Research Article

L. Naneva, M. Nedyalkova*, S. Madurga, F. Mas, V. Simeonov

Applying Discriminant and Cluster Analyses to Separate Allergenic from Non-allergenic Proteins

https://doi.org/10.1515/chem-2019-0045
received December 26, 2018; accepted January 30, 2019.

Abstract: As a result of increased healthcare requirements and the introduction of genetically modified foods, the problem of allergies is becoming a growing health problem. The concept of allergies has prompted the use of new methods such as genomics and proteomics to uncover the nature of allergies. In the present study, a selection of 1400 food proteins was analysed by PLS-DA (Partial Least Square-based Discriminant Analysis) after suitable transformation of structural parameters into uniform vectors. Then, the resulting strings of different length were converted into vectors with equal length by Auto and Cross-Covariance (ACC) analysis. Hierarchical and non-hierarchical (K-means) Cluster Analysis (CA) was also performed in order to reach a certain level of separation within a small training set of plant proteins (16 allergenic and 16 non-allergenic) using a new three-dimensional descriptor based on surface protein properties in combination with amino acid hydrophobicity scales. The novelty of the approach in protein differentiation into allergenic and non-allergenic classes is described in the article.

The general goal of the present study was to show the effectiveness of a traditional chemometric method for classification (PLS–DA) and the options of Cluster Analysis (CA) to separate by multivariate statistical methods allergenic from non-allergenic proteins.

Keywords: amino acid descriptors; ACC transformation; Cluster Analysis (CA); Partial Least Squares-based Discriminant Analysis (PLS-DA)

1 Introduction

Allergies represent one of the most important health problems faced by humanity. Allergic reactions are caused by various food sources such as eggs, soybeans, fruits, vegetables, marine and dairy products [1-5]. The introduction of genetically modified foods have made allergies an even more concerning problem. The term “allergy” was introduced in 1906 by the Austrian pediatrician Clemens Pirquet to indicate the altered reaction in some children injected prophylactically with an anti-infiltrating vaccine. In an allergic response, the body’s reactivity to the effects of certain factors called allergens has been altered or impaired. Allergens provoke the body to produce neutralizing antibodies. Initially, the reaction of the interaction between allergens and antibodies can go unnoticed.

Food allergy is a condition in which the body reacts negatively to food due to a response of the immune system to nutritional protein. Food allergy differs from other bodily reactions to food such as food intolerance, drug intolerance, and toxin-mediated reactions. Food intolerance is the inability of the body to process a nutrient properly, usually due to the lack of an enzyme, but in food allergy, the immune system generates antibody responses to the absorbed food [6]. An allergic reaction occurs when the susceptible organism is exposed to a specific protein. Because the body perceives this protein (allergen) as a threat, it begins to produce T-helper lymphocytes (Th2) that release interleukins. Interleukins increase the production of antibodies called immunoglobulins E (IgE) from B-cells. The body reacts by producing a large amount of these antibodies. The latter binds to mast cells in the blood. Upon reintroduction into the body of the same allergen, it binds to the antibodies located on the mast cells. As a result of the antigen-antibody response, mast

***Corresponding author: M. Nedyalkova,** Faculty of Chemistry and Pharmacy, University of Sofia ”St. Kl. Okhridski”, J. Bourchier Blvd. 1, 1164 Sofia, Bulgaria, E-mail: mici345@yahoo.com
L. Naneva, V. Simeonov: Faculty of Chemistry and Pharmacy, University of Sofia ”St. Kl. Okhridski”, J. Bourchier Blvd. 1, 1164 Sofia, Bulgaria
S. Madurga, F. Mas: Physical Chemistry Unit, Materials Science and Physical Chemistry Department & Research Institute of Theoretical and Computational Chemistry (IQTCUB) of Barcelona University (UB), Barcelona (Catalonia, Spain)

cells release histamine that causes the allergic symptoms: including redness, swelling, and itching [7].

Recognition of allergenic proteins is important because of increasing usage of modified proteins in foods, medicines, household chemicals, and other products. [8]. According to the Food and Agriculture Organization of the United Nations (FAO) and the World Health Organization (WHO), a protein is a potential allergen if it has in its structure 6 to 8 consecutive amino acids or 35% similarity within 80 amino acid residues of already known allergens.

