ORIGINAL PAPER

AllerTOP v.2—a server for in silico prediction of allergens

Ivan Dimitrov · Ivan Bangov · Darren R. Flower · Irini Doytchinova

Received: 11 December 2013 / Accepted: 25 April 2014

Abstract Allergy is an overreaction by the immune system to a previously encountered, ordinarily harmless substance — typically proteins—resulting in skin rash, swelling of mucous membranes, sneezing or wheezing, or other abnormal conditions. The use of modified proteins is increasingly widespread: their presence in food, commercial products, such as washing powder, and medical therapeutics and diagnostics, makes predicting and identifying potential allergens a crucial societal issue. The prediction of allergens has been explored widely using bioinformatics, with many tools being developed in the last decade; many of these are freely available online. Here, we report a set of novel models for allergen prediction utilizing amino acid *E*-descriptors, auto- and cross-covariance transformation, and several machine learning methods for classification, including logistic regression (LR), decision tree (DT), naïve Bayes (NB), random forest (RF), multilayer perceptron (MLP) and *k* nearest neighbours (*k*NN). The best performing method was *k*NN with 85.3 % accuracy at 5-fold cross-validation. The resulting model has been implemented in a revised version of the AllerTOP server (http://www.ddg-pharmfac.net/AllerTOP).

This paper belongs to Topical Collection MIB 2013 (Modeling Interactions in Biomolecules VI)

I. Dimitrov · I. Doytchinova (✉)
Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., 1000, Sofia, Bulgaria
e-mail: idoytchinova@pharmfac.net

I. Bangov
Faculty of Natural Sciences, Konstantin Preslavski Shumen University, 115 Universitetska Str., 9712 Shumen, Bulgaria

D. R. Flower
Life and Health Sciences, Aston University, Aston Triangle, Birmingham, B4 7ET, UK

Published online: 31 May 2014

Keywords Allergen prediction · *E*-descriptors · ACC transformation · Logistic regression · Decision tree · Naïve bayes · Random forest · Multilayer perceptrone · *k* Nearest neighbours

Introduction

Allergy is an overreaction of the immune system to a previously encountered, ordinarily harmless substance, usually a protein. Allergic reactions can result in skin rash, swelling of mucous membranes, sneezing or wheezing, or other abnormal conditions. Drugs, foods, environmental irritants, microorganisms, or physical conditions, such as temperature extremes, act as antigenic stimuli, provoking a significant immunological response involving the release of inflammatory substances, such as histamine, in the body. Allergies may be innate or acquired in genetically predisposed individuals. The increasingly widespread use of modified proteins in food; commercial products, such as latex gloves or washing powder; and medical therapeutics and diagnostics, makes the identification of allergens an important issue for manufacturers and consumers alike.

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 [1, 2]. According to these guidelines, a protein is a potential allergen if it has either an identity of 6 to 8 consecutive amino acids or greater than 35 % overall sequence similarity over a window of 80 amino acids when compared with known allergens [3].

Allergen prediction has been explored extensively by bioinformatics. Many bioinformatics tools have been developed during the last 10 years; many of them are available free online. Aller Hunter (http://tiger.dbs.nus.edu.sg/AllerHunter) is a cross-reactive allergen prediction program built on a combination of Support Vector Machine (SVM) and pair wise sequence similarity [4]. 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)]. The APPEL tool (Allergen Protein Prediction E-Lab) (http://jing.cz3.nus.edu.sg/cgi-bin/APPEL) finds novel allergen proteins using SVM and sequence-derived structural and physicochemical properties of a whole proteins [5]. 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 [6].

We have recently developed two powerful new allergen prediction methods. AllerTOP (http://www.pharmfac.net/allertop) is the first proper alignment-free server for in silico prediction of allergens based on the physicochemical properties of protein sequences [7]. AllerTOP v.1 utilizes a model based on amino acid z-descriptors, ACC protein sequence transformation, and k nearest neighbors (*k*NN) clustering. AllergenFP (http://www.ddg-pharmfac.net/AllergenFP) is another alignment-free tool for allergenicity prediction, which uses a variety of amino acid principal properties, such as hydrophobicity and β-strand forming propensities: potential allergens are first transformed into descriptor-based fingerprints and then compared using the Tanimoto coefficient [8].

In the present study, we describe the development of a set of novel models for allergen prediction that utilizes knowledge derived post hoc from analysis of our two recently developed servers: AllerTOP v.1 and AllergenFP. In the approach explored below, 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- and cross-covariance (ACC) transformation [9]. Finally, computational machine learning methods are applied to classify and differentiate allergens from non-allergens. Several methods for classification were tested and the best performing model was uploaded into a wholly new version of the server AllerTOP.

