BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 846–851
doi:10.1093/bioinformatics/btt619

Systems biology

Advance Access publication October 27, 2013

AllergenFP: allergenicity prediction by descriptor fingerprints

Ivan Dimitrov[1], Lyudmila Naneva[2], Irini Doytchinova[1,*] and Ivan Bangov[2,*]

[1]Medical University of Sofia, Faculty of Pharmacy, 2 Dunav st., 1000 Sofia and [2]Konstantin Preslavski Shumen University, Faculty of Natural Sciences, 115 Universitetska st., 9712 Shumen, Bulgaria

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Allergenicity, like antigenicity and immunogenicity, is a property encoded linearly and non-linearly, and therefore the alignment-based approaches are not able to identify this property unambiguously. A novel alignment-free descriptor-based fingerprint approach is presented here and applied to identify allergens and non-allergens. The approach was implemented into a four step algorithm. Initially, the protein sequences are described by amino acid principal properties as hydrophobicity, size, relative abundance, helix and β-strand forming propensities. Then, the generated strings of different length are converted into vectors with equal length by auto- and cross-covariance (ACC). The vectors were transformed into binary fingerprints and compared in terms of Tanimoto coefficient.

Results: The approach was applied to a set of 2427 known allergens and 2427 non-allergens and identified correctly 88% of them with *Matthews correlation coefficient* of 0.759. The descriptor fingerprint approach presented here is universal. It could be applied for any classification problem in computational biology. The set of E-descriptors is 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 conversion of protein ACC values into binary descriptor fingerprints allows similarity search and classification.

Availability and implementation: The algorithm described in the present study was implemented in a specially designed Web site, named AllergenFP (FP stands for FingerPrint). AllergenFP is written in Python, with GIU in HTML. It is freely accessible at http://ddg-pharmfac.net/Allergen FP.

Contact: idoytchinova@pharmfac.net or ivanbangov@shu-bg.net

Received on July 9, 2013; revised on September 30, 2013; accepted on October 21, 2013

1 INTRODUCTION

Allergy is a growing health problem of modern life. The prevalence of allergic diseases worldwide is rising dramatically in both developed and developing countries (Pawankar *et al.*, 2011). Allergy involves a complex series of reactions to external and internal factors, called allergens, which contribute to the development of diseases like rhinitis, asthma, atopic dermatitis and skin sensitization. Severe reactions as acute and fatal anaphylactic shock also may occur.

An allergic reaction occurs when a susceptible organism is re-exposed to a specific allergen. The allergen-specific Th2 cells drive the B cells to produce IgE, which binds to mast cells, basophils and activated eosinophils. When the same allergens reenter the body, they bind to these IgEs and activate the cells to release stored mediators, which give rise to inflammation and tissue damage (Cooper, 2004; Huby *et al.*, 2000; Rusznak *et al.*, 1998).

Although there is no consensus on the allergen structure, the United Nations Food and Agriculture Organization (FAO) and the World Health Organization (WHO) have developed guidelines for evaluating the potential allergenicity of novel food proteins (FAO/WHO Agriculture and Consumer Protection, 2001; FAO/WHO Codex Alimentarius Commission, 2003). According to these guidelines, a protein is a potential allergen, if it has either an identity of 6–8 consecutive amino acids or >35% sequence similarity over a window of 80 amino acids when compared with known allergens (Stadler and Stadler, 2003).

Bioinformatics provides two types of approaches for allergen prediction. The first approach follows FAO/WHO guidelines and seeks for sequence similarity. Extensive databases of known allergens like Structural Database of Allergenic Proteins (SDAP) (Ivanciuc *et al.*, 2003), Allermatch (Fiers *et al.*, 2004) and AllerTool (Zhang *et al.*, 2007) are used for similarity searches. This approach has a good ability to identify allergens, but also generates a large number of false allergens. Moreover, the discovery of novel structurally different allergens is limited by the lack of similarity to known allergens.