In this study, we describe two methods for predicting allergenicity based on either linear sequence of amino acids or spatial distribution of amino acids on the surface. In the first method, descriptors [9] using Auto-Cross-Covariance (ACC) transformation of protein sequences in universal vectors of the same length [10] are used in a big data set of allergenic proteins. ACC was used for a Structure-Activity-Quantification (QSAR) peptide studies, protein classification and prediction of immunogenicity [11-13]. In the present study, the transformed protein data are used in PLS-DA for reliable classification of allergenic proteins [15,16]

In the second method, the crystallographic structure of the allergenic protein is required. Cluster analysis (hierarchical and non-hierarchical) is here used as classification methodology for separation of allergic from non-allergic proteins.

The aim of this study is to demonstrate the ability of different chemometric methods using amino acid sequence information or spatial distribution of surface amino acids to separate allergic from non-allergic proteins.

2 Datasets and Methods

2.1 Protein datasets

A dataset of 700 food allergens and 700 non-allergens was collected from databases CSL (Central Science Laboratory) (http://allergen.csl.gov.uk), FARRP (Food Allergen Research and Resource Program) (http://www.allergenonline.org) and SDAP (Structural Database of Allergenic Proteins) (http://fermi.utmb.edu/SDAP/sdap_man. html). The non-allergens were selected from the same species using a BLAST (Basic Local Alignment Search Tool) search with 0% identity to allergens at E-value 0.001 [15]. The final set of allergens contained 1400 proteins.

Additionally, a training set of 32 plant proteins was selected for checking the ability of cluster analysis (hierarchical and nonhierarchical mode) to correctly separate proteins into allergenic and non-allergenic classes based on surface protein descriptors. A data set of 16 allergenic proteins related with foods has been selected. Those proteins are classified as allergens by Protein Data Bank (PDB) (https://www.rcsb.org/) or/and by the Structural Database of Allergenic Proteins (SDAP) (http://fermi.utmb.edu/). These allergenic proteins are in one of the following foods: apple, barley, castor bean, cattle, coconut, fungi, legumin, maize, papaya, peach, peanut, olive or soybean. A complementary data set of 16 structures of non allergenic proteins has been also selected from proteins that are constituents of the following foods: apple, barley, castor bean, cattle, maize, papaya, peanut, or soybean. In the description of the proteins, these are not indicated to be either allergenic or related to cancer.

2.2 Protein sequences by E-descriptors and ACC transformation

The E-descriptors for the 20 naturally occurring amino acids, defined by Venkatarajan and Braun [14], were derived by Principal Components Analysis (PCA) of a data matrix consisting of 237 physicochemical properties. The first principal component (E1) reflects the hydrophobicity of the amino acid, the second (E2) reflects its size, the third (E3) reflects its helix-forming propensity, the fourth (E4) correlates with the relative abundance of the amino acid, and the fifth (E5) describes its strand forming propensity. In the present study, the five E-descriptors were used to describe the protein sequences.

The values for the five E-descriptors used in the present study to describe the protein sequences are given in Table 1. To make the length of the proteins uniform, an Auto- and Cross Covariance (ACC) transformation was used [10]. Auto-covariance Ajj(l) and cross-covariance Cjk(l) were calculated according to the following equations:

$$A_{jj}(l) = \sum_{i}^{n-l} \frac{E_{j,i} \times E_{j,i+1}}{n-l} \qquad C_{jk}(l) = \sum_{i}^{n-l} \frac{E_{j,i} \times E_{k,i+1}}{n-l}$$

where indices j and k refer to the E-descriptors (j = 1-5, k = 1-5, $j \neq k$), n is the number of amino acids in a sequence, index i points the amino acid position (i = 1, 2,.. , n) and l is the lag (l = 1, 2, ..., L). Short lags (L = 8) were chosen, as only the influence of close amino acid proximity was investigated. The subsets of antigens and non-antigens were transformed into matrices with 200 variables (5x5x8) each.

