Special Issue Article

Journal of CHEMOMETRICS

Received: 03 November 2013, Revised: 23 December 2013, Accepted: 29 December 2013, Published online in Wiley Online Library: 27 January 2014

(wileyonlinelibrary.com) DOI: 10.1002/cem.2597

Allergenicity prediction by artificial neural networks

Ivan Dimitrov[a], Lyudmila Naneva[b], Ivan Bangov[b] and Irini Doytchinova[a]*

Two artificial neural network (ANN)-based algorithms for allergenicity prediction were developed and tested. The first algorithm consists of three steps. Initially, the protein sequences are described by amino acid principal properties as hydrophobicity, size, relative abundance, helix and β-strand forming propensities. Second, the generated strings of different length are converted into vectors with equal length by auto-covariance and cross-covariance (ACC). At the third step, ANN is applied to discriminate between allergens and non-allergens. The second algorithm consists of four steps. It has one additional step before the final ANN modeling. At this step, the ACC vectors are transformed into binary fingerprints. The algorithms were applied to a set of 2427 known allergens and 2427 non-allergens and compared in terms of predictive ability. The three-step algorithm performed better than the four-step one identifying 82% versus 76% of the allergens and non-allergens. The ANN algorithms presented here are universal. They could be applied for any classification problem in computational biology. The amino acid descriptors are able to capture the main structural and physicochemical properties of amino acids building the proteins. The ACC transformation overcomes the main problem in the alignment-based comparative studies arising from the different length of the aligned protein sequences. The uniform-length vectors allow similarity search and classification by different computational methods. Optionally, the ACC vectors could be converted into binary descriptor fingerprints. The comparative study on several Web tools for allergenicity prediction showed that the usage of more than one predictor is reasonable and recommendable because some of the tools recognize better the allergens, some of them—the non-allergens, but none of them—both. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: amino acid descriptors; ACC transformation; descriptor fingerprint; artificial neural networks

1. INTRODUCTION

Allergy is a form of hypersensitivity to normally innocuous substances as dust, pollen, foods or drugs. Allergens are small antigens that commonly provoke an immunoglobulin E (IgE) antibody response [1]. Such antigens normally enter the body at very low doses by diffusion across mucosal surfaces and trigger a Th2 response. The allergen-specific Th2 cells drive allergen-specific B cells to produce IgE, which binds to the high-affinity surface receptor on mast cells, basophils and activated eosinophils. On activation, these cells release stored mediators, which cause inflammation and tissue damage manifested by different symptoms. Inhalant allergens cause rhinitis, conjunctivitis and asthmatic symptoms, while food allergens lead to abdominal pain, bloating, vomiting and diarrhea.

Although there is no consensus allergen structure, the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) have produced Codex Alimentarius guidelines for evaluating potential allergenicity for any novel food protein [2]. According to these guidelines, a query protein is potentially allergenic if it has either an identity of 6–8 contiguous amino acids or >35% sequence similarity over a window of 80 amino acids when compared with known allergens.

Nowadays, two bioinformatics approaches exist to deal with allergen prediction. The first approach follows FAO/WHO guidelines and searches for sequence similarity. Structural Database of Allergenic Proteins [3] and Allermatch [4] contain extensive databases of known allergen proteins and use them as references in sequence-alignment search of query protein. These methods are characterized by high sensitivity, but produce many false positives and low precision. Besides, discovery of novel antigens is restricted by their lack of similarity to known allergens. The second approach is based on identification of conserved allergenicity-related linear motifs. These methods use different techniques for identification, representation and analysis of allergenicity-related motifs. Stadler and Stadler used MEME motif discovery tool to define 52 allergen motifs [5]. Li and colleagues used cluster analysis, wavelet analysis and hidden Markov model profiles to identify allergen motifs [6]. Bjorklund and colleagues have developed an Automated Selection of Allergen-Representative Peptides (ARP) protocol [7]. AlgPred is a server for allergen prediction based on four methods: support vector machines (SVM), program MEME/MAST, IgE epitopes and ARP [8].

Both approaches are based on the assumption that the allergenicity is a linearly coded property. However, in order to act as an allergen, a protein should contain epitopes for Th2 and B cells. The Th2 epitopes are linear, but the B-cell epitopes are

* *Correspondence to: Irini Doytchinova, Faculty of Pharmacy, Medical University of Sofia, 2 Dunav St, 1000 Sofia, Bulgaria. E-mail: idoytchinova@pharmfac.net*

a *I. Dimitrov, I. Doytchinova*
Faculty of Pharmacy, Medical University of Sofia, 2 Dunav St 1000 Sofia, Bulgaria

b *L. Naneva, I. Bangov*
Faculty of Natural Sciences, Konstantin Preslavski Shumen University, 115 Universitetska St 9712 Shumen, Bulgaria

conformational [9]. Apart from the alignment-based approaches, the allergen prediction requires the development and employment of alignment-free approaches.

