ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Sequence-Based Prediction of Plant Allergenic Proteins: Machine Learning Classification Approach

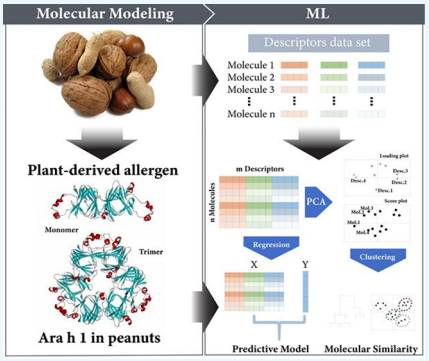
Miroslava Nedyalkova,* Mahdi Vasighi, Amirreza Azmoon, Ludmila Naneva, and Vasil Simeonov

Cite This: *ACS Omega* 2023, 8, 3698−3704

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: This Article proposes a novel chemometric approach to understanding and exploring the allergenic nature of food proteins. Using machine learning methods (supervised and unsupervised), this work aims to predict the allergenicity of plant proteins. The strategy is based on scoring descriptors and testing their classification performance. Partitioning was based on support vector machines (SVM), and a *k*-nearest neighbor (KNN) classifier was applied. A fivefold cross-validation approach was used to validate the KNN classifier in the variable selection step as well as the final classifier. To overcome the problem of food allergies, a robust and efficient method for protein classification is needed.



INTRODUCTION

The mechanisms behind food allergies are not fully understood. At the same time, food allergies are a significant social health problem. No one person is protected from developing a food allergy at any time of life. Eight foods cause the most common allergic reactions: milk, eggs, peanuts, tree nuts, soy, wheat, fish, and shellfish. In the work of Nagler et al., the question of why one person tolerates a food while another is allergenic was outlined,[1] and the microbiome was pointed to as a reason for allergenicity. In recent years, the microbiome has been explored through many different studies, which has led to the application of machine learning data to reveal patterns and trends.[2] Many unanswered questions make it difficult for researchers to develop a proper treatment with broad applicability against food allergies.

Food allergens are water-soluble proteins that differ significantly according to their origin, such as whether they are animal- or plant-derived.[3−8] The production of antigen-specific IgE antibody responses indicates food allergy and atopic disease.

The allergen-induced activation process occurs on epithelial barrier surfaces and is caused by the alarmins thymic stromal lymphopoietin (TSLP), interleukin 33 (IL-33), and interleukin 25 (IL-2). The produced alarmins invoke the production of type 2 innate lymphoid cells (ILC 2s), which generate Th2 cytokines and prime DCs to stimulate an allergen-specific immune reaction.[3] Methods of allergen identification include serology and cytology approaches as well as in vivo and in vitro techniques. Although considered reliable, all these methods are complicated, expensive, and relatively slow. Therefore, more attention has recently been given to bioinformatics and machine learning strategies as potential tools for detecting and classifying food allergens. Among the great variety of methods, intelligence neural networks, supervised learning, support vector machines with linear kernel functions, and different classifiers such as *k*-nearest neighbor are used as reliable options for identifying, modeling, and predicting allergenic properties.[8−11] Wang et al.[12] developed a new deep learning model (transformer with a self-attention mechanism combining the learning models Light Gradient Boosting Machine [LightGBM] and eXtreme Gradient Boosting [XGBoost]) for the prediction of food allergens. Machine learning is proving to be a tremendously helpful solution in this field. Comprehensive proteomic analysis still cannot be completely avoided, but combining it with complementary techniques will aid the development of a future coherent model.[13]

It is worth mentioning that the reported efficiency of an applied prediction from machine learning is often better than those of real in vivo and in vitro experiments. According to recent studies,[14,15] classical chemometric approaches can be used as strategies for the interpretation, modeling, classification, and prediction of the allergenic nature of food proteins. They are characterized by simplicity, rapidity, and sufficient efficacy.

Received: May 10, 2022
Accepted: November 21, 2022
Published: January 20, 2023

Figure 1. Basic computational framework of the proposed method.

Cluster analysis (CA) is an option that provides a calculation algorithm for grouping a set of objects characterized by various descriptors (variables) into patterns of similarity. The goal of CA is to reach an optimal grouping of the objects (or descriptors) corresponding to the selected measure of similarity between members of each group (cluster) and to ensure a difference between identified clusters. In hierarchical cluster analysis, the grouping of objects is spontaneous (using an unsupervised pattern recognition method), and the number of clusters to be formed is not known in advance. On the other hand, nonhierarchical clustering (most often known as K-means clustering) is a supervised pattern recognition method in which the number of clusters is retrieved from previous steps. The clustering algorithm requires the initial normalization of the raw data (autoscaling or Z-transformation), the introduction of Euclidean distances as the measure of similarity, and the selection of appropriate linkages between the clusters. A tree-like plot (dendrogram) usually illustrates the hierarchical relationship between clusters. No hierarchical relation exists in nonhierarchical clustering.

