Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 567529, 5 pages
http://dx.doi.org/10.1155/2013/567529

Research Article

Identification of Antioxidants from Sequence Information Using Naïve Bayes

Peng-Mian Feng,[1] Hao Lin,[2] and Wei Chen[3]

[1] *School of Public Health, Hebei United University, Tangshan 063000, China*
[2] *Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[3] *Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China*

Correspondence should be addressed to Hao Lin; hlin@uestc.edu.cn and Wei Chen; greatchen@heuu.edu.cn

Received 3 May 2013; Revised 20 July 2013; Accepted 22 July 2013

Academic Editor: Elisabeth Tillier

Antioxidant proteins are substances that protect cells from the damage caused by free radicals. Accurate identification of new antioxidant proteins is important in understanding their roles in delaying aging. Therefore, it is highly desirable to develop computational methods to identify antioxidant proteins. In this study, a Naïve Bayes-based method was proposed to predict antioxidant proteins using amino acid compositions and dipeptide compositions. In order to remove redundant information, a novel feature selection technique was employed to single out optimized features. In the jackknife test, the proposed method achieved an accuracy of 66.88% for the discrimination between antioxidant and nonantioxidant proteins, which is superior to that of other state-of-the-art classifiers. These results suggest that the proposed method could be an effective and promising high-throughput method for antioxidant protein identification.

1. Introduction

Oxidation is a chemical reaction that transfers electrons or hydrogen from a substance to an oxidizing agent. Oxidation reactions can produce free radicals. In turn, these radicals can start chain reactions. When the chain reaction occurs in a cell, it can cause damage or death to the cell. Moreover, oxidative stress is also the cause and the consequence of disease. Antioxidants are protein molecules that terminate these chain reactions by removing free radical intermediates and inhibit other oxidation reactions. They do this by being oxidized themselves, so antioxidants are often reducing agents such as thiols, ascorbic acid, or polyphenols [1].

Antioxidants are widely used in dietary supplements and have been investigated for the prevention of diseases such as cancer, coronary heart disease, and even altitude sickness. Plants and animals maintain complex systems of multiple types of antioxidants, such as glutathione, vitamin A, vitamin C, and vitamin E, as well as enzymes such as catalase, superoxide dismutase, and various peroxidases. Insufficient levels of antioxidants or inhibition of the antioxidant enzymes can cause oxidative stress and may damage or kill the cells.

As oxidative stress appears to be an important part of many human diseases, the use of antioxidants in pharmacology is intensively studied, particularly as treatments for stroke and neurodegenerative diseases. Recently, Fernandez-Blanco et al. reported a computational model to identify antioxidant proteins based on star graph topological indices [2]. However, by analyzing Fernandez-Blanco et al.'s dataset, we found that sequences in their dataset share high-sequence similarities; some sequences in their dataset even share 100% sequences identity. It has been demonstrated that the predictive accuracy is closely related to sequence identity [3, 4], and high-sequence similarity can surely lead to the overestimation of predictive performance. Therefore, their results are not credible. There is an urgent need to develop efficient computational tools for antioxidant proteins identification.

In the current study, we propose a Naïve Bayes-based computational model for predicting antioxidant proteins using amino acid compositions and dipeptide compositions. The correlation-based feature subset selection algorithm [5] was introduced to find the optimal feature set. By using the optimized features, the proposed model was evaluated in a benchmark dataset in the jackknife test.

According to some recent comprehensive reviews [6, 7] a series of recent publications [8–13], to establish a really useful statistical predictor, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) and establish a user-friendly webserver for the predictor that is accessible to the public. Below, let us describe how to deal with these steps one by one.

2. Materials and Methods

2.1. Dataset. Fernandez-Blanco et al. have constructed a dataset containing 324 proteins with antioxidant activity and 1657 proteins without [2]. However, sequences in their dataset share high-sequence identity. The predictive accuracy is closely related to the sequence identity [3, 4], and high-sequence similarity can surely lead to the overestimation of predictive performance.

In order to prepare a reliable benchmark dataset, we first extracted proteins with antioxidant activities from the UniProt [14] according to the following steps: (i) only proteins with the experimentally confirmed antioxidant activities were included; (ii) the proteins which are fragments of other proteins were dislodged; (iii) and proteins containing nonstandard letters, that is, "B", "X", or "Z", were excluded as their meanings are ambiguous. After following the aforementioned strict screening procedures, we obtained 686 proteins with antioxidant activity and obtained a new raw dataset by merging the 686 proteins into Fernandez-Blanco et al.'s dataset [2].