Methods

Protein datasets

The dataset used in the present study consisted of 2,427 allergens and 2,427 non-allergens. Initially, allergens were 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 the Allergome database (http://www.allergome.org/). Next, only allergens annotated with the label “evidence for the existence of protein—evidence at protein level” in Swiss-Prot database (http://www.uniprot.org) were selected. Duplicates were removed. This set is curated manually and contains only known allergens with evidence at protein level.

The non-allergens were collected from widely used food species: *Solanum lycopersicum* (tomato), *Capsicum annuum* (pepper), *Solanum tuberosum* (potato), *Triticum aestivum* (bread wheat) and *Oryza sativa* (Asian rice) and *Oryza glaberrima* (African rice) using searches in Swiss-Prot for proteins annotated with the label “evidence for the existence of protein—evidence at protein level”. Proteins containing the keyword “allergen” in their description were excluded. 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*, again labeled 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.

E-descriptors

The protein sequences of allergens and non-allergens were described by five *E*-descriptors [10]. They 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 propensity for β-strand formation.

Auto- and cross-covariance transformation

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

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

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

Table 1 *E*-Descriptors of amino acids [10]

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	−0.280	−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

acid position (i=1, 2, n). At the end of this step, the proteins were converted into strings of $5^2 \times lag$ ACC values.

Methods for classification used in the study

A variety of machine learning methods for classification were used in the present study. These methods were logistic regression (LR), decision tree (DT), naïve Bayes (NB), random forest (RF), multilayer perceptron (MLP) and *k* nearest neighbours (*k*NN). *k*NN and LR algorithms were applied as implemented in python scripts based on the Bio python module [11]. DT, NB, RF and MLP algorithms were applied using WEKA Data Mining Software [12].

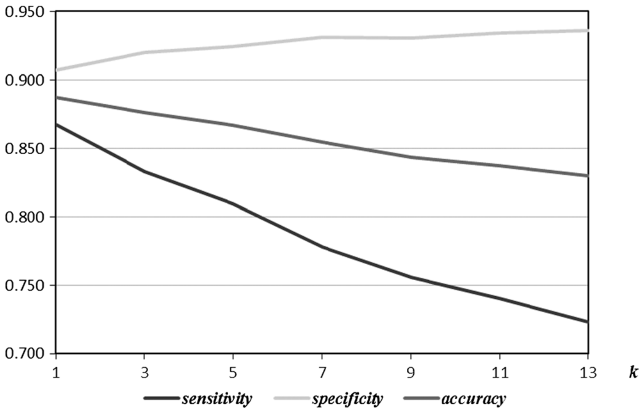

Fig. 1 Optimization of *k* nearest neighbours (*k*NN) method

Evaluation of performance

The allergen prediction algorithms developed in the present study were evaluated using 10-fold cross-validation (10CV). Correctly predicted allergens and non-allergens were defined as true positives (TP) and true negatives (TN), respectively. Incorrectly predicted allergens and non-allergens were defined as false negatives (FN) and false positives (FP), respectively. Using these values, *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 x TN – FP × FN)/((TP + FP)(TP + FN)(TN + FP)(TN + FN))$^{1/2}$] were calculated. Additionally, the best performing algorithm was cross-validated in five groups (5CV).

Results and discussion

A set of 2,427 allergens and 2,427 non-allergens was collected from different databases. Each amino acid in the protein sequences was presented by five *E*-descriptors, reflecting its hydrophobicity (*E1*), size (*E2*), helix-forming propensity (*E3*), relative abundance (*E4*) and β-strand forming propensity (*E5*). Venkatarajan and co-workers [10, 16] derived these descriptors by multidimensional scaling of 237 physical-

Table 2 Evaluation of the performance of six methods for allergen classification

Method	Sensitivity	Specificity	Accuracy	ppv	F1	MCC
Logistic regression (LR)	0.685	0.764	0.725	0.744	0.713	0.451
Decision tree (DT)	0.756	0.711	0.733	0.723	0.739	0.467
Naïve bayes (NB)	0.583	0.827	0.705	0.773	0.665	0.423
Random forest (RF)	0.754	0.868	0.811	0.851	0.799	0.625
Multilayer perceptrone (MLP)	0.715	0.773	0.744	0.759	0.736	0.489
k Nearest neighbours (*k*NN) with *k* =1	0.867	0.907	0.887	0.903	0.885	0.775

