symmetry

MDPI

Article

Multivariate Chemometrics as a Strategy to Predict the Allergenic Nature of Food Proteins

Miroslava Nedyalkova [1] and Vasil Simeonov [2,*]

[1] Department of Inorganic Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia, 1 James Bourchier Blvd., 1164 Sofia, Bulgaria; nhmn@chem.uni-sofia.bg

[2] Department of Analytical Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia, 1 James Bourchier Blvd., 1164 Sofia, Bulgaria

* Correspondence: vsimeonov@chem.uni-sofia.bg

Received: 3 September 2020; Accepted: 21 September 2020; Published: 29 September 2020

check for updates

Abstract: The purpose of the present study is to develop a simple method for the classification of food proteins with respect to their allerginicity. The methods applied to solve the problem are well-known multivariate statistical approaches (hierarchical and non-hierarchical cluster analysis, two-way clustering, principal components and factor analysis) being a substantial part of modern exploratory data analysis (chemometrics). The methods were applied to a data set consisting of 18 food proteins (allergenic and non-allergenic). The results obtained convincingly showed that a successful separation of the two types of food proteins could be easily achieved with the selection of simple and accessible physicochemical and structural descriptors. The results from the present study could be of significant importance for distinguishing allergenic from non-allergenic food proteins without engaging complicated software methods and resources. The present study corresponds entirely to the concept of the journal and of the Special issue for searching of advanced chemometric strategies in solving structural problems of biomolecules.

Keywords: food proteins; allergenicity; multivariate statistics; structural and physicochemical descriptors; classification

1. Introduction

Food allergy is an atypical immunological reaction to food proteins, which causes an adverse clinical reaction. According to the data, approximately 5% of adults and 8% of children have a food allergy [1–3]. Allergy to cow's milk, eggs, wheat, soy, peanut, tree nuts, fish and shellfish comprises the majority of food allergy reactions. A study from 2001 indicated that peanut allergies are a mainstream cause for anaphylaxis, and fatal outcomes due to food allergies are 63%–67% of the deaths toll [4]. It has been proposed recently that the thermal treatment (boiled or in fried form) of peanuts leads to fewer allergenic products than roasting. In the work of Maleki et al., the digestibility of the major allergens in peanut during boiling, frying or roasting and in refined form was considered [5,6]. Despite the social importance of this issue, there is still no valuable methodology for prediction of allergenic structure or a proper methodology for the treatment of food allergies. The structural features of proteins could possibly contribute towards their allergenicity prediction. Developing such an *in silico* classification model may validate an appropriate approach for assisting in the allergenic potential of novel proteins. The exploratory methods based on a three-dimensional structure of allergens is of significance importance to make prediction models for allergenicity, which would allow the interpretation of the possible reasons for allergenicity of the proteins by combination of experimental and theoretical approaches

Why do proteins become allergens? This is a question that has triggered scientists to investigate what unique molecular features and properties make proteins become allergens. Information that the scientific community requires for allergy assessment should be developed by multiscale approaches and strategies with implications for different methods.

Only strategies based on multilayer approaches can boost early potential allergenicity detection. In the study of Naneva et. al., where an effort for prediction of allergic properties of a large number of proteins (over 700) is done by the use of linear sequence or surface spacial distribution of amino acids is made. The partial least square-based discriminant analysis (PLS-DA) classification model based on allergenic transformed protein data were constructed. A cluster analysis (hierarchical) was tried as a classification method for the separation of allergic from non-allergic proteins based on protein sequences to meet the needs of biologists in terms of phylogenetic analysis and prediction of biological functions like allergenicity [7–11]. A new type of molecular descriptor based on surface properties has been used. This approach for generating a database of is based on introduction of new types of descriptors able to reach classification of the proteins using only protein amino acids surface properties. To define the amino acids like polar, non-polar or charged, a set of hydrophobic scales were applied to explore these properties [11].

In the study of Guarino and Sciarrillo [12], deals with the proteome variation to different red strawberry species in order to clarify changes in allergen content and proteome variation for the different plant species. The detected allergens of strawberry were mapped on a 2-DE plot, and they were matched with spots recognized by a series of patients with different allergic patterns. By this approach, the authors identified the allergen proteins in *Fragaria ananassa* Duch, a variety of strawberry, by application of proteomic strategy compared to traditional approaches including protein isolation processes for discovering the binding between a patient's IgE (immunoglobulin E) and separated plant allergenic protein on a membrane. The obtained results revealed that the application of proteomics analyses enhanced identification of multiple allergens in plants in contrast to the well-known techniques. Therefore, besides for the conventional methods for noticing allergens, the use of the proteomic method has wide advantage and practical value in allergens studies concerning their detection and characterization. It is expected that a combination of proteomics and biological assays could substantially contribute to better understanding of the function the IgE-binding proteins. The cheminformatics methods have been widely established as a proper approach for drug discovery applications [8,9]. The principles used in drug development can also be useful in food chemistry as well [10]. The application of chemoinformatics for studies focused on food proteins reveals the broad spectrum of application of chemical information to elucidate structure–property relationships towards an object from food science in combination with data mining. This study reveals a workflow based on chemometric methods for prediction of the allergenicity nature of the most important food allergens causing IgE-mediated food allergy that is believed to be responsible for most immediate-type, food-induced hypersensitive reactions. Analysis and classification of the allergenic molecules by the proper choice of descriptors in the chemical space provide an overall model for prediction of the pattern for recognition of allergenic and non-allergenic proteins based on their properties.

The aim of this study is to demonstrate the ability of different chemometric methods, using mainly easily assessable structural descriptors for protein molecules, to separate allergic from non-allergic proteins.

2. Materials and Methods

In this stage of the work an attempt was made to separate (classify) allergenic from non-allergenic proteins using a set of descriptors of a molecular structural nature (totally 27) based on chemometric algorithms described in detail in [11–18]. The number of proteins involved in the separation procedure was 19–13 allergenic and 6 non allergenic proteins (from data sets with proven quality and correct qualification). Thus, the data set treated had dimensions 18 x 27.

The following multivariate statistical methods for data mining were used: hierarchical cluster analysis (HCA) using standardized input data, Ward's method of linkage, squared Euclidean distances as similarity measures and Sneath's test for cluster significance.

Nonhierarchical (K-means) clustering was applied as a supervised pattern recognition method used as confirmation of the results from hierarchical clustering.

Two-way joining (clustering) was included for finding correspondence between the objects and variables of interest represented on a single plot.

Principal component analysis and factor analysis were used for studying the data set structure, for reducing the number of the initial variables and for projection of the data on bivariate plots.

The key goal of the data mining was to reveal patterns of similarity between the objects of study (allergenic and non-allergenic proteins) and reach reliable separation between both classes, to determine of the descriptors responsible for the formation of both classes and establish similarity patterns between the descriptors used for possible further selection of reduced number of descriptors for classification and to explain the significance of the descriptors for the classification procedure.

All chemometric calculations were performed by the software package STATISTICA 8.0.

2.1. Data Preparation

The set of plant proteins included in the present study is shown in Figure 1. Those proteins were classified as allergens by the Protein Data Bank (PDB) (https://www.rcsb.org/) or/and by the Structural Database of Allergenic Proteins (SDAP) (http://fermi.utmb.edu/). Each protein was described by a set of descriptors. The following types of descriptors were computed (using MOE and Alvadesc software) and are presented in Table 1.

Allergenic group:

