Bulgarian Chemical Communications, Volume 46, Number 2(pp. 389 – 396) 2014

Allergenicity prediction by partial least squares-based discriminant analysis

L. H. Naneva[1], I. D. Dimitrov[2], I. P. Bangov[1], I. A. Doytchinova[2*]

[1]*Konstantin Preslavski Shumen University, Faculty of Natural Sciences, General Chemistry Chair, 115 Universitetska Str., 9712 Shumen, Bulgaria*

[2]*Medical University of Sofia, Faculty of Pharmacy, 2 Dunav Str., 1000 Sofia, Bulgaria*

Received November 30, 2012; Revised February 5, 2013

Allergenicity of food proteins is a crucial problem associated with the widespread usage of new foods, supplements and herbs, many of them having known or unknown genetically modified origin. Allergenicity is a subtle, non-linearly coded property. Most of the existing methods for allergenicity prediction are based on structural similarity of novel proteins to known allergens. Thus, the identification of a novel, structurally diverse allergen 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 as hydrophobicity, size and electronic structure. Proteins are transformed into uniform vectors and analyzed by PLS-based discriminant analysis. The preliminary model derived on the basis of a small set of 120 allergens and 120 non-allergens identified correctly 73% of the proteins included in the external test set of 1,164 allergens and non-allergens. The extended model based on a set of 1,404 proteins (702 allergens and 702 non-allergens) showed 70% accuracy in the cross-validations.

Key words: allergens, non-allergens, *z*-descriptors, auto- and cross-covariance, discriminant analysis, PLS

INTRODUCTION

Allergy is a growing health problem of modern life. Food allergies affect 10 – 15% of infants and young children [1]. They are caused by different sources: milk, eggs, peanuts, soy, shellfish, fruits, etc. [2-4]. Allergy involves complex reactions to external factors that contribute to the development of diseases characterized by symptoms such as rhinitis, asthma, atopic dermatitis, skin sensitization. In some cases, severe reactions such as acute and fatal anaphylactic shock may occur.

The term "allergy" was introduced in 1906 by the Austrian pediatrician Clemens von Pirquet to denote the modified reaction to smallpox vaccine [5]. Allergy is an altered capacity of the body to react to a foreign substance called allergen. When potential allergens enter the body for the first time, allergen-specific IgE antibodies are produced, which stay around long after the initial allergen is cleared from the body. Most of the antibodies are caught by Fcε receptors, which are IgE-specific receptors that are exposed on the surface of mast cells, basophils and activated eosinophils. These cells are then primed to react the next time the body encounters the allergen. They release stored mediators, which give rise to inflammation and tissue damage causing a variety of symptoms [6-9].

Although there is no consensus on the structure of the allergen, the United Nations Food and Agriculture Organization (FAO) and World Health Organization (WHO) have developed *Codex alimentarius* guidelines for assessing the potential allergenicity of novel proteins [10-11]. According to these guidelines, protein is a potential allergen, if it has an identity of 6 to 8 contiguous amino acids or greater than 35% similarity within a window of 80 amino acids when compared to known allergens.

Currently, two bioinformatic approaches exist for allergen prediction. The first approach follows the guidelines of FAO and WHO and searches for sequence similarity. There are databases that contain extensive information on known allergens, which are used for sequence similarity search. Such databases are Structural Database of Allergenic Proteins (SDAP) [12], Allermatch [13] and AllerTool [14]. This approach has a good allergen prediction, but generates a large number of false allergens. Moreover, the discovery of new structurally different allergens is limited by the lack of similarity to already known allergens.

The second approach is based on identification of linear motifs for allergenicity. The motif is a sequence of amino acids responsible for a particular activity of the protein. Stadler and Stadler (2003) defined 52 allergenic motifs by comparing

* To whom all correspondence should be sent:
E-mail: doytchinova@gmail.com

allergens to non-allergens [15]. Li and colleagues (2004) identified motifs for allergenicity using clustering of known allergens by hidden Markov model (HMM) [16]. Bjorklund and colleagues (2005) developed a method for identifying allergens by detecting allergenic peptides (allergen-representative peptides, ARP) [17]. AlgPred is a server for predicting allergenic protein that combines four motif search methods: support vector machines (SVM), MEME/MAST program, IgE epitopes and ARP [18].