Table 1: E-descriptors of amino acids.

amino acid	*E1*	E2	E3	E4	E5
Alanine (A)	0.008	0.134	-0.475	-0.039	0.181
Arginine (R)	0.171	-0.361	0.107	-0.258	-0.364
Asparagine (N)	0.255	0.038	0.117	0.118	-0.055
Aspartic acid (D)	0.303	-0.057	-0.014	0.225	0.156
Cysteine (C)	-0.132	0.174	0.07	0.565	-0.374
Glutamate (Q)	0.149	-0.184	0.03	0.035	-0.112
Glutamic acid (E)	0.221	-0.28	-0.315	0.157	0.303
Glycine (G)	0.218	0.562	-0.024	0.018	0.106
Histidine (H)	0.023	-0.177	0.041	0.28	-0.021
Isoleucine (I)	-0.353	0.071	-0.088	-0.195	-0.107
Leucine (L)	-0.267	0.018	-0.265	-0.274	0.206
Lysine (K)	0.243	-0.339	-0.044	-0.325	-0.027
Methionine (M)	-0.239	-0.141	-0.155	0.321	0.077
Phenylalanine (F)	-0.329	-0.023	0.072	-0.002	0.208
Proline (P)	0.173	0.286	0.407	-0.215	0.384
Serine (S)	0.199	0.238	-0.015	-0.068	-0.196
Threonine (T)	0.068	0.147	-0.015	-0.132	-0.274
Tryptophan (W)	-0.296	-0.186	0.389	0.083	0.297
Tyrosine (Y)	-0.141	-0.057	0.425	-0.096	-0.091
Valine (V)	-0.274	0.136	-0.187	-0.196	-0.299

2.3 Surface Descriptor for proteins

In the present study, a new type of molecular descriptor based on surface properties is defined. This descriptor is based on the characterization of the environment of any type of amino acid present on the surface of the proteins. The idea of this descriptor is to be able to classify proteins in terms of surface properties instead of global composition of the protein. In order to be able to characterize the amino acids chemically (polar, non-polar or charged) a set of hydrophobic scales is used [11].

The descriptor, $d_{s,r}$, of scale, *s*, for the residue *r* (any of the 20 amino acids), is obtained from

$$d_{s,r} = \frac{1}{n_{sc}(r)} \sum_{r'=1}^{n_{sc}(r)} s_{r'}$$

where nsc (r) contains all the surface contacts established between *r* and *r'* residues. sr is the value of the hydrophobicity scale for the residue *r'*.

A surface contact is defined when two residues are on the surface with a distance of separation between alpha carbons less than 8 angstroms. To determine that a residue is on the surface, the Residue Depth module of the BioPython package is used to determine its distance with respect to the surface. An average surface distance below of 2.5 angstroms is used to determine that a residue belongs to the protein surface. If a residue *r* is not present in the surface a value of 0 is assigned to $d_{s,r}$ descriptor.

2.4 Selection of Hydrophobicity Scales

In order to calculate surface descriptors for any amino acid, *r*, it must have different scales, *s*. These scales have to be able to separate the amino acids according to the nonpolar, polar or ionic character of their side chains. For that reason, an initial selection of 20 experimentally determined hydrophobicity scales were used [11]. These scales are mainly derived from partition coefficients found experimentally from measurements of amino acid solubility in water and in organic solvents. Depending on the measuring technique, organic solvent, chromatographic column and experimental procedure, different hydrophobicity scales are obtained.

2.5 Partial Least Squares–based Discriminant Analysis (PLS-DA)

Discriminant Analysis (DA) is a method for data classification based on a linear combination of explanatory variables (Ligand-based design manual, Sybyl [12]). Partial Least Squares (PLS)–based DA was used in the present study. PLS forms new X variables, called Principal Components (PC), as linear combinations of old variables, and then uses them to predict class membership. The optimum number of PCs was selected by adding components until the next added component explained less than 10% of the variance. In the present study, PLS-DA was performed by Soft Independent Modeling of Class Analogy (SIMCA) P-8.0 [13].

