Analytical Biochemistry 442 (2013) 118–125

Contents lists available at SciVerse ScienceDirect

Analytical Biochemistry

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

iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition

Peng-Mian Feng [a], Wei Chen [b,c,*], Hao Lin [d,*], Kuo-Chen Chou [c,e]

[a] School of Public Health, Hebei United University, Tangshan 063000, China
[b] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China
[c] Gordon Life Science Institute, Belmont, MA 02478, USA
[d] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 12 April 2013
Received in revised form 21 May 2013
Accepted 22 May 2013
Available online 10 June 2013

Keywords:
Heat shock protein
Reduced amino acid alphabet
n-Peptide composition
PseAAC
SVM
Web server

ABSTRACT

Heat shock proteins (HSPs) are a type of functionally related proteins present in all living organisms, both prokaryotes and eukaryotes. They play essential roles in protein–protein interactions such as folding and assisting in the establishment of proper protein conformation and prevention of unwanted protein aggregation. Their dysfunction may cause various life-threatening disorders, such as Parkinson's, Alzheimer's, and cardiovascular diseases. Based on their functions, HSPs are usually classified into six families: (i) HSP20 or sHSP, (ii) HSP40 or J-class proteins, (iii) HSP60 or GroEL/ES, (iv) HSP70, (v) HSP90, and (vi) HSP100. Although considerable progress has been achieved in discriminating HSPs from other proteins, it is still a big challenge to identify HSPs among their six different functional types according to their sequence information alone. With the avalanche of protein sequences generated in the post-genomic age, it is highly desirable to develop a high-throughput computational tool in this regard. To take up such a challenge, a predictor called *iHSP-PseRAAAC* has been developed by incorporating the reduced amino acid alphabet information into the general form of pseudo amino acid composition. One of the remarkable advantages of introducing the reduced amino acid alphabet is being able to avoid the notorious dimension disaster or overfitting problem in statistical prediction. It was observed that the overall success rate achieved by iHSP-PseRAAAC in identifying the functional types of HSPs among the aforementioned six types was more than 87%, which was derived by the jackknife test on a stringent benchmark dataset in which none of HSPs included has ⩾40% pairwise sequence identity to any other in the same subset. It has not escaped our notice that the reduced amino acid alphabet approach can also be used to investigate other protein classification problems. As a user-friendly web server, iHSP-PseRAAAC is accessible to the public at http://lin.uestc.edu.cn/server/iHSP-PseRAAAC.

Heat shock proteins (HSPs),[1] first discovered in 1962 [1], are a set of functionally related proteins involved in the folding and unfolding of other proteins. HSPs are ubiquitously expressed in virtually all living organisms from bacteria to humans and function as intracellular chaperones for other proteins. Their expression is increased when cells are exposed to a wide variety of physiological and environmental stress conditions such as elevated temperature, infection, and inflammation [2–4]. They play essential roles in protein–protein interactions such as folding and assisting for establishing proper protein conformation and in prevention of unwanted protein aggregation [5,6]. In addition, HSPs are key determinants of quality control and play a critical role in maintaining the overall cellular protein homeostasis [7]. Their dysfunction is implicated in life-threatening disorders, including Parkinson's, Alzheimer's, and cardiovascular diseases [8–10]. The diversified nature of HSPs and their vast repertoire of functions have drawn considerable attention among researchers, and extensive studies are in progress to deduce the intricate cellular functional networks among these proteins [11].

Based on their different functions (Fig. 1), HSPs are generally classified into the following six families: (i) HSP20 or sHSP, (ii) HSP40 or J-class proteins, (iii) HSP60 or GroEL/ES, (iv) HSP70, (v) HSP90, and (vi) HSP100 [11]. Although considerable progress has

[*] Corresponding authors. Address: Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China. Fax: +86 315 372 5715 (W. Chen).

E-mail addresses: wchen@gordonlifescience.org (W. Chen), hlin@uestc.edu.cn (H. Lin).

[1] *Abbreviations used:* HSP, heat shock protein; AAC, amino acid composition; PseAAC, pseudo amino acid composition; RAAA, reduced amino acid alphabet; PB, Protein Block; SVM, support vector machine; OVO, one versus one; OVR, one versus rest; Sn, sensitivity; Sp, specificity; MCC, Mathew's correlation coefficient; OA, overall accuracy; CRG, completely random guess; WRG, weighted random guess.

0003-2697/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ab.2013.05.024

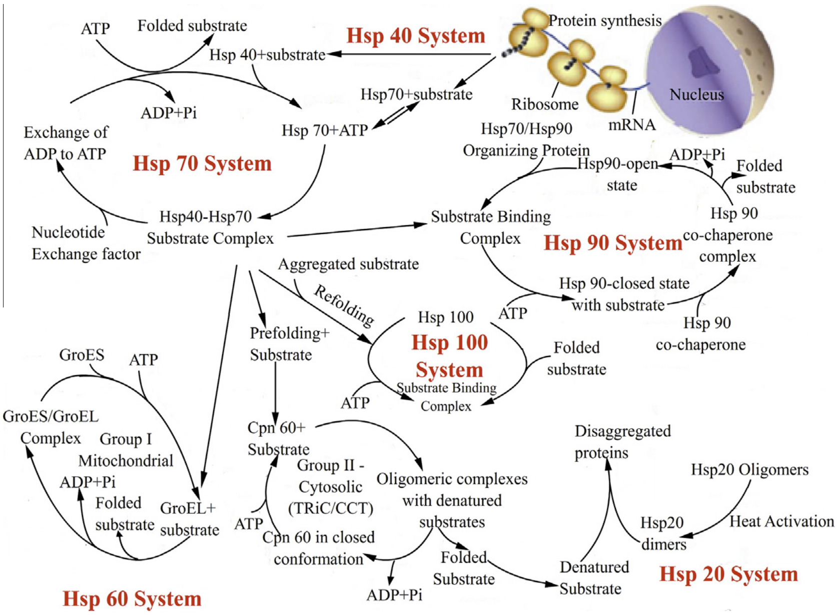


Fig.1. Schematic illustration to show the different biological processes and functions of HSPs from six different families: HSP20, HSP40, HSP60, HSP70, HSP90, and HSP100. The HSP family members and their associated cofactors function together in complexes, acting in concert as molecular chaperones to facilitate the proper folding and activation of many cellular proteins. See text for further explanation.

been achieved in identifying HSPs from other proteins, it is still a big challenge to identify the families of HSPs according to their sequence information alone. With the explosive growth of protein sequences generated in the post-genomic age, it is highly desirable to develop automated methods for timely and reliably annotating their functional types. In view of this, the current study was initiated in an attempt to develop a new predictor by which one can easily identify the functional types or families of HSPs based on their sequence information alone.

