Journal of Cell and Molecular Research (2016) 8(2), 65-70

Research Article

Unravelling Over-Represented Amino Acids in Protein Structure of Allergen Proteins; a Large-Scale Study

Nassim Rahmani[1], Esmaeil Ebrahimie[1,2,3,4*], Ali Niazi[1], Najaf Allahyari Fard[5], Bijan Bambai[5], Zarrin Minuchehr[5], Mansour Ebrahimi[6]

[1] *Institute of Biotechnology, Shiraz University, Shiraz, Iran*
[2] *Division of Information Technology, Engineering and the Environment, School of Information Technology and Mathematical Sciences University of South Australia, Adelaide, Australia*
[3] *School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, Australia*
[4] *Schools of Medicine, The University of Adelaide, Adelaide, Australia*
[5] *National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran*
[6] *Department of Biology and Bioinformatics Research Group, University of Qom, Qom, Iran*

Received 12 November 2016 Accepted 18 December 2016

Abstract

Allergens are proteins or glycoproteins which make widespread disorders that can lead to a systemic anaphylactic shock and even death within a short period of time. Understanding the protein features that are involved in allergenicity is important in developing future treatments as well as engineering proteins in genetic transformation projects. A big dataset of 1439 protein features from 761 plant allergens and 7815 non-allergen proteins was constructed. Thereafter, 10 different attribute weighting algorithms were utilized to find the key characteristics differentiating allergens and non-allergen proteins. The frequency of Leu, Arg and Gln selected by different attribute weighting algorithms with more than 50% confidence, including attribute weighting by Weight_Info Gain, Weight Chi Squared, Weight_Gini Index and Weight Relief. High amount of Gln and low percentage of Leu and Arg discriminate plant allergens from non-allergens

***Keywords*:** Plant allergens, Attribute weighting algorithms, Amino acid

Introduction

Allergy is an over-reaction of the immune system stimulated by allergens (Taylor and Hefle, 2006). Allergic reactions occur when allergenic proteins are detected by the antibody immunoglobulin E (IgE) (Van Gasse et al., 2015). During contact with the allergens, the immune system of allergic patients shows hypersensitive reactions (Ross and Montoya, 2015). Most plant allergens belong to only 10 proteins families such as Bi-functional inhibitor/lipid-transfer protein/seed storage 2S albumin, Profilin-like, Cupin, Bet V-1 like and etc. (our recent research- Unpublished). It indicated that conserved structure and biological activities play a role in determining and promoting allergenic properties (Breiteneder and Mills, 2005). Therefore, with finding the special characteristics of allergens the solution for reducing allergen properties could be offered. Investigating amino acid composition of nine different pollen allergens showed that the frequency of some amino acids such as Proline and Aspartic Acid were more than Arginine and Leucine (Mondal et al., 1998). Ara h 2, a peanut 2S albumin, has tough allergic reactions, but a homologous protein, soybean 2S albumin, is not introduced as an important allergen. Study of Structural difference between these proteins determined that some amino acids with a large side chain such as glutamine, and tyrosine were highly recognizable by the immune system (Youngshin et al., 2016).

Due to exponential increase in bioinformatics tools and techniques, huge amount of information from protein sequences can be obtained by computational biology offering a new vista for protein modeling such as various supervised (decision tree and neural network) and Unsupervised (with using the operators like K-Means, K-Medoids, Support Vector Clustering (SVC) and Expectation Maximization (EM)) machine learning algorithms and attribute weighting approaches. Weighting algorithm is a very simple method to determine important characteristics in a large database, by saving the time. According to weighted attributes we have

Corresponding authors E-mail:
* esmaeil.ebrahimie@adelaide.edu.au

better prediction and better decision. The benefit of attribute weighting for allergens is the fastest and easiest reactions to the new protein. Nowadays Attribute weighting algorithms are used in many researches to obtain original knowledge about the investigated traits such as influenza A virus (Ebrahimi et al., 2014), thermostable enzymes (Ebrahimi et al., 2011) and α-linolenic acid (ALA) (Zinati et al., 2014).

The aim of the present study was determined as amino acids frequency role in plant allergens and non-allergen proteins and find which amino acids lead the protein to be an allergen.

Materials and Methods

The structural protein attributes based on allergen and non-allergen proteins Sequence were extracted. Two Databases were created, one for plant allergens, next for Non- allergen proteins.

Data Collection

The plant allergens were collected individually according to the latest statistics data in allergen databases (SDAP, Allergenonlin, Allergome, and ADFS).

