

GENOME-WIDE PREDICTION AND EXPRESSION ANALYSIS OF FOOD ALLERGENS

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

in

Biotechnology

By

Bhupender Singh

Registration Number: 11617871

Supervised By

Dr Neeta Raj Sharma (11840)
Department of Biotechnology (Dean)
Lovely Professional University

Co-Supervised by

Dr Atul Kumar Upadhyay
Department of Biotechnology (Assistant Professor)
Thapar Institute of Engineering and Technology

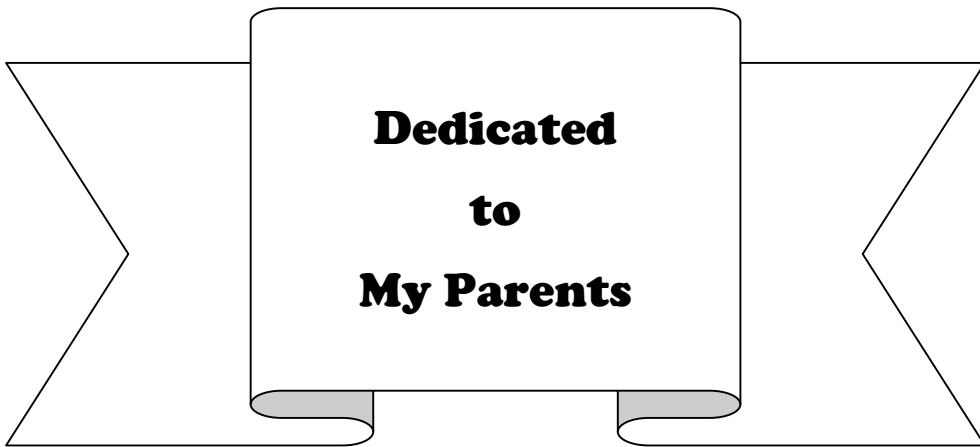
Dr Vijayalakshmi Ahanathapillai
Department of Biomedical Engineering (Lecturer)
Birmingham City University



LOVELY PROFESSIONAL UNIVERSITY

PUNJAB

2022



Declaration

I hereby proclaim that this thesis delineates my own work which has been carried out following enrolment for the degree of PhD at Lovely Professional University, Punjab, India, and no part of this work has been produced earlier for the award of any degree, diploma or other qualification at this or any other institutions.

The objectives of the study were accomplished under the superintendence of Prof. Neeta Raj Sharma, School of Bioengineering & Biosciences, Lovely Professional University, India, as supervisor and Dr. Vijayalakshmi Ahanathapillai, School of Health Sciences, Birmingham City University, United Kingdom, and Dr. Atul Kumar Upadhyay, Department of Biotechnology, Thapar Institute of Engineering and Technology, India as co-supervisors respectively.



Bhupender Singh

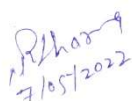
School of Bioengineering & Biosciences,

Lovely Professional University,


Punjab, India.

Certificate

We hereby substantiate that the thesis entitled “GENOME-WIDE PREDICTION AND EXPRESSION ANALYSIS OF FOOD ALLERGENS”, submitted for the degree of PhD by Mr Bhupender Singh is the proof of research work carried out by him under our supervision, and no part of this work has been previously submitted for the award of any degree/certificate from this or any other higher institutions.



Neeta Raj Sharma
7/10/2022



Prof Neeta Raj Sharma
Supervisor
School of Bioengineering
& Biosciences
Lovely Professional University
India

Dr Vijayalakshmi Ahanathapillai
Co-Supervisor
School of Health Sciences
Birmingham City University
United Kingdom

Dr Atul Kumar Upadhyay
Co-Supervisor
Department of Biotechnology
Thapar Institute of Engineering and
Technology
India