For balancing the number of samples and providing a significant statistics, sequences which have >60% sequence similarity were removed from the new raw dataset using CD-HIT program [15]. If the sequence identity cutoff is set to a stringent threshold of 25%, the results will be more objective and reliable. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the number of antioxidant proteins would be too low to have statistical significance. Finally, a benchmark dataset containing 254 antioxidant and 1567 nonantioxidant proteins was constructed and can be found in the online Supporting Information S1 available online at http://dx.doi.org/10.1155/2013/567529. For further estimating the performance of the method, we also collected 20 antioxidant proteins (supporting information S2) which are independent from the training set.

2.2. Feature Vector. One of the most important parts for identifying protein attributes is to generate a set of proper informative parameters to encode protein sequences. The amino acid composition and dipeptide composition are the most important and effective parameters which have been widely applied in the realm of protein prediction [10–13]. Hence, every protein sequence in the benchmark dataset was encoded in a discrete vector as follows:

$$\mathbf{F} = [f_1, f_2, \ldots, f_{420}]^{\mathbf{T}}, \tag{1}$$

where f_i are the normalized occurrence frequencies of the 20 amino acids ($i = 1, 2, \ldots, 20$) and the 400 dipeptides ($i = 21, 22, \ldots, 420$) in the protein sequence, respectively. $\mathbf{T}$ is the transposing operator.

2.3. Feature Selection. Inclusion of redundant and noisy features in the model building process would cause poor predictive performance and increased computation time. Feature selection is the process of removing irrelevant features and is extremely useful in reducing the dimensionality of the data and improving the predictive accuracy. To reduce the dimension of the feature space and improve the predictive accuracy, the filter method Correlation-based Feature Selection [5] combined with best-first search strategy was used in the process of feature selection in the current work.

The process starts with an empty set of features and generates all possible single-feature expansions. The subset with the highest accuracy is chosen and expanded in the same way by adding single features. If when expanding a subset the accuracy does not maximize, the search drops back to the next best unexpanded subset and continues from there until all features are added. The subset with the highest accuracy will be selected as the final optimized feature set [16].

2.4. Naïve Bayes. Naïve Bayes is an effective statistical classification algorithm [17] and has been successfully used in the realm of bioinformatics [18–20]. The theory of Naïve Bayes is to assume the attribute variables to be independent from each other given the outcome. This assumption greatly simplifies the calculation of conditional probabilities.

In the Naïve Bayes framework, a classification problem can be seen as the problem of finding the outcome with maximum probability given a set of observed variables. Given the protein example described by its feature vector $\mathbf{F} = (f_1, f_2, \ldots, f_n)$, we need to look for a class $\mathbf{C}$ that maximizes the likelihood $\mathbf{P}(\mathbf{F} \mid \mathbf{C}) = \mathbf{P}(f_1, f_2, \ldots, f_n \mid \mathbf{C})$. Since the current work is intended to classify antioxidant and nonantioxidant proteins, a binary class $\mathbf{C} \in (0, 1)$ was generated, where 1 denotes the sample that was predicted as an antioxidant protein and 0 denotes nonantioxidant protein. For the binary classification, the class for the protein sample could be determined by comparing two posteriors as follows:

$$\frac{P(C = 1 \mid F = f_1, f_2, \ldots, f_n)}{P(C = 0 \mid F = f_1, f_2, \ldots, f_n)} = \frac{P(C = 1) \prod_{i=1}^{n} P_i(f_i \mid C = 1)}{P(C = 0) \prod_{i=1}^{n} P_i(f_i \mid C = 0)}. \tag{2}$$

Taking the logarithm of (2), we have

$$\log \frac{P(C=1 \mid F=f_1, f_2, \ldots f_n)}{P(C=0 \mid F=f_1, f_2, \ldots f_n)} = \log \frac{P(C=1)}{P(C=0)} + \sum_{i=1}^{n} \log \frac{P_i(f_i \mid C=1)}{P_i(f_i \mid C=0)}. \quad (3)$$

Hence, the sample will be predicted as 1 (antioxidant protein) if

$$\log \frac{P(C=1 \mid F=f_1, f_2, \ldots, f_n)}{P(C=0 \mid F=f_1, f_2, \ldots, f_n)} \geq \theta, \quad (4)$$

and 0 (nonantioxidant protein) for otherwise. θ is the threshold determining the tradeoff between sensitivity and specificity and can be trained on the training dataset to maximize the prediction performance.

2.5. Performance Evaluation. The performance of the proposed model was evaluated using sensitivity, specificity (Garmer, Sperling, and Forsberg), and accuracy (Acc), which are expressed as follows:

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad \text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}. \quad (5)$$

TP, TN, FP, and FN represent the number of the correctly recognized antioxidant proteins, the number of the correctly recognized nonantioxidant proteins, the number of nonantioxidant proteins recognized as antioxidant proteins, and the number of antioxidant proteins recognized as nonantioxidant proteins, respectively.