Both approaches are based on the assumption that allergenicity is a linearly encoded property. To act as an allergen, a protein must contain epitopes for both Th2 cells and B lymphocytes [7]. Epitope is part of the protein that interacts with another protein. The epitopes for Th2 are linear, but the B-cell epitopes are non-linear, conformational patches on the protein surface. Obviously, allergenicity, like immunogenicity and antigenicity, is a property coded linearly as well as nonlinearly. Therefore, the alignment-based approaches are not able to identify such property in an unambiguous manner.

In the present study, we develop and validate an alignment-free method for allergenicity prediction, based on the principal amino acid properties as hydrophobicity, size and electronic properties. Partial least squares-based discriminant analysis is used to develop models for food allergenicity prediction. The models were validated by internal and external test sets of allergens and non-allergens.

DATASETS AND METHODS

Allergens and non-allergens

A dataset of 702 food allergens and 702 non-allergens was collected from the databases CSL (Central Science Laboratory) [19], FARRP (Food Allergen Research and Resource Program) [20] and SDAP (Structural Database of Allergenic Proteins) [21]. The non-allergens were selected from the same species using a BLAST search with 0% identity to allergens at E-value 0.001.

Descriptors of the protein structures

The *z*-descriptors describe principal properties of amino acids. They are derived by applying principal component analysis on a set of 29 molecular properties of amino acids [22]. The first principal component (1PC), named z_1, is dominated by the hydrophobicity of amino acids. The second principal component (2PC), named z_2, relates best to amino acid size. The third principal component (3PC), named z_3, explains the electronic properties of amino acids. The scores of these components define the set of *z*-descriptors for each amino acid. In the present study, the three *z*-descriptors were used to describe the amino acid sequences of allergens and non-allergens.

The proteins used in the study were of different length. In order to convert them into uniform vectors, the method of auto- and cross-covariance (auto- and cross-covariance, ACC) transformation was used [23]. Auto-covariance (A_{jj}) and cross-covariance (C_{jk}) were calculated by the following formulas:

$$A_{jj}(L) = \sum_{i}^{n-L} \frac{Z_{j,i} \times Z_{j,i+L}}{n-L}$$

$$C_{jk}(L) = \sum_{i}^{n-L} \frac{Z_{j,i} \times Z_{k,i+L}}{n-L}$$

The index *j* refers to the *z*-descriptors (*j* = 1, 2, 3); the index *i* indicates the position of amino acid (*i* = 1, 2, 3 ... *n*); *n* is the number of amino acids in protein; *l* is the lag (*L* = 1, 2, ..., *l*). Lag is the length of the frame of contiguous amino acids, for which A_{jj} and C_{jk} are calculated. In the present study, a short lag (*L* = 5) was chosen, as the influence of neighboring amino acids was investigated. Each protein was transformed into a string of 45 elements (3^2 x 5).

Partial least squares-based discriminant analysis (PLS-DA)

The discriminant analysis (DA) is a method for data classification based on a linear combination of explanatory variables [24]. Partial least squares (PLS)-based DA was used in the present study. PLS forms new X variables named 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 predictive ability of the model increases. In the present study, PLS-DA was performed by SIMCA P-8.0 [24].

The projection of the *i*-th protein on the plane formed by two PSs is called score. Proteins with similar descriptors are projected close to each other and form a cluster. The loading of the *i*-th descriptor on a PC equals cos α, where α is the angle between the axis of descriptor X_i and the plane formed by two PCs. As more distant is a descriptor from the origin, as higher loading has this descriptor on the corresponding PC.

Receiver Operating Characteristics (ROC) statistics

The predictive ability of the derived final model was assessed by Receiver Operating Characteristic (*ROC*) statistics [25]. 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*) (A_{ROC}) was calculated. A_{ROC} is a quantitative measure of predictive ability and varies from 0.5 for random prediction to 1.0 for perfect prediction.

