Original Research

ALLERDET: A novel web app for prediction of protein allergenicity

Francisco M. Garcia-Moreno [a,*], Miguel A. Gutiérrez-Naranjo [b]

[a] Department of Software Engineering, Computer Science School, University of Granada, C/ Periodista Daniel Saucedo Aranda, s/n, Granada, 18014, Spain
[b] Department of Computer Sciences and Artificial Intelligence, University of Sevilla, Avda. Reina Mercedes, s/n, Sevilla, 41012, Spain

ARTICLE INFO

MSC:
68-04
92-04

Keywords:
ALLERDET
Allergen detection
Food allergy
Pairwise sequence alignment
FASTA
Restricted Boltzmann Machines

ABSTRACT

Allergic diseases are increasing around the world with unprecedented complexity and severity. One of the reasons is that genetically modified crops produce new potentially allergenic proteins. From this starting point, many researchers have paid attention to the development of tools to predict the allergenicity of new proteins. In this study, a novel approach is introduced for the prediction of food allergens based on Artificial Intelligence techniques: a pairwise sequence alignment with the FASTA program for feature extraction and the use of the Deep Learning technique known as Restricted Boltzmann Machines in combination with the Decision Tree method for the prediction process. The developed tool, called ALLERDET (publicly available at http://allerdet.frangam.com), overcomes the state-of-the-art methods. The performance of our method is: 98.46% sensitivity, 94.37% specificity and 97.26% accuracy), on a data set built from several publicly available sources.

1. Introduction

Allergic diseases are increasing worldwide with unprecedented complexity and severity. Allergic reactions affect more than 30% of the world's population. In particular, food allergies affect 6% of the pediatric population and 4% of adults [1–3]. Parallel to this increase, the number of new proteins has grown rapidly in recent years. They are used in therapeutics, food, household products, and pharmaceuticals. These modified proteins are a source of potential allergenicity for humans, and the development of accurate methods capable of detecting true allergens is crucial to ensure safety. The usual way to test for potential risk is by comparing the potential allergen with a database of labeled proteins. Predicting allergenicity is a hard task, since similarity does not always take into account cross-reactivity in protein folds [4].

The current joint recommendation by the World Health Organization (WHO) and the Food and Agriculture Organization (FAO) to test allergenicity is based on a Decision Tree schema [5]. This schema compares the similarity of a protein with a set of proteins of known allergenicity, and if they are 35% identical over a linear window of 80 residues, then this protein is declared potentially allergen [6]. In the literature, different Machine Learning techniques can be found to classify new proteins that are not based on Decision Tree schemata. These techniques involve the k-Nearest Neighbor classifier [7], Fourier transform [8], linear / quasi-Gaussian classifier [9], SVM methods [10], allergen-representative peptides [11], global protein descriptors [12], and a combination of several methods known as *hybrid techniques* [13].

Researchers are interested in using tools to predict allergenicity in different domains, such as the discovery of the allergenicity pattern of vaccine and drug development [14–17]. Several applications have been developed to deal with the prediction of allergenicity. For example, AllerCatPro 2.0 [18] is a web server that predicts protein allergenicity with 84% accuracy, based on amino acid sequences and the similarity of the 3D protein structure. Other tools do not exceed 90% accuracy, such as AllergenFP [19], which reaches 88% accuracy and AllerTOP [20] with 94% sensitivity, but none of the accuracy is mentioned. However, existing tools are imprecise in terms of sensitivity, specificity and accuracy. Some of them are high in sensitivity but low in specificity, resulting in inconsistent and unbalanced approaches to accuracy. Except AlgPred 2.0 [21], the current approaches have balanced metrics and all of them are below 90%.

In this paper, a novel Machine Learning approach is presented to discriminate between proteins that are capable or not of causing allergic reactions. Our proposal combines classical Decision Tree methods with a Deep Learning technique known as Restricted Boltzmann Machines (RBM). RBM have been successfully applied to other research areas such as collaborative filtering [22], modeling documents [23], modeling natural images [23], or even modeling human motion [24]. They have also been applied in the health domain, namely, to model electronic medical records (EMRs) [25]. However, to the best of our knowledge, this is the first time that RBM has been applied to the study of allergenicity. The intuition behind the use of RBM is that this deep

[*] Corresponding author.
E-mail address: fmgarmor@ugr.es (F.M. Garcia-Moreno).

https://doi.org/10.1016/j.jbi.2022.104217
Received 12 July 2022; Received in revised form 21 September 2022; Accepted 30 September 2022

Available online 13 October 2022
1532-0464/

learning model is capable of producing a compact representation of the relevant information, which is provided as input to the Decision Tree method. Our model to detect the potential allergenicity of protein sequences was trained and validated using 2000 well-known allergens and 2000 non-allergens, overcoming the state-of-the-art (SOTA) methods, maintaining balanced metrics performances of sensitivity, specificity and accuracy, and is publicly available at http://allerdet.frangam.com.

To resume, the main contributions of this paper can be summarized as follows:

- A new sequence database for training models of supervised learning methods has been constructed by integrating many other publicly available databases. To our knowledge, it is currently the most comprehensive database for this purpose and is publicly available to the scientific community.
- We empirically demonstrate that the use of Restricted Boltzmann Machines, a Deep Learning technique, is a useful tool for preprocessing information before using Decision Tree methods, as they provide a more compact (and more efficient) way of encoding information.
- A new machine learning method based on the combination of RBM and decision tree has been implemented and trained on the newly constructed database. The results obtained outperform the scores of state-of-the-art methods and can be publicly accessed via web.
- Our tool has balanced metrics in terms of sensitivity, specificity and accuracy, all of which are greater than 98%.