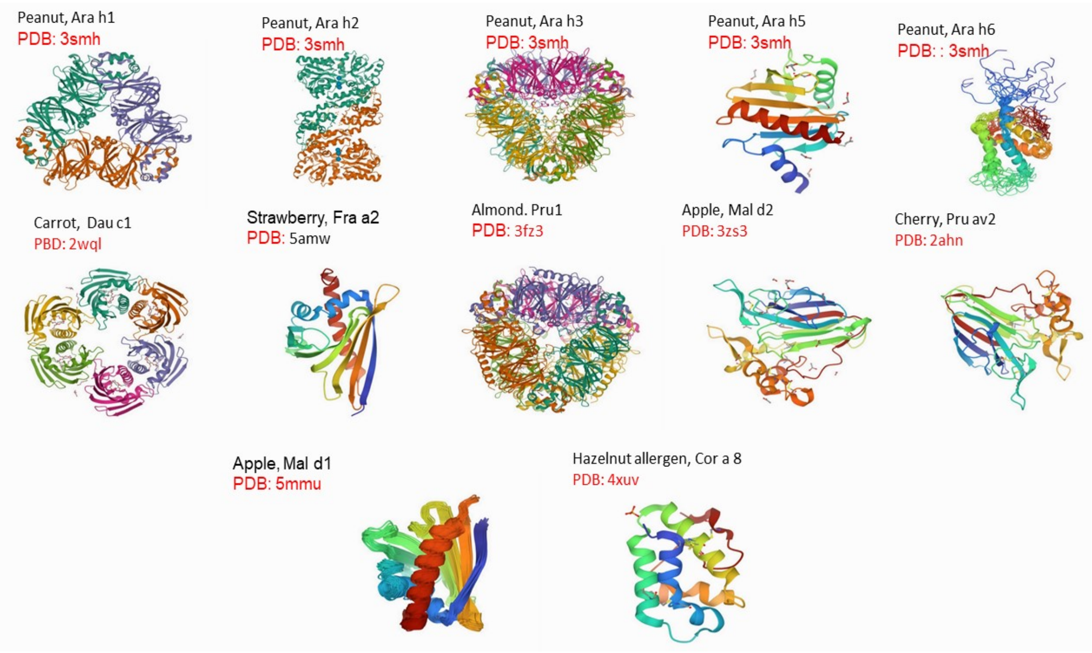

Figure 1. *Cont.*

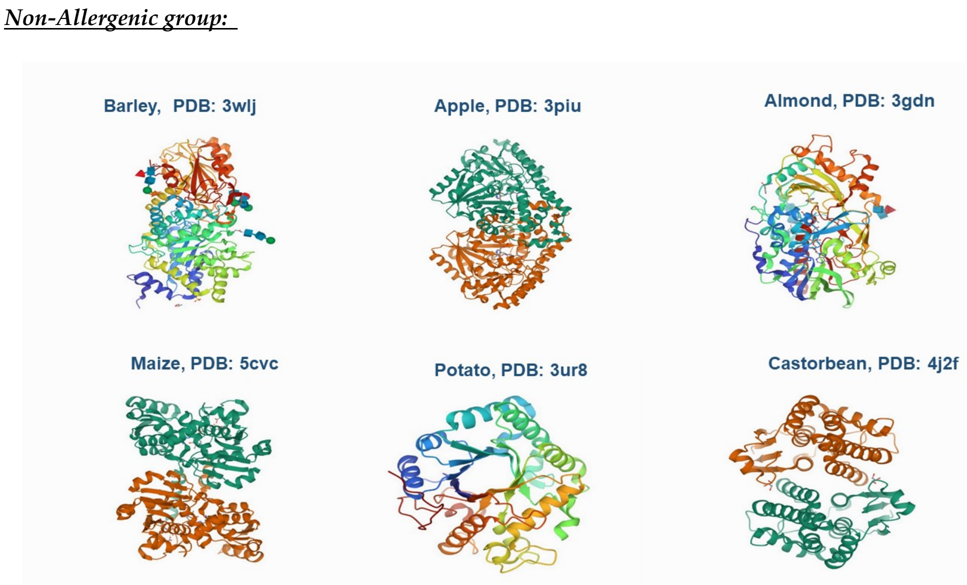

Figure 1. Allergenic and non-allergenic protein structures.

Table 1. Molecular descriptors used and their explanation.

CODE	DESCRIPTION
pro_asa_hph	Water accessible surface area of all hydrophobic ($\|q_i\| < 0.2$) atoms.
pro_asa_hyd	Water accessible surface area of all hydrophilic ($\|q_i\| < 0.2$) atoms.
pro_vdw	Van der Waals surface area
pro_dipole_moment	Dipole moment calculated from the partial charges of the molecule.
dens	Mass density: molecular weight divided by van der Waals volume as calculated in the vol descriptor.
pro_mobility	mobility
pro_charge	Total charge of the molecule (sum of formal charges).
pro_r_gyr	Radius of gyration.
pro_r_solv	Radius of cross-section
pro_volume	van der Waals volume calculated using a grid approximation (spacing 0.75 A).
a_acc	Hydrogen bond acceptor atoms (number)
a_acid	Number of acidic atoms.
a_aro	Number of aromatic atoms.
a_base	Number of basic atoms.
a_don	Number of hydrogen bond donor atoms (not counting basic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH).
b_1rotR	Fraction of rotatable single bonds: b_1rotN divided by b_heavy.
KierFlex	Kier molecular flexibility index: (KierA1) (KierA2) / n
rings	The number of rings.
SlogP	Log of the octanol/water partition coefficient (including implicit hydrogens).
TPSA (topological polar surface area based on fragments)	Polar surface area (Å2) calculated using group contributions to approximate the polar surface area from connection table information only.
Weight	Molecular weight (including implicit hydrogens)
weinerPath	Wiener path number
weinerPol	Wiener polarity number
Zagreb index	Zagreb index: the sum of di2 over all heavy atoms i.

2.2. Cluster Analysis for Protein Separation

Cluster analysis (CA) is a method to find optimal groupings of observations or their descriptive variables in such a way that the members of a cluster are similar to each other and the clusters formed are different from each other. Hierarchical clustering is a type of unsupervised machine learning algorithm. In unsupervised learning mode, the learner algorithm can be used to group the data, since the non-hierarchical clustering as a supervised pattern recognition method requires a priori determination of the number of groups for data interpretation.

In order to interpret the data structure, a similarity measure should be introduced, such as Euclidean distance. Unwanted data rotations in the data structure are avoided by different data transformations, the most applied one being auto scaling or z-transformation. The graphical output of the analysis is known as a dendrogram plot.

The next important step after auto scaling and distance determination is the linkage algorithm. There are many options, but hierarchical clustering relies often on Ward's method of linkage and non-hierarchical on K-means mode.

It must be mentioned that in non-hierarchical clustering, all a priori required clusters are simultaneously obtained, and this grouping does not possess hierarchy.

2.3. Principal Component Analysis for Protein Separation

Principal component analysis (PCA) is a powerful mathematical technique used to reduce the dimensionality of the parameter space [19]. PCA was first carried out with the aim of reducing the input variables. Varimax rotation mode was used. Three latent factors explaining over 80% of the total variance were selected for estimation of the variable relationships by factor loadings. Only statistically significant loadings (higher than 0.7) were considered for interpretation purposes.

3. Results and Discussion

3.1. Cluster Analysis for Protein Classification Based on 2D and 3D Molecular Descriptors

The essential findings presented in this study could be defined as a pattern recognition classifying an unknown pattern into one of two predefined categories (allergenic and non-allergenic groups). Cluster analysis can be considered an important strategy in pattern recognition aiming putting a set of patterns into classes (categories). The cluster analysis approach is a highly applicable sampling method. Sampling by clusters happens over multiple stages, and the resulting process is a defined by similarity path in data and pattern space. The clustering was performed using the linkage option of the method of Ward, which was found to be most suitable as it creates a small number of clusters. In order to check the option for separation of proteins into allergenic and non-allergenic classes using structural descriptors of the proteins, a data set of 18 food proteins was prepared (12 allergenic and 6 non-allergenic). All descriptors were z-score normalized prior to the analysis so that they were on the same dimensionless scale.

The results and the comments of the data interpretation can be summarized as follows.

In Figure 2 the hierarchical dendrogram for linkage of 27 descriptors is presented.

Two very well expressed clusters were formed. One of them mainly included descriptors for surface area, volume and shape descriptors. These descriptors depend on the structure connectivity and conformation (dimensions are measured in Å). The second group of descriptors was based on physical properties. The resulting physical properties could be calculated from the connection table (with no dependence on conformation states of the molecules) of a molecule and the Kier molecular flexibility index descriptor was also defined in this cluster.

Next, Figure 3 represents the linkage between 18 proteins, and the distinctive separation of the proteins in "allergenic" and "non allergenic" classes was obvious.

The upper cluster consisted entirely of allergenic proteins, and the lower one of non-allergenic ones.

In order to confirm the clustering found by HCA, nonhierarchical K-means clustering was also performed. The stated hypothesis was that both descriptors and protein should be separated into two predetermined clusters.