Comparing CA and more traditional classification methods such as partial least-squares discriminant analysis (PLS-DA), Zimmerman et al.[13] found that the classification results reached by CA for the separation of allergenic from nonallergenic proteins are of the same level of efficiency as those reached by PLS-DA.

Another work[14] presented a simple way to classify food proteins by allergenicity. The methods used for problem solving were well-established multivariate statistical strategies (hierarchical and nonhierarchical cluster analysis, two-way clustering, principal component analysis, and factor analysis), which are an essential part of exploratory data analysis (chemometrics). The methods were applied to a data set of 18 food proteins (allergenic and nonallergenic). They convincingly showed that the classification of two types of food proteins could be easily achieved by selecting simple and accessible physicochemical and structural descriptors. Optimal descriptors were selected by applying principal component analysis and factor analysis through the successive reduction of initial descriptor numbers and checking the resolving power for the chosen descriptors. The results may be of significant importance for building a model for partitioning allergenic from nonallergenic food proteins without engaging complicated software methods and resources.

Our study proposes a concept for the still-open niche of detecting allergenicity in food proteins through chemometrics, which is a part of the broader field of solving structural problems of biomolecules through chemometric approaches.

MATERIALS AND METHODS

Structure of the Data Set. A data set of 954 food allergens and nonallergens was collected from the databases of CSL (Central Science Laboratory) (http://allergen.csl.gov.uk), FARRP (Food Allergen Research and Resource Program) (http://www.allergenonline.org), SDAP (Structural Database of Allergenic Proteins(http://Fermi.utmb.edu/SDAP/sdap_manhtml). The data set and generated descriptors are available in the Supporting Information.

COMPUTATIONAL FRAMEWORK

The main parts of the computational framework of the proposed method are shown in Figure 1. After the allergenic and nonallergenic protein sequences were collected and the data set was prepared in FASTA format, sequence information was incorporated to extract compositional and AAindex-based properties, i.e., the 2 g exchange group frequency (TGR) and the radius of gyration (GYR). Protein sequences were encoded in information-rich descriptor vectors and used to build a classifier to discriminate allergenic from nonallergenic proteins. To find the most important and valuable descriptors, a sequential forward-selection strategy with a KNN classifier was used. The resulting best subset of descriptors was used to build the final classifier. A fivefold cross-validation (5-fold CV) approach was used to validate the KNN classifier in the variable selection step and the final classifier. To perform a 5-fold CV, protein data sets were divided into five subsets. At each step of the CV, one of the subsets was left out as the test set to validate the classification model, which was trained using the remaining four subsets. These important steps of the proposed approach will be explained in the next sections.

Feature Extraction. Each amino acid has a unique chemical structure based on its side chain, and only 20 different amino acids are known in eukaryotes. A protein consists of a chain of multiple amino acids linked together by peptide bonds. Different proteins have different numbers and orders of amino acids, which results in proteins having unique folding and functionality characteristics in the native state. A key step in developing an efficient classifier for discriminating allergens and nonallergens based on sequence information is encoding the sequences into informative numerical descriptors. According to previous classification studies and studies on allergenic proteins, the proposed method utilizes sequence-based properties as well as the physicochemical properties of amino acids. The properties of allergenic and nonallergenic proteins have also been considered in the selection of features, which is described as follows.

Amino Acid Composition. The amino acid composition (AAC) is represented by a feature vector consisting of 20 values (AAC1−AAC20), each of which is the frequency of occurrence for an amino acid in the protein sequence. AAC contains general information about the proteins in terms of the amino acid content and does not take the order of amino acids into account.

Hydrophobic Group. Amino acids can be classified into hydrophobic and hydrophilic groups based on side-chain structure and characteristics. Hydrophobic amino acids are {A, C, F, I, L, M, P, V, W, Y}, and hydrophilic amino acids are {D, E, G, H, K, N, Q, R, S, T}. By calculating the frequency of hydrophobic and hydrophilic amino acids along the sequence, two numerical descriptors can be obtained for the hydrophobic and hydrophilic frequency, which are abbreviated as HYD1 and HYD2, respectively.

https://doi.org/10.1021/acsomega.2c02842
ACS Omega 2023, 8, 3698−3704

Electronic Group. This property describes the electronegativity of an amino acid. Amino acids have different electronic properties according to their side-chain structure. Amino acids were classified into six electrical groups (Table 1). By listing the amino acids in the protein sequence and calculating the amino acid frequency for each category, six numerical descriptors (ELC1−ELC6) can be obtained.