2. Materials and methods

2.1. Datasets

The dataset was built by combining data from several publicly available sources: UniProt database [26], AllerHunter Training Dataset [10, 27], AllerTop Training Dataset [20,28], COMPARE [29] and AllergenOnline [30].

On the one hand, the allergens collected from UnitProt [26] were obtained following *Zorzet et al.* [7], allergenic sequences were collected from UniProt database searching for matches with allergen terms (see Appendix. An amount of 2952 sequences were collected from SwissProt records (sequences are manually annotated by experts), and 1125 different allergen sequences were collected from AllerHunter (from its train/test/independent sets) and 2427 from AllerTop, 2463 from COMPARE, 2233 from AllergenOnline and 8060 from AlgPred 2.0. Obviously, since among these sets there are duplications, we avoided these repetitions and obtained a final result of 4670 sequences of allergens in total. Furthermore, for comparing state-of-the-art methods, we apply the validation hold-out set of AlgPred 2.0 (2015 sequences) used in benchmarks [21,18]. Although, we found it had some duplications and the final holdout-set had 763 allergens, which were never seen in training dataset. All of these sequences are in FASTA format [31].

On the other hand, the dataset of non-allergen sequences was obtained from the UniProt database, including the exclusion filters proposed in [7,9]. Two exclusion filters were used. Some technical details can be found in Appendix. These searches reported 1227 non-allergen sequences from SwissProt records, and a total of 3977 sequences from AllerHunter, 2427 from AllerTop and 8060 from AlgPred 2.0 non-allergen. However, although the total amount of sequences resulted in 15,369, we selected the same value as allergens to get balanced the dataset (4670). Again, we used the validation hold-out set of AlgPred 2.0 (763 non-allergens, get classes balanced), avoiding also duplications with our training dataset.

The training dataset is available to the research community for allergens and non-allergens at [32] and [33], respectively.

2.2. Pairwise sequence alignment

Pairwise Sequence Alignment (PSA) is a widely used technique in bioinformatics to extract useful information on structural, functional and evolutionary relationships between two biological sequences, based on the identification of similarity regions between amino acids [34]. In this article, PSA was performed with the FASTA 3.6 tool [31,35]. The parameters used were the BLOSUM50 substitution matrix, gap opening penalty with a value of -12 and the extension gap penalty value set to -2. All sequences were aligned with the allergen training data set. The allergen training set was aligned against itself (ATD) and the non-allergen training set was aligned against the allergen training set (NTD). The rationale for this alignment is to obtain a measure of similarity, since similar sequences in proteins are commonly accepted that imply a similar function and structure [36].

From these alignments, four features returned by the FASTA program were selected to evaluate them: alignment score (Smith-Waterman score), alignment length, identity and similarity percentages over 35% (following FAO/WHO criteria [5]), init1 (the highest scoring alignment without gaps), initn (the score which combines consistent non-overlapping runs without gaps, z-score and bits scores [31,35]. Furthermore, the best m alignments reported by FASTA were considered. Next, all alignments were put into a feature matrix to feed Machine Learning models, and each row was labeled as an allergen or non-allergen, depending on whether these features correspond to ATD or NTD, respectively (Fig. 1).

2.3. Restricted Boltzmann machines

One of the key points in our study is the use of RBM. Next, we provide a short introduction to this Deep Learning model.

Boltzmann Machines (BM) are bidirectionally connected networks of stochastic processing units. Such devices can be considered as artificial neural network models [37]. They are Machine Learning tools that can learn an unknown probability distribution from samples of this distribution [38]. RBM are BM in which neurons are placed on the nodes of a bipartite graph. These neurons are distributed in two layers. Two nodes from the same layer are not connected, and each of the two nodes from different layers is connected by a symmetric edge [39]. In this way, an RBM can be seen as an artificial neural network with a single-layer architecture for unsupervised feature learning.

One of the layers is called the visible variable layer v_i, $i \in \{1, \ldots, m\}$, corresponding to the input data, and the second one is called the hidden variable layer h_j, $j \in \{1, \ldots, g\}$, corresponding to feature detectors. Each edge between neurons in the visible and hidden layers has a weight associated with it. Let $\mathbf{W} = (W_{ij})_{m \times g}$ be a matrix representing the weight parameter settings, where W_{ij} represents the connection between the variables v_i and h_j.

