Springer

Dear Author,

Here are the proofs of your article.

- You can submit your corrections **online**, via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For fax submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/ corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI].
If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information go to: http://www.springerlink.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

ArticleTitle	Quat-2L: a web-server for predicting protein quaternary structural attributes	
Article Sub-Title		
Article CopyRight	Springer Science+Business Media B.V. (This will be the copyright line in the final PDF)	
Journal Name	Molecular Diversity	
Corresponding Author	Family Name	**Xiao**
	Particle	
	Given Name	**Xuan**
	Suffix	
	Division	Computer Department
	Organization	Jing-De-Zhen Ceramic Institute
	Address	333001, Jing-De-Zhen, China
	Division	
	Organization	Gordon Life Science Institute
	Address	13784 Torrey Del Mar Drive, San Diego, CA, CA 92130, USA
	Email	xxiao@gordonlifescience.org
Author	Family Name	**Wang**
	Particle	
	Given Name	**Pu**
	Suffix	
	Division	Computer Department
	Organization	Jing-De-Zhen Ceramic Institute
	Address	333001, Jing-De-Zhen, China
	Email	
Author	Family Name	**Chou**
	Particle	
	Given Name	**Kuo-Chen**
	Suffix	
	Division	
	Organization	Gordon Life Science Institute
	Address	13784 Torrey Del Mar Drive, San Diego, CA, CA 92130, USA
	Email	

Schedule	Received	13 November 2009
	Revised	
	Accepted	21 January 2010

Abstract By hybridizing the functional-domain and sequence-correlated pseudo amino acid composition approaches, a 2-layer predictor called "Quat-2L" was developed for predicting the quaternary structural attribute of a protein according to its sequence information alone. The 1st layer is for identifying the query protein as monomer, homo-oligomer, or hetero-oligomer. If the result thus obtained turns out to be homo-oligomer or hetero-oligomer, then the prediction will be automatically continued to further identify it as belonging to which one of the following six subtypes: (1) dimer, (2) trimer, (3) tetramer, (4) pentamer, (5) hexamer, and (6) octamer. The overall success rate by Quat-2L for the 1st layer identification was 71.14%; while the overall success rates by the 2nd layer for homo-oligomers and hetero-oligomers were 76.91 and 82.52%, respectively.

	These rates were derived by the jackknife cross-validation tests on the stringent benchmark data set in which none of proteins has ≥**60%** pairwise sequence identity to any other in the same subset. As a web-server, Quat-2L is freely accessible to the public via http://icpr.jci.jx.cn/bioinfo/Quat-2L, by which one can get the desired 2-level results in about 15 s.
Keywords (separated by '-')	SMART - Function domain composition - Pseudo amino acid composition - Complexity measure factor - Fuzzy K nearest neighbor
Footnote Information	Electronic supplementary material The online version of thisarticle (doi:10.1007/s11030-010-9227-8) contains supplementarymaterial, which is available to authorized users.

Metadata of the article that will be visualized in OnlineAlone

Electronic Supplementary Material	The Below is the Electronic Supplementary Material. **MOESM1:** ESM 1 (DOC 131 kb). **MOESM2:** ESM 1 (PDF 2035 kb).

Journal: 11030
Article: 9227

Springer
the language of science

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
Affiliation	Please confirm the affiliation.	