Table 1. Electronic Groups of Amino Acids

electronic group	descriptor abbreviation	amino acids
electron donor	ELC1	D, E, P, A
weak electron donor	ELC2	V, L, I
electron acceptor	ELC3	K, N, R
weak electron acceptor	ELC4	F, Y, M, T, Q
neutral	ELC5	G, H, W, S
special AA	ELC6	C

R Group. Amino acids can be clustered into different groups according to other characteristics of their side chains. In this study, the clusters provided by Kedarisetti et al.[15] were used, which divided the amino acids into five groups (Table 2). The frequency of occurrence for each group was calculated along the protein sequence, and five numerical descriptors (RGR1−RGR5) were obtained for each protein sequence.

Table 2. Groups of Amino Acids Based on the Side Chain Characteristic

residue group	description abbreviation	amino acids
nonpolar aliphatic	RGR1	A, L, I, V
glycine	RGR2	G
nonpolar	RGR3	F, M, P, W
polar uncharged	RGR4	C, N, Q, S, T, V
charged	RGR5	D, E, H, K, R

2 g Exchange Group Frequency. To calculate this feature, the protein is converted into a sequence of the same length with reduced alphabets (a, b, g, d, e, and z) based on evolutionary information and the PAM matrix (Table 3).

Table 3. Groups of Amino Acids Based on Evolutionary Information

reduced alphabet	amino acids
α	K, H, R
β	D, E, N, Q
γ	C
δ	A, G, P, S, T
ε	I, L, M, V
ζ	F, Y, W

For example, a sequence of KRDQGKIH will be reduced to $\delta\delta\beta\beta\delta\beta\delta$. The occurrence frequency of all possible 2 g words (combinations of two consecutive reduced letters) is calculated. In this study, the method proposed by Eghbal et al.[16] was used to reduce the amino acid alphabet to six letters. Hence, 36 numerical descriptors (TGR1−TGR36) are obtained for each sequence. Both the order of amino acids and the relative mutability during evolution are reflected in this feature.

Groups of amino acids bases on evolutionary information

AAindex-Based Descriptors. To incorporate the compositional property of protein sequences in the computational model, several physicochemical properties of amino acids were used to calculate AAindex-based descriptors. Hydrophobicity is one of the most important physicochemical factors in the folding of proteins and the formation of their three-dimensional structures. Amino acids have different degrees of hydrophobicity based on their side chains, and this property can be measured using different hydrophobic indicators. In this study, the normalized Eisenberg hydrophobicity index[17] has been used (Table 4). The sum of this index is calculated for all

Table 4. List of Descriptors Sets, Number of Numerical Descriptors for Each Set, and Corresponding Abbreviations

descriptor set	abbreviation	number of descriptors
amino acid composition	AAC	20
electronic group	ELC	6
hydrophobic group	HYD	2
sum of hydrophobicity	HYS	1
R-group	RGR	5
2 g exchange group frequency	TGR	36
accessible surface area	ASA	2
sum of the normalized van der Waals volume	VAN	1
sum of the radius of gyration	GYR	1
number of hydrogen bond donors	HDN	1
sum of the partition coefficient	PRT	1
sum of average flexibility indices	FLX	1
net charge	CHR	1
number of rings	RNG	1

amino acids in the sequence and gives a numerical descriptor abbreviated as HYS. Similarly, we have calculated indices for the sum of the accessible surface area[18] for hydrophobic and hydrophilic residues (ASA1 and ASA2), the sum of the normalized van der Waals volume (VAN), the radius of gyration (GYR), the partition coefficient (PRT), the sum of the average flexibility (FLX), the sum of the net charge (CHR), and the total number of rings (RNG) over all amino acids in the protein sequence. All mentioned physicochemical and biochemical properties of amino acids were extracted from the AAindex database (version 9.2)[19] and are summarized in Table S1 and S2 in the Supporting Information.

Given this set of sequence-related descriptors, the feature vector of each protein sequence has 79 elements (presented in the Table S2). The list of descriptor sets and the corresponding abbreviations are given in Table 4.

Feature Selection. Feature extraction methods can produce vast numbers of features that may be noise or may be redundant or irrelevant for modeling purposes. Hence, using all extracted features in the data classification process can increase both the computational cost and the chance of incorrect label prediction, which would decrease the model performance. Feature selection methods can reduce the dimensions of the feature space by finding the most suitable subset including the top-ranked information-rich features.[20] In this study, sequential forward selection[21] was used as the search strategy and the non-error rate (NER) of the cross-validated KNN model was used to measure the performance of the subsets during the selection process. Additionally, the random forest (RF) classifier was used as an embedded feature

https://doi.org/10.1021/acsomega.2c02842
ACS Omega 2023, 8, 3698−3704

selection method to evaluate variable importance and select a reduced subset for classification.

Classification. To build a predictive model, the extracted features were presented to several classification methods and the results were compared to decide the best performing model. In this study, we used three supervised machine learning methods called *k*-nearest neighbors (KNN), support-vector machine (SVM), and the naïve Bayes classifier (NB) to build a model for the correct classification of allergenic and nonallergenic proteins.

