BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 16 2012, pages 2178–2179
doi:10.1093/bioinformatics/bts326

Sequence analysis

Advance Access publication June 12, 2012

SORTALLER: predicting allergens using substantially optimized algorithm on allergen family featured peptides

Lida Zhang[1,2,†], Yuyi Huang[2,†], Zehong Zou[2], Ying He[2], Ximo Chen[2] and Ailin Tao[2,*]

[1]Plant Biotechnology Research Center, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200030 and [2]Guangzhou Municipal Key Laboratory of Allergy & Clinical Immunology, Allergy Research Branch of the State Key Laboratory of Respiratory Disease, the Second Affiliated Hospital of Guangzhou Medical University, Guangzhou 510260, China

Associate Editor: Martin Bishop

ABSTRACT

Summary: SORTALLER is an online allergen classifier based on allergen family featured peptide (AFFP) dataset and normalized BLAST E-values, which establish the featured vectors for support vector machine (SVM). AFFPs are allergen-specific peptides panned from irredundant allergens and harbor perfect information with noise fragments eliminated because of their similarity to non-allergens. SORTALLER performed significantly better than other existing software and reached a perfect balance with high specificity (98.4%) and sensitivity (98.6%) for discriminating allergenic proteins from several independent datasets of protein sequences of diverse sources, also highlighting with the Matthews correlation coefficient (MCC) as high as 0.970, fast running speed and rapidly predicting a batch of amino acid sequences with a single click.

Availability and implementation: http://sortaller.gzhmc.edu.cn/.

Contact: taoailin@gzhmc.edu.cn

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 19, 2012; revised on May 30, 2012; accepted on May 31, 2012

1 INTRODUCTION

Allergy symptoms affect more than 25% of the population in industrialized countries (Ball *et al*., 2009). The allergy mechanism is an immunological reaction against an allergen that is often IgE mediated, but non-IgE mediated (i.e. cell mediated) or mixed IgE and non-IgE mediated reaction may be more important in food allergy (Boyce *et al.*, 2010). Biotechnology is rapidly developing, which involves the introduction of novel proteins into drugs or foods, and those proteins can be allergenic (Nestle, 1996). It is essential to avoid transferring a gene that encodes a major allergenic protein (from any source) into a drug/food crop that did not previously contain that protein. To accurately discriminate candidate genes from allergens before transferring those into a drug/food organism would aid preventive efforts to curb the rising incidence of allergies.

Sequence comparison of novel proteins for similarity to known allergens is a critical part of the weight of evidence approach used to ascertain the safety of the gene(s) to be transformed. Over the past 5 years, several sophisticated bioinformatic methods have recorded substantial achievements in allergen prediction (Barrio *et al.*, 2007; Fiers *et al.*, 2004; Muh *et al.*, 2009; Saha and Raghava, 2006; Zhang *et al.*, 2007). Although these tools are effective for predicting allergenic proteins from sequence segments that show homology with known allergens, they are less accurate for novel proteins with low similarity to any known allergens, especially for the undeveloped exogenous genes with desirable traits.

In this study, we developed an allergen prediction method, SORTALLER, which predicts allergens by using a novel algorithm on allergen family featured peptides (AFFPs), which were substantially optimized on most of the SVM-based (Webb-Robertson *et al.*, 2010) classifier parameters. SORTALLER outperformed other methods and achieved a perfect balance between high sensitivity and high specificity for discriminating allergenic proteins from diverse sources. This method could be used as the first step of programmed allergenicity assessment, and a web server of SORTALLER has been developed to allow for simultaneously and rapidly predicting a batch of amino acid sequences.

2 METHODS AND RESULTS

2.1 Datasets

A total of 2359 allergenic protein sequences were firstly obtained from the Allergome database (Mari *et al.*, 2006), the Swiss-Prot Allergen Index (http://www.uniprot.org/docs/allergen.txt), the Food Allergy Research and Resource Program allergen protein database (http://www.allergenonline.org/) and the Allergen Nomenclature database of the International Union of Immunological Societies (http://www.allergen.org/). After filtered by negative results of IgE binding, 2290 allergenic protein sequences were retained.

The non-allergenic protein sequences were excerpted from commonly consumed commodities and human by searching in the UniProt/Swiss-Prot protein database (Bairoch *et al.*, 2005). 234 760 non-allergenic protein sequences were retained by a filtering process (detailed in Supplementary material).

2.2 Allergen classifier

After analyzing the AFFP rationale, we have panned 444–556 AFFPs from 211 known families and ungrouped allergenic proteins by adopting a specific screening procedure with different sliding window sizes. Of all allergen families, more than half of the families contained only one AFFP, and most AFFPs were shorter than 200 aa (detailed in Supplementary material).

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

*To whom correspondence should be addressed.