Data Pooling

A comparison of the information in different allergen sequences was performed. The information contained in the databases did not have the same formats as some of the allergens listed in some databases. All plant allergens collected in the primary secondary database were 2,424.

Data Cleaning

All plant allergens were cleaned and duplicated, while redundant and incomplete data were excluded using Clustal Omega, BioEdite and CD-HIT softwares. Finally, a total of 761 plant allergens remained.

Plant Allergens Database Creation

With more researches, attempts were made to prepare some common complete characteristics and features for all plant allergens data and eventually made our secondary database for plant allergens.

Non- allergen proteins Database Creation

Non- allergen proteins were collected randomly with omitting the isoforms, in uniprot database.

Protein Feature Extraction

The protein sequences were submitted in PROFEAT (Protein Features) webserver. It computed features from amino acid sequences and scored each attribute. PROFEAT is a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequences and determine more than 1400 features for one protein. Eventually we performed a large scale functional analysis of 20 Amino Acid Composition between 8576 proteins (plant allergens and non-allergens).

Attribute Weighting

Ten different attribute weighting algorithms such as weighting by PCA, weighting by SVM, weighting by Relief, weighting by Uncertainty, weighting by Gini index, weighting by Chi Squared, weighting by Deviation, weighting by Rule, weighting by Information Gain Ratio, and weighting by Information Gain were used to find the amino acid composition role in separating plant allergens and non-allergen proteins. Attribute weighting algorithms find the characteristics which differ in protein structure between plant allergen proteins and the non-allergen proteins. The protein attributes which were assumed to be important by most attribute weighting algorithms (intersection of different weighting methods) were assumed as the key distinguishing features of allergen proteins from the non-allergen proteins.

Weighting by PCA (Principal Component Analysis)

This operator creates attribute weights of the Example Set by using a component created by the PCA. This operator behaves exactly the same way as if a PCA model is given to the Weight by Component Model operator.

Weighting by SVM (Support Vector Machine)

This operator calculates the relevance of the attributes by computing for each attribute of the input Example Set based on the weight with respect to the class attribute. The coefficients of a hyperplane calculated by an SVM are set as attribute weights.

Weighting by Relief

This operator calculates the relevance of the attributes by Relief. The key idea of Relief is to estimate the quality of features according to how well their values distinguish between the instances of the same and different classes that are close to each other.

Weighting by Uncertainty

This operator calculates the relevance of attributes

of the given Example Set by measuring the symmetrical uncertainty with respect to the class.

Weighting by Gini Index

This operator calculates the relevance of the attributes of the given Example Set based on the Gini impurity index.

Weighting by Chi Squared

This operator calculates the relevance of the attributes by computing for each attribute of the input Example Set based on the value of the chi-squared statistic with respect to the class attribute.

Weighting by Deviation

This operator calculates the relevance of attributes of the given Example Set based on the (normalized) standard deviation of the attributes.

Weighting by Rule: This operator calculates the relevance of the attributes of the given Example Set by constructing a single rule for each attribute and calculating the errors.

Weighting by Information Gain Ratio

This operator calculates the relevance of the attributes based on the information gain ratio and assigns weights to them accordingly.

Weighting by Information Gain

This operator calculates the relevance of the attributes based on information gain and assigns weights to them accordingly.

Results

Data were normalized before running the models; consequently all weights, regardless of the employed model, were between 0 and 1. The most important amino acids that were confirmed by 10 different weighting models to be involved in differentiation of plant allergens and non-allergen proteins are shown in Table 1. In this table, each weight shows the importance of each attribute regarding the target label based on its attribute weighting model. If one amino acid composition received weight higher than 0.5 (>0.5) by a certain attribute weighting algorithm, it was assumed to be an important amino acid supplement as shown in Table 1.

Weighting by PCA

In this model all attributes weighed a value of 0.0.

Weighting by SVM

Amino Acid Composition (%) L weighted more than 0.6 by this model.

Table 1. The most important Amino Acid Composition selected by different attribute weighting algorithms in discriminating allergen proteins from non-allergen proteins.

Amino Acid Composition	The number of attribute weighting algorithms that indicated the attribute as important
Leucine	5
Arginine	4
Glutamine	4

This table presents the number of algorithms that selected the attribute. Weighting algorithms were respectively PCA, SVM, Relief, Uncertainty, Gini index, Chi Squared, Deviation, Rule, Information Gain Ratio, and Information Gain.