As the performance of the current classifier depends on the threshold θ as given in (4), the receiver operating characteristic (ROC) curve was employed. Therefore, the quality of a classifier can be objectively evaluated by measuring the area under the receiver operating characteristic curve (auROC). The value of auROC score ranges from 0 to 1, with a score of 0.5 corresponding to a random guess and a score of 1.0 indicating a perfect separation.

3. Results and Discussion

Three cross-validation methods, namely, subsampling test, independent dataset test, and jackknife test, are often employed to evaluate the predictive capability of a predictor. Among the three methods, the jackknife test is deemed the most objective and rigorous one that can always yield a unique outcome as demonstrated by a penetrating analysis in a recent comprehensive review [21], and hence has been widely and increasingly adopted by investigators to examine the quality of various predictors (see, e.g., [8, 22–28]). Accordingly, the jackknife test was used to examine the performance of the model proposed in the current study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample and all the rule parameters are calculated without including the one being identified.

Table 1: Predictive performance of Naïve Bayes based on different features.

Feature dimensions	Sn (%)	Sp (%)	Acc (%)	auROC
420	75.59	52.65	55.85	0.680
44	72.04	66.05	66.88	0.855

3.1. Prediction of Antioxidant Proteins. We trained the Naïve Bayes classifier using Waikato Environment for Knowledge Analysis (WEKA) [29] on the benchmark dataset. As shown in Table 1, in the jackknife test, an auROC score of 0.68 and an accuracy of 55.85% with an average sensitivity of 75.59% and an average specificity of 52.65% were obtained for the classification of antioxidant and nonantioxidant proteins by using all the 420 features, that is, 20 amino acid compositions and 400 dipeptide compositions.

For saving computing time, cross-validation methods (fivefold or tenfold) are widely used for feature selection in computational proteomics [16, 30]. In order to identify prominent features that can distinguish between antioxidant and nonantioxidant proteins, feature selection method was also carried out to eliminate the redundant features using WEKA in a ten-fold cross-validation approach on the benchmark dataset. In the ten-fold cross-validation, the benchmark dataset is split into ten pieces, and cross validation is performed using each of these ten pieces as the testing set. Thus, the training process is performed ten times, each of which uses the data obtained by deleting the testing set from the whole dataset. We found that the proposed method achieved a maximum accuracy of 66.89% and auROC of 0.762 when the feature dimension reduced to 44 (i.e., C, G, FP, FW, LK, LS, IE, VL, VH, VC, VW, MS, PD, AP, AY, YQ, YE, YR, HE, HG, QA, KA, KH, DF, DK, DR, EF, EM, EY, ER, CP, CN, CG, WC, RT, RD, RW, SV, SD, GV, GY, GK, GC).

The jackknife test results of the Naïve Bayes classifier based on the 44 optimized features for identifying antioxidant proteins were listed in Table 1. As it can be seen from Table 1, the current method yielded a better auROC score of 0.855 and a predictive accuracy of 66.88% with an average sensitivity of 72.04% and an average specificity of 66.05% (Table 1). Both predictive accuracy and auROC are higher than those of the model based on the 420 features.

Moreover, for the purpose of evaluating the performance of the proposed method, we used the 20 experimentally-confirmed antioxidant proteins (in Supporting Information S2) to examine the method. As a result, 16 antioxidant proteins were correctly predicted by the proposed method; see Table 2. This result demonstrates the excellent performance of our model.

3.2. Comparison with Other Methods. In order to further testify its superiority, we compared the capability of the present model with that of other models based on different kinds of algorithms such as BayesNet, J48 tree, and Random

Table 2: Predictive results based on the independent dataset.

UniProt ID	Predictive result
Q148E0	Antioxidant
Q7RTV5	Antioxidant
Q9D1A0	Antioxidant
P80239	Antioxidant
P0AE08	Antioxidant
Q7BHK8	Nonantioxidant
P0A251	Antioxidant
P0A5N4	Antioxidant
Q8L5E0	Antioxidant
P06728	Antioxidant
Q03247	Antioxidant
P23529	Nonantioxidant
P30041	Antioxidant
O19097	Antioxidant
P23345	Antioxidant
P23346	Antioxidant
O65198	Nonantioxidant
P93407	Antioxidant
P11964	Nonantioxidant
P10792	Antioxidant

Table 3: Comparison of Naïve Bayes with other methods by using optimized features.

Classifier	Sn (%)	Sp (%)	Acc (%)	auROC
BayesNet	42.12	92.53	85.50	0.800
J48 tree	26.37	90.81	81.82	0.565
Random Forest	28.35	97.64	87.97	0.797
Naïve Bayes	72.04	66.05	66.88	0.855

forest. All the classifiers were compared on the benchmark dataset based on the optimized features (i.e., C, G, FP, FW, LK, LS, IE, VL, VH, VC, VW, MS, PD, AP, AY, YQ, YE, YR, HE, HG, QA, KA, KH, DF, DK, DR, EF, EM, EY, ER, CP, CN, CG, WC, RT, RD, RW, SV, SD, GV, GY, GK, GC). Their best predictive results from jackknife test were shown in Table 3.