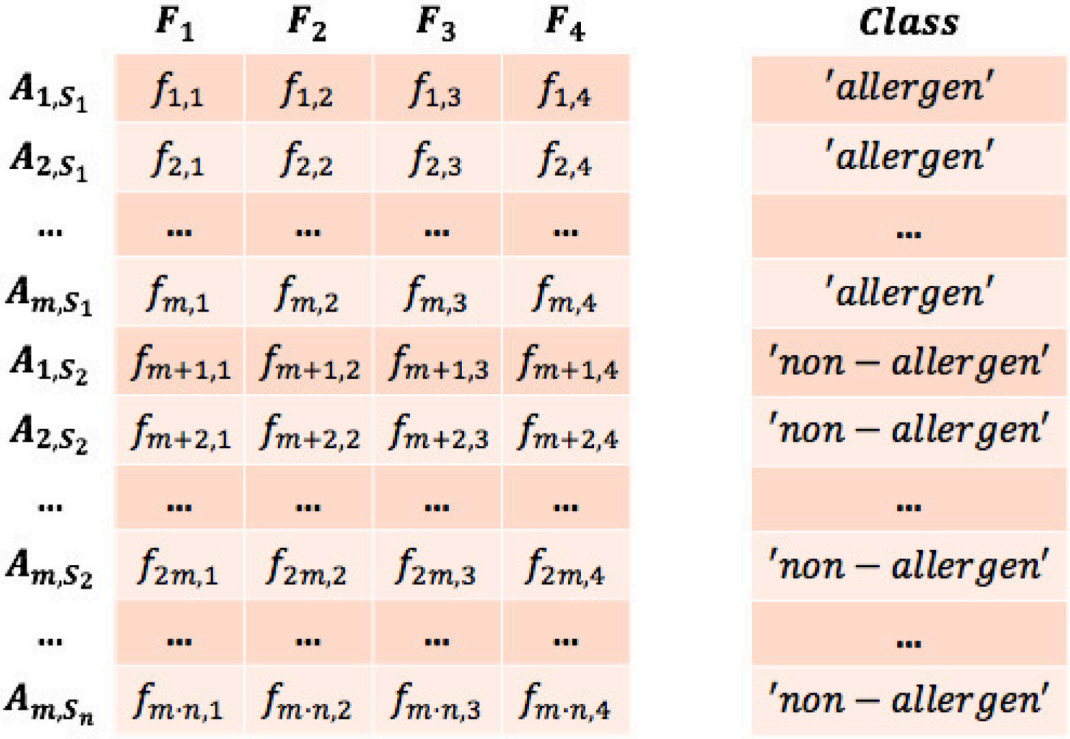
Let $\mathbf{a} = (a_1, \ldots, a_m)$ and $\mathbf{b} = (b_1, \ldots, b_g)$ be the bias vectors, where a_i and b_j are the biases associated with variables v_i and h_j, respectively. Let $\mathbf{v} = (v_1, \ldots, v_m)$ and $\mathbf{h} = (h_1, \ldots, h_g)$ be two vectors representing the state of variables in visible and hidden layers and denote by $(\mathbf{v}, \mathbf{h})$ the configuration of the entire RBM. Given this configuration, the energy function of the model is defined as follows.

$$E(v, h; \theta) = \mathbf{a}^T \mathbf{v} + \mathbf{b}^T \mathbf{h} + v^T \mathbf{W} \mathbf{h} \tag{1}$$

where $\theta = (\mathbf{a}, \mathbf{b}; \mathbf{W})$ is the setting of the model parameters. Since there are no links between hidden variables, the marginal distribution of visible variables can easily be computed.

$$p(v) = \frac{1}{Z} \sum_h p(v, h) = \frac{1}{Z} \sum_h e^{-E(v,h;\theta)} \tag{2}$$

where Z is the normalization constant. The learning process in RBM tries to maximize the product of probabilities in a training set V [40]. The usual training algorithm is the contrastive divergence algorithm [31], which tries to maximize the probability of visible



	F_1	F_2	F_3	F_4	$Class$
A_{1,S_1}	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$'allergen'$
A_{2,S_1}	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	$'allergen'$
...	...	...	...	...	...
A_{m,S_1}	$f_{m,1}$	$f_{m,2}$	$f_{m,3}$	$f_{m,4}$	$'allergen'$
A_{1,S_2}	$f_{m+1,1}$	$f_{m+1,2}$	$f_{m+1,3}$	$f_{m+1,4}$	$'non-allergen'$
A_{2,S_2}	$f_{m+2,1}$	$f_{m+2,2}$	$f_{m+2,3}$	$f_{m+2,4}$	$'non-allergen'$
...	...	...	...	...	...
A_{m,S_2}	$f_{2m,1}$	$f_{2m,2}$	$f_{2m,3}$	$f_{2m,4}$	$'non-allergen'$
...	...	...	...	...	...
A_{m,S_n}	$f_{m\cdot n,1}$	$f_{m\cdot n,2}$	$f_{m\cdot n,3}$	$f_{m\cdot n,4}$	$'non-allergen'$

Fig. 1. Features matrix and classes vector.

variables using. Its objective is to approximate the probability density function of the data set using an unsupervised approach, that is, without any kind of labeled information. Once the parameters of the model have been fixed after the learning process, the model is ready to obtain the likely samples. The method is used inside a gradient descent technique, computing weight update and performing the Gibbs sampling. Although the most common RBM has binary-valued hidden/visible units [41], in this paper, we use real-valued units.

2.4. Evaluation metrics

The performance of the classifiers is evaluated using the following measures:

- Accuracy (ACC): total number of true classifications divided by the size of the test set.
- Sensitivity (SE): the ability to identify allergen proteins correctly.
- Specificity (SP): the ability to identify non-allergen proteins correctly.

The cross-validation method [42] was used to evaluate the tool.

2.5. Stratified k-fold cross-validation

The cross-validation method [42] was used to evaluate the tool in order to ensure the same proportion of two classes in every fold. It is also well known. The database D is divided into mutually exclusive subsets of random samples, also known as folds $(D_1, D_2, \dots, D_k)$. Approximately, such a fold has equal size. The classifier is then trained in k iterations, each of them using the set $(D\backslash D_i)$ for training and D_i for testing $i \in \{1, \dots, k\}$. Cross-validation accuracy is the total number of true predictions divided by the total instances. The cross-validation estimate k is calculated as follows:

$$acc_{cv} = \frac{1}{n} \sum_{(v_i, y_i) \in D} \delta(I(D\backslash D_{(i),v_i}), y_i) \tag{3}$$

where $(I(D\backslash D_{(i)}, v_i), y_i)$ corresponds to the label assigned by I to an instance v_i with the dataset $D\backslash D_{(i)}$; y_i is the classification of the instance ; n is the size of D; and $\delta(i, j)$ is the Kronecker delta. In this paper, a stratified k-fold validation was used. Approximately, this means that the folds contain the same proportion of classes as in the full dataset.

