mathbf{P}_m^k$ we have $D(\mathbf{P},\mathbf{P}_m^k)=0$, indicating the "distance" between these two proteins is zero and hence they have perfect or 100% similarity. In using the KNN rule, the predicted result will depend on the selection of the parameter K, the number of the nearest neighbors to the query protein $\mathbf{P}$, as described below.

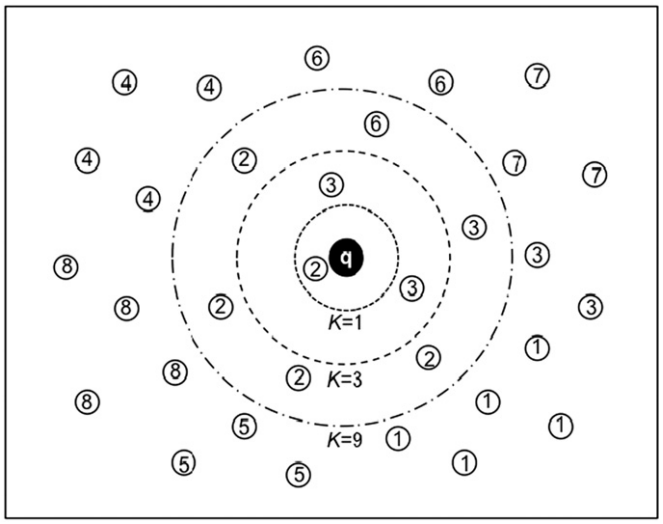


Fig. 5. Illustration to show how the KNN classifier depends on the selection of parameter K in identifying the attribute category of a query protein, where the query protein $\mathbf{P}$ is represented by the character **q** with a filled circle, proteins belonging to subset $\mathbb{S}_1$ (category 1) are represented by the open circle with number 1, proteins of $\mathbb{S}_2$ by the open circle with number 2, and so forth. When $K=1$, the query protein is predicted belonging to category 2 as its nearest protein does; when $K=3$, the query protein is predicted belonging to category 3 because two of its three nearest proteins belong to that category; when $K=9$, the query protein is predicted belonging to category 2 again because the majority of its nine nearest proteins belong to category 2.

4.1. Nearest neighbor classifier

The nearest neighbor classifier (Cover and Hart, 1967), also called NN classifier, is a special case of KNN classifier with $K=1$ (Fig. 5). With the NN classifier, the protein $\mathbf{P}$ will be predicted belonging to the same attribute category of the protein in the learning dataset that has the shortest "distance" to $\mathbf{P}$, i.e., the query protein will be classified in the μ-th attribute category if

$$\mu=\arg\ \min_m\left\{\min_{\mathbf{P}_m^k\in\mathbb{S}_m}\left[D(\mathbf{P},\mathbf{P}_m^k)\right]\right\},\quad (m=1,2,\ldots,M) \tag{17}$$

where $\min_{\mathbf{P}_m^k\in\mathbb{S}_m}$ means taking the minimum value of $D(\mathbf{P},\mathbf{P}_m^k)$ for the proteins in the subset $\mathbb{S}_m$ (cf. Eqs. (1) and (16)), and the operator $arg\ \min_m$ means taking the argument of m that minimizes the quantity right after the operator. In other words, μ in Eq. (17) is equal to the argument of m that minimizes $\left\{\min_{\mathbf{P}_m^k\in\mathbb{S}_m}\left[D(\mathbf{P},\mathbf{P}_m^k)\right]\right\}$. If there are two and more arguments leading to the same minimum value, the query protein will be randomly assigned to one of the subsets associated with these arguments although this kind of tie case rarely happens. Owing to its simplicity and apparent efficiency, the NN classifier is still a favorite method used by many investigators (see, e.g., Chen et al., 2010; He et al., 2010; Huang et al., 2010).

4.2. KNN classifier

With the KNN classifier when $K>1$, the attribute of the query protein $\mathbf{P}$ will be determined by the majority of its K nearest neighbors via a vote (Fig. 5), as can be formulated as follows. Suppose ($\mathbf{P}_1^*,\mathbf{P}_2^*,\ldots,\mathbf{P}_K^*$) are the K proteins in $\mathbb{S}$ that have the closest distances to $\mathbf{P}$, the query protein will be predicted belonging to

Author's personal copy

the μ-th subset (attribute category) if

$$\mu = \arg\ \max_m \left\{ \sum_{i=1}^{K} \Delta(\mathbf{P}_i^*, \mathbb{S}_m) \right\}, \quad (m = 1,2,\ldots,M) \tag{18}$$

where μ is the argument of m that maximize $\left\{ \sum_{i=1}^{K} \Delta(\mathbf{P}_i^*, \mathbb{S}_m) \right\}$ and

$$\Delta(\mathbf{P}_i^*, \mathbb{S}_m) = \begin{cases} 1, & \text{if } \mathbf{P}_i^* \in \mathbb{S}_m \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

where $\in$ is a symbol in the set theory meaning "member of". If there is a tie for the voting results, the query protein will be randomly assigned to one of the locations associated with the tie case. Generally speaking, the greater the K (the number of the nearest neighbors counted), the less likely the tie case occurs.

As mentioned above, the sequential evolution PseAAC mode of Eq. (13) contains a parameter λ, which is associated with what tier of sequence correlation is taken into account for the PseAAC. As we can see from Eq. (14), the only constraint to λ is that it must be smaller than L, the number of the amino acids in the protein concerned. Suppose the length of the shortest protein investigated is 50, then λ can be any of the following 50 numbers: 0, 1, 2,…,49. Although in principle we can include all these possibilities for λ by enlarging the dimension of the PseAAC to contain $20 \times 50 = 1000$ components, it may cause various unfavorable problems for statistical prediction, such as "high dimension disaster" and "overfitting redundancy" (Wang et al., 2008a). Actually, it may reduce the cluster-tolerant capacity (Chou, 1999) and lower down the success rate of cross-validation if the PseAAC contains too many trivial components. Accordingly, for a given training dataset, there is an optimal number for λ. However, it would be time-consuming and tedious to find the optimal λ by changing its value and doing tests one-by-one.

Likewise, the KNN classifier (cf. Eq. (18)) also contains a parameter K, the number of the nearest neighbors to a query protein (Fig. 5). It will affect the predicted result by choosing a different value for K. In other words, for a given training dataset, there is an optimal value for K as well.

