

A Comparative Analysis of Allergen Proteins between Plants and Animals Using Several Computational Tools and Chou's PseAAC Concept

Mandana Behbahani Parisa Rabiei Hassan Mohabatkar

Department of Biotechnology, Faculty of Biological Science and Technology, University of Isfahan, Isfahan, Iran

Keywords

Allergen · Physiochemical properties · Secondary structure · Epitope prediction · PseAAC

Abstract

Background: A large number of allergens are derived from plant and animal proteins. A major challenge for researchers is to study the possible allergenic properties of proteins. The aim of this study was in silico analysis and comparison of several physiochemical and structural features of plant- and animal-derived allergen proteins, as well as classifying these proteins based on Chou's pseudo-amino acid composition (PseAAC) concept combined with bioinformatics algorithms.

Methods: The physiochemical properties and secondary structure of plant and animal allergens were studied. The classification of the sequences was done using the PseAAC concept incorporated with the deep learning algorithm. Conserved motifs of plant and animal proteins were discovered using the MEME tool. B-cell and T-cell epitopes of the proteins were predicted in conserved motifs. Allergenicity and amino acid composition of epitopes were also analyzed via bioinformatics servers. **Results:** In comparison of physiochemical features of animal and plant allergens, extinction coefficient was different significantly. Secondary structure prediction showed more random coiled structure in plant

allergen proteins compared with animal proteins. Classification of proteins based on PseAAC achieved 88.24% accuracy. The amino acid composition study of predicted B- and T-cell epitopes revealed more aliphatic index in plant-derived epitopes. **Conclusions:** The results indicated that bioinformatics-based studies could be useful in comparing plant and animal allergens.

© 2020 S. Karger AG, Basel

Introduction

Allergy can be emerged by several inflammatory symptoms such as rhinitis, urticaria, asthma, and anaphylactic immediate and late reactions [1, 2]. Mast cells and basophils have critical roles in allergy procedure and are subsequently activated by immune system components: antibodies of IgE class and T_H2-mediated pathways [3, 4]. Typically, an allergen source releases a complex mixture of potentially allergenic macromolecules such as carbohydrates, complex lipids, polypeptides, glycoproteins, and peptidoglycans. Protein allergens as the major group of allergens are intended subject of this assessment. Answering to the questions what makes a protein an allergen

Edited by: O. Palomares, Madrid.

and what differences are between allergens' properties derived from different sources is one of the major research focuses worldwide [5, 6]. The ability of an allergen to eliciting the immune responses directly refers to molecular characteristics of allergens, especially allergen epitopes located in conserved sequences [7, 8]. In the scientific literature, common features of allergenic proteins such as molecular size, stability, and solubility have been repeatedly studied [9, 10]. While these properties are not adequate and unique in allergens [9], more investigations into the specific molecular property and a comparative study on different source-originated allergens are required.

In an overall classification based on allergen sources, these proteins are categorized into 2 groups: plant-derived and animal-derived allergens. It is demonstrated that animal-derived allergens belong to certain protein family clusters, including lipocalins, caseins, serum albumins, and tropomyosins [11–13]. Plant allergens are mainly classified in protein families of the plant defense system [14] such as chitinases and lipid transfer proteins [15, 16]. Since the diversity of animal and plant allergens has not been investigated yet, investigating the common properties of these proteins for understanding the nature of their allergenicity is indispensable.

To detect the allergens, 2 common approaches are used currently: (1) protein-based methods such as enzyme-linked immunosorbent assay (ELISA), mass spectrometry, and 2D electrophoresis; (2) DNA-based methods such as polymerase chain reaction (PCR) and real-time PCR [17]. Since all of these experimental methods are time consuming and laborious, studying the allergen characteristics, structure, sequence, and mechanism using bioinformatics tools is helpful for developing the novel diagnostic and therapeutic strategies. The aim of this research was to study the animal and plant allergen proteins using bioinformatics tools and to compare these 2 datasets to find out similarities and differences in their molecular and structural characters.

Materials and Methods

Data Collection

Amino acid sequences of animal and plant allergens that belong to different protein families were randomly collected from NCBI (<http://www.ncbi.nlm.gov>). The dataset contained 270 sequences for plant allergens and 123 for animal ones. Then, the data were analyzed with the CD-HIT tool with a threshold of 90% to decrease redundancy. The number of animal and plant allergen sequences reduced to 110 and 68, respectively. The new datasets were used in next analyses.

Pseudo-Amino Acid Composition

Pseudo-amino acid composition (PseAAC) was originally introduced in 2001 by Chou [18] to improve the protein classification accuracy [19]. Based on the PseAAC, a protein sequence can be formulated by mathematical concepts avoiding losing the information hidden in sequence order of proteins and peptides [20]. Because of progression of PseAAC for protein sequences, this approach was expanded to implement in DNA and RNA sequences [21, 22]. So far, PseAAC and PseKNC concepts have been used in a great number of research studies, for instance, for identifying C-terminal cysteine prenylation sites in proteins [23], predicting the posttranslational modification (PTM) sites [20, 23], predicting hydroxyproline and hydroxylysine in proteins [24], identifying protein-protein interactions [25], identifying phosphothreonine sites [26], and identifying DNA replication origins [27]. In this research, the PseAAC concept was used to incorporate the sequence order, composition, and physicochemical properties of plant and animal allergen proteins in prediction.