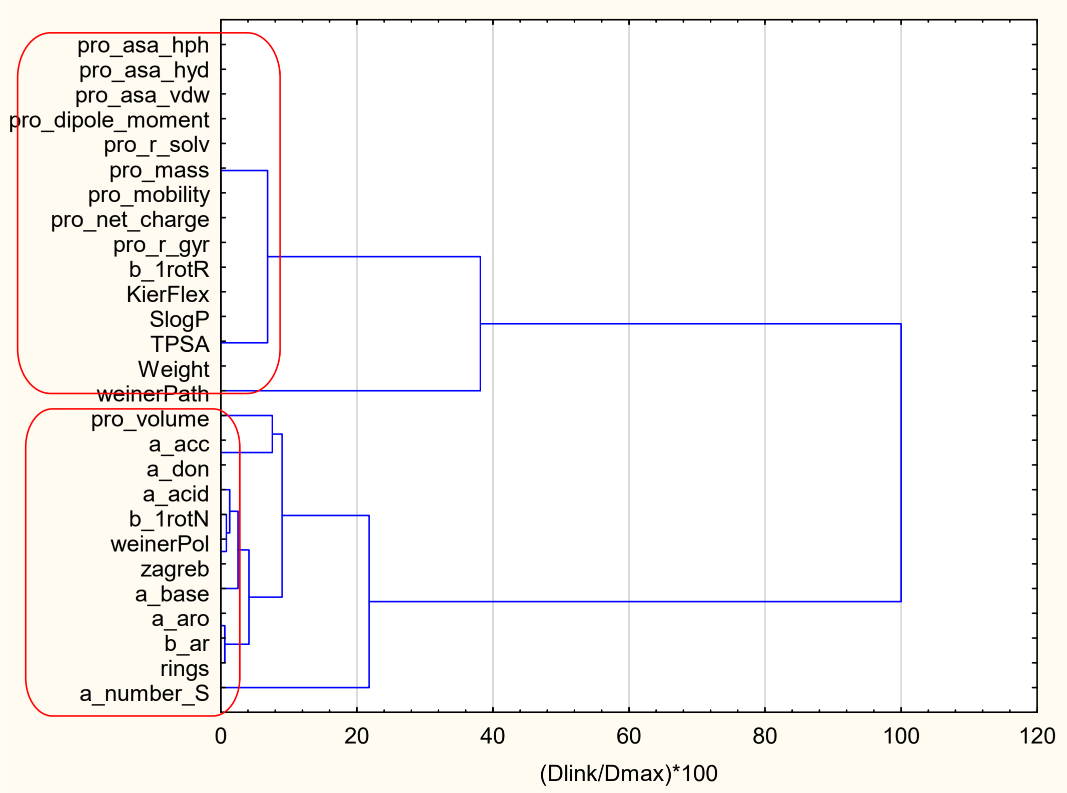

Figure 2. Hierarchical dendrogram for linkage of 27 descriptors.

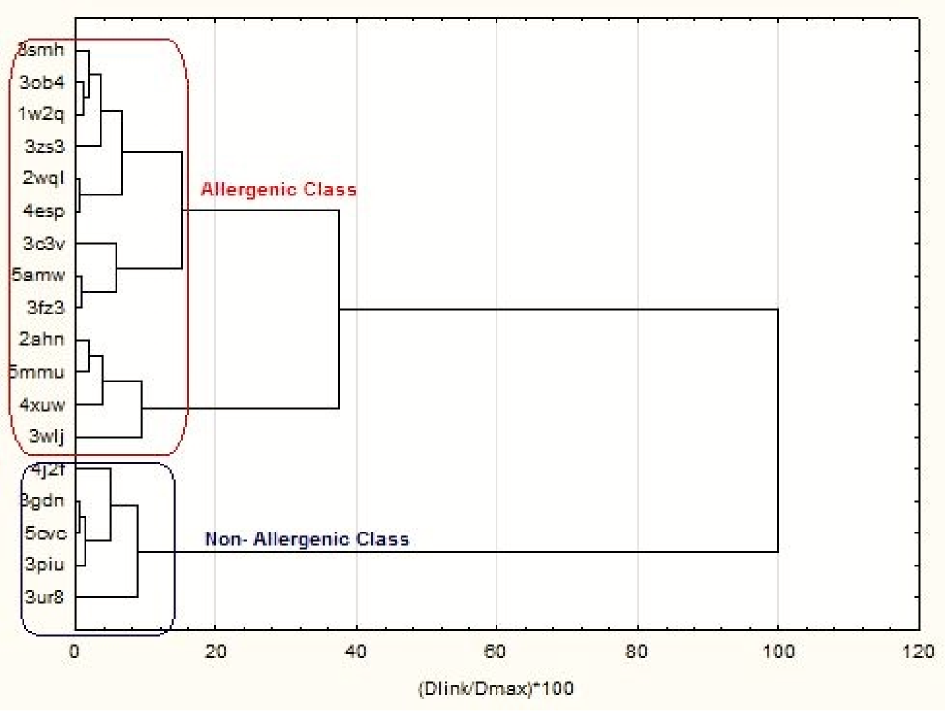

Figure 3. Hierarchical dendrogram for clustering of the 18 proteins.

In Table 2, the members of both clusters for 27 descriptors are presented, and in Table 2, the members of two clusters are presented for 18 proteins.

Table 2. Members of cluster 1 and cluster 2 for 27 descriptors.

Variable (2D Descriptors)	Members of Cluster Number 1 and Distances from Respective Cluster Center Cluster Contains 15 Variables
	Distance
pro_asa_hph	0.149240
pro_asahyd	0.149240
pro_asa_vdw	0.149240
pro_dipole_moment	0.149240
pro_mass	0.149240
pro_mobility	0.149240
pro_net_charge	0.149240
pro_r_gyr	0.149240
pro_r_solv	0.149240
B_1rotR	0.243133
KierFlex	0.243133
SlogP	0.243133
TPSA	0.243133
Variable (2D Descriptors)	**Members of Cluster Number 2 and Distances from Respective Cluster Center Cluster Contains 12 Variables**
	Distance
pro_volume	0.649643
a_acc	0.413425
a_acid	0.382602
a_aro	0,230073
a_base	0.469422
a_don	0.418075
a_number_S	1.026357
b_1rotN	0.322417
b_ar	0.231530
rings	0.236831
weinerPol	0.180442
zagreb	0.178149

As could be seen, the separation between the descriptors (variables) followed entirely that reached by HCA.

The obtained results for the distribution profiles for each class of proteins based on a descriptor set were classified as allergenic and non-allergenic. Clearly, this particular selection of descriptors provided a pure dissimilarity in the profiles of allergenic and non-allergenic compounds.

Tables 3 and 4 indicate the clustering of the proteins into two classes. Again, the similarity with HCA was almost perfect, as the only exception was protein barley defined as non-allergenic protein (PDB ID: 3wlj.pdb) belonging in K-means classification to the cluster of non-allergenic proteins rather than to the cluster of allergenic as expected.

Table 3. Members of cluster 1 for 18 proteins.

Proteins	**Members of Cluster Number 1 and Distances from Respective Cluster Center Cluster Contains 6 Cases** **Distance**
3wlj	0.624337
3ur8	0.919784
3gdn	0.250043
2v3f	0.266641
3piu	0.357769
4j2f	0.929623

Table 4. Members of cluster 2 for 18 proteins.

Proteins	**Members of Cluster Number 2 and Distances from Respective Cluster Center Cluster Contains 12 Cases** **Distance**
3smh	0.745996
3c3v	0.895197
2c3b	0.491340
1w2q	0.492303
2wql	0.569632
3vor	0.343808
5amw	0.490781
3fz3	0.660688
3zs3	0.633154
2ahn	0.616859
5mmu	0.672782
4cpv	0.928246

An important step in the chemometric analysis was to try to determine the role of the separate descriptors in the separation procedure. From the plot of means (Figure 4) for the mean value of each descriptor for each of the identified clusters of proteins, it is readily seen that each protein cluster is described by different values of the descriptors.

It was readily seen that almost all descriptors were very well separated from each other for each identified cluster of proteins. Except for the descriptor weinerPa, all descriptors for non-allergic proteins had higher average (standardized) values as compared to those for allergic proteins (cluster 2). In Figure 5, the plot of means for each protein (object) for each identified cluster of descriptors is presented.

