

Improving Allergen Prediction in Main Crops Using a Weighted Integrative Method

Jing Li¹ · Jing Wang² · Jing Li¹

Received: 29 July 2016 / Revised: 29 September 2016 / Accepted: 3 October 2016

© International Association of Scientists in the Interdisciplinary Areas and Springer-Verlag Berlin Heidelberg 2016

Abstract As a public health problem, food allergy is frequently caused by food allergy proteins, which trigger a type-I hypersensitivity reaction in the immune system of atopic individuals. The food allergens in our daily lives are mainly from crops including rice, wheat, soybean and maize. However, allergens in these main crops are far from fully uncovered. Although some bioinformatics tools or methods predicting the potential allergenicity of proteins have been proposed, each method has their limitation. In this paper, we built a novel algorithm PREAL^W, which integrated PREAL, FAO/WHO criteria and motif-based method by a weighted average score, to benefit the advantages of different methods. Our results illustrated PREAL^W has better performance significantly in the crops' allergen prediction. This integrative allergen prediction algorithm could be useful for critical food safety matters. The PREAL^W could be accessed at <http://lilab.life.sjtu.edu.cn:8080/prealw>.

Keywords Allergen · Crops · Integrative prediction · Genome-wide screen

1 Introduction

The growing evidence shows the number of people with the food allergy is increasing throughout the world over the past few decades [1]. Food allergic reaction is an abnormal immune response caused by naturally-occurring proteins in foods, which affect nearly 5 % of adults and 8 % of children [2–4]. The food allergens in our daily lives are mainly from crops. For example, hundreds of proteins in wheat, rice, soybean and other cereal crops were identified as allergens that could cause allergic responses in susceptible individuals [5–8]. However, the genome-wide investigation has not been performed in these cereal crops until now, even though they are cultured worldwide and provide food for almost half of the global population.

Currently, there are three types of bioinformatic method for allergen prediction. The first approach follows the FAO/WHO guidelines and compares the sequence similarity with the known allergens [9]. In the FAO/WHO rules, a potential allergenic protein is defined with an exact match of at least six consecutive amino acids (rule 1) or the identity within any 80 consecutive amino acids is over 35 % in comparison with any known allergen (rule 2). Because of the risk of a high positive rate of sequence-based approach, the second approach was proposed based on identifying conserved allergenicity-related linear motif later. The motif-based method focused on protein's secondary structure (motif) instead of primary structure. Twenty-seven allergen motifs were reported by comparing allergens to non-allergens using MEME/MAST [10]. The motif-based approach controls the false positive following the loss of sensitivity. The third approach is a statistical learning method. Saha et al. [11] developed methods based on support vector machine (SVM) using amino acid composition and achieved an accuracy of about 85 %. An

✉ Jing Li
jing.li@sjtu.edu.cn

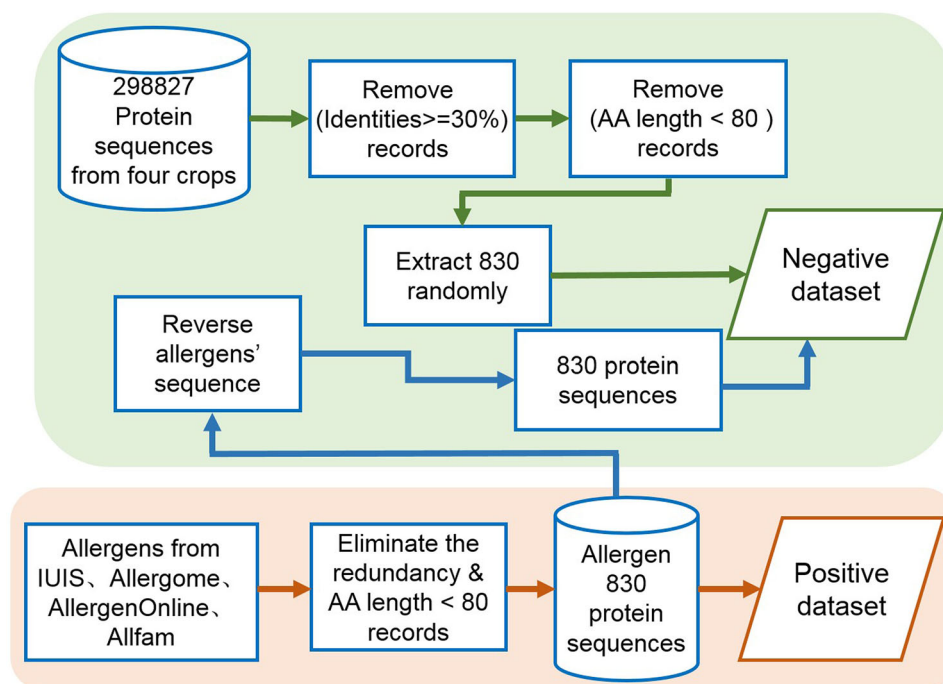
Jing Li
lijing910409@sjtu.edu.cn

Jing Wang
jing.wang@vanderbilt.edu

¹ Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, People's Republic of China