Recently, we developed a three-step alignment-free algorithm for allergenicity prediction. Initially, the protein sequences of allergens and non-allergens are described by amino acid descriptors, then the different-length strings are converted into uniform, equal-length vectors by auto-covariance and cross-covariance (ACC) transformation [10] and finally, a computational method for allergen/non-allergen discrimination is applied. The algorithm was applied to amino acid *z*-descriptors [11], and *k* nearest neighbors (*k*NN) clustering method and showed *accuracy* of 94% [12]. Next, one additional step, converting the ACC values into a binary descriptor fingerprint, was added, and the algorithm was applied to amino acid *E*-descriptors [13] and Tanimoto coefficient similarity search [14]. Accuracy of 88% was achieved [15].

Here, we further develop our algorithm by applying artificial neuronal networks (ANN) as a method for allergen/non-allergen discrimination. ANN is a widely used method for dealing with nonlinearity in datasets [16]. Two ANN-based algorithms were developed and tested in the present study. The algorithms were compared to existing servers for allergenicity prediction in terms of ability to predict allergens and non-allergens.

2. MATERIALS AND METHODS

2.1. Datasets

The training dataset used in the present study consists of 2427 allergens and 2427 non-allergens. Allergens were collected from the Central Science Laboratory allergen database (http://allergen.csl.gov.uk), the Food Allergen Research and Resource Program allergen database (http://www.allergenonline.org), Structural Database of Allergenic Proteins (http://fermi.utmb.edu/SDAP/sdap_man.html) and Allergome database (http://www.allergome.org/). The allergen proteins were searched in UniProt database (http://www.uniprot.org), and only sequences with "evidence for the existence of protein—evidence at protein level" were selected. Duplicates were removed. The non-allergens were collected from widely used food species as *Solanum lycopersicum* (tomato), *Capsicum annuum* (pepper), *Solanum tuberosum* (potato), *Triticum aestivum* (bread wheat), *Oryza sativa* (Asian rice) and *Oryza glaberrima* (African rice) after search in Swiss-Prot for proteins with "evidence for the existence of protein—evidence at protein level" and exclusion of proteins containing the keyword "allergen" in their description. The resulting set consisted of 950 non-allergens. Additionally, a set of non-allergens was collected from UniProt to include proteins from *Homo sapiens* species with "evidence for the existence of protein—evidence at protein level." The proteins with keywords "allergen" and "cancer" in their description as well as proteins with unidentified amino acids in their sequences were excluded. The set of allergens and non-allergens used in the present study is freely accessible at http://www.ddg-pharmfac.net/AllergenFP/data.html. This set is manually curated and contain only known allergens with evidence at protein level.

An external validation set of allergens containing proteins not included in the training set was compiled from UniProt database with keywords "allergen" and "evidence for the existence of protein—evidence at protein level." Three hundred and three allergens were identified. The same number of non-allergens was selected randomly from UniProt to include proteins from *H. sapiens* species with "evidence for the existence of protein—evidence at protein level" excluding proteins containing the keywords "allergen" and "cancer" in their descriptions. The external validation set was used to test the predictive ability of the tested algorithms.

2.2. *E*-descriptors

The protein sequences of allergens and non-allergens were described by five *E*-descriptors [13]. They were derived by principal component analysis of a data matrix consisting of 237 physicochemical properties. The first principal component (*E*1) reflects the hydrophobicity of amino acids; the second (*E*2) reflects their size; the third (*E*3) reflects their helix-forming propensity; the forth (*E*4) correlates with the relative abundance of amino acids; and the fifth (*E*5) is dominated by the β-strand forming propensity.

2.3. ACC transformation

The ACC transformation turns the different-length strings of *E*-descriptors into uniform equal-length vectors. Auto-covariance $ACC_{jj}(lag)$ and cross-covariance $ACC_{jk}(lag)$ were calculated according to the following equations:

$$ACC_{jj}(lag) = \sum_{i}^{n-lag} \frac{E_{j,i} \times E_{j,i+lag}}{n - lag} \qquad ACC_{jk_{j\neq k}}(lag) = \sum_{i}^{n-lag} \frac{E_{j,i} \times E_{k,i+lag}}{n - lag}$$

where index *j* refers to the *E*-descriptors ($j = 1$–3), *n* is the number of amino acids in a sequence, index *i* points the amino acid position ($i = 1, 2, \ldots, n$) and $lag = 1, 2, \ldots, 20$. In our previous study, it was found that $lag = 15$ gives maximum accuracy [15]. Thus, at the end of this step, the proteins were converted into strings of 375 ($5^2 \times 15$) ACC values.

