

Databases and ontologies

WebAllergen: a web server for predicting allergenic proteins

Tariq Riaz, Hen Ley Hor, Arun Krishnan, Francis Tang and Kuo-Bin Li*

Bioinformatics Institute, 30 Biopolis Street, Singapore 138671

Received on November 17, 2004; revised on February 20, 2005; accepted on February 24, 2005

Advance Access publication March 3, 2005

ABSTRACT

Summary: WebAllergen is a web server that predicts the potential allergenicity of proteins. The query protein will be compared against a set of prebuilt allergenic motifs that have been obtained from 664 known allergen proteins. The query will also be compared with known allergens that do not have detectable allergenic motifs. Moreover, users are allowed to upload their own allergens as alternative training sequences on which a new set of allergenic motifs will be built. The query sequences can also be compared with these motifs.

Availability: <http://weballergen.bii.a-star.edu.sg/>

Contact: kuobin@bii.a-star.edu.sg

INTRODUCTION

Different types of food, pollen or dust mites are common sources of allergens. The ability to predict allergens becomes important these days because of the introduction of genetically modified foods (Taylor, 2002) and new modified proteins aimed at therapeutic usage (Lee and Sinko, 2000). The key step to a protein specific allergic reaction is the binding of IgE antibodies to an allergen (Bredehorst and David, 2001). Here the immune system recognizes peptide units within allergens. We introduce a prediction server that extracts protein motifs that are important for immune responses from known allergens. The server then compares user's query proteins with these motifs to determine the potential allergenicity.

World Health Organization (WHO) and the Food and Agriculture Organization (FAO) proposed guidelines to assess the potential allergenicity of proteins (FAO/WHO, 2001, <http://www.fao.org/es/ESN/food/pdf/allergygm.pdf>; 2003, <http://www.codexalimentarius.net/download/report/46/A10334ae.pdf>). This guideline includes biological and bioinformatics tests. In practice, the bioinformatics section of the guideline says that a protein is potentially allergenic if it has either an identity of at least 6 contiguous amino acids or a minimum 35% sequence similarity over a window of 80 amino acids when compared with known allergens (FAO/WHO, 2001; Taylor, 2002). The 6-amino acid identity rule has been shown to produce many false positives (Hileman *et al.*, 2002; Bjorklund *et al.*, 2005) while Stadler and Stadler (2003) claimed that the 35% over 80-residue rule might be too conservative since allergenic cross-reactivity typically requires >70% identity across the entire proteins (Aalberse, 2000). Several other computational prediction systems of allergenicity either based solely on comparison between allergen signature amino acid sequences or in conjunction with structural properties of such proteins (alternatively in combination with non-allergenic proteins),

have been reported (Gendel, 2002; Hileman *et al.*, 2002; Fiers *et al.*, 2004; Bjorklund *et al.*, 2005).

WebAllergen is a web server that predicts allergenic proteins by evaluating similarities in the underlying physicochemical properties between the query protein and the allergenicity-related protein motifs. In addition to using the 62 allergenic motifs obtained from 664 allergens (Li *et al.*, 2004), WebAllergen allows users to upload their own allergen sequences and will output a list of potential allergenic motifs. The motifs can be sent to the users via email.

METHODS

To predict allergenicity, WebAllergen needs a set of known allergens as training sequences. Potential allergenic motifs and unique sequences—those that do not share detectable motifs with others—will be generated based on the known allergens. This is a one-time process. Users are allowed to upload their own allergens. WebAllergen will process these allergens ‘on the fly’ and either email the detected motifs back to the users or use them to make prediction.

The flow of the prediction server can be found in Figure 1 and in a previous paper (Li *et al.*, 2004). The central component of the server is a program to detect conserved protein motifs. The motifs are conserved over some known allergens and hence are considered to be related to the allergenicity. The motif-detection program (Krishnan *et al.*, 2004) uses wavelet analysis to identify protein motifs that may have a sequence similarity too low to be detected by general methods.

The web server offers two main services: prediction of allergenic proteins and generation of motif profiles for allergen proteins. The two services are independent of each other and are invoked according to the input specified by the user on the web server. To predict the potential allergenicity of proteins, users can select one of three options provided on the web server. The users can either use pregenerated motif profiles obtained from a set of 664 known allergen proteins, or upload their own motif profiles; alternatively, they can also upload allergen proteins that will be used to build the motif profiles. The motif profiles generated from user-supplied allergens are available to be downloaded for future use. The prediction results list the *E*-values [reported by HMMer (Eddy, 1998)] between the query protein and the allergenicity-related motifs. The matched regions on the query proteins are also shown. The prediction server is developed in Python and is running on a Linux cluster.

ACKNOWLEDGEMENTS

We thank Prof. Chew Fook Tim of National University of Singapore for the useful discussions and Stephen Wong of Bioinformatics Institute for hardware help.

REFERENCES

Aalberse, R.C. (2000) Structural biology of allergens. *J. Allergy Clin. Immunol.*, **106**, 228–238.

