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

Peng-Mian Feng, Wei Chen, Hao Lin, Kuo-Chen Chou

Abstract

Heat shock proteins (HSPs), first discovered in 1962 [1], 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.

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

Heat shock proteins (HSPs), first discovered in 1962 [1], are a type 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
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 \( P \geq 40\% \) pairwise sequence identity to any other in the same subset. Finally, we obtained a dataset \( \mathcal{S} \) of 2225 HSPs classified into six families, as can be formulated by

\[
\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4 \cup \mathcal{S}_5 \cup \mathcal{S}_6, \tag{1}
\]

where the subset \( \mathcal{S}_1 \) contains 357 HSP20 sequences, \( \mathcal{S}_2 \) contains 1279 HSP40 sequences, \( \mathcal{S}_3 \) contains 163 HSP60 sequences, \( \mathcal{S}_4 \) contains 283 HSP70 sequences, \( \mathcal{S}_5 \) contains 58 HSP90 sequences, and \( \mathcal{S}_6 \) contains 85 HSP100 sequences (Table 1) and where \( \cup \) 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

\[
P = R_1 R_2 R_3 R_4 R_5 R_6 \ldots R_L. \tag{2}
\]

where \( R_1 \) represents the 1st residue of the protein \( P \), \( R_2 \) represents the 2nd residue of the protein \( P \), 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>
<thead>
<tr>
<th>Dataset</th>
<th>Family</th>
<th>Number of HSP samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{S}_1 )</td>
<td>HSP20</td>
<td>357</td>
</tr>
<tr>
<td>( \mathcal{S}_2 )</td>
<td>HSP40</td>
<td>1279</td>
</tr>
<tr>
<td>( \mathcal{S}_3 )</td>
<td>HSP60</td>
<td>163</td>
</tr>
<tr>
<td>( \mathcal{S}_4 )</td>
<td>HSP70</td>
<td>283</td>
</tr>
<tr>
<td>( \mathcal{S}_5 )</td>
<td>HSP90</td>
<td>58</td>
</tr>
<tr>
<td>( \mathcal{S}_6 )</td>
<td>HSP100</td>
<td>85</td>
</tr>
<tr>
<td>( \mathcal{S} )</td>
<td>Overall</td>
<td>2225</td>
</tr>
</tbody>
</table>
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

\[
P = \begin{bmatrix} f_1^{(1)} & f_2^{(1)} & \ldots & f_{20}^{(1)} \\ f_1^{(2)} & f_2^{(2)} & \ldots & f_{20}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(20)} & f_2^{(20)} & \ldots & f_{20}^{(20)} \end{bmatrix},
\]

(3)

where \( f_u^{(v)} \) (\( u = 1, 2, \ldots, 20 \)) 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 viral 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 sub mitochondria locations [36], identifying risk type of human papillomaviruses [37], identifying cyclin proteins [38], predicting GABA\(_\alpha\) 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

\[
P = [\Psi_1 \ \Psi_2 \ \ldots \ \Psi_s \ \ldots \ \Psi_{u}]^T,
\]

(4)

where the subscript \( \Omega \) is an integer and its value, as well as the components \( \Psi_u \) (\( u = 1, 2, \ldots, \Omega \)), 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 \( \% \) 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 \( n \gg 2 \), 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 Brevem 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 \((\Psi, \Psi)\) dihedral angles [63]. The labels \( m \) and \( d \) of PBs are prototypes for the central region of \( \alpha \)-helix and \( \beta \)-strand, respectively; labels \( a \) through \( c \) primarily represent the N-cap of \( \beta \)-strand; labels \( g \) through \( j \) are specific to coils; labels \( k \) and \( l \) correspond to \( N \)-cap of \( \alpha \)-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—CP(13), CP(11), CP(9), CP(8), and CP(5)—as formulated below:

\[
\begin{align*}
CP(13) &= \{G; IV; FYW; A; L; M; E; QRK; P; ND; HS; T; C\} \\
CP(11) &= \{G; IV; FYW; A; LM; EQRK; P; ND; HS; T; C\} \\
CP(9) &= \{G; IV; FYW; ALM; EQRK; P; ND; HS; T; C\} \\
CP(8) &= \{G; IV; FYW; ALM; EQRK; P; ND; HS; T; C\} \\
CP(5) &= \{G; IV FYW; ALMEQKR; P; NDHSCT\}
\end{align*}
\]

(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

\[
\Psi_u = f_u^{(1)}, \ u = 1, 2, \ldots, \Omega,
\]