KNN is a supervised machine learning method that has been widely used for classification problems in bioinformatics and data science.[22] The KNN method is a sample-based classification method that predicts the label of the test sample based on the majority vote of its nearest neighbors.[23,24] In this study, the Euclidean distance was used to measure the distance of the nearest neighbor and 10-fold cross-validation was used to decide the best value for the number of neighbors (k).

SVM is one of the widely supervised methods used for classification in many research fields[25−27] and has also been adopted for the classification of proteins and peptides because of its ability to handle nonlinear and complex conditions.[28−30] The concept behind SVM is to map the data into higher dimensions using a kernel function and construct a maximum marginal hyperplane to discriminate classes in high-dimensional feature space. In this study, the radial base function (RBF) was selected to train the SVM model and the regularization parameter (C) and the kernel width parameter (g) were optimized through a grid-search strategy. More detail about SVM can be found in the literature[31]

Naïve Bayes is one of the most popular statistical approaches for solving classification problems and has been successfully used to classify biological sequences.[32−34] Naïve Bayes assumes that the probability distributions of variables are independent of each other. Hence, the calculation of conditional probabilities can be simplified significantly. In this way, the prediction of the class label can be seen as finding the outcome of maximum probability given a set of calculated descriptors.

Among the ensemble learning methods used for classification, RF is a popular technique that is widely used in the field of computational biology.[35,36] The decision trees are the building blocks of the RF classifier and operate as an ensemble. Each decision tree predicts a class membership, and the class with the most votes decides the final class predicted by the RF model. By randomly splitting features into different decision trees, relatively uncorrelated trees operating as an ensemble can outperform the individual decision trees.

Performance Evaluation. The classification performance of the classifiers was evaluated by various criteria, including sensitivity (Sn), specificity (Sp), accuracy (Acc), nonerror rate (NER), and Matthew's correlation coefficient (MCC), as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{3}$$

$$\mathrm{NER} = \frac{\mathrm{Sn} + \mathrm{Sp}}{2} \tag{4}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{5}$$

TP, TN, FP, and FN represent the number of the correctly classified allergenic proteins, the number of the correctly classified nonallergenic proteins, the number of nonallergenic proteins recognized as allergenic, and the number of allergenic proteins recognized as nonallergenic proteins, respectively. Furthermore, the area under the curve (AUC) from the receiver operating characteristics (ROC) is also reported.[37]

RESULTS AND DISCUSSION

To investigate the performance of the classifiers and the effect of variable selection on the prediction result, several experiments were conducted. The prepared data set of allergen and nonallergen proteins was analyzed following the provided computational framework and the classification models adopted, and fivefold cross-validation was used to evaluate the performance of the model at different conditions. Classifier performance was assessed using the performance measures presented in the previous section according to two general approaches: (i) using all 79 calculated descriptors to define feature space and (ii) applying feature selection to find an optimal feature set to reduce the dimensionality of the feature space and improve the classification model performance.

Performance Analysis Using All Descriptors. This section examines the results obtained by different classifiers using four different descriptor sets. In the first case, the descriptor set includes only AAC, ELC, HYD, and RNG features, which are composition-based features. In the second case, only the TGR descriptors are included as evolutionary based descriptors. The third descriptor set includes the AAindex-based feature descriptors HYS, ASA, VAN, GYR, HDN, PRT, FLX, CHR, and RNG. In the fourth case, all 79 properties are used in the construction of the classification model. The hyperparameters of all classifiers were optimized to ensure they performed at their optimal setting. The efficiency of the KNN classifier was evaluated at different numbers of neighborhoods k; k = 1 was the optimal number of neighborhoods and performed significantly better than the other cases. The best performance for the SVM classifier was obtained using a radial basis function as the kernel function, and the box constraint and kernel scale parameters were set at 14.09 and 3, respectively. The optimal kernel for the NB classifier was the normal distribution function with a width of 0.0346. The performance of the classifiers with different descriptor sets is summarized in Table S2. The results compare the overall performance of the KNN, SVM, and NB classifiers for the highest-performing descriptor sets in terms of different classification measures. For allergenic and nonallergenic sequences, it can be seen that KNN performs better than classifiers that use composition- and evolution-based descriptors. The classification results of KNN using all 79 descriptors are also comparable to the SVM classifier, which performs better than classifier feature spaces.

