



An Alignment-Independent Platform for Allergenicity Prediction

Ivan Dimitrov and Irini Doytchinova

Abstract

A great number of novel proteins have been generated from new sources and genetically modified foods during the last decade. As the allergenicity of these proteins is of particular importance for their safe usage, fast and reliable screening strategies for allergenicity assessment are required. The WHO/FAO guidance directs to structural similarities between the novel proteins and known allergens detected by sequence alignment. However, the allergic response involves conformational IgE epitopes that are undetectable by sequence alignment. Here, we present a protocol for allergenicity prediction based on a platform of three alignment-independent servers developed in our lab: AllerTOP v.1, AllerTOP v.2, and AllergenFP. The servers use similar datasets but different chemical descriptors and methods to derive models for allergenicity prediction. The platform is freely accessible and user-friendly. The protocol is demonstrated stepwise on a randomly chosen query protein.

Key words Allergenicity, Allergenicity prediction, Physicochemical properties of amino acids, Alignment-independent methods

1 Introduction

Allergy is a type of hypersensitivity of the human immune system to normally innocuous substances, such as dust, pollen, foods, or drugs. Allergens are small antigens that commonly provoke human immunoglobulin E (IgE) antibody response. As the number of novel proteins generated from new sources and genetically modified foods increased significantly during the last decade, a fast and reliable screening strategy for allergenicity assessment is required. According to the guidance of the World Health Organization (WHO) and Food and Agriculture Organization (FAO) of the United Nations, the allergenicity assessment is based on structural similarities between the novel proteins and known allergens detected by sequence alignment [1]. However, the allergic response involves conformational IgE epitopes that are undetectable by sequence alignment.

Recently, we developed three alignment-independent models for allergenicity prediction. They were implemented in three modular servers: AllerTOP v.1 [2], AllerTOP v.2 [3], and AllergenFP [4]. AllerTOP v.1 uses a dataset of 2395 known allergens as a positive training set and a set of 2395 non-allergens from the same species as a negative training set. The protein sequences are encoded by three amino acids z -scores [5], transformed in uniform vectors by auto- and cross-covariance (ACC) [6] and analyzed by k nearest neighbor (k NN) algorithm. The model recognizes 94% of the allergens and 94% of the non-allergens in the external test set [2].

AllerTOP v.2 is an updated version of AllerTOP v.1. It uses an updated set of 2427 known allergens and set of 2427 non-allergenic proteins from widely used foods species and non-immunogenic human proteins. The data processing is similar to that used in AllerTOP v.1: protein sequences are encoded by five amino acid E -descriptors [7], transformed by ACC and analyzed by k NN classifier. The model recognizes 87% of the allergens and 91% of the non-allergens in the external test set.

The AllergenFP model is derived on the AllerTOP v.2 training set of transformed protein sequences represented as binary fingerprints. The predictions are made by comparing the fingerprint of the tested protein to the fingerprints of the proteins from the training set. The similarity is assessed by Tanimoto index [8]. This model recognizes 87% of the allergens and 89% of the non-allergens in the external test set.

Here, we utilize the three modular servers to assess the allergenicity of a randomly selected allergenic protein and demonstrate a protocol for fast and reliable *in silico* prediction. The protocol could be automated by a simple script and used for *in silico* screening of novel proteins.

2 Databases and Servers

2.1 *Swiss-Prot*

Swiss-Prot is the manually annotated and reviewed database of the Universal Protein Knowledgebase (UniProtKB) (<http://www.uniprot.org>). It contains more than 50,000 protein sequences (July 2019).

2.2 *NCBI (National Center for Biotechnology Information) Protein Database*

The NCBI Protein database (<https://www.ncbi.nlm.nih.gov/>) is a collection of sequences from several sources including translations from annotated coding regions in other NCBI databases like GenBank, RefSeq, and TPA as well as records from other databases like SwissProt, PIR, PRF, and PDB.

2.3 AllerTOP v.1

AllerTOP is an alignment-independent server for in silico prediction of allergens based on the main physicochemical properties of the amino acid residues in proteins. The dataset of allergens consists of 84 food, 1156 inhalant, and 555 toxin (venom or salivary) allergenic proteins collected from several databases. The set of non-allergens consists of proteins from the same species selected by BLAST search tailored to mirror the allergen set.