Mol Divers
DOI 10.1007/s11030-010-9227-8

FULL LENGTH PAPER

Quat-2L: a web-server for predicting protein quaternary structural attributes

Xuan Xiao · Pu Wang · Kuo-Chen Chou

Received: 13 November 2009 / Accepted: 21 January 2010

Abstract By hybridizing the functional-domain and sequence-correlated pseudo amino acid composition approaches, a 2-layer predictor called "Quat-2L" was developed for predicting the quaternary structural attribute of a protein according to its sequence information alone. The 1st layer is for identifying the query protein as monomer, homo-oligomer, or hetero-oligomer. If the result thus obtained turns out to be homo-oligomer or hetero-oligomer, then the prediction will be automatically continued to further identify it as belonging to which one of the following six subtypes: (1) dimer, (2) trimer, (3) tetramer, (4) pentamer, (5) hexamer, and (6) octamer. The overall success rate by Quat-2L for the 1st layer identification was 71.14%; while the overall success rates by the 2nd layer for homo-oligomers and hetero-oligomers were 76.91 and 82.52%, respectively. These rates were derived by the jackknife cross-validation tests on the stringent benchmark data set in which none of proteins has ≥**60**% pairwise sequence identity to any other in the same subset. As a web-server, Quat-2L is freely accessible to the public via http://icpr.jci.jx.cn/bioinfo/Quat-2L, by which one can get the desired 2-level results in about 15 s.

Keywords SMART · Function domain composition · Pseudo amino acid composition · Complexity measure factor · Fuzzy K nearest neighbor

Introduction

The biological functions of proteins are closely related with their structures. According to their structural hierarchy, proteins are generally classified into four levels: primary, secondary, tertiary, and quaternary (see, e.g., [1]).

This study is focused on predicting the quaternary structure attributes of proteins based on their sequences because information thus obtained is very useful for screening the candidates of proteins for their 3D (dimensional) structure determination by the X-ray cryptography technique and for drug design as elaborated in [2].

Although a number of computational methods [3–8] have been developed for predicting protein quaternary structure attributes in the last eight years or so, all these methods were either confined to homo-oligomers [3–7] or established based on the training data sets that contained many uncertain annotation terms [8], such as "probable," "potential," and "by similarity," which might affect the reliability of the predicted results.

In a previous study, we constructed a new data set in which all the protein samples with uncertain annotations were removed. Based on such a stringent data set and by hybridizing functional domain composition [9] and pseudo amino acid composition [10,11], we developed a user-friendly web-server predictor called "PQSA-Pred" [2,12], by which one can easily identify a query protein as belonging to monomer, homo-oligomer, or hetero-oligomer according to its sequence information alone.

Electronic supplementary material The online version of this article (doi:10.1007/s11030-010-9227-8) contains supplementary material, which is available to authorized users.

X. Xiao · P. Wang
Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333001, China
e-mail: xxiao@gordonlifescien[illegible]

X. Xiao (✉) · K.-C. Chou
Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA
e-mail: xiaoxuan0326@yahoo.com

However, the following in-depth questions might be raised: Whether the homo-oligomer belongs to homo-dimer, homo-trimer, homo-tetramer, homo-pentamer, homo-hexamer, or homo-octamer? Whether the hetero-oligomer belongs to hetero-dimer, hetero-trimer, hetero-tetramer, hetero-pentamer, hetero-hexamer, or hetero-octamer?

This study was initiated in an attempt to address this kind of in-depth questions.

Materials and methods

In order to develop an effective method for predicting protein attributes, the following three things are indispensable: a valid benchmark data set, an effective mathematical expression for the samples that can truly reflect their intrinsic correlation with the target concerned [11], and a powerful prediction algorithm (or engine). Now, let us work on the 1st necessity.

Benchmark data sets

Protein sequences were taken from the Release 15.6 of UniProtKB. The detailed procedures are as follows. (1) Open the web-page at http://www.uniprot.org/. (2) Click the button "*Fields*," followed by selecting "General annotation [CC]" for Fields; "Subunit structure" for Topic; typing in "monomer," "homodimer," "homotrimer," "homotetramer," "homopentamer," 'homohexamer," "homoheptamer," "homo-octamer," "heterodimer," "heterotrimer," "heterotetramer," "heteropentamers," "heterohexamer," "heteroheptamer," or "hetero-octamer" for Term; and selecting "Experimental" for Confidence. (3) Click it Add & Search.

Thus, we collected a total of 12,839 protein sequences from 13 different quaternary structural attributes. In order to ensure a consistent and high-quality calibrated data set, all the collected sequences were screened strictly according to the following criteria: (1) Sequences annotated with "fragment" were excluded. also, sequences with less than 50 amino acid residues were removed because they might just be fragments. (2) Sequences which contain irregular amino acid characters such as "X" or "Z" were removed. (3) Those entries which annotated with more than one quaternary attribute were removed because of lacking uniqueness. (4) In order to reduce redundancy or homology bias, the program called "Cd-hit" [13] was utilized to winnow those sequences which have $\geq 60\%$ sequence identity to any other in a same subset.