According to a recent review [12], to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (v) establish a user-friendly web server for the predictor that is freely accessible to the public. We elaborate how to deal with these procedures one by one below.

Materials and methods

Benchmark dataset

The sequences of HSPs were taken from the HSPIR database (http://pdslab.biochem.iisc.ernet.in/hspir), which currently contains 9902 protein sequences encompassing 277 genomes ranging from prokaryotes to eukaryotes [11]. To reduce homologous bias and redundancy, the program CD–HIT [13] was used to remove those HSPs that have ⩾40% pairwise sequence identity to any other in the same subset. Finally, we obtained a dataset $\mathbb{S}$ of 2225 HSPs classified into six families, as can be formulated by

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \cup \mathbb{S}_6, \qquad (1)$$

where the subset $\mathbb{S}_1$ contains 357 HSP20 sequences, $\mathbb{S}_2$ contains 1279 HSP40 sequences, $\mathbb{S}_3$ contains 163 HSP60 sequences, $\mathbb{S}_4$ contains 283 HSP70 sequences, $\mathbb{S}_5$ contains 58 HSP90 sequences, and $\mathbb{S}_6$ contains 85 HSP100 sequences (Table 1) and where ∪ represents the symbol for union in the set theory. For readers' convenience, the sequences of the 2225 HSPs and their codes are given in the Supplementary material.

Pseudo amino acid composition and reduced amino acid alphabet

To develop a sequence-based predictor for identifying the attribute of a protein, one of the keys is to formulate its sequence with an effective mathematical expression that can truly reflect the intrinsic correlation with the attribute to be predicted [14]. The most straightforward method to formulate the sample of a protein with L residues is to use its entire amino acid sequence, as can be formulated by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\ldots\mathrm{R}_L, \qquad (2)$$

where R_1 represents the 1st residue of the protein $\mathbf{P}$, R_2 represents the 2nd residue of the protein $\mathbf{P}$, and so forth. To identify its attribute, the tools for computing amino acid sequence similarity, such as BLAST [15,16], were used to search the database for those targets that have high sequence similarity to the query protein. Subsequently, the attribute annotations of the target proteins found in this way were used to infer the attribute of the query protein. Although this kind of straightforward sequential model contains the entire sequence information, unfortunately it failed to work when the query protein did not have any significant sequence similarity to the attribute known proteins [12,17]. To overcome the

Table 1
Breakdown of the 2225 HSPs in the benchmark dataset $\mathbb{S}$ according to their six subfamilies.

Dataset	Family	Number of HSP samples
$\mathbb{S}_1$	HSP20	357
$\mathbb{S}_2$	HSP40	1279
$\mathbb{S}_3$	HSP60	163
$\mathbb{S}_4$	HSP70	283
$\mathbb{S}_5$	HSP90	58
$\mathbb{S}_6$	HSP100	85
$\mathbb{S}$	Overall	2225

above difficulty, which is inherent to the sequential model, various nonsequential or discrete models to formulate protein samples have been proposed.

Among the discrete models, the simplest one for a protein was based on its amino acid composition (AAC) as defined by

$$\mathbf{P} = [f_1^{(1)} \quad f_2^{(1)} \quad \ldots \quad f_{20}^{(1)}]^{\mathbf{T}}, \tag{3}$$

where $f_u^{(1)}(u = 1, 2, \ldots, 20)$ are the normalized occurrence frequencies of the 20 native amino acids [18–20] in the protein **P** and **T** is the transposing operator. The AAC discrete model was widely used for identifying various attributes of proteins. However, as can be seen from Eq. (3), all of the sequence order effects were lost by using the AAC discrete model. To completely avoid losing the sequence order information, the pseudo amino acid composition (PseAAC) was proposed [21,22] to replace the simple AAC model for representing the sample of a protein. Ever since the concept of PseAAC was proposed in 2001 [21], it has penetrated into nearly all of the areas of protein attribute prediction such as identifying bacterial virulent proteins [23], predicting supersecondary structure [24], predicting protein subcellular location [25–27], predicting membrane protein types [28], discriminating outer membrane proteins [29], identifying antibacterial peptides [30], identifying allergenic proteins [31], predicting metalloproteinase family [32], predicting protein structural class [33], identifying G-protein-coupled receptors (GPCRs) and their types [34], identifying protein quaternary structural attributes [35], predicting protein submitochondria locations [36], identifying risk type of human papillomaviruses [37], identifying cyclin proteins [38], predicting GABA_A receptor proteins [39], predicting subchloroplast locations [40], and classifying amino acids [41], among many others (see a long list of articles cited in the References section of Ref. [12]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [42,43] as well as other biological samples (see, e.g., Refs. [44–46]). Because it has been widely and increasingly used, recently two powerful software programs, called *PseAAC-Builder* [47] and *propy* [48], were established for generating various special Chou's pseudo amino acid compositions in addition to the web server *PseAAC* [49] built in 2008.

According to a recent review [12], the general form of PseAAC for a protein **P** is formulated by

$$\mathbf{P} = [\Psi_1 \quad \Psi_2 \quad \ldots \quad \Psi_u \quad \ldots \quad \Psi_\Omega]^{\mathbf{T}}, \tag{4}$$

where the subscript Ω is an integer and its value, as well as the components Ψ_u $(u = 1, 2, \ldots, \Omega)$, will depend on how to extract the desired information from the amino acid sequence of **P** (cf. Eq. (2)). Here we describe how to extract useful information from the benchmark dataset $\mathbb{S}$ to define the components in Eq. (4) for the protein samples concerned in this study.