Feature Selection Results. To determine the most-relevant features and decrease feature redundancy, we used a sequential forward-selection strategy to find the best subset of descriptors. The search strategy starts with an empty set and adds descriptors to the subset sequentially to minimize an objective criterion. The objective criterion used here is 1 − MCC, which is calculated using the fivefold cross validation of

https://doi.org/10.1021/acsomega.2c02842
ACS Omega 2023, 8, 3698−3704

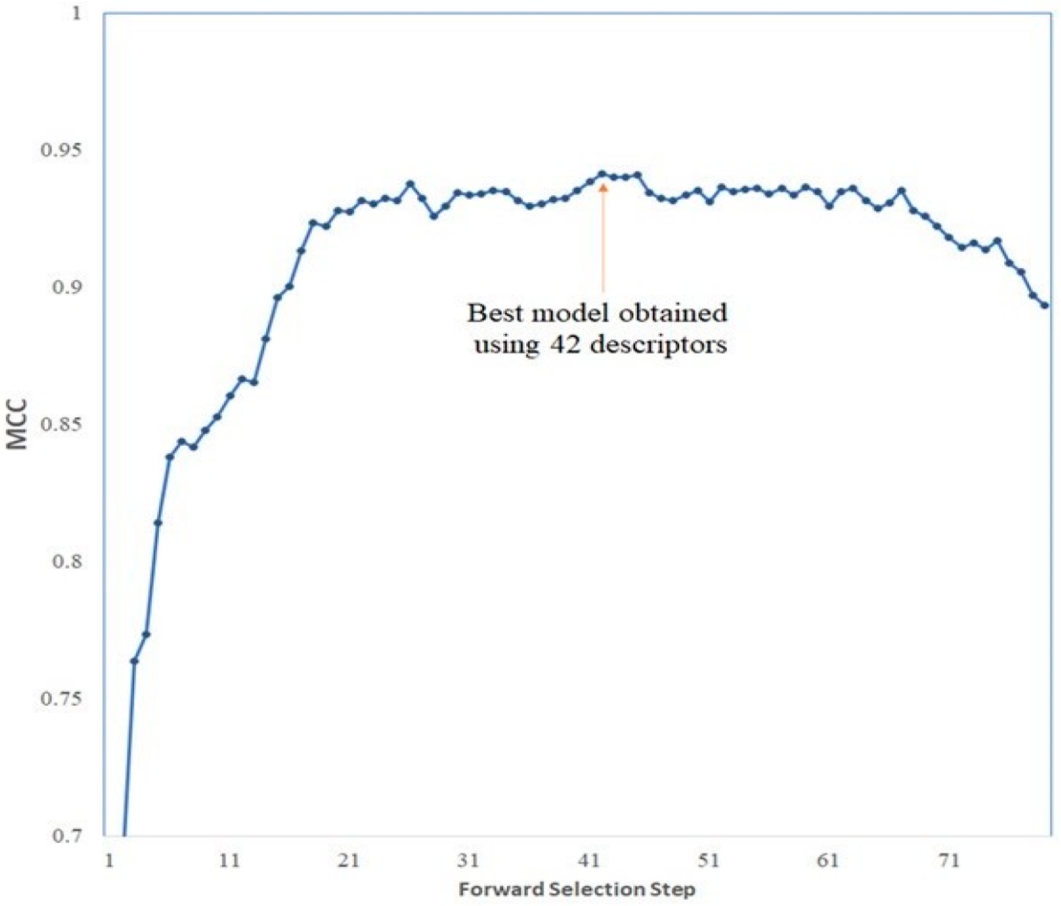


Figure 2. Objective criterion (1 − MCC) at the steps of the forward selection strategy.

the KNN classifier on the reduced set of descriptors at each step. Figure 2 shows the objective values at each forward-selection step. As shown in the figure, the best model was obtained using 42 descriptors (Table S2). The descriptors added at each step of the sequential feature selection, including the objective criterion value and a short description of the descriptor, are summarized in Table S4. Representing the evolution-based descriptors, the 2 g exchange group frequency (TGR) has the largest share among the 42 selected descriptors. Taking a closer look at the selected TGR descriptors, the ($\varepsilon\alpha$) exchange group (TGR5) is contributed the most frequently. The selection of the TGR group will be an improvement for the obtained classification accuracy based on an appropriate feature selection technique.

This means the adjacency of the amino acids {I, L, M, V} followed by {K, H, R} has the most effect on the classifier performance. Additionally, according to the selection of TGR9, TGR13, TGR15, TGR16, TGR17, and TGR21, the presence of g group (cysteine) and its adjacency with other exchange groups plays a significant role in the discrimination of allergenic and nonallergenic proteins. The selection of AAC5 (cysteine) as an important descriptor in the fifth step of variable selection confirms the role of cysteine, as claimed previously. The normalized van der Waals volume (VAN) and partition coefficient (PRT) of the amino acids in the protein sequence are AAindex-based descriptors and have a significant impact on the classification performance. Glutamine (AAC6) and arginine (AAC2) are other relevant compositional descriptors that significantly improve the model performance. The presence of electron-donor amino acids is also another important factor, as demonstrated by the selection of the ELC1 descriptor.