2.6. Webserver implementation

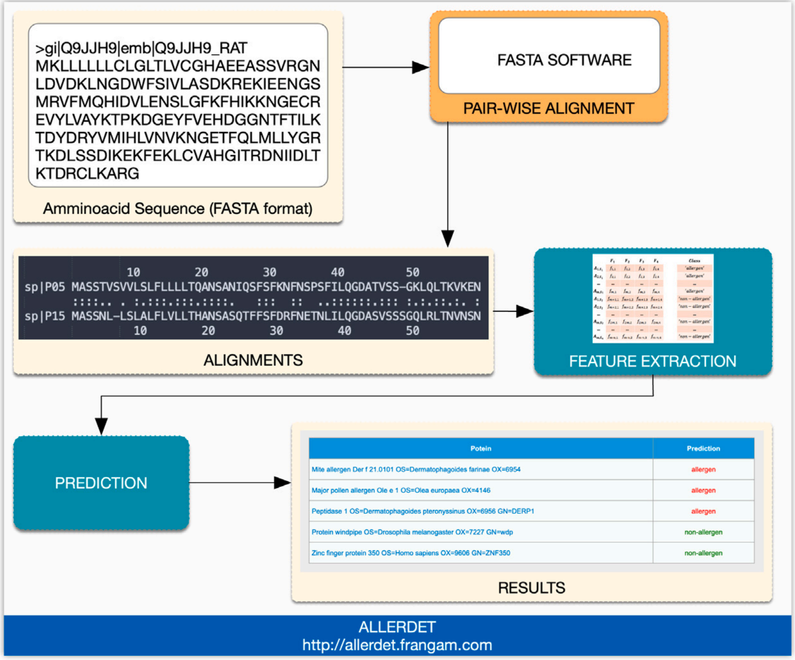
The software architecture of ALLERDET is presented in Fig. 2. The input of this application consists of one or several protein sequences in FASTA format (Fig. 3). Then, when the user click on “Predict” button, the back-end launches FASTA program in order to perform the pair-wise alignment between the input sequences and our curate list of allergens. Following, we get a file with the alignments and our method performs a feature extraction process and, then the final allergen prediction.

Finally, the output is an interactive table (Fig. 4) with the results of allergenicity prediction, which presents links for each protein to view more details in the corresponding database.

3. Results

Fig. 5 (left) shows the pairwise sequence alignments of the nonallergenic sequences against the allergens dataset (NTD). The result of aligning the allergenic sequences against themselves (ATD) is presented in Fig. 5 (right). On the one hand, the nonallergen alignment presents a low alignment score and a relatively short alignment length, in general. On the other hand, an important point is that allergen alignment shows a high alignment score on average.

Three-fold cross-validation with stratification was applied to validate the quality of each model. In this way, a search was performed for the best classification accuracy, considering several permutations of different values of the best m alignments and the four different alignment features selected.

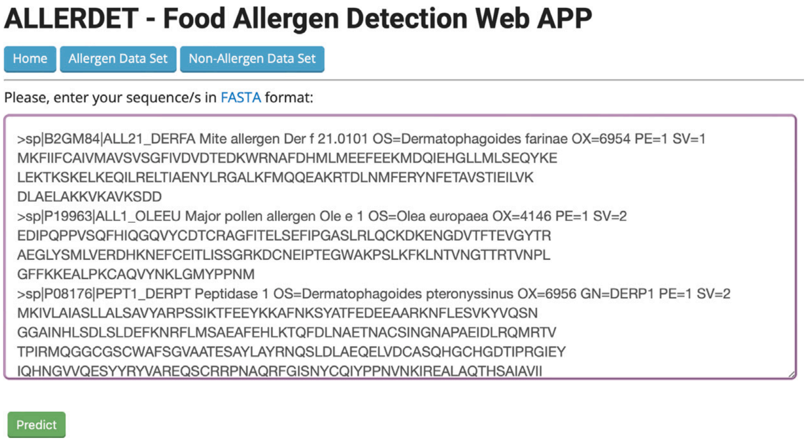


```
>gi|Q9JJH9|emb|Q9JJH9_RAT
MKLLLLLLLCLGLTLVCGHAEEASSVRGN
LDVDKLNGDWFSIVLASDKREKIEENGS
MRVFMQHIDVLENSLGFKFHIKKNGECR
EVYLVAYKTPKDGEYFVEHDGGNTFTILK
TDYDRYVMIHLVNVKNGETFQLMLLYGR
TKDLSSDIKEKFEKLCVAHGITRDNIIDLT
KTDRCLKARG
```

```
                 10        20        30        40         50
sp|P05 MASSTVSVVLSLFLLLLTQANSANIQSFSFKNFNSPSFILQGDATVSS-GKLQLTKVKEN
       ::::.. . :.:::.:::.::::.    :::  ::  ..::::::.::: :.:.::.:. :
sp|P15 MASSNL-LSLALFLVLLTHANSASQTFFSFDRFNETNLILQGDASVSSSGQLRLTNVNSN
                  10        20        30        40        50
```

Potein	Prediction
Mite allergen Der f 21.0101 OS=Dermatophagoides farinae OX=6954	allergen
Major pollen allergen Ole e 1 OS=Olea europaea OX=4146	allergen
Peptidase 1 OS=Dermatophagoides pteronyssinus OX=6956 GN=DERP1	allergen
Protein windpipe OS=Drosophila melanogaster OX=7227 GN=wdp	non-allergen
Zinc finger protein 350 OS=Homo sapiens OX=9606 GN=ZNF350	non-allergen

Fig. 2. ALLERDET software architecture.