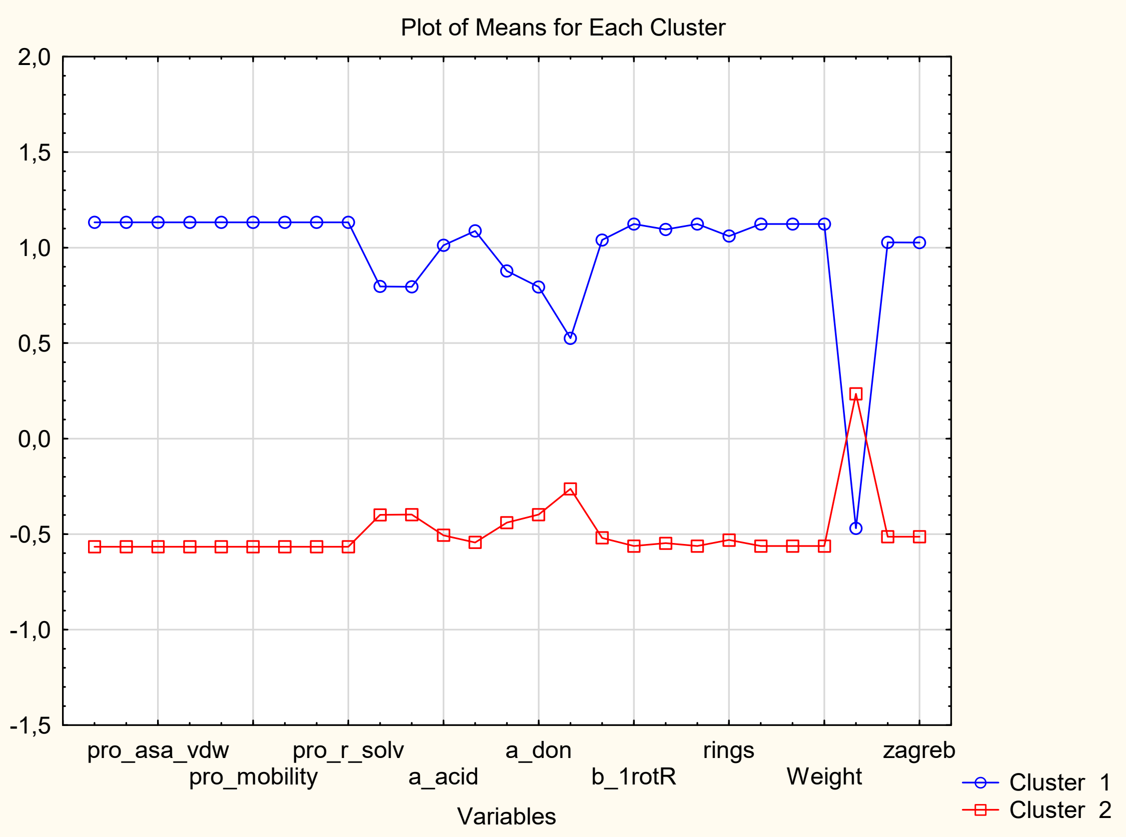

Figure 4. Plot of means for each descriptor for each identified cluster of proteins.

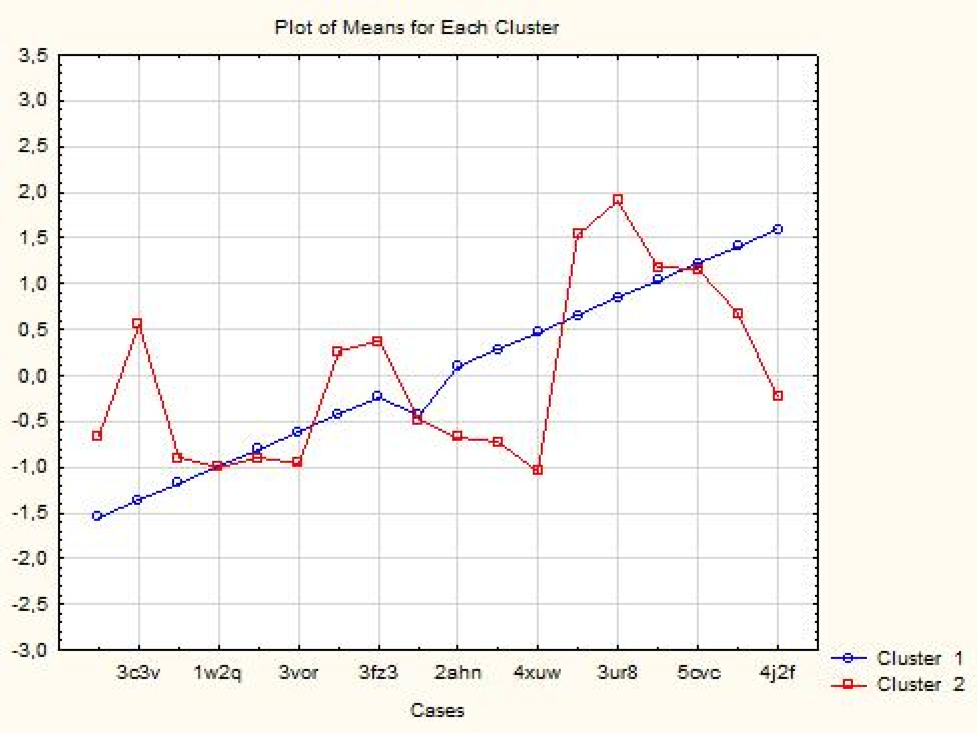

Figure 5. Plot of means (standardized values) of each protein for each identified cluster of descriptors.

These results implied that cluster dendrograms obtained by HCA can be reliably used to identify and gain deeper insights in the presented results from K-means clustering. The obtained linearity profile, related to the descriptors of proteins obtained as well as with HCA, since a second cluster with HCA shows a non-linearly trend in each descriptor space. (to be deleted)

The results in Figure 5 indicate that the clusters of descriptors identified by K-means clustering separate very well allergenic from non-allergenic proteins. It should be noted that protein 3wlj marked in advance as non-allergenic is correctly classified by K-means clustering since in hierarchical clustering its position is doubtful.

3.1.1. Two-Way Clustering

In the next step of the data mining two-way clustering approach was applied. In Figure 6 the correspondence between the groups of objects and groups of descriptors is presented.

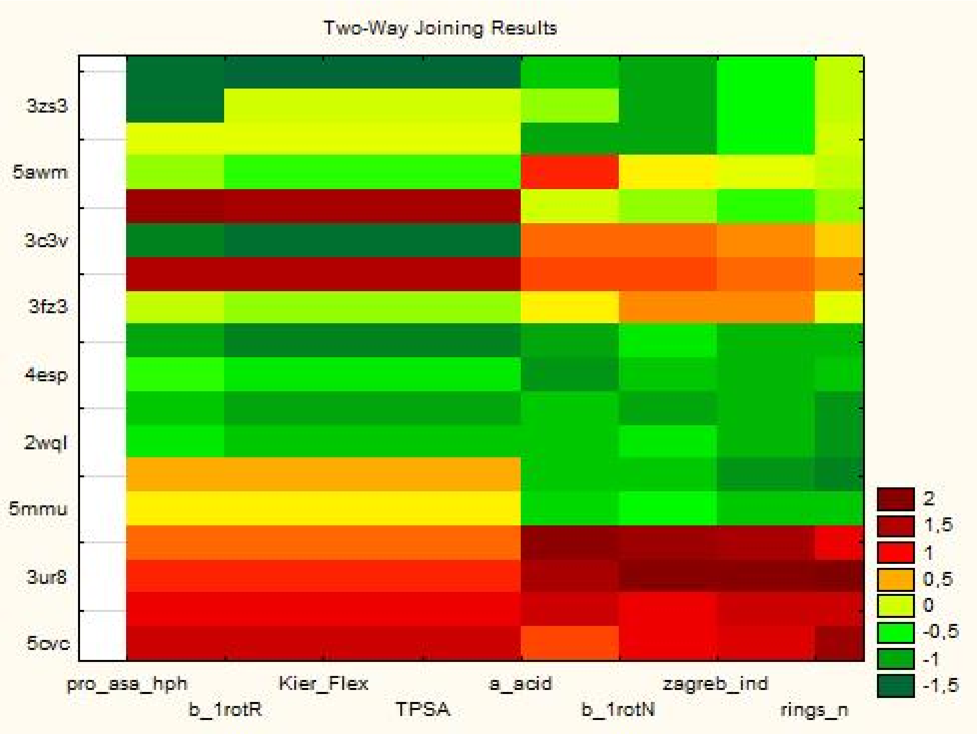

Figure 6. Two-way clustering of proteins and descriptors.

The output of the K-means clustering was confirmed: for non-allergic proteins, the descriptor values were higher than those for allergenic proteins.

3.1.2. Principal Components and Factor Analysis

Principal components analysis of the data to correct the heavy skewness in some variables was performed.

In Table 5, the factor loadings for two latent factors explaining over 85% of the total variance are presented.

Two latent factors explain over 85% of the total variance of the system. The first latent factor includes high factor loadings for the groups of descriptors for physical properties and surface area, volume and shape and the second one includes high factor loadings from the descriptors connected to groups atom counts and bond counts, Kier Hall connectivity and Kappa Shape Indices in general, it confirms the results from cluster analysis.

From this table is possible to select a reduced number of variables (descriptors) to try even better separation of the proteins into two classes.

Table 5. Factor Loadings.

Variable	Factor Loadings (Varimax Normalized) Extraction: Principal Components (Marked Loadings are >0.700000)	
	Factor 1	Factor 2
pro_mass	0.96464	0.236248
pro_mobility	0.96464	0.236248
pro_net_charge	0.96464	0.236248
pro_r_gyr	0.96464	0.236248
pro_r_solv	0.96464	0.236248
pro_volume	0.17264	0.728891
a_acc	0.18398	0.890058
a_acid	0.31796	0.867425
a_aro	0.38478	0.907553
a_base	0.17497	0.864993
a_don	0.18896	0.886082
a_number_S	0.04313	0.346071
b_1rotN	0.35525	0.884013
b_1rotR	0.94226	0.244491
b_ar	0.38999	0.905283
KierFlex	0.94226	0.244491
rings	0.31002	0.924282
SlogP	0.94226	0.244491
TPSA	0.94226	0.244491
Weight	0.94226	0.244491
weinerPath	−0.00924	−0.568203
weinerPol	0.27742	0.948778
zagreb	0.27791	0.949411
Expl. Var %	50.83	36.79

Graphically, the separation of the descriptor into two major groups and the special position of the descriptor Weiner Pa is well illustrated on Figure 7.