Generation of PseAAC Number

Several web servers are available to generate various modes of pseudo-amino acid/DNA/RNA composition such as PseAAC (<http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/>) [19], PseKNC (<http://lin.uestc.edu.cn/pseknc/default.aspx>) [28], Pse-in-One (<http://bioinformatics.hitsz.edu.cn/Pse-in-One/server/>) [29], and BioSeq Analysis (<http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/server/>). However, the physicochemical properties which are available in these servers are different.

In this work, the PseAAC components were achieved for collected animal and plant allergen sequences by the PseAAC server in the Type 1 mode and $\lambda = 3$ and $w = 0.05$. There are 6 different properties to be selected by users including hydrophobicity, hydrophilicity, pI, pK1, pK2, and mass. Sixty three different combinations of these properties were incorporated in PseAAC generation. In result part, $20 + \lambda$ numbers are produced by the web server. Subsequently, for $\lambda = 3$, $20 + 3$ numbers were achieved in which first 20 numbers represent the values of the conventional sequence compositions and the last 3 ones represent the values of sequence-order correlation. In this work, the last 3 numbers for all protein samples were collected for processing in a data mining program.

Classification Based on PseAAC

In this step, several algorithms were used for classification of the obtained data. K-nearest neighbors (k-NN), gradient boosted tree, and deep learning are such of these algorithms. Data mining processes was run in RapidMiner studio software version 9.3 [30] (downloadable at <https://rapidminer.com/get-started/>). Among plant and animal proteins, one group was selected as the positive dataset and the other one as the negative dataset. Then, several mentioned algorithms were chosen to be applied on the datasets. In the last section, the potency of each algorithm was evaluated by applying the 10-fold cross-validation test. Performance precision of algorithms is quantified by 4 parameters, sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC), in the cross-validation test. It is demonstrated that if the values for Sn, Sp, and Acc achieved >80%, the classification method is acceptable. Hence, all of these parameters were calculated for tested algorithms and the most validated algorithm was selected as our dataset classifier.

Physiochemical Properties

In this study, ProtParam (<https://web.expasy.org/protparam/>) [31] as a reliable tool from ExPASy resource portal was used to compute some of the physicochemical properties of proteins. The parameters scored by ProtParam include theoretical pI, amino acid composition, atomic composition, molecular weight, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for a protein which is carried out using a spectrophotometer while purifying it. It has been shown that it is possible to estimate the molar extinction coefficient of a protein from knowledge of its amino acid composition. More details about calculating the extinction coefficient is available at <https://web.expasy.org/protparam/protparam-doc.html>. The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids divided by the number of residues in the sequence. The larger the number is, the more hydrophobic the amino acid. The animal and plant datasets were analyzed using the ProtParam tool to achieve the aforementioned protein properties.

Secondary Structure Prediction

The secondary structure has 3 regular forms such as helical, extended, and loops which are also known as α helices, β sheets, and random coils, respectively [32]. The GOR method is one of the successful methods to predict protein secondary structure from the amino acid sequence [33]. This method is based on the analysis of residues' occurrence frequencies and a sum of information from single residues and pairs of residues [32, 34]. In a protein structure prediction, the inputs are the amino acid sequences and the output is the predicted structures which is a combination of α helices, β sheets, and loops [2, 32, 35]. The GOR IV tool [36] available at https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=C2%BCnpsa_gor4.htm was used to predict plant and animal allergens' secondary structures.

Statistical Analysis

In silico comparison of secondary structures and physiochemical features of plant and animal allergen proteins was evaluated using receiver operating characteristic (ROC) curve analysis. The ROC curve is a graphical plot of the true positive rate against the false positive rate. The ROC curve shows the test accuracy and relationship between specificity and sensitivity. When the graph is closer to the top and left-hand borders, the test is more accurate. Test accuracy is also shown as the area under the curve (AUC), that is, the greater the AUC, the more the accuracy [37].

Epitope Identification

MEME Motif Discovery

As the function of a protein depends on its structure, finding the short patterns in a protein structure may aid in discovering the role of protein in the cell. Since the role of a protein and its interactions in the cell are dependent on the protein structure, it is important to find out the structural motifs in a protein. To study the sequence motifs in animal and plant allergens, MEME Suite web server available at <http://meme-suite.org/tools/meme> was used with the following parameters: the number of motifs which the server should find, 8; distribution of motif sites, 0 or 1 occurrence per sequence; minimum and maximum width of motifs; 6 and 50, respectively.

Table 1. Results of PseAAC-based classification of datasets via the 10-fold cross-validation test

Method	ACC, ¹ %	Sp, ² %	Sn, ³ %	MCC ⁴
PseAAC + k-nearest neighbors	87.12	79.41	91.82	0.71
PseAAC + gradient boosted trees	84.35	96.36	64.71	0.66
PseAAC + deep learning	88.24	90.91	83.82	0.74

The most valuable algorithm for distinguishing between plant and animal allergens is shown in bold. PseAAC, pseudo-amino acid composition. ¹Accuracy. ²Specificity. ³Sensitivity. ⁴Matthews correlation coefficient.