Variable influence on projection (VIP)

The parameter *VIP* (variable influence on projection) was introduced by Wold in 1993 [26] to describe the importance of each independent variable on the dependent one. It is calculated by the following formula:

$$VIP_{ak} = \sqrt{\left(\sum_{a=1}^{A}(w_{ak}^2(SSY_{a-1} - SSY_a))\frac{K}{SSY_0 - SSY_A}\right)}$$

where w_{ak} is the weight (coefficient) of the variable k on the component a, and SSY_a is the explained variance of Y by the component a. Variables with *VIP*, greater than 1, are the most relevant for explaining Y. *VIP* parameters are calculated by SIMCA-P 8.0.

Model validation

The models derived in the present study were validated by cross-validation and by external test set. The cross-validation (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 – 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 external test set containing allergens and non-allergens not included in the training set. The predictive ability of the models was estimated by the parameters *sensitivity*, *specificity*, *accuracy* and A_{ROC}.

RESULTS

Preliminary model for allergenicity prediction

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 1,440 proteins used in the study. The structure of proteins was described by the three *z*-descriptors and each protein was transformed into a string of 45 variables, applying ACC-transformation, as described in "Materials and Methods". The two-class matrix consisting of 240 proteins and 45 variables was subjected to PLS-DA with varying number of principal components from 1 to 5. The models were evaluated using the parameters *sensitivity*, *specificity* and *accuracy* at threshold 0.5. The area under the curve A_{ROC} also was recorded. The results are shown in Figure 1.

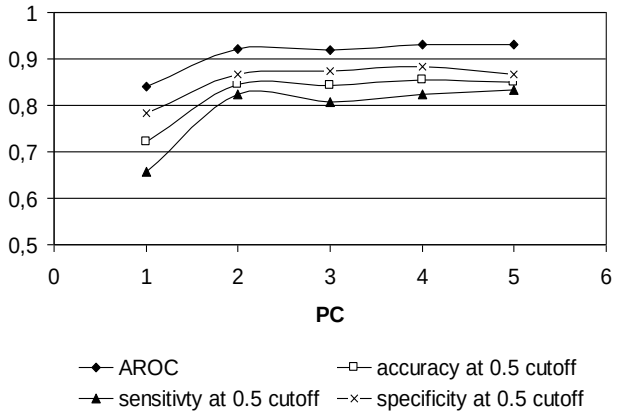


Fig. 1. *Sensitivity*, *specificity* and *accuracy* at threshold 0.5, and A_{ROC} for the preliminary models for allergenicity prediction with different number of PCs.

The results showed that the addition of a second PC significantly increases all parameters of the model. Further addition of PCs initially decreases slightly the parameters, and then increases them slightly. Thus, two PCs was the optimal number of PCs for this model.

The preliminary model for allergenicity prediction is shown in Table 1. The assignment of ACC variables is as follows: the first digit corresponds to the number of *z*-descriptor for the *i*-th amino acid in the protein; the second digit corresponds to the number of *z*-descriptor for the *j*-th amino acid; the third digit shows the lag. For

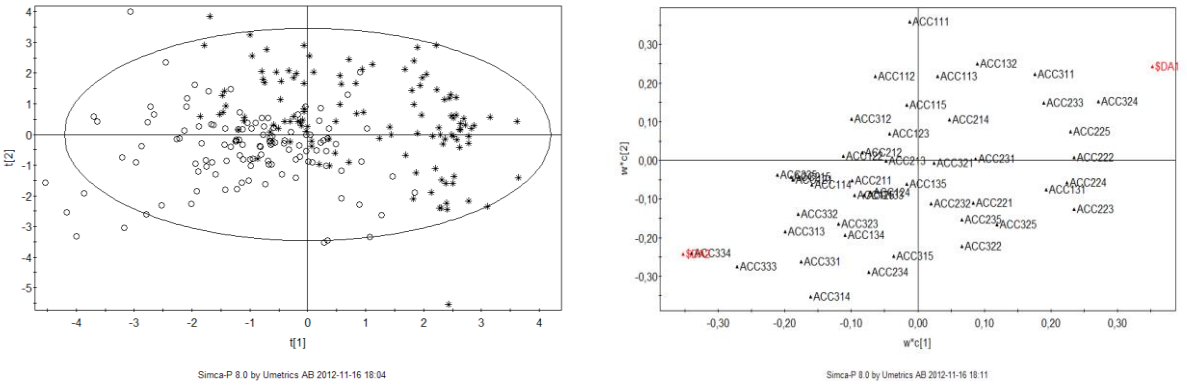
Table 1 .VIP values and coefficients of the preliminary model for allergenicity prediction. The constant of the model is 0.998. Variables with VIP > 1.5 and coefficients > |0.100| are given in bold.