2.6 Receiver Operating Characteristics (ROC) statistics

The predictive ability of the derived final model was assessed by Receiver Operating Characteristic (ROC) statistics [14]. Four outcomes are possible in ROC-statistics: true positives (TP, true binders predicted as binders); true negatives (TN, true non-binders predicted as non-binders); false positives (FP, true non-binders predicted as binders); and false negatives (FN, true binders predicted as non-binders). Three classification functions were used in the present study: sensitivity (true positives/total positives), specificity (true negatives/total negatives) and accuracy (true positives and negatives/total). Sensitivity, specificity and accuracy were calculated at different thresholds and the area under the ROC curve (sensitivity/1-spesificity) (ROC) was calculated. AROC is a quantitative measure of predictive ability and varies from 0.5 for random prediction to 1.0 for perfect prediction.

2.7 Model validation

The models derived in the present study were validated by Cross-Validation (CV) and by an external test set. CV is a procedure for testing the predictive ability of models. The training set is divided into several groups with approximately equal numbers of members in each group. One group is defined as a test set and the rest form a new training set. The training set is used to derive a model, the test set, in order to test its predictivity. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

The derived models are validated also by an external test set containing allergens and non-allergens not included in the training set. The predictive ability of the models was estimated by the parameter's sensitivity, specificity, accuracy and AROC.

2.8 Cluster analysis for protein separation

Cluster Analysis (CA) is a general term to indicate series of calculation procedures used for classification and grouping of objects or variables describing the objects [17,18]. The major goal of CA is 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. In hierarchical clustering, the number of groups is preliminarily unknown since the non-hierarchical clustering as a supervised pattern recognition method requires *a priori* determination of the number of groups for data interpretation.

Each object in the data set could be presented by an object vector *Xi*. In order to interpret the data structure a similarity measure should be introduced like Euclidean distances [17]. Unwanted data rotations in the data structure are avoided by different data transformations the most applied one being the autoscaling or z – transformation [17]. The graphical output of the analysis is known as dendrogram plot.

Next important step after autoscaling and distance determination is the linkage algorithm. There are many options but hierarchical clustering relays often on Ward's method of linkage and the non-hierarchical – on K – means mode.

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

Ethical approval: The conducted research is not related to either human or animal use.

3 Results and Discussion

In order to derive a preliminary model for allergenicity prediction, a small set of 120 allergens and 120 non-allergens was compiled randomly from the set of 1400 proteins used in the study. The structure of proteins was described by the five E-descriptors and each protein was transformed into a string of 200 variables, applying ACC-transformation, as described in "Datasets and Methods". The two-class matrix consisting of 240 proteins and 200 variables was subjected to PLS-DA with numbers of principal components varying from 1 to 4. The models were evaluated according to sensitivity, specificity and accuracy at threshold 0.5. The area under the ROC curve (AROC) also was recorded. The results are shown in Figure 1.

The preliminary model for allergenicity prediction is shown in Table 2. The assignment of ACC variables is as follows: the first digit corresponds to the E-descriptor for the *i*-th amino acid in the protein; the second digit corresponds to the E-descriptor for the *j*-th amino acid; and the third digit shows the lag. For example, ACC121 assigns the sum of ACC values calculated using E1 and E2 scales with a lag of 1 (first and second, second and third, third and fourth, etc. The variables in the model are ordered by their Variables Importance in Projection (VIP) values. Variables with VIP > 2.0 are essential to the model. Nineteen variables (32.5%) in the model have a VIP > 2.0. To differentiate between the most important, the threshold for VIP was increased to 1.500. Variables that meet this threshold include ACC121, ACC447, ACC444, ACC 228, ACC222, ACC141, ACC 243 and ACC 246. ACC444, ACC228, ACC222, ACC141, ACC243 have positive coefficients, while ACC121, ACC447 and ACC246 have negative ones. This means that proteins having negative ACC121, ACC447,

Table 2: VIP values and coefficients of the preliminary model for allergenicity prediction. The constant of the model is 0.998. Variables with VIP > 2.0 and coefficients > |0.100| are given in bold.