The second approach is based on identification of patterns for allergenicity, called motifs. The motif is a combination of amino acids responsible for a specific protein activity. Stadler and Stadler (2003) defined 52 allergen motifs by comparing allergens and non-allergens using software MEME. Li *et al.* (2004) used cluster analysis, wavelet analysis and hidden Markov model profiles to identify allergen motifs. Björklund *et al.* (2005) developed an Automated Selection of Allergen Representative Peptides (ARP) protocol. AlgPred is a server for allergen prediction based on four methods: support vector machines (SVM), program MEME/MAST, IgE epitopes and ARP (Saha and Raghava, 2006). Ivanciuc *et al.* (2009) found that some of the allergen motifs coincided with IgE epitopes. Motif-generating algorithms may be used for identifying major IgE binding structures of coiled-coil proteins (Marti *et al.*, 2009).

Both approaches are sequence based, i.e. they require initial sets of allergens and non-allergens to be trained. Both approaches assume that allergenicity is a linearly coded property. To act as an allergen, a protein should contain epitopes for both Th-2 and B cells. The Th-2 epitopes are linear, but B-cell epitopes are conformational patches on the protein surface. Furmonaviciene *et al.* (2005) found allergen-specific conformational epitopes, consisting mainly of hydrophobic residues on

*To whom correspondence should be addressed.

the surface. This finding is consistent with the fact that the innate immune system is able to detect hydrophobic parts of immunogenic proteins containing aliphatic or aromatic amino acids (Seong and Matzinger, 2004).

Obviously, allergenicity, like antigenicity and immunogenicity, is a property encoded linearly and non-linearly, and therefore the alignment-based approaches are not able to identify this property unambiguously. In the present study, we describe an alignment-free method for allergenicity prediction based on amino acid *E*-descriptors (Venkatarajan and Braun, 2001) and auto- and cross-covariance (ACC) transformation of protein sequences into uniform equal-length vectors (Nyström *et al*., 2000). ACC was used for quantitative structure–activity relationship (QSAR) studies of peptides (Nyström *et al*., 2000), protein classification (Lapinsh *et al*., 2002) and immunogenicity prediction (Doytchinova and Flower, 2007). Recently, we applied this method in combination with k nearest neighbors method to predict allergenicity (Dimitrov *et al*., 2013). Here, we improve the method by using a new set of amino acid descriptors and a novel descriptor fingerprint approach based on Tanimoto coefficient similarity search (Tanimoto, 1958).

The *E*-descriptors for the 20 naturally occurring amino acids, defined by Venkatarajan and Braun (2001), were derived by principal component analysis of a data matrix consisting of 237 physicochemical properties. The first principal component (*E1*) reflects the hydrophobicity of amino acids; the second (*E2*)—their size; the third (*E3*)—their helix-forming propensity; the forth (*E4*) correlates with the relative abundance of amino acids; and the fifth (*E5*) is dominated by the β-strand forming propensity. In the present study, the five *E*-descriptors were used to describe the protein sequences.

The fingerprints in chemoinformatics are developed to describe the 2D chemical structures (Barnard, 2003; Kochev *et al*., 2003; Tomczak, 2003; Willett, 2003). They are generated in two ways. In the most popular way, the structure of a compound is divided into substructural fragments and assembled into a binary string—if the fragment exists in the structure the corresponding string element takes 1, otherwise, it takes 0. Binary structural fingerprints are especially useful, as there are highly efficient in computer algorithms working with binary strings. Daylight has developed another approach (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html). The Daylight fingerprints are generated as patterns from the environment of each atom (up to seven atoms away). Each pattern serves as a seed to a pseudo-random number generator (it is 'hashed'), the output of which is a set of bits (typically 4 or 5 bits per pattern); the set of bits is added (with a logical OR) to the fingerprint.

Here, we present a novel descriptor-based fingerprint approach. The fingerprint consists of descriptors instead of structural fragments. This enlarges the area of fingerprint application outside the chemical structure description, even outside chemistry. Moreover, it allows physical and/or biological information to be included in the similarity search process. Obviously, the basic problem here is, as in other quantitative structure–activity relationship methods, the selection of proper descriptors. They could be discrete, i.e. Boolean values, e.g. the presence or absence of given physical, chemical and/or biological features. The fingerprint is binary coded—when the given feature presents, the fingerprint takes 1 at the proper position, otherwise, it equals 0. Many descriptors have continuous values, e.g. a nuclear magnetic resonance or infrared spectra, logP, quantum chemistry-based descriptors and so forth. These descriptors also could be converted into discrete values by scaling and dividing into intervals with regular size. If a descriptor value falls within an interval, the corresponding fingerprint element takes 1; otherwise, it takes 0. Obviously, the length of the fingerprint depends on the number of intervals: as many they are, as longer is the fingerprint. The longer fingerprints are able to capture more molecular features than the shorter ones but in the same time are redundant in information. A balance between fingerprint length and information content should be found in any particular study.