One of the most simple PseAAC modes is the so-called *n*-peptide composition: when $n = 1$, it is reduced to AAC; when $n = 2$, it is reduced to dipeptide composition [28,30,50–52]; when $n = 3$, it is reduced to tripeptide composition [53]; and so forth. Although the *n*-peptide composition can incorporate some sort of sequence order information when $n \geqslant 2$, the dimension of PseAAC formed in this way will increase rapidly. For instance, the PseAAC formed by the dipeptide composition would be a $20^2 = 400$-D vector [30,52,54], that formed by the tripeptide composition would be a $20^3 = 8000$-D vector [53], and that formed by the *n*-peptide composition would be a 20^n-D vector [54,55]. Accordingly, we might face the high-dimension disaster problem [56] or machine breakdown problem [57]. To alleviate the problem of geometric increase in dimension, we consider the following approach.

Based on the physiochemical properties, the 20 native amino acids can be clustered into a smaller number of representative residues called reduced amino acid alphabet (RAAA) [58–60]. Compared with the traditional amino acid composition, RAAA not only simplifies the complexity of protein system but also improves the ability to find structurally conserved regions and structural similarity of entire proteins.

One common way to design RAAA is by clustering amino acids into groups according to sequence or structure information. Recently, a structural alphabet called Protein Blocks (PBs) was proposed by de Brevern and coworkers [61,62] and has been widely used in computational proteomics, as indicated in a review [63]. PBs contain a set of 16 local structures or prototypes, labeled from *a* to *p*, of five residues length described based on the (Φ, Ψ) dihedral angles [63]. The labels *m* and *d* of PBs are prototypes for the central region of α-helix and β-strand, respectively; labels *a* through *c* primarily represent the *N*-cap of β-strand; labels *g* through *j* are specific to coils; labels *k* and *l* correspond to *N*-cap of α-helix; labels *e* and *f*, as well as labels *n* through *p*, correspond to the C-caps [64].

To aid the design of mutations, PBs have been used to define RAAA by Etchebest and coworkers [65]. Recently, it was demonstrated that the RAAA defined by these authors is quite useful for protein family classification [66–68]. According to different optimization procedures as elaborated by Etchebest and coworkers [65], the 20 native amino acids may have five different cluster profiles—$\mathbb{CP}(13)$, $\mathbb{CP}(11)$, $\mathbb{CP}(9)$, $\mathbb{CP}(8)$, and $\mathbb{CP}(5)$—as formulated below:

$$\begin{cases} \mathbb{CP}(13) = \{\text{G; IV; FYW; A; L; M; E; QRK; P; ND; HS; T; C}\} \\ \mathbb{CP}(11) = \{\text{G; IV; FYW; A; LM; EQRK; P; ND; HS; T; C}\} \\ \mathbb{CP}(9) = \{\text{G; IV; FYW; ALM; EQRK; P; ND; HS; TC}\} \\ \mathbb{CP}(8) = \{\text{G; IV; FYW; ALM; EQRK; P; ND; HSTC}\} \\ \mathbb{CP}(5) = \{\text{G; IV FYW; ALMEQRK; P; NDHSTC}\} \end{cases}, \tag{5}$$

where the single letters without a semicolon (;) to separate them mean belonging to a same cluster.

Thus, for the *n*-peptide composition with various cluster profiles, the corresponding components and dimensions will be different. For example, for the single amino acid composition, the *u*th component in Eq. (4) will be formulated as

$$\Psi_u = f_u^{(1)}, \ u = 1, 2, \ldots, \Omega, \tag{6}$$

where $f_u^{(1)}$ is the occurrence frequency of the *u*th amino acid in protein **P** (cf. Eq. (2)) and the corresponding dimension of PseAAC is given by

$$\Omega = \begin{cases} 13^1 = 13 & \text{for } \mathbb{CP}(13) \text{ cluster} \\ 11^1 = 11 & \text{for } \mathbb{CP}(11) \text{ cluster} \\ 9^1 = 9 & \text{for } \mathbb{CP}(9) \text{ cluster} \\ 8^1 = 8 & \text{for } \mathbb{CP}(8) \text{ cluster} \\ 5^1 = 5 & \text{for } \mathbb{CP}(5) \text{ cluster} \end{cases}. \tag{7}$$

For the dipeptide composition, the *u*th component in Eq. (4) will be formulated as

$$\Psi_u = f_u^{(2)}, u = 1, 2, \ldots, \Omega, \tag{8}$$

where $f_u^{(2)}$ is the occurrence frequency of the *u*th dipeptide in protein **P** (cf. Eq. (2)) and the corresponding dimension of PseAAC is given by

$$\Omega = \begin{cases} 13^2 = 169 & \text{for } \mathbb{CP}(13) \text{ cluster} \\ 11^2 = 121 & \text{for } \mathbb{CP}(11) \text{ cluster} \\ 9^2 = 81 & \text{for } \mathbb{CP}(9) \text{ cluster} \\ 8^2 = 64 & \text{for } \mathbb{CP}(8) \text{ cluster} \\ 5^2 = 25 & \text{for } \mathbb{CP}(5) \text{ cluster} \end{cases}. \tag{9}$$

For the tripeptide composition, the *u*th component in Eq. (4) will be formulated as

$$\Psi_u = f_u^{(3)},\ u = 1,\ 2, \ldots, \Omega, \tag{10}$$

where $f_u^{(3)}$ is the occurrence frequency of the *u*th tripeptide in protein **P** (cf. Eq. (2)) and the corresponding dimension of PseAAC is given by

$$\Omega = \begin{cases} 13^3 = 2197 & \text{for } \mathbb{CP}(13) \text{ cluster} \\ 11^3 = 1331 & \text{for } \mathbb{CP}(11) \text{ cluster} \\ 9^3 = 729 & \text{for } \mathbb{CP}(9) \text{ cluster} \\ 8^3 = 512 & \text{for } \mathbb{CP}(8) \text{ cluster} \\ 5^3 = 125 & \text{for } \mathbb{CP}(5) \text{ cluster} \end{cases}, \tag{11}$$

and so forth.