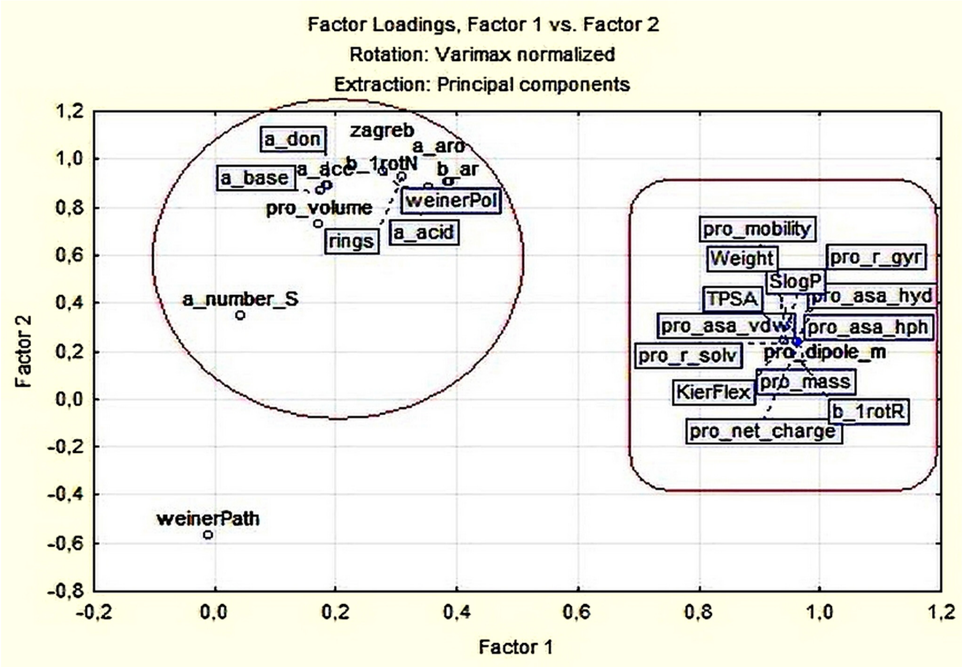

Figure 7. Biplot PC1 vs. PC 2 for grouping of descriptors.

In Figure 8, the same distribution is shown.

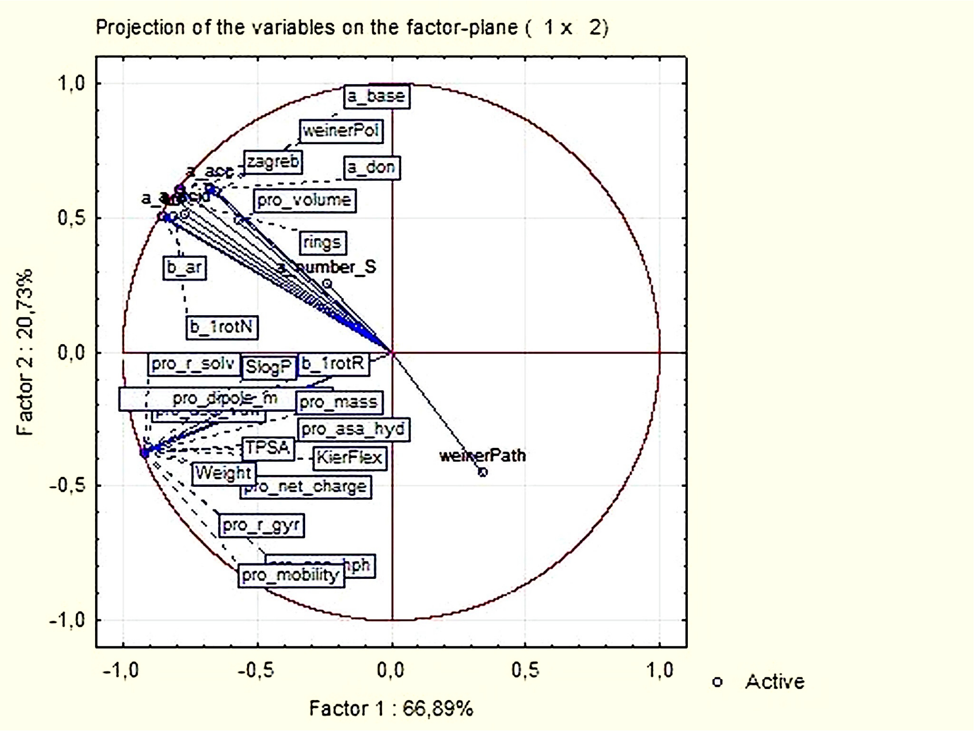

Figure 8. Projection of descriptors on the plane of the first two latent factors.

The separation of the objects (proteins) is illustrated additionally in Figure 9.

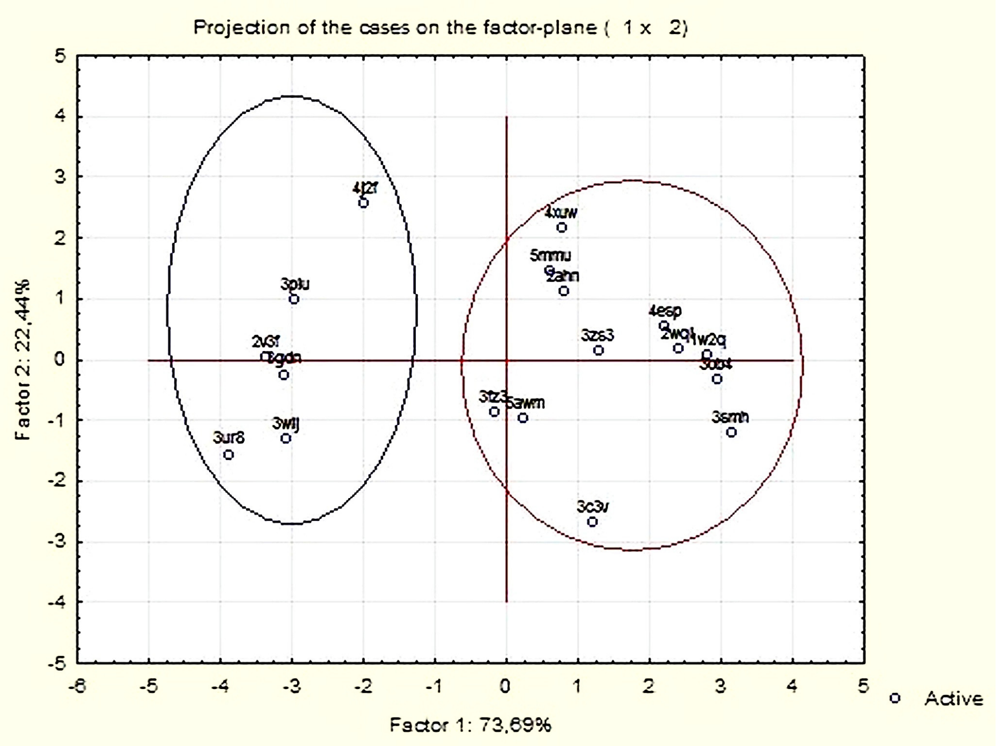

Figure 9. Projection of objects (proteins) on the plane of the first two latent factors.

There was a clear separation between allergenic and non-allergenic in each descriptor space and also showed a particularly good separation between two protein groups.

It was convincingly shown that the objects (proteins) were well separated into two classes: left side—non-allergic proteins, and right side—allergic proteins. Again, an exception was found concerning the classification of the expectedly allergic protein 3wlj.pdb as non-allergic.

3.1.3. Data Mining with Reduced Number of Descriptors

The data mining procedure was carried out once more with significantly reduced number of descriptors. The selection of descriptors was done on the basis of the factor loadings table, and out of several descriptors with high loadings (strong correlation), only single representatives were used: pro_asa; a_acid; b_1rotN, b_1rotR, Kier Flex, rings, TPSA, Zagreb

In general, the results of all multivariate statistical analyses applied (as in previous case) were the same:

Two clusters were formed, representative of the larger groups conditionally marked as a_descriptors and pro_descriptors (Figure 10).

Two clusters were formed, namely the upper part—allergenic—and the lower part—non allergenic. This way of feature reduction improved the allocation of protein 3wlj in the cluster of non-allergenic (Figure 11).

