Interdiscip Sci Comput Life Sci
DOI 10.1007/s12539-015-0124-9

ORIGINAL RESEARCH ARTICLE

Identifying Antioxidant Proteins by Using Optimal Dipeptide Compositions

Pengmian Feng[1] · Wei Chen[2] · Hao Lin[3]

Received: 26 February 2015 / Revised: 29 June 2015 / Accepted: 29 August 2015

Abstract Antioxidant proteins are a kind of molecules that can terminate cellular and DNA damages caused by free radical intermediates. The use of antioxidant proteins for prevention of diseases has been intensively studied in recent years. Thus, accurate identification of antioxidant proteins is essential for understanding their roles in pharmacology. In this study, a support vector machine-based predictor called **AodPred** was developed for identifying antioxidant proteins. In this predictor, the sequence was formulated by using the optimal 3-gap dipeptides obtained by using feature selection method. It was observed by jackknife cross-validation test that **AodPred** can achieve an overall accuracy of 74.79 % in identifying antioxidant proteins. As a user-friendly tool, **AodPred** is freely accessible at http://lin.uestc.edu.cn/server/AntioxiPred. To maximize the convenience of the vast majority of experimental scientists, a step-by-step guide is provided on how to use the web server to obtain the desired results.

✉ Wei Chen
greatchen@heuu.edu.cn

✉ Hao Lin
hlin@uestc.edu.cn

[1] School of Public Health, North China University of Science and Technology, Tangshan 063000, China

[2] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China

[3] 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

Keywords Antioxidant protein · AodPred · *g*-gap dipeptides composition · Support vector machine

1 Introduction

A healthy cell membrane that made primarily of fat is selectively permeable. It not only allows water and oxygen to flow freely into the cell, but also permits carbon dioxide and other waste products to flow freely out of the cell. However, cells in our body are assaulted by free radicals all day long. Free radicals are unstable molecules desperately seeking electrons from surrounding atoms for stability. Free radicals can be derived either from normal essential metabolic processes in the body or from external sources such as exposures to X-rays, ozone, cigarette smoking, air pollutants, and industrial chemicals [1]. Once created, they will set off a chain reaction that begins a cycle of oxidative damage and then alter or destroy cells. In addition, they can also cause cellular and DNA damages [2], which in turn contribute to aging and the onset of various diseases [3–5].

Antioxidant proteins are a kind of molecules that can interact with and neutralize free radicals. By donating an electron to rampaging free radicals, antioxidant proteins can terminate the chain reactions caused by free radical intermediates. This biological process protects the cells from further damage or death. The application of antioxidant proteins in pharmacology is intensively studied in recent years. Antioxidant proteins have been investigated for the prevention of diseases such as cancer [6], coronary heart disease [7], and even altitude sickness [8]. Antioxidant proteins also attracted considerable attention in relation to longevity [9].

There are several enzyme systems within the body that can scavenge free radicals, and the principle micronutrient

Published online: 07 September 2015

(vitamins) antioxidant proteins are vitamin E (α-tocopherol), vitamin C (ascorbic acid), and B-carotene [1]. However, our body cannot manufacture these micronutrients, so they must be supplied in the dietary supplements. Therefore, it is urgent to search for effective, nontoxic natural compounds with anti-oxidative activities.

Although biochemical experiment is an objective method to identify antioxidant proteins, it is time-consuming. With the avalanche of protein sequences generated in recent years, it is highly desirable to develop computational methods to accurately identify antioxidant proteins. Recently, a computational model based on star graph topological indices was proposed to identify antioxidant proteins [5]. However, sequences in their dataset share high-sequence similarities and some sequences in their dataset even with 100 % sequences identity. It has been demonstrated that the predictive accuracy is closely related to sequence identity [10], and high-sequence similarity can surely lead to the overestimation of prediction performance. Later on, Feng et al. [11] proposed a naive Bayes model to predict antioxidant proteins based on optimal dipeptides and obtained an accuracy of 66.88 % in the jackknife test. However, the predictive accuracy is still unsatisfactory.

