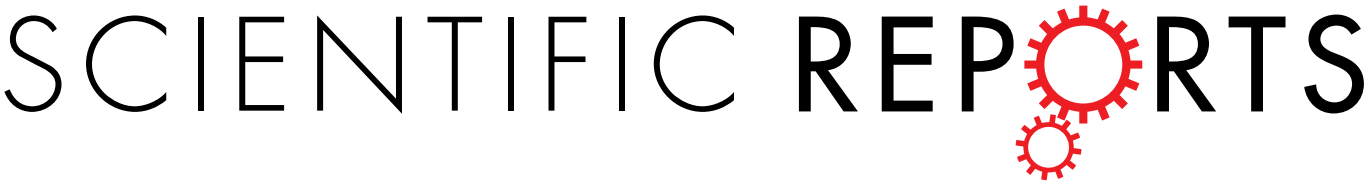

OPEN

Identification of apolipoprotein using feature selection technique

Hua Tang[1], Ping Zou[1], Chunmei Zhang[1], Rong Chen[1], Wei Chen[2,3] & Hao Lin[2]

Received: 25 April 2016
Accepted: 01 July 2016
Published: 22 July 2016

Apolipoprotein is a kind of protein which can transport the lipids through the lymphatic and circulatory systems. The abnormal expression level of apolipoprotein always causes angiocardiopathy. Thus, correct recognition of apolipoprotein from proteomic data is very crucial to the comprehension of cardiovascular system and drug design. This study is to develop a computational model to predict apolipoproteins. In the model, the apolipoproteins and non-apolipoproteins were collected to form benchmark dataset. On the basis of the dataset, we extracted the *g*-gap dipeptide composition information from residue sequences to formulate protein samples. To exclude redundant information or noise, the analysis of various (ANOVA)-based feature selection technique was proposed to find out the best feature subset. The support vector machine (SVM) was selected as discrimination algorithm. Results show that 96.2% of sensitivity and 99.3% of specificity were achieved in five-fold cross-validation. These findings open new perspectives to improve apolipoproteins prediction by considering the specific dipeptides. We expect that these findings will help to improve drug development in anti-angiocardiopathy disease.

Apolipoproteins (Apo) are proteins that bind lipids to form lipoproteins, whose main function is to transport the lipids through the lymphatic and circulatory systems[1]. There are two major types of apolipoproteins. One type has mostly beta-sheet structure and associates with lipid droplets irreversibly. They can form low-density lipoprotein. The other type consists of alpha-helices and associates with lipid droplets reversibly. Most proteins of the second type can form high-density lipoprotein particles. Several studies have demonstrated that the apolipoproteins have important functions in cardiovascular system, digestive system, and etc. For example, ApoE mediates the transport and uptake of cholesterol and lipid by way of its high affinity interaction with different cellular receptors[2]. ApoA1 is thought to act primarily in intestinal lipid absorption[1]. Thus, accurate identification of the apolipoprotein is very crucial to the comprehension of cardiovascular and digestive system as well as drug design.

The available of huge amounts of proteins generated in postgenomic age provides us an opportunity to design computational methods to timely and precisely predict protein functions. In fact, in the past two decades, a great deal of works have been focused on protein structure and function prediction by using machine learning methods, such as support vector machine (SVM)[3–5], random forest (RF)[6], Increment of diversity (ID)[7], the Mahalanobis discriminant[8], ensemble classifiers[9,10], feature selection techniques[11], etc. Amino acid composition[12] was always selected as the key feature to formulate protein sequence. However, the residue-order information was completely lost. To improve the description about protein samples, the *n*-peptide information[13], N-terminal amino acid sequence[14], secondary structure involved features[15,16] were proposed. Evolution information generated by PSI-BLAST was also considered by several studies[17–20]. To incorporate the physicochemical properties with residue-order information, the pseudo amino acid composition (PseAAC) was developed to efficiently improve prediction quantity[21–23].