The parameters such as λ and K are called uncertain parameters. The number of the uncertain parameters depends on which model is used to represent the protein samples and what classifier is used for the prediction engine. It can be seen from Eqs. (9), (11), (13), and (18) that one uncertain parameter, K, needs to be determined if using KNN classifier based on the FunD (or GO) mode of PseAAC, and that two uncertain parameters, K and λ, need to be determined if using KNN classifier based on the sequential evolution mode. It would be much more tedious and time-consuming to determine the optimal values for two uncertain parameters. To deal with this kind of uncertain parameters, let us introduce the fusion approach.

4.3. One-dimensional fusion

For most cases in using the KNN classifier to predict protein attributes, when $K > 20$, the success rate by the KNN classifier would decrease remarkably. Therefore, the basic individual classifiers to be considered can be generally expressed as

$$\mathbb{KNN} \rhd \mathbf{P} = \mathfrak{M}_1(K,\mathbf{P}) \in \mathbb{S}, \quad \text{when } \mathbf{P} \text{ is in the } \mathbf{P}_{\text{FunD}} \text{ or } \mathbf{P}_{\text{GO}} \text{ mode} \quad (K = 1,2,\ldots,20) \tag{20}$$

where $\mathbb{KNN}$ represents the KNN classifier that is a function of K, the symbol $\rhd$ is the identification operator meaning using $\mathbb{KNN}$ to identify the attribute of the query protein $\mathbf{P}$ among the M subsets of $\mathbb{S}$ in Eq. (1). Suppose the accumulated score thus obtained (with K=1,2,…,20) for the protein $\mathbf{P}$ belonging to the m-th subset $\mathbb{S}_m \in \mathbb{S}$ is given by

$$Y_m^{(1)}(\mathbf{P}) = \sum_{K=1}^{20} \Delta\left[\mathfrak{M}_1(K,\mathbf{P}),\ \mathbb{S}_m\right], \quad (m = 1,2,\ldots,M) \tag{21}$$

where

$$\Delta\left[\mathfrak{M}_1(K,\mathbf{P}),\ \mathbb{S}_m\right] = \begin{cases} 1, & \text{if } \mathfrak{M}_1(K,\mathbf{P}) \in \mathbb{S}_m \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

Thus the query protein $\mathbf{P}$ is predicted belonging to the subset with which its score of Eq. (21) is the highest, i.e., the query protein $\mathbf{P}$ is identified as belonging to the μ-th subset if

$$\mu = \arg\ \max_m \{Y_m^{(1)}(\mathbf{P})\}, \quad (m = 1,2,\ldots,M) \tag{23}$$

where μ is the argument of m that maximizes the score function $Y_m^{(1)}$ of Eq. (21). If there are two and more arguments leading to the same maximum value, the query protein will be randomly assigned to one of the subset associated with these arguments although this kind of tie case rarely happens.

4.4. Two-dimensional fusion

When the KNN classifier is operated on the query protein formulated with the sequential evolution mode (cf. Eq. (13)), we are facing a problem with two uncertain parameters, K and λ. In general, the shortest protein sequence investigated is 50 amino acids (Chou and Shen, 2008a; Chou and Shen, 2010c), hence we can set the maximum value allowed for λ is 49. Thus, the basic individual classifiers to be considered would become as follows:

$$\mathbb{KNN} \rhd \mathbf{P} = \mathfrak{M}_2(K,\lambda,\mathbf{P}) \in \mathbb{S}, \quad \text{when } \mathbf{P} \text{ is in the } \mathbf{P}_{\text{Evo}}^{\lambda} \text{ mode} \quad (K = 1,2,\ldots,20;\ \lambda = 0,1,2,\ldots,49) \tag{24}$$

and the corresponding accumulated score for the query protein $\mathbf{P}_{\text{Evo}}^{\lambda}$ belonging to the m-th subset $\mathbb{S}_m \in \mathbb{S}$ is given by

$$Y_m^{(2)}(\mathbf{P}) = \sum_{\lambda=0}^{49} \sum_{K=1}^{20} \Delta\left[\mathfrak{M}_2(K,\lambda,\mathbf{P}),\ \mathbb{S}_m\right], \quad (m = 1,2,\ldots,M) \tag{25}$$

where

$$\Delta\left[\mathfrak{M}_2(K,\lambda,\mathbf{P}),\ \mathbb{S}_m\right] = \begin{cases} 1, & \text{if } \mathfrak{M}_2(K,\lambda,\mathbf{P}) \in \mathbb{S}_m \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

and the query protein $\mathbf{P}_{\text{Evo}}^{\lambda}$ is predicted belonging to the subset with which its score of Eq. (25) is the highest, i.e., the query protein $\mathbf{P}$ is identified as belonging to the μ-th subset if

$$\mu = \arg\ \max_m \{Y_m^{(2)}(\mathbf{P})\}, \quad (m = 1,2,\ldots,M) \tag{27}$$

where μ is the argument of m that maximizes the score function $Y_m^{(2)}(\mathbf{P})$ of Eq. (25). If there are two and more arguments leading to the same maximum value, the query protein will be randomly assigned to one of the subcellular locations associated with these arguments although this kind of tie case rarely happens.

If a basic individual classifier involves with three or more uncertain parameters, by following the similar procedures as described above, we can perform three or higher dimensional fusion.

5. Cross-validation test

After a prediction method has been developed, a subsequent and natural question to ask is: What is its accuracy?

In statistical prediction, it would be meaningless to simply say a success rate of a predictor without specifying what cross-validation method and benchmark dataset were used to test its accuracy.

Author's personal copy

In literatures, the following three cross-validation methods are generally used for examining the effectiveness of a statistical prediction method: (1) the independent dataset test, (2) the subsampling (Γ-fold such as 5- or 10-fold cross-validation) test, and (3) the jackknife test (Chou and Zhang, 1995).

For the independent dataset test, although all the proteins used to test the predictor are outside the training dataset used to train it so as to exclude the "memory" effect or bias, the way of how to select the independent proteins to test the predictor could be quite arbitrary unless the number of independent proteins is sufficiently large. This kind of arbitrariness might result in completely different conclusions. For instance, a predictor achieving a higher success rate than the other predictor for a given independent testing dataset might fail to keep so when tested by another independent testing dataset (Chou and Zhang, 1995). Accordingly, the independent dataset test is not a fairly objective test method although it was often used to demonstrate the practical application of a predictor (see, e.g., Cedano et al., 1997; Chou and Elrod, 1999; Chou and Shen, 2006c; Chou and Shen, 2007a).