2.4. Descriptor fingerprints

The ACC values were scaled by a factor of 100, divided into *K* intervals each and converted into $25 \times 15 \times K$-digit binary fingerprints. If a given ACC value falls into a given interval, the corresponding digit in the fingerprint takes 1; otherwise, it takes 0. The ACC values in our dataset range from −10 to +11. In our previous study, it was found that the optimum resolution step is 2, which results in 11 intervals (10 regular intervals with step 2 and 1 shorter interval with step 1) for each ACC [15]. Thus, each protein in the present study had a unique binary fingerprint consisting of 375 units and 3750 nulls.

2.5. Artificial neural networks

The ANNs used in the present study were built in WEKA [17] using the multilayer perceptron (MLP) function. MLP is a classifier that uses backpropagation to classify instances. It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. The nodes in the network are all sigmoid. Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. The learning rate and momentum of the network were set to 0.3 and 0.2, respectively. Although the training time was fixed at 500 epochs, a 20% validation set was used. The training of the network continues until it is observed that the error of the validation set has been consistently getting worse or if the training time was reached.

2.6. Evaluation of performance

The algorithms for allergenicity prediction developed in the present study were evaluated by 10-fold cross-validation and randomly selected test sets comprising 20%, 25% and 33% of the initial set. The correctly predicted allergens and non-allergens were defined as true positives (TP) and true negatives (TN), respectively. The incorrectly predicted allergens and non-allergens were defined as false negatives (FN) and false positives (FP), respectively. *Sensitivity* [TP/(TP + FN)], *specificity* [TN/(TN + FP)], *positive predictive value* (*ppv*) [TP/(TP + FP)], *F*1 score [2**sensitivity***ppv*/(*sensitivity* + *ppv*)] and *area under curve sensitivity*/1-*specificity* (*AUC*) were calculated.

2.7. Available tools for allergenicity prediction

The external validation set was used to compare the predictive ability of the algorithms developed in the present study to five servers for allergenicity prediction freely available in the Web. These servers were AllerHunter, AlgPred, ProAp, AllerTOP and AllergenFP.

AllerHunter (http://tiger.dbs.nus.edu.sg/AllerHunter) is a cross-reactive allergen prediction program built on a combination of SVM and pairwise sequence similarity [18].

AlgPred (http://imtech.res.in/raghava/algpred) predicts allergens by applying four different methods: MEME/MAST motif search, SVM-based classification of allergens and non-allergens by single amino acid composition (AlgPred(SVM_single_aa)) and by dipeptide composition (AlgPred(SVM_dipeptide), and BLAST search against allergen representative peptides (AlgPred (ARP)). MEME is a tool for discovering motifs in a group of related protein sequences. MAST searches in biological sequence databases for sequences that contain one or more groups of known motifs. Single amino acid composition gives the fraction of each amino acid in a protein. Dipeptide composition is used to encapsulate the global information about each protein sequence and gives a fixed pattern length of 400 (20 × 20). The BLAST search is performed against a set containing 24 amino acid long peptides, so-called ARP, and finds proteins with high similarity to allergenic proteins [19].

ProAp (http://gmobl.sjtu.edu.cn/proAP/main.html) is a Web-based application that integrates and optimizes sequence-based, motif-based (ProAp(motif)) and SVM-based (ProAp(SVM)) allergen prediction approaches for determination of cross-reactivities between potential allergens and known allergens [20]. The applied SVM method takes amino acid composition as protein features.

AllerTOP (http://www.pharmfac.net/allertop) is the first alignment-free server for *in silico* prediction of allergens based on the main physicochemical properties of proteins [12]. AllerTOP utilizes a model based on amino acid *z*-descriptors, ACC protein transformation and *k*NN clustering. The protein sequences are uploaded in plain format. The results page returns the allergen status: "Probable Allergen" or "Probable Non-allergen." It also returns the *k*NN in the training set. On this basis, AllerTOP defines the most probable route of exposure of tested proteins predicted as an allergen: food, inhalant or toxin.

AllergenFP (http://www.ddg-pharmfac.net/AllergenFP) also is an alignment-free server for allergenicity prediction. It is based on comparison of protein descriptor fingerprints in terms of Tanimoto similarity coefficient [15].