Weighting by Relief

When this model is applied to the dataset, 8 Amino Acid Composition (%) showed weights higher than 0.5. The Amino Acid Composition (%) L and R both weighed value of 0.9.

Weighting by Uncertainty

No attributes resulted in weights higher than 0.5.

Weighting by Gini index

Again Amino Acid Composition (%) L, R and Q were weighted higher than 0.5.

Weighting by Chi Squared

5 Amino Acid Composition (%) were weighted higher than 0.5. The Amino Acid Composition (%) L and Q were selected by this model with weight of 0.9. The Amino Acid Composition (%) R weighed the highest possible weights of 1.0.

Weighting by Deviation

In this model all attributes weighed a value of 0.0.

Weighting by Rule

No attribute weighted more than 0.6 by this model.

Weighting by Information Gain Ratio

When this algorithm was applied to the dataset, 3 Amino Acid Composition (%) had weights higher than 0.5. and The Amino Acid Composition (%) Q weighed the highest possible weights of 1.0.

Weighting by Information Gain: In this model 4 Amino Acid Composition (%) weighed higher than 0.5. The Amino Acid Composition (%) L and R both weighed a value of 1.0. The result of 10 attribute weighting models to define important Amino Acid Composition (%) were different as shown in Table 2.

Research Article

Table 2. The different algorithms weights for 20 Amino Acid Composition (%) in plant allergens and non-allergen Proteins

Weighting algorithm / Amino Acid Composition	PCA	SVM	Relief	Uncertainty	Gini Index	Chi Squared	Deviation	Rule	Info Gain Ratio	Info Gain
Alanine	0.0	0.0	0.6	0.1	0.1	0.2	0.0	0.0	0.5	0.1
Leucine	0.0	0.6	0.9	0.4	0.6	0.9	0.0	0.0	0.1	1.0
Methionine	0.0	0.1	0.4	0.1	0.2	0.1	0.0	0.2	0.7	0.2
Asparagine	0.0	0.1	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.1
Proline	0.0	0.0	0.4	0.1	0.2	0.3	0.0	0.0	0.2	0.3
Glutamine	0.0	0.3	0.5	0.4	0.7	0.9	0.0	0.4	1.0	0.6
Arginine	0.0	0.2	0.9	0.4	0.8	1.0	0.0	0.0	0.2	1.0
Serine	0.0	0.1	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Threonine	0.0	0.3	0.3	0.0	0.1	0.1	0.0	0.0	0.0	0.1
Valine	0.0	0.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tryptophan	0.0	0.1	0.4	0.0	0.0	0.1	0.0	0.0	0.0	0.1
Cysteine	0.0	0.4	0.3	0.4	0.4	0.5	0.0	0.0	0.1	0.5
Tyrosine	0.0	0.1	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Aspartic Acid	0.0	0.0	0.6	0.1	0.2	0.2	0.0	0.0	0.2	0.2
Glutamic Acid	0.0	0.1	0.6	0.1	0.2	0.3	0.0	0.0	0.0	0.3
Phenylalanine	0.0	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Glycine	0.0	0.2	0.4	0.3	0.4	0.5	0.0	0.0	0.2	0.4
Histidine	0.0	0.4	0.3	0.1	0.2	0.1	0.0	0.0	0.0	0.2
Isoleucine	0.0	0.0	0.4	0.1	0.1	0.1	0.0	0.0	0.0	0.1
Lysine	0.0	0.1	0.5	0.0	0.1	0.1	0.0	0.0	0.1	0.2

This table presents the attribute weights resulted from different algorithms

Altogether, the number of Amino Acid Composition (%) that gained weights higher than 0.5 in each weighting model were as follows: PCA (0 attribute), SVM (1 attributes), Relief (8 attributes), Uncertainty (0 attributes), Gini index (3 attributes), Chi squared (5 attributes), Deviation (0 attribute), Rule (0 attributes), Info Gain ratio (3 attributes) and Info Gain (4 attributes). The most important Amino Acid Compositions (%) that were confirmed by different weighting algorithms between plant allergens and non-allergen proteins are shown in Table 1.Three Amino Acid Composition of L, R and Q were selected by some different attribute weighting models. The average of Amino Acid Composition of L in plant allergens was 7.3 while in non-allergen proteins was 9.6. The mean value of Amino Acid Composition of Q L in plant allergens was 6.3 but in non-allergen proteins was 3.8. The average of Amino Acid Composition of R in plant allergens was 3.8 whereas in non-allergen proteins was 5.4 as showed in figure 1.