Based on these methods, many protein prediction issues such as protein subcellular localization[22], protein structural classes[24,25], enzyme classification[26,27] have been studied. However, to the best of our knowledge, no computational method was proposed to predict apolipoproteins. The appearance of a great number of protein data provides us an opportunity to statistically study and predict apolipoproteins. Thus, this study aims to develop a computational method to identify apolipoproteins. High quality dataset was constructed to train and test the proposed method. Informative features were optimized by using feature selection technique and then inputted

[1]Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China. [2]Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China. [3]Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009 China. Correspondence and requests for materials should be addressed to H.T. (email: tanghua771211@aliyun.com) or W.C. (email: greatchen@heuu.edu.cn) or H.L. (email: hlin@uestc.edu.cn)

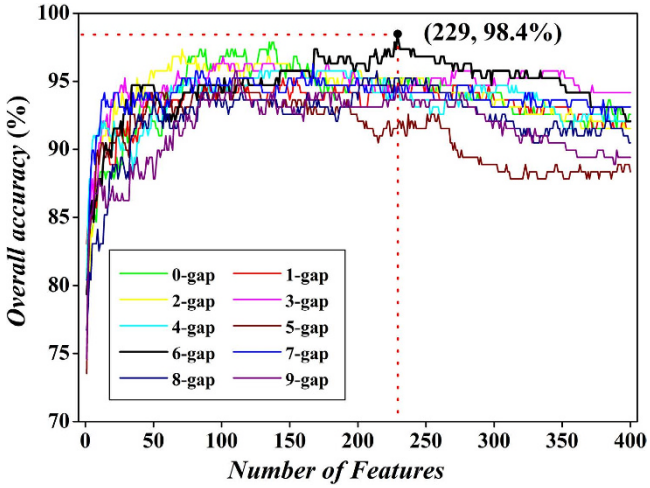

Figure 1. A plot showing the IFS procedure for discriminating apolipoproteins from non-apolipoproteins. When the top 229 6-gap dipeptides were used to perform prediction, the overall success rate reaches an IFS peak of 98.4% in five-fold cross-validation.

into SVM for performing prediction. Finally, based on the optimal model, we established a webserver called **ApoliPred** which can be freely used by all scholars. The following section will introduce these steps in detailed.

Results

Prediction Performance. The predictive performances of *g*-gap dipeptide composition were investigated by using SVM with five-fold cross-validation test. To investigate whether a specific type of *g*-gap dipeptides is mostly contributable for apolipoproteins identification, we varied the interval residue parameter *g* from 0 to 9. After checking the overall accuracies (*OA*s) obtained by different *g*-gap dipeptide composition, we found that the *OA* of 3-gap dipeptide composition is 94.2% in five-fold cross-validation which is higher than that of other *g*-gap dipeptide compositions.

The above results implied that the information of apolipoproteins mainly stores in the correlation of two residues with 3 residues interval. However, the noises or redundant information maybe result in the poor predictive capabilities of other *g*-gap dipeptide compositions. Thus, feature selection method *F*-score as defined in Eq. 3 was used to exclude noise and redundant information. For an arbitrary *g*-gap dipeptide composition, it has 400 features. On the basis of Eq. 3, a total of 400 *F*-scores were calculated for the 400 features. Subsequently, the 400 features were ranked according to their *F*-scores. The incremental feature selection (IFS) is used to determine the optimal number of features according to the following steps. Firstly, the feature with the maximum *F*-score was selected as the input of SVM. The *OA* was calculated to evaluate the performance of this feature. Secondly, the feature with the second maximum *F*-score was combined with the first feature to form a new feature subset. The *OA* was still used to estimate the performance of the new feature subset by using SVM. This process was repeated until 400 *OA*s were calculated. The best feature subset is the subset that can produce the maximum *OA*. By setting dimension of feature subset (the number of features) as abscissa and the *OA* as ordinate, we plotted 10 curves as shown in Fig. 1. From the figure, we noticed that the maximum *OA* of 98.4% can be achieved by 229 6-gap dipeptides which are regarded as the optimal feature subset. Thus, the final model was constructed by these features. The sensitivity (*Sn*) and specificity (*Sp*) are 96.2% and 99.3%, respectively. We also investigated the *OA* of optimal feature subset by using jackknife cross-validation. Results showed that *OA* is 97.35%, which also demonstrates that the optimal model is powerful.

By comparing the results of original and optimal *g*-gap dipeptide composition, we may draw a conclusion that feature selection not only improves the model's performance, but also find the potential correlation between two residues. In fact, the above results show that, in apolipoprotein prediction, the correlation information mainly exists in 6-gap dipeptides. This demonstrates that the feature selection is a very valid way to discover the intrinsic characteristics of apolipoproteins.