For the subsampling test, the concrete procedure usually used in literatures is the 5-fold, 7-fold, or 10-fold cross-validation. The problem with the Γ-fold cross-validation test as such is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset. This is because for a benchmark dataset as formulated in Eq. (1), the number of possible combinations of taking one Γ-th or $1/\Gamma$ proteins from each of the subsets in Eq. (1) will be

$$\Pi = \Pi_1 \cdot \Pi_2 \cdots \Pi_m \cdots \Pi_M \tag{28}$$

where

$$\Pi_m = \frac{N_m!}{[N_m - \mathrm{Int}(N_m/\Gamma)]!\ \mathrm{Int}(N_m/\Gamma)!}, \quad (m = 1,2,\ldots,M) \tag{29}$$

where N_m is the number of proteins in the m-th subset $\mathbb{S}_m$, and the symbol Int is the integer-truncating operator meaning to take the integer part for the number in the brackets right after it.

For example, without losing generality let us consider the case of 5-fold cross-validation (i.e., $\Gamma = 5$) for a very simple benchmark dataset that contains 250 proteins, of which $N_1 = 65$ belongs to subset $\mathbb{S}_1$, $N_2 = 60$ to subset $\mathbb{S}_2$, $N_3 = 55$ to subset $\mathbb{S}_3$, and $N_4 = 70$ to subset $\mathbb{S}_4$. Substituting these figures into Eqs. (28–29), we have that the number of possible combinations of taking one-fifth proteins from each of the four subsets will be

$$\begin{aligned}\Pi &= \Pi_1 \cdot \Pi_2 \cdot \Pi_3 \cdot \Pi_4 \\ &= \frac{65!}{(65-13)!13!} \cdot \frac{60!}{(60-12)!12!} \cdot \frac{55!}{(55-11)!11!} \cdot \frac{70!}{(70-14)!14!} \\ &> 5.3135 \times 10^{50}\end{aligned} \tag{30}$$

indicating that for such a simple and small benchmark dataset, the number of possible combinations of taking one-fifth proteins from each of the four subsets for 5-fold cross-validation will be an astronomical number.

Now let us consider a moderate-size dataset that consists of 640 proteins classified into $M = 8$ subsets with each containing 80 proteins, i.e., $N_1 = N_2 = \cdots = N_8 = 80$. According to Eqs. (28–29), the number of possible combinations of taking one-fifth proteins from each of the 8 subsets for 5-fold-cross-validation will be

$$\Pi = \Pi_1 \cdot \Pi_2 \cdot \Pi_3 \cdots \Pi_8 = \left(\frac{80!}{(80-16)!16!}\right)^8 > 2.7907 \times 10^{131} \tag{31}$$

If the above benchmark dataset is slightly larger and complicated, i.e., the number of proteins is increased from 640 to 800, and the number of subsets from 8 to 10 with each still containing 80 proteins, then the number of possible combinations of taking one-fifth proteins from each of the 10 subsets for 5-fold-cross-validation will be

$$\begin{aligned}\Pi = \Pi_1 \cdot \Pi_2 \cdot \Pi_3 \cdots \Pi_{10} &= \left(\frac{80!}{(80-16)!16!}\right)^{10} \\ &> \text{the maximum number allowed to be calulated in a computer}\end{aligned} \tag{32}$$

Actually, many typical benchmark datasets contain more than 1000 proteins (see, e.g., Chou and Shen, 2008a; Chou and Shen, 2010a; Chou and Shen, 2010c). Therefore, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark dataset and a same predictor, the subsampling test (such as 5-fold cross-validation) cannot avoid the arbitrariness either. A test method unable to yield a unique outcome cannot be deemed as an ideal one.

In the jackknife test, all the proteins in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining protein samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each protein sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the arbitrariness problem as mentioned above for the independent dataset test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. As for the possible overestimation in success rate by jackknife test because of only one sample being singled out at a time for testing, the answer is that as long as the jackknife test is performed on a stringent benchmark dataset in which none of proteins has $\geq 25\%$ pairwise sequence identity to any other in a same subset such as those mentioned in the Section 2, it is highly unlikely to yield an overestimated rate compared with the actual success rate in practical applications, as demonstrated in Chou and Shen (2010c) and Shen and Chou (2010). Besides, when the jackknife test was used to compare two predictors, even if there was some overestimate due to using a less stringent benchmark dataset for one predictor, the same overestimate would exist for the other as long as they were both tested by the same dataset.

Accordingly, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors (see, e.g., Anand and Suganthan, 2009; Cai et al., 2010; Chen et al., 2008a; Chen et al., 2008b; Chen and Han, 2009; Du and Li, 2008; Du et al., 2009; Fang et al., 2008; Feng and Luo, 2008; Gu and Chen, 2009; Gu et al., 2010a; Jahandideh et al., 2007a; Jahandideh et al., 2007b; Jahandideh et al., 2009; Ji et al., 2010; Kannan et al., 2008; Li et al., 2009; Lin, 2008; Lin et al., 2009a; Liu et al., 2010a; Munteanu et al., 2008; Nanni and Lumini, 2008; Nanni and Lumini, 2009; Rezaei et al., 2008; Shao et al., 2009; Shi et al., 2008; Shi and Hu, 2010; Vilar et al., 2009; Wang and Yang, 2010; Wang et al., 2010a; Wang et al., 2008b; Yang and Jiang, 2010; Yang et al., 2009; Yang et al., 2010; Zhao et al., 2008; Zhou et al., 2008).

However, even if using the jackknife approach for cross-validation, the same predictor may still generate obviously different success rates when tested by different benchmark datasets. This is because the more the stringent of a benchmark dataset in excluding homologous and high similarity sequences, the more the difficult for a predictor to achieve a high overall success rate (Chou and Shen, 2010a). Also, the more the number of subsets (attribute categories) a benchmark dataset covers, the more the difficult to achieve a high overall success rate. This can be easily conceivable via the following consideration. Suppose a benchmark dataset consists of two subsets (attribute categories) with each containing the same number of proteins. The overall success rate in identifying their attribute categories by random assignment would be $1/2 = 50\%$. However, for a benchmark dataset consisting of 20

Author's personal copy

subsets, the corresponding overall success rate by the random assignment would be 1/20 = 5%, which is only one-tenth of the former.