Abstract

Food allergy is an exponentially pullulating health concern affecting developed as well as developing countries. Consumption of antigenic food protein leads to sensitisation by production of distinct Immunoglobulin type-E (IgE) antibodies. IgE mounted immune response symptoms generally arises after two hours of allergenic food consumption thereby distressing skin, respiratory and gastrointestinal system causing erythema, pruritus, wheezing, sneezing, nausea, diarrhoea, unconsciousness, and anaphylaxis. Moreover, the individuals experiencing food allergic conditions are more susceptible to acquire respiratory ailments like asthma. The main objective of the study was to dynamically explore the pharmacological space to screen out drug-like pharmacophores against the food allergen profilins and further develop the machine learning based allergenicity assessment model based on exhaustive physicochemical space mining of these allergens. Profilin, a ubiquitously found protein in plants, animals and viruses have been associated with IgE cross-reactivity and are responsible for oral allergy syndrome and pollen food allergy syndrome. The multiple sequence alignment profiles of human versus food allergen profilins revealed a very low level of sequence identity, signifying that the imposed hypersensitive reactions against allergen profilins is by virtue of difference in their amino acid composition make-up. The allergenic profilin protein from apple, pineapple, wheat, and soybean were subjected to homology modelling and molecular dynamic analysis, which revealed that their conserved structural property having equal number of helices, sheets, and loops is responsible for their IgE cross-reactivity and classification as pan allergens. Pharmacophores were screened against these allergen profilins by virtual screening and molecular docking studies unveiled their efficient binding dynamics. Bioavailability studies of these pharmacophores was also in accordance with suitable therapeutics and thus qualifies them to act as lead molecules for drug designing against these allergens. The investigations pertaining to utilisation of physicochemical space for computationally assisted allergenicity assessment is scarce and therefore profilin gene family was extensively explored based on relative amino acid usage and correspondence analysis which unravelled interesting findings. Correspondence analysis based on amino acid usage of allergen and non-allergen

profilins revealed discrete clusters among them, signifying differential patterns of amino acid usage. Correlation analysis revealed that physicochemical features like protein disorder, trypsin digestion and solubility differed significantly among the allergen and non-allergen profilins, thus supporting the observations from correspondence analysis. An in-depth structural analysis revealed that the over-represented amino acids in allergen profilins have a propensity of being exposed on the surface, which may be attributed to their distinct allergenic characteristics. These distinguished physicochemical features along with other parameters constituted the descriptor space and ZeroR, Support Vector Machine (SVM), and Random Forest models which were developed on a manually curated non-redundant dataset and RF model was evaluated as the best classifier. The developed RF model in comparison to existing classifiers outperformed them in terms of their accuracy, Mathew's Correlation coefficient (MCC) value and Receiver Operating Characteristic (ROC) area, suggesting the gravity of employed descriptors extracted by differential amino acid usage analysis in achieving superior accuracy for computationally assisted allergenicity assignment of query protein instances. Further the classification of validated allergens as "predicted allergens" by the developed RF model with higher prediction probability value on an independent dataset of sesame proteome is an indicative of true positive classification capture accuracy of the model. Along with this, the developed model also classified the profilin instances in sesame proteome as potential allergen with significant prediction probability values. The expressional analysis of the sesame transcriptomic data revealed higher expression of validated allergens in sesame. Although the expression profile of profilin was lower but considerable as compare to the validated allergens and thus necessitates to conduct clinical examinations pertaining to allergenicity assessment of profilin protein in sesame. Overall, the study provides scientifically approved inputs to perform pharmacophore validation against food allergen profilins and on the other hand, allows the food industries to modify, validate and utilise the developed RF model for allergenicity assessment in a time and resource-saving manner.

Preface

The present investigation was carried out considering the global epidemiology of food mediated allergic responses and its life-threatening effect on suffering population. The thesis presents the results pertaining to computational analysis of food allergen profilins (apple, pineapple, wheat, and soybean) out of which suitable pharmacophores have been identified. Further, in the second part physicochemical space mining of the food allergen profilins were carried out employing differential amino acid usage analysis for the first time and those finding were translated into development of machine-learning based allergenicity assessment model. The research was carried out at School of Bioengineering & Biosciences, Lovely Professional University, for the award of PhD degree.

I would like to initiate my acknowledgement by translating my pre-eminent gratitude towards the deities of Himachal Pradesh for showering their blessings which helped me to accomplish this project.