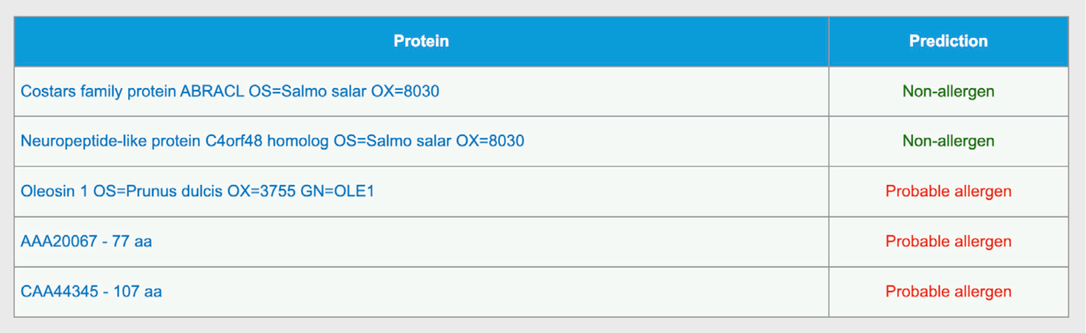


ALLERDET - Food Allergen Detection Web APP

Please, enter your sequence/s in FASTA format:

```
>sp|B2GM84|ALL21_DERFA Mite allergen Der f 21.0101 OS=Dermatophagoides farinae OX=6954 PE=1 SV=1
MKFIIFCAIVMAVSVSGFIVDVDTEDKWRNAFDHMLMEEFEEKMDQIEHGLLMLSEQYKE
LEKTKSKELKEQILRELTIAENYLRGALKFMQQEAKRTDLNMFERYNFETAVSTIEILVK
DLAELAKKVKAVKSDD
>sp|P19963|ALL1_OLEEU Major pollen allergen Ole e 1 OS=Olea europaea OX=4146 PE=1 SV=2
EDIPQPPVSQFHIQGQVYCDTCRAGFITELSEFIPGASLRLQCKDKENGDVTFTEVGYTR
AEGLYSMLVERDHKNEFCEITLISSGRKDCNEIPTEGWAKPSLKFKLNTVNGTTRTVNPL
GFFKKEALPKCAQVYNKLGMYPPNM
>sp|P08176|PEPT1_DERPT Peptidase 1 OS=Dermatophagoides pteronyssinus OX=6956 GN=DERP1 PE=1 SV=2
MKIVLAIASLLALSAVYARPSSIKTFEEYKKAFNKSYATFEDEEAARKNFLESVKYVQSN
GGAINHLSDLSLDEFKNRFLMSAEAFEHLKTQFDLNAETNACSINGNAPAEIDLRQMRTV
TPIRMQGGCGSCWAFSGVAATESAYLAYRNQSLDLAEQELVDCASQHGCHGDTIPRGIEY
IQHNGVVQESYYRYVAREQSCRRPNAQRFGISNYCQIYPPNVNKIREALAQTHSAIAVII
```

Fig. 3. Interface of ALLERDET for submitting one or several protein sequences in FASTA format.



Protein	Prediction
Costars family protein ABRACL OS=Salmo salar OX=8030	Non-allergen
Neuropeptide-like protein C4orf48 homolog OS=Salmo salar OX=8030	Non-allergen
Oleosin 1 OS=Prunus dulcis OX=3755 GN=OLE1	Probable allergen
AAA20067 - 77 aa	Probable allergen
CAA44345 - 107 aa	Probable allergen

Fig. 4. Interface of ALLERDET showing the results of allergenicity detection in an interactive table, with links to protein databases for each entry.

In such a way, training of an RBM was performed, and then the outputs were used as inputs of the Decision Tree model, which we called RBM + DT and deployed in ALLERDET. Therefore, the best performing method was RBM + DT with relatively higher performance

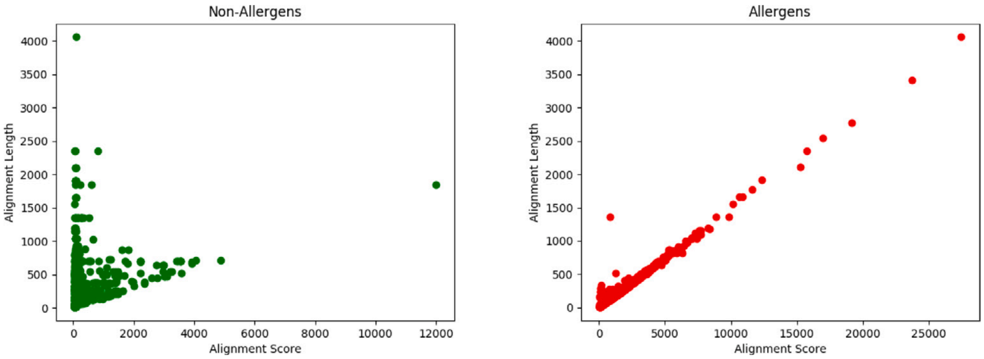


Fig. 5. Non-allergen alignment (left) and allergen alignment (right).

Table 1
Comparison of ALLERDET with existing tools as reported in the literature.