Finally, we obtained a data set $\mathbb{S}$ containing 5,495 sequences of which 1,223 belong to monomers, 3,036 to homo-oligomers, and 1,236 to hetero-oligomers, as can be represented by three main subsets: i.e.,

$$\mathbb{S} = \mathbb{S}^{\text{mono}} \cup \mathbb{S}^{\text{homo}} \cup \mathbb{S}^{\text{hetero}} \tag{1}$$

where $\cup$ is the symbol for union in the set theory, $\mathbb{S}^{\text{mono}}$ is the subset containing monomers only, $\mathbb{S}^{\text{homo}}$ containing homo-oligomers only, and $\mathbb{S}^{\text{hetero}}$ containing hetero-oligomers only. Moreover, according to their experimental annotations, the protein sequences in $\mathbb{S}^{\text{homo}}$ and $\mathbb{S}^{\text{hetero}}$ can be further classified into six sub-subsets respectively: i.e.,

$$\begin{cases} \mathbb{S}^{\text{homo}} = \mathbb{S}_2^{\text{homo}} \cup \mathbb{S}_3^{\text{homo}} \cup \mathbb{S}_4^{\text{homo}} \cup \mathbb{S}_5^{\text{homo}} \cup \mathbb{S}_6^{\text{homo}} \cup \mathbb{S}_8^{\text{homo}} \\ \mathbb{S}^{\text{hetero}} = \mathbb{S}_2^{\text{hetero}} \cup \mathbb{S}_3^{\text{hetero}} \cup \mathbb{S}_4^{\text{hetero}} \cup \mathbb{S}_5^{\text{hetero}} \cup \mathbb{S}_6^{\text{hetero}} \cup \mathbb{S}_8^{\text{hetero}} \end{cases} \tag{2}$$

where $\mathbb{S}_2^{\text{homo}}$ is the sub-subset containing the homodimers only, $\mathbb{S}_3^{\text{homo}}$ containing the homotrimers only, and so forth; while $\mathbb{S}_2^{\text{hetero}}$ is the subset containing the heterodimers only, $\mathbb{S}_3^{\text{hetero}}$ containing the heterotrimers only, and so forth.

The number of protein sequences in each of the subsets or sub-subsets is given in Table 1, and the 5,495 sequences classified according to Eqs. 1 and 2 are given in the Online Supplementary Information A.

Below, let us address the second necessity, i.e., how to define the descriptor of a protein sequence that can catch the core feature for predicting its quaternary structural attribute. As summarized in a recent comprehensive review [14], the following two strategies are often adopted to represent protein sequences: the sequential mode and the discrete mode. The most straightforward sequential mode for a protein chain is to use its entire amino acid sequence. Its advantage is its ability to contain the most complete information. In order to get the desired results via this approach, the sequence similarity search-based tools, such as BLAST [15], are usually applied to conduct the prediction. However, this kind of approach failed to work when the query protein did not have significant sequence similarity to those of known attributes. Thus, various discrete models were proposed. The simplest one was to use the amino acid composition of a protein to represent it [16]. However, using the amino acid composition mode to represent a protein, all its sequence-order information would be missing.

In order to overcome such a shortcoming, the pseudo amino acid composition approaches [10] based on the functional domain mode and the sequence-correlated mode will be adopted in this study, as described in the next section.