All these works could yield quite encouraging results, and each of them did play a role in simulating the development of antioxidant identification. Unfortunately, to the best of our knowledge, no web server whatsoever was provided for these methods, and hence, their usage is quite limited, particularly for the broad experimental scientists. Therefore, in the present study, we proposed a novel method to identify antioxidant proteins based on the sequence information. A feature selection technique was used to pick out a number of informative features. On the basis of the optimal features, the support vector machine was performed to establish the prediction model. Results of jackknife cross-validation test demonstrate that the proposed method is reliable. Based on this method, a free online server called **AodPred** was built to provide a useful tool for identifying antioxidant proteins.

2 Materials and Methods

2.1 Benchmark Dataset

Sequences of antioxidant proteins were collected from the UniProt database (release 2014_02) with the keyword "antioxidant." In order to prepare a reliable dataset, the following steps were performed: (i) only proteins with the experimentally confirmed anti-oxidative activities were included, and (ii) proteins containing nonstandard letters, i.e., "B," "X," or "Z," were excluded as their meanings are ambiguous. Therefore, we obtained 710 protein sequences with experimentally proven/confirmed anti-oxidative activity and set them as the positive samples. The negative samples consist of 1567 experimentally proved non-antioxidant proteins that have been used in our previous work [11]. As elaborated in [10], a benchmark dataset containing many redundant samples with high similarity would lack statistical representativeness. A predictor, if trained and tested by such a biased benchmark dataset, might yield misleading results with overestimated accuracy. To remove the homologous sequences from the benchmark dataset, the CD-HIT program [12] was used to eliminate proteins with >60 % identity in positive and negative datasets. After such a screening procedure, we finally obtained a benchmark dataset containing 253 antioxidant proteins and 1552 non-antioxidant proteins to build the prediction model.

2.2 Support Vector Machine

Support vector machine (SVM) is an effective method for supervised pattern recognition and has been widely used in the realm of bioinformatics [13–16]. The basic idea of SVM is to transform the data into a high-dimensional feature space and then determine the optimal separating hyperplane. Because of its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the best classification hyperplane. The SVM implementation was based on the freely available package LIBSVM written by Chang and Lin, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The regularization parameter C and the kernel width parameter γ were tuned via the grid search method in the fivefold cross-validation.

2.3 Sequences Representation

The proximate dipeptide composition has been widely used in computational proteomics. However, the intrinsic properties of protein sequences usually deposit in higher tier correlation of residues because of the hydrogen bonding in the secondary structure [17, 18]. Instead of the proximate dipeptide composition, the g-gap dipeptide composition describing the long-range correlations between two residues was proposed and has demonstrated its effectiveness in the realm of protein identifications [18, 19]. Therefore, in the present work, the g-gap dipeptide composition was used to encode the proteins in benchmark dataset.

Suppose a protein sequence **P** with L amino acid residues as follows:

$$\mathbf{P} = R_1R_2R_3R_4...R_{L-2}R_{L-1}R_L \tag{1}$$

where R_1 represents the first residue in the protein sequence, R_2 represents the second residue, and so forth. For the g-gap dipeptide, the feature vector contains $20 \times 20 = 400$ components and can be formulated as,

$$\mathbf{P} = \left[f_1^g \quad f_2^g \quad \cdots f_i^g \cdots \quad f_{400}^g \right]^{\mathbf{T}} \tag{2}$$

where the symbol $\mathbf{T}$ denotes the transposition of the vector; f_i^g denotes the frequency of the ith g-gap dipeptide in the protein sequence and is defined as,

$$f_i^g = \frac{n_i^g}{\sum_{i=1}^{400} n_i^g} = \frac{n_i^g}{(L - g - 1)} \tag{3}$$

where n_i^g denotes the number of the ith g-gap dipeptide and g is an integral number within the range of [0, 9] with a step of 1. $g = 0$ indicates the correlation of two proximate residues; $g = 1$ describes the correlation between two residues with one residue interval; and $g = 2$ indicates the correlation between two residues with the interval of two residues and so forth.