(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

\[
\Omega = \begin{bmatrix}
13^1 = 13 & \text{for CP(13) cluster} \\
11^1 = 11 & \text{for CP(11) cluster} \\
9^1 = 9 & \text{for CP(9) cluster} \\
8^1 = 8 & \text{for CP(8) cluster} \\
5^1 = 5 & \text{for CP(5) cluster}
\end{bmatrix}
\]

(7)

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

\[
\Psi_u = f_u^{(2)}, \ u = 1, 2, \ldots, \Omega.
\]

(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

\[
\Omega = \begin{bmatrix}
13^2 = 169 & \text{for CP(13) cluster} \\
11^2 = 121 & \text{for CP(11) cluster} \\
9^2 = 81 & \text{for CP(9) cluster} \\
8^2 = 64 & \text{for CP(8) cluster} \\
5^2 = 25 & \text{for CP(5) cluster}
\end{bmatrix}
\]

(9)
For the tripeptide composition, the \( u \)th component in Eq. (4) will be formulated as

\[
\Psi_u = \delta_{u}^{(3)}, \ u = 1, 2, \ldots, \Omega, \tag{10}
\]

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

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

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.

\[\text{Support vector machine}\]

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 \( \gamma \) 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 \( \gamma = 0.125 \).

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.”

\[\text{Results and discussion}\]

\[\text{Criteria for performance evaluation}\]

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 \( S_i \) \((i = 1, 2, \ldots, 6)\) and those not belonging to the subset \( S_i \) are respectively defined by

\[
\begin{align*}
\Lambda^+(i) &= \frac{N^+(i) - N^{-}(i)}{N^+(i)} \\
\Lambda^-(i) &= \frac{N^-(i) - N^{+}(i)}{N^-(i)}
\end{align*}
\tag{12}
\]

where \( N^+(i) \) is the total number of the investigated HSP samples in the subset \( S_i \), whereas \( N^{-}(i) \) is the number of HSP samples in \( S_i \) 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 correctly predicted belonging to \( S_i \). The overall sub-success prediction rate for each of the subsets is given by [77]

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

It is obvious from Eqs. (12) and (13) that, if and only if all of the samples in the subset \( S_i \) (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:

\[
\begin{align*}
\text{Sn}(i) &= \frac{TP(i)}{TP(i)+FN(i)} \\
\text{Sp}(i) &= \frac{TN(i)}{TN(i)+FP(i)} \\
\text{MCC}(i) &= \sqrt{\frac{TP(i) \times TN(i) - FP(i) \times FN(i)}{TP(i) + FP(i) + FN(i) + TN(i)}} \\
\text{OA} &= \frac{1}{N} \sum_{i=1}^{N} \text{TP}(i)
\end{align*}
\tag{14}
\]

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 \( S \).

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

\[
\begin{align*}
\text{TP}(i) &= N^+(i) - N^{-}(i) \\
\text{TN}(i) &= N^{-}(i) - N^{+}(i) \\
\text{FP}(i) &= N^{+}(i) \\
\text{FN}(i) &= N^{-}(i)
\end{align*}
\tag{15}
\]

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

\[
\begin{align*}
\text{Sn}(i) &= 1 - \frac{N^{-}(i)}{N^{+}(i)} \\
\text{Sp}(i) &= 1 - \frac{N^{+}(i)}{N^{-}(i)} \\
\text{MCC}(i) &= \sqrt{\frac{1 - \frac{N^{+}(i) \times N^{-}(i)}{N^{+}(i) + N^{-}(i)}}{1 + \frac{N^{+}(i) \times N^{-}(i)}{N^{+}(i) + N^{-}(i)}}} \\
\text{OA} &= \frac{1}{N} \sum_{i=1}^{N} N^+(i) - N^{-}(i)
\end{align*}
\tag{16}
\]