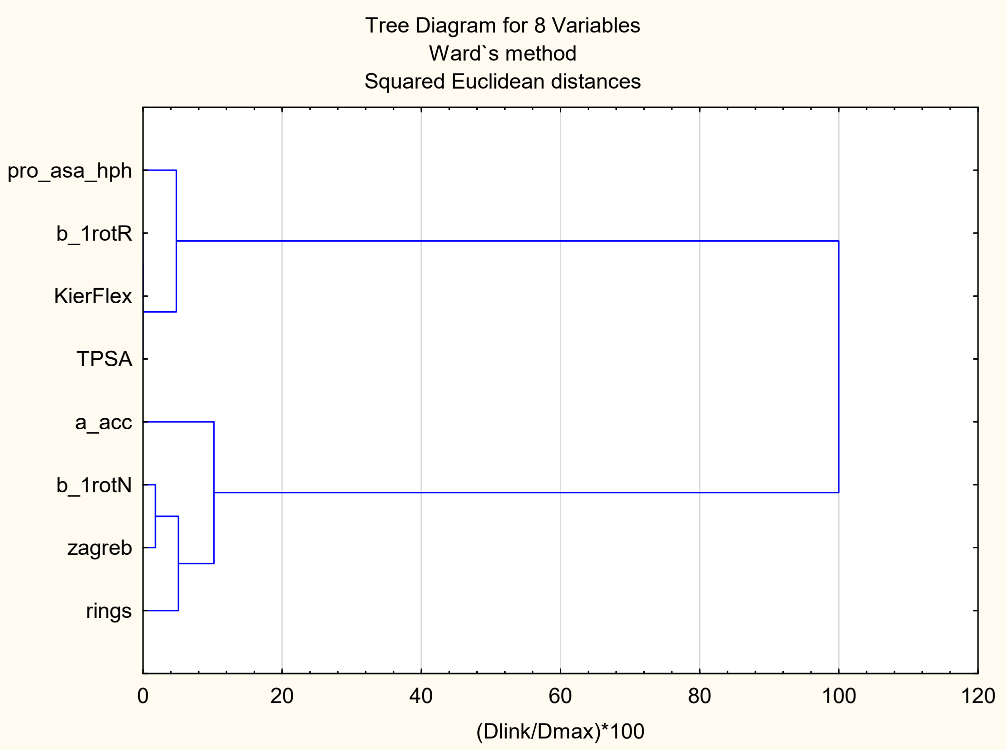

Figure 10. Hierarchical dendrogram for the clustering of 8 descriptors.

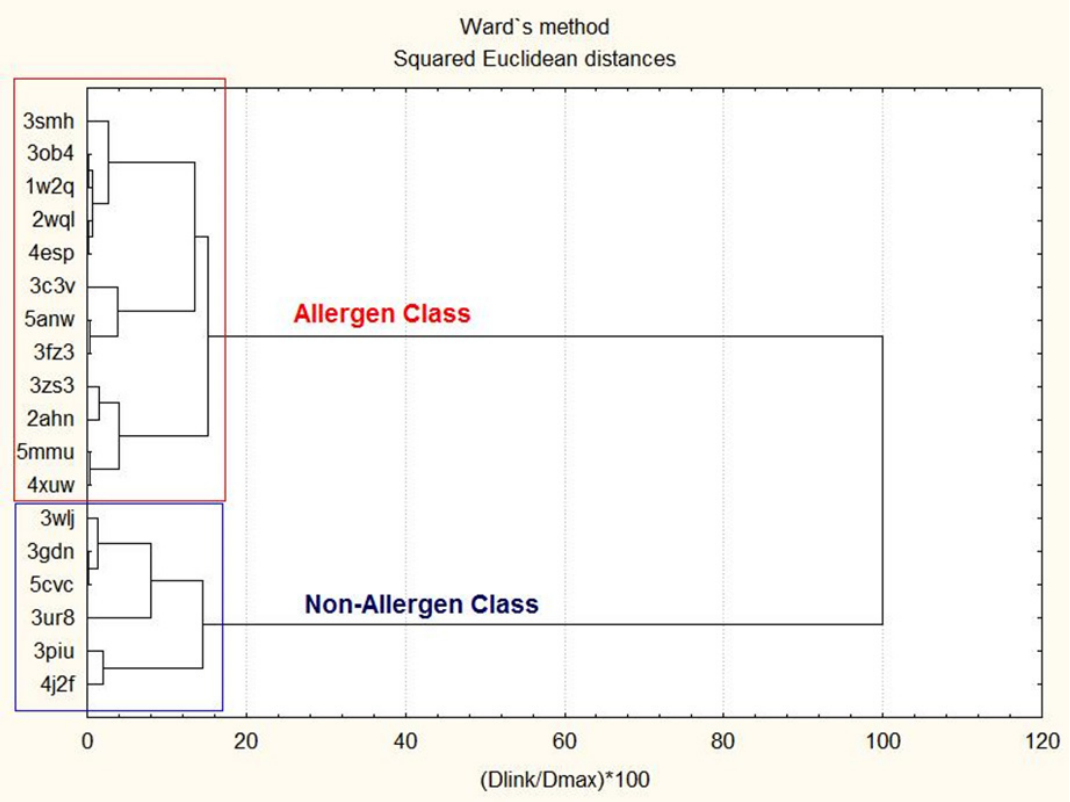

Figure 11. Hierarchical dendrogram for the clustering of 18 proteins.

The same separation of the descriptors into two clusters depicted in Figure 12 was observed with the reduced number of descriptors.

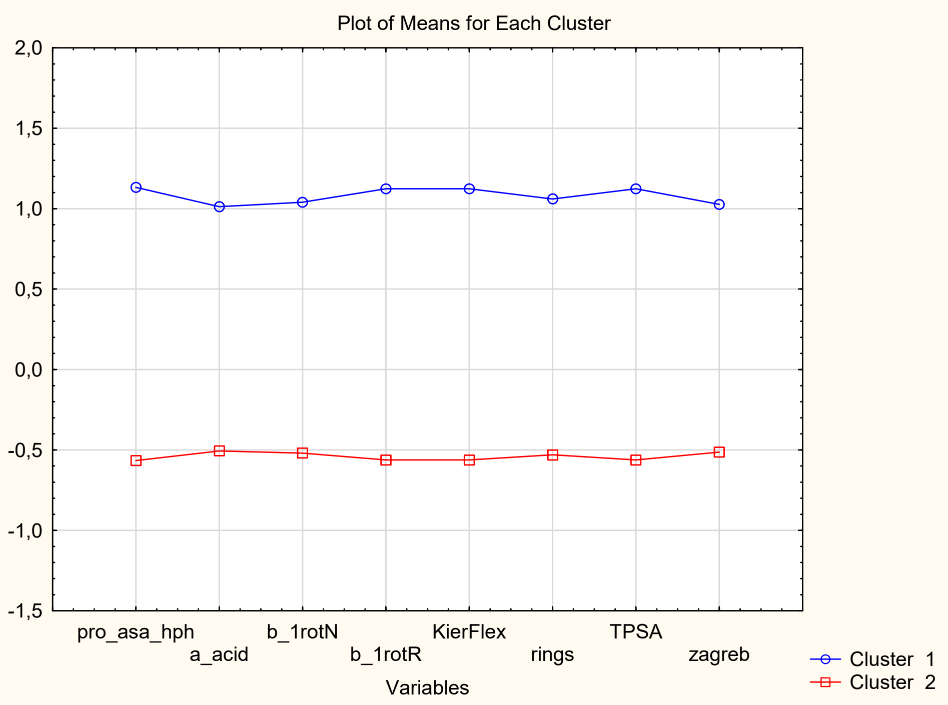

Figure 12. Plot of means (standardized) for each descriptor for each identified cluster of proteins.

The separation of the descriptors in this case was even better compared to the plot of all 27 descriptors. We did not present here the members of the two clusters of descriptors, but it resembled the outputs for 27 descriptors. More interestingly, the plot of means depicted in Figure 13 for each protein for each identified cluster of descriptors had the same output as the case with all descriptors.

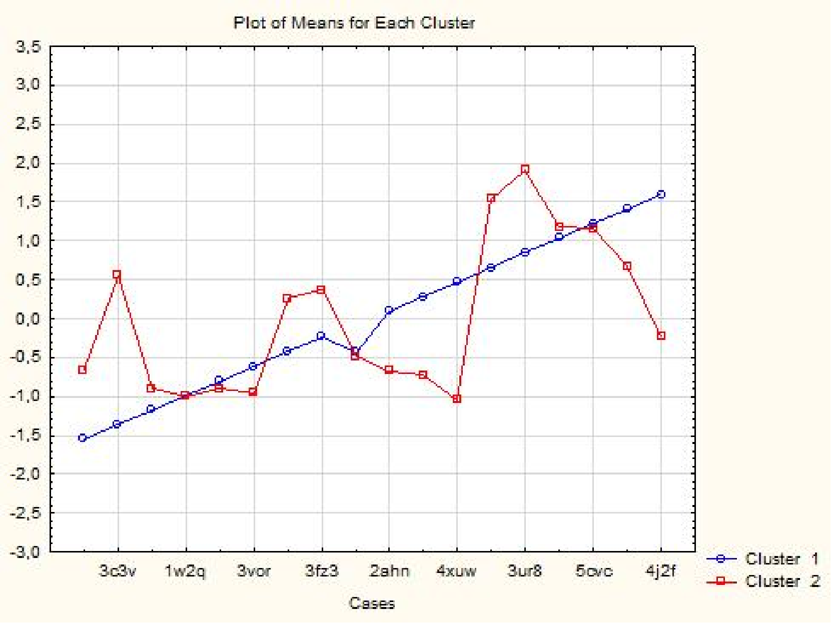

Figure 13. Plot of means (standardized) for each descriptor for each identified cluster of proteins.