Variable	VIP	coef.	Variable	VIP	coef.	Variable	VIP	coef.
ACC121	**2.759**	**-0.178**	ACC518	1.635	-0.078	ACC117	1.439	0.003
ACC447	**2.554**	**-0.162**	ACC344	1.639	-0.078	ACC414	1.436	-0.036
ACC444	**2.448**	**0.32**	ACC335	1.605	-0.097	ACC451	1.426	-0.047
ACC228	**2.289**	**0.52**	ACC427	1.589	0.001	ACC255	1.404	-0.055
ACC222	**2.198**	**0.84**	ACC353	1.587	-0.096	ACC245	1.379	-0.024
ACC141	**2.136**	**0.65**	ACC118	1.583	-0.071	ACC146	1.375	0.032
ACC243	**2.115**	**0.099**	ACC147	1.566	-0.031	ACC252	1.369	0.029
ACC246	**2.093**	**-0.142**	ACC442	1.552	0.030	ACC244	1.369	-0.051
ACC323	1.961	0.050	ACC523	1.543	0.092	ACC242	1.367	-0.045
ACC122	1.929	-0.083	ACC342	1.519	-0.073	ACC311	1.361	-0.020
ACC144	1.897	0.083	ACC418	1.492	-0.085	ACC526	1.357	0.043
ACC128	1.799	-0.115	ACC438	1.484	-0.081	ACC345	1.338	0.002
ACC211	1.748	-0.125	ACC114	1.482	-0.009	ACC426	1.325	-0.017
ACC544	1.716	0.116	ACC251	1.468	-0.014	ACC124	1.322	-0.021
ACC324	1.710	0.103	ACC443	1.466	0.064	ACC358	1.264	0.007

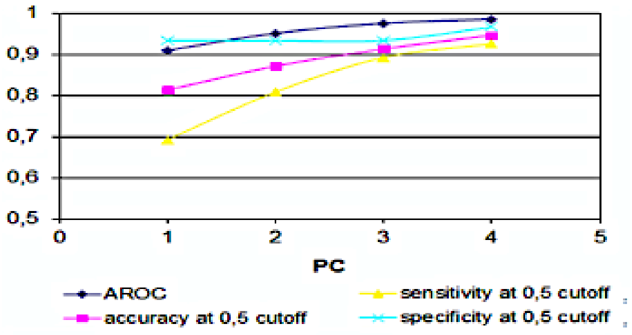

Figure 1: Sensitivity, specificity and accuracy at threshold 0.5, and AROC for the preliminary model for allergenicity prediction with different number of PCs.

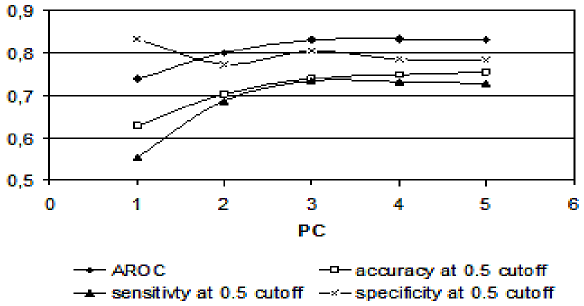

Figure 2: Sensitivity, specificity and accuracy at threshold 0.5, and AROC for the extended model for allergenicity prediction with different number of PCs.

ACC246, and positive ACC444, ACC228, ACC222, ACC141 and ACC243 are likely to act as allergens.

Further, the preliminary model was used to predict the allergenicity of an external test set of 580 allergens and 580 non-allergens. It recognized 68% of the allergens and 77% of the non-allergens with 73% total accuracy at threshold 0.5. The AROC value was 0.785.

Encouraged by the good predictability of the preliminary model, we derived an extended model for allergenicity prediction based on 700 food allergens and 700 non-allergens. The structure of proteins was described by the three z-descriptors and ACC-transformed into strings of 200 variables. The two-class matrix consisting of 1,400 proteins and 200 variables was subjected to PLS-DA with number of PC varying from 1 to 4. The models were evaluated according to sensitivity, specificity and accuracy at threshold 0.5. AROC also was recorded. The results are shown in Figure 2.