6. Web-server

Even if a powerful predictor has been developed by accomplishing the above four procedures, namely constructing a valid benchmark dataset, formulating protein samples with PseAAC to successfully catch their essential and core features, introducing a powerful and efficient algorithm or engine to operate the prediction, and achieving a high overall success rate by jackknife test on a stringent dataset in which none of the proteins included has $\geq 25\%$ pairwise sequence identity to any other in the same subset (attribute category), it does not mean that the predictor has been really completed. This is because we are living in the Internet Age. To make a new prediction method really useful for the majority of people, it is an important direction or necessary procedure to provide a user-friendly and publicly accessible web-server for the method (Chou and Shen, 2009). Technically speaking, a web-server means a computer program that is responsible for accepting Hypertext Transfer Protocol (HTTP) requests from clients. By means of web-servers, many computational prediction methods, regardless how difficult their mathematics or how complicated their algorithms are, can be easily used by the vast majority of scientists to generate their desired data without the need to understand the mathematical details.

7. Conclusion and perspectives

In order to timely utilize the huge amount of newly discovered protein sequences generated in the postgenomic era for basic research and drug development, scientists are anxious to know their biological attributes. Many studies from various research laboratories around the world have indicated that mathematical analysis, computational modeling, and introducing novel physical concept to biology and medicine, such as graphical analysis (Andraos, 2008; Myers and Palmer, 1985; Zhou and Deng, 1984), modeling three-dimensional structures of targeted proteins/peptides for drug design (Sharma et al., 2008; Zhou and Troy, 2003; Zhou and Troy, 2005a; Zhou and Troy, 2005b; Zhou et al., 2004), diffusion-controlled reaction simulation (Zhou et al., 1981; Zhou and Zhong, 1982; Zhou et al., 1983), cellular responding kinetics (Qi et al., 2007), and biological functions of solitons in DNA (Zhou, 1989) can provide useful insights for both basic research and drug design and hence are widely welcome by science community. In view of this, it is highly desirable to develop automated methods by introducing new concepts and approaches for fast and accurately predicting the attributes of uncharacterized proteins based on their sequence information alone. During the past two decades or so, many statistical methods for predicting various protein attributes have been proposed. In this review, the key steps for establishing a powerful predictor in this regard have been analyzed in hopes that the points raised here may help stimulate the further development of new and more powerful predictors in this area. It is anticipated that the general form of PseAAC as formulated in this review may further stimulate the efforts to find various new modes of optimal PseAAC, which is one of the most important future directions we should focus on in order to substantially improve the power of predicting protein attributes.

Acknowledgements

The author would like to thank Professor Denise Kirschner and Dr. Dale Seaton of Elseivier for inviting him to write this review article. The author would also like to thank the two anonymous reviewers for their constructive comments in strengthening the presentation of this review.

References

Altschul, S.F., 1997. Evaluating the statistical significance of multiple distinct local alignments. In: Suhai, S. (Ed.), Theoretical and Computational Methods in Genome Research. Plenum, New York, pp. 1–14.
Anand, A., Suganthan, P.N., 2009. Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates. J. Theor. Biol. 259, 533–540.
Andraos, J., 2008. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can. J. Chem. 86, 342–357.
Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.
Cai, Y.D., Chou, K.C., 2003. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem. Biophys. Res. Commun. 305, 407–411.
Cai, Y.D., Chou, K.C., 2005. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J. Proteome Res. 4, 967–971.
Cai, Y.D., Chou, K.C., 2006. Predicting membrane protein type by functional domain composition and pseudo amino acid composition. J. Theor. Biol. 238, 395–400.
Cai, Y.D., Li, Y.X., Chou, K.C., 2000. Using neural networks for prediction of domain structural classes. Biochim. Biophys. Acta 1476, 1–2.
Cai, Y.D., Liu, X.J., Chou, K.C., 2001. Artificial neural network model for predicting membrane protein types. J. Biomol. Struct. Dynam. 18, 607–610.
Cai, Y.D., Zhou, G.P., Chou, K.C., 2003. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys. J. 84, 3257–3263.
Cai, Y.D., Zhou, G.P., Chou, K.C., 2005. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J. Theor. Biol. 234, 145–149.
Cai, Y.D., Pong-Wong, R., Feng, K., Jen, J.C.H., Chou, K.C., 2004. Application of SVM to predict membrane protein types. J. Theor. Biol. 226, 373–376.
Cai, Y.D., He, J., Li, X., Feng, K., Lu, L., Kong, X., Lu, W., 2010. Predicting protein subcellular locations with feature selection and analysis. Protein Pept. Lett. 17, 464–472.
Call, M.E., Wucherpfennig, K.W., Chou, J.J., 2010. The structural basis for intramembrane assembly of an activating immunoreceptor complex. Nat. Immunol. 11, 1023–1029.
Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R., 2004. The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. Nucl. Acids Res. 32, D262-6.
Cedano, J., Aloy, P., P'erez-Pons, J.A., Querol, E., 1997. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. 266, 594–600.
Chen, C., Chen, L.X., Zou, X.Y., Cai, P.X., 2008a. Predicting protein structural class based on multi-features fusion. J. Theor. Biol. 253, 388–392.
Chen, C., Chen, L., Zou, X., Cai, P., 2009. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept. Lett. 16, 27–31.
Chen, C., Zhou, X., Tian, Y., Zou, X., Cai, P., 2006a. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal. Biochem. 357, 116–121.
Chen, C., Tian, Y.X., Zou, X.Y., Cai, P.X., Mo, J.Y., 2006b. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J. Theor. Biol. 243, 444–448.
Chen, K., Kurgan, L.A., Ruan, J., 2008b. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J. Comput. Chem. 29, 1596–1604.
Chen, L., Feng, K.Y., Cai, Y.D., Chou, K.C., Li, H.P., 2010. Predicting the network of substrate–enzyme-product triads by combining compound similarity and functional domain composition. BMC Bioinform. 11, 293.
Chen, Y., Han, K., 2009. BSFINDER: finding binding sites of HCV proteins using a support vector machine. Protein Pept. Lett. 16, 373–382.
Chen, Y.L., Li, Q.Z., 2007. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J. Theor. Biol. 248, 377–381.
Chou, K.C., 1995a. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Struct. Funct. Genet. 21, 319–344.
Chou, K.C., 1995b. The convergence–divergence duality in lectin domains of the selectin family and its implications. FEBS Lett. 363, 123–126.
Chou, K.C., 1999. A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun. 264, 216–224.
Chou, K.C., 2000. Review: prediction of protein structural classes and subcellular locations. Curr. Protein Pept. Sci. 1, 171–208.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. 43, 246–255 (Erratum: ibid., 2001, vol. 44, p. 60).
Chou, K.C., 2004a. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochem. Biophys. Res. Commun. 316, 636–642.