A heat map analysis. To provide an overall and intuitive view for understanding the contribution of features, the preference of 6-gap dipeptide composition in both positive and negative datasets was investigated by using Eq. 4. If $F^0(x) > 0$, the x-th 6-gap dipeptide prefers apolipoprotein, otherwise it prefers non-apolipoprotein. Based on Eq. 4, a heat map was drawn in Fig. 2. The first and second residues of the 6-gap dipeptides were, respectively, listed in row and column of the heat map. Thus, each element in the heat map represents one of the 400 6-gap dipeptides. The color of the element was drawn according to its *F*-score. The features in red and blue boxes are positively and negatively correlated with apolipoproteins, respectively. It is obvious that the redder the element is, the more highly relevant with apolipoproteins it is, and vice versa. From the figure, we found that the residues Leu (L), Phe (F), Ile (I) and Trp (W) as well as their 6-gap correlations are abundant in apolipoproteins compare to non-apolipoproteins, whereas the residues Cys (C), Glu (E), K and P (blue) as well as their 6-gap correlations exhibit the opposite behavior. Thus, we further used 20 6-gap dipeptides with maximum *F*-scores to discriminate apolipoproteins from non-apolipoproteins. It shows that the overall accuracy reaches 91.0% in five-fold

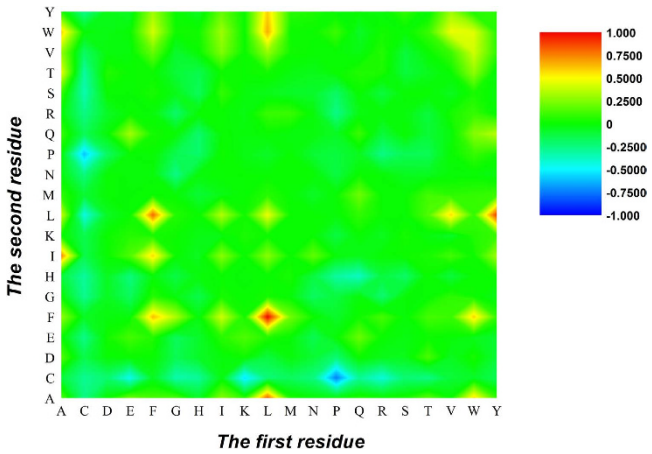

Figure 2. A heat map or chromaticity diagram for the *F*-scores of the 400 6-gap dipeptides. The blue boxes indicate that the features are enriched in apolipoproteins, while the red boxes indicate that the features are enriched in non-apolipoproteins.

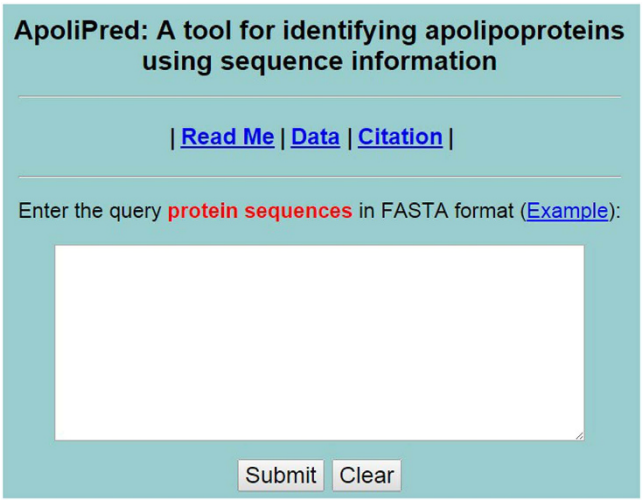

Figure 3. A semi-screenshot to show the top page of the ApoliPred webserver. Its website address is http://lin.uestc.edu.cn/server/apoliPred.

cross-validation. These results indicate that Leu (L), Phe (F), Ile (I), Trp (W), Cys (C), Glu (E), Lys (K) and Pro (P) (blue) are key features for apolipoprotein identification.

Web-server guide. To provide the convenience for vast majority of scholars, a user-friendly web-server called **ApoliPred** was established. The server is freely access at http://lin.uestc.edu.cn/server/ApoliPred. A step-by-step guide on how to use the webserver is given below.