The results showed that the highest values of the parameters are obtained by three PCs. The model with 4 PCs and the VIP-values of the variables are shown in Figure 2. Variables that have a VIP> 2.0 have the greatest significance for the model and coincide with the variables

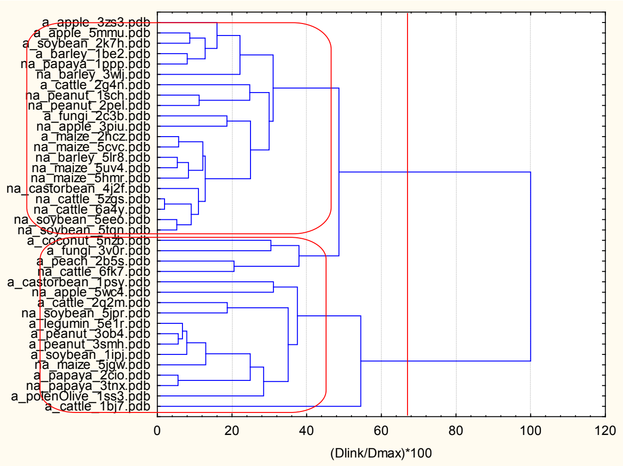

Figure 3: Hierarchical dendrogram for separation between allergenic (a) from non-allergenic (na) proteins.

from the original model. The concept of the variables found in the preliminary model is confirmed here.

3.1 Cluster Analysis (CA) as separation tool for allergenicity of using a surface protein descriptor

In order to check the option for separation of proteins into allergenicity and non-allergenicity classes using the surface properties of proteins, a data set of 32 food proteins was prepared (16 allergenic and 16 non-allergenic). A new set of descriptors, $d_{s,r}$ was created based on the reference data for hydrophobicity of the amino acid components.

Values of 20 experimentally determined hydrophobicity scales [11] were used after the transformation. Thus, a training set of 38 proteins (19 allergic and 19 non-allergic) described by 98 most significant descriptors out of totally 400 was treated by Cluster Analysis (CA). The variable reduction was performed by the use of Principal Components Analysis (PCA).

In Figure 3 the hierarchical dendrogram for separation of the proteins into classes of allergenicity and non-allergenicity is shown.

As seen in Figure 3 the separation between allergenic (a) and non-allergenic (na) classes of food proteins is well expressed. Two major clusters are formed:

- K1 (lower left) with a total of 17 members, including 12 (a) and 5 (na), which could be conditionally named *allergenic protein cluster*. Correctly classified are 12 allergic proteins out of a total of 17 members (70%). Five out of a total of 17 members (30%) were wrongly classified as allergenic proteins;
- K2 (upper left) with a total of 21 members, including 14 (na) and 7 (a), which could be conditionally named non-*allergenic protein cluster*. Correctly classified are 14 non allergenic proteins out of 21 members (67%), while 7 out of 21 members (33%) were wrongly classified non-allergenic proteins.

The non-hierarchical clustering (K-means mode) gave the same results after checking an *a prior*i stated hypothesis of separation of all 38 objects (proteins) into two clusters.

The results obtained by Cluster Analysis (CA) are of the same level of efficiency reached by the other classification approach, PLS–DA.

4 Conclusions

Allergenicity of food proteins is a crucial problem associated with the widespread usage of new foods, supplements and herbs, many of which may be of genetically modified origin. Allergenicity is a subtle, non-linearly coded property. Most of the existing methods for allergenicity prediction are based on structural similarities of novel proteins to known allergens. Thus, the identification of a novel, structurally diverse allergens could not be predicted by these methods.

In the present study, we propose an alignment-free method for allergenicity prediction, based on the amino acid principal properties of hydrophobicity, size and electronic structure. Proteins are transformed into uniform vectors and analyzed by PLS-DA. Initially, a preliminary model was derived based on a small set of 120 allergenics and 120 non-allergenics. The model was tested by Cross-Validation and external test set and recognized correctly 73% of the proteins from the external test set. Then, the dataset was extended to 1,400 proteins (700 allergenics and 700 non-allergenics) and a new model was derived. The Cross-Validation study showed that the extended model is able to identify correctly 70% of the tested proteins.

The food allergens involved in the present study have diverse structure, composition and origin, which imply great variance in the set. By increasing the number of proteins in the training set, the number of PCs needed to explain this variance was increased. In the small initial set used to derive the preliminary model, two PCs were sufficient to obtain a model with good predictive ability. In the extended set of proteins used in the extended model, it was necessary to include a third PC. The model with 4 PCs had the highest predictive ability.