3. RESULTS AND DISCUSSION

3.1. ANN-based algorithms for allergenicity prediction

Two ANN-based algorithms for allergenicity prediction were developed in the present study. The first algorithm consists of three steps: description of proteins by amino acid *E*-descriptors; ACC transformation into uniform vectors; and development of ANN-based model for allergen/non-allergen discrimination. The second algorithm consists of four-steps. One additional step before the final ANN modeling was included converting the ACC vectors into binary fingerprints.

3.1. Optimization of ANN

The networks were trained and tested at one, two and three hidden layers. The first hidden layer consists of 188 nodes ((375 ACC + 1)/2); the second layer is a half of the first (94 nodes); and the third one is a half of the second (47 nodes). In order to decrease the number of elements in the fingerprints generated in the four-step algorithm, the binary strings were presented as sets of numbers indicating the positions of units. For example, the fingerprint "…|00010000000|00000010000|00000000001|…" was presented as "…|4|7|11|…." This allowed us to use the same NN architecture for both algorithms.

Maximum predictive ability was obtained with a single hidden layer for both algorithms (Table I). The three-step algorithm yields 82.4% sensitivity, 82.4% specificity, 82.5% ppv, 0.824 F1 score and 0.873 AUC, when tested on 20% test set. The corresponding values for the four-step algorithm were 75.6% sensitivity, 75.6% specificity, 75.7% ppv, 0.756 F1 score and 0.829 AUC.

Table I. Predictive ability of the ANN-based algorithms at different number of hidden layers tested on 20% test set

Number of hidden layers	Sensitivity %	Specificity %	PPV %	F1	AUC
Three-step algorithm					
One	82.4	82.4	82.5	0.824	0.873
Two	78.1	78.1	78.1	0.781	0.844
Three	78.2	78.2	78.3	0.782	0.844
Four-step algorithm					
One	75.6	75.6	75.7	0.756	0.829
Two	74.4	74.4	74.4	0.743	0.812
Three	75.1	75.1	75.1	0.750	0.812

Table II. Predictive ability of the ANN-based models tested by 10-fold cross-validation and different-sized test sets

Test set	Sensitivity %	Specificity %	PPV %	F1	AUC
Three-step algorithm					
10-fold cross-validation	82.0	82.0	82.1	0.820	0.876
20%	82.4	82.4	82.5	0.825	0.873
25%	81.3	81.3	81.3	0.813	0.865
33%	82.3	82.5	82.6	0.824	0.871
Three-step algorithm					
10-fold cross-validation	77.6	77.6	77.7	0.776	0.84
20%	75.6	75.6	75.7	0.756	0.829
25%	77.5	77.5	77.5	0.775	0.838
33%	74.7	74.7	74.9	0.747	0.827

Table III. Evaluation of the performance of the algorithms developed in the present study and five freely accessible Web servers for allergenicity prediction

Server	Sensitivity %	Specificity %	Accuracy %
AllerHunter	35.3	99.0	67.2
AlgPred(SVM_single_aa)	79.5	76.9	78.2
AlgPred(SVM_dipeptide)	71.0	81.2	76.1
AlgPred(ARP)	36.3	98.0	67.2
ProAp(motif)	100.0	0.0	50.0
ProAp(SVM)	60.0	92.4	81.2
AllerTOP	36.3	91.4	63.9
AllergenFP	57.4	83.5	70.5
Three-step ANN algorithm	79.2	81.9	80.5
Four-step ANN algorithm	68.7	71.6	70.1

3.2. ANN-based models for allergenicity prediction

The algorithms developed and optimized in the present study were applied to the dataset of 2427 allergens and 2427 non-allergens and tested by 10-fold cross-validation. The predictive ability of the ANN-based models is presented in Table II. It is evident that the three-step algorithm performed better than the four-step one identifying 82.0% of the allergens and 77.6% of the non-allergens. Additionally, the initial set of proteins was divided randomly into training and test sets. Three test sets were selected to include 20%, 25% and 33% of the initial set. Again, the three-step algorithm showed higher predictive ability than the four-step one with sensitivity and specificity between 81.3% and 82.4%. The corresponding values for the four-step algorithm ranged between 74.7% and 75.6%. Both models were quite stable as indicated by the close predictive abilities tested on the different-sized test sets.

3.3. External validation and comparison to available tools for allergenicity prediction

An external validation set of 303 allergens and 303 non-allergens not included in the training set was used to compare the predictive ability of the algorithms developed in the present study to five servers for allergenicity prediction freely available in the Web. These servers were AllerHunter, AlgPred, ProAp, AllerTOP and AllergenFP. The predictive ability of the tools was evaluated by sensitivity, that is, percent of identified allergens, specificity, that is, percent of identified non-allergens, and accuracy. The results are given in Table III.