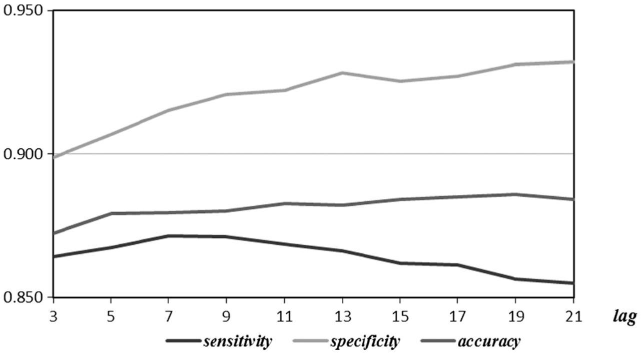

Fig. 2 *Lag* optimization

chemical properties of amino acids and compressing them to a five-dimensional property space. The *E*-descriptors correlate well with similarity indices derived from substitution matrices as PAM 250 [13] and BLOSUM 62 [14]. They are used widely for property-based motif search in proteins [15–20]. In the present study, the proteins were presented as *E*-descriptor-based numerical strings with different length. An ACC transformation was applied to convert these strings into uniform vectors as described in Methods. Further, the ACC matrix was used to derive models for allergen prediction.

Choice of a method for allergen prediction

Six methods for classification were applied to the protein dataset (allergens + non-allergens) to derive models for allergen prediction. The models were validated by 10-fold cross-validation. The methods used in the study were logistic regression (LR), decision tree (DT), naïve Bayes (NB), random forest (RF), multilayer perceptron (MLP) and *k* nearest neighbours (*k*NN). Here, we used *kNN* with *k*=1 and *lag* value for ACC 5 (Table 1).

The performance of the derived models was assessed by 10CV using *sensitivity*, *specificity*, *ppv*, *MCC*, and the *F1* score, with a threshold of 0.5. See Table 2. The best performing model was *k*NN. It recognized 86.7 % of the allergens, 90.7 % of the non-allergens yielding *ppv*=0.903, *F1*=0.885 and *MCC*=0.775. Table 2 indicates significant superiority for the *k*NN method for all measures.

Optimization of *k*NN method

Further, *k*NN models with different values for the number of nearest neighbours *k* were derived and tested using 10CV. Values for *k* were odd numbers in the range 1 to 13. The results are shown in Fig. 1. As *k* increases, the *sensitivity* decreases, *specificity* slightly increases and the overall *accuracy* decreases. Thus, the optimal value for *k* was found to be 1.

Lag optimization

The *k*NN model with *k*=1 was optimized in terms of *lag* length changing *lag* from 3 to 21 with a step of 2. The performances of the resulting models were compared using *sensitivity*, *specificity* and *accuracy* (Fig. 2). *Sensitivity* initially increases, reaches a maximum at *lag*=7 and then decreases slightly. *Specificity* slightly increases as the lag increases. The overall *accuracy* increases slightly up to *lag*=5, and then only increases marginally. Thus, *lag*=5 was chosen as an optimum value.

Cross-validation

The *k*NN model with *k*=1 and *lag*=5 was cross-validated in five groups. The whole set of allergens and non-allergens was divided into five groups. Four of them formed a training set, the fifth was used as a test set. The average values for *sensitivity*, *specificity* and *accuracy* were 0.825, 0.881 and 0.853, respectively.

AllerTOP v.2 server

A model based on the total set of allergens and non-allergens was derived by the *k*NN algorithm with *k*=1 and *lag*=5, and

Table 3 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	0.748
AlgPred(SVM_single_aa)	0.894	0.657	0.775	0.723	0.799	0.575
AlgPred(SVM_dipeptide)	0.866	0.726	0.796	0.760	0.809	0.612
AlgPred(ARP)	0.730	0.953	0.842	0.940	0.822	0.717
APPEL	0.653	0.914	0.783	0.883	0.751	0.613
ProAp(motif)	0.938	0.072	0.505	0.503	0.655	−0.006
ProAp(SVM)	0.813	0.874	0.843	0.866	0.839	0.703
AllerTOP v.1	0.876	0.780	0.828	0.799	0.836	0.671
AllergenFP	0.868	0.891	0.879	0.889	0.878	0.765
AllerTOP v.2	0.867	0.907	0.887	0.903	0.885	0.775

made freely accessible via a revised version of the server AllerTOP. AllerTOP v.2 is implemented in Python, with a GUI written in HTML. 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 neighbour in the training set. On this basis, AllerTOP v.2 defines the most probable route of exposure of tested proteins predicted as an allergen. AllerTOP v.2 can be accessed via the URL: http://www.ddg-pharmfac.net/AllerTOP.

Comparison of AllerTOP v.2 to existing servers for allergen prediction

The performance of AllerTOP v.2 was compared to the nine freely available web servers using the total set of 2,427 allergens and 2,427 non-allergens (Table 3). Servers were compared in terms of *sensitivity*, *specificity*, *ppv*, *F1* and *MCC* after 10CV.