2 DATASETS AND METHODS

2.1 Protein datasets

2.1.1 Allergens An initial allergen dataset was collected from the CSL (Central Science Laboratory) allergen database (http://allergen.csl.gov.uk), the FARRP (Food Allergen Research and Resource Program) allergen database (http://www.allergenonline.org), SDAP (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 Swiss-Prot 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 final set of allergens contained 2427 proteins.

2.1.2 Non-allergens A set of proteins from species *Solanum lycopersicum* (tomato), *Capsicum annuum* (pepper), *Solanum tuberosum* (potato), *Triticum aestivum* (bread wheat) and *Oryza sativa* (Asian rice) and *Oryza glaberrima* (African rice) was collected after search in Swiss-Prot for proteins with 'evidence for the existence of protein-evidence at protein level' and exclusion of proteins containing the key word 'allergen' in their description. The resulting set consisted of 950 non-allergens. Additionally, a set of non-allergens was collected from Swiss-Prot to include proteins from *Homo sapiens* species with 'evidence for the existence of protein-evidence at protein level'. The proteins with key words 'allergen' and 'cancer' in their description as well as proteins with unidentified amino acids in their sequences were excluded. The final set of non-allergens contained 2427 proteins.

The set of allergens and non-allergens used in the present study is freely accessible from AllergenFP Web site. This set could be used as a uniform reference set in future studies on allergenicity prediction as it is manually curated and contains known allergens with evidence at protein level.

2.2 Presentation of protein sequences by *E*-descriptors and ACC transformation

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-cross covariance (ACC) transformation was used (Nyström *et al*., 2000).

Table 1. *E*-descriptors of amino acids (Venkatarajan and Braun, 2001)

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

Auto-covariance $Ajj(L)$ and cross-covariance $Cjk(L)$ were calculated according to the following equations:

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

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

Indices j and k refer to the E-descriptors ($j = 1\text{-}5$, $k = 1\text{-}5$, $j \neq k$), n is the number of amino acids in a sequence, index i points the amino acid position ($i = 1, 2, \ldots, n$) and L is the lag ($L = 1, 2, \ldots, L$). Short lags ($L = 5 \div 20$) were chosen, as only the influence of close amino acid proximity was investigated. $A_{jj}(L)$ and $C_{jk}(L)$ values were assigned further in the article as ACC values. The subsets of allergens and non-allergens were transformed into matrices with $25 \times L$ variables ($5^2 \times L$) each.

2.3 Descriptor fingerprints approach based on Tanimoto coefficient similarity search

In the present study, ACC-based descriptor fingerprints were generated. The $25 \times L$ ACC descriptors were scaled by a factor of 100, divided into K intervals each and converted into $25 \times L \times K$-digit binary fingerprints. A digit in the fingerprint equals 1, if the ACC value falls into the corresponding interval, otherwise, it takes 0. Thus, each protein has a unique binary fingerprint consisting of $25 \times L$ units and ($25 \times L \times K$–$25 \times L$) nulls.

The fingerprints for each pair proteins were compared according to Tanimoto coefficient:

$$T(A, B) = \frac{N_C}{N_A + N_B - N_C}$$

where N_A and N_B are the number of digits in the fingerprints of structure A and structure B, respectively, and N_C is the number of common digits between them. Tanimoto coefficient takes value between 0 and 1. The larger is the value, the more similar are the two structures. Thus, two structures with Tanimoto coefficient $T = 0.98$ are much more similar than two structures with Tanimoto coefficient $T = 0.56$.