Pseudo amino acid (PseAA) composition descriptors

According to its original definition [10], the PseAA composition or PseAAC is actually a set of discrete number as long as it is different from the classical amino acid (AA) composition or AAC [11]. One of the very effective PseAAC models is the functional domain (FunD) descriptor. As is well known, proteins usually consist of several domains or modules, with each having a distinct evolutionary origin and function. Actually, several functional domain (FunD) databases were developed, such as COG [17], KOG [17], Pfam [18],

Journal: 11030-MODI Article No.: 9227 MS Code: MODI438.4 TYPESET DISK LE CP Disp.:2010/2/3 Pages: 7 Layout: Large

Table 1 Breakdown of the monomers, homo-oligomers, and hetero-oligomers of the benchmark data set obtained in Materials and methods section

Attribute	Description	Number of sequences
Monomer	$\mathbb{S}^{\text{mono}}$	1223
Homo-oligomer	Homodimer $\mathbb{S}_2^{\text{homo}}$	1965
	Homotrimer $\mathbb{S}_3^{\text{homo}}$	252
	Homotetramer $\mathbb{S}_4^{\text{homo}}$	593
	Homopentamer $\mathbb{S}_5^{\text{homo}}$	20
	Homohexamer $\mathbb{S}_6^{\text{homo}}$	151
	Homo-octamer $\mathbb{S}_8^{\text{homo}}$	55
	Overall $\mathbb{S}^{\text{homo}}$	3036
Hetero-oligomer	Heterodimer $\mathbb{S}_2^{\text{hetero}}$	795
	Heterotrimer $\mathbb{S}_3^{\text{hetero}}$	84
	Heterotetramer $\mathbb{S}_4^{\text{homo}}$	250
	Heteropentamer $\mathbb{S}_5^{\text{homo}}$	11
	Heterohexamer $\mathbb{S}_6^{\text{homo}}$	71
	Hetero-octamer $\mathbb{S}_8^{\text{homo}}$	25
	Overall $\mathbb{S}^{\text{homo}}$	1236

SMART [19], and CDD [20]. Here, let us use the SMART database [19] to formulate the FunD descriptor for a protein sample. The SMART contains 11,124 domain entries. Thus, by following the same procedures as described in [9,21], the protein sample can be formulated as

$$\mathbf{P}_{\text{PseAA}}^{\text{FunD}} = \left[\delta_1^{\text{Fun}} \quad \delta_2^{\text{Fun}} \quad \cdots \delta_i^{\text{Fun}} \quad \cdots \quad \delta_{11124}^{\text{Fun}}\right]^{\mathbf{T}} \tag{3}$$

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

$$\delta_i^{\text{Fun}} = \begin{cases} 1, & \text{when a hit is found for } \mathbf{P} \text{ in SAMRT} \\ 0, & \text{otherwise} \end{cases} \quad (i=1, 2, \ldots, 11124) \tag{4}$$

Thus, rather than a 20D vector as defined in the classic AAC space [16], the protein $\mathbf{P}$ is now corresponding to an 11124D (dimensional) vector with its components defined, respectively, by the sequence patterns of the 11,124 functional domains in the SMART database [19]. By doing so, not only many sequence pattern features but also considerable function-related information is naturally incorporated in the descriptor of Eq. 3.

If no hit whatsoever was found in Eq. 4, then the protein sample $\mathbf{P}_{\text{FunD}}$ as defined in Eq. 3 would become a naught vector that would be meaningless. In order to deal with this kind of situation, we would instead use the sequence-correlated PseAAC to express the protein sample. The sequence-correlated PseAAC was originally introduced for predicting protein subcellular localization and membrane protein type [10]. It was also used for predicting enzyme functional class later [22]. According to its original definition, the sequence-correlated PseAAC for a given protein sample is expressed by a set of $20 + \lambda$ discrete numbers, where the first 20 numbers represent the 20 components of the classical AAC while the additional λ numbers incorporate some of its sequence-order information via various different kinds of correlating modes. Since it was introduced [10], the sequence-correlated PseAAC and its various different modes were proposed to improve the prediction quality of various protein attributes (see, e.g., [23–36]).