2.4 Performance Evaluation

In statistical prediction, three cross-validation methods, namely the independent dataset test, the subsampling (e.g., five- or tenfold cross-validation) test, and the jackknife test, are often used to evaluate the performance of the predicted methods in practical application. Among the three test methods, the jackknife test is the least arbitrary and can yield a unique result for a given benchmark dataset [10] and hence has been increasingly and widely adopted by investigators to examine the power of various prediction methods [20, 21]. Accordingly, we used jackknife cross-validation in this study to examine the anticipated success rates of the predictor. In the process of feature selection, for reducing the computational time, the fivefold cross-validation approach was used to deal with the parameter optimization.

For a binary classification problem, the sensitivity (*Sn*), specificity (*Sp*), and accuracy (*Acc*) were often used to measure the prediction quality and they are expressed as [13–21]

$$\begin{cases} Sn = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\ Sp = \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \\ Acc = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \end{cases} \tag{4}$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively.

Let N^+ is the total number of the antioxidant proteins investigated, while N_-^+ is the number of antioxidant proteins incorrectly predicted to be non-antioxidant proteins; N^- is the total number of the non-antioxidant proteins investigated, while N_+^- is the number of the non-antioxidant proteins incorrectly predicted to be antioxidant proteins. We have

$$\begin{cases} \mathrm{TP} = N^+ - N_-^+ \\ \mathrm{TN} = N^- - N_+^- \\ \mathrm{FP} = N_+^- \\ \mathrm{FN} = N_-^+ \end{cases} \tag{5}$$

Substituting Eq. (5) into Eq. (4), we obtain

$$\begin{cases} Sn = 1 - \dfrac{N_-^+}{N^+} \\ Sp = 1 - \dfrac{N_+^-}{N^-} \\ Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} \end{cases} \tag{6}$$

Equation (6) has the same meanings as Eq. (4), while it is more intuitive and easier to understand [14, 23]. When $N_-^+ = 0$ meaning that none of the antioxidant proteins was mispredicted to be non-antioxidant proteins, we have the sensitivity $\mathrm{Sn} = 1$; while $N_-^+ = N^+$ meaning that all the antioxidant proteins were mispredicted to be non-antioxidant proteins, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_+^- = 0$ meaning that none of the non-antioxidant proteins was mispredicted to be antioxidant proteins, we have the specificity $\mathrm{Sp} = 1$; while $N_+^- = N^-$ meaning that all the non-antioxidant proteins were incorrectly predicted as antioxidant proteins, we have the specificity $\mathrm{Sp} = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of the antioxidant proteins and none of the non-antioxidant proteins was incorrectly predicted, we have the overall accuracy $Acc = 1$; while $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the antioxidant proteins and all the non-antioxidant proteins were mispredicted, we have the overall accuracy $Acc = 0$.

2.5 Feature Selection

Inclusion of redundant and noisy information would cause poor predictive results. To improve the prediction quality, the analysis of variance (ANOVA) was performed to select the optimal features derived from g-gap dipeptide compositions. ANOVA has been widely used for feature selection in computational proteomics [18, 19]. The principle of ANOVA is to measure the feature variances by calculating the ratio (F-value) of features between groups and within groups [23]. The F-value of the ξth g-gap dipeptide in benchmark dataset is defined as,

$$F(\xi) = \frac{s_B^2(\xi)}{s_W^2(\xi)} \tag{7}$$

where $s_B^2(\xi)$ and $s_W^2(\xi)$ denote the sample variance between groups (also called mean square between, MSB) and sample variance within groups (also called Mean Square Within, MSW), respectively, and are calculated by

$$s_B^2(\xi) = \sum_{i=1}^{K} m_i \left(\frac{\sum_{j=1}^{m_i} f_\xi^g(i,j)}{m_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_\xi^g(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 \Bigg/ df_B \tag{8}$$

$$s_W^2(\xi) = \sum_{i=1}^{K}\sum_{j=1}^{m_i} \left(f_\xi^g(i,j) - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_\xi^g(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 \Bigg/ df_W \tag{9}$$

where $df_B = K - 1$ and $df_W = M - K$ are degrees of freedom for MSB and MSW, respectively. K and M represent the number of groups (here $K = 2$) and total number of samples (here $M = 1805$), respectively. $f_\xi^g(i,j)$ indicates the frequency of the ξth g-gap dipeptide of the jth sample in the ith group. m_i indicates the number of samples in the ith group ($m_1 = 253$, $m_2 = 1552$).