It proves that the reduction of the number of descriptors does not change the existing tendency.

Similar tendency without observed biases were reported for the results of the two-way clustering presented in Figure 14. The same trend was observed for all descriptors.

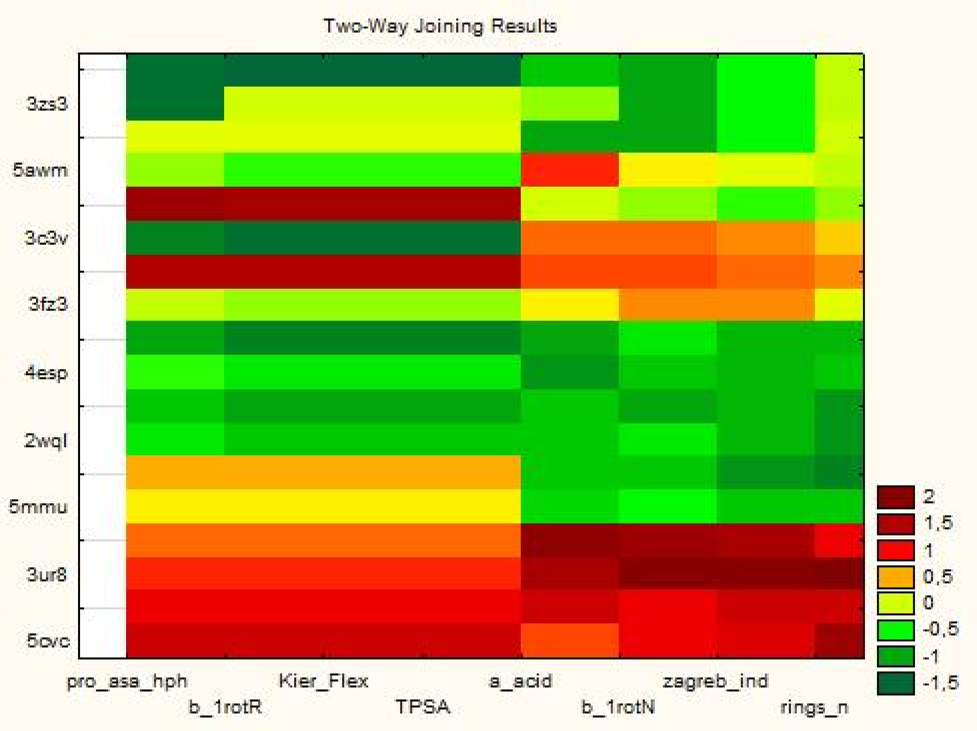

Figure 14. Two-way clustering of proteins and descriptors.

This is also confirmed on the biplot presented in Figure 15 above and on the projection plot for descriptors below as indicated in Figure 16 and Table 6 for the factor loadings.

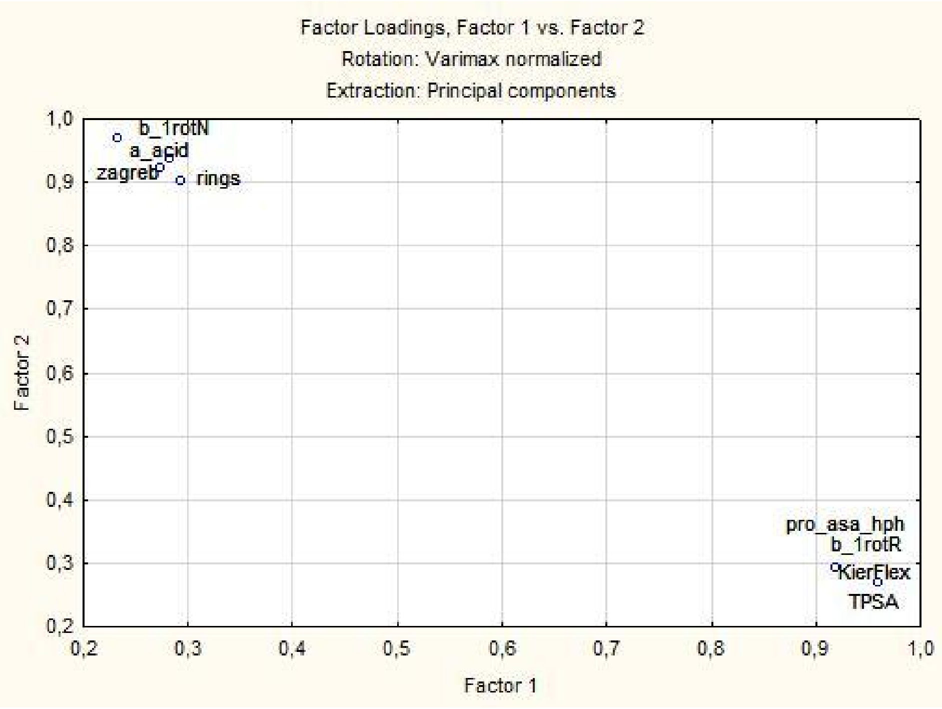

Figure 15. Projection of objects (2D descriptors) on the plane of the first two latent factors.

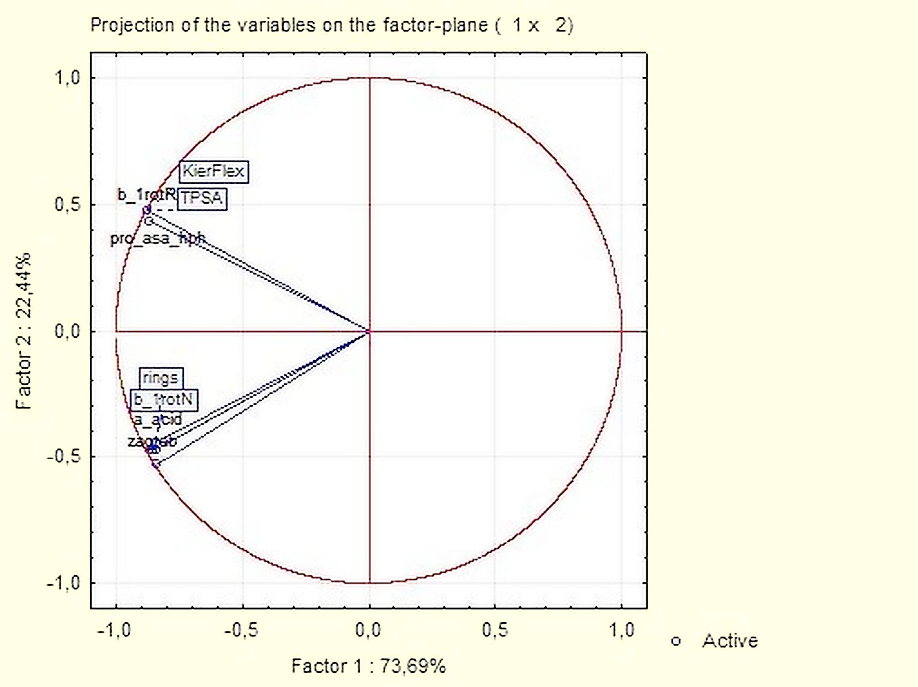

Figure 16. Projection plot of the variables.

This is also confirmed on the biplot presented in Figure 15 and on the projection plot for descriptors, as indicated in Figure 16.

The projection plot for the objects (proteins) proves the separation into two classes as in the case with all descriptors as seen in Figure 17.