*To whom correspondence should be addressed.

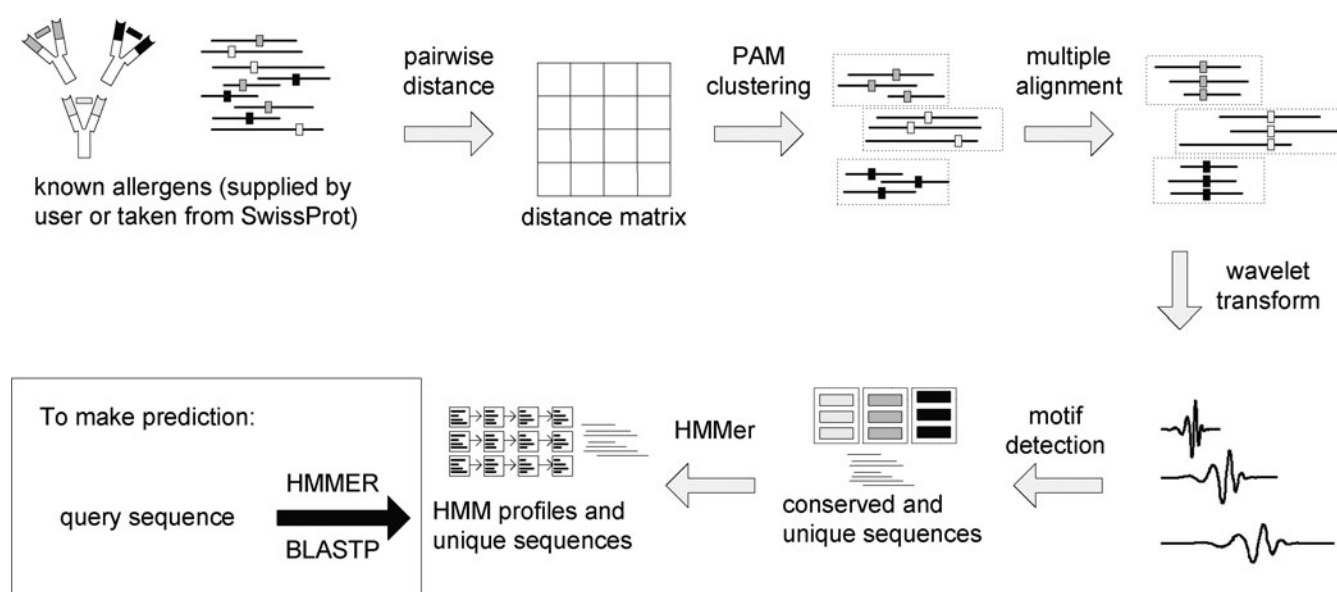


Fig. 1. The flow of the WebAllergen prediction server. Before any prediction can be done, a set of known allergen sequences must be collected. Their pairwise similarities will be computed by global-alignment. A clustering algorithm, partitioning around medoids (PAMs) in this case, will construct various clusters containing similar sequences. Each cluster will then undergo a multiple alignment stage. Sequences within each cluster go through a wavelet analysis program whereby conserved sequence motifs (with length >30 amino acids) will be detected. Those sequences that do not share any detectable motif with others are referred to as unique sequences. The prediction is made by comparing the query sequence with the HMM profiles built from the conserved motifs and the unique sequences.

- Bjorklund, A.K. *et al.* (2005) Supervised identification of allergen-representative peptides for *in silico* detection of potentially allergenic proteins. *Bioinformatics*, **21**, 39–50.
- Bredehorst, R. and David, K. (2001) What establishes a protein as an allergen? *J. Chromatogr. B*, **756**, 33–40.
- Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- FAO/WHO. (2001) Evaluation of allergenicity of genetically modified foods. *Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology*.
- FAO/WHO. (2003) *Report of the Fourth Session of The Codex ad hoc Intergovernmental Task Force on Foods Derived from Biotechnology*.
- 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.

- Gendel, S.M. (2002) Sequence analysis for assessing potential allergenicity. *Ann. N.Y. Acad. Sci.*, **964**, 87–98.
- Hileman, R.E. *et al.* (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.*, **128**, 280–291.
- Krishnan, A. *et al.* (2004) Rapid detection of conserved regions in protein sequences using wavelets. *In Silico Biol.*, **4**, 133–148.
- Lee, Y.H. and Sinko, P.J. (2000) Oral delivery of salmon calcitonin. *Adv. Drug Deliv. Rev.*, **42**, 225–238.
- Li, K.-B. *et al.* (2004) Predicting allergenic proteins using wavelet transform. *Bioinformatics*, **20**, 2572–2578.
- Stadler, M.B. and Stadler, B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J.*, **17**, 1141–1143.
- Taylor, S.L. (2002) Protein allergenicity assessment of foods produced through agricultural biotechnology. *Annu. Rev. Pharmacol. Toxicol.*, **42**, 99–112.