The $F(\xi)$-value in Eq. (7) reveals the correlation between the ξth feature and the group variables. The $F(\xi)$ will become large as the MSB becomes increasingly greater than the MSW. In the absence of differences between groups, the $F(\xi)$ will be near to 1. In other words, the feature with a larger $F(\xi)$ indicates that it is a more relevant one for the target to be predicted. Accordingly, we ranked the 400 g-gap dipeptides according to their $F(\xi)$ values. And then based on the ranked g-gap dipeptides, we performed the Incremental Feature Selection (IFS) strategy to find an optimal subset of features that gives the highest predictive accuracy. During the IFS procedure, the feature subset starts with one feature with the highest F-score. A new feature subset was composed when one more feature with the second highest F-score was added. By adding these features sequentially from the higher to lower ranks, 400 feature sets will be obtained. The τth feature set can be formulated as

$$S_\tau = \{f_1, f_2, \cdots, f_\tau\} \quad (1 \le \tau \le 400) \tag{10}$$

For each of the 400 feature sets, a support vector machine-based model was constructed and examined using the fivefold cross-validation test on the benchmark dataset. By doing so, we can obtain an IFS curve in a 2D Cartesian coordinate system with index τ as its abscissa (or X-coordinate) and the overall accuracy as its ordinate (or Y-coordinate). The optimal feature set is expressed as

$$S_\Theta = \{f_1, f_2, \cdots, f_\Theta\} \tag{11}$$

with which the IFS curve reaches its peak. In other words, in the 2D coordinate system, when X = Θ, the value of accuracy is the maximum.

3 Results and Discussions

3.1 Prediction Performance

In order to determine whether a specific type of dipeptides is mostly contributable for antioxidant protein identification, we considered ten different kinds of g-gap dipeptides ($g = 0, 1, 2, \ldots, 9$). For each kind of g-gap dipeptides, we ranked the 400 dipeptides according to their F-scores and plotted its IFS curve in Fig. 1. Among the ten IFS curves, we found that the best predictive result ($Acc = 90.7\ \%$) was obtained when $g = 3$ and $\Theta = 158$. It means that the IFS curve reaches its peak with the optimal feature set S_{158}, in which the features are the top 158 ranked 3-gap dipeptides. Therefore, the top 158 ranked 3-gap dipeptides were used to build the SVM model for antioxidant protein predictions. The model thus formed is called **AodPred**. In the jackknife test, the AodPred obtained an accuracy of 74.79 % for identifying antioxidant proteins.

3.2 Comparison with Existing Prediction Tools

It is necessary to compare the proposed methods with other existing methods. Therefore, we compared the predictive results of **AodPred** with that of the existing methods. Recently, a computational model to identify antioxidant proteins based on star graph topological indices was proposed [5]. However, sequences in their dataset share high-sequence similarities and some sequences in their dataset even with 100 % sequences identity. To overcome this shortage, we constructed a non-redundant dataset and

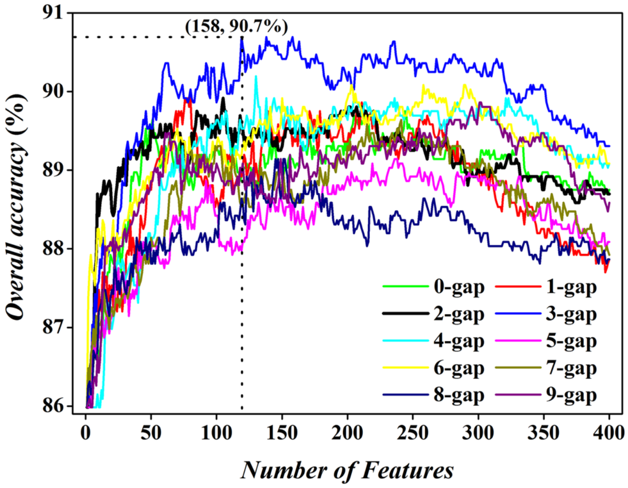


Fig. 1 Plot to show the IFS curves for different g-gap dipeptides ($g = 0, 1, 2, \ldots, 9$), where the abscissa and ordinate axis denote the number of features and the overall accuracy, respectively. As shown in the figure, the value of overall accuracy reached its peak (90.7 %) when the top 158 ranked 3-gap dipeptides are taken into account

proposed a naive Bayes model based on optimal dipeptides and obtained an accuracy of 66.88 % in jackknife test [11]. In contrast, the current predictor **AodPred** obtained an accuracy of 74.79 % for identifying antioxidant proteins in the jackknife test, which is higher than that of existing methods [5, 11].