variable	*VIP*	*coefficient*	*variable*	*VIP*	*coefficient*	*variable*	*VIP*	*coefficient*
ACC334	**2.003**	**-0.178**	ACC215	1.121	-0.078	ACC221	0.693	0.003
ACC333	**1.656**	**-0.162**	ACC121	1.101	-0.078	ACC122	0.685	-0.036
ACC324	**1.586**	**0.132**	ACC332	1.061	-0.097	ACC211	0.575	-0.047
ACC223	1.578	0.052	ACC325	1.011	0.001	ACC125	0.573	-0.055
ACC222	1.413	0.084	ACC234	0.950	-0.096	ACC212	0.526	-0.024
ACC224	1.413	0.065	ACC114	0.928	-0.071	ACC231	0.520	0.032
ACC225	1.344	0.099	ACC322	0.924	-0.031	ACC115	0.510	0.029
ACC314	1.315	-0.142	ACC112	0.896	0.030	ACC133	0.503	-0.051
ACC131	1.252	0.050	ACC132	0.867	0.092	ACC124	0.447	-0.045
ACC335	1.248	-0.083	ACC315	0.793	-0.073	ACC232	0.418	-0.020
ACC111	1.210	0.083	ACC134	0.792	-0.085	ACC214	0.393	0.043
ACC313	1.200	-0.115	ACC323	0.779	-0.081	ACC123	0.381	0.002
ACC331	1.185	-0.125	ACC312	0.770	-0.009	ACC213	0.284	-0.017
ACC311	1.128	0.116	ACC235	0.725	-0.014	ACC135	0.201	-0.021
ACC233	1.123	0.103	ACC113	0.699	0.064	ACC321	0.153	0.007

example, ACC324 assigns the sum of ACC values calculated as z3*z2 for each pair amino acids with lag 4 (first and fourth, second and fifth, third and sixth, etc.). The variables in the model are ordered according to their VIP values. Variables with VIP > 1 are essential to the model. Nineteen variables (42%) in the model have a VIP > 1. To differentiate between the most important, the threshold for VIP was increased to 1.500. Only four variables have VIP > 1.500 and coefficient > |0.100|. These are ACC334, ACC333 and ACC324. ACC324 has positive coefficient, ACC334 and ACC333 have negative ones. This means that proteins having negative ACC334 and ACC333, and positive ACC324 are likely to act as allergens. Figure 2A shows the scores of the proteins from the initial set, and Figure 2B gives the loadings of ACC variables. The model distinguishes relatively well allergens (top right, Figure 2A) from non-allergens (bottom left), despite the lack of a clear boundary between the two clusters. The variable ACC324 is situated most distantly from the origin in the upper right quadrant close to the allergenicity variable (assigned as DA1), while variables ACC334 and ACC333 variables are situated most distantly from the origin in the lower left quadrant close to the non-allergenicity variable (assigned as DA2). The model was tested for *sensitivity*, *specificity* and *accuracy* at threshold 0.5. It detects 83% of allergens, 87% of non-allergens and 85% correctly identified proteins from the initial set. The A_{ROC} value is 0.922, indicating for the excellent predictivity of the model.

The initial model for allergenicity prediction was cross-validated in 6 groups. The initial set of 120 allergens and 120 non-allergens was divided into 6 subsets of 20 allergens and 20 non-allergens. Five subsets were united in a training set; the sixth subset was a test. The training set was used to derive the model; the test set was used to validate it. The procedure was repeated six times, so any protein acts five times as a trainer and one time – as a tester. The average values for the test subsets from the cross-validation are: 77% *sensitivity*, 79% *specificity* and 78% *accuracy* at threshold 0.5, and A_{ROC} = 0.856. The cross-validation showed that the preliminary model has a good predictive ability, independent of the training set composition.

