Sequence analysis

AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins

Zong Hong Zhang*, Judice L. Y. Koh, Guang Lan Zhang, Khar Heng Choo, Martti T. Tammi and Joo Chuan Tong

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 and 1Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

Received on August 21, 2006; revised on November 23, 2006; accepted on December 5, 2006
Advance Access publication December 6, 2006
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). A 10-fold cross-validation results showed that the area under the receiver operating curve (AUC) of SVM models is 0.90 with 86.00% sensitivity (SE) at specificity (SP) of 86.00%.

Availability: AllerTool is freely available at http://research.i2r.a-star.edu.sg/AllerTool/
Contact: zhzhang@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-reacts with the human immune system. The main purpose of AllerTool is the support of potential cross-reactivity information of a query sequence through a graphical representation of the cross-reactivity network of known allergens and (3) a modification of BLAST that displays similar proteins. The main purpose of AllerTool is the support of 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 (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).


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—(1) a sequence similarity search tool for assessment of allergenicity in accordance to FAO/WHO Codex alimentarius guidelines; (2) a SVM-based method for prediction of protein cross-reactivity with little or no similarity to known allergens and (3) a modification of BLAST that displays cross-reactive allergens. In addition, AllerTool also provides potential cross-reactivity information of a query sequence through a graphical representation of the cross-reactivity network of the similar proteins. The main purpose of AllerTool is the support of...
molecular studies of allergens, the assessment of allergic responses and of allergic cross-reactivity.

**SYSTEM DESCRIPTION**

**Data**

Allergen data were extracted from the International Union of Immunological Societies (IUIS) Allergens website (http://www.ALLERGEN.org) and stored in the ALLERDB database (Zhang et al., manuscript in preparation; http://antigen.i2r.a-star.edu.sg/Templar/DB/Allergen/). The dataset consists of all IUIS allergens and isoallergens that have protein sequences available in the public sequence databases or publication references. The dataset consists of 373 allergens, 260 isoallergens and 128 instances of reported cross-reactivity collected from the literature and verified using a text-mining tool ABK (Miotto et al., 2005).

**Analysis tools**

AllerTool and web interface are written in C/C++ and Perl and run on a SunOS 5.9 UNIX system with Apache web server. It comprises of four integrated tools for assessing the potential allergenicity of protein sequences—XR-BLAST, XR-Graph, ALR-SCAN and ALR-SVM.

XR-BLAST (Koh et al., 2004) is a local sequence comparison tool based on BLAST2.2.3 (Altschul et al., 1997) that outputs information on allergens that have reported cross-reactivity with the individual matches. A sample output of XR-BLAST is given in Figure 1.

XR-Graph (http://antigen.i2r.a-star.edu.sg/Templar/DB/Allergen/) is a visualization tool for graphical representation of allergen cross-reactivity information. Each graph displays allergens (boxes) that are related by reported cross-reactivity (links). This visual tool enables user to establish possible allergen cross-reactivity relationships not reported before. This tool has potential uses in the development of novel allergy diagnostics approaches. A sample output of XR-Graph is shown in Figure 2.

ALR-SCAN (Koh et al., 2004) is a sequence similarity search tool that reports sequence similarity in accordance to the current FAO/WHO recommendation for the assessment of allergenicity. Both the six contiguous amino acids identity rule, and >35% identity over a stretch of 80 amino acids are implemented. Users can submit the protein of interest to ALR-SCAN, which will return the list of matches that satisfy either of the rules. Sample query and output is shown in Figure 3.

ALR-SVM is a useful tool for predicting protein allergenicity based on global description of amino acid sequence using SVM as the prediction engine (Cui et al., 2006; Fig. 4). The training dataset consists of 460 allergens and 560 non-allergens, while the testing dataset includes 114 allergens and 140 non-allergens derived from

---

Fig. 1. An example of XR-BLAST output.

Fig. 2. An example of XR-Graph output for Aln g 1. Reported cross-reactivity patterns are represented by links. Possible cross-reactivity relationships may be inferred by missing links.

Fig. 3. An example of ALR-SCAN output.

Fig. 4. An example of ALR-SVM output.
http://www.slv.se/templates/SLV_Page.aspx?id=9343 (Björklund et al., 2005) selected using a debiasing strategy based on sequence similarity of protein sequences commonly found in consumed food with no records in existing allergen databases (Saha and Raghuva, 2006). The percentage of allergens represents ~45% of the testing dataset, while non-allergens represent the remaining 55%. Different kernel functions (linear, polynomial, radial and sigmoid) were explored to improve the prediction accuracy of the SVM models. ALR-SVM is based on the third degree polynomial kernel function encoded using descriptors derived from amino acid composition. The $A_{ROC}$ value is 0.90 (http://antigen.i2r.a-star.edu.sg/Templar/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 Raghuva (2006) (SE = 85.02%, SP = 84.00%) and allergen-representative peptides (SE = 81.00%, SP = 90.00%; Björklund et al., 2005), and outperform motif-based approach using MEME/MAST software (SE = 93.94%, SP = 33.34%; Saha and Raghuva, 2006). For the predicted allergenic sequences, a list of high-similarity IUIS allergen sequences and reported cross-reactivity information is provided.

CONCLUSION

With the advent of genetically modified proteins in foods, therapeutics and biopharmaceuticals (Saha and Raghuva, 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.

Conflict of Interest: none declared.

REFERENCES