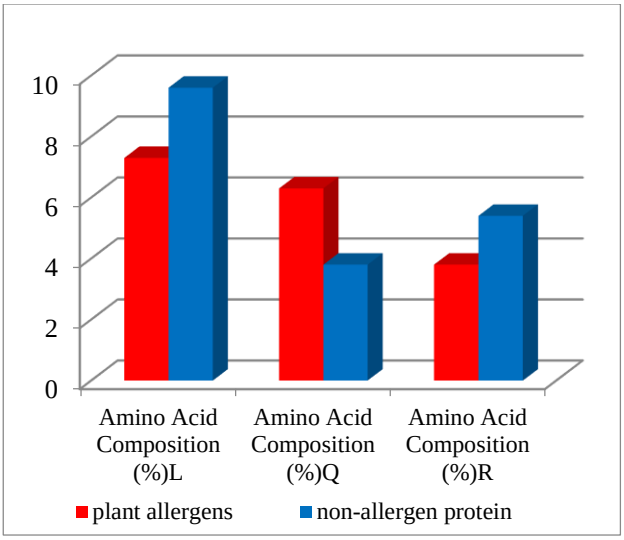

Figure 1. The mean value of three amino acids selected by attribute weighting algorithms in plant allergens and non-allergen proteins
1: Amino Acid Composition (%) L, 2: Amino Acid Composition (%) Q, 3: Amino Acid Composition (%) R

Discussion

The prevalence of allergic diseases and impact of allergic diseases are increasing worldwide. According to the World Health Organization, the number of patients having asthma is 300 million and with the rising trends it is expected to increase to 400 million, by 2025 (Pawankar et al., 2013). Thus any attempt to eliminate this problem is essential. Recently, a great attention has been paid to supervised machine learning methods implementing diverse amino acid composition and physic-chemical properties to unravel the underlying layers of protein function (Hosseinzadeh et al., 2012; Kumar et al., 2011). Mining of structural amino acid features have the potential to reflect these differences and help us to know specific changes which make a considerable impact on protein structure (Ebrahimi et al., 2014).

Each attribute weighting algorithm uses a specific pattern to define the most important features. Therefore, the results were different (Baumgartner et al., 2010; Bijanzadeh et al., 2010; Ebrahimi and Ebrahimie, 2010; Ebrahimi et al., 2011; Ebrahimie et al., 2011) as showed below: PCA, Relief, Gini Index, Chi Squared and Info Gain weighting models selected amino acid composition (%) L, with more than 50% confidence. Amino Acid Composition (%) Q Was selected by Relief, Gini Index, Chi Squared, Info Gain Ratio and Info Gain weighting algorithms with more than 50% confidence. Relief, Gini Index, Chi Squared, and Info Gain weighting models selected Amino Acid Composition (%) R, with more than 50% confidence.

Several research groups have studied and identified many different allergen proteins. Glutamine is one of the most abundant allergen protein. The studies reported that it has a great role in wheat allergens (Tanabe, 2001), peanut allergens (Youngshin et al., 2016), and also exists in pollen (Mondal et al., 1998). The Prolamin (an allergen superfamily) characterized by their high contents of glutamine (Shewry et al., 2002). This may explain why the frequency of Gln had higher weights in the applied attribute weighting algorithms we used in this study. Our results showed, more amount of Gln in plant allergens vs non-allergen proteins.

Study of homologous proteins, Ara h 2 allergen protein of Peanut with Soy Al 3 and Soy Al 1 of soybean, indicated a high level of glutamine in Ara h 2 epitopes in compared of corresponding areas in Soy Al 1 and Soy Al 3. So it can be concluded that may be glutamine is most recognizable by the immune system (Youngshin et al., 2016).

Our study demonstrated that the frequency of Leu was different in plant allergens and non-allergen proteins. The percentage of Leu in non-allergen proteins was more than plant allergens.

The investigation of 9 different allergen pollens composition of amino acids revealed an extremely low composition of leucine in only two of them (Mondal et al., 1998).

Our results determined the frequency of Arg was also weighted as a distinguishable effect on separating allergens from non-allergen proteins. The amount of this amino acid in plant allergens was lower than non-allergen proteins.