Users may open the web server at http://lin.uestc.edu.cn/server/ApoliPred and the top page of **ApoliPred** will appear on user's computer screen, as shown in Fig. 3. The Read Me button provides a brief introduction about the predictor and the caveat when using it. The Data button provides the benchmark dataset used in this study to train and test the **ApoliPred** predictor. The Citation button given the relevant papers documenting the detailed development and algorithm of **ApoliPred**.

Users may type or copy/paste the query residue sequences into the input box at the center of Fig. 3. The input sequence must be in the FASTA format. By clicking on Example button, users can look at the example sequences in FASTA format. By clicking on the Submit button, the predicted results will appear.

Discussion

The aim of this work is to build a predictive model to identify apolipoproteins. In fact, the programs such as BLAST and FASTA have been widely used in genomic and proteomic analysis or prediction based on similarity search. However, these programs are helplessness as facing low-similar sequences. Especially, more and more orphan genes were found. Few of them can be functional annotated in GenBank by similarity search. With the

appearance of more and more orphan genes or low-similar proteins, it is urgent to develop a statistical predictive model. Thus, the study is very meaningful and important.

All results and models are derived from a strict and objective benchmark dataset. All data in benchmark dataset have been confirmed by biochemical experiments. Thus, the information extracted from such dataset is precise and reliable. Moreover, many papers have reported that the proposed model will be overestimated or bias if they are trained and tested by high similar sequences[22,28]. In this study, we have removed the redundant sequences by setting sequence identity to 40%. Thus, our models are reliable and harmonious.

Correlation of nucleotides or residues is the key carrier of genetic information. Therefore, we used *g*-gap dipeptide composition as features in prediction. However, the performances of models are far from satisfactory based on such fundamental information. To improve the accuracies and find out the real correlations hiding in protein sequences, a feature selection technique was applied to select optimal features. Results demonstrate that the technique can pick out informative features, dramatically improve the predictive performance and enhance the generalization abilities of the proposed models. Based on the correlation information and feature selection technique, our models did achieve promising results for recognition of apolipoproteins.

Conclusion

In this study, a new tool, called **ApoliPred**, was established for accurate identification of potential novel apolipoproteins. In **ApoliPred**, a high-quality benchmark dataset was constructed by setting a series of standards, which can guarantee the reliable of the tool. Thus, the dataset has the potential to become a standard dataset in the development of computational methods in theoretical study of apolipoproteins. Moreover, a feature selection technique has been successfully applied to improve the performance. The special dipeptide distribution were discovered in apolipoproteins. Our results indicated that the proposed model can provide a high discriminative accuracy. We expect that these findings will help to improve drug development in anti-angiocardiopathy disease. The method proposed in this work can also be used in other field of bioinformatics and computational biology.

Materials and Method

Benchmark dataset. A strict and objective benchmark dataset can guarantee the reliable of the prediction model. Thus, all apolipoproteins were downloaded from the Universal Protein Resource (UniProt)[29]. For the purpose of obtaining a high quality data, the following steps were performed.

(1) Choose the apolipoproteins which have been annotated in Swiss-Prot.
(2) Exclude the proteins whose sequences contain illegal characters such as "B", "J", "O", "U", "X" and "Z".
(3) Select the proteins whose function (namely the Gene Ontology (GO)) have been annotated.
(4) Remove homologous sequences by setting the cutoff threshold of CD-HIT to 0.4.

After following the above processes, we obtained a total of 53 apolipoproteins and 136 non-apolipoproteins which can be freely downloaded from http://lin.uestc.edu.cn/server/ApoliPred.

Protein Feature Extraction. It is widely accepted that the functions of proteins correlate with their three-dimensional structures. Thus, we extracted features from the primary sequence of apolipoproteins and non-apolipoproteins. Dipeptide composition has been widely applied in protein classification and has achieved encouraging results[30,31]. However, it can only reflect the short-range correlation. In most cases, non-adjoining residues in primary sequence might be proximate in three-dimensional space. For example, in alpha helix and beta sheet, the hydrogen bonds are in charge of the connection of two non-adjoining residues. Thus, we used the *g*-gap dipeptide composition to describe the correlation of residues in protein primary sequence. Thus, a given protein **P** can be formulated by a 400-Dimension vector and defined as:

$$\mathbf{P} = [f_1^g, f_2^g, \cdots, f_\varepsilon^g, \cdots, f_{400}^g]^{\mathbf{T}} \quad (1)$$

where **T** is the transposing operator, the f_ε^g denotes the frequency of the ε-th type ($\varepsilon = 1, 2, \ldots, 400$) of the *g*-gap dipeptide in protein **P** and can be calculated by:

$$f_\varepsilon^g = n_\varepsilon^g/(L - g - 1) \quad (2)$$

where the n_ε^g is the occurrence number of the ε-th type ($\varepsilon = 1, 2, \ldots, 400$) of *g*-gap dipeptide in the protein **P**. The *L* in the denominator of Eq. 2 is the length of the protein **P**. If the parameter *g* is set to 0, the formulation degrades into the adjoining dipeptide composition which describes the correlation of two proximate residues. Thus, the *g*-gap dipeptide composition represents the direct correlation between two residues with *g* residues interval.

Support vector machine (SVM). Based on the statistical learning theory, Vapnik and his colleagues have developed a powerful and popular machine learning method called SVM. It is a supervised learning method which projects samples with low-dimension feature into a high-dimension Hilbert space and constructs a hyperplane to perform classification. Due to its good capability for non-linear classification and small sample classification, SVM has been widely applied in protein structure and function classification as well as DNA motif prediction[3,13,21,22,26,28,32,33]. Thus, we also used the SVM to perform the classification. A free software package LibSVM (version 3.2) was used to implement the SVM. The radial basis function (RBF) was chosen as the kernel function because it is more suitable for nonlinear classification than other kernel functions. To obtain the best performance, the grid search approach was applied to optimize regularization parameter *C* and the kernel width parameter γ by using 5-fold cross-validation.

Feature selection technique. Generally, in statistical learning problem, not every feature has a positive contribution to the classification, especially for high dimension data. Some features are noise or redundant information which will reduce the predictive performance of classification models. Thus, it is very important to develop a method to evaluate the contribution of every feature to the classification. Based on this consideration, we proposed an *F*-score to describe the contribution of each feature.

Based on the hypothesis that if the sample variance of a feature between groups is larger than sample variance within groups[3], the feature is suitable for classification, we defined the *F*-score of a feature *x* as follow.

$$F(x) = \frac{m_p(\overline{x_p} - \overline{x})^2 + m_n(\overline{x_n} - \overline{x})^2}{\left[\sum_{i=1}^{m_p}(x_i - \overline{x_p})^2 + \sum_{i=1}^{m_n}(x_i - \overline{x_n})^2\right]/(m_p + m_n - 2)} \tag{3}$$

where $\overline{x}$, $\overline{x_p}$ and $\overline{x_n}$ are the means of feature *x* in all samples, positive samples and negative samples, respectively. m_p and m_n are the number of samples in positive dataset or negative dataset, respectively. Thus, the numerator and denominator in Eq. 3 denote the variances between groups and within groups, respectively. It is obvious that the larger the F(x) is, the better capability the feature *x* has.

We used the following normalized function to scale the $F(x)$ of the feature *x* as follow

$$F^0(x) = \frac{F(x) - F_{min}}{F_{max} - F_{min}} \times \mathbf{sgn}\left(\overline{f^g_{\varepsilon,p}} - \overline{f^g_{\varepsilon,n}}\right) \tag{4}$$

where $F_{\min}$ and $F_{\max}$ are the minimum and maximum F values of the 400 *g*-gap dipeptides. The $\overline{f^g_{\varepsilon,p}}$ and $\overline{f^g_{\varepsilon,n}}$ are the average frequencies of the ε-th *g*-gap dipeptide in positive sample dataset and negative sample set, respectively; **sgn** is the sign function. Thus, the upper limit and lower limit of $F^0(x)$ are 1 and −1, respectively.

Performance evaluation. In statistical prediction, several test methods such as *n*-fold cross-validation test, jackknife cross-validation test, independent data test can be used to estimate the predictive performance of proposed method[34]. Jackknife cross-validation test is usually more suitable for small sample problem and always yields a unique results for a given benchmark dataset[28,35,36], however, it is time-consuming. Thus, we used five-fold cross-validation in this study to evaluate the performance of our model.