Once the feature vectors for protein samples are defined via PseAAC of Eq. (4), the next thing we need to consider is an effective algorithm or engine to operate the classification.

Support vector machine

Support vector machine (SVM) is a powerful and popular method for pattern recognition that has been widely used in the realm of bioinformatics (see, e.g., Refs. [42,50,69–73]). The basic idea of SVM is to transform the data into a high-dimensional feature space and then determine the optimal separating hyperplane using a kernel function. To handle a multi-class problem, one versus one (OVO) and one versus rest (OVR) are generally applied to extend the traditional SVM. For a brief formulation of SVM and how it works, see Refs. [69,70]. For more details about SVM, see Ref. [74].

In the current study, the LIBSVM 2.84 package [75] was used as an implementation of SVM, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. The OVO strategy was employed for making predictions using the popular radial basis function (RBF). The regularization parameter *C* and the kernel width parameter γ were determined via an optimization procedure using a grid search approach, and their actual values obtained in this way for the current study were *C* = 2.0 and γ = 0.125.

The predictor obtained via the aforementioned procedure is called *iHSP-PseRAAAC*, where "i" stands for "identify," "HSP" stands for "heat shock protein," "Pse" stands for "pseudo," "R" stands for "reduced," "AAA" stands for "amino acid alphabet," and "C" stands for "composition."

Results and discussion

Criteria for performance evaluation

One of the important procedures in developing a useful statistical predictor [12] is to objectively evaluate its performance or anticipated success rate. Now we address this problem.

To provide a more intuitive and easier to understand method to measure the prediction quality, here the criterion proposed in Ref. [76] was adopted. According to that criterion, the rates of correct predictions for the HSP samples in the subset $\mathbb{S}_i$ $(i = 1,\ 2, \ldots, 6)$ and those not belonging to the subset $\mathbb{S}_i$ are respectively defined by

$$\begin{cases} \Lambda^+(i) = \dfrac{N^+(i) - N_-^+(i)}{N^+(i)} \\ \Lambda^-(i) = \dfrac{N^-(i) - N_+^-(i)}{N^-(i)} \end{cases}, \tag{12}$$

where $N^+(i)$ is the total number of the investigated HSP samples in the subset $\mathbb{S}_i$, whereas $N_-^+(i)$ is the number of HSP samples in $\mathbb{S}_i$ that were incorrectly predicted belonging to the other subsets, and $N^-(i)$ is the total number of the HSP samples in all of the other subsets, whereas $N_+^-(i)$ is the number of the HSP samples that were incorrectly predicted belonging to $\mathbb{S}_i$. The overall sub-success prediction rate for each of the subsets is given by [77]

$$\Lambda(i) = \frac{\Lambda^+(i)N^+(i) + \Lambda^-(i)N^-(i)}{N^+(i) + N^-(i)} = 1 - \frac{N_-^+(i) + N_+^-(i)}{N^+(i) + N^-(i)},\ (i = 1,\ 2, \ldots, 6). \tag{13}$$

It is obvious from Eqs. (12) and (13) that, if and only if all of the samples in the subset $\mathbb{S}_i$ (cf. Eq. (1)) are perfectly correctly predicted without any underprediction or overprediction (i.e., $N_-^+(i) = N_+^-(i) = 0$ and $\Lambda^+(i) = \Lambda^-(i) = 1$), we have $\Lambda(i) = 1$; otherwise, $\Lambda(i)$ would be smaller than 1.

On the other hand, it is instructive to point out that the following metrics are often used in the literature for examining the performance quality of a predictor:

$$\begin{cases} \mathrm{Sn}(i) = \dfrac{\mathrm{TP}(i)}{\mathrm{TP}(i) + FN(i)} \\ \mathrm{Sp}(i) = \dfrac{\mathrm{TN}(i)}{\mathrm{TN}(i) + FP(i)} \\ \mathrm{MCC}(i) = \dfrac{\mathrm{TP}(i) \times \mathrm{TN}(i) - \mathrm{FP}(i) \times \mathrm{FN}(i)}{\sqrt{[\mathrm{TP}(i) + FP(i)][\mathrm{TP}(i) + FN(i)][\mathrm{TN}(i) + FP(i)][\mathrm{TN}(i) + FN(i)]}} \\ \mathrm{OA} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{M} \mathrm{TP}(i) \end{cases}, \tag{14}$$

where TP represents the true positive, TN represents the true negative, FP represents the false positive, FN represents the false negative, Sn represents the sensitivity, Sp represents the specificity, MCC represents the Mathew's correlation coefficient, OA represents the overall accuracy, *M* = 6 is the number of subsets (cf. Eq. (1)), and *N* is the number of the total samples in $\mathbb{S}$.

The relations between the symbols in Eqs. (13) and (14) are given by

$$\begin{cases} \mathrm{TP}(i) = N^+(i) - N_-^+(i) \\ \mathrm{TN}(i) = N^-(i) - N_+^-(i) \\ \mathrm{FP}(i) = N_+^-(i) \\ \mathrm{FN}(i) = N_-^+(i) \end{cases}. \tag{15}$$

Substituting Eq. (15) into Eq. (14) and also noting Eq. (13), we obtain

$$\begin{cases} \mathrm{Sn}(i) = 1 - \dfrac{N_-^+(i)}{N^+(i)} \\ \mathrm{Sp}(i) = 1 - \dfrac{N_+^-(i)}{N^-(i)} \\ \mathrm{MCC}(i) = \dfrac{1 - (\frac{N_-^+(i)}{N^+(i)} + \frac{N_+^-(i)}{N^-(i)})}{\sqrt{(1 + \frac{N_+^-(i) - N_-^+(i)}{N^+(i)})(1 + \frac{N_-^+(i) - N_+^-(i)}{N^-(i)})}} \\ \mathrm{OA} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{M}[N^+(i) - N_-^+(i)] \end{cases}. \tag{16}$$