Author's personal copy

Chou, K.C., 2004b. Review: structural bioinformatics and its impact to biomedical science. Curr. Med. Chem. 11, 2105–2134.
Chou, K.C., 2005a. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
Chou, K.C., 2005b. Prediction of G-protein-coupled receptor classes. J. Proteome Res. 4, 1413–1418.
Chou, K.C., Zhang, C.T., 1992. A correlation coefficient method to predicting protein structural classes from amino acid compositions. Eur. J. Biochem. 207, 429–433.
Chou, K.C., Zhang, C.T., 1994. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J. Biol. Chem. 269, 22014–22020.
Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.
Chou, K.C., Maggiora, G.M., 1998. Domain structural class prediction. Protein Eng. 11, 523–538.
Chou, K.C., Elrod, D.W., 1999. Protein subcellular location prediction. Protein Eng. 12, 107–118.
Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. 277, 45765–45769.
Chou, K.C., Elrod, D.W., 2002. Bioinformatical analysis of G-protein-coupled receptors. J. Proteome Res. 1, 429–433.
Chou, K.C., Cai, Y.D., 2003a. Predicting protein quaternary structure by pseudo amino acid composition. Proteins: Struct. Funct. Genet. 53, 282–289.
Chou, K.C., Cai, Y.D., 2003b. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem. Biophys. Res. Commun. 311, 743–747.
Chou, K.C., Cai, Y.D., 2003c. Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J. Cellul. Biochem. 90, 1250–1260 (Addendum, ibid. 2004, vol. 91, p. 1085).
Chou, K.C., Cai, Y.D., 2004a. Predicting enzyme family class in a hybridization space. Protein Sci. 13, 2857–2863.
Chou, K.C., Cai, Y.D., 2004b. Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. 321, 1007–1009 (Corrigendum: ibid., 2005, vol. 329, p. 1362).
Chou, K.C., Cai, Y.D., 2004c. Using GO-PseAA predictor to predict enzyme sub-class. Biochem. Biophys. Res. Commun. 325, 506–509.
Chou, K.C., Cai, Y.D., 2004d. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem. Biophys. Res. Commun. 320, 1236–1239.
Chou, K.C., Cai, Y.D., 2004e. Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J. Cell Biochem. 91, 1197–1203.
Chou, K.C., Cai, Y.D., 2005. Using GO-PseAA predictor to identify membrane proteins and their types. Biochem. Biophys. Res. Comm. 327, 845–847.
Chou, K.C., Shen, H.B., 2006a. Large-scale predictions of Gram-negative bacterial protein subcellular locations. J. Proteome Res. 5, 3420–3428.
Chou, K.C., Shen, H.B., 2006b. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *K*-nearest neighbor classifiers. J. Proteome Res. 5, 1888–1897.
Chou, K.C., Shen, H.B., 2006c. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.
Chou, K.C., Shen, H.B., 2006d. Predicting protein subcellular location by fusing multiple classifiers. J. Cell Biochem. 99, 517–527.
Chou, K.C., Cai, Y.D., 2006. Predicting protein–protein interactions from sequences in a hybridization space. J. Proteome Res. 5, 316–322.
Chou, K.C., Shen, H.B., 2007a. Large-scale plant protein subcellular location prediction. J. Cell Biochem. 100, 665–678.
Chou, K.C., Shen, H.B., 2007b. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. 6, 1728–1734.
Chou, K.C., Shen, H.B., 2007c. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Chou, K.C., Shen, H.B., 2007d. MemType-2L: A WEB server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Commun. 360, 339–345.
Chou, K.C., Shen, H.B., 2008a. ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem. Biophys. Res. Commun. 376, 321–325.
Chou, K.C., Shen, H.B., 2008b. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc. 3, 153–162.
Chou, K.C., Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Natur. Sci. 2, 63–92 (openly accessible at http://www.scirp.org/journal/NS/).
Chou, K.C., Shen, H.B., 2010a. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE 5, e9931.
Chou, K.C., Shen, H.B., 2010b. Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natur. Sci. 2, 1090–1103 (openly accessible at http://www.scirp.org/journal/NS/).
Chou, K.C., Shen, H.B., 2010c. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS ONE 5, e11335.
Chou, K.C., Liu, W., Maggiora, G.M., Zhang, C.T., 1998. Prediction and classification of domain structural classes. Proteins: Struct. Funct. Genet. 31, 97–103.
Chou, P.Y., 1989. Prediction of protein structural classes from amino acid composition. In: Fasman, G.D. (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation. Plenum Press, New York, pp. 549–586.
Cover, T.M., Hart, P.E., 1967. Nearest neighbour pattern classification. IEEE Trans. Inform. Theor. IT-13, 21–27.
Denoeux, T., 1995. A *K*-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. Syst. Man Cybernet. 25, 804–813.
Diao, Y., Ma, D., Wen, Z., Yin, J., Xiang, J., Li, M., 2008. Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. Amino Acids 34, 111–117.
Ding, H., Luo, L., Lin, H., 2009. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept. Lett. 16, 351–355.
Ding, Y.S., Zhang, T.L., Chou, K.C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept. Lett. 14, 811–815.
Du, P., Li, Y., 2006. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinform. 7, 518.
Du, P., Li, Y., 2008. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. J. Theor. Biol. 253, 579–589.
Du, P., Cao, S., Li, Y., 2009. SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic *K*-nearest neighbor (ET-KNN) algorithm. J. Theor. Biol. 261, 330–335.
Du, Q.S., Jiang, Z.Q., He, W.Z., Li, D.P., Chou, K.C., 2006. Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. J. Biomol. Struct. Dynam. 23, 635–640.
Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.
Fang, Y., Guo, Y., Feng, Y., Li, M., 2008. Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34, 103–109.
Feng, K.Y., Cai, Y.D., Chou, K.C., 2005. Boosting classifier for predicting protein domain structural class. Biochem. Biophys. Res. Commun. 334, 213–217.
Feng, Y.E., Luo, L.F., 2008. Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids 35, 607–614.
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., Bateman, A., 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247-51.
Gao, Q.B., Ye, X.F., Jin, Z.C., He, J., 2010. Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition. Anal. Biochem. 398, 52–59.
Gao, Q.B., Jin, Z.C., Ye, X.F., Wu, C., He, J., 2009. Prediction of nuclear receptors with optimal pseudo amino acid composition. Anal. Biochem. 387, 54–59.
Gao, Y., Shao, S.H., Xiao, X., Ding, Y.S., Huang, Y.S., Huang, Z.D., Chou, K.C., 2005. Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28, 373–376.
Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., Torres, A., 2009. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J. Theor. Biol. 257, 17–26.
Gu, F., Chen, H., 2009. Evaluating long-term relationship of protein sequence by use of d-Interval conditional probability and its impact on protein structural class prediction. Protein Pept. Lett. 16, 1267–1276.
Gu, Q., Ding, Y.S., Zhang, T.L., 2010a. Prediction of G-Protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein Pept. Lett. 17, 559–567.
Gu, Q., Ding, Y., Zhang, T., Shen, Y., 2010b. Prediction of G-protein-coupled receptor classes with pseudo amino acid composition. Shengwu Yixue Gongchengxue Zazhi 27, 500–504.
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., White, R., 2004. The gene ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258-61.
He, Z.S., Zhang, J., Shi, X.H., Hu, L.L., Kong, X.G., Cai, Y.D., Chou, K.C., 2010. Predicting drug–target interaction networks based on functional groups and biological features. PLoS ONE 5, e9603.
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., 2010. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE 5, e10972.
Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Asadabadi, E.B., 2007a. Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys. Chem. 128, 87–93.
Jahandideh, S., Sarvestani, A.S., Abdolmaleki, P., Jahandideh, M., Barfeie, M., 2007b. Gamma-turn types prediction in proteins using the support vector machines. J. Theor. Biol. 249, 785–790.
Jahandideh, S., Hoseini, S., Jahandideh, M., Hoseini, A., Disfani, F.M., 2009. Gamma-turn types prediction in proteins using the two-stage hybrid neural discriminant model. J. Theor. Biol. 259, 517–522.