To quantitatively evaluate the performance of models, the following three indexes called Sensitivity (*Sn*), Specificity (*Sp*) and Overall Accuracy (*OA*) were used and can be defined as:

$$Sn = \frac{n^+}{N^+} \tag{5}$$

$$Sp = \frac{n^-}{N^-} \tag{6}$$

$$OA = \frac{n^+ + n^-}{N^+ + N^-} \tag{7}$$

where n^+ and n^- are the number of the correctly identified positive samples (i.e. apolipoproteins) and the number of the correctly identified negative samples (i.e. non-apolipoproteins), respectively; N^+ and N^- are the number of the positive samples and the number of negative samples in the benchmark dataset, respectively.

References

1. Saito, H., Lund-Katz, S. & Phillips, M. C. Contributions of domain structure and lipid interaction to the functionality of exchangeable human apolipoproteins. *Progress in lipid research* **43,** 350–380, doi: 10.1016/j.plipres.2004.05.002 (2004).
2. Holtzman, D. M., Herz, J. & Bu, G. Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer disease. *Cold Spring Harbor perspectives in medicine* **2,** a006312, doi: 10.1101/cshperspect.a006312 (2012).
3. Lin, H. *et al.* Predicting cancerlectins by the optimal g-gap dipeptides. *Scientific reports* **5,** 16964, doi: 10.1038/srep16964 (2015).
4. Chen, W. & Lin, H. Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information. *Biochemical and biophysical research communications* **401,** 382–384, doi: 10.1016/j.bbrc.2010.09.061 (2010).
5. Chen, W. & Lin, H. Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. *Computers in biology and medicine* **42,** 504–507, doi: 10.1016/j.compbiomed.2012.01.003 (2012).
6. Li, K. *et al.* Prediction and identification of the effectors of heterotrimeric G proteins in rice (Oryza sativa L.). *Briefings in bioinformatics*, doi: 10.1093/bib/bbw021 (2016).
7. Zuo, Y. C. *et al.* Discrimination of membrane transporter protein types using K-nearest neighbor method derived from the similarity distance of total diversity measure. *Molecular bioSystems* **11,** 950–957, doi: 10.1039/c4mb00681j (2015).
8. Liu, B., Wang, X., Lin, L., Dong, Q. & Wang, X. A Discriminative Method for Protein Remote Homology Detection and Fold Recognition Combining Top-n-grams and Latent Semantic Analysis. *BMC Bioinformatics* **9,** 510 (2008).
9. Lin, C. *et al.* LibD3C: Ensemble Classifiers with a Clustering and Dynamic Selection Strategy. *Neurocomputing* **123,** 424–435 (2014).
10. Song, L. *et al.* nDNA-prot: Identification of DNA-binding Proteins Based on Unbalanced Classification. *BMC Bioinformatics* **15,** 298 (2014).
11. Zou, Q., Zeng, J., Cao, L. & Ji, R. A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. *Neurocomputing* **173,** 346–354 (2016).
12. Sharma, A., Gupta, P., Kumar, R. & Bhardwaj, A. dPABBs: A Novel in silico Approach for Predicting and Designing Anti-biofilm Peptides. *Scientific reports* **6,** 21839, doi: 10.1038/srep21839 (2016).
13. Lin, H., Chen, W., Yuan, L. F., Li, Z. Q. & Ding, H. Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta biotheoretica* **61,** 259–268, doi: 10.1007/s10441-013-9181-9 (2013).