The protein sequences were encoded by three amino acid z -scores accounting for hydrophobicity, size, and polarity. As the proteins were of different length, ACC transformation was applied to derive a set of uniform vectors. The set was analyzed by k NN algorithm and the model was optimized in terms of lag length and k neighbors and tested on external set. The highest accuracy of 94% was achieved at lag of 5 and k of 3 [2]. Additionally, AllerTOP is able to predict the route of allergen exposure: food, inhalant, or toxin.

2.4 AllerTOP v.2

AllerTOP v.2 model was developed on an updated dataset of 2427 allergens and 2427 non-allergens. The set of allergens was collected from several databases containing allergens of different origin and route of exposure. The set of non-allergens was collected from widely used food species like tomato, pepper, potato, bread wheat, rice, and human non-immunogenic proteins.

The protein sequences were encoded by five E -descriptors accounting for hydrophobicity, size, helix- or β -strand forming propensities, and relative abundance. After ACC transformation, the dataset was analyzed by k NN algorithm. The best performing model showed accuracy of 85% at $k = 1$ [3].

2.5 AllergenFP

AllergenFP uses the same ACC transformed dataset as AlerTOP v.2. Each vector from the dataset is encoded as a binary fingerprint. The tested protein is presented as a string of E -descriptors, transformed by ACC and coded binary to generate a fingerprint. The fingerprint is compared to the protein fingerprints from the AllergenFP dataset by Tanimoto similarity index. The protein showing the highest similarity with the tested protein predicts the allergenicity/non-allergenicity of the tested protein. The external validation of the model gave 88% accuracy [4].

3 Methods

In the present protocol, the three modular servers described above are used to predict the allergenicity of a randomly chosen query protein. As a query protein is selected the allergen **AHZ10957.1** (Fra a 1.02) from *Fragaria x ananassa subsp. ananassa* (strawberry). It contains 160 residues, is proven experimentally to be an allergen [9], and does not present in the datasets used to derive the models in the three servers.

3.1 Allergenicity Prediction by AllertOP v.1

1. Retrieve the protein sequence for Fra a 1.02 from UniProtKB (entry: A0A024B3G5) or GenBank (**AHZ10957.1**).
2. Download or copy the sequence in FASTA format.
3. Open AllertOP v.1 (URL: <http://www.ddg-pharmfac.net/allertop>).
4. Enter the protein sequence in plain format (removing the title line: delimiter >).
5. Press “Get the Result” button.
6. The AllertOP v.1. RESULTS page lists the statement: “PROBABLE ALLERGEN” (Fig. 1) and a list of the three nearest proteins presented by accession number, hyperlink to UniProtKB or NCBI protein database, and class allergen or non-allergen. In our case, the three nearest neighbors are:
 - (a) 60280861 defined as allergen—major allergen Mal d 1.03F from *Malus domestica* (apple) [10] with accession number in the NCBI protein database