Epitope Prediction

Through the IgE-mediated and IgE-independent allergenicity reactions, an allergen can induce the humoral (B cell) and adaptive (T cell) immune response [38, 39]. To investigate the power of an allergen in stimulating the immune cells and their reactions' mechanism, finding the epitopes of an allergen is necessary. As the structural motifs correspond to vital, functional, and mainly exposed sections in a protein, B-cell and T-cell epitopes were merely studied on motifs discovered by using the MEME suit tool. For T-cell epitopes, MHC I and MHC II binding epitopes were predicted by the following servers: <http://tools.immuneepitope.org/mhci/> and <http://tools.immuneepitope.org/mhcii/>. In addition, linear B-cell epitopes were predicted using Bcepred web server available at <http://crdd.osdd.net/raghava/bcepred/>.

Allergenicity of the Epitopes

To identify the allergenicity of the predicted epitopes, AlgPred (<http://crdd.osdd.net/raghava/algpred/>) [40] was used. There are 6 different approaches to select for analyzing the data. In this research, the "SVM module based on amino acid composition" approach was chosen. Since the input sequences of AlgPred should have the minimum length of 11, the length of epitopes was set to 11 in epitope prediction tools.

Amino Acid Composition of Epitopes

Amino acid composition of B-cell and T-cell epitopes with positive allergenicity was analyzed with the ProtParam tool. Then, the percentage of each amino acid in total epitopes was calculated.

Results

Cross-Validation of PseAAC-Based Classification

Table 1 shows the results of 3 algorithms used in classifying the datasets. The percentages of Acc, Sn, Sp, and MCC were achieved by the 10-fold cross-validation test. The result demonstrated that the deep learning algorithm was more powerful to compare these datasets.

Table 2. ROC curve analysis results of 6 types of physiochemical properties achieved by ProtParam server for comparison of plant and animal allergens

Physicochemical properties	Extinction coefficient	Molecular weight	Isoelectric point	Aliphatic index	GRAVY
Acc ¹	0.809	0.794	0.668	0.616	0.645
AUC ²	0.750	0.750	0.694	0.666	0.722

ROC curve, receiver operating characteristic curve; GRAVY, grand average of hydropathicity. ¹ Accuracy. ² Area under the curve.

Table 3. ROC curve analysis results of the GOR IV prediction server

2D structure	Alpha helix	Extended strand	Random coil
Acc ¹	0.652	0.652	0.913
AUC ²	0.583	0.621	0.931

Accuracy and area under the curve for 3 components of 2D structure (α helix, extended strand, and random coil) were calculated for comparing plant and animal allergens. ROC curve, receiver operating characteristic curve. ¹ Accuracy. ² Area under the curve.

Physiochemical Properties

Five parameters of ProtParam results were analyzed using the ROC curve. Despite of inconsiderable difference in Acc values for molecular weight, isoelectric point, aliphatic index, and GRAVY, the ACC value of extinction coefficient showed the difference between plant and animal allergens (Table 2).

Secondary Structure Analysis

The results of ROC curve analysis of 3 forms of secondary structure, α helices, extended strands, and random coils, is demonstrated in Table 3. Animal and plant allergen proteins showed significant difference in the percentage of random coils. However, the Acc values of α helices and extended strands are not different significantly.

Epitope Analysis

MEME Motif Discovery

The 8 motifs in plant allergen proteins and 8 motifs in animal allergen proteins which were achieved by MEME

Suite web server are shown in Figure 1. The more the e-value comes close to zero, the more valuable the motifs. These motifs revealed the conserved parts in the protein sequence. Subsequently, the exposed plant and animal motifs were applied in epitope prediction tools as input sequences. As it is shown in Figure 1, the majority of motifs were more than forty amino acids in length and e-values of all discovered motifs were near to zero.

Epitope Prediction

The aforementioned motifs were inputted in B- and T-cell servers. The resulted 11-mer epitopes are displayed in Tables 4–6. To interpret the predicted T-cell epitopes, 2 values are considered, IC₅₀ and percentile rank. Among predicted epitopes, ones with lower value of IC₅₀ and percentile rank have more affinity to target the MHC molecule. Then, the epitopes were sorted using IC₅₀ and percentile rank. In case of B-cell binding epitopes, 4 epitopes for animal sequences and 5 epitopes for plant sequences were achieved. In addition, 8 epitopes for plants and 8 for animals were predicted as MHC I and MHC II binding epitopes. In the next step, the predicted 11-mer epitopes which are investigated by B- and T-cell servers were analyzed by the allergenicity server.

Allergenicity

The allergenicity of all chosen predicted epitopes was analyzed by the AlgPred web server. The sequences with the score of ≥ -0.4 (threshold) is considered as an allergen. Among B-cell epitopes, except for one peptide, all other epitopes were detected as allergens. In addition, all 16 T-cell epitopes in both plants and animals were also predicted as allergens (Tables 4–6).