In this study, we used the following PseAAC mode to represent the protein sample $\mathbf{P}$:

$$\mathbf{P}_{\text{PseAA}}^{\text{SqCo}} = (p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots, p_{30})^{\mathbf{T}} \tag{5}$$

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

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{j=20+1}^{30}\theta_j}, & (1 \le k \le 20) \\ \dfrac{w\theta_k}{\sum_{i=1}^{20} f_i + w\sum_{j=20+1}^{30}\theta_j}, & (20+1 \le k \le 30) \end{cases} \tag{6}$$

where f_i $(i = 1, 2, \ldots, 20)$ are the occurrence frequencies of the 20 native amino acids in a protein [37], w is the weight factor, and θ_j $(j = 1, 2, \ldots, 10)$ are the complexity measure factors [38,39] for the protein sequence concerned, which were obtained as follows: (1) converting the original protein sequence to 0–1 string; (2) evolving 10 times by cellular automation [40] and calculating the complexity measurement factor from every evolving string. For the details as to how to generate the 10 complexity measure factors, refer to Appendix A. In this study, we chose $w = 1/3000$ to make the results of Eq. 6 within the range easier to be handled (w can, of course, be assigned with other values, but this would not make a significant difference to the final results [10]).

Journal: 11030-MODI Article No.: 9227 MS Code: MODI438.4 ☐TYPESET ☐DISK ☐LE ☐CP Disp.:2010/2/3 Pages: 7 Layout: Large

Table 2 The jackknife success rates in identifying the monomer, homo-oligomer, and hetero-oligomer with the first layer of Quat-2L predictor on the benchmark data set in the Online Supporting Information A

Quaternary attribute	Number of proteins	Number of correct predictions	Success rate (%)
Monomer	1223	843	68.93
Homo-oligomer	3036	2291	75.46
Hetero-oligomer	1236	775	62.70
Overall	5495	3909	71.14

Table 3 The jackknife success rates in identifying the six subtypes of homo-oligomers and hetero-oligomers with the second layer of Quat-2L predictor on the benchmark data set in the Online Supporting Information A

Subtype	Number of proteins	Number of correct prediction	Success rate (%)
Homo-oligomer			
Homodimer	1965	1864	94.86
Homotrimer	252	120	47.62
Homotetramer	593	286	48.23
Homopentamer	20	4	20.00
Homohexamer	151	39	25.83
Homo-octamer	55	22	40.00
Overall	3036	2335	76.91
Hetero-oligomer			
Heterodimer	795	771	96.98
Heterotrimer	84	43	51.19
Heterotetramer	250	153	61.20
Heteropentamer	11	2	18.18
Heterohexamer	71	47	66.20
Hetero-octamer	25	4	16.00
Overall	1236	1020	82.52

The fuzzy K-nearest neighbor (K-NN classifier)

Here, let us address the third necessity, i.e., what prediction algorithm (or engine) should be used in this study? The K-nearest neighbor (K-NN) rule [41] is one of the oldest and simplest methods for performing nonparametric classification. The main idea of K-NN rule can be stated as follows: Given a test sample with unknown label, its label is assigned according to the labels of its K nearest neighbors in the training set. Recently, the K-NN classifier has been successfully used to predict protein subcellular localization [42–44], membrane protein type [45], protease type [21], and many other protein attributes [12].

Fuzzy K-NN classification method [46] is a special variation of the K-NN classification family. Instead of roughly assigning the label based on a voting from the nearest neighbors, it attempts to estimate the membership values that indicate how much degree the query sample belong to the classes concerned, as briefly described in the following.

Suppose $\{\mathbf{P}_1, \mathbf{P}_1, \ldots, \mathbf{P}_N\}$ is a set of vectors representing N proteins in the training set which has been classified to M classes: $\{C_1, C_2, \ldots, C_M\}$, where C_i denotes the i-th class. Thus, for a query protein $\mathbf{P}$, its fuzzy membership value for the i-th class is given by

$$\mu_i(\mathbf{P}) = \frac{\sum_{j=1}^{K} \mu_i(\mathbf{P}_j) \mathrm{d}(\mathbf{P}, \mathbf{P}_j)^{-2/(\varphi-1)}}{\sum_{j=1}^{K} \mathrm{d}(\mathbf{P}, \mathbf{P}_j)^{-2/(\varphi-1)}} \quad (7)$$