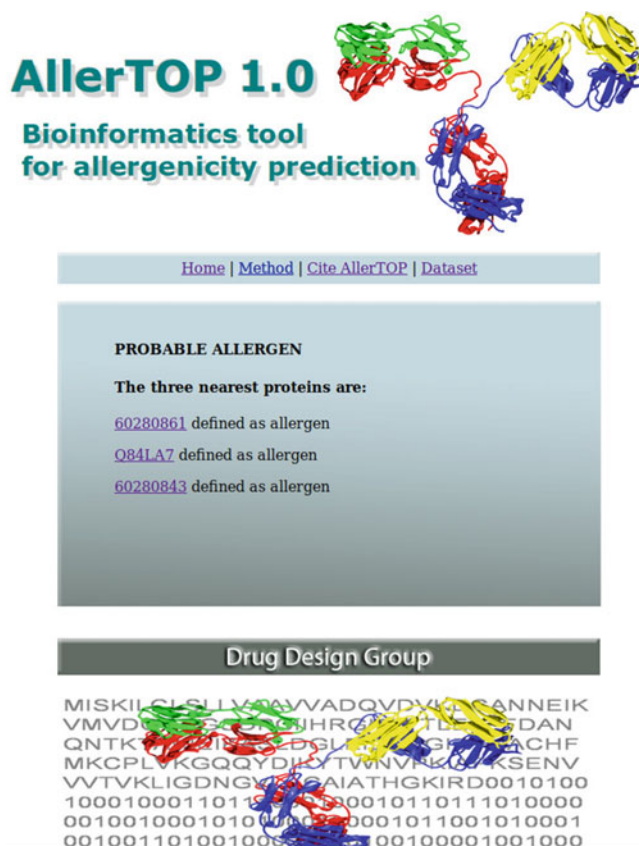


Fig. 1 The AllertOP v.1. RESULTS page for the query protein **AHZ10957.1** from *Fragaria x ananassa subsp. ananassa* (strawberry)

- (b) Q84LA7 defined as allergen—major allergen Mal d 1 from *Malus domestica* (apple) [11] with accession number in the UniprotKB database.
- (c) 60280843 defined as allergen—major allergen Mal d 1B from *Malus domestica* (apple) [10] with accession number in the NCBI protein database.

3.2 Allergenicity Prediction by AllertOP v.2

1. Retrieve the protein sequence for Fra a 1.02 from UniProtKB (entry: A0A024B3G5) or GenBank (AHZ10957.1).
2. Download or copy the sequence in FASTA format.
3. Go to AllertOP v.2 homepage (URL: <http://www.ddg-pharmfac.net/AllertOP>).
4. Enter the protein sequence in plain format (removing the title line: delimiter >).
5. Press “Get the Result” button.
6. The AllertOP v.2. RESULTS page lists the statement: “Your sequence is PROBABLE ALLERGEN. The nearest protein is NCBI gi number 60280861 (hyperlink) defined as allergen” (Fig. 2). Here again, the nearest neighbor is the major allergen Mal d 1.03F from *Malus domestica* (apple) [10].

3.3 Allergenicity Prediction by AllergenFP

1. Retrieve the protein sequence for Fra a 1.02 from UniProtKB (entry: A0A024B3G5) or GenBank (AHZ10957.1).
2. Download or copy the sequence in FASTA format.

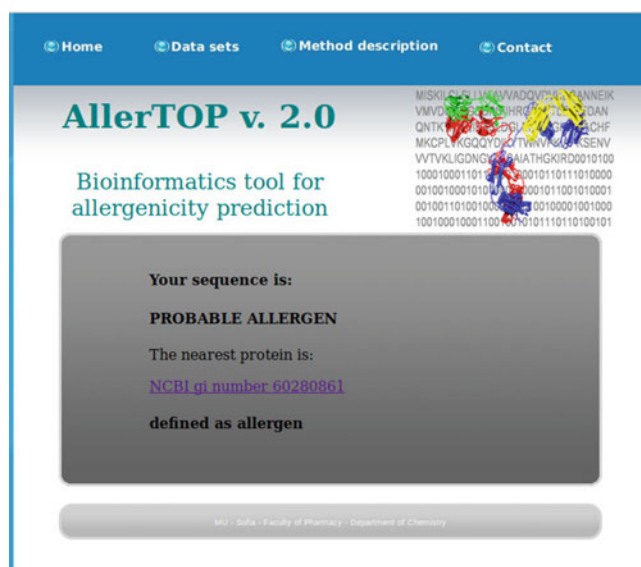


Fig. 2 The AllertOP v.2. RESULTS page for the query protein AHZ10957.1 from *Fragaria x ananassa* subsp. *ananassa* (strawberry)

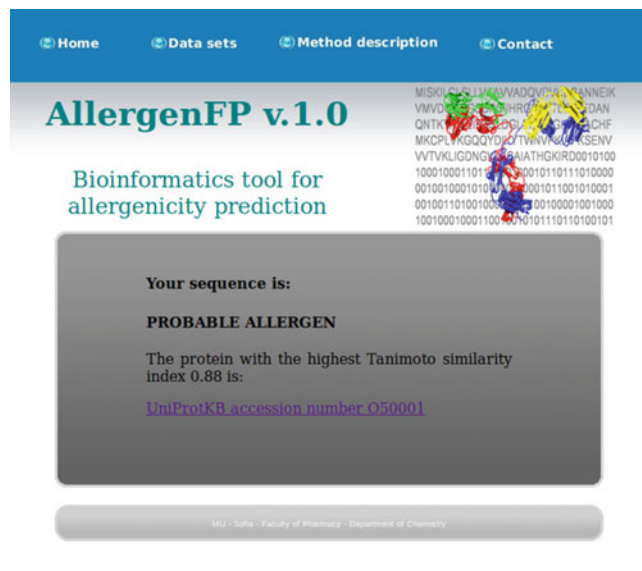


Fig. 3 The AllergenFP RESULTS page for the query protein **AHZ10957.1** from *Fragaria x ananassa subsp. ananassa* (strawberry)