2.4 Evaluation of performance

The method for allergenicity prediction developed in the present study was evaluated by leave-one-out cross-validation (LOO-CV) and by 10-fold cross-validation (10-fold-CV). In the LOO-CV procedure, each protein was excluded from the dataset and compared with the remaining *n-1* proteins by Tanimoto coefficient. The protein with the highest value for $T(A,B)$ determined the class identity of the tested protein (allergen or non-allergen). In the 10-fold-CV procedure, the whole set is divided into 10 groups. Each group was treated as a test set, the remaining nine as a training 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* [FP/(TN + FP)], *positive predictive value* (*ppv*) [TP/(TP + FP)], *F1* score [2**sensitivity***ppv*/(*sensitivity* + *ppv*)] and *Matthews correlation coefficient* (*MCC*) [(TP × TN − FP × FN)/((TP + FP)(TP + FN)(TN + FP)(TN + FN))$^{1/2}$] were calculated. For the 10-fold-CV, average values of the 10 runs were given.

2.5 Web servers for allergenicity prediction

The method developed in the present study was compared to five servers for allergenicity prediction freely available in the web. These were AllerHunter, AlgPred, APPEL, ProAp and AllerTOP.

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 (Zorzet *et al.*, 2002). Each proteins sequence in the training set is vectorized by performing sequence alignment and Basic Local Alignment Search Tool (BLAST) against all other members of the training set. The protein sequences are represented as vectors consisted of similarity scores for each pair of proteins in the training set.

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 (Saha and Raghava, 2006).

The APPEL tool (Allergen Protein Prediction E-Lab) (http://jing.cz3. nus. edu.sg/cgi-bin/APPEL) uses SVM to identify novel allergen proteins from the sequence-derived structural and physicochemical properties of a whole protein (Cui *et al.*, 2007). It is based on a statistical method and has the potential to discover novel allergen proteins.

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 (Wang *et al.*, 2013). 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 (Dimitrov *et al.*, 2013). AllerTOP uses a model based on amino acid z-descriptors, ACC protein transformation and *k* nearest neighbors 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* nearest neighbours 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.

3 ALGORITHM

The algorithm used in the present study is described in Figure 1. Initially, the amino acids in the protein sequences were described by the five *E*-descriptors and the strings were transformed into uniform vectors by ACC. The derived matrix consisted of 4854 rows (2427 allergens and 2427 non-allergens) and $25 \times L$ columns. Each column was divided into K intervals and a $25 \times L \times K$-digit binary fingerprint was generated for each protein. A digit in the fingerprint equals 1, if the ACC value falls into the corresponding interval, otherwise, it takes 0. Thus, each protein has a unique binary fingerprint consisted of $25 \times L$ units and $(25 \times L \times K – 25 \times L)$ nulls. Tanimoto coefficients were calculated for all protein pairs in the set. A protein was classified as allergen or non-allergen according to the protein from the pair with the highest Tanimoto coefficient.

The algorithm was optimized in terms of lag length l changing L from 5 to 20 with step of 5. The performances were compared by *sensitivity*, *specificity*, *accuracy*, *ppv* and *F1* value (Fig. 2). *Sensitivity* initially decreases with increasing of L, reaches minimum at $L = 10$, then increases, reaches maximum at $L = 15$ and again decreases. *Specificity*, *accuracy*, *ppv* and *F1* increase gradually, reach maximum at $L = 15$ and then decrease. Thus, $L = 15$ was chosen as an optimum value.

The ACC values range from −10 to +11. Each ACC was divided into K regular intervals. Short intervals generate long fingerprints requiring long time for calculation. Longer fingerprints capture more structural information. Some of this information brings more 'noise' than 'signal'. A balance between fingerprint length and information content should be found in any particular study. In the present study, the optimal number of regular intervals was found by changing the resolution step and