ProAp(motif) identified all 303 allergens, followed by AlgPred (SVM_single_aa) with 241 identified allergens (79.5% sensitivity) and the three-step ANN algorithm with 240 identified allergens (79.2% sensitivity). AllerHunter identified 300 non-allergens (99.0% specificity), followed by AlgPred(APR) with 297 identified non-allergens (98.0% specificity) and ProAp(SVM) with 280 identified non-allergens (92.4% specificity). The most accurate predictor was ProAp(SVM) with 492 correctly identified allergens and non-allergens (81.2% accuracy), followed by the three-step ANN algorithm with 488 correct predictions (80.5% accuracy) and AlgPred(SVM_single_aa) with 474 correct predictions (78.2% accuracy).

It is evident that some of the available tools for allergenicity prediction predict well allergens, some of them—non-allergens, but none of them—both allergens and non-allergens. In this sense, we recommend more than one predictor to be used when allergenic proteins are searched.

4. CONCLUSIONS

The ANN algorithms presented here are universal. They could be applied for any classification problem in computational biology. The amino acid descriptors are able to capture the main structural and physicochemical properties of amino acids building the proteins. The ACC transformation overcomes the main problem in the alignment-based comparative studies arising from the different length of the aligned protein sequences. The uniform-length vectors allow similarity search and classification

by different computational methods. Optionally, the ACC vectors could be converted into binary descriptor fingerprints.

The comparative study on several Web tools for allergenicity prediction showed that all servers compromise between sensitivity and specificity, that is, some of them predict well allergens, some of them—non-allergens, but none of them—both. In this sense, the usage of more than one predictor is reasonable and recommendable.

Acknowledgements

We acknowledge the Bulgarian Science Fund for the financial support (grants DCVNP 02-1/2009 and IO1/7).

REFERENCES

1. Cooper PJ. Intestinal worms and human allergy. *Parasite Immunol.* 2004; **26**: 455-467.
2. FAO/WHO Codex Alimentarius Commission: Codex principles and guidelines on foods derived from biotechnology. *Joint FAO/WHO Food Standards Programme*. Rome, Italy, 2003.
3. Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.* 2003; **31**: 359–362.
4. Fiers MWEJ, Kleter GA, Nijland H, Peijnenburg AA, Nap JP, van Ham RC. Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics* 2004; **5**: 133.
5. Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. *FASEB J.* 2003; **17**: 1141–1143.
6. Li KB, Issac P, Krishnan A. Predicting allergenic proteins using wavelet transform. *Bioinformatics* 2004; **20**: 2572–2578.
7. Björklund AK, Soeria-Atmadja D, Zorzet A, Hammerling U, Gustafsson MG. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics* 2005; **21**: 39–50.
8. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* 2006; **34**: W202–W209.
9. Furmonaviciene R, Sutton BJ, Glaser F, Laughton CA, Jones N, Sewell HF, Shakib F. An attempt to define allergen-specific molecular surface features: a bioinformatic approach. *Bioinformatics* 2005; **21**: 4201–4204.
10. Nyström Å, Andersson PM, Lundstedt T. Multivariate data analysis of topographically modified alpha-melanotropin analogues using auto and cross auto covariances (ACC). *Quant. Struct.-Act. Relat.* 2000; **19**: 264–269.
11. Hellberg S, Sjöström M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* 1987; **14**: 1126–1135.
12. Dimitrov I, Flower DR, Doytchinova I. AllerTOP – a server for *in silico* prediction of allergens. *BMC Bioinformatics* 2013; **14**(Suppl. 6): S4.
13. Venkatarajan MS, 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.
14. Tanimoto TT. *An Elementary Mathematical Theory of Classification and Prediction*. IBM Research Yorktown Heights: New York, 1958.
15. Dimitrov I, Naneva L, Doytchinova I, Bangov I. Allergenicity prediction by descriptor fingerprints. *Bioinformatics* 2013. DOI: 10.1093/bioinformatics/btt619
16. Somers MJ, Casal JC. Using artificial neural networks to model nonlinearity. *Organ. Res. Meth.* 2009; **12**: 403–417.
17. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update; SIGKDD explorations, 2009; **11**.
18. Zorzet A, Gustafsson M, Hammerling U. Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol.* 2002; **2**: 525–534.
19. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* 2006; **34**: W202–W209.
20. Wang J, Yu Y, Zhao Y, Zhang D, Li J. Evaluation and integration of existing methods for computational prediction of allergens. *BMC Bioinformatics*, 2013; **14**(Suppl. 4), S1.