This work has not been possible by the invaluable support provided by my parents throughout this project. I am indebted to all they have done as words will fall short to match their level of commitment and acting as a backbone for me to triumph over all the challenges faced. I am obliged to the love and support rendered by my elder brother Mr Pankaj Thakur his wife Mrs Reena Thakur and my dearest nephew Prayan Thakur.

Academically, I would like to express my thanks towards Prof Neeta Raj Sharma, Dean, School of Bioengineering & Biosciences, Lovely Professional University for her peer scientific and technical inputs which helped me to complete this work by acting as the supervisor. I am highly thankful to her adept suggestions and consistent motivation throughout this endeavour. I appreciate Madam's effort to find out time from her busy schedule and review the work progress in order to achieve the targets on time.

I am obliged to have Dr. Vijayalakshmi Ahanathapillai, School of Health Sciences, Birmingham City University, United Kingdom as my co-supervisor for this project. This project would not have been possible without the scheduled meetings which

provided me with deep understanding about the machine learning processes. I am thankful for her encouraging discussions and prompt response whenever I seek out for her scientific advice.

I would like to express my deep sense of gratitude towards Dr Ayan Roy, Research Scientist, Columbia University for his invaluable support to this project. The expertise of Sir and his scientific temperament helped me to think analytically and translate the research observations into knowledgeable findings. I am obliged to the constant support provided by Sir throughout this project. Sir will always be an inspiration for me in this scientific journey.

I would like to thank Dr Atul Kumar Upadhyay, Department of Biotechnology, Thapar Institute of Engineering and Technology, for acting as co-supervisor for this project. I am thankful to the guidance offered by the Sir in this project and encouraging me to try new solutions/approach with reference to this project. I am obliged to have his expertise and vital suggestions throughout this project.

With this, I would like to thank my colleague Sadaf Jan for polishing my skills which helped me to present my work effectively.

Last but not the least, I would like to acknowledge Lovely Professional University for providing the necessary infrastructure to complete this project.



Bhupender Singh

Date: 07-05-2022

Table of Contents

<i>Chapter</i>	<i>Description</i>	<i>Page Number</i>
	Declaration	ii
	Certificate	iii
	Abstract	iv-v
	Preface	vi-vii
	List of Tables	ix
	List of Figures	x-xi
	List of Appendices	xii
1	Introduction	1-7
2	Review of Literature	8-22
3	Research Gap	23
4	Research Objectives	24
5	Materials and Methods	25-36
6	Results and Discussion	37-79
7	Summary and Conclusions	80-82
	Bibliography	83-99
	Index	100-102
	Appendix	A-M
	List of publications	N

List of Tables

<i>Table Number</i>	<i>Title</i>	<i>Page Number</i>
Table 2.1	Various cross-reactive food and aeroallergen groups responsible for IgE-cross-reactivity	12-13
Table 2.2	Various syndromes associated with cross-reactive food and aeroallergens	13
Table 5.1	List of profilins considered in the study along with their Uniprot accessions and templated PDB ID used for homology modelling	27
Table 5.2	Hit screening parameters employed at Pharmit with reference to Lipinski's rule of five and Veber's rule	27-28
Table 5.3	Descriptors considered in the study along with their AAIndex accessions	32-33
Table 6.1	Percent identity profile of the aligned human and food allergen profilins respectively	39
Table 6.2	Top ranked inhibitors screened by the Pharmit web server against ZINC (Purchasable) database for the allergen profilins	44
Table 6.3	Interaction map of the best docked conformations of profilins from apple, pineapple, wheat, and soybean	47-48
Table 6.4	Bioavailability analysis of the pharmacophores ZINC000524729534, ZINC000000041632, ZINC000065529251 and ZINC000257349595 respectively	51-52
Table 6.5	Over/under-represented amino acids based on normalised RAAU of the allergen and non-allergen profilins.	53-54
Table 6.6	Pearson's correlation coefficient values of the physicochemical features of profilins with Axes 1 and 2 of RAAU data	56
Table 6.7	The surface exposed residues found in the allergen profilin in various organisms. These amino acids were over-represented by the allergen profilins on comparison to non-allergen profilins	59-60
Table 6.8	Evaluation of the developed classifiers based upon various parameters	64