Obviously, when $N_-^+(i) = 0$, meaning that none of the HSP samples in subset $\mathbb{S}_i$ was mispredicted belonging to other subsets, we have the sensitivity $\mathrm{Sn}(i) = 1$, whereas when $N_-^+(i) = N^+(i)$, meaning that all of the HSP samples in subset $\mathbb{S}_i$ were mispredicted belonging to the other subsets, we have the sensitivity $\mathrm{Sn}(i) = 0$. Likewise, when $N_+^-(i) = 0$, meaning that none of the HSP samples in the other subsets was incorrectly predicted belonging to the subset $\mathbb{S}_i$, we have the specificity $\mathrm{Sp}(i) = 1$, whereas when $N_+^-(i) = N^-(i)$, meaning that all of the HSP samples in the other subsets were incorrectly predicted belonging to the subset $\mathbb{S}_i$, we have the specificity $\mathrm{Sp}(i) = 0$. When $N_-^+(i) = N_+^-(i) = 0$ $(i = 1, 2, \ldots, 6)$, meaning that none of the HSP samples in all of the subsets of $\mathbb{S}$ (cf. Eq. (1)) was incorrectly predicted, we have the overall accuracy $\mathrm{OA} = 1$, whereas when $N_-^+(i) = N^+(i)$ and $N_+^-(i) = N^-(i)$ $(i = 1, 2, \ldots, 6)$, meaning that the HSP samples in all of the subsets of $\mathbb{S}$ were mispredicted, we have the overall accuracy $\mathrm{OA} = 0$. When $N_-^+(i) = N_+^-(i) = 0$, meaning that none of the HSP samples in the subset $\mathbb{S}_i$ was mispredicted, we have $\mathrm{MCC}(i) = 1$; when $N_-^+(i) = N^+(i)/2$ and $N_+^-(i) = N^-(i)/2$, we have $\mathrm{MCC}(i) = 0$, meaning no better than random prediction for the HSP samples in the subset $\mathbb{S}_i$; when $N_-^+(i) = N^+(i)$ and $N_+^-(i) = N^-(i)$, we have $\mathrm{MCC}(i) = -1$, meaning total disagreement between prediction and observation for the HSP samples in the subset $\mathbb{S}_i$. As we can see from the above discussion, it is much more intuitive and easier to understand when using Eq. (16) to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient.

Cross-validation

Three cross-validation methods, namely sub-sampling (or K-fold cross-validation) test, independent dataset test, and jackknife test, are often used to evaluate the quality of a predictor [78]. Among the three methods, however, the jackknife test is deemed the least arbitrary and most objective, as elucidated in Ref. [79] and demonstrated by Eqs. (28) to (32) of Ref. [12], and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (see, e.g., Refs. [25,27–29,34,37–39,42,80,81]). Accordingly, the jackknife test was used to examine the performance of the model proposed in the current study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample, and all of the rule parameters are calculated without including the one being identified.

Listed in Table 2 are the jackknifing results obtained by iHSP-PseRAAAC on the benchmark dataset $\mathbb{S}$ (cf. Supporting Information S1 in supplementary material) based on the five different cluster profiles (Eq. 5) for the dipeptide case (i.e., $n = 2$ with Eqs. (8) and (9)). For facilitating comparison, the results calculated for the single amino acid case (i.e., $n = 1$ with Eqs. (6) and (7)) and the tripeptide case (i.e., $n = 3$ with Eqs. (10) and (11)) are given in Supporting Information S2 of the supplementary material, from which we can see that the corresponding success rates are obviously lower than those for the case of $n = 2$ (Table 2). Although in principle we could enlarge the feature vector dimension Ω (cf. Eq. (4)) by further increasing n, it would cause the following two problems. One is that the computational time would be significantly longer, and the other is that the results might be even worse due to the so-called "overfitting" [82] or "high-dimension disaster" [56] problem in statistical prediction. Accordingly, for the current benchmark dataset, the optimal value for n was 2. Furthermore, as we can see from Table 2, when the predictions were based on $\mathbb{CP}(11)$ with $\Omega = 121$, the best overall success rate was achieved. In other words, when the general form of PseAAC (Eq. (4)) for the HSP samples was formulated by

$$\mathbf{P} = \begin{bmatrix} f_1^{(2)} & f_2^{(2)} & \ldots & f_u^{(2)} & \ldots & f_{121}^{(2)} \end{bmatrix}^{\mathbf{T}}, \tag{17}$$

where $f_u^{(2)}$ $(u = 1, 2, \ldots, 121)$ has the same meaning as that of Eq. (8), the best prediction quality was obtained by iHSP-PseRAAAC in identifying the HSP functional types.

In addition, to our best knowledge, so far there is no existing predictor whatsoever that could be used to identify the functional types of HSPs according to their sequence information alone, and hence no comparison could be made in this study for iHSP-PseRAAAC with its counterparts. However, it would be instructive to make a comparison of the overall success rate achieved by iHSP-PseRAAAC with those achieved by completely random guess (CRG) and weighted random guess (WRG) [83]. Obviously, the overall success rate OA (cf. Eq. (16)) in identifying the HSPs among their six functional types by CRG is given by

$$\mathrm{OA(CRG)} = \frac{1}{6} \approx 16.67\%, \tag{18}$$

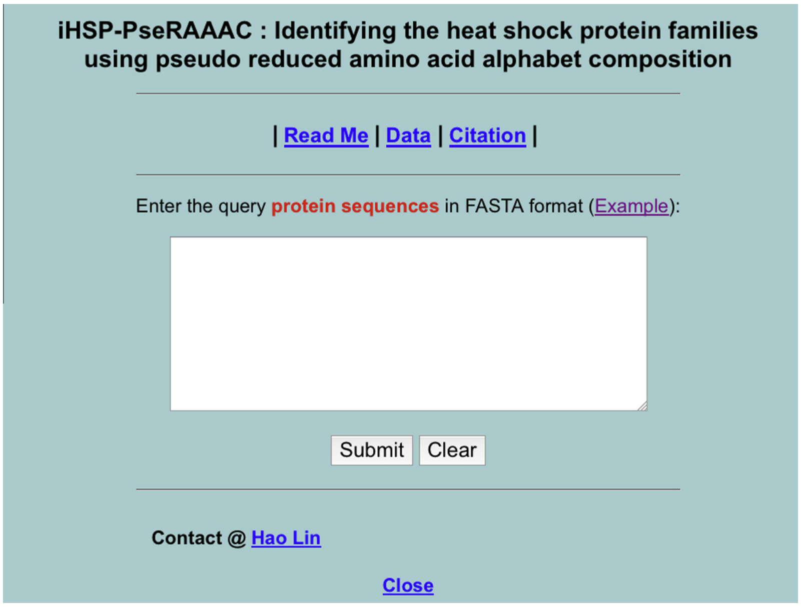
whereas that by WRG is given by [84]

$$\mathrm{OA(WRG)} = \frac{(N_1)^2 + (N_2)^2 + (N_3)^2 + (N_4)^2 + (N_5)^2 + (N_6)^2}{(N)^2}, \tag{19}$$

where N is the number of HSPs in the benchmark dataset $\mathbb{S}$, N_1 is the number of HSPs in the subset $\mathbb{S}_1$, N_2 is the number of HSPs in