Table 5 summarizes the classifiers that were rebuilt using the set of 42 descriptors chosen by feature selection and classification measures. Similarly, fivefold cross validation has been employed to make a fair comparison among all the classifiers. It is clear that the SVM classifier, which uses a reduced descriptor set, in general performs better than KNN, RF, and NB for all metrics. The result obtained using SVM and variable solution is superior to that obtained in the case of using all descriptors for classification.

Table 5. Performance of Classifiers on Test Set Using Selected Descriptors and Fivefold Cross Validation

	performance measures					
classifier	Sen	Spec	Acc	NER	MCC	AUC
KNN	0.91	0.92	0.92	0.93	0.84	0.96
RF	0.89	0.92	0.91	0.93	0.82	0.97
RF (embedded)[a]	0.88	0.91	0.90	0.93	0.80	0.97
SVM	0.90	0.95	0.93	0.93	0.86	0.98
NB	0.47	0.86	0.66	0.71	0.36	0.78

[a]The reduced feature set was selected using RF.

In order to show the efficiency of the feature selection result, the RF classifier method was used to determine the importance of features in an embedded way. In this manner, the features with importance higher than the average importance of all

https://doi.org/10.1021/acsomega.2c02842
ACS Omega 2023, 8, 3698−3704

features were used as the selected feature set to construct the classification model using random forest. The selected feature set that used RF as an embedded method included 41 descriptors and was used to build a different RF classifier.

Clearly, the subset selected by RF shares many features with the subset selected with forward selection, and the selection of several AAC-related and TGR variables as highly important descriptors is in accordance with the result of feature selection by forward selection. Hence, regarding the classification performance, particularly the MCC value and the area under curve (AUC) for the test set using the reduced feature set, it can be concluded that the SVM classifier outperforms the others. Figure 3 compares the classification performance of the classifiers for the test set.

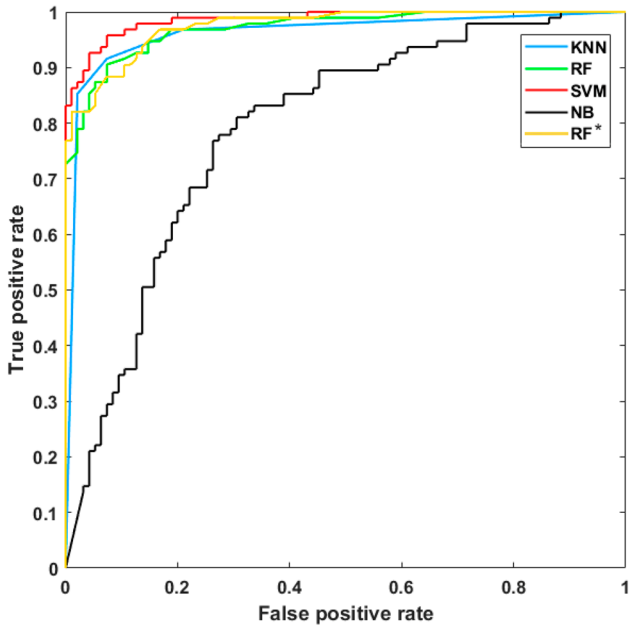


Figure 3. ROC curves showing the performance of the classifiers for the test set.

CONCLUSIONS

With this work, we want to bring a sustainable and coherent approach to building models for predicting the allergenic nature of proteins. The presented model can be a relevant and valuable tool that combines different stages to analyze allergens using the descriptor set. The level of predictor accuracy is close to those of other described methods. The model's main chemometric procedure for partitioning the data set into three specific clusters is based on knowledge obtained from these previous methods.

It appears that the complete separation of the objects into A and NA protein patterns depends not only on these specific allergenicity descriptors but also on other structural and physicochemical parameters. This possibility is well-illustrated by the formation of three patterns of proteins encompassing not only the "allergenic" and "non-allergenic" groups but also a "mixed" cluster.

ASSOCIATED CONTENT

Data Availability Statement

Data will be made available in the Supporting Information section.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c02842.

Photophysicochemical and biochemical properties of amino acids extracted from the AAindex database and descriptions of the used features (XLSX)

AUTHOR INFORMATION

Corresponding Author

Miroslava Nedyalkova − *Department of Chemistry, University of Fribourg, CH-1700 Fribourg, Switzerland; Faculty of Chemistry and Pharmacy, Inorganic Chemistry, University of Sofia, 1172 Sofia, Bulgaria;* orcid.org/0000-0003-0793-3340; Email: Miroslava.nedyalkova@unifr.ch

Authors

Mahdi Vasighi − *Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137, Iran*

Amirreza Azmoon − *Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137, Iran*

Ludmila Naneva − *Medical University, 9002 Varna, Bulgaria*

Vasil Simeonov − *Department of Inorganic Chemistry, University of Sofia, 1172 Sofia, Bulgaria*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c02842

Funding

This work was supported in part by the Bulgarian Science Fund (Grant K-06-KO/17−16.12.2020).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author M.N. is grateful for the support by the project COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies" and the project with Grant 80-10-152/16.04.2019 funded by the Bulgarian Science Fund.