Amino Acid Composition of Epitopes

The results of amino acid composition analysis represented similarities and differences between amino acid content in plant and animal epitopes. In terms of similarity, in both animal and plant epitopes the percentage of Lys was high, but the Trp and Met were not seen. From a difference perspective, Cys was only detected in animal epitopes, while Arg was only seen in plant epitopes. Furthermore, the amounts of Ala and Leu were significantly high in plant epitopes. However, animal epitopes were rich in Asp and Glu (data not shown). According to the results of amino acid composition analysis, the average percentage of most aliphatic amino acids (i.e., Ala and Val) was more in plant B-cell and T-cell epitopes compared with animal epitopes (Fig. 2).

Fig. 1. The motifs discovered in animal (a) and plant (b) allergen proteins using the MEME Suite web server. Eight motifs were predicted for plant and animal allergens, separately. The e-value shows the accuracy of the predicted motif; the less the e-value, the more the possible motif. Width parameter reveals the sequence length of the identified motif.

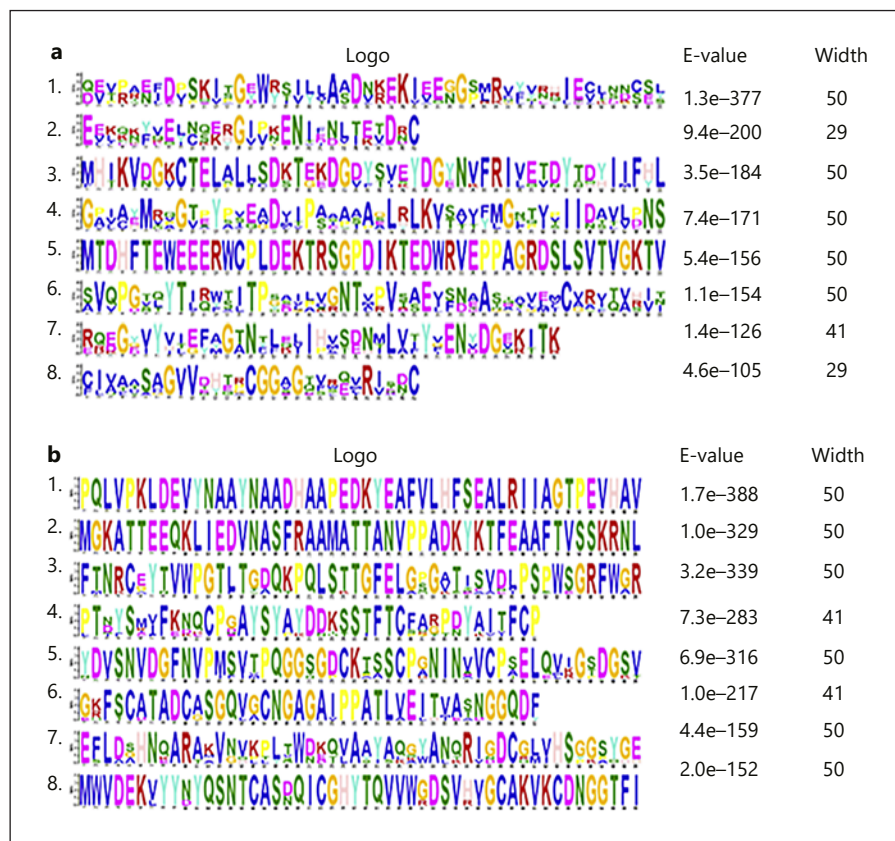


Table 4. Predicted allergenicity of animal and plant B-cell epitopes using the AlgPred server

Animal epitopes	Score	Allergenicity	Plant epitopes	Score	Allergenicity
ASDVREKIEEN	-0.091958881	+	DHAAPEDKYE	-0.31895262	+
LSDKTEKDGDY	-0.19467195	+	GKATTEEQKLI	-0.063545462	+
CPLDEKTRSGP	-0.35290419	+	AAFTVSSKRNL	-0.19867409	+
VEPPAGRDSLS	-0.73778407	-	FLDSHNQARAK	-0.043317506	+
			GTLTGDQKPQL	-0.2093773	+

Table 5. Predicted allergenicity of animal and plant MHC I binding epitopes using the AlgPred server

Animal epitopes	Score	Allergenicity	Plant epitopes	Score	Allergenicity
ISGYWYSVLLA	-0.28983018	+	AADHAAPEDKY	-0.26350853	+
FVEYCQKHGIV	0.1201639	+	TTANVPPADKY	-0.26668651	+
LSDKTEKDGDY	-0.194672	+	VTISVDLPSPW	-0.22549735	+
PIAHMVVGTPY	-0.27600228	+	STFTCFARPDY	-0.20621452	+
MTDHFTEWEEE	-0.24688564	+	NVDGFNVPMSTV	-0.20425316	+
ISDAASATVLM	-0.079121637	+	ITVASNGGQDF	0.56863889	+
EVIHVSENMLV	-0.22642765	+	AAQAQGYANQR	-0.25081337	+
CSAGVVEHQPC	-0.083914599	+	YTQVVWGDVSH	-0.24093932	+

Fig. 2. The average percentage of most aliphatic amino acids (Val and Ala) in B-cell epitope and MHC I and MHC II binding epitopes in animals and plants are compared. All epitopes mentioned in Tables 4, 5, and 6 were used for aliphatic index analysis.