Further, the preliminary model was used to predict the allergenicity of external test set of 582 allergens and 582 non-allergens. It recognized 68% of the allergens and 77% of the non-allergens with 73% total *accuracy* at threshold 0.5. The A_{ROC} value was 0.785.



A **B**

Fig. 2. Scores (A, allergens are given as stars, non-allergens – as blank circles) and loadings (B) according to the preliminary model for allergenicity prediction.

Extended model for allergenicity prediction

Encouraged by the good predictivity of the preliminary model, we derived an extended model for allergenicity prediction based on 702 food allergens and 702 non-allergens. The structure of proteins was described by the three *z*-descriptors and ACC-transformed into strings of 45 variables. The two-class matrix consisting of 1,404 proteins and 45 variables was subjected to PLS-DA with varying number of principal components from 1 to 5. The models were evaluated using the parameters *sensitivity*, *specificity* and *accuracy* at threshold 0.5. The area under the curve A_{ROC} also was recorded. The results are shown in Figure 3. The highest values of the parameters are obtained by three PCs.

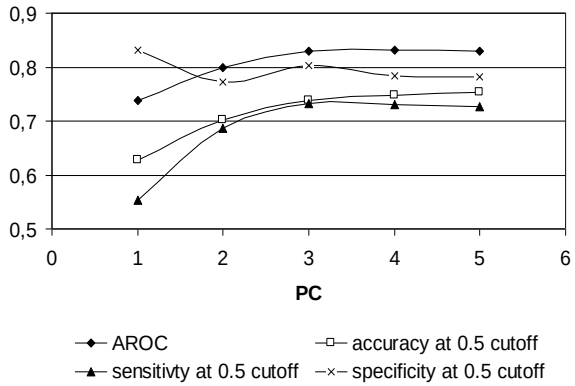
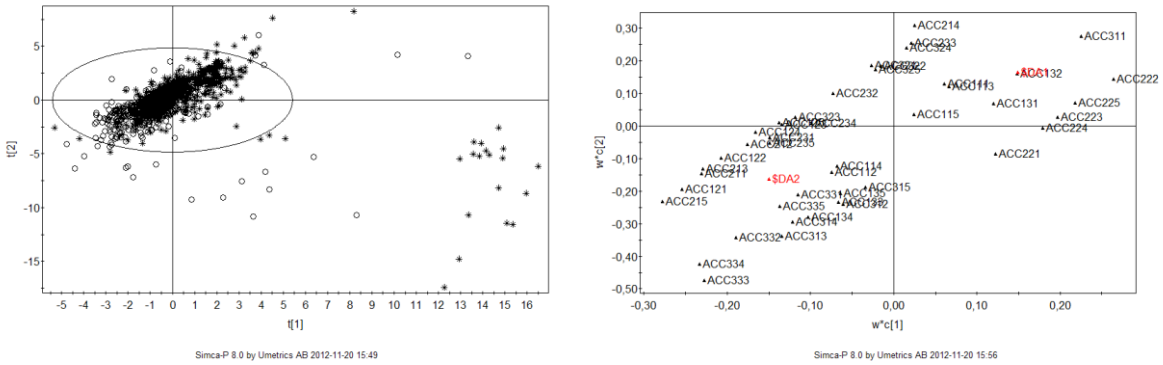


Fig. 3. *Sensitivity*, *specificity* and *accuracy* at threshold 0.5, and A_{ROC} for the extended models for allergenicity prediction with different number of PCs

The model with 3 PCs and the VIP-values of the variables are shown in Table 2. Three variables have VIP > 1.300 and coefficient > |0.100|. These are ACC333, ACC214 and ACC334. One of them has a positive coefficient (ACC214); the other two are negative (ACC333 and ACC334). The significance of variables and ACC333 ACC334, found in the preliminary model is confirmed here. Figure 4 shows the scores and the loading according to the extended model projected on the plane of the first two PCs. There is one outlier (non-allergen with GI: 315113274) (not shown). The model detects 73% of allergens, 80% of non-allergens and 77% correctly identified proteins from the whole set. The A_{ROC} value is 0.830.