² Center for Quantitative Sciences, Vanderbilt University School of Medicine, Nashville, TN 37203, USA

Fig. 1 Flow diagram of dataset collection. The setup processes of the positive dataset were displayed in the *orange area* and that of the negative dataset were in *green*. And a reversed negative dataset was built as the flow in *blue line* for evaluating FAO/WHO rule 2 specifically



SVM-based predictive model using variable-length peptide extraction of allergens was also proposed [12]. In our previous study, we presented a comprehensive model in 128 features space, named PREAL, integrating various properties of proteins for the prediction of allergenicity.

In this study, we collected up-to-date known allergens in rice, wheat, maize and soybean and built a novel weighted average score prediction algorithm PREAL^W, which integrated PREAL, FAO/WHO criteria and motif-based method by a weighted average score, so as to benefit the advantages of different methods. Then, we compared its performance with three existing methods. This work would be of help in uncovering the molecular properties of food allergens and developing powerful approaches to control or reduce the allergenic risk.

2 Materials and Methods

2.1 Training Data Set

The known allergens (allergenic molecules) which identified causing an IgE-mediated disease with international scientific journals support were collected from WHO/IUIS Allergen Nomenclature (<http://www.allergen.org/>), Allergome (<http://www.allergome.org/>), FARRP AllergenOnline (<http://www.allergenonline.org/>) and Allfam (<http://www.meduniwien.ac.at/allergens/allfam/>). In total, 830 known allergens in rice, maize, soybean and wheat were collected from different data sources. The redundant proteins were removed as well as the proteins that were less than 80 amino acids.

In order to build the negative dataset, we downloaded 298,827 protein sequences of rice, maize, soybean and wheat from NCBI-GenPept and RefSeq (<http://www.ncbi.nlm.nih.gov/>), since few of wheat proteins are included in RefSeq. Then, we removed the proteins whose identities are more than 30 % with any known allergen and the proteins whose sequence length is less than 80. Finally, 830 records were randomly sampled from the remaining sequences to build the negative dataset.

Since we removed the proteins with more than 30 % sequence identities with any known allergen, the same negative dataset cannot be used to assess the performance of FAO/WHO rule 2 to control the risk of overfitting. Therefore, same as the way adopted by Stadler et al. [13], we used the reversed sequences of all allergens as the negative dataset for evaluating the performance of FAO/WHO rule 2. The workflow of the construction of training datasets is summarized in Fig. 1.

2.2 Software for Development

The sequence similarity was measured using NCBI-BLAST (version 2.2.23) (<ftp.ncbi.nlm.nih.gov/blast/>) [14]. To distinguish a protein is more likely to be allergen or non-allergen, we adopt the SVM method through LIBSVM software (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>, v3.0). MEME/MAST 4.9.1 (<http://alternate.meme-suite.org/>) was acquired for scanning motif and 27 allergy-related motifs obtained from previous research [10, 15].

The Web-based server of PREAL^W was developed under the environment of LAMP. The detailed versions of

these softwares are: Linux (CentOS_5.7); Apache 2.2.3; MySQL 5.0.77; and PHP_5.1.6.

2.3 Integrative Algorithm

The weighted average of the normalized scores (WAS) initially was used to evaluate the single nucleotide variant (SNVs) nonsynonymous substitution whether it was benign or malignant [16]. Here WAS algorithm was used in PREAL^W to integrate the FAO/WHO rules, the motif-based approach and the SVM-based method for allergen prediction [9, 10, 17]. The weights which are according to the reference cumulative distributions of true positives and true negatives reflected the probability that allergens with higher scores were not false positives [16]. More details of these three methods can be found in references [10, 17]. Through the comparative analysis, we already found each method has its advantage [10, 13, 17]. To take advantage of the benefit of each method, we calculated the integrated score WAS of a candidate protein using the scores given by the three allergen prediction methods as below:

$$\text{WAS} = \frac{\sum_i S_i * W_i}{\sum_i W_i}$$

where $C_i = 1$, $W_i = 1 - P_{ni}$

$C_i = 0$, $W_i = 1 - P_{pi}$

S_i is the normalized score as calculated by the i^{th} tool; $C_i = 1$ if the candidate protein is predicted as allergen in the i^{th} tool, otherwise $C_i = 0$; W_i is the weight for the given classification. Note that the weight of individual method W_i is calculated on the basis of the probabilities of true negative P_{ni} or true positive P_{pi} , respectively, with a score greater than S_i . The score given by FAO/WHO criteria and motif-based method is the Boolean value. In this case, the term P_{ni} corresponds to the probability of true negative in the data set when S_i is a value of true, while P_{pi} is the probability of true positive when the value is false. The cutoff of PREAL^W was chosen by using an R package “Daim,” which selects the score at the peak of the sensitivity–accuracy curve in the tenfold cross-validation [18].

2.4 Cross-Validation

The performances of the allergen prediction methods presented in our study were assessed using tenfold cross-validation. We divided the dataset into ten groups randomly, where each group contains 83 of allergens and non-allergens, respectively. One group would be retained as testing data, while the remaining nine groups were used as training data for training. To reduce variability, ten rounds of cross-validation are performed. And then, the validation results are averaged over the rounds.

2.5 Performance Evaluation

We used the following statistics measures to evaluate the performance of the methods of allergen prediction in our study: Sensitivity measures the proportion of allergens that are correctly predicted.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity measures the proportion of non-allergens (negative) that are correctly identified.

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

Accuracy is the percentage of correctly predicted both allergen and non-allergen proteins.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

In the aforementioned formulas, TP and FN refer to true positives and false negatives, while TN and FP refer to true negatives and false positives.

3 Results

3.1 Performance of PREAL^W Method

Our PREAL^W method allows the evaluation of the potential allergenicity of protein by integrating and combining the advantages of three kinds of widespread allergen prediction methods. Using the method described by Wang et al. [10], we constructed the positive and negative datasets including 830 entries in each one, for the crops' allergen prediction using the protein sequences in rice, soybean, maize and wheat proteome. The tenfold cross-validation showed that PREAL^W receives the AUC value of 0.87 and the accuracy of 85.9 %, which outperforms either the FAO/WHO criteria or PREAL (Table 1), and Venkatraman's test shows our new method has a significant improvement (p value $< 2.2\text{e}-4$).

3.2 Genome-Wide Allergen Prediction in the Crops

Using PREAL^W method, we scanned the whole proteome sequences and predicted the potential allergens in soybean, rice, wheat and maize. In the results, 299, 927, 984 and 932 putative allergens were identified in soybean, rice, wheat and maize confidently. The statistical summary is displayed in Table 2.

3.3 Online Server

A Web site named PREAL^W (<http://lilab.life.sjtu.edu.cn:8080/mocha/tool.html>) has been developed that allows users to predict the potential allergens online using our new algorithm. PREAL^W not only provides the query of single protein, also supports batch

Table 1 Performance comparison of allergen prediction methods

	FAO/WHO rule 1 (%)	FAO/WHO rule 2 (%)	Motif-based (%)	PREAL (%)	PREAL ^w (%)
Sensitivity	74.6	77.8	79.8	70.1	77.2
Specificity	90.3	89.8	60.5	91.1	94.6
Accuracy	82.5	83.9	70.1	80.6	85.9

Table 2 Statistics of allergens in major crops

	Verified	Predicted	Total
Soybean	92	299	391
Rice	151	927	1078
Wheat	482	984	1466
Maize	121	932	1053
Total	846	3142	3988

prediction. A screenshot of the prediction result page of PREAL^w is shown in Fig. 2. Moreover, the data download of known and putative allergens of four crops is also available.