Tool	Sensitivity	Specificity	Accuracy	Alive Web
ALLERDET (3-fold)	99.62%	99.74%	99.48%	
ALLERDET (AlgPred 2.0 test)	98.46%	94.37%	97.26%	Yes
AlgPred 2.0	93.1%	95.36%	94.23%	Yes
AlgPred (2003)	88.87%	81.86%	85.02%	Yes
AllerCatPro 2.0 (AlgPred 2.0 test)	93.2%	98.8%	96.0%	Yes
AllerCatPro 1.7 (AlgPred 2.0 test)	91.1%	94.8%	93.0%	Yes
AllerHunter	83.7%	96.4%	95.3%	No
AllerTOP v.2 (kNN, k = 1)	86.7%	90.7%	88.7%	Yes
AllerTOP v.1	87.6%	78%	82.8%	No
Zorzet et al. (kNN, k = 9)	81%	98%	89.5%	–
AllergenFP (SVM)	86.8%	89.1%	87.9%	
FAO/WHO	97.8%	27.8%	20.9%	–

correctly recognizes 99.66% allergens, 98.89% of non-allergens (accuracy=99.3%, $m = 1$ best alignments considered and all four alignment features extracted).

In addition, using the independent hold-out test dataset (AlgPred 2.0 validation test of 2015 allergens), our model was evaluated resulting in a recognition of 98.46% allergens, 94.37% of non-allergens. Furthermore, we also tested successfully some known allergens from different plant species, such as oleosins from Arachis hypogea or Prunus dulcis [43].

The performance of ALLERDET was compared with five tools for allergenicity prediction as shown in Table 1. All of the results have been performed on a computer with an Intel Core i7 processor with 2.4 GHz and 16 GB of RAM.

With respect to sensitivity (allergen detection rate), our model reaches 98.46%. It is the best score, followed by AlgPred 2.0, which reaches 93.1%.

Finally, the accuracy column of Table 1 shows a comparison of the accuracy of several methods. In this case, our ALLERDET method clearly obtains the best score.

4. Ablation study

The proposed classification method is a combination of RBM with the Decision Tree method. The use of RBM in order to obtain a compact representation of the relevant information is one of the main novelty of this paper, and the use of Decision Tree methods is justified because it is the current technology currently used by WHO and FAO.

Nevertheless, in order to complete our study, we have also considered other machine learning methods to compare the results. In this way, the application of RBM was performed with a pipeline based on the Scikit-Learn [44] library: first, a pre-training of an RBM model was made, second, the output of RBM model was used as the input of several classification tools. To achieve all that, a variety of scripts were written in Python version 3 [45]. All of them were included in a webserver application called ALLERDET and were developed using the Flask library for web development in Python.

Table 2
Evaluation of the performance of four methods before RBM.

Method	Sensibility	Specificity	Accuracy
Decision tree	99.53%	98.59%	99.06%
Naïve Bayes	99.36%	98.33%	98.80%
kNN (k = 3)	99.22%	97.37%	98.34%
Multilayer perceptron	98.21%	98.39%	98.30%

Firstly, in order to get the best classifier, we performed different Machine Learning methods without RBM pre-training: Decision Tree, k-Nearest Neighbors (kNN), Naive Bayes (NB), and Multilayer Perceptron (MLP). They were evaluated to obtain one of them with the best accuracy performance. Simultaneously, the GridSearch method from the Scikit-Learn library was used to establish the best parameter setup of each model. Table 1 shows the percentages of sensitivity, specificity, and accuracy for each model obtained for the best setup. Based on the results presented in Table 1, the four models considered achieve an accuracy greater than 98% and the best model evaluated was the Decision Tree with an accuracy of 99.06%.

Second, taking this result into account, we then applied an RBM model as the input of Decision Tree. The result was an increase in the final accuracy (Table 2): 99.3% accuracy, 99.66% sensitivity and 98.89% specificity. In addition, the best m alignments resulted in $m = 1$ and the relevant features of the model were: alignment score (Smith-Waterman score), alignment length, identity, and similarity percentages over 35% (following FAO/WHO criteria [5]).

5. Conclusions

In this work, we have presented a novel approach to food allergen prediction based on pairwise sequence alignments performed with FASTA, which uses a combination of two well-known Machine Learning models, namely RBM and Decision Trees. Our proposed method is developed as a webserver tool called ALLERDET and is publicly available. ALLERDET performance has the highest performance (98.46% sensitivity, 94.37% specificity and 97.26% accuracy) compared to the state-of-the-art tools/methods evaluated. Our tool can be a useful tool for researchers and can contribute to improve the prediction of food allergens.

CRediT authorship contribution statement

Francisco M. Garcia-Moreno: Investigation, Conceptualization, Methodology, Software, Data curation, Visualization, Validation, Writing – original draft, Editing. **Miguel A. Gutiérrez-Naranjo:** Supervision, Methodology, Validation, Writing – reviewing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

To build allergen dataset from UnitProt, all searches include the following allergen search:

- (keyword:KW-0020) OR allergen OR allergome OR allergy OR atopy OR atopic OR allergenic OR allergens OR allergies OR allergen*

The exclusion filters used to build a set of nonallergen sequences were the following:

All searches include these two filters of exclusion:

- NOT: sequence length between 1 and 50.
- NOT all: the previous allergen search (to exclude allergenic sequences).

First, to collect nonallergenic vegetable proteins, the following search criteria were added:

- Organism: Daucuscarota, Lycopersicon, Malus, Prunus, Spinacia oleracea.
- NOT all (text): bet v 1, chitinase, germin, Kunitz, legumin, lipid transfer protein, lipid-transfer protein, papain-like cysteine protease, profilin, thau-matin OR vicilin. They are widespread in food allergies.

Second, to add nonallergenic cow milk proteins, the following search criteria are performed:

- Organism: Bos Taurus.
- AND all (text): casein, lactalbumin, lactoferrin, lactoperoxidase, milk, proteose–peptone OR xanthine dehydrogenase.

Third, the following search criteria were added to get nonallergenic chicken egg proteins:

- Organism: Gallus gallus.
- AND all (text): egg.