Obviously, when \( N^+(i) = 0 \), meaning that none of the HSP samples in subset \( S_i \) was mispredicted belonging to other subsets, we have the sensitivity \( \text{Sn}(i) = 1 \), whereas when \( N^{-}(i) = N^{+}(i) \), meaning that all of the HSP samples in subset \( S_i \) was mispredicted belonging to the other subsets, we have the sensitivity \( \text{Sn}(i) = 0 \). Likewise, when \( N^{-}(i) = 0 \), meaning that none of the HSP samples in the other subsets was incorrectly predicted belonging to the subset \( S_i \), we have the specificity \( \text{Sp}(i) = 1 \), whereas when \( N^{+}(i) = N^{-}(i) \), meaning that all of the HSP samples in the other subsets were incorrectly predicted belonging to the other subsets.
predicted belonging to the subset $S_i$, we have the specificity $Sp(i) = 0$. When $N^*_i(i) > N^*_i(i) = 0$ ($i = 1, 2, \ldots, 6$), meaning that none of the HSP samples in all of the subsets of $S$ (cf. Eq. (1)) was incorrectly predicted, we have the overall accuracy $OA = 1$, whereas when $N^*_i(i) = N^*_i(i) = N^*_i(i) = N^*_i(i)$ ($i = 1, 2, \ldots, 6$), meaning that the HSP samples in all of the subsets of $S$ were mispredicted, we have the overall accuracy $OA = 0$. When $N^*_i(i) = N^*_i(i) = 0$, meaning that none of the HSP samples in the subset $S_i$ was mispredicted, we have $MCC(i) = 1$; when $N^*_i(i) = N^*_i(i)/2$ and $N^*_i(i) = N^*_i(i)/2$, we have $MCC(i) = 0$, meaning no better than random prediction for the HSP samples in the subset $S_i$; when $N^*_i(i) = N^*_i(i)$ and $N^*_i(i) = N^*_i(i)$, we have $MCC(i) = -1$, meaning total disagreement between prediction and observation for the HSP samples in the subset $S_i$. 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 Matthew’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 $S$ (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 $\Omega$ (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 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 \[ P = \left[ f_1^{(2)} f_2^{(2)} \ldots f_N^{(2)} \right]^T, \]

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

\[ OA(CRG) = \frac{1}{6} \approx 16.67\%, \]

whereas that by WRG is given by [84]

\[ OA(WRG) = \frac{(N_1)^2 + (N_2)^2 + (N_3)^2 + (N_4)^2 + (N_5)^2 + (N_6)^2}{N^2}, \]

where $N$ is the number of HSPs in the benchmark dataset $S$, $N_i$ is the number of HSPs in the subset $S_i$, and $N_2$ is the number of HSPs in

<table>
<thead>
<tr>
<th>HSP family</th>
<th>Subset</th>
<th>Metrics (Eq. (16))</th>
<th>Cluster profile (Eq. (5))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CP(13)</td>
<td>CP(11)</td>
</tr>
<tr>
<td>HSP20</td>
<td>$S_1$</td>
<td>Sn(1)</td>
<td>84.87%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp(1)</td>
<td>96.82%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCC(1)</td>
<td>0.82</td>
</tr>
<tr>
<td>HSP40</td>
<td>$S_2$</td>
<td>Sn(2)</td>
<td>94.84%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp(2)</td>
<td>84.82%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCC(2)</td>
<td>0.97</td>
</tr>
<tr>
<td>HSP60</td>
<td>$S_3$</td>
<td>Sn(3)</td>
<td>69.94%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp(3)</td>
<td>98.28%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCC(3)</td>
<td>0.69</td>
</tr>
<tr>
<td>HSP70</td>
<td>$S_4$</td>
<td>Sn(4)</td>
<td>79.86%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp(4)</td>
<td>86.77%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCC(4)</td>
<td>0.55</td>
</tr>
<tr>
<td>HSP90</td>
<td>$S_5$</td>
<td>Sn(5)</td>
<td>55.17%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp(5)</td>
<td>99.58%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCC(5)</td>
<td>0.27</td>
</tr>
<tr>
<td>HSP100</td>
<td>$S_6$</td>
<td>Sn(6)</td>
<td>67.06%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sp(6)</td>
<td>99.37%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MCC(6)</td>
<td>0.76</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>OA</td>
<td>87.42%</td>
</tr>
</tbody>
</table>
the subset \( S_2 \), and so forth (see Eq. (1) and Table 1). Substituting these data in Table 1 and Eq. (19), we obtain

\[
\text{OA(WRG)} = \frac{(357)^2 + (1279)^2 + (163)^2 + (283)^2 + (58)^2 + (85)^2}{(2225)^2}
\]

\[
\approx 37.99\%.
\]

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 (19) 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-RAAAC 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-RAAAC web server to get their desired results.

Step 1
Open the web server at http://lin.uestc.edu.cn/server/iHSP-RAAAC, and you will see the top page of iHSP-RAAAC 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 “>” 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-RAAAC.

Step 5
Click on the Data button to download the benchmark datasets used to train and test the iHSP-RAAAC 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