4 Discussion

Although many allergen-related databases and bioinformatics tools are online, which are dedicated to instore or predict allergens, only few allergens in rice, wheat, maize

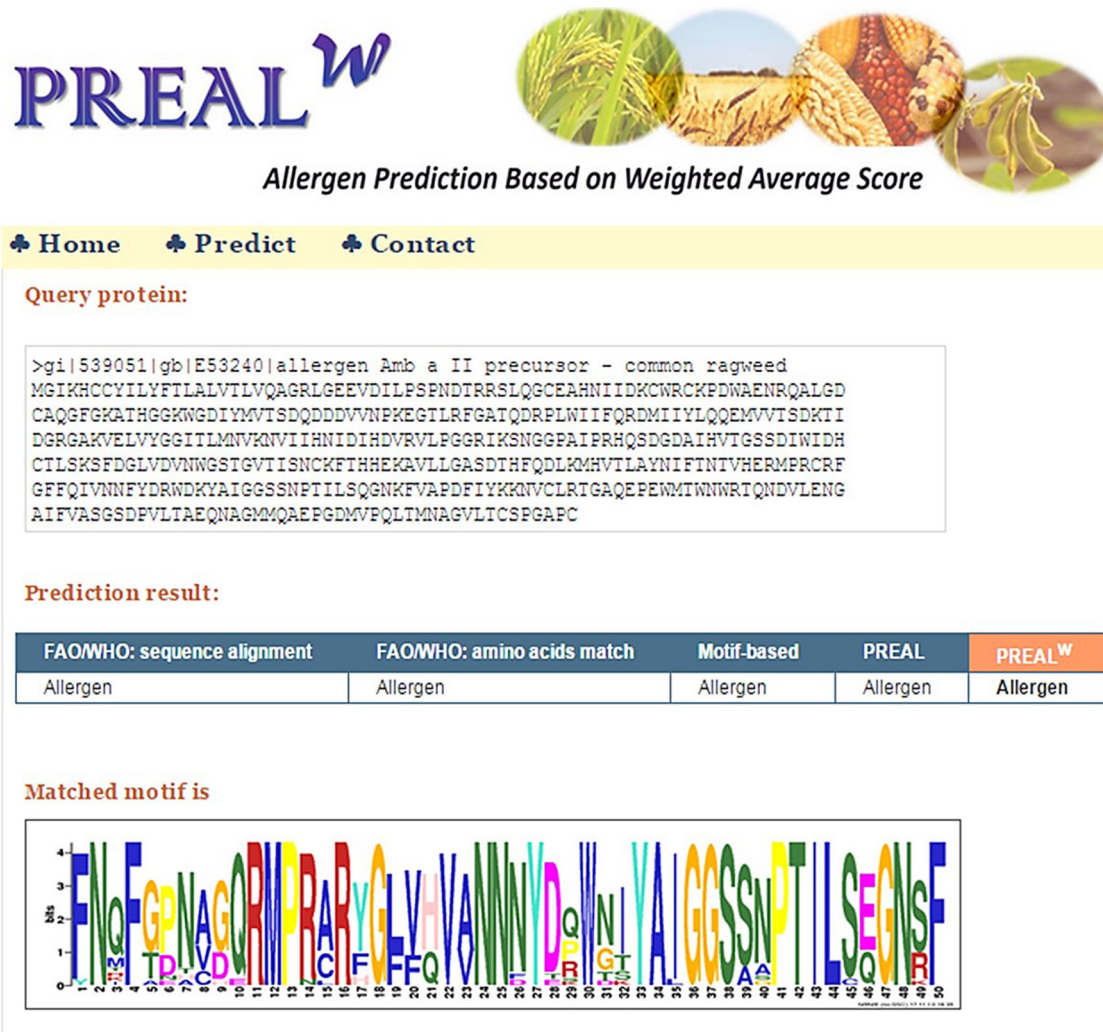


Fig. 2 Snapshot of the prediction result page of PREAL^w. It showed the prediction results of FAO/WHO criteria, motif-based method, PREAL and our new method PREAL^w in a table format with matched motif picture

and soybean were uncovered so far; even these crops provide food for almost half the global population. Some crops were genetically modified (GM) to improve productivity or increase nutrients, and the allergen is one of the major concerns about these GM crops.

In this study, a weighted integrative algorithm PREAL^W was proposed and used in allergen prediction here. As we reported previously, the FAO/WHO criteria show high sensitivity followed by high risk of false positive; the motif-based method can identify the conserved motif, and the machine learning method PREAL obtains good accuracy, which depends on the training dataset of enough size. The new algorithm PREAL^W is designed to integrate the PREAL, FAO/WHO criteria and the motif-based method by the weighted average score and take the advantages of different methods. The results have illustrated that the algorithm PREAL^W has better specificity and overall measurement accuracy. In accordance with the previous study, the rule 2 of FAO/WHO criteria and motif-based method have a higher sensitivity, but the high risk of false positive should be concerned [19]. In addition, even using the rules of FAO/WHO criteria with high sensitivity, some known allergen still cannot be identified. For example, a protein named “Alpha Purothionin” (UniProtKB—Q9T0P1) had been identified as an allergen protein in wheat’s seed. Our integrated method PREAL^W, PREAL and motif-based method do predict it as an allergen, while the FAO/WHO criteria do not. That suggests that only considering the sequence similarity is insufficient in allergen prediction and the combination of multiple methods would be more powerful.