Table 1. Comparison of different prediction methods

Methods	SE (%)	SP (%)	ACC (%)	MCC
FAO/WHO[a]	99.2	9.6	54.4	0.198
EVALLER[b]	86.6	99.0	92.8	0.863
AlgPred (amino acid)[c]	92.4	80.2	86.3	0.731
AlgPred (dipeptide)[c]	88.8	88.2	88.5	0.770
AlgPred (ARPs BLAST)[c]	81.8	98.0	89.9	0.809
AllerHunter[d]	82.2	99.2	90.7	0.826
SORTALLER[e]	**98.6**	**98.4**	**98.5**	**0.970**

SE, sensitivity; SP, specificity; ACC, accuracy; MCC, Matthews correlation coefficient.
[a]Allergen sorting was conducted on a Perl script referenced in the FAO/WHO guidelines: an identity of at least six contiguous amino acids or a minimum of 35% sequence identity over a window of 80 amino acids with an allergen would deduce cross-reactivity.
[b]http://www.slv.se/en-gb/Group1/Food-Safety/e-Testing-of-protein-allergenicity/e-Test-allergenicity/.
[c]http://www.imtech.res.in/raghava/algpred/.
[d]http://tiger.dbs.nus.edu.sg/AllerHunter/.
[e]http://sortaller.gzhmc.edu.cn/.

A SVM-based classifier was constructed and optimized in concert on two major parameters, i.e. sliding windows of different peptide lengths for AFFP identification and the constant ***C*** value in sigmoid function for raw BLAST E-value scaling. It was shown that 20 amino acids was the optimal peptide length for AFFP screening and the ***C*** value of 4 in the sigmoid function contributed to the optimum effects on allergen prediction. Matthews correlation coefficient (MCC; Mizianty *et al.*, 2010) value as high as 0.969 manifested the best prediction result acquired under this circumstance (see Supplementary Table S2 and Fig. S2).

2.3 Benchmarking performance in comparison with existing methods

The independent dataset of 1000 sequences (including 500 allergens and 500 non-allergenic proteins) and 14 arduous proteins (including at least four allergens referenced by IgE experiments) were used to compare SORTALLER with the following methods that can be accessible: (i) the FAO/WHO evaluation scheme, which is based on the identity of six or more contiguous amino acids or a minimum 35% sequence identity over a window of 80 amino acids; (ii) EVALLER (Barrio *et al.*, 2007); (iii) three different prediction methods of AlgPred (Saha and Raghava, 2006); (iv) AllerHunter (Muh *et al.*, 2009); (v) APPEL (Cui *et al.*, 2007) and (vi) Allermatch (Fiers *et al.*, 2004). As shown in Table 1 and Supplementary Table S3, the SORTALLER software significantly outperformed the other methods and achieved a perfect balance between substantially higher specificity and sensitivity in discriminating allergens from diverse sources.

The following aspects may contribute to the superiority of SORTALLER, such as MCC used as the performance measuring end point, which orchestrated two parameters, specificity and sensitivity, to express an unbiased accuracy indicator; the appropriate amount of AFFPs bearing enough information of allergens and synergetically optimized sigmoid function; BLAST used for fast and reliable detection of the similarities of query sequence with AFFPs (detailed in Supplementary material).

3 CONCLUSION

SORTALLER is unique since it is founded on preferential panning of specific AFFPs harboring perfect information for allergens sorting based on a thoroughly optimized procedure. The novel classifier outperformed other methods and exhibited highly balanced sensitivity and specificity and higher accuracy. Moreover, the AFFP dataset incorporated in SORTALLER would be useful for regimens of component-resolved diagnosis and peptide immunotherapy.

ACKNOWLEDGEMENTS

We thank 20 members from Guangzhou Municipal Key Laboratory of Allergy and Clinical Immunology for their assistance on allergen sequence collecting and software testing.

Funding: This research was supported by the Great Project (2011ZX08011-005) and Key Project (2009ZX08011-004B) from the Major Program of National Science and Technology of China; National Natural Science Foundation of China (30771240); and Key Program from Guangdong Provincial Natural Science Foundation (8251018201000002).

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
Ball,T. *et al.* (2009) Reducing allergenicity by altering allergen fold: a mosaic protein of Phl p 1 for allergy vaccination. *Allergy*, **64**, 569–580.
Barrio,A.M. *et al.* (2007) EVALLER: a web server for *in silico* assessment of potential protein allergenicity. *Nucleic Acids Res.*, **35**, W694–W700.
Boyce,J.A. *et al.* (2010) Guidelines for the diagnosis and management of food allergy in the United States: report of the NIAID-sponsored expert panel. *J. Allergy Clin. Immunol.*, **126**, S1–S58.
Cui,J. *et al.* (2007) Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol. Immunol.*, **44**, 514–520.
Fiers,M. *et al.* (2004) Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics*, **5**, 133.
Mari, A. *et al.* (2006) Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. *Cell Immunol.*, **244**, 97–100.
Mizianty,M.J. *et al.* (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.
Muh,H. *et al.* (2009) AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PLoS One*, **4**, e5861.
Nestle,M. (1996) Allergies to transgenic foods—questions of policy. *N. Engl. J. Med.*, **334**, 726–728.
Saha,S. and Raghava,G.P.S. (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.*, **34**, W202–W209.
Webb-Robertson,B.-J.M. *et al.* (2010) A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics*, **26**, 1677–1683.
Zhang,Z.H. *et al.* (2007) AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics*, **23**, 504–506.