

BIOINFORMATICS APPLICATIONS NOTE

AllerTool: A Web Server for Predicting Allergenicity and Allergic Cross-Reactivity in Proteins

Zong Hong Zhang^{1,*}, Judice L.Y. Koh¹, Guang Lan Zhang¹, Khar Heng Choo¹, Martti T. Tammi², Joo Chuan Tong¹

¹Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

²Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

Associate Editor: John Quackenbush

ABSTRACT

Summary: Assessment of potential allergenicity and patterns of cross-reactivity is necessary whenever novel proteins are introduced into human food chain. Current bioinformatic methods in allergology focus mainly on the prediction of allergenic proteins, with no information on cross-reactivity patterns among known allergens. In this study, we present AllerTool, a web server with essential tools for the assessment of predicted as well as published cross-reactivity patterns of allergens. The analysis tools include graphical representation of allergen cross-reactivity information; a local sequence comparison tool that displays information of known cross-reactive allergens; a sequence similarity search tool for assessment of cross-reactivity in accordance to FAO/WHO *Codex alimentarius* guidelines; and a method based on support vector machine (SVM). 10-fold cross-validation results showed that the area under the receiver operating curve (A_{ROC}) of SVM models is 0.90 with 86.00% sensitivity at specificity of 86.00%.

Availability: AllerTool is freely available at <http://research.i2r.a-star.edu.sg/AllerTool/>.

Contact: zhzhong@i2r.a-star.edu.sg

INTRODUCTION

Atopic allergy and other hypersensitivity reactions are major causes of chronic ill health in affluent industrial nations, affecting up to 25% of the general population (Mekori, 1996; Nieuwenhuizen and Lopata, 2005). Allergy is caused by adverse immunological reaction to causative agents known as allergens that are otherwise innocuous in nature. The acute symptoms of allergy are usually due to the release of inflammatory mediators when an allergen cross-links immunoglobulin E (IgE) antibodies on mast cells or basophils (Sutton and Gould, 1993). This may be followed by a late-phase reaction characterized by the influx of T-cells, eosinophils and monocytes (Gould *et al.*, 2003). Atopic individuals may have one or more manifestations of the disease including asthma, conjunctivitis, dermatitis (eczema), rhinitis (hay fever) and some experience life-threatening severe anaphylaxis.

Methods for assessing potential allergenicity are essential whenever new proteins are brought into contact with humans, either through food, or other modes of exposure. The current joint recommendation by the World Health Organization (WHO) and

Food and Agriculture Organization (FAO) is a scheme based on a decision tree which compares local sequence similarity of a query protein against known allergenic proteins (FAO/WHO, 2003). Two decision criteria have been proposed for the assessment of allergenic potential: identity of six or more contiguous amino acids, or minimum 35% sequence similarity over a window of 80 amino acids. Several research groups, including Gendel *et al.* (1998, 2002), Stadler and Stadler (2003), and Fiers *et al.* (2004) developed computational tools to scan sequences that satisfy these criteria. While these tools are useful for standardized prediction of potential allergenicity of proteins according to the current recommendations of the FAO/WHO Expert Consultation, more complex techniques are needed as the six amino acid rule is non-specific and the minimum of 35% sequence similarity is too stringent to find most true allergens (Li *et al.*, 2004; Hileman *et al.*, 2002; Stadler and Stadler, 2003; Silvanovich *et al.*, 2006).

More sophisticated bioinformatic tools for detecting motifs among allergenic sequences have been recently described. Zorzet *et al.* (2002) combined FASTA3 algorithm with *k*-Nearest-Neighbour (kNN) classifier to assess potential food protein allergenicity. Soeria-Atmadja *et al.* (2004) extended the study on a larger set of allergens using a combination of kNN classifier, Bayesian linear Gaussian classifier and Bayesian quadratic Gaussian classifier. Li *et al.* (2004) demonstrated the use of wavelet transform to predict potential allergens. Björklund *et al.* (2005) introduced the use of allergen-representative peptides for detection of potentially allergenic proteins. Cui *et al.* (2006) as well as Saha and Raghava (2006) reported the use of support vector machine (SVM) for the prediction of novel allergen proteins.

In this paper, we present AllerTool, a web server providing essential tools for assessing predicted as well as published allergic cross-reactivity patterns of clinically relevant protein allergens. Three different programs are available for assessing the potential allergenicity of protein sequences – (i) a sequence similarity search tool for assessment of allergenicity in accordance to FAO/WHO *Codex alimentarius* guidelines; (ii) a SVM-based method for prediction of protein cross-reactivity with little or no similarity to known allergens; and (iii) a modification of BLAST that displays cross-reactive allergens. In addition, AllerTool also provides potential cross-reactivity information of a query sequence through a

*To whom correspondence should be addressed.