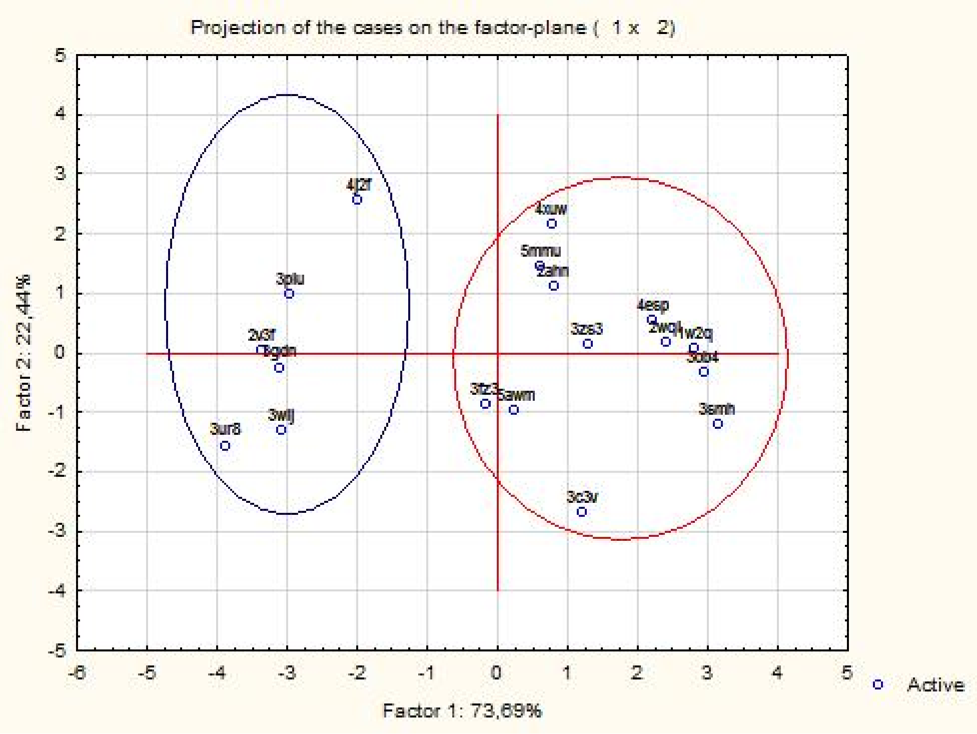

Figure 17. Projection of the cases on the factors.

Table 6. Factor loadings.

Variable	**Factor Loadings (Varimax Normalized) Extraction: Principal Components (Marked Loadings are >0.700000)**	
	Factor 1	**Factor 2**
pro_asa_hph	0.920048	0.292616
a_acid	0.272915	0.923270
b_1rotN	0.282057	0.936526
b_1rotR	0.959852	0.268366
KierFlex	0.959852	0.268366
TPSA	0.959852	0.268366
rings	0.294110	0.903245
zagreb	0.232500	0.968740
Expl. Var %	48.81	47.32

This study relies on the hypothesis that if we could describe in a simple way the proteins with a proper set of 2D molecular descriptors defined as numerical properties that can be calculated from the connection table representation of a molecule (e.g., elements, formal charges and bonds, but not atomic coordinates) it would be beneficial for differentiating and finding a pattern for prediction of allergenic proteins based on 2D.

Our ultimate goal is the development of highly accurate models for the prediction of the allergenicity toward plant proteins. In this context, analyzing the ability and proper choice of each molecular descriptor set to distinguish both forms is crucial. The results suggest that each descriptor set contains exceptional and complementary information to describe the final classification model. Obviously, variable selection methods based on PCA reduction are needed to identify the best descriptors from each descriptor set. These observations support the perspective that the combination of many different classes of chemical descriptors based on physical properties, subdivided surface areas, atom count and bond count descriptors, which are functions of the counts of atoms and bonds that are based on an approximate accessible van der Waals surface or pharmacophore feature descriptors, should be considered when constructing a classification model for the case of chosen proteins of families of these proteins. Using only descriptors from one type may result in a large loss of information and lack of a pattern. These results demonstrate how chemometric tools can provide us with an added layer of key information on such a complicated task, such as allergenicity prediction of food proteins. More effort is needed to validate and test for comprehension of this result; we will go further with a larger data set. In particular, the PCA-based approach has revealed a clear segregation between the groups as well as the obtained similarity pattern obtained by the HCA.

In this paper, we propose a new mapping method for classification of allergenic plant proteins that incorporates a simple scheme based on molecular descriptors. Therefore, the proposed simple model is based on the protein structure.

4. Conclusions

This study demonstrates the potential of exploiting chemometrics methods for separation and prediction between allergenic and non-allergenic food proteins. The complexity of the problem with a food allergy and especially the peanut allergies causing the majority of the annual emergency room admissions due to food allergies and approximately 63%–67% of deaths due to anaphylaxis with allergenicity nowadays is well documented. Our case study on the stated problem shows that generated descriptors could help in discriminating the groups in proteins and within the descriptors as well. Therefore, the workflow for the characterization of molecules could boost the prediction performances of models developed by PCA and HCA by a combination of different descriptors. This study paves the way to predictive abilities of the PCA and HCA models involving classical 2D molecular descriptors without a need for conducting more complicated model studies.

Author Contributions: Conceptualization, M.N.; methodology, M.N.; software, M.N.; validation, M.N. and V.S. formal analysis, M.N. and V.S.; investigation, M.N.; resources, M.N.; data curation, M.N.; writing—original draft preparation, M.N. and V.S.; writing—review and editing, M.N. and V.S.; visualization, M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "Information and Communication Technologies for a Single Digital Market in Science, Education and Security" of the Scientific Research Center, grant number NIS-3317 and National roadmaps for research infrastructures (RIs) grant number NIS-3318.

Acknowledgments: The author M.N. is grateful for the additional support by the project "Information and Communication Technologies for a Single Digital Market in Science, Education and Security" of the Scientific Research Center, NIS-3317 and National roadmaps for research infrastructures (RIs) grant number NIS-3318. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anderson, A.; Shah, S.; Nurruzzaman, F. Increasing anaphylaxis hospitalizations in the first 2 decades of life: New York State, 1990–2006. *Ann. Allergy Asthma Immunol.* **2008**, *101*, 387–393.
2. Branum, A.; Lukacs, S. Food Allergy Among Children in the United States. *Pediatrics* **2009**, *124*, 1549–1555. [CrossRef] [PubMed]
3. Gupta, R.; Kim, J.; Springston, E.; Pongracic, J.; Wang, X.; Holl, J. Development of the chicago food allergy research surveys: Assessing knowledge, attitudes, and beliefs of parents, physicians, and the general public. *BMC Health Serv Res.* **2009**, *9*, 142. [CrossRef] [PubMed]
4. Bock, S.; Munoz-Furlong, A.; Sampson, H.A. Fatalities due to anaphylactic reactions to foods. *J. Allergy Clin. Immunol.* **2001**, *107*, 191–193. [CrossRef] [PubMed]
5. Maleki, S.J.; Schmitt, D.A.; Galeano, M.; Hurlburt, B.K. Comparison of the Digestibility of the Major Peanut Allergens in Thermally Processed Peanuts and in Pure Form. *Foods* **2014**, *3*, 290–303. [CrossRef] [PubMed]
6. Dyer, S.; Nesbit, J.B.; Cabanillas, B.; Cheng, H.; Hurlburt, B.K.; Maleki, S.J. Contribution of chemical modifications and conformational epitopes to ige binding by ara h 3. *Foods* **2018**, *7*, 189. [CrossRef] [PubMed]

7. 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*, 401–407. [CrossRef]
8. Krause, A.; Stoye, J.; Vingron, M. Large scale hierarchical clustering of protein sequences. *BMC* **2005**, *6*, 15–26.
9. Paccanaro, A.; Casbon, J.A.; Saqi, M.A. Spectral clustering of protein sequences. *Nucleic Acids Res.* **2006**, *34*, 1571–1580. [CrossRef] [PubMed]
10. Kelil, A.; Wang, S.; Brzezinski, R.; Fleury, A. CLUSS: Clustering of protein sequences based on a new similarity measure. *BMC Bioinform.* **2007**, *8*, 286. [CrossRef] [PubMed]
11. Yu, C.; Deng, M.; Cheng, S.Y.; Yau, S.C.; He, R.L.; Yau, S.S. Protein space: A natural method for realizing the nature of protein universe. *J. Theor. Biol.* **2013**, *318*, 197–204. [CrossRef] [PubMed]
12. Steinegger, M.; Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **2018**, *9*, 2542. [CrossRef] [PubMed]
13. Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*, 151. [CrossRef] [PubMed]
14. Engel, T. Basic Overview of Chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46*, 2267–2277. [CrossRef] [PubMed]
15. Peña-Castillo, A.; Méndez-Lucio, O.; Owen, J.R.; Martínez-Mayorga, K.; Medina-Franco, J.L. Chemoinformatics in Food Science. In *Applied Chemoinformatics*; Engel, T., Gasteiger, J., Eds.; John Wiley and Sons: Hoboken, NJ, USA, 2020.
16. Massart, D.L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*; John Wiley and Sons: Hoboken, NJ, USA, 1989.
17. Vandeginste, B.; Massart, D.; De Jong, S.; Massaart, D.; Buydens, L. *Handbook of Chemometrics And Qualimetrics: Part B*; Elsevier: Amsterdam, The Netherlands, 1998.
18. Bartholomew, D.J. Principal Components Analysis. In *International Encyclopedia of Education*, 3rd ed.; Peterson, P., Baker, E., Mc Gaw, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2010; pp. 374–377. ISBN 9780080448947. [CrossRef]
19. Guarino, C.; Sciarrillo, R. The identification of allergen proteins in two different varieties of strawberry by two different approaches: Proteomic and western blotting method. *Ann. Agric. Sci.* **2018**, *63*, 181–189. [CrossRef]