REFERENCES

(1) Iweala, O. I.; Nagler, C. R. The Microbiome and Food Allergy Annu. *Rev. Immunol* **2019**, *37* (1), 377−403.

(2) Moreno-Indias, I.; Lahti, L.; Nedyalkova, M.; Elbere, I.; Roshchupkin, G.; Adilovic, M.; Aydemir, O.; Bakir-Gungor, B.; Santa Pau, E. C.; D'Elia, D.; Desai, M. S.; Falquet, L.; Gundogdu, A.; Hron, K.; Klammsteiner, T.; Lopes, M. B.; Marcos-Zambrano, L. J.; Marques, C.; Mason, M.; May, P.; Pašić, L.; Pio, G.; Pongor, S.; Promponas, V. J.; Przymus, P.; Saez-Rodriguez, J.; Sampri, A.; Shigdel, R.; Stres, B.; Suharoschi, R.; Truu, J.; Truică, C.-O.; Vilne, B.; Vlachakis, D.; Yilmaz, E.; Zeller, G.; Zomer, A. L.; Gómez-Cabrero, D.; Claesson, M. J. Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Front. Microbiol.* **2021**, *12*, No. 635781.

(3) Peterson, L. W.; Artis, D. Intestinal Epithelial Cells: Regulators of Barrier Function and Immune Homeostasis. *Nat. Rev. Immunol.* **2014**, *14* (3), 141−153.