List of Figures

<i>Figure Number</i>	<i>Title</i>	<i>Page Number</i>
Figure 1.1	Generalised graphical overview of Ig-E mediated allergic response by the food-antigens	2
Figure 2.1	Classification of various allergen families by Pfam database. The horizontal axis represents various allergen families whereas vertical axis represents number of allergens in that family	14
Figure 5.1	Command line execution window of Trimmomatic tool	34
Figure 5.2	STAR command line execution window	36
Figure 5.3	Schematic representation of proteome wide prediction and RNASeq data analysis for sesame	36
Figure 6.1	Multiple sequence alignment profile of human versus food allergen profilins	38
Figure 6.2	Homology modelled 3D structures of profilin from apple, pineapple wheat and soybean	41
Figure 6.3	Ramachandran plot of the modelled profilins from apple, pineapple, wheat, and soybean	41
Figure 6.4	RMSD plot of the modelled allergen profilins from Mal d 4 (apple), Ana c 1 (pineapple), Tri a 12 (wheat) and Gly m 3 (soybean)	43
Figure 6.5	Docked conformations of the allergen profilins from apple, pineapple, wheat, and soybean along with their non-covalent interactions are referred to (a), (b), (c) and (d) respectively	47
Figure 6.6	Correspondence analysis of the profilin gene family plotted against Axis 1 and Axis 2 of RAAU data	53
Figure 6.7	Multiple sequence alignment profile of the profilin allergens performed by MEGAX	57
Figure 6.8	Surface exposed residues of allergen profilins depicted from their pdb structures by PyMol	60-61

Figure 6.9	Confusion matrix for the ZeroR, LibSVM and Random Forest classifiers	63
Figure 6.10	Clustered column based graphical evaluation of the ZeroR, LibSVM and RF classifiers	65
Figure 6.11	ROC curves for the ZeroR, LibSVM and Random Forest classifiers. The horizontal axis represents FP rate whereas vertical axis denotes TP rate	66
Figure 6.12	Comparative analysis of the developed RF model with existing classifiers	68
Figure 6.13	ARFF structured dataset for the sesame proteome	70
Figure 6.14	RF model predictions log file on the sesame dataset	70
Figure 6.15	Pie chart for the overall prediction summary of sesame proteome. Different colours were used to distinguish the predicted allergen and non-allergen classes	71
Figure 6.16	Number of instances classified as allergens based on prediction probability range for sesame proteome dataset	73
Figure 6.17	FastQC report of the raw sequencing data	75
Figure 6.18	FastQC report of the processed sequencing data	76
Figure 6.19	Screenshot of the output log file of STAR module	77
Figure 6.20	Number of mapped reads per gene for the sesame transcriptomic data	79

List of Appendices

<i>Appendix Number</i>	<i>Title</i>	<i>Page Number</i>
<i>1</i>	List of Uniprot Accessions with identifiers considered in the study for differential amino acid usage analysis	A-H
<i>2</i>	The z-score table of the generated alignment for eight allergen profilins from <i>Hevea brasiliensis</i> (Hev b 8), <i>Artemisia vulgaris</i> (Art v 4), <i>Betula verrucosa</i> (Bet v 2), <i>Cucumis melo</i> (Cuc m 2), <i>Phleum pratense</i> (Phl p 12), <i>Zea mays</i> (Zea m 12), <i>Arachis hypogaea</i> (Ara h 5) and <i>Ambrosia artemisiifolia</i> (Amb a 8) respectively	H-M