where ~~K is~~ the number of the nearest neighbors counted; $\mu_i(\mathbf{P}_j)$ is the fuzzy membership value of the protein $\mathbf{P}_j$ to the i-th class (it is set to 1 if the real label of $\mathbf{P}_j$ is C_i; otherwise, zero); $\mathrm{d}(\mathbf{P}, \mathbf{P}_j)$ is the distance between the query protein $\mathbf{P}$ and its j-th nearest protein $\mathbf{P}_j$ in the training data set; and $\varphi(> 1)$ is the fuzzy coefficient for determining how heavily the distance is weighted when calculating each nearest neighbor's contribution to the membership value. Various metrics can be chosen for $\mathrm{d}(\mathbf{P}, \mathbf{P}_j)$, such as Euclidean distance, Hamming distance [47], and Mahalanobis distance [16,48]. In this article, the Euclidean metric was used, and the values of φ and K will be mentioned later. After calculating all the memberships for a query protein, it is assigned to the class with which it has the highest membership value; i.e., the predicted class for the query protein $\mathbf{P}$ should be

Journal: 11030-MODI Article No.: 9227 MS Code: MODI438.4 ☐TYPESET ☐DISK ☐LE ☐CP Disp.:2010/2/3 Pages: 7 Layout: Large

$$C_u = \mathbf{argmax}_i \{\mu_i(\mathbf{P})\} \quad (8)$$

where u is the argument of i that maximizes $\mu_i(\mathbf{P})$.

Now, we have all the three necessities available for predicting the quaternary structural attributes of proteins. The predictor thus established is called **Quat-2L**, where "Quat" means for predicting the quaternary structural attribute of protein, and "2L" means the prediction consisting of two layers. The 1st layer prediction engine is to identify a query protein as monomer, homo-oligomer or hetero-oligomer; if it is a homo-oligomer or hetero-oligomer, then the process will be automatically continued with the second-layer prediction engine to further identify its corresponding attribute among the following six categories: (1) dimer, (2) trimer, (3) tetramer, (4) pentamer, (5) hexamer, and (6) octamer.

It is instructive to stress that the following principle should be complied during the prediction.

The self-consistency principle

During the prediction process, the following "self-consistency principle" must be observed: the query protein **P** and the proteins used to train the prediction engine should be defined in a same infrastructural frame having exactly the same dimensions no matter which kind of descriptor is used to represent the protein samples. For example, if the query protein is defined in the 11124D PseAA FunD composition space (cf. Eq. 3), then the prediction should be performed using all those proteins in the training set that can be defined in the exactly same 11124D FunD PseAAC space as well. If the query protein in the 11124D FunD PseAAC space is a naught vector and, hence, must be defined instead in the (20+10)D sequence-correlated PseAAC space as shown in Eq. 5, then all the proteins in the training data set must also be formulated in the same (20+10)D SqCo PseAAC space for training the predictor.

Results and discussion

Of the 5,495 protein sequences in the benchmark data set $\mathbb{S}$ (see Online Supporting Information A), 4,305 were found with at least one hit in searching the SMART functional domain database [19] and hence can be formulated as $\mathbf{P}^{\text{FunD}}_{\text{PseAA}}$, the functional domain PseAAC of Eq. 3. The remaining $(5495 - 4305) = 1{,}190$ sequences were without any hit, and, hence, formulated by $\mathbf{P}^{\text{SqCo}}_{\text{PseAA}}$, the sequence-correlated PseAAC of Eq. 5.

Three cross-validation test methods are often used for examining the accuracy of a statistical predictor, i.e., independent data set test, subsampling (or K-fold cross-over validation) test, and jackknife test (Chou & Zhang 1995). Of these three, however, the jackknife test is deemed the most objective as elucidated in a recent comprehensive review [14], and it has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g., [21,24,25,28,30,32,35,49–57]). Accordingly, in this study, we also used the jackknife crossover validation to examine the prediction quality of the current method.

The values of φ and K used in Eq. 7 were determined by optimizing the overall jackknife success rate through a 2D search.

The jackknife cross-validation results obtained with Quat-2L on the benchmark data set (cf. Online Supplementary Information A) are given in Tables 2 and 3. It can be seen from Table 2 that the overall success rate by the 1st layer of the predictor among the types of monomer, homo-oligomer, and hetero-oligomer is 71.14%. Meanwhile, as shown in Table 3, the overall success rate by the second layer of the predictor among the six subtypes of homo-oligomers and six subtypes of hetero-oligomers are 76.91 and 85.52%, respectively.