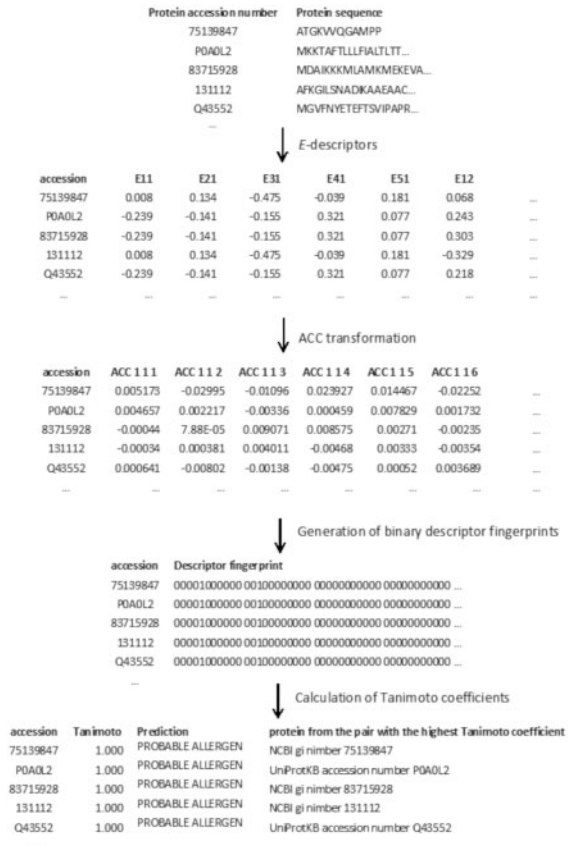

Fig. 1. Flowchart of the algorithm

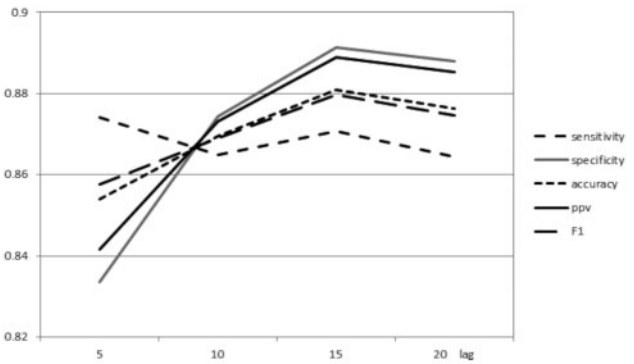

Fig. 2. The impact of lag length on the predictive ability of the algorithm

evaluating the algorithm performance by the parameters *sensitivity*, *specificity*, *accuracy*, *ppv* and *F1* value. Six different steps were tested: 0.1, 0.5, 1.0, 1.5, 2.0 and 2.5 (Fig. 3). Zigzag curves were generated with maximum at step 2. Only the *sensitivity* slightly increases at step 2.5. Thus, step 2 was selected as an optimum resolution step that results in 11 intervals (10 regular intervals with step 2 and 1 shorter interval with step 1) for each ACC.

The final matrix consisted of 4854 rows (2427 allergens and 2427 non-allergens) and 4125 columns ($25 \times 15 \times 11$). Every

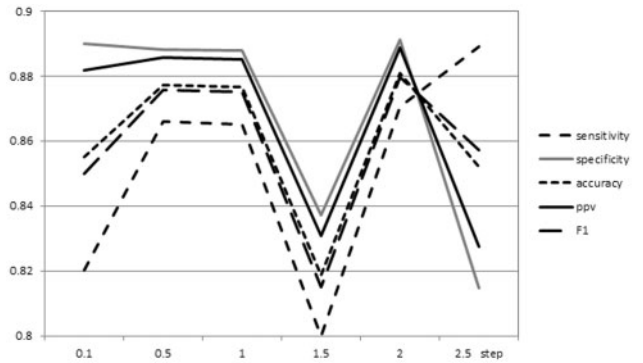

Fig. 3. The impact of resolution step on the predictive ability of the algorithm

protein is presented by a descriptor fingerprint containing 375 units (25×15) and 3750 nulls ($25 \times 15 \times 11 - 25 \times 15$). The LOO-CV tests gave 87% *sensitivity*, 89% *specificity*, 88% *accuracy*, 89% *ppv*, $F1 = 0.88$ and $MCC = 0.76$. The 10-fold-CV tests gave slightly lower predictions: 85% *sensitivity*, 85% *specificity*, 85% *accuracy*, 85% *ppv*, $F1 = 0.85$ and $MCC = 0.70$.

4 IMPLEMENTATION

The algorithm described in the present study was implemented in a specially designed Web site, named AllergenFP (FP stands for **F**inger**P**rint). AllergenFP is written in Python, with GIU in HTML. It is freely accessible at http://ddg-pharmfac.net/AllergenFP. Single protein sequence is uploaded in plain format. The results page returns the predicted protein status: 'Probable Allergen' or 'Probable Non-allergen'. It also returns the protein from the pair with the highest value of Tanimoto coefficient.