Lastly, for nonallergenic salmon proteins, this search criterion was introduced:

- Organism: Salmo salar.

References

[1] A. Hjern, Chapter 5.8: Major public health problems — allergic disorders, Scand. J. Public Health 34 (67_suppl) (2006) 125–131, http://dx.doi.org/10.1080/14034950600677139.

[2] L.M. Taussig, A.L. Wright, C.J. Holberg, M. Halonen, W.J. Morgan, F.D. Martinez, Tucson children's respiratory study: 1980 to present, J. Allergy Clin. Immunol. 111 (4) (2003) 661–675, http://dx.doi.org/10.1067/mai.2003.162.

[3] R. Gupta, A. Sheikh, D.P. Strachan, H.R. Anderson, Time trends in allergic disorders in the UK, Thorax 62 (1) (2007) 91–96, http://dx.doi.org/10.1136/thx.2004.038844.

[4] R.C. Aalberse, Structural biology of allergens, J. Allergy Clin. Immunol. 106 (2) (2000) 228–238, http://dx.doi.org/10.1067/mai.2000.108434.

[5] FAO/WHO, Codex Principles and Guidelines on Foods Derived from Biotechnology, Tech. rep., FAO/WHO, 2003.

[6] FAO/WHO, Evaluation of Allergenicity of Genetically Modified Foods, Biotechnology, Tech. rep., FAO/WHO, 2001.

[7] A. Zorzet, M. Gustafsson, U. Hammerling, Prediction of food protein allergenicity: A bioinformatic learning systems approach, In Silico Biol. 2 (4) (2002) 525–534.

[8] K.B. Li, P. Issac, A. Krishnan, Predicting allergenic proteins using wavelet transform, Bioinformatics 20 (16) (2004) 2572–2578, http://dx.doi.org/10.1093/bioinformatics/bth286.

[9] D. Soeria-atmadja, Statistical Evaluation of Local Alignment Features for Prediction of Protein Allergenicity using Supervised Classification Algorithms (Ph.D. thesis), Uppsala University School of Engineering, 2003.

[10] H.C. Muh, J.C. Tong, M.T. Tammi, AllerHunter: A SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins, PLoS One 4 (6) (2009) 2–6, http://dx.doi.org/10.1371/journal.pone.0005861.

[11] Å.K. Björklund, D. Soeria-Atmadja, A. Zorzet, U. Hammerling, M.G. Gustafsson, Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins, Bioinform. Orig. Pap. 21 (1) (2005) 39–5010, http://dx.doi.org/10.1093/bioinformatics/bth477.

[12] J. Cui, L.Y. Han, H. Li, C.Y. Ung, Z.Q. Tang, C.J. Zheng, Z.W. Cao, Y.Z. Chen, Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties, Mol. Immunol. 44 (4) (2007) 514–520, http://dx.doi.org/10.1016/j.molimm.2006.02.010.

[13] S. Saha, G.P.S. Raghava, AlgPred: Prediction of allergenic proteins and mapping of IgE epitopes, Nucleic Acids Res. 34 (WEB. SERV. ISS.) (2006) http://dx.doi.org/10.1093/nar/gkl343.

[14] J. Dey, S.R. Mahapatra, S. Patnaik, S. Lata, G.S. Kushwaha, R.K. Panda, N. Misra, M. Suar, Molecular characterization and designing of a novel multiepitope vaccine construct against *Pseudomonas aeruginosa*, Int. J. Pept. Res. Ther. 28 (2) (2022) 49, http://dx.doi.org/10.1007/s10989-021-10356-z, URL https://link.springer.com/10.1007/s10989-021-10356-z.

[15] A. Banik, S. Sinha, S.R. Ahmed, M.M.H. Chowdhury, S. Mukta, N. Ahmed, N.A. Rani, Immunoinformatics approach for designing a universal multiepitope vaccine against chandipura virus, Microb. Pathog. 162 (2022) 105358, http://dx.doi.org/10.1016/j.micpath.2021.105358, URL https://linkinghub.elsevier.com/retrieve/pii/S088240102100632X.

[16] H. Darsaraei, S. Ghovvati, In silico methods for secretory production of a fungal hydrophobin (HYPAI) in yeast to serve as a promising target for drug delivery, IInt. J. Pept. Res. Ther. 28 (1) (2022) 23, http://dx.doi.org/10.1007/s10989-021-10327-4, URL https://link.springer.com/10.1007/s10989-021-10327-4.

[17] A. Raza, M. Asif Rasheed, S. Raza, M. Tariq Navid, A. Afzal, F. Jamil, Prediction and analysis of multi epitope based vaccine against Newcastle disease virus based on haemagglutinin neuraminidase protein, Saudi J. Biol. Sci. (2022) http://dx.doi.org/10.1016/j.sjbs.2022.01.036, URL https://linkinghub.elsevier.com/retrieve/pii/S1319562X22000365.

[18] M.N. Nguyen, N.L. Krutz, V. Limviphuvadh, A.L. Lopata, G.F. Gerberick, S. Maurer-Stroh, AllerCatPro 2.0: A web server for predicting protein allergenicity potential, Nucleic Acids Res. 50 (W1) (2022) W36–W43, http://dx.doi.org/10.1093/nar/gkac446, arXiv:https://academic.oup.com/nar/article-pdf/50/W1/W36/44378679/gkac446.pdf.

[19] I. Dimitrov, L. Naneva, I. Doytchinova, I. Bangov, AllergenFP: Allergenicity prediction by descriptor fingerprints, Bioinformatics 30 (6) (2014) 846–851, http://dx.doi.org/10.1093/bioinformatics/btt619, URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt619.