(4) Lee, M. M.; Chan, M. K.; Bundschuh, R. Simple Is Beautiful: A Straightforward Approach to Improve the Delineation of True and False Positives in PSI-BLAST Searches. *Bioinformatics* **2008**, *24* (11), 1339−1343.
(5) Guarneri, F.; Guarneri, C.; Benvenga, S. Identification of Potentially Cross-Reactive Peanut-Lupine Proteins by Computer-Assisted Search for Amino Acid Sequence Homology. *Int. Arch. Allergy Immunol.* **2005**, *138* (4), 273−277.
(6) Goodman, R. E. Practical and Predictive Bioinformatics Methods for the Identification of Potentially Cross-Reactive Protein Matches. *Mol. Nutr. Food Res.* **2006**, *50* (7), 655−660.
(7) Hayes, M.; Rougé, P.; Barre, A.; Herouet-Guicheney, C.; Roggen, E. L. In Silico Tools for Exploring Potential Human Allergy to Proteins. *Drug Discovery Today Dis. Models* **2015**, *17*−*18*, 3−11.
(8) Soeria-Atmadja, D.; Lundell, T.; Gustafsson, M. G.; Hammerling, U. Computational Detection of Al- Lergenic Proteins Attains a New Level of Accuracy with in Silico Variable- Length Peptide Extraction and Machine Learning. *Nucleic Acids Res.* **2006**, *34* (1), 3779−3793.
(9) Soeria-Atmadja, D.; Zorzet, A.; Gustafsson, M. G.; Hammerling, U. Statistical Evaluation of Local Alignment Features Predicting Allergenicity Using Supervised Classification Algorithms. *Int. Arch. Allergy Immunol.* **2004**, *133* (2), 101−112.
(10) Mohabatkar, H.; Mohammad Beigi, M.; Abdolahi, K.; Mohsenzadeh, S. Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach. *Med. Chem.* **2013**, *9* (1), 133−137.
(11) Behbahani, M.; Rabiei, P.; Mohabatkar, H. A Comparative Analysis of Allergen Proteins between Plants and Animals Using Several Computational Tools and Chou's PseAAC Concept. *Int. Arch. Allergy Immunol.* **2020**, *181* (11), 813−821.
(12) Wang, L.; Niu, D.; Zhao, X.; Wang, X.; Hao, M.; Che, H. A Comparative Analysis of Novel Deep Learning and Ensemble Learning Models to Predict the Allergenicity of Food Proteins. *Foods* **2021**, *10* (4), 809.
(13) Zimmermann, J.; Hubel, P.; Pfannstiel, J.; Afzal, M.; Longin, C. F. H.; Hitzmann, B.; Götz, H.; Bischoff, S. C. Comprehensive Proteome Analysis of Bread Deciphering the Allergenic Potential of Bread Wheat, Spelt and Rye. *J. Proteomics* **2021**, *247*, No. 104318.
(14) Naneva, L.; Nedyalkova, M.; Madurga, S.; Mas, F.; Simeonov, V. Applying Discriminant and Cluster Analyses to Separate Allergenic from Non-Allergenic Proteins. *Open Chem.* **2019**, *17* (1), 401−407.
(15) Kedarisetti, K.; Kurgan, L.; Dick, S. Classifier Ensemble for Protien Structural Class Prediction with Varying Homolgoy. *Biochem. Biophys. Res. Commun.* **2006**, *348* (3), 981−988.
(16) Mansoori, E.; Zolghadri, M.; Katebi, S. Protien Superfamily Classification Using Fuzzy Rule-Based Classifier. *IEEE Trans. Nanobiosci* **2009**, *8* (1), 92−99.
(17) Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. Analysis of Membrane and Surface Protein Sequences with the Hydrophobic Moment Plot. *J. Mol. Biol.* **1984**, *179* (1), 125−142.
(18) Radzicka, A.; Wolfenden, R. Comparing the Polarities of the Amino Acids: Side-Chain Distribution Coef- Ficients between the Vapor Phase, Cyclohexane, 1-Octanol, and Neutral Aqueous Solution. *Biochemistry* **1988**, *27* (1), 1664−1670.
(19) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Res.* **2007**, *36* (suppl_1), D202−D205.
(20) Sharma, A.; Imoto, S.; Miyano, S. A Top-r Feature Selection Algorithm for Microarray Gene Expression Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 754−764.
(21) Zhang, L.; Kong, L. A Novel Amino Acid Properties Selection Method for Protein Fold Classification. *Protein Pept. Lett.* **2020**, *27*, 287−294.
(22) Chen, P.; Liu, C.; Burge, L.; Mahmood, M.; Southerland, W.; Gloster, C. Protein fold classification with genetic algorithms and feature selection. *J. Bioinform. Comput. Biol.* **2009**, *07*, 773−788.
(23) Ahmad, S.; Kabir, M.; Hayat, M. Identification of Heat Shock Protein Families and J-Protein Types by Incorporating Dipeptide Composition into Chou's General PseAAC. Computer Methods and Programs in Biomedicine. Identification of Heat Shock Protein Families and J-Protein Types by Incorporating Dipeptide Composition into Chou's General PseAAC. *Comput. Methods Programs Biomed.* **2015**, *122*, 165−174, DOI: 10.1016/j.cmpb.2015.07.005.
(24) Yang, L.; Xia, J.-F.; Gui, J. Prediction of Protein-Protein Interactions from Protein Sequence Using Local Descriptors. *Protein Pept. Lett.* **2010**, *17*, 1085−1090.
(25) Huang, S.; Cai, N.; Pacheco, P. P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics* **2018**, *15*, 41−51.
(26) Noble, W. S. What Is a Support Vector Machine? *Nat. Biotechnol.* **2006**, *24*, 1565−1567.
(27) Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing* **2020**, *408*, 189−215.
(28) Xu, L.; Liang, G.; Shi, S.; Liao, C. SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins. *Int. J. Mol. Sci.* **2018**, *19*, 1773.
(29) Muh, H. C.; Tong, J. C.; Tammi, M. T. AllerHunter: A SVM-Pairwise System for Assessment of Aller- Genicity and Allergic Cross-Reactivity in Proteins. *PloS one* **2009**, *4*, e5861.
(30) Das, S.; Chakrabarti, S. Classification and Prediction of Protein-Protein Interaction Interface Using Machine Learning Algorithm. *Sci. Rep.* **2021**, *11*, 1761.
(31) Cortes, C.; Vapnik, V. Support-vector networks. *Mach Learn* **1995**, *20*, 273−297.
(32) Geng, H.; Lu, T.; Lin, X.; Liu, Y.; Yan, F. Prediction of Protein-Protein Interaction Sites Based on Naive Bayes Classifier. *Biochem. Res. Int.* **2015**, *2015*, No. 978193.
(33) He, B.; Mortuza, S. M.; Wang, Y.; Shen, H.-B.; Zhang, Y. NeBcon: Protein Contact Map Prediction Using Neural Network Training Coupled with Naïve Bayes Classifiers. *Bioinformatics* **2017**, *33* (15), 2296−2306.
(34) Sharma, N.; Patiyal, S.; Dhall, A.; Devi, N. L.; Raghava, G. P. S. ChAlPred: A Web Server for Prediction of Allergenicity of Chemical Compounds. *Comput. Biol. Med.* **2021**, *136*, No. 104746.
(35) Qi, Y.Random Forest for Bioinformatics. In *Ensemble Machine Learning*; Zhang, C., Ma, Y., Eds.; Springer: Boston, MA, 2012; pp 307−323.
(36) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(37) Gribskov, M.; Robinson, N. L. Use of Receiver Operating Characteristic (ROC) Analysis to Evaluate Sequence Matching. *Comput. Chem.* **1996**, *20* (1), 25−33.

https://doi.org/10.1021/acsomega.2c02842
ACS Omega 2023, 8, 3698−3704