5 COMPARISON WITH EXISTING METHODS FOR ALLERGEN PREDICTION

AllergenFP was compared to five freely available web servers for allergenicity prediction. The set of known allergens and non-allergens compiled in the present study was used to test the servers. The results are given in Table 2.

ProAp(motif) showed the highest *sensitivity*. It recognized 94% of the allergens in the set, followed by AlgPred(SVM_single_aa) (89%) and AllerTOP (88%). The highest *specificity* belongs to AllerHunter. It recognized 96% of the non-allergens, closely followed by AlgPred(ARP) (95%) and APPEL (91%). AllergenFP is the most accurate predictor identifying 88% of both allergens and non-allergens. Close to it is AllerHunter with 87% *accuracy*. The parameter *ppv* accounts for the fraction of TP among all predicted positives. Ninety-five percent of the predicted allergens by AllerHunter are true allergens. Next to it is AlgPred(ARP) with 94% *ppv*. *F1* is a weighted average of *sensitivity* and *ppv*. AllergenFP has the highest value for *F1* (0.878), followed by AllerHunter (0.858). By definition, *MCC* is a correlation coefficient between the observed and predicted binary classifications. AllergenFP showed the highest *MCC* (0.759) closely followed by AllerHunter (0.754).

Table 2. Evaluation of the performance of six freely accessible web servers for allergenicity prediction

Server	*Sensitivity*	*Specificity*	*Accuracy*	*ppv*	*F1*	*MCC*
AllerHunter	0.782	0.960	0.871	0.951	0.858	0.754
AlgPred(SVM_single_aa)	0.894	0.657	0.775	0.723	0.799	0.567
AlgPred(SVM_dipeptide)	0.866	0.726	0.796	0.760	0.809	0.598
AlgPred (ARP)	0.730	0.953	0.842	0.940	0.822	0.701
APPEL	0.653	0.914	0.783	0.883	0.751	0.587
ProAp(motif)	0.938	0.072	0.505	0.503	0.655	0.020
ProAp(SVM)	0.813	0.874	0.843	0.866	0.839	0.688
AllerTOP	0.876	0.780	0.828	0.799	0.836	0.659
AllergenFP	0.868	0.891	0.879	0.889	0.878	0.759

6 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 amino acid principal properties as hydrophobicity, size, relative abundance, helix and β-strand forming propensities. Proteins are transformed into descriptor-based fingerprints and compared by Tanimoto coefficient. The algorithm was optimized in terms of lag length and resolution step and cross-validated by a set of 2427 known allergens and 2427 non-allergens. It recognized 87% of the allergens and 89% of the non-allergens. The algorithm and the set of proteins used in the study were implemented in a specially designed Web site, named AllergenFP and freely accessible at http://ddg-pharmfac.net/AllergenFP. AllergenFP was compared with five freely available web servers for allergenicity prediction and showed the highest predictive ability.

For example, recently Liao *et al*. (2013) identified a novel allergenic protein Tyr p 8 from *Tyrophagus putrescentiae* that cross-reacts with Der p 8 from *Dermatophagoides pteronyssinus*. Both proteins are not available in our database. AllergenFP identifies them as allergens with $T = 0.84$.

The descriptor fingerprint approach based on Tanimoto coefficient similarity search used in the present study to discriminate between allergens and non-allergens is universal. It could be applied for any classification problem in computational biology. The set of *E*-descriptors is able to capture the main structural and physicochemical properties of amino acids. The ACC transformation overcomes the main problem in the alignment-based comparative studies arising from the different length of the aligned protein sequences. The conversion of protein ACC values into a binary descriptor fingerprint is a computational novelty, described here for the first time.

The comparison between the freely available servers for allergenicity prediction showed that most of them performed well and gave reliable predictions. Some of them, like ProAp(motif), AlgPred(SVM_single_aa) and AllerTOP, identify allergens better than non-allergens. Others, like AlgPred(ARP) and APPEL, are good in identifying non-allergens. Only AllergenFP and AllerHunter gave balanced predictions of both allergens and non-allergens. Furthermore, AllergenFP showed the best performance in terms of *F1* and *MCC*.

Funding: Bulgarian Science Fund (grants DCVNP 02-1/2009 and IO 7/1).

Conflict of Interest: none declared.

