J.Natn.Sci.Foundation Sri Lanka 2018 46 (2): 233 - 236
DOI: http://dx.doi.org/10.4038/jnsfsr.v46i2.8424

RESEARCH COMMUNICATION

WebAllergen: a web-based database for protein allergenicity prediction

So Youn Won[1], Jeong-Ho Baek[2], Jae-Hyeon Oh[1], Gang-Seob Lee[3], Yong-Hwan Kim[4] and Chang-Kug Kim[1*]

[1] *Genomics Division, National Institute of Agricultural Sciences, Rural Development Administration, Jeonju 54874, Korea.*
[2] *Gene Engineering Division, National Institute of Agricultural Sciences, Rural Development Administration, Korea.*
[3] *Biosafety Division, National Institute of Agricultural Sciences, Rural Development Administration, Korea.*
[4] *Department of Crop Science and Biotechnology, Dankook University, Cheonan 31116, Korea.*

Revised: 30 October 2017; Accepted: 18 January 2018

Abstract: Allergies are an important health problem. In the present study, a web-based allergen platform was developed with an allergen database and allergenicity prediction functions. Drawing from the literature and public databases, 2,939 allergens were identified and categorised according to their origin and known information. This platform provides a function to search allergenic proteins through formats such as keywords, FASTA, BLAST, and provides sequence-based, motif-based and epitope-based methods for allergenicity prediction. Using specific sequence or allergen predictions, the user can find summarised allergen information to link the UniProtKB and the PDB databases.

Keywords: Allergen database, allergenicity prediction, WebAllergen.

INTRODUCTION

Allergy (or allergies) refers to a condition in which a person exhibits a hypersensitive immune response to allergens. Allergen databases provide allergen information and the bioinformatics tools for allergenicity prediction (Ladics *et al*., 2011). These databases contain the criteria for including allergens, data shown for each allergen, and the availability of bioinformatics tools. Many tools for allergen prediction have been developed (Mari *et al*., 2009). The guidelines of FAO/WHO suggest that allergenicity should be determined by the degree of amino acid sequence similarity between a suspected and known allergen (FAO/WHO, 2001). The prediction methods rely on sequence similarity to allergen-specific peptide/motifs, and conserved structural patterns as evidenced by IgE binding epitopes and cross reactivity (Sircar *et al*., 2014; Radauer, 2017). In this study, information on allergens was collected from literature and public allergen databases, and a tool for allergenicity prediction was developed. WebAllergen (*http://nabic.rda.go.kr/NabicAllergen*) provides a platform of allergen information and makes it practical for allergen prediction in applications such as biotechnology-derived gene screening and allergen discovery from specific sequences.

METHODOLOGY

Allergen information was collected from international journal literature and public allergen databases (Goodman *et al*., 2016). All the data were curated by the Allergen Expert Committee of Rural Development Administration (RDA, *http://www.rda.go.kr/*). A total of 2,939 allergens were classified into 3 groups based on whether they were of animal, microbial, or plant origin, and then into 13 subgroups (Table 1). The architecture of the platform consists of three sub-platforms, specifically: User interface, Front-end, and Database. The platform has a multi-layer architecture and is connected to other local database layers (Figure 1).

---

[*] Corresponding author (chang@korea.kr; https://orcid.org/0000-0002-8770-1736)