3. Go to AllergenFP homepage (URL: <http://www.ddg-pharmfac.net/AllergenFP>).
4. Enter the protein sequence in plain format (removing the title line: delimiter >).
5. Press “Get the Result” button.
6. The AllergenFP RESULTS page lists the statement: “Your sequence is PROBABLE ALLERGEN. The protein with the highest Tanimoto similarity index 0.88 is UniProtKB accession number O50001 (hyperlink)” (Fig. 3). Here, the protein with the highest Tanimoto index is the major allergen Pru ar 1 from *Prunus armeniaca* (Apricot) (*Armeniaca vulgaris*) [12].

3.4 Conclusion

The decision on the allergenicity/non-allergenicity of the query protein is based on the predictions made by the three servers. In our case, all the nearest proteins are defined as allergens. Hence, the query protein is predicted to be an allergen as well. In the general case, if only one of the nearest proteins is defined as allergen, the tested protein should be considered as a probable allergen.

The platform for allergenicity assessment presented here is freely accessible and user-friendly. The three servers work complementary and the platform could be considered as an expert system for prediction of protein allergenicity.

Acknowledgments

This work has been accomplished with the financial support by the Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014-2020), and co-financed by the European Union through the European structural and investment funds.

References

1. FAO/WHO (2009) Foods derived from modern biotechnology (annex: assessment of possible allergenicity, Codex Alimentarius—joint FAO/WHO food standards), 2nd edn. WHO, Geneva
2. Dimitrov I, Flower DR, Doytchinova I (2013) AllerTOP—a server for in silico prediction of allergens. *BMC Bioinformatics* 14(Suppl.6):S4
3. Dimitrov I, Bangov I, Flower DR, Doytchinova I (2014) AllerTOP v.2—a server for in silico prediction of allergens. *J Mol Model* 20:2278
4. Dimitrov I, Naneva L, Bangov I, Doytchinova I (2014) AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics* 30:846–851
5. Hellberg S, Sjöström M, Skagerberg B, Wold S (1987) Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem* 30:1126–1135
6. Wold S, Jonsson J, Sjöström M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal components analysis and partial least squares projections to latent structures. *Anal Chim Acta* 277:239–253
7. Venkatarajan MS, 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
8. Tanimoto TT (1958) An elementary mathematical theory of classification and prediction. IBM Research Yorktown Heights, New York
9. Franz-Oberdorf K, Eberlein B, Edelmann K, Hücherig S, Besbes F, Darsow U, Ring J, Schwab W (2016) Fra a 1.02 is the most potent isoform of the bet v 1-like allergen in strawberry fruit. *J Agric Food Chem* 64:3688–3696
10. Gao ZS, van de Weg WE, Schaart JG, Schouten HJ, Tran DH, Kodde LP, van der Meer IM, van der Geest AHM, Kodde J, Breiteneder H, Hoffmann-Sommergruber K, Bosch D, Gilissen LJWJ (2005) Genomic cloning and linkage mapping of the mal d 1 (PR-10) gene family in apple (*Malus domestica*). *Theor Appl Genet* 111:171–183
11. Hsieh LS, Moos M, Lin Y (1995) Characterization of apple 18 and 31 kd allergens by microsequencing and evaluation of their content during storage and ripening. *J Allergy Clin Immunol* 96:960–970
12. Scheurer S, Pastorello EA, Wangorsch A, Kästner M, Haustein D, Vieths S (2001) Recombinant allergens Pru av 1 and Pru av 4 and a newly identified lipid transfer protein in the in vitro diagnosis of cherry allergy. *J Allergy Clin Immunol* 107:724–731