To further verify the power of **AodPred**, we also compared its performance with other classifiers such as Bayes Net, Logistic, RBFNetwork, J48, and Random forest. Bayes Net, Logistic, RBFNetwork, J48, and Random forest models were tested on the benchmark dataset and implemented in WEKA [24]. The jackknife test results of these classifiers based on the optimal features are reported in Table 1. As indicated in Table 1, although the other four classifiers yielded higher predictive accuracies than **AodPred**, their sensitivities are all much lower than that of **AodPred**. These results indicate that the **AodPred** proposed in this paper is quite promising and holds a potential to become a useful tool in identifying antioxidant protein, or at least can play a complementary role to the existing method in this area.

4 Web Server

For the convenience of experimental scientists, based on the model proposed in the present work, a free web server called **AodPred** was provided. A step-by-step guide on how to use the web server was given below:

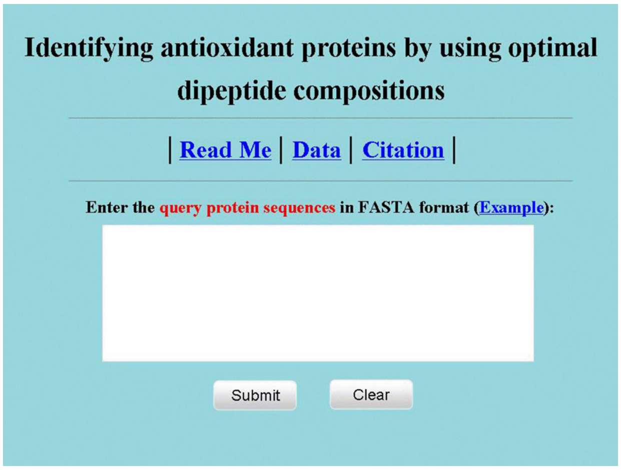
Step 1. Open the web server at http://lin.uestc.edu.cn/server/AntioxiPred and you will see the top page of **AodPred** on your computer screen, as shown in Fig. 2. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

Step 2. Either type or copy/paste the query protein sequences into the input box at the center of Fig. 2. The input sequence should be in the FASTA format. 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 instance, if using the two example sequences as an input and clicking the *Submit* button, you will see the following shown on the screen of your computer: The outcome for the first query example is "antioxidant" with a probability of 0.75; the outcome for the second query sample is "non-antioxidant" with a probability of 0.98.

Table 1 Comparison of AodPred with other methods by using optimal features

Classifier	*Sn* (%)	*Sp* (%)	*Acc* (%)
Bayes net	41.27	90.63	83.16
Logistic	35.67	89.32	80.22
J48 tree	29.45	89.56	80.32
Random forest	28.09	93.12	80.34
AodPred	75.09	74.48	74.79


Identifying antioxidant proteins by using optimal dipeptide compositions

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

Fig. 2 Semi-screenshot to show the top page of the AodPred. Its Web site address is at http://lin.uestc.edu.cn/server/AntioxiPred

Step 4. Click on the *Data* button to download the benchmark datasets used to train and test the AodPred predictor.

Caveats. The input query sequences must be formed by the single-letter codes of the 20 native amino acids; any other characters such as "B," "X," and "Z" are invalid and should not be part of the query sequence.

5 Conclusions

The role of antioxidant proteins in neutralizing free radicals and preventing them from causing damage or death to cells is well known. Unfortunately, the number of molecules with antioxidant properties in nature is quite low. Therefore, it is highly desirable to develop computational methods for identifying antioxidant proteins, so as to help speed up researches on antioxidant proteins.