Chapter 1

Introduction

Food serves as a medium of vital nutrients for human beings and on the other side it allows the invasion of toxic substances and pathogens inside the human system. The human system, over the years, has efficiently evolved its sensory machineries to differentiate among the useful and unhealthy food constituents. Allergen mediated immunological responses are accountable for maintaining the food quality by imposing immune reactions against the food-antigens. Allergic reactions generally imposed by the immune system for removing the harmful constituents from the body are recognised by rhinitis, sneeze, cough, vomit, and diarrhoea and in worst conditions these reactions become fatal by leading to conditions like edema, hives, and anaphylaxis (Florsheim *et al.*, 2021). Food allergy is termed as an individual's inappropriate immune response against the food-antigens. The list of major food allergens currently recognised by the worldwide organisations are egg, wheat, milk, soy, peanut, tree nuts, fish, shellfish, and sesame. The mounted immune response as classified by National Institute of Allergy and Infectious Diseases (NIAID) can be triggered by either IgE, without IgE or by both. IgE mounted immune response symptoms generally arises after two hours of allergenic food consumption and thereby distressing skin, respiratory and gastrointestinal system causing erythema, pruritus, wheezing, sneezing, nausea, diarrhoea, unconsciousness, and anaphylaxis (Anvari *et al.*, 2019). Sensitisation process takes place by the production of IgE antibodies (specific to the food allergen) by plasma cells which differentiates from B-lymphocytes. These antibodies attach to the cell surface of mast cells and basophils and when individual is exposed to same allergen for second time, the antigenic part of the food allergen attaches to these antibodies which further results in release of histamine and leukotriene signalling molecules (Burks *et al.*, 2012; Tedner *et al.*, 2022). A generalised graphical representation of Ig-E mediated allergic responses are shown in Figure 1.1 below (Tedner *et al.*, 2022).

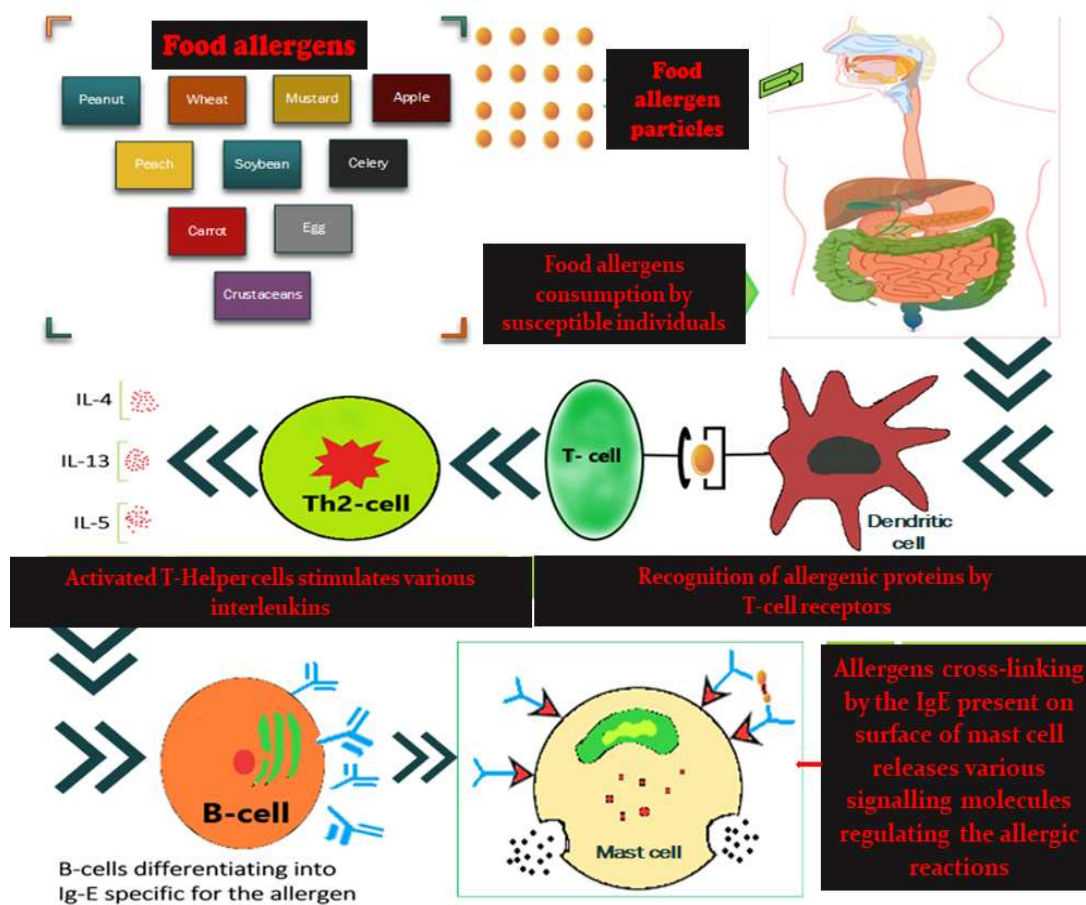


Figure 1.1: Generalised graphical overview of Ig-E mediated allergic response by the food-antigens.