Author's personal copy

Ji, G., Wu, X., Shen, Y., Huang, J., Quinn, Li, Q., 2010. A classification-based prediction model of messenger RNA polyadenylation sites. J. Theor. Biol. 265, 287–296.
Jiang, X., Wei, R., Zhao, Y., Zhang, T., 2008a. Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34, 669–675.
Jiang, X., Wei, R., Zhang, T.L., Gu, Q., 2008b. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept. Lett. 15, 392–396.
Kandaswamy, K.K., Pugalenthi, G., Moller, S., Hartmann, E., Kalies, K.U., Suganthan, P.N., Martinetz, T., 2010. Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein Pept. Lett. 17, 1473–1479.
Kannan, S., Hauth, A.M., Burger, G., 2008. Function prediction of hypothetical proteins without sequence similarity to proteins of known function. Protein Pept. Lett. 15, 1107–1116.
Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy *k*-nearest neighbours algorithm. IEEE Trans. Syst. Man Cybern. 15, 580–585.
Klein, P., 1986. Prediction of protein structural class by discriminant analysis. Biochim. Biophys. Acta 874, 205–215.
Klein, P., Delisi, C., 1986. Prediction of protein structural class from amino acid sequence. Biopolymers 25, 1659–1672.
Lee, V., Camon, E., Dimmer, E., Barrell, D., Apweiler, R., 2005. 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. 5, 5–8.
Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P., 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 34, D257-60.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–557.
Li, F.M., Li, Q.Z., 2008a. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34, 119–125.
Li, F.M., Li, Q.Z., 2008b. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept. Lett. 15, 612–616.
Li, Z.C., Zhou, X.B., Dai, Z., Zou, X.Y., 2009. Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37, 415–425.
Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.
Lin, H., Li, Q.Z., 2007a. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem. Biophys. Res. Commun. 354, 548–551.
Lin, H., Li, Q.Z., 2007b. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J. Comput. Chem. 28, 1463–1466.
Lin, H., Ding, H., Feng-Biao Guo, F.B., Zhang, A.Y., Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept. Lett. 15, 739–744.
Lin, H., Wang, H., Ding, H., Chen, Y.L., Li, Q.Z., 2009a. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor. 57, 321–330.
Lin, W.Z., Xiao, X., Chou, K.C., 2009b. GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. Protein Eng. Des. Sel 22, 699–705.
Liu, H., Yang, J., Wang, M., Xue, L., Chou, K.C., 2005. Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. The Protein J. 24, 385–389.
Liu, L., He, D., Yang, S., Xu, Y., 2010a. Applying chemometrics approaches to model and predict the binding affinities between the human amphiphysin SH3 domain and its peptide ligands. Protein Pept. Lett. 17, 246–253.
Liu, T., Zheng, X., Wang, C., Wang, J., 2010b. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept. Lett. 17, 1263–1269.
Liu, W., Chou, K.C., 1998. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J. Protein Chem. 17, 209–217.
Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2, 49–55.
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., Bryant, S.H., 2007. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35, D237-40.
Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis: Chapter 11 Discriminant Analysis; Chapter 12 Multivariate Analysis of Variance; Chapter 13 Cluster Analysis. Academic Press, London, pp. 322–381.
Metfessel, B.A., Saurugger, P.N., Connelly, D.P., Rich, S.T., 1993. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci. 2, 1171–1182.
Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept. Lett. 17, 1207–1214.
Mondal, S., Bhavna, R., Mohan Babu, R., Ramakumar, S., 2006. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J. Theor. Biol. 243, 252–260.
Munteanu, C.B., Gonzalez-Diaz, H., Magalhaes, A.L., 2008. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J. Theor. Biol. 254, 476–482.
Murvai, J., Vlahovicek, K., Barta, E., Pongor, S., 2001. The SBASE protein domain library, release 8.0: a collection of annotated protein sequence segments. Nucleic Acids Res. 29, 58–60.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of protein database for the investigation of sequence and structures. J. Mol. Biol. 247, 536–540.
Myers, D., Palmer, G., 1985. Microcomputer tools for steady-state enzyme kinetics. Bioinformatics (original: Computer Applied Bioscience) 1, 105–110.
Nakashima, H., Nishikawa, K., 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol 238, 54–61.
Nakashima, H., Nishikawa, K., Ooi, T., 1986. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99, 152–162.
Nanni, L., Lumini, A., 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34, 653–660.
Nanni, L., Lumini, A., 2009. A further step toward an optimal ensemble of classifiers for peptide classification, a case study: HIV protease. Protein Pept. Lett. 16, 163–167.
Niu, B., Cai, Y.D., Lu, W.C., Zheng, G.Y., Chou, K.C., 2006. Predicting protein structural class with AdaBoost learner. Protein Pept. Lett. 13, 489–492.
Pan, Y.X., Zhang, Z.Z., Guo, Z.M., Feng, G.Y., Huang, Z.D., He, L., 2003. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J. Protein Chem. 22, 395–402.
Pielak, R.M., Chou, J.J., 2010. Solution NMR structure of the V27A drug resistant mutant of influenza A M2 channel. Biochem. Biophys. Res. Commun. 401, 58–63.
Pillai, K.C.S., 1985. Mahalanobis D2. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences, vol. 5. John Wiley & Sons., New York, pp. 176–181 This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics.
Qi, J.P., Shao, S.H., Li, D.D., Zhou, G.P., 2007. A dynamic model for the p53 stress response networks under ion radiation. Amino Acids 33, 75–83.
Qiu, J.D., Huang, J.H., Liang, R.P., Lu, X.Q., 2009. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal Biochem. 390, 68–73.
Qiu, J.D., Huang, J.H., Shi, S.P., Liang, R.P., 2010. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept. Lett. 17, 715–722.
Reinhardt, A., Hubbard, T., 1998. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. 26, 2230–2236.
Rezaei, M.A., Abdolmaleki, P., Karami, Z., Asadabadi, E.B., Sherafat, M.A., Abrishami-Moghaddam, H., Fadaie, M., Forouzanfar, M., 2008. Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks. J. Theor. Biol. 254, 817–820.
Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F., 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29, 2994–3005.
Schnell, J.R., Chou, J.J., 2008. Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451, 591–595.
Shao, X., Tian, Y., Wu, L., Wang, Y., Jing, L., Deng, N., 2009. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J. Theor. Biol. 258, 289–293.
Sharma, A.K., Zhou, G.P., Kupferman, J., Surks, H.K., Christensen, E.N., Chou, J.J., Mendelsohn, M.E., Rigby, A.C., 2008. Probing the interaction between the coiled coil leucine zipper of cGMP-dependent protein kinase Ialpha and the C terminus of the myosin binding subunit of the myosin light chain phosphatase. J. Biol. Chem. 283, 32860–32869.
Shen, H.B., Chou, K.C., 2005a. Using optimized evidence-theoretic *K*-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem. Biophys. Res. Commun. 334, 288–292.
Shen, H.B., Chou, K.C., 2005b. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem. Biophys. Res. Commun. 337, 752–756.
Shen, H.B., Chou, K.C., 2006. Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.
Shen, H.B., Chou, K.C., 2007a. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem. Biophys. Res. Commun. 364, 53–59.
Shen, H.B., Chou, K.C., 2007b. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng. Des. Sel. 20, 39–46.
Shen, H.B., Chou, K.C., 2007c. Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85, 233–240.
Shen, H.B., Chou, K.C., 2007d. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem. Biophys. Res. Commun. 355, 1006–1011.
Shen, H.B., Chou, K.C., 2008. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373, 386–388.
Shen, H.B., Chou, K.C., 2009a. Predicting protein fold pattern with functional domain and sequential evolution information. J. Theor. Biol. 256, 441–446.