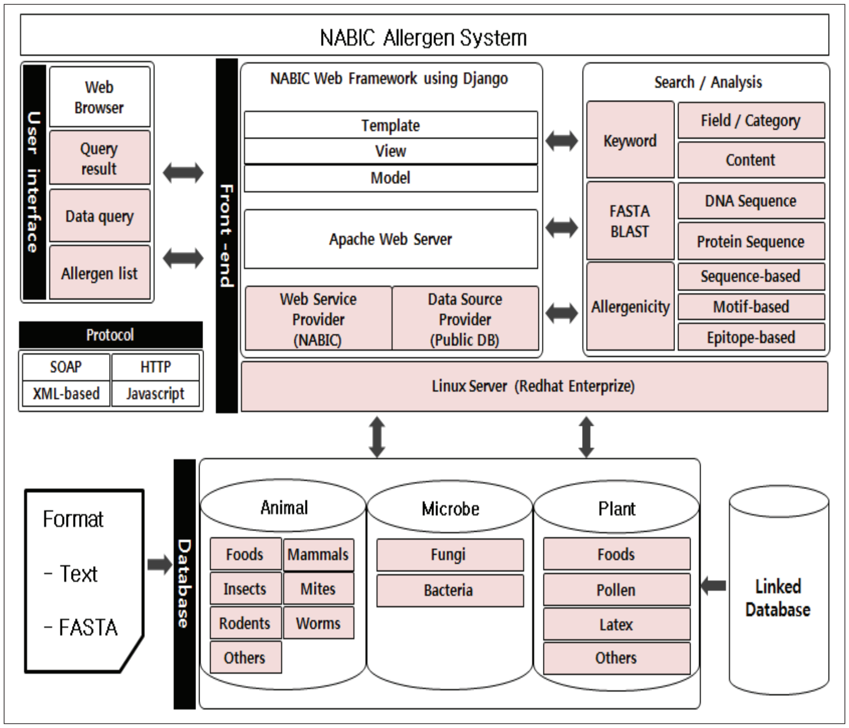


Figure 1: Overview of the allergen search and allergenicity platform. The platform consists of three systems including User interface, Front-end, and Database. The platform architecture was designed using various open standard protocols. SOAP: simple object access protocol; HTTP: hyper text transfer protocol; XML: extensible markup language

Table 1: The overview of classification of allergens in the WebAllergen database

Group	Category	Number of allergens
Animal	Foods	173
	Mammals	27
	Insects	187
	Mites	207
	Rodents	39
	Worms	68
	Others	152
Microbe	Fungi	223
	Bacteria	499
Plant	Foods	751
	Pollen	461
	Latex	40
	Others	112
Total		2,939

(Source: *http://nabic.rda.go.kr/NabicAllergen*)

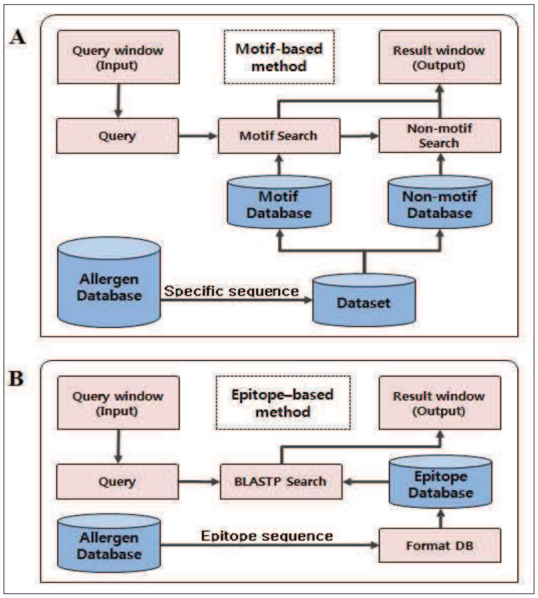


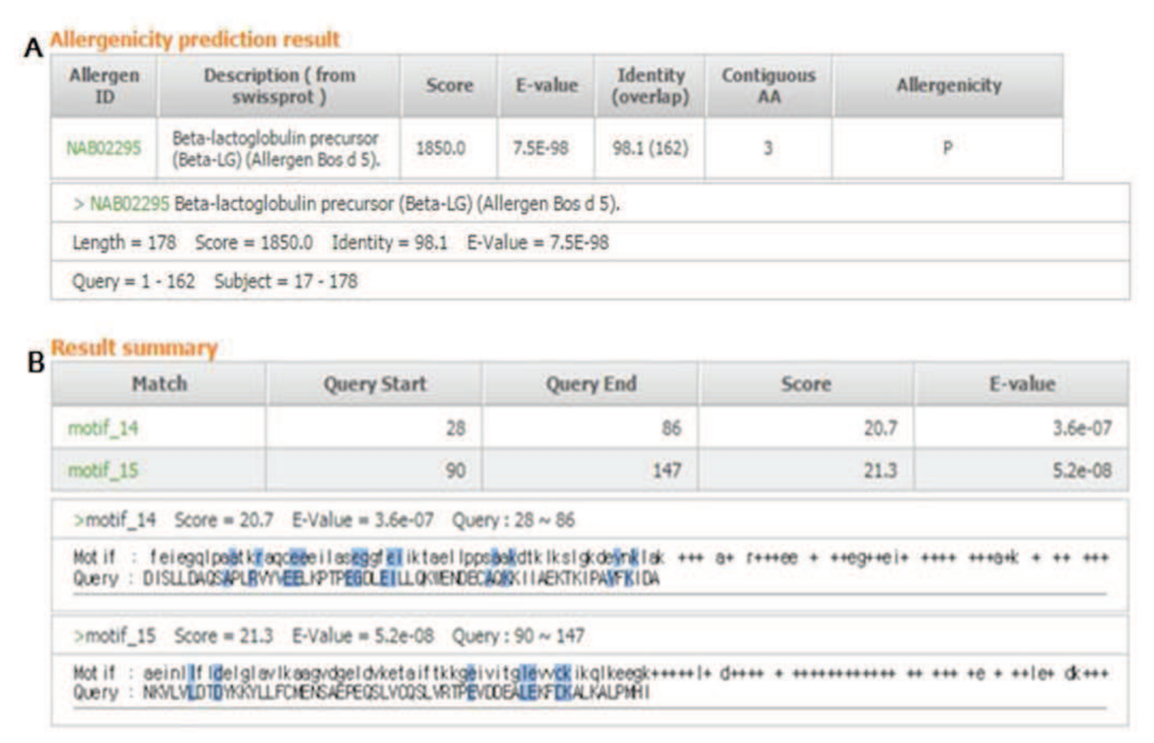
Figure 2: Overview of the allergen prediction methods. (A) Motif-based method; (B) epitope-based method. BLASTP: protein-protein BLAST