The extended model for allergenicity prediction was cross-validated in 7 groups. The whole set of 702 allergens and 702 non-allergens was divided into 7 subsets of 100 or 101 allergens and 100 or 101 non-allergens. Six subsets were united in a training set; the seventh subset was a test. The training set was used to derive the model; the test set was used to validate it. The procedure was repeated seven times, so any protein acts six times as a trainer and one time – as a tester. The average values for the test subsets from the cross-validation are: 60% *sensitivity*, 79% *specificity* and 70% *accuracy* at threshold 0.5, and A_{ROC} = 0.746. The cross-validation showed that the extended model has a lower predictive ability than the preliminary one, but still independent of the training set composition.



A **B**

Fig. 4 . Scores (A, allergens are given as stars, non-allergens – as blank circles) and loadings (B) according to the extended model for allergenicity prediction.

Table 2. VIP values and coefficients of the extended model for allergenicity prediction. The constant of the model is 1.000. Variables with VIP > 1.300 and coefficients > |0.100| are given in bold.

variable	*VIP*	*coefficient*	*variable*	*VIP*	*coefficient*	*variable*	*VIP*	*coefficient*
ACC333	**1.505**	**-0.158**	ACC211	1.044	-0.058	ACC232	0.901	-0.013
ACC214	**1.499**	**0.129**	ACC224	1.032	0.030	ACC231	0.895	-0.058
ACC334	**1.387**	**-0.124**	ACC311	1.015	0.107	ACC323	0.895	-0.038
ACC335	1.236	-0.030	ACC235	1.008	-0.074	ACC325	0.885	0.020
ACC222	1.227	0.092	ACC233	1.000	0.059	ACC113	0.877	0.083
ACC332	1.215	-0.085	ACC324	0.983	0.048	ACC315	0.871	-0.081
ACC215	1.190	-0.093	ACC124	0.979	-0.007	ACC123	0.833	-0.002
ACC313	1.149	-0.125	ACC314	0.964	-0.103	ACC133	0.790	-0.071
ACC121	1.105	-0.077	ACC221	0.963	-0.001	ACC331	0.758	-0.051
ACC225	1.101	0.066	ACC134	0.958	-0.102	ACC115	0.697	0.049
ACC234	1.078	-0.062	ACC212	0.934	-0.020	ACC135	0.697	-0.067
ACC122	1.077	-0.028	ACC111	0.930	0.086	ACC132	0.681	0.074
ACC312	1.054	-0.104	ACC125	0.929	0.007	ACC112	0.656	-0.021
ACC213	1.050	-0.054	ACC321	0.924	0.026	ACC131	0.590	0.022
ACC223	1.049	0.042	ACC322	0.918	0.022	ACC114	0.471	-0.027

DISCUSSION

Allergenicity of food proteins is a crucial problem associated with the widespread usage of new foods, supplements and herbs, many of them having known or unknown genetically modified origin. Allergenicity is a subtle, non-linearly coded property. Most of the existing methods for allergenicity prediction are based on structural similarity of novel proteins to known allergens. Thus, the identification of a novel, structurally diverse allergen 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 as 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 allergens and 120 non-allergens. 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 1,404 proteins (702 allergens and 702 non-allergens) 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 increases the number of PCs needed to explain this variance. 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 3 PCs had the highest predictive ability.

Both models point the importance of the variables ACC333 and ACC334. These variables account for the electronic structure of amino acids located in close proximity but not next to each other. This 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.

Acknowledgements: *This study is supported by the National Research Fund of the Ministry of Education and Science, Bulgaria, Grant 02-1/2009 and Grant FFNNIPO/12-00985.*