14. Chang, E. *et al.* N-Terminal Amino Acid Sequence Determination of Proteins by N-Terminal Dimethyl Labeling: Pitfalls and Advantages When Compared with Edman Degradation Sequence Analysis. *Journal of biomolecular techniques: JBT*, doi: 10.7171/jbt.16-2702-002 (2016).
15. Wei, L., Liao, M., Gao, X. & Zou, Q. An Improved Protein Structural Prediction Method by Incorporating Both Sequence and Structure Information. *IEEE Transactions on Nanobioscience* **14,** 339–349 (2015).
16. Wei, L., Liao, M., Gao, X. & Zou, Q. Enhanced Protein Fold Prediction Method through a Novel Feature Extraction Technique. *IEEE Transactions on Nanobioscience* **14,** 649–659 (2015).
17. Bui, V. M. *et al.* SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. *BMC genomics* **17** Suppl 1, 9, doi: 10.1186/s12864-015-2299-1 (2016).
18. Huang, C. H. *et al.* UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines. *BMC systems biology* **10** Suppl 1, 6, doi: 10.1186/s12918-015-0246-z (2016).
19. Shen, H. S. *et al.* HIV coreceptor tropism determination and mutational pattern identification. *Scientific reports* **6,** 21280, doi: 10.1038/srep21280 (2016).
20. Zou, Q., Hu, Q., Guo, M. & Wang, G. HAlign: Fast Multiple Similar DNA/RNA Sequence Alignment Based on the Centre Star Strategy. *Bioinformatics* **31,** 2475–2481, doi: 10.1093/bioinformatics/btv177 (2015).
21. Tang, H., Chen, W. & Lin, H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Molecular bioSystems* **12,** 1269–1275, doi: 10.1039/c5mb00883b (2016).
22. Zhu, P. P. *et al.* Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. *Molecular bioSystems* **11,** 558–563, doi: 10.1039/c4mb00645c (2015).
23. Liu, B. *et al.* Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research* **43,** W65–W71 (2015).
24. Ding, H. *et al.* Prediction of protein structural classes based on feature selection technique. *Interdisciplinary sciences, computational life sciences* **6,** 235–240, doi: 10.1007/s12539-013-0205-6 (2014).
25. Li, D., Ju, Y. & Zou, Q. Protein Folds Prediction with Hierarchical Structured SVM. *Current Proteomics* **13,** 79–85 (2016).
26. Lin, H., Chen, W. & Ding, H. AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. *PloS one* **8,** e75726, doi: 10.1371/journal.pone.0075726 (2013).
27. Cheng, X.-Y. *et al.* A global characterization and identification of multifunctional enzymes. *PloS one* **7,** e38979 (2012).
28. Lin, H. *et al.* The prediction of protein structural class using averaged chemical shifts. *Journal of biomolecular structure & dynamics* **29,** 643–649, doi: 10.1080/07391102.2011.672628 (2012).
29. Breuza, L. *et al.* The UniProtKB guide to the human proteome. *Database: the journal of biological databases and curation* **2016,** doi: 10.1093/database/bav120 (2016).
30. Ahmad, K., Waris, M. & Hayat, M. Prediction of Protein Submitochondrial Locations by Incorporating Dipeptide Composition into Chou's General Pseudo Amino Acid Composition. *The Journal of membrane biology*, doi: 10.1007/s00232-015-9868-8 (2016).
31. Liou, Y. F. *et al.* SCMMTP: identifying and characterizing membrane transport proteins using propensity scores of dipeptides. *BMC genomics* **16** Suppl 12, S6, doi: 10.1186/1471-2164-16-S12-S6 (2015).
32. Liu, B. *et al.* Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **30,** 472–479 (2014).
33. Wang, R., Xu, Y. & Liu, B. Recombination spot identification Based on gapped k-mers. *Scientific reports* **6,** 23934 (2016).
34. Chen, J., Wang, X. & Liu, B. iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Scientific reports* **6,** 19062 (2016).
35. Chen, W., Feng, P. & Lin, H. Prediction of ketoacyl synthase family using reduced amino acid alphabets. *Journal of industrial microbiology & biotechnology* **39,** 579–584, doi: 10.1007/s10295-011-1047-z (2012).
36. Feng, P., Lin, H., Chen, W. & Zuo, Y. Predicting the types of J-proteins using clustered amino acids. *BioMed research international* **2014,** 935719, doi: 10.1155/2014/935719 (2014).

Acknowledgements

This work was supported by the Applied Basic Research Program of Sichuan Province (No. 2015JY0100 and LZ-LY-45), the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122), the Nature Scientific Foundation of Hebei Province (No. C2013209105), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015J144; ZYGX2015Z006) and Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

Author Contributions

H.T. and H.L. conceived and designed the experiments; H.T., P.Z., C.Z., R.C. and H.L. analyzed the data; W.C. and H.L. implemented SVM and created the web-server; H.T., W.C. and H.L. performed the analysis and wrote the paper. All authors read and approved the final manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Tang, H. *et al.* Identification of apolipoprotein using feature selection technique. *Sci. Rep.* **6**, 30441; doi: 10.1038/srep30441 (2016).