The quaternary structure of proteins is an extremely complicated problem. Even though, a quite decent success rate in predicting their quaternary structural attributes can be achieved by hybridizing the FunD PseAAC and sequence-correlated PseAAC approaches. It is anticipated that with more experimental data available in protein quaternary structures, the prediction quality will be further improved.

Conclusion

A protein may be formed by a single peptide chain or multiple peptide chains. Multiple-chain proteins may be formed as a homo-oligomer by identical peptide chains (subunits) or as a hetero-oligomer formed by different peptide chains. Furthermore, according to the number of the constituent chains, the homo-oligomer or hetero-oligomer can be respectively classified as dimer, trimer, tetramer, pentamer, hexamer, octamer, and so forth.

The function of a protein is closely relevant to its quaternary structure and hence very useful information can be obtained by studying the relation of the quaternary structural attribute of a protein chain and its sequence feature.

By hybridizing two different forms of PseAAC modes, a two-layer predictor called Quat-2L was developed. It can be used to predict the quaternary structural attribute of a protein chain according to its sequence information alone with quite decent success rate.

Quat-2L is available as a web-server at http://icpr.jci.jx.cn/bioinfo/Quat-2L.

Acknowledgements The authors wish to thank the two anonymous reviewers whose constructive comments were very helpful for enhancing the quality of the presentation of this article. This study was supported by the grants from the National Natural Science Foundation of China (No. 60961003), the Department of Education of JiangXi Province (No.GJJ09271), and the Plan for Training Youth Scientists (stars of Jing-Cang) of Jiangxi Province.

References

1. Voet D, Voet JG (1995) Biochemistry, 2. Wiley, New York 180–185
2. Xiao X, Wang P, Chou KC (2009) Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. J Appl Crystallogr 42: 169–173. doi:10.1107/S0021889809002751
3. Garian R (2001) Prediction of quaternary structure from primary structure. Bioinformatics 17: 551–556. doi:10.1093/bioinformatics/17.6.551
4. Chou KC, Cai YD (2003) Predicting protein quaternary structure by pseudo amino acid composition. Proteins: Struct Funct Genet 53: 282–289. doi:10.1002/prot.10500
5. Zhang SW, Pan Q, Zhang HC, Zhang YL, Wang HY (2003) Classification of protein quaternary structure with support vector machine. Bioinformatics 19: 2390–2396. doi:10.1093/bioinformatics/btg331
6. Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (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. doi:10.1007/s00726-006-0263-8
7. Zhang SW, Chen W, Yang F, Pan Q (2008) Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. Amino Acids 35: 591–598. doi:10.1007/s00726-008-0086-x
8. Carugo O (2007) A structural proteomics filter: prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences. Appl Crystallogr 40: 986–989. doi:10.1107/S0021889807041076
9. Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769. doi:10.1074/jbc.M204161200
10. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct Funct Genet (Erratum: ibid, 2001, vol 44, 60) 43: 246–255. doi:10.1002/prot.1035
11. Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteom 6: 262–274. doi:10.2174/157016409789973707
12. Chou KC, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. Nat Sci 2:63–92. http://www.scirp.org/journal/NS/. doi:10.4236/ns.2009.12011
13. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659. doi:10.1093/bioinformatics/btl158
14. Chou KC, Shen HB (2007) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370: 1–16. doi:10.1016/j.ab.2007.07.006
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402. doi:gka562[pii]
16. Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Struct Funct Genet 21: 319–344. doi:10.1002/prot.340210406
17. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform 4:41. doi:10.1186/1471-2105-4-41
18. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T et al (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247–D251. doi:10.1093/nar/gkj149
19. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34: D257–D260. doi:10.1093/nar/gkj079
20. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M et al (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35: D237–D240. doi:10.1093/nar/gkl951
21. Chou KC, Shen HB (2008) 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. doi:10.1016/j.bbrc.2008.08.125
22. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19. doi:10.1093/bioinformatics/bth466
23. 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. doi:10.1016/j.jtbi.2006.06.025
24. Ding YS, Zhang TL (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognit Lett 29: 1887–1892. doi:10.1016/j.patrec.2008.06.007
25. 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. doi:10.2174/092986609787848045
26. Georgiou DN, Karakasidis TE, Nieto JJ, 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. doi:10.1016/j.jtbi.2008.11.003
27. Jiang X, Wei R, Zhang TL, Gu Q (2008) 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. doi:10.2174/092986608784246443
28. Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15: 612–616. doi:10.2174/092986608784966930
29. 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. doi:10.1016/j.jtbi.2008.02.004
30. Lin H, Ding H, Feng-Biao Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15: 739–744
31. Lin H, Wang H, Ding H, Chen YL, Li QZ (2009) Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor 57: 321–330. doi:10.1007/s10441-008-9067-4
32. Qiu JD, Huang JH, Liang RP, Lu XQ (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. doi:10.1016/j.ab.2009.04.009
33. Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, Li ML (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. doi:10.1016/j.jtbi.2009.03.028
34. Zhang GY, Fang BS (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. doi:10.1016/j.jtbi.2008.03.015