Three public algorithms were used for allergenicity prediction. The first method was in accord with the guideline proposed by FAO-WHO consultation group in 2001 (FAO/WHO, 2001). This guideline determines allergenicity based on the similarity of the primary structure of a protein with the primary structure of known allergenic proteins. The second method was based on the conserved motif approach, also called the motif-based method. The pipeline was developed to detect motifs of more than 30 amino acids that were present in well-known allergen proteins, and to identify these motifs using Hidden Markov model profiles (Figure 2A). Finally, an epitope-based method was developed to compare epitope information and query sequence with BLASTP (Figure 2B).

RESULTS AND DISCUSSION

WebAllergen provides a summarised list, three search methods (i.e., keyword, FASTA and BLAST) and three prediction methods (i.e., sequence, motif and epitope). The allergen list provides the summarised table of allergen information with structure and epitopes. For allergen search, users can search antigens through three formats. In the keyword menu, users can input antigen identification, the accession number, name of the gene or allergen, description, or species; and the search can be restricted to the 13 categories. In the other menu, users can use the FASTA format or BLAST programme to identify homologous allergenic proteins and can extend the search by setting an e-value for each programme.

In order to predict allergenicity, three menus have been provided such as the sequence, motif and epitope methods. In browsing the results table, users can predict the allergenic structure of query protein by changing related parameters and can access the individual characterisation information tables by clicking a link to the respective hypertext. The sequence-based method provides the function to find the query protein for an 80 amino acids sliding window according to the FAO/WHO guideline. This method yields a summarised



Allergenicity prediction result

Allergen ID	Description (from swissprot)	Score	E-value	Identity (overlap)	Contiguous AA	Allergenicity
NAB02295	Beta-lactoglobulin precursor (Beta-LG) (Allergen Bos d 5).	1850.0	7.5E-98	98.1 (162)	3	P

> NAB02295 Beta-lactoglobulin precursor (Beta-LG) (Allergen Bos d 5).
Length = 178 Score = 1850.0 Identity = 98.1 E-Value = 7.5E-98
Query = 1 - 162 Subject = 17 - 178

Result summary

Match	Query Start	Query End	Score	E-value
motif_14	28	86	20.7	3.6e-07
motif_15	90	147	21.3	5.2e-08

```
>motif_14   Score = 20.7   E-Value = 3.6e-07   Query : 28 ~ 86
Motif  :  feiegqlpsatkfaqceeeilasegqfelikteellppsaskdtklkslgkdesnklak  +++  a+  r+++ee  +  ++eg++ei+  ++++  +++a+k  +  ++  +++
Query  :  DISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENDECAQKKIIAEKTKIPAVFKIDA

>motif_15   Score = 21.3   E-Value = 5.2e-08   Query : 90 ~ 147
Motif  :  aeinllflldelglavlkaegvdgeldvketaiftkkgeivitglevvckikqlkeegk+++++l+  d++++  +  +++++++++++  ++  +++  +e  +  ++le+  dk+++
Query  :  NKVLVLDTDYKKYLLFCMENSAEPEQSLVCQCLVRTPEVDDEALEKFDKALKALPMHI
```