Table 2
Results obtained by iHSP-PseRAAAC in identifying heat shock protein families with dipeptide or *n*-peptide ($n = 2$) composition based on different reduced amino acid alphabet approaches.

HSP family	Subset	Metrics (Eq. (16))	Cluster profile (Eq. (5))				
			$\mathbb{CP}(13)$	$\mathbb{CP}(11)$	$\mathbb{CP}(9)$	$\mathbb{CP}(8)$	$\mathbb{CP}(5)$
			Dimension Ω when $n = 2$ (Eqs. (8) and (9))				
			169	121	81	64	25
HSP20	$\mathbb{S}_1$	Sn(1)	84.87%	87.68%	82.63%	81.51%	63.02%
		Sp(1)	96.82%	96.36%	97.01%	95.88%	95.65%
		MCC(1)	0.82	0.82	0.81	0.77	0.64
HSP40	$\mathbb{S}_2$	Sn(2)	94.84%	95.31%	95.39%	95.46%	90.38%
		Sp(2)	84.82%	84.87%	81.49%	78.90%	55.22%
		MCC(2)	0.97	0.99	0.96	0.93	0.63
HSP60	$\mathbb{S}_3$	Sn(3)	69.94%	66.87%	64.42%	60.12%	36.19%
		Sp(3)	98.28%	98.93%	98.24%	99.12%	98.07%
		MCC(3)	0.69	0.69	0.64	0.66	0.39
HSP70	$\mathbb{S}_4$	Sn(4)	79.86%	79.15%	74.91%	72.44%	54.06%
		Sp(4)	86.77%	86.54%	86.45%	87.36%	86.64%
		MCC(4)	0.55	0.54	0.52	0.51	0.39
HSP90	$\mathbb{S}_5$	Sn(5)	55.17%	51.72%	43.10%	48.28%	20.69%
		Sp(5)	99.58%	99.89%	99.89%	99.79%	99.45%
		MCC(5)	0.27	0.30	0.28	0.28	0.16
HSP100	$\mathbb{S}_6$	Sn(6)	67.06%	69.41%	69.41%	64.70%	31.76%
		Sp(6)	99.37%	99.84%	99.62%	99.73%	98.83%
		MCC(6)	0.76	0.83	0.77	0.79	0.40
Overall		OA	87.42%	87.82%	86.11%	85.30%	73.35%



iHSP-PseRAAAC : Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition

Enter the query protein sequences in FASTA format (Example):

Fig.2. A semi-screenshot to show the top page of the iHSP-PseRAAAC web server. Its website address is at http://lin.uestc.edu.cn/server/iHSP-PseRAAAC.

the subset $\mathbb{S}_2$, and so forth (see Eq. (1) and Table 1). Substituting these data in Table 1 and Eq. (19), we obtain

$$\begin{aligned}\mathrm{OA(WRG)} &= \frac{(357)^2+(1279)^2+(163)^2+(283)^2+(58)^2+(85)^2}{(2225)^2}\\ &\simeq 37.99\%.\end{aligned} \tag{20}$$

In contrast, the overall success rate achieved by iHSP-RAAAC was 87.82 (cf. Table 2). Comparing it with the results in Eqs. (18) and (20) indicates that the overall success rate by the current predictor is more than 70% higher than that by the CRG and approximately 50% higher than that by the WRG, indicating that iHSP-PseRAAAC may at least become an easy and useful tool for timely identifying the functional types of HSPs. More important, we hope that this study can play the role of "cast a brick to attract jade," as is often quoted in a Chinese proverb, to stimulate more in-depth studies in this area.

Web server guide

For the convenience of the vast majority of experimental scientists, below we give a step-by-step guide on how to use the iHSP-PseRAAAC web server to get their desired results.

Step 1

Open the web server at http://lin.uestc.edu.cn/server/iHSP-PseRAAAC, and you will see the top page of iHSP-PseRAAAC on your computer screen, as shown in Fig. 2. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

Step 2

Either type or copy/paste the query heat shock protein sequence into the input box at the center of Fig. 2. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a "greater than" symbol (>) in the first column, followed by lines of sequence data. The words right after the ">" symbol in the single initial line are optional and only used for the purpose of identification and description. All lines should be no longer than 120 characters and usually do not exceed 80 characters. The sequence ends if another line starting with a ">" symbol appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

Step 3

Click on the Submit button to see the predicted result. For example, if you use the six query HSP sequences in the Example window as the input, after clicking the Submit button you will see the following shown on the screen of your computer: the outcome for the 1st query sample is "HSP100"; the outcome for the 2nd query sample is "HSP90"; the outcome for the 3rd query sample is "HSP70"; the outcome for the 4th query sample is "HSP60"; the outcome for the 5th query sample is "HSP40"; the outcome for the 6th query sample is "HSP20". All of these results are fully consistent with the experimental observations as summarized in Supporting Information S1 of the supplementary material. It takes a few seconds for the above computation before the predicted result appears on your computer screen; the greater number of query sequences and the longer each sequence, the more time that is usually needed.

Step 4

Click on the Citation button to find the relevant articles that document the detailed development and algorithm of iHSP-PseRAAAC.

Step 5

Click on the Data button to download the benchmark datasets used to train and test the iHSP-PseRAAAC predictor.