DB/AllerTool/Algorithms.html). Using amino acid composition as input for training and testing ALR-SVM, the system can predict allergenic proteins with sensitivity (SE) of 86.00% and specificity (SP) of 86.00% respectively. These values are comparable to the SVM approach by Saha and Raghava (2006) (SE=85.02%, SP=84.00%) and allergen representative peptides (SE=81.00%, SP=90.00%; Bjorklund *et al.*, 2005), and outperform motif-based approach using MEME/MAST software (SE=93.94%, SP=33.34%; Saha and Raghava, 2006). For the predicted allergenic sequences, a list of high similarity IUIS allergen sequences and reported cross-reactivity information is provided.

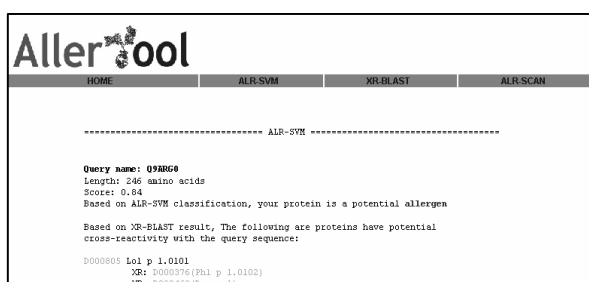


Fig 4. An example of ALR-SVM output.

CONCLUSION

With the advent of genetically modified proteins in foods, therapeutics and biopharmaceuticals (Saha and Raghava, 2006), AllerTool provides a new service for the assessment of predicted as well as published cross-reactivity patterns of novel proteins. ALR-SCAN is useful for assessment of allergenicity in proteins according to the FAO/WHO Codex alimentarius guidelines. However, concerns have been raised about the validity of the current FAO/WHO guidelines (Li *et al.*, 2004; Hileman *et al.*, 2002). Various groups including Silvanovich *et al.* (2006) and Stadler and Stadler (2003) reported that short sequence searches of six contiguous amino acids to identify allergenic proteins is a product of chance and adds little value to allergy assessments for newly expressed proteins. There is a need for more sophisticated techniques for screening of allergenicity in proteins. ALR-SVM has been developed to capture non-linear characteristics that may be encapsulated within allergenic protein sequences. Future work will focus on the development of other supplementary methods to support and refine the prediction of cross-reactivity patterns.

ACKNOWLEDGEMENTS

Authors are thankful to Prof. Vladimir Brusic (UQ, Australia) for critically reading the manuscript.

REFERENCES

- Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
- Bjorklund, A.K., *et al.* (2005) Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics*, **21**, 39-50.
- Cui, J., *et al.* (2006) Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol. Immunol.*, In press.
- FAO/WHO (2003) Codex Principles and Guidelines on Foods Derived from Biotechnology. Rome, Italy, Joint FAO/WHO Food Standards Programme.
- Fiers, M.W., *et al.* (2004) Allermatch, a webtool for the prediction of potential Allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics*, **5**, 133-138.
- Gendel, S.M. (1998) The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv. Food Nutr. Res.*, **42**, 45-62.
- Gendel, S.M. (2002) Sequence analysis for assessing potential allergenicity. *Ann. N. Y. Acad. Sci.*, USA, **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.
- Koh, J.L.Y., *et al.* (2004) BioWare: A framework for bioinformatics data retrieval, annotation and publishing, in *ACM SIGIR Workshop on Search and Discovery in Bioinformatics (SIGIRBIO)*, Sheffield, UK, July 2004.
- Li, G.B., *et al.* (2004) Predicting allergenic proteins using wavelet transform. *Bioinformatics*, **20**, 2572-2578.
- Mekori, Y.A. (1996) Introduction to allergic diseases. *Crit. Rev. Food Sci. Nutr.*, **36**, S1-18.
- Miotto, O., *et al.* (2005) Supporting the curation of biological databases with reusable text mining. *Genome Inform.*, **16**, 32-44.
- Nieuwenhuizen, N.E. and Lopata, A.L. (2005) Fighting food allergy. Current approaches. *Ann. N.Y. Acad. Sci.*, **1056**, 30-45.
- Saha, S. and Raghava, G.P.S. (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.*, **34**, W202-W209.
- Silvanovich, A., *et al.* (2006) The Value of Short Amino Acid Sequence Matches for Prediction of Protein Allergenicity. *Toxicol Sci.*, **90**, 252-258.
- Soeria-Atmadja, D., *et al.* (2004) Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int. Arch. Allergy Immunol.*, **133**, 101-112.
- Stadler, M.B. and Stadler, B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J*, **17**, 1141-1143.
- Stothard, P. (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*, **28**, 1102-1104.
- Sutton, B.J. and Gould, H.J. (1993) The human IgE network. *Nature*, **366**, 421-428.
- Zorzet, A., *et al.* (2002) Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol.*, **2**, 525-534.