Author's personal copy

Shen, H.B., Chou, K.C., 2009b. QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J. Proteome Res. 8, 1577–1584.
Shen, H.B., Chou, K.C., 2009c. Identification of proteases and their types. Anal. Biochem. 385, 153–160.
Shen, H.B., Chou, K.C., 2009d. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Anal Biochem. 394, 269–274.
Shen, H.B., Chou, K.C., 2010. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. J. Theor. Biol. 264, 326–333.
Shen, H.B., Yang, J., Chou, K.C., 2006. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J. Theor. Biol. 240, 9–13.
Shen, H.B., Yang, J., Chou, K.C., 2007. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33, 57–67.
Shi, J.Y., Zhang, S.W., Pan, Q., Zhou, G.P., 2008. Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. Amino Acids 35, 321–327.
Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.-M., Xie, J., 2007. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33, 69–74.
Shi, R., Hu, X., 2010. Predicting enzyme subclasses by using support vector machine with composite vectors. Protein Pept. Lett. 17, 599–604.
Small, I., Peeters, N., Legeai, F., Lurin, C., 2004. Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4, 1581–1590.
Spiess, M., 1995. Heads or tails - what determines the orientation of proteins in the membrane. FEBS Lett 369, 76–79.
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., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinform. 4, 41.
Vilar, S., Gonzalez-Diaz, H., Santana, L., Uriarte, E., 2009. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. J. Theor. Biol. 261, 449–458.
Wang, J., Pielak, R.M., McClintock, M.A., Chou, J.J., 2009. Solution structure and functional analysis of the influenza B proton channel. Nat. Struct. Mol. Biol. 16, 1267–1271.
Wang, M., Yang, J., Liu, G.P., Xu, Z.J., Chou, K.C., 2004. Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng. Des. Sel. 17, 509–516.
Wang, S.Q., Yang, J., Chou, K.C., 2006. Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J. Theor. Biol. 242, 941–946.
Wang, T., Yang, J., 2010. Predicting subcellular localization of Gram-negative bacterial proteins by linear dimensionality reduction method. Protein Pept. Lett. 17, 32–37.
Wang, T., Xia, T., Hu, X.M., 2010a. Geometry preserving projections algorithm for predicting membrane protein types. J. Theor. Biol. 262, 208–213.
Wang, T., Yang, J., Shen, H.B., Chou, K.C., 2008a. Predicting membrane protein types by the LLDA algorithm. Protein Pept. Lett. 15, 915–921.
Wang, Y., Xue, Z., Shen, G., Xu., J., 2008b. PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. Amino Acids 35, 295–302.
Wang, Y.C., Wang, X.B., Yang, Z.X., Deng, N.Y., 2010b. Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. Protein Pept. Lett. 17, 1441–1449.
Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17, 149–163.
Wu, J., Li, M.L., Yu, L.Z., Wang, C., 2010. An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition. Protein J. 29, 62–67.
Xiao, X., Lin, W.Z., Chou, K.C., 2008a. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J. Comput. Chem. 29, 2018–2024.
Xiao, X., Wang, P., Chou, K.C., 2008b. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J. Theor. Biol. 254, 691–696.
Xiao, X., Wang, P., Chou, K.C., 2009a. Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. J. Appl. Crystallogr. 42, 169–173.
Xiao, X., Wang, P., Chou, K.C., 2009b. GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J. Comput. Chem. 30, 1414–1423.
Xiao, X., Wang, P., Chou, K.C., 2010a. Quat-2L: a web-server for predicting protein quaternary structural attributes. Mol. Diversity. doi:10.1007/s11030-010-9227-8.
Xiao, X., Wang, P., Chou, K.C., 2010b. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol. Biosyst.. doi:10.1039/C0MB00170H.
Xiao, X., Shao, S.H., Huang, Z.D., Chou, K.C., 2006a. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J. Comput. Chem. 27, 478–482.
Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C., 2006b. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30, 49–54.
Xiao, X., Shao, S., Ding, Y., Huang, Z., Huang, Y., Chou, K.C., 2005. Using complexity measure factor to predict protein subcellular location. Amino Acids 28, 57–61.