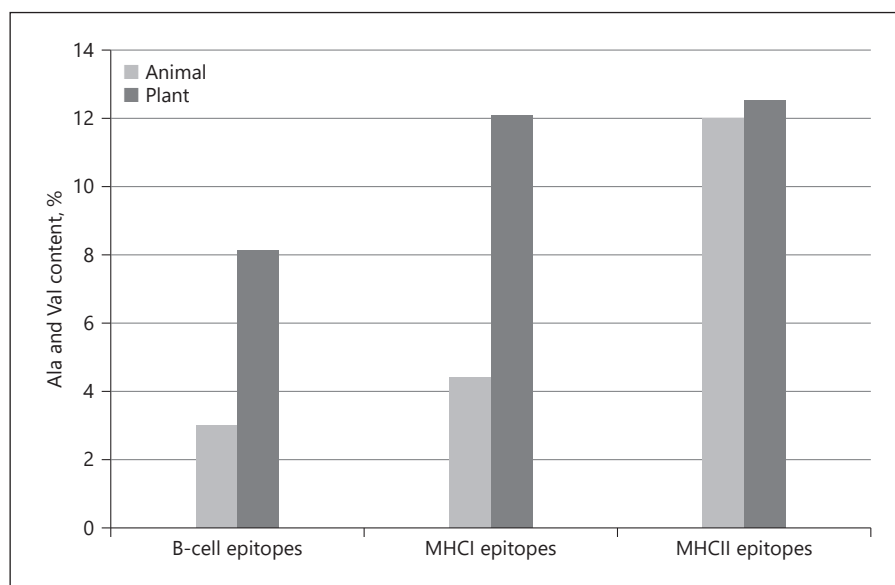


Table 6. Predicted allergenicity of animal and plant MHC II binding epitopes using the AlgPred server

Animal epitopes	Score	Allergenicity	Plant epitopes	Score	Allergenicity
YWYSVLLASDV	-0.27583172	+	AFVLHFSEALR	-0.24241872	+
GIVKENIFNLT	0.10736328	+	ASFRAAMATTA	-0.26145603	+
NVFRIVETDYT	-0.08176071	+	TGFELGSGVTI	-0.2530809	+
TVFMGQTVVII	-0.24921068	+	MIFKNQCPGAY	0.1278998	+
EDWRVEPPAGR	-0.45353397	-	INVVCPSSELQV	-0.13850561	+
ATVLMCARITA	-0.27189321	+	ATLVEITVASN	0.065508558	+
EVIHVSENMLV	-0.22642765	+	LTWQDKQVAAYA	-0.14642929	+
CIVACSAGVVE	-0.19413594	+	YYNYQSNTCAS	-0.23700282	+

Discussion

Allergy is an immunologically mediated hypersensitivity reaction to a foreign substance that is not typically harmful to the body. In the current study, randomly selected plant and animal allergen proteins were analyzed and then compared via several approaches: (1) machine learning-based methods to distinguish these 2 allergen sources, (2) exhibited physiochemical properties, (3) secondary structure content, and (4) allergenic B- and T-cell epitopes and amino acid composition of the aforementioned epitopes. In recent research studies, the structure and physiochemical features of all kind of allergen proteins have been studied [16], but the comparison study between animal and plant allergens have been done here for the first time. The ProtParam tool was utilized to study the physiochemical properties of the mentioned al-

lergens. Except for the extinction coefficient parameter, the ProtParam results showed no notable difference between plant and animal allergens in physiochemical properties. The extinction coefficient of animal allergens was more than plant allergens with the Acc of 80.9%. This result demonstrates that animal allergens have more light adsorption at a certain wavelength, which is referred to different amino acid compositions of animal allergens than plant allergens. In fact, the more the extinction coefficient value of animal allergens is, the more the Tyr, Cys, and Trp amino acids (aromatic amino acids) in their sequence order. The computed extinction coefficient values are useful in the quantitative study of protein-protein and protein-ligand interactions in solution and are proportional to the binding strength [1, 41, 42].

Furthermore, plant allergens exhibited more random coils in secondary structure analysis by the GOR IV serv-

er. Since the random coils in a protein structure is mainly located in the surface domains, these random coiled parts have more likelihood ratio to include B-cell epitopes [43, 44]. This matter is also confirmed through our epitope prediction results, in which the numbers of plant B-cell epitopes were significantly more than animal ones. In addition, plant epitopes showed more Val and Ala content, in overall. Ala and Val are known as most aliphatic amino acids [45].

According to the previous studies, both linear and conformational B-cell epitopes and T-cell epitopes, especially MHC II binding ones, involve in allergenicity of a protein [46–51]. In this work, the B- and T-cell epitopes were predicted in conserved motifs, and then, the IgE binding of predicted epitopes were analyzed in silico. Ebrahimi et al. [52] identified B-cell epitopes of 25 allergen proteins of mites in silico. Also, they found that the most numbers of predicted epitopes are dominantly located in the conserved sequences [52]. Furthermore, Chen et al. [53] research on pollen allergens suggested that allergenic epitopes were located in conserved functional domains of proteins.