By encoding sequences using the optimal *g*-gap dipeptide composition, a SVM-based predictor was developed to identify antioxidant proteins. The new predictor is promising as reflected by the high success rates obtained by the rigorous jackknife tests. It is instructive to point out that the accuracy can be further improved with future accumulation of knowledge regarding antioxidant proteins and antioxidant protein collections in the benchmark dataset.

Since publicly accessible web servers represent the direction for developing practically more useful predictor, a user-friendly web server for **AodPred** has been established at http://lin.uestc.edu.cn/server/AntioxiPred, by which users can easily obtain their desired results.

It is anticipated that **AodPred** may become a useful tool for identifying antioxidant proteins, or, at the very least, it can play a complementary role to the existing methods in this area.

Acknowledgments This work was supported by National Nature Scientific Foundation of China (Nos. 61100092 and 61202256), Nature Scientific Foundation of Hebei Province (No. C2013209105), and Foundation of Science and Technology Department of Hebei Province (No. 132777133).

Compliance with ethical standards

Conflict of interest The authors have declared that no competing interests exist.

References

1. Lobo V, Patil A, Phatak A, Chandra N (2010) Free radicals, antioxidants and functional foods: impact on human health. Pharmacogn Rev 4:118–126
2. Barbusinski K (2009) Fenton reaction-controversy concerning the chemistry. Ecol Chem Eng S 16:347–358
3. Shah AM, Channon KM (2004) Free radicals and redox signalling in cardiovascular disease. Heart 90:486–487
4. Pham-Huy LA, He H, Pham-Huy C (2008) Free radicals, antioxidants in disease and health. Int J Biomed Sci 4:89–96
5. Fernandez-Blanco E, Aguiar-Pulido V, Munteanu CR, Dorado J (2013) Random forest classification based on star graph topological indices for antioxidant proteins. J Theor Biol 317:331–337
6. Dreher D, Junod AF (1996) Role of oxygen free radicals in cancer development. Eur J Cancer 32A:30–38
7. Maxwell SR (2000) Coronary artery disease-free radical damage, antioxidant protection and the role of homocysteine. Basic Res Cardiol 95:65–71
8. Bailey DM, Evans KA, James PE, McEneny J, Young IS, Fall L, Gutowski M, Kewley E, McCord JM, Moller K, Ainslie PN (2009) Altered free radical metabolism in acute mountain sickness: implications for dynamic cerebral autoregulation and blood-brain barrier function. J Physiol 587:73–85
9. Mecocci P, Polidori MC, Troiano L, Cherubini A, Cecchetti R, Pini G, Straatman M, Monti D, Stahl W, Sies H, Franceschi C, Senin U (2000) Plasma antioxidants and longevity: a study on healthy centenarians. Free Radic Biol Med 28:1243–1248
10. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247
11. Feng PM, Lin H, Chen W (2013) Identification of antioxidants from sequence information using naive Bayes. Comput Math Methods Med 2013:567529
12. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152
13. Ding C, Yuan LF, Guo SH, Lin H, Chen W (2012) Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. J Proteom 77:321–328
14. Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41:e68
15. Chen W, Feng PM, Lin H (2012) Prediction of replication origins by calculating DNA structural properties. FEBS Lett 586:934–938
16. Liu B, Wang X, Lin L, Dong Q, Wang X (2009) Exploiting three kinds of interface propensities to identify protein binding sites. Comput Biol Chem 33:303–311
17. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43:246–255
18. Ding H, Guo SH, Deng EZ, Yuan LF, Guo FB, Huang J, Rao NN, Chen W, Lin H (2013) Prediction of Golgi-resident protein types by using feature selection technique. Chemometr Intell Lab Syst 124:9–13
19. Lin H, Chen W, Ding H (2013) AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PLoS ONE 8:e75726
20. Feng PM, Ding H, Chen W, Lin H (2013) Naive Bayes classifier with feature selection to identify phage virion proteins. Comput Math Methods Med 2013:530696
21. Feng PM, Chen W, Lin H, Chou KC (2013) iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem 442:118–125
22. Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30:1522–1529
23. Lin H, Ding H (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 269:64–69
24. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using weka. Bioinformatics 20:2479–2481