Acknowledgments

The authors are very much indebted to the two anonymous reviewers for their valuable comments, which were very helpful for strengthening the presentation of this article. Many thanks are also due to Patrick D. Silva and his colleagues at the Indian Institute of Technology for their helpful suggestions in using HSPIR.

This work was supported by grants from the National Nature Scientific Foundation of China (61100092 and 61202256).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ab.2013.05.024.

References

[1] P. Ritossa, Problems of prophylactic vaccinations of infants, Riv. Ist. Sieroter. Ital. 37 (1962) 79–108.

[2] Z. Chen, T. Zhou, X. Wu, Y. Hong, Z. Fan, H. Li, Influence of cytoplasmic heat shock protein 70 on viral infection of *Nicotiana benthamiana*, Mol. Plant Pathol. 9 (2008) 809–817.

[3] J.L. Edwards, P.J. Hansen, Elevated temperature increases heat shock protein 70 synthesis in bovine two-cell embryos and compromises function of maturing oocytes, Biol. Reprod. 55 (1996) 341–346.

[4] M.G. Goldstein, Z. Li, Heat-shock proteins in infection-mediated inflammation-induced tumorigenesis, J. Hematol. Oncol. 2 (2009) 5.

[5] T.J. Hubbard, C. Sander, The role of heat-shock and chaperone proteins in protein folding: Possible molecular mechanisms, Protein Eng. 4 (1991) 711–717.

[6] X.C. Zeng, S. Bhasin, X. Wu, J.G. Lee, S. Maffi, C.J. Nichols, K.J. Lee, J.P. Taylor, L.E. Greene, E. Eisenberg, Hsp70 dynamics in vivo: Effect of heat shock and protein aggregation, J. Cell Sci. 117 (2004) 4991–5000.

[7] Y. Mallouk, M. Vayssier-Taussat, J.V. Bonventre, B.S. Polla, Heat shock protein 70 and ATP as partners in cell homeostasis [review], Int. J. Mol. Med. 4 (1999) 463–474.

[8] J.E. Hamos, B. Oblas, D. Pulaski-Salo, W.J. Welch, D.G. Bole, D.A. Drachman, Expression of heat shock proteins in Alzheimer's disease, Neurology 41 (1991) 345–350.

[9] A.G. Pockley, Heat shock proteins, inflammation, and cardiovascular disease, Circulation 105 (2002) 1012–1017.

[10] Y.R. Wu, C.K. Wang, C.M. Chen, Y. Hsu, S.J. Lin, Y.Y. Lin, H.C. Fung, K.H. Chang, G.J. Lee-Chen, Analysis of heat-shock protein 70 gene polymorphisms and the risk of Parkinson's disease, Hum. Genet. 114 (2004) 236–241.

[11] K.R. Ratheesh, N.S. Nagarajan, S.P. Arunraj, D. Sinha, V.B. Veedin Rajan, V.K. Esthaki, P. D'Silva, HSPIR: A manually annotated heat shock protein information resource, Bioinformatics 28 (2012) 2853–2855.

[12] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition [50th anniversary year review], J. Theor. Biol. 273 (2011) 236–247.

[13] W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

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

[15] S.F. Altschul, Evaluating the statistical significance of multiple distinct local alignments, in: S. Suhai (Ed.), Theoretical and Computational Methods in Genome Research, Plenum, New York, 1997, pp. 1–14.

[16] J.C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, Comput. Chem. 17 (1993) 149–163.

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

[18] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99 (1986) 152–162.

[19] K.C. Chou, C.T. Zhang, Predicting protein folding types by distance functions that make allowances for amino acid interactions, J. Biol. Chem. 269 (1994) 22014–22020.

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

[21] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255 (erratum: vol. 44, p. 60).

[22] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.

[23] L. Nanni, A. Lumini, D. Gupta, A. Garg, Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (2012) 467–475.

[24] D. Zou, Z. He, J. He, Y. Xia, Supersecondary structure prediction using Chou's pseudo amino acid composition, J. Comput. Chem. 32 (2011) 271–278.

[25] S.W. Zhang, Y.L. Zhang, H.F. Yang, C.H. Zhao, Q. Pan, 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 (2008) 565–572.

[26] K.K. Kandaswamy, G. Pugalenthi, S. Moller, E. Hartmann, K.U. Kalies, P.N. Suganthan, T. Martinetz, 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 (2010) 1473–1479.

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

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

[29] M. Hayat, A. Khan, Discriminating outer membrane proteins with fuzzy *K*-nearest neighbor algorithms based on the general form of Chou's PseAAC, Protein Pept. Lett. 19 (2012) 411–421.

[30] M. Khosravian, F.K. Faramarzi, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods, Protein Pept. Lett. 20 (2012) 180–186.

[31] H. Mohabatkar, M.M. Beigi, K. Abdolahi, S. Mohsenzadeh, Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach, Med. Chem. 9 (2013) 133–137.

[32] M. Mohammad Beigi, M. Behjati, H. Mohabatkar, Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach, J. Struct. Funct. Genomics 12 (2011) 191–197.

[33] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.

[34] R. Zia Ur, A. Khan, Identifying GPCRs and their types with Chou's pseudo amino acid composition: An approach from multi-scale energy representation and position specific scoring matrix, Protein Pept. Lett. 19 (2012) 890–903.

[35] X.Y. Sun, S.P. Shi, J.D. Qiu, S.B. Suo, S.Y. Huang, R.P. Liang, Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform, Mol. Biosyst. 8 (2012) 3178–3184.

[36] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization, Amino Acids 34 (2008) 653–660.

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

[38] H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition, Protein Pept. Lett. 17 (2010) 1207–1214.

[39] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of $GABA_A$ receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 281 (2011) 18–23.

[40] H. Lin, C. Ding, L.F. Yuan, W. Chen, H. Ding, Z.Q. Li, F.B. Guo, J. Hung, N.N. Rao, Predicting subchloroplast locations of proteins based on the general form of Chou's pseudo amino acid composition: Approached from optimal tripeptide composition, Int. J. Biomath. 6 (2013). article 1350003.

[41] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, 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 (2009) 17–26.