Moreover, in this research, a combined approach of the PseAAC concept and machine learning-based technique was utilized to compare the plant and animal allergens. In 2013, Mohabatkar et al. [54] used the PseAAC to predict the allergen proteins from nonallergen ones. They achieved the accuracy of 91.19% via the support-vector machines (SVM) algorithm, which was a significant success. In addition, very recently, Westerhout et al. [55] published a paper on predicting the allergenicity of novel and modified proteins using random forest algorithm based on protein physicochemical and biochemical properties. Their results showed robust model performance with an accuracy of $\geq 85\%$. However, it should be noted that in our research, a method was developed to distinguish between 2 allergen groups, plant- and animal-derived allergens based on PseAAC. Among applied algorithms, the most valuable performance belonged to deep learning algorithm with the accuracy of 88.24%. In measuring the efficiency of a predicting method, the accuracy, sensitivity, and specificity percentage more than 80 indicates the reliable and valuable performance of the predictor. In addition, the Matthews correlation coefficient, a continuous variable between -1 and $+1$, is an important parameter in interpretation of the results, such that -1 , $+1$, and 0 indicate the weakest performance, the best performance, and random prediction, respectively. The resulted parameters in our developed method, Acc = 88.24%, Sp = 90.91%, Sn = 83.82%, and MCC = 0.74, reveal the reliability of this method.

Conclusion

In this research, plant and animal allergen proteins were compared successfully using bioinformatics tools and servers. It was found that the PseAAC concept incorporated with deep learning algorithm is a promising method to classify plant and animal allergens. The assessment of similarities and differences discovered in physicochemical properties of the aforementioned allergens, as well as predicted B- and T-cell epitopes and their amino acid compositions, could be useful to understand the essence of allergen molecules and design the more targeted drugs. Furthermore, comparative studies similar to the present research can pave the way for developing in silico methods which not only can predict possible allergenicity of a molecule but also identify its source.

Acknowledgement

Support of this work by the University of Isfahan is highly acknowledged.

Statement of Ethics

Since this is an in silico analysis, no statement of ethics is needed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

The authors did not receive any funding.

Author Contributions

M. Behbahani designed and performed the bioinformatics analysis and interpreted the results. P. Rabiei performed the bioinformatics analysis and interpreted the results and was a major contributor in writing the manuscript. H. Mohabatkar designed and performed the bioinformatics analysis and interpreted the results. All authors read and approved the final manuscript.

Data Availability Statement

The datasets used during the current study are available from the corresponding author on reasonable request.