Xiaohui, N., Nana, L., Feng, S., Xuehai, H., Jingbo, X., Huijuan, X., 2010. Predicting protein solubility with a hybrid approach by pseudo amino acid composition. Protein Pept. Lett. 17, 1466–1472.
Yang, J., Jiang, X.F., 2010. A novel approach to predict protein–protein interactions related to Alzheimer's disease based on complex network. Protein Pept. Lett. 17, 356–366.
Yang, J.Y., Peng, Z.L., Yu, Z.G., Zhang, R.J., Anh, V., Wang, D., 2009. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J. Theor. Biol. 257, 618–626.
Yang, X.Y., Shi, X.H., Meng, X., Li, X.L., Lin, K., Qian, Z.L., Feng, K.Y., Kong, X.Y., Cai, Y.D., 2010. Classification of transcription factors using protein primary structure. Protein Pept. Lett. 17, 899–908.
Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., Xiong, W., Qin, W., 2010. SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. J. Theor. Biol. 267, 1–6.
Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., Li, M.L., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259, 366–372.
Zhang, G.Y., Fang, B.S., 2008. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J. Theor. Biol. 253, 310–315.
Zhang, G.Y., Li, H.C., Gao, J.Q., Fang, B.S., 2008a. Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Pept. Lett. 15, 1132–1137.
Zhang, S.W., Chen, W., Yang, F., Pan, Q., 2008b. Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. Amino Acids 35, 591–598.
Zhang, S.W., Pan, Q., Zhang, H.C., Shao, Z.C., Shi, J.Y., 2006. Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30, 461–468.
Zhang, S.W., Zhang, Y.L., Yang, H.F., Zhao, C.H., Pan, Q., 2008c. 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 34, 565–572.
Zhang, T.L., Ding, Y.S., 2007. Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33, 623–629.
Zhang, T.L., Ding, Y.S., Chou, K.C., 2008d. Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. J. Theor. Biol. 250, 186–193.
Zhao, X.M., Chen, L., Aihara, K., 2008. Protein function prediction with high-throughput data. Amino Acids 35, 517–530.
Zhou, G.P., 1989. Biological functions of soliton and extra electron motion in DNA structure. Phys. Scr. 40, 698–701.
Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.
Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochem. J. 222, 169–176.
Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins: Struct. Funct. Genet. 44, 57–59.
Zhou, G.P., Troy 2nd, F.A., 2003. Characterization by NMR and molecular modeling of the binding of polyisoprenols and polyisoprenyl recognition sequence peptides: 3D structure of the complexes reveals sites of specific interactions. Glycobiology 13, 51–71.
Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct. Funct. Genet. 50, 44–48.
Zhou, G.P., Troy, F.A., 2005a. NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. Curr. Protein Pept. Sci. 6, 399–411.
Zhou, G.P., Troy 2nd, F.A., 2005b. NMR study of the preferred membrane orientation of polyisoprenols (dolichol) and the impact of their complex with polyisoprenyl recognition sequence peptides on membrane structure. Glycobiology 15, 347–359.
Zhou, G.P., Cai, Y.D., 2006. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. PROTEINS: Struct. Funct. Bioinform. 63, 681–684.
Zhou, G.P., Li, T.T., Chou, K.C., 1981. The flexibility during the juxtaposition of reacting groups and the upper limits of enzyme reactions. Biophys. Chem. 14, 277–281.
Zhou, G.P., Surks, H.K., Schnell, J.R., Chou, J.J., Mendelsohn, M.E., Rigby, A.C., 2004. The three-dimensional structure of the cGMP-dependent protein kinase I-α leucine zipper domain and its interaction with the myosin binding subunit. Blood 104, 963a.
Zhou, G.Q., Zhong, W.Z., 1982. Diffusion-controlled reactions of enzymes. A comparison between Chou's model and Alberty–Hammes–Eigen's model. Eur. J. Biochem. 128, 383–387.
Zhou, G.Z., Wong, M.T., Zhou, G.Q., 1983. Diffusion-controlled reactions of enzymes. An approximate analytic solution of Chou's model. Biophys. Chem. 18, 125–132.
Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2007. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J. Theor. Biol. 248, 546–551.
Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2008. Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids 35, 383–388.
Zou, D., He, Z., He, J., Xia, Y., 2011. Supersecondary structure prediction using Chou's pseudo amino acid composition. J. Comput. Chem. 32, 271–278.