[42] W. Chen, P.M. Feng, H. Lin, K.C. Chou, IRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68 (open access at http://nar.oxfordjournals.org/content/41/6/e68.long).

[43] W. Chen, H. Lin, P.M. Feng, C. Ding, Y.C. Zuo, K.C. Chou, INuc-PhysChem: A sequence-based predictor for identifying nucleosomes via physicochemical properties, PLoS One 7 (2012) e47843.

[44] B.Q. Li, T. Huang, L. Liu, Y.D. Cai, K.C. Chou, Identification of colorectal cancer related genes with mRMR and shortest path in protein–protein interaction network, PLoS One 7 (2012) e33393.

[45] T. Huang, J. Wang, Y.D. Cai, H. Yu, K.C. Chou, Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma, PLoS One 7 (2012) e34460.

[46] Y. Jiang, T. Huang, C. Lei, Y.F. Gao, Y.D. Cai, K.C. Chou, Signal propagation in protein interaction network during colorectal cancer progression, Biomed. Res. Int. 2013 (2013) 287019 (open access at http://www.hindawi.com/journals/bmri/2013/287019/2013).

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

[48] D.S. Cao, Q.S. Xu, Y.Z. Liang, Propy: A tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.

[49] H.B. Shen, K.C. Chou, PseAAC: A flexible web-server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.

[50] H. Lin, H. Ding, Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition, J. Theor. Biol. 269 (2011) 64–69.

[51] H. Lin, Q.Z. Li, Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components, J. Comput. Chem. 28 (2007) 1463–1466.

[52] M. Shu, X. Cheng, Y. Zhang, Y. Wang, Y. Lin, L. Wang, Z. Lin, Predicting the activity of ACE inhibitory peptides with a novel mode of pseudo amino acid composition, Protein Pept. Lett. 18 (2011) 1233–1243.

[53] B. Liao, J.B. Jiang, Q.G. Zeng, W. Zhu, Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition, Protein Pept. Lett. 18 (2011) 1086–1092.

[54] W. Liu, K.C. Chou, Protein secondary structural content prediction, Protein Eng. 12 (1999) 1041–1050.
[55] K.C. Chou, Using pair-coupled amino acid composition to predict protein secondary structure content, J. Protein Chem. 18 (1999) 473–480.
[56] T. Wang, J. Yang, H.B. Shen, K.C. Chou, Predicting membrane protein types by the LLDA algorithm, Protein Pept. Lett. 15 (2008) 915–921.
[57] W.Z. Lin, J.A. Fang, X. Xiao, K.C. Chou, ILoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins, Mol. Biosyst. 9 (2013) 634–644.
[58] P.D. Thomas, K.A. Dill, An iterative method for extracting energy-like quantities from protein structures, Proc. Natl. Acad. Sci. USA 93 (1996) 11628–11633.
[59] L.A. Mirny, E.I. Shakhnovich, Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics, and function, J. Mol. Biol. 291 (1999) 177–196.
[60] A.D. Solis, S. Rackovsky, Optimized representations and maximal information in proteins, Proteins 38 (2000) 149–164.
[61] A.G. de Brevern, C. Etchebest, S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks, Proteins 41 (2000) 271–287.
[62] A.G. de Brevern, New assessment of a structural alphabet, In Silico Biol. 5 (2005) 283–289.
[63] A.P. Joseph, G. Agarwal, S. Mahajan, J.C. Gelly, L.S. Swapna, B. Offmann, F. Cadet, A. Bornot, M. Tyagi, H. Valadie, B. Schneider, C. Etchebest, N. Srinivasan, A.G. de Brevern, A short survey on protein blocks, Biophys. Rev. 2 (2010) 137–147.
[64] A.P. Joseph, N. Srinivasan, A.G. de Brevern, Improvement of protein structure comparison using a structural alphabet, Biochimie 93 (2011) 1434–1445.
[65] C. Etchebest, C. Benros, A. Bornot, A.C. Camproux, A.G. de Brevern, A reduced amino acid alphabet for understanding and designing protein adaptation to mutation, Eur. Biophys. J. 36 (2007) 1059–1069.
[66] W. Chen, P. Feng, H. Lin, Prediction of ketoacyl synthase family using reduced amino acid alphabets, J. Ind. Microbiol. Biotechnol. 39 (2012) 579–584.
[67] Y.C. Zuo, Q.Z. Li, Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet, Peptides 30 (2009) 1788–1793.
[68] Y.L. Chen, Q.Z. Li, L.Q. Zhang, Using increment of diversity to predict mitochondrial proteins of malaria parasite: Integrating pseudo-amino acid composition and structural alphabet, Amino Acids 42 (2012) 1309–1316.
[69] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[70] Y.D. Cai, G.P. Zhou, K.C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (2003) 3257–3263.
[71] W. Chen, H. Lin, Prediction of midbody, centrosome, and kinetochore proteins based on gene ontology information, Biochem. Biophys. Res. Commun. 401 (2010) 382–384.
[72] M. Hayat, A. Khan, Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, J. Theor. Biol. 271 (2011) 10–17.
[73] X. Xiao, P. Wang, K.C. Chou, INR-PhysChem: A sequence-based predictor for identifying nuclear receptors and their subfamilies via physical–chemical property matrix, PLoS One 7 (2012) e30869.
[74] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.
[75] C.C. Chang, C.J. Lin, LIBSVM: A library for support vector machines, 2001 (software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm).
[76] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[77] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973–1979.
[78] K.C. Chou, C.T. Zhang, Review: Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[79] K.C. Chou, H.B. Shen, Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2008) 153–162 (updated version: Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, Nat. Sci. 2 (2010) 1090–1103).
[80] K.C. Chou, Z.C. Wu, X. Xiao, ILoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, PLoS One 6 (2011) e18258.
[81] K.C. Chou, Z.C. Wu, X. Xiao, ILoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. Biosyst. 8 (2012) 629–641.
[82] K.C. Chou, A key driving force in determination of protein structural classes, Biochem. Biophys. Res. Commun. 264 (1999) 216–224.
[83] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.
[84] K.C. Chou, D.W. Elrod, Protein subcellular location prediction, Protein Eng. 12 (1999) 107–118.