The highest *sensitivity* was achieved by ProAp(motif) (93.8 %), followed by AlgPred(SVM_single_aa) (89.4 %) and AllerTOP v.1 (87.6 %). Aller Hunter has the highest *specificity* (96.0 %), closely followed by AlgPred(ARP) (95.3 %) and APPEL (91.4 %). Importantly, the crucial statistic of overall *accuracy* confirms that AllerTOP v.2 is the best performing method (88.7 %), followed by AllergenFP (87.9 %) and Aller Hunter (87.1 %). In terms of *ppv*, Aller Hunter is the best (0.951), followed by AlgPred(ARP) (0.940) and AllerTOP v.2 (0.903). Using the *F1* statistic, AllerTOP v.2 is again the best performing method (0.885), followed by AllergenFP (0.878) and Aller hunter (0.858 %)). Finally, the *MCC* values show that AllerTOP v.2 is the best performing (0.775), followed by AllergenFP (0.765) and Aller Hunter (0.748).

While AllerTOP v.2 is not first in each statistic, it performs well for each measure and does not exhibit the high variability that characterizes most of the other servers. Given that our evaluation was undertaken using 10CV, the encouraging consistency demonstrated by AllerTOP v.2 is likely indicative of probable robustness in the face of new data, and its potential utility as an allergen prediction engine. AllerTOP v.2 also demonstrates considerable improvement over version 1. Comparing the two versions of AllerTOP, it is clear for example that the second version is better in terms of *accuracy*, *ppv*, *F1* and *MCC*. Immunonformatic prediction methods typically combine biological data (here, an allergen versus non-allergen classification, plus sequence data) with an appropriate data representation (here, conversion of protein sequences into numerical values using *E*-descriptors, plus normalization using ACC transformation), and model induction (here, machine learning approaches). It is well established that model induction methods are so numerous and of such quality that data representation is the key area for advancement—a sentiment borne out by the current study, where adopting *E*-descriptors has led to a significant improvement in performance. It is likely that future development will increase both the scope and range of biological data and enhance further data representation as a means to develop even more accurate allergen prediction methods.

Conclusions

In the present study we derived a model for allergen prediction based on amino acid descriptors, accounting for residue hydrophobicity, size, abundance, helix- and β-strand forming propensities. The protein strings were transformed into uniform vectors by auto- and cross-covariance and a machine learning method using *k* nearest neighbours was used to classify allergens and non-allergens. The model is implemented in a fully revised version of the AllerTOP server, which is freely accessible at http://www.ddg-pharmfac.net/AllerTOP. The comparison between several servers for allergen prediction indicates that AllerTOP v.2 has the highest *accuracy*. AllerTOP v.2 offers a useful, robust, and strongly complimentary approach to allergen prediction that should provide researchers with important and persuasive new approach to identifying allergens in both existing and newly developed materials.

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

References

FAO/WHO Agriculture and Consumer Protection (2001) Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, Rome

FAO/WHO Codex Alimentarius Commission (2003) Codex principles and guidelines on foods derived from biotechnology. Joint FAO/WHO Food Standards Programme, Rome

Stadler MB, Stadler BM (2003) FASEB J 17:1141–1143

Zorzet A, Gustafsson M, Hammerling U (2002) In Silico Biol 2:525–534

Cui J, Han LY, Li H, Ung CY, Tang ZQ, Zheng CJ, Cao ZW, Chen YZ (2007) Mol Immunol 44:514–520

Wang J, Yu Y, Zhao Y, Zhang D, Li J (2013) BMC Bioinforma 14(4):S1

Dimitrov I, Flower DR, Doytchinova I (2013) BMC Bioinforma 14(6):S4

Dimitrov I, Naneva L, Doytchinova I, Bangov I (2014) Bioinformatics 30(6):846–851

Nyström Å, Andersson PM, Lundstedt T (2000) Quant Struct-Act Relat 19:264–269

Venkatarajan MS, Braun W (2001) J Mol Model 7:445–453

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ (2009) Bioinformatics 25:1422–1423

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) SIGKDD Explorations 11:10–18

Dayhoff MO, Schwartz RM, Orcutt BC (1978) In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, DC, pp 345–352

Henikoff S, Henikoff J (1992) Proc Natl Acad Sci USA 89:10915–10919

Schein CH, Ozgun N, Izumi T, Braun W (2002) BMC Bionformatics 3:37

Venkatarajan MS, Schein CH, Braun W (2003) Bioinformatics 19:1381–1390

Schein CH, Zhou B, Braun W (2005a) Virol J 2:40

Schein CH, Zhou B, Oezguen N, Mathura VS, Braun W (2005b) Proteins 58:200–210

Negi SS, Braun W (2007) J Mol Model 13:1157–1167

Ivanciuc O, Braun W (2007) Protein Pept Lett 14:903–916