REFERENCES

Barnard,J.M. (2003) Representation of molecular structures. overview. In: Gastejger,J. (ed.) *Handbook of Chemoinformatics*. Vol. 1, Welley-VCH, Weinheim, Germany, pp. 27–50.

Björklund,A.K. *et al.* (2005) Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics*, **21**, 39–50.

Cooper,P.J. (2004) Intestinal worms and human allergy. *Parasite Immunol.*, **26**, 455–467.

Cui,J. *et al.* (2007) Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol. Immunol.*, **44**, 514–520.

FAO/WHO Agriculture and Consumer Protection. (2001) Evaluation of Allergenicity of Genetically Modified Foods. *Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology*. Rome, Italy.

FAO/WHO Codex Alimentarius Commission. (2003) Codex Principles and Guidelines on Foods Derived from Biotechnology. *Joint FAO/WHO Food Standards Programme*. Rome, Italy.

Dimitrov,I. *et al.* (2013) AllerTOP – a server for *in silico* prediction of allergens. *BMC Bioinformatics*, **14** (**Suppl. 6**), S4.

Doytchinova,I.A. and Flower,D.R. (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, **8**, 4.

Fiers,M.W.E.J. *et al.* (2004) Allermatch, a webtool for the prediction of potential allergenicity according to current fao/who codex alimentarius guidelines. *BMC Bioinformatics*, **5**, 133.

Furmonaviciene,R. *et al.* (2005) An attempt to define allergen-specific molecular surface features: a bioinformatic approach. *Bioinformatics*, **21**, 4201–4204.

Huby,R.D.J. *et al.* (2000) Why are some proteins allergens. *Toxicol. Sci.*, **55**, 235–246.

Ivanciuc,O. *et al.* (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.*, **31**, 359–362.

Ivanciuc,O. *et al.* (2009) Characteristic motifs for families of allergenic proteins. *Mol. Immunol.*, **46**, 559–568.

Kochev,N. *et al.* (2003) Searching Chemical Structures. In: Engel,T. and Gastejger,J. (eds) *Chemoinformatics. A Textbook*. Welley-VCH, Weinheim, Germany, pp. 291–318.

Lapinsh,M. *et al.* (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci.*, **11**, 795–805.

Li,K.B. *et al.* (2004) Predicting allergenic proteins using wavelet transform. *Bioinformatics*, **20**, 2572–2578.

Marti,P. *et al.* (2007) Allergen motifs and the prediction of allergenicity. *Immunol. Lett.*, **109**, 47–55.

Nyström,Å. *et al.* (2000) Multivariate data analysis of topographically modified á-melanotropin analoques using auto and cross auto covariances (ACC). *Quant. Struct.-Act. Relat.*, **19**, 264–269.

Pawankar,R. *et al.* (2011) WAO White book on allergy 2011 – 2012: Executive summary. World Allergy Organization.

Rusznak,C. *et al.* (1998) ABC of allergies. Diagnosing allergy. *BMJ*, **316**, 686–689.

Saha,S. and Raghava,G.P.S. (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.*, **34**, W202–W209.

Seong,S.Y. and Matzinger,P. (2004) Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses. *Nat. Rev. Immunol.*, **4**, 469.

Stadler,M.B. and Stadler,B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J.*, **17**, 1141–1143.

Tanimoto,T.T. (1958) *An Elementary Mathematical Theory of Classification and Prediction*. IBM Research Yorktown Heights, New York.

Tomczak,J. (2003) DataTypes. In: Gastejger,J. (ed.) *Handbook of Chemoinformatics*. Vol. 2, Welley-VCH, Weinheim, Germany, pp. 392–409.

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

Wang,J. *et al.* (2013) Evaluation and integration of existing methods for computational prediction of allergens. *BMC Bioinformatics*, **14** (**Suppl. 4**), S1.

Willett,P. (2003) Similarity searching in chemical databases. In: Gastejger,J. (ed.) *Handbook of Chemoinformatics*. Vol. 2, Welley-VCH, Weinheim, Germany, pp. 904–915.

Zhang,Z.H. *et al.* (2007) AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics*, **23**, 504–506.

Zorzet,A. *et al.* (2002) Prediction of food protein allergenicity: a bio-informatic learning systems approach. *In Silico Biol.*, **2**, 525–534.