[20] I. Dimitrov, I. Bangov, D.R. Flower, I. Doytchinova, AllerTOP v.2-a server for in silico prediction of allergens, J. Mol. Model. 20 (6) (2014) 2278, http://dx.doi.org/10.1007/s00894-014-2278-5, URL http://www.ncbi.nlm.nih.gov/pubmed/24878803 https://www.researchgate.net/publication/275887095_AllerTOP_v2_-_A_server_for_in_silico_prediction_of_allergens.

[21] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, G.P.S. Raghava, AlgPred 2.0: An improved method for predicting allergenic proteins and mapping of IgE epitopes, Brief. Bioinform. 22 (4) (2020) http://dx.doi.org/10.1093/bib/bbaa294, arXiv:https://academic.oup.com/bib/article-pdf/22/4/bbaa294/39140559/bbaa294.pdf, bbaa294.

[22] R. Salakhutdinov, G. Hinton, Deep Boltzmann machines, in: D. van Dyk, M. Welling (Eds.), Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 5, PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009, pp. 448–455, URL https://proceedings.mlr.press/v5/salakhutdinov09a.html.

[23] N. Srivastava, R. Salakhutdinov, G.E. Hinton, Modeling documents with deep Boltzmann machines, 2013, URL http://arxiv.org/abs/1309.6865.

[24] G.W. Taylor, G.E. Hinton, S.T. Roweis, Modeling human motion using binary latent variables, in: P.B. Schölkopf, J.C. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems 19, MIT Press, 2007, pp. 1345–1352.

[25] T. Tran, T.D. Nguyen, D. Phung, S. Venkatesh, Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM), J. Biomed. Inform. 54 (2015) 96–105, http://dx.doi.org/10.1016/j.jbi.2015.01.012.

[26] UniProt Consortium, UniProtKB, protein knowledgebase, 2002, URL http://www.uniprot.org/uniprot/.

[27] H.C. Muh, J.C. Tong, M.T. Tammi, AllerHunter data sets, PLoS One 4 (6) (2009) 2–6, http://dx.doi.org/10.1371/journal.pone.0005861, URL http://tiger.dbs.nus.edu.sg/AllerHunter/.

[28] I. Dimitrov, I. Bangov, D.R. Flower, I. Doytchinova, AllerTop v.2 data sets, 2014, URL http://www.ddg-pharmfac.net/AllerTOP/data.html.

[29] HESI Global, COMPARE: Comprehensive protein allerger resource, 2022, URL https://comparedatabase.org/.

[30] University of Nebraska-Lincoln, AllergenOnline, 2022, URL http://www.allergenonline.org/.

[31] W.R. Pearson, D.J. Lipmant, Improved tools for biological sequence comparison, Biochemistry 85 (1988) 2444–2448, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013/pdf/pnas00260-0036.pdf.

[32] F.M. Garcia-Moreno, ALLERDET allegern train dataset, 2022, URL http://allerdet.frangam.com/train-allergens.

[33] F.M. Garcia-Moreno, ALLERDET non-allergen train dataset, 2022, URL http://allerdet.frangam.com/train-non-allergens.

[34] V.S. Mathura, P. Kangueane, Bioinformatics: A Concept-Based Introduction, Springer US, 2009, http://dx.doi.org/10.1007/978-0-387-84870-9.

[35] W.R. Pearson, D.J. Lipman, UVA FASTA downloads, 2017, URL http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml.

[36] C.A. Wilson, J. Kreychman, M. Gerstein, Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores, J. Mol. Biol. 297 (2000) 233–249, http://dx.doi.org/10.1006/jmbi.2000.3550.

[37] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, A learning algorithm for Boltzmann machines, Cogn. Sci. 9 (1985) 147–169, http://dx.doi.org/10.1207/s15516709cog0901_7, URL https://pdfs.semanticscholar.org/2e3e/09e48a7a62dc30efd8ef7fc4665a53e84d7a.pdf.

[38] A. Fischer, C. Igel, Training restricted Boltzmann machines: An introduction, Pattern Recognit. 47 (2013) 25–39, http://dx.doi.org/10.1016/j.patcog.2013.05.025.

[39] P. Smolensky, Information processing in dynamical systems: Foundations of harmony theory, in: D.E. Rumelhart, J.L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volumen 1: Foundations, MIT Press, 1986, pp. 194–281 (Chapter 6).

[40] J.-W. Perng, I.-H. Kao, Y.-W. Chen, Y.-H. Lai, C.-M. Su, S.-C. Hung, M.S. Lee, C.-T. Kung, Analysis of the 72-h mortality of emergency room septic patients based on a deep belief network, IEEE Access 6 (2018) 76820–76830, http://dx.doi.org/10.1109/ACCESS.2018.2884509, URL https://ieeexplore.ieee.org/document/8558682/.

[41] S. Jian, J. Jiang, K. Lu, Y. Zhang, SEU-tolerant restricted Boltzmann machine learning on DSP-based fault detection, in: 2014 12th International Conference on Signal Processing, ICSP, 2014, pp. 1503–1506, http://dx.doi.org/10.1109/ICOSP.2014.7015250.

[42] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI'95: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995, pp. 1137–1143.

[43] E. Majsiak, M. Choina, K. Miśkiewicz, Z. Doniec, R. Kurzawa, Oleosins: A short allergy review, in: M. Pokorski (Ed.), Medical Research and Innovation, Springer International Publishing, Cham, 2021, pp. 51–55, http://dx.doi.org/10.1007/5584_2020_579.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, URL http://scikit-learn.org/stable/.

[45] Python Software Foundation, Numpy, 2001, URL https://pypi.python.org/pypi/numpy/.