Journal: 11030-MODI Article No.: 9227 MS Code: MODI438.4 ☐TYPESET ☐DISK ☐LE ☐CP Disp.:2010/2/3 Pages: 7 Layout: Large

35. Zhang GY, Li HC, Fang BS (2008) Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Pept Lett 15: 1132–1137. doi:10.2174/092986608786071184
36. Zhou XB, Chen C, Li ZC, Zou XY (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. doi:10.1016/j.jtbi.2007.06.001
37. Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269:22014–22020
38. Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61. doi:10.1007/s00726-004-0148-7
39. Xiao X, Shao SH, Huang ZD, Chou KC (2006) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482. doi:10.1002/jcc.20354
40. Wolfram S (1984) Cellular automation as models of complexity. Nature 311:419–424
41. Cover TM, Hart PE (1967) Nearest neighbour pattern classification. IEEE Trans Inform Theory (IT) 13:21–27
42. Chou KC, Shen HB (2006) Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157. doi:10.1016/j.bbrc.2006.06.059
43. Chou KC, Shen HB (2007) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678. doi:10.1002/jcb.21096
44. Chou KC, Shen HB (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteom Res 6: 1728–1734. doi:10.1021/pr060635i
45. Chou KC, Shen HB (2007) 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. doi:10.1021/pr060635i
46. Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbours algorithm. IEEE Trans Syst Man Cybern 15:580–585
47. Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis: chapter 11 discriminant analysis; chapter 12 multivariate analysis of variance; chapter 13 cluster analysis. Academic Press, London 322–381
48. Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2:49–55
49. Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738. doi:10.1023/A:1020713915365
50. Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins: Struct Funct Genet 44: 57–59. doi:10.1002/prot.1071
51. Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins: Struct Funct Genet 50: 44–48. doi:10.1002/prot.10251
52. Wang T, Yang J, Shen HB, Chou KC (2008) Predicting membrane protein types by the LLDA algorithm. Protein Pept Lett 15: 915–921. doi:10.2174/092986608785849308
53. Chen K, Kurgan M, Kurgan L (2008) Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values. J Biomed Sci Eng (JBiSE) 1: 1–9. http://www.srpublishing.org/journal/jbise/. doi:10.4236/jbise.2008.11001
54. Shen HB, Song JN, Chou KC (2009) Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J Biomed Sci Eng (JBiSE) 2:136–143. http://www.srpublishing.org/journal/jbise/. doi:10.4236/jbise.2009.23024
55. Chou KC, Shen HB (2009) FoldRate: a web-server for predicting protein folding rates from primary sequence. Open Bioinform J 3:31–50. http://www.bentham.org/open/tobioij/
56. 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. doi:10.1016/j.jtbi.2009.08.004
57. 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. doi:10.1016/j.jtbi.2009.07.031

Journal: 11030-MODI Article No.: 9227 MS Code: MODI438.4 ☐TYPESET ☐DISK ☐LE ☐CP Disp.:2010/2/3 Pages: 7 Layout: Large