REFERENCES

1. W. Cookson, *Nat. Rev. Immunol*., **4**, 978 (2004).
2. H.A. Sampson, *J. Allergy Clin. Immunol*., **103**, 717 (1999).
3. H.A. Sampson, *J. Allergy Clin. Immunol*., **103**, 981 (1999).
4. H.A. Sampson, *J. Allergy Clin. Immunol*., **115**, 139 (2005).
5. S.C. Bukantz, *J. Allergy Clin. Immunol*., **109**, 724 (2002).
6. P.J. Cooper, *Parasite Immunol*., **26**, 455 (2004).
7. C.A. Janeway, P. Travers, M. Walport, J. D. Capra, Immunobiology: the immune system in health and disease, Current Biology Publications, London, 1999.
8. C. Rusznak, R. J. Davies, *BMJ*, **316**, 686 (1998).
9. R.D.J. Huby, R.J. Dearman, I. Kimber, *Toxicological Sci.*, **55**, 235 (2000).
10. FAO/WHO Agriculture and Consumer Protection, Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, Rome, Italy, 2001.
11. FAO/WHO Codex Alimentarius Commission. Joint FAO/WHO Food Standards Programme, Rome, Italy, 2003.
12. O. Ivanciuc, C. H. Schein, W. Braun, *Nucleic Acids Res.,* **31**, 359 (2003).
13. M.W.E.J. Fiers, G.A. Kleter, H. Nijland, A.A.C.M. Peijnenburg, J.P. Nap, R.C.H.J. van Ham, *BMC Bioinformatics*, **5,** 133 (2004).
14. Z. H. Zhang, J.L. Koh, G.L. Zhang, K.H. Choo, M.T. Tammi, J.C. Tong, *Bioinformatics*, **23**, 504 (2007).
15. M.B. Stadler, B.M. Stadler, *FASEB J*., **17**, 1141 (2003).
16. K.B. Li, P. Isaac, P. Krishnan, *Bioinformatic*s, 20, 2572 (2004).
17. A.K. Björklund, D Soeria-Atmadja, A Zorzet, U Hammerling, MG Gustafsson, Bioinformatics, 21, 39 (2005).
18. S. Saha, G. P. S. Raghava, Nucleic Acids Res., 34, W202 (2006).
19. http://allergen.csl.gov.uk
20. http://www.allergenonline.org
21. http://fermi.utmb.edu/SDAP/sdap_man. html
22. S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, J. *Med. Chem*., **30**, 1126 (1987).
23. S. Wold, J. Jonsson, M. Sjöström, M. Sandberg, S. Rännar, *Anal. Chim. Acta*, **277**, 239 (1993).
24. SIMCA-P 8.0. Umetrics UK Ltd., Wokingham Road, RG42 1PL, Bracknell, UK.
25. A.P. Bradley, *Pattern Recogn*., **30**, 1145 (1997).
26. L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, in: Multi- and Megavariate Data Analysis. Principles and Applications. Umetrics Academy, Umea, 2001, p.21.

ОЦЕНКА НА АЛЕРГЕННОСТ ЧРЕЗ ДИСКРИМИНАНТЕН АНАЛИЗ ПО МЕТОДА НА ПАРЦИАЛНИТЕ НАЙ-МАЛКИ КВАДРАТИ

Л. Х. Нанева[1], И. Д. Димитров[2], И. П. Бангов[1], И. А. Дойчинова[2]*

[1]*Шуменски университет „Епископ Константин Преславски“, Факултет по природни науки, ул. „Университетска“ 115, Шумен 9712, България*

[2]*Медицински университет – София, Фармацевтичен факултет, ул. „Дунав“ 2, София 1000, България*

Постъпила на 30 ноември 2012 г.; коригирана на 5 февруари 2013 г.

(Резюме)

Алергенността на хранителните протеини е важен проблем, свързан с широкото използване на нови храни, хранителни добавки и билки, много от които съдържат известни или неизвестни генетично модифицирани протеини. Алергенността е скрито, нелинейно кодирано свойство. Повечето от съществуващите методи за оценка на алергенност се основават на наслагване на секвенции и търсене на структурни прилики с известни алергени. Следователно, идентифицирането на нови, структурно различни алергени не може да бъде осъществено чрез тези методи. В настоящото изследване ние предлагаме нов метод за оценка на алергенност, който не се базира на наслагване на секвенции, а на сравняване на основни свойства на аминокиселините като хидрофобност, размер и електронната структура. Протеините се трансформират във вектори с еднаква дължина и се анализират чрез дискриминантен анализ по метода на парциалните най-малки квадрати. Предварителният модел, получен въз основа на обучаваща група от 120 алергена и 120 неалергена, идентифицира правилно 73% от протеините, включени във външна тестова група от 1164 алергени и неалергени. Разширеният модел, получен въз основа на обучаваща група от 1404 протеина (702 алергена и 702 неалергена) показа 70% точност при кръстосаното валидиране в 7 групи.