Although the accuracies of BayesNet, J48 tree, and Random forest are higher than those of Naïve Bayes, their auROC scores and sensitivities are all much lower than those of Naïve Bayes. These results indicate that the proposed Naïve Bayes model can be effectively used to classify antioxidant and nonantioxidant proteins.

4. Conclusions

In this study, the Naïve Bayes classifier with feature selection method is presented to identity antioxidant proteins based on the primary sequence information. By using Correlation-based Feature Subset Selection algorithm, the feature dimensions were reduced to 44 prominent features that could remarkably improve the predictive accuracies. However, the detailed analyses of the selected features are required to provide more information about their roles in biological activity. It is expected that the presented model will provide novel insights into the research on antioxidants. Since user-friendly and publicly accessible webservers represent the future direction for developing practically more useful predictors [31], we shall make efforts in our future work to provide a webserver for the method presented in this paper.

Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (Nos. 61100092, 61202256) and the Fundamental Research Funds for the Central Universities (ZYGX2012J113).

References

[1] H. Sies, "Oxidative stress: oxidants and antioxidants," *Experimental Physiology*, vol. 82, no. 2, pp. 291–295, 1997.

[2] E. Fernandez-Blanco, V. Aguiar-Pulido, C. R. Munteanu, and J. Dorado, "Random Forest classification based on star graph topological indices for antioxidant proteins," *Journal of Theoretical Biolology*, vol. 317, pp. 331–337, 2013.

[3] R. Nair and B. Rost, "Sequence conserved for subcellular localization," *Protein Science*, vol. 11, no. 12, pp. 2836–2847, 2002.

[4] C.-S. Yu, Y.-C. Chen, C.-H. Lu, and J.-K. Hwang, "Prediction of protein subcellular localization," *Proteins*, vol. 64, no. 3, pp. 643–651, 2006.

[5] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning: Data Mining, Inference and Prediction*, Springer, Berlin, Germany, 2008.

[6] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review)," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[7] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013.

[8] W. Chen, P. M. Feng, H. Lin, K. C. Chou et al., "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, p. e68, 2013.

[9] W. Chen, H. Lin, P. M. Feng et al., "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS One*, vol. 7, no. 10, Article ID e47843, 2012.

[10] K.-C. Chou, Z.-C. Wu, and X. Xiao, "ILoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular BioSystems*, vol. 8, no. 2, pp. 629–641, 2012.

[11] W. Z. Lin, J. A. Fang, X. Xiao, and K.-C. Chou, "iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins," *Molecular BioSystems*, vol. 9, no. 9, pp. 634–644, 2013.

[12] X. Xiao, P. Wang, W. Z. Lin, J.-H. Jiaa, and K.-C. Choub, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

[13] Y. Xu, J. Ding, L. Y. Wu, and K.-C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS One*, vol. 8, no. 2, Article ID e55844, 2013.

[14] UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, pp. D71–D75, 2012.

[15] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[16] W. Chen and H. Lin, "Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 504–507, 2012.

[17] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 1997.

[18] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido, and S. Ahmad, "A naïve Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins," *Bioinformatics*, vol. 19, no. 2, pp. 234–240, 2003.

[19] M. Yousef, S. Jung, A. V. Kossenkov, L. C. Showe, and M. K. Showe, "Naïve Bayes for microRNA target predictions—machine learning for microRNA targets," *Bioinformatics*, vol. 23, no. 22, pp. 2987–2992, 2007.

[20] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites," *Bioinformatics*, vol. 26, no. 15, pp. 1841–1848, 2010.

[21] F. Sambo, E. Trifoglio, B. di Camillo, G. M. Toffolo, and C. Cobelli, "Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data," *BMC Bioinformatics*, vol. 13, p. S2, 2012.

[22] K. C. Chou and H. B. Shen, "Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Natural Science*, vol. 2, no. 10, pp. 1090–1103, 2010.

[23] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS One*, vol. 6, no. 3, Article ID e18258, 2011.

[24] M. Hayat and A. Khan, "MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM," *Journal of Theoretical Biology*, vol. 292, pp. 93–102, 2012.

[25] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 411–421, 2012.

[26] S. Mei, "Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning," *Journal of Theoretical Biology*, vol. 310, pp. 80–87, 2012.

[27] H. Mohabatkar, "Prediction of cyclin proteins using chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 10, pp. 1207–1214, 2010.

[28] X. Xiao, Z.-C. Wu, and K.-C. Chou, "iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 42–51, 2011.

[29] H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2005.

[30] C. Ding, L. F. Yuan, S. H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[31] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 2, pp. 63–92, 2009.