References

- Adikaram PR, Beckett D. Protein:protein interactions in control of a transcriptional switch. *J Mol Biol*. 2013;425(22):4584–94.
- Bordoloi H, Sarma KK. Protein structure prediction using artificial neural network. *IJCA* 2011;3:22–5.
- Sicherer SH, Sampson HA. Food allergy: a review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *J Allergy Clin Immunol*. 2018;141(1):41–58.
- Maazi H, Banie H, Muench GRA, Patel N, Wang B, Sankaranarayanan I, et al. Activated plasmacytoid dendritic cells regulate type 2 innate lymphoid cell-mediated airway hyper-reactivity. *J Allergy Clin Immunol*. 2018; 141(3):893–905. e6.
- Pekar J, Ret D, Untertsmayr E. Stability of allergens. *Mol Immunol*. 2018;100:14–20.
- Cui J, Han LY, Li H, Ung CY, Tang ZQ, Zheng CJ, et al. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol Immunol*. 2007;44(4):514–20.
- Masilamani M, Commisins S, Shreffler W. Determinants of food allergy. *Immunol Allergy Clin North Am*. 2012;32(1):11–33.
- Willeroider M, Fuchs H, Ballmer-Weber BK, Focke M, Susani M, Thalhamer J, et al. Cloning and molecular and immunological characterisation of two new food allergens, Cap a 2 and Lye e 1, profilins from bell pepper (*Capsicum annuum*) and tomato (*Lycopersicon esculentum*). *Int Arch Allergy Immunol*. 2003;131(4):245–55.
- Aalberse RC. Structural biology of allergens. *J Allergy Clin Immunol*. 2000;106(2):228–38.
- Huby RD, Dearman RJ, Kimber I. Why are some proteins allergens? *Toxicol Sci*. 2000; 55(2):235–46.
- Kuehn A, Hilger C. Animal allergens: common protein characteristics featuring their allergenicity. *Front Immunol*. 2015;6:40.
- Dall'Antonia F, Pavkov-Keller T, Zangger K, Keller W. Structure of allergens and structure based epitope predictions. *Methods*. 2014; 66(1):3–21.
- Radauer C, Bublin M, Wagner S, Mari A, Breiteneder H. Allergens are distributed into few protein families and possess a restricted number of biochemical functions. *J Allergy Clin Immunol*. 2008;121(4):847–52.e7.
- Breiteneder H, Radauer C. A classification of plant food allergens. *J Allergy Clin Immunol*. 2004;113(5):821–30.
- Breiteneder H, Ebner C. Molecular and biochemical classification of plant-derived food allergens. *J Allergy Clin Immunol*. 2000; 106(1, Pt 1):27–36.
- Garino C, Coisson JD, Arlorio M. In silico allergenicity prediction of several lipid transfer proteins. *Comput Biol Chem*. 2016;60:32–42.
- Rizou K. Detection of animal allergens in foods. In: *Food allergen testing*. John Wiley & Sons; 2014. p. 67–103.
- Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*. 2001;43(3):246–55.
- Shen H-B, Chou K-C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal Biochem*. 2008;373(2):386–8.
- Mohabatkar H, Rabiei P, Alamdaran M. New achievements in bioinformatics prediction of post translational modification of proteins. *Curr Top Med Chem*. 2017;17(21):2381–92.
- Zhou X, Li Z, Dai Z, Zou X. Predicting methylation status of human DNA sequences by pseudo-trinucleotide composition. *Talanta*. 2011;85(2):1143–7.
- Chen W, Feng P-M, Lin H, Chou K-C. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Res Int*. 2014;2014:623149.
- Xu Y, Wang Z, Li C, Chou K-C. iPrePseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med Chem*. 2017;13(6):544–51.
- Xu Y, Wen X, Shao X-J, Deng N-Y, Chou K-C. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int J Mol Sci*. 2014;15(5):7594–610.
- Jia J, Li X, Qiu W, Xiao X, Chou K-C. PPI-PseAAC (CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC. *J Theor Biol*. 2019; 460:195–203.
- Khan YD, Rasool N, Hussain W, Khan SA, Chou K-C. iPhosT-PseAAC: identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. *Anal Biochem*. 2018;550:109–16.
- Liu B, Weng F, Huang D-S, Chou K-C. iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. *Bioinformatics*. 2018;34(18):3086–93.
- Chen W, Lei T-Y, Jin D-C, Lin H, Chou K-C. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal Biochem*. 2014;456:53–60.
- Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*. 2015;43(W1):W65–71.
- Mierswa I, Klinkenberg R. Rapidminer studio. 2019.
- Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. *The proteomics protocols handbook*. Totowa, NJ, USA: Springer; 2005. p. 571–607.
- Rout SB, Mishra S, Mishra S. A review on application of artificial neural network (ANN) on protein secondary structure prediction. Paper presented at: 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India; 2017.
- Kloczkowski A, Ting K-L, Jernigan R, Garnier J. Protein secondary structure prediction based on the GOR algorithm incorporating multiple sequence alignment information. *Polymer*. 2002;43(2):441–9.
- Lee J, Freddolino PL, Zhang Y. Ab initio protein structure prediction. In: *Rigden DJ, editor. From protein structure to function with bioinformatics*. Dordrecht, Netherlands: Springer; 2017. p. 3–35.
- Mitra S, Hayashi Y. Bioinformatics with soft computing. *IEEE Trans Syst Man Cybern C Appl Rev*. 2006;36(5):616–35.
- Garnier J. GOR secondary structure prediction method version IV. *Meth Enzym, RF Doolittle Ed*. 1998;266:540–53.
- Carter JV, Pan J, Rai SN, Galandiuk S. ROC-ing along: evaluation and interpretation of receiver operating characteristic curves. *Sur-gery*. 2016;159(6):1638–45.
- Nauta AJ, Engels F, Knippen LM, Garssen J, Nijkamp FP, Redegeld FA. Mechanisms of allergy and asthma. *Eur J Pharmacol*. 2008; 585(2–3):354–60.
- Shirinbak S, Taher YA, Maazi H, Gras R, van Esch BC, Henricks PA, et al. Suppression of Th2-driven airway inflammation by allergen immunotherapy is independent of B cell and Ig responses in mice. *J Immunol*. 2010;185(7): 3857–65.
- Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res*. 2006;34(Suppl_2): W202–9.
- Kurrey K, Paramanik V. Identification and physicochemical analysis of ERK interacting proteins using bio-computational tools. *World*. 2018;8:303–13.
- Jabalia N, Bansal H, Mishra P, Chaudhary N. In silico comparative analysis of papain family cysteine protease using computational tools and servers. *Int J Basic Appl Eng Res*. 2015;2(5):310–4.
- Li Y, Liu X, Zhu Y, Zhou X, Cao C, Hu X, et al. Bioinformatic prediction of epitopes in the Emy162 antigen of *Echinococcus multilocularis*. *Exp Ther Med*. 2013;6(2):335–40.
- Tahmoorepour M, Nazifi N, Pirkhezranian Z. In silico prediction of B-cell and T-cell epitopes of protective antigen of *Bacillus anthracis* in development of vaccines against anthrax. *Iran J Appl Anim Sci*. 2017;7(3):429–36.
- IKAI A. Thermostability and aliphatic index of globular proteins. *J Biochem*. 1980;88(6): 1895–8.