Evaluation of amino acids in pollen of 9 different allergen plants showed a small amount of arginine in only two of them (Mondal et al., 1998). Compere of Ara h 2, Soy Al 1 and Soy Al 3 revealed more percentage of Arg in some Ara h 2 epitopes, rather than Soy Al 1 and Soy Al 3(Youngshin et al., 2016), But the study had some limitations, by just focusing on Ara h 2, the result might not be enough to be generalized to other allergens also more than 30% of Ara h 2 epitopes doesn't have any Arginine in their sequence (http://www.uwm.edu.pl/biochemia/index.php/en/biopep), and because of lacking in identification of epitopes that are associated with clinical reactions; the interaction, structure and composition study of antigen–antibody are needed.

Conclusion

The findings of this study indicate that the weighting models can be efficiently used for understanding the protein. According to the result of attribute weighting algorithms in a large-scale study, high amount of Gln and low percentage of Leu and Arg discriminate plant allergens from non-allergens.

Acknowledgments

The authors are grateful to the Institute of Biotechnology of Shiraz University, National Institute of Genetic Engineering and Biotechnology (NIGEB) and Department of Biology & Bioinformatics Research Group, University of Qom for their useful help during this study.

References

1. Baumgartner C., Lewis G. D., Netzer M., Pfeifer B. and Gerszten R. E. (2010) A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after

myocardial injury. Bioinformatics 26:1745–1751.
2. Bijanzadeh E., Emam Y. and Ebrahimie E. (2010) Determining the most important features contributing to wheat grain yield using supervised feature selection mode. Australian Journal of Crop Science 4:402-407.
3. Breiteneder H. and Mills C. (2005) Plant food allergens—structural and functional aspects of allergenicity. Biotechnology Advances 23:395-399.
4. Ebrahimi M., Aghagolzadeh P., Shamabadi N., Tahmasebi A., Alsharifi M., Adelson D. L., Hemmatzadeh F. and Ebrahimie E. (2014) Understanding the underlying mechanism of HA subtyping in the level of physic-chemical characteristics of Protein. PLOS ONE 9.
5. Ebrahimi M. and Ebrahimie E. (2010) Sequence-based prediction of enzyme thermostability through bioinformatics algorithms. Current Bioinformatics 5.
6. Ebrahimi M., Lakizadeh A., AghaGolzadeh P., Ebrahimie E. and Ebrahimi M. (2011) Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. PLoS One 6.
7. Ebrahimie E., Ebrahimi M. and Sarvestani N. R. (2011) Protein attributes contribute to halo _ stability, bioinformatics approach. Saline Systems 7.
8. Hosseinzadeh F., Ebrahimi M., Goliaei B. and Shamabadi N. (2012) Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. PLoS One 7.
9. Kumar M., Gromiha M. M. and Raghava G. P. (2011) SVM based prediction of RNA binding proteins using binding residues and evolutionary information. Journal of Molecular Recognition 24:303-313.
10. Mondal A. K., Parui S. and Mandal S. (1998) Analysis of the free amino acid content in pollen of nine Asteraceae species of known allergenic activity. Annals of Agricultural and Environmental Medicine 5:17-20.
11. Pawankar R., Canonica G. W., Holgate S. T., Lockey R. F. and Blaiss M. (2013) The WAO White Book on Allergy.
12. Ross S. M. and Montoya M. I. (2015) Allergic reactions. *In* Basic Clinical Anesthesia. Springer. 197-202.
13. Shewry P. R., Beaudoin F., Jenkins J., Griffiths S. and Mills E. (2002) Plant protein families and their relationships to food allergy. Biochemical Society Transactions 30:906-910.
14. Tanabe S. (2001) Identification of wheat allergens. Internet Symposium on Food Allergens 3:163_170.
15. Taylor S. L. and Hefle S. L. (2006) Food allergen labeling in the USA and Europe. Current Opinion in Allergy and Clinical Immunology 6:186–190.
16. Van Gasse A., Mangodt E., Faber M., Sabato V., Bridts C. and Ebo D. (2015) Molecular allergy diagnosis: Status anno 2015. Clinica Chimica Acta 444:54-61.
17. Youngshin H., Jing L., Ludmilla B., Galina A. G., Chaeyoon L., Won H. S. and Hugh A. S. (2016) What characteristics confer proteins the ability to induce allergic responses? IgE epitope mapping and comparison of the structure of soybean 2S albumins and ara h 2. Molecules 21.
18. Zinati Z., Zamansani F., KayvanJoo A., Ebrahimi M., Ebrahimi M., Ebrahimie E. and MohammadiDehcheshmeh M. (2014) New layers in understanding and predicting α-linolenic acid content in plants using amino acid characteristics of omega-3 fatty acid desaturase. Computers in Biology and Medicine 54:14-23.