However, our algorithm still has some limits that need to overcome. The existing allergen predictors mainly focused on the high similarity of sequence, so they are less effective when facing a sequence with low overall similarity. Some new methods were proposed in the latest protein classification works, which might be adopted in allergen prediction in future [20–22].

Acknowledgments This work was supported by the National Transgenic Plant Special Fund (2016ZX08011006-004) and the Major Program of the National Natural Science Foundation of China (Grant No. 31230058).

References

- Allergen Bureau Labelling Review Survey (2009) Food Aust 62(2010) 344–345
- Caffarelli C (2001) A revised nomenclature for allergy: an EAACI position statement from the EAACI nomenclature task force (vol 56, pg 813, 2001). *Allergy* 56:1229
- Johansson SGO, Bieber T, Dahl R, Friedmann PS, Lanier BQ, Lockey RF, Motala C, Martell JAO, Platts-Mills TAE, Ring J, Thien F, Van Cauwenberge P, Williams HC (2003) Revised nomenclature for allergy for global use: report of the nomenclature review committee of the world allergy organization. *J Allergy Clin Immunol* 113(2004):832–836
- Sicherer SH, Sampson HA (2014) Food allergy: epidemiology, pathogenesis, diagnosis, and treatment. *J Allergy Clin Immunol* 133:291–307
- Armentia A, Sanchez-Monge R, Gomez L, Barber D, Salcedo G (1993) In vivo allergenic activities of eleven purified members of a major allergen family from wheat and barley flour. *Clin Exp Allergy* 23:410–415
- Izumi H, Adachi T, Fujii N, Matsuda T, Nakamura R, Tanaka K, Urisu A, Kurosawa Y (1992) Nucleotide sequence of a cDNA clone encoding a major allergenic protein in rice seeds Homology of the deduced amino acid sequence with members of alpha-amylase/trypsin inhibitor family. *FEBS Lett* 302:213–216
- Gu XL, Beardslee T, Zeece M, Sarath G, Markwell J (2001) Identification of IgE-binding proteins in soy lecithin. *Int Arch Allergy Immunol* 126:218–225
- Petersen A, Dresselhaus T, Grobe K, Becker WM (2006) Proteome analysis of maize pollen for allergy-relevant components. *Proteomics* 6:6317–6325
- FWCA Commission (2003) Codex principles and guidelines on foods derived from biotechnology, joint FAO/WHO Food Standards Programme
- Wang J, Yu Y, Zhao Y, Zhang D, Li J (2013) Evaluation and integration of existing methods for computational prediction of allergens. *BMC Bioinformatics* 14(Suppl 4):S1
- Saha S, Raghava GP (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* 34:W202–W209
- Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U (2006) Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning. *Nucleic Acids Res* 34:3779–3793
- Stadler MB, Stadler BM (2003) Allergenicity prediction by protein sequence. *FASEB J* 17:1141–1143
- Johnson M, Zaretskaya I, Raytselis Y, Merezukh Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res* 36:W5–W9
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren JY, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37:W202–W208
- Gonzalez-Perez A, Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 88:440–449
- Wang J, Zhang D, Li J (2013) PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection. *BMC Syst Biol* 7(Suppl 5):S9
- Adler W, Lausen B (2009) Bootstrap estimated true and false positive rates and ROC curve. *Comput Stat Data Anal* 53:718–729
- Stadler MB, Stadler BM (2003) Allergenicity prediction by protein sequence. *FASEB J* 17:1141–1143
- Song L, Li D, Zeng X, Wu Y, Guo L, Zou Q (2014) nDNA-Prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformatics* 15:298
- Wei L, Liao M, Gao X, Zou Q (2015) Enhanced protein fold prediction method through a novel feature extraction technique. *IEEE Trans Nanobioscience* 14:649–659
- Lin C, Zou Y, Qin J, Liu X, Jiang Y, Ke C, Zou Q (2013) Hierarchical classification of protein folds using a novel ensemble classifier. *PLoS One* 8:e56499