- 46 Zeiler T, Virtanen T. Mapping of human T-cell epitopes of allergens. In: [Jones MG, Lympany P, editors. Allergy methods and protocols](#). Totowa, New Jersey: Springer, Humana Press; 2008. p. 51–56.
- 47 Pomés A. Relevant B cell epitopes in allergic disease. [Int Arch Allergy Immunol](#). 2010; 152(1):1–11.
- 48 Larché M. T cell epitope-based allergy vaccines. In: [Valenta R, Coffman RL, editors. Vaccines against allergies](#). Berlin/Heidelberg: Springer; 2011. p. 107–119.
- 49 Westernberg L, Schulten V, Greenbaum JA, Natali S, Tripple V, McKinney DM, et al. T-cell epitope conservation across allergen species is a major determinant of immunogenicity. [J Allergy Clin Immunol](#). 2016;138(2):571–8.e7.
- 50 Mishra A, Jain A, Arora N. Mapping B-cell epitopes of major and minor peanut allergens and identifying residues contributing to IgE binding. [J Sci Food Agr](#). 2016;96(2):539–47.
- 51 Schulten V. [Strategies to study T cells and T cell targets in allergic disease](#). Rijeka, Croatia: IntechOpen; 2017. p. 123.
- 52 Ebrahimi N, Nezafat N, Esmaeilzadeh H, Ghasemi Y, Nabavizadeh SH, Alyasin S. In silico prediction of B-cell epitopes for twenty-five mite allergens: the therapeutic potentials for immunotherapy. [Mol Cell Probes](#). 2019; 46:101408.
- 53 Chen M, Xu J, Devis D, Shi J, Ren K, Searle I, et al. Origin and functional prediction of pollen allergens in plants. [Plant Physiol](#). 2016; 172(1):341–57.
- 54 Mohabatkar H, Mohammad Beigi M, Abdolahi K, Mohsenzadeh S. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. [Med Chem](#). 2013; 9(1):133–7.
- 55 Westerhout J, Krone T, Snippe A, Babé L, McClain S, Ladics GS, et al. Allergenicity prediction of novel and modified proteins: not a mission impossible! development of a random forest allergenicity prediction model. [Regul Toxicol Pharmacol](#). 2019;107: 104422.

Accession number and IDs of Animal allergen proteins				
pdb 4WfV	XP_012630565.1	XP_012630564.1	XP_008706188.1	BAE45865.1
KDR17360.1	XP_020727845.1	XP_019822297.1	XP_008706187.1	BAE45865.1
NP_001035360.1	XP_020727844.1	OAF64053.1	XP_008706186.1	OXA60863.1
KDR14384.1	XP_020725501.1	KXJ13686.1	XP_008702524.1	OXA62391.1
KFM82824.1	XP_020725500.1	XP_012804513.1	CBY93636.1	OXA58983.1
pdb 5WEE	XP_020770229.1	XP_004653898.1	XP_003369246.1	OXA54094.1
pdb 5WEE C	XP_020769597.1	XP_008706190.1	XP_003378212.1	OXA45999.1
pdb 5WEE D	XP_020145028.1	XP_008706189.1	BAD74060.2	OXA44542.1
XP_020730842.1	XP_020730841.1	XP_020730840.1	XP_020724594.1	EFV56543.1
AAK12274.1	OXA50472.1	OXA43322.1	OXA42154.1	XP_020730846.1
OXA64129.1	OXA48685.1	OXA42509.1	OXA41534.1	XP_020730844.1
OXA63906.1	OXA48175.1	OXA40806.1	OXA39940.1	EFV51118.1
OXA60315.1	OXA43669.1	OXA39518.1	XP_020731239.1	
XP_020730843.1	OXA43665.1	OXA37363.1	OXA36818.1	

Accession number and IDs of plant allergen proteins				
BAA06172.1	sp P81295.1	AEE34450.1	OMO72675.1	PNY01836.1
AAA32671.1	sp Q7Y1X1.1	AED91856.1	KVI11789.1	PNY01836.1
CAA55854.1	pdb 3OB4 A	NP_196818.1	AED92204.1	PNX80673.1
sp P18632.3	BAD09208.1	NP_176774.1	AEC06520.1	PNX71771.1
sp E1U332.1	BAD09204.1	OMP03616.1	NP_179254.2	OTF87078.1
NP_567562.2	OMP12302.1	OMP08613.1	OMO92202.1	OMO92200.1
OMO92165.1	pdb 4N42 A	AAD20920.1	CAC05582.1	AES74729.1

ABA98863.1	AAF25553.1	AAO50557.1	AAO41979.1	CAA01911.1
PWA96184.1	PWA62909.1	PWA62858.1	PWA47474.1	PON91843.1
PON51874.1	ANM64155.1	NP_001326202.1	JAV45574.1	ONM56233.1
AQK61269.1	AQK57370.1	AQK57369.1	ONM29876.1	AQK39769.1
ONM07031.1	EOY28217.1	EOY17706.1	EOY13248.1	EOY12438.1
EOY08036.1	EOY06552.1	EOY03811.1	EOY03810.1	EOY03777.1
EOX99716.1	EOX97748.1	EOX95162.1	EOX95161.1	EOX95160.1
EOX95158.1	EOX95156.1	EOX95155.1	KVI12392.1	JAG89064.1
JAG87204.1	JAG86409.1	KEH42147.1	KEH42145.1	KEH42142.1
KEH42137.1	KEH42124.1	KEH42122.1	KEH42121.1	AES67433.2
AES67417.2	KEH38354.1	AES81924.2	AES88673.1	AES67432.1
AES67429.1	AES67426.1	AES67425.1	AES67418.1	AES67413.1
pdb 2QIM A	pdb 1NLX I	pdb 1NLX H	pdb 1NLX F	pdb 1NLX E
pdb 1NLX D	pdb 1NLX C	pdb 1NLX B	pdb 1NLX A	sp P19656.1
sp P14946.2	sp P10414.2	sp P81370.2	PNY01836.1	