Figure 3: The results of allergenicity prediction with the beta-lactoglobulin protein in the animals (UniProtKB: P02754). The individual panels show results from the analysis by (A) the sequence-based method and (B) the motif-based method.

table, which includes the allergen ID, description, score, e-value, identity, contiguous amino acids (AA), and allergenicity (Figure 3A). Motif-based prediction provides the function to compare the two sequences with the specific profiles between the conserved motifs and the unique sequences (Riaz *et al*., 2005). If a specific protein is predicted to be allergenic, query results will be displayed to compare against a set of well-known allergenic motifs (Figure 3B). The epitope-based method provides the function to compare the epitope database and query sequence. If a given sequence has a region identical to known epitopes, the protein will be identified as a potential allergen.

For comparison of the three prediction methods, 100 randomly selected proteins from the literature were tested. In previous studies, Li *et al*. (2004), and Saha and Raghara (2006) have investigated the performance of prediction methods and observed that it is difficult to compare these methods. Although the present study did not find a significant difference between the three methods due to small sample, the results show that the sequence-based method had higher allergen precision than the other two methods. However, the motif-based and epitope-based methods detected the most positive allergenicity on the specific sequence. Therefore, the sequence-based method identified the strongest allergen, and the other two methods are more efficient in the case of specific proteins as motif and antigen's amino acid sequence.

CONCLUSION

WebAllergen is a platform to help biologists conduct efficient research in the field of allergy. Users can navigate the metadata and allergen structure by searching a specific allergen with keywords or sequences of DNA or protein. In addition, users can find summarised allergen information to link the UniProtKB (UniProt Knowledgebase, *http://www.uniprot.org/uniprot/*) and the PDB (Protein Data Bank, *http://www.rcsb.org/pdb/*) databases. Our platform may provide the basis for future selection strategies, potentially breeding for the development of allergen-free varieties. In future studies, it will also be broadly applicable for areas as diverse as the development of vaccines to genetically modified crops.

Acknowledgement

This study was conducted with support from the Research Programme for Agricultural Science and Technology Development (Project No. PJ010112) and the Next-Generation BioGreen 21 Programme (SSAC, Grant No. PJ011650), RDA, Republic of Korea.

REFERENCES

1. FAO/WHO. (2001). Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, Rome, Italy.
2. Goodman R.E. *et al*. (11 authors) (2016). AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Molecular Nutrition and Food Research* **60**(5): 1183 – 1198. DOI: https://doi.org/10.1002/mnfr.201500769
3. Ladics G.S., Cressman R.F., Herouet-Guicheney C., Herman R.A., Privalle L., Song P., Ward J.M. & McClain S. (2011). Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. *Regulatory Toxicology and Pharmacology* **60**(1): 46 – 53. DOI: https://doi.org/10.1016/j.yrtph.2011.02.004
4. Li K.B., Issac P. & Krishnan A. (2004). Predicting allergenic proteins using wavelet transform. *Bioinformatics* **20**(16): 2572 – 2578. DOI: https://doi.org/10.1093/bioinformatics/bth286
5. Mari A., Rasi C., Palazzo P. & Scala E. (2009). Allergen databases: current status and perspectives. *Current Allergy and Asthma Reports* **9**(5): 376 – 383. DOI: https://doi.org/10.1007/s11882-009-0055-9
6. Radauer C. (2017). Navigating through the jungle of allergens: features and applications of allergen databases. *International Archives of Allergy and Immunology* **173**(1): 1 – 11. DOI: https://doi.org/10.1159/000471806
7. Riaz T., Hor H.L., Krishnan A., Tang F. & Li K.B. (2005). WebAllergen: a web server for predicting allergenic proteins. *Bioinformatics* **21**(10): 2570 – 2571. DOI: https://doi.org/10.1093/bioinformatics/bti356
8. Saha S. & Raghava G. (2006). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Research* **34**(2): W202 – W209. DOI: https://doi.org/10.1093/nar/gkl343
9. Sircar G., Sarkar D., Bhattacharya S.G. & Saha S. (2014). Allergen databases. *Methods in Molecular Biology* **1184**: 165 – 181. DOI: https://doi.org/10.1007/978-1-4939-1115-8_9