Both models point to the importance of variables ACC121, ACC447 and ACC246. These variables account for the electronic structure of amino acids located in close proximity but not next to each other. In addition, hierarchical and non-hierarchical (K-means) clustering using a surface protein descriptor reached an important level of separation within a small training set of allergenic and non-allergenic food proteins.

These results once again shows that the allergenicity is a hidden, complex property, depending on many factors, some of which are encoded in the primary structure of proteins and others in the spatial distribution of amino acids on the protein surface.

Acknowledgement: Author M. Nedyalkova is grateful to the National Scientific Program ICT in SES, financed by the Ministry of Education and Science. The author M.Nedyalkova is gratefully acknowledged to the L'Oréal Program for Woman in Science. Financial support from Generalitat de Catalunya (Grant 2017SGR1033) and Spanish Structures of Excellence María de Maeztu program through grant MDM- 2017–0767 is fully acknowledged.

Conflict of interest: Authors declare no conflict of interest.

References

[1] Sampson H.A., Food allergy. Part 2: diagnosis and management, J. Allergy Clin. Immunol., 1999, 103, 981-989.

[2] Sampson H.A., Food allergy. Part 1: immunopathogenesis and clinical disorders, J. Allergy Clin. Immunol., 1999, 103, 717-728.

[3] Sampson H.A., Food allergy: when mucosal immunity goes wrong, J. Allergy Clin. Immunol., 2005, 115, 139-141.

[4] Vanekkrebitz M., Hoffmannsommergruber K., Machado M.L.D., Susani M., Ebner C., Kraft D., et al., Cloning and Sequencing of Mal d 1, the Major Allergen from Apple (Malus domestica), and Its Immunological Relationship to Bet v 1, the Major Birch Pollen Allergen, Biochem. Biophys. Res. Commun., 1995, 214, 538-551.

[5] Scheurer S., Son D.Y., Boehm M., Karamloo F., Franke S., Hoffmann A., et al., Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen, Mol. Immunol., 1999, 36, 155-167.

[6] Glaspole I.N., de Leon M.P., Rolland J.M., O'Hehir R.E., Characterization of the T-cell epitopes of a major peanut allergen, Ara h 2, Allergy, 2005, 60, 35-40.

[7-8] Fitch W.L., McGregor M., Katritzky A.R., Lomaka A., Petrukhin R., Karelson M., Prediction of Ultraviolet Spectral Absorbance Using Quantitative Structure–Property Relationships, J. Chem. Inf. Comput. Sci., 2002, 42, 830-840.

[9] Venkatarajan M.S., Braun W., New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties, J. Mol. Model., 2001, 7, 445-453.

[10] Nyström Å., Andersson P.M., Lundstedt T., Multivariate Data Analysis of Topographically Modified α-Melanotropin Analogues using Auto and Cross Auto Covariances (ACC), Quant. Struct.-Act. Relat., 2000, 19, 264-269.

[11] Palecz B., Enthalpic Homogeneous Pair Interaction Coefficients of l-α-Amino Acids as a Hydrophobicity Parameter of Amino Acid Side Chains, J. Am. Chem. Soc., 2002, 124, 6003–6008

[12] Eriksson L., Umetrics, Multi- and megavariate data analysis : basic principles and applications, MKS Umetrics, 2013.

[13] SIMCA-P 8.0. Umetrics UK Ltd., Wokingham Road, RG42 1PL, Bracknell, UK.

[14] Bradley A.P., The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, 1997, 30, 1145-1159.

[15] Ivanciuc O., Schein C.H. Braun W., SDAP: database and computational tools for allergenic proteins, Nucleic Acids Res., 2003, 31, 359-362.

[16] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., Basic local alignment search tool, J. Mol. Biol., 1990, 215, 403-410.

[17] Massart D.L., Kaufman L., The interpretation of analytical chemical data by the use of cluster analysis, John Wiley and Sons, 1989.

[18] Vandeginste B., Massart D., De Jong S., Massaart D., Buydens L., Handbook of chemometrics and qualimetrics: Part B, Elsevier, 1998.

[19] Simeonov V., Classification: Encyclopedia of environmetrics, J. Wiley & Sons, 2001.

[20] Feldman R., Sanger J., The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2006.

[21] Leskovec J., Rajaraman A., Ullman J.D., Mining of massive datasets, Cambridge university press, 2014.