Epidemiology of food allergy in last 20 years have increased in industrialised countries and this rise has been observed in developing countries with upsurged economy (Florsheim *et al.*, 2021). The contemporary factors like alterations in the food-habit, use of processed food, changed gut microbiome and intestine immune system by the influence of less-exposure to parasites are associated with this prevalence (Florsheim *et al.*, 2021). Recently, a study published in the Journal of American Medical Association (JAMA) concluded that more than 26 million adults in the United States are under the influence of food allergy complications (Gupta *et al.*, 2019). In European countries, the incidence of self-reported food-allergy ranged from 6.5% to 24.6%, whereas, the food sensitisation percentage was in the range of 11% to 28.7% (Lyons *et al.*, 2020). A study conducted by the European Union in 2019, 19.1% (n=5677) the Indian population was found to be IgE sensitised (Li *et al.*,

2020). The prevalence data shown by the report in India may not be taken-up as a reference to provide overall food-allergy prevalence as only a small-subset of population from a region is considered in the study. Another major gap in analysing the epidemiology is the implication of non-standard methodology in food-allergy assessment, which makes it difficult to analyse significantly (Li *et al.*, 2020).

According to FAO (Food and Agriculture Organisation), the world population in 2050 will be around 9.1 billion and to feed this huge population food supplies must increase by 70 percent (FAO, 2009). As per the latest reports by (FAO, 2021), around 720 to 811 million people experienced malnutrition in 2020 which is an increase of 118 to 161 million as compared to 2019 malnutrition index. Current practices to meet the food supply demands include introduction of foreign proteins to produce genetically engineered crops with desired traits which certainly requires an expertise safety assessment analysis before its commercialisation (Westerhout *et al.*, 2019). At present in India, there are more than 85 genetically engineered plant species under research and development phase including rice, cotton, tomato, brinjal, maize, wheat, banana, apple, mango, sesame, papaya, and guava (Warrier and Pande, 2016).

Currently no single experiment/assay/technique is sufficient to assess the food allergy in humans, therefore the expert committees have regulated fusion of numerous approaches for potential allergenicity assessment in “weight of evidence” manner (CODEX, 2009; EFSA, 2010; EFSA, 2017). The major pathways suggested in these guidelines are the use of sequence homology approach with known allergen, enzymatic digestion assays and IgE attachment assays and in special cases can be extended to animal studies to recognise the allergenic potential of foreign protein (Remington *et al.*, 2018). However, the applicability of animal and *in-vitro* T-cell assays for allergenicity assessment have not been proven efficient (Ladics *et al.*, 2009; Remington *et al.*, 2018; Westerhout *et al.*, 2019).

Presently, computational biology approaches have been utilised extensively for the allergenicity assessment of foreign proteins (Wang *et al.*, 2021). Numerous web-servers and models have been developed recently which are based on homology search with known allergen, motif-based approach, identification of IgE epitopes, support vector machine, and pseudo amino-acid composition, and artificial neural

networks for allergenicity assessment of proteins (Ivanciuc *et al.*, 2003; Saha and Raghava, 2006; Kumar and Shelokar, 2008; Mohabatkar *et al.*, 2013; Dimitrov *et al.*, 2014). Some studies have also embarked upon the physicochemical features like aliphatic index, grand average of hydropathy (GRAVY), molecular weight, polarity, amino acid composition and hydrophobicity as input features which resulted in improved efficiency of the model (Li and Wang, 2017; Wang *et al.*, 2021). However, systematic exploration of the physicochemical space and utilising this information to develop food allergenicity assessment model is lacking from the contemporary scenario.

The profilin gene family will be taken as a model dataset for the current study. Profilin is found in plants, protozoa, fungus, animals, and viruses (Santos and Van Ree, 2011). It was firstly brought into the picture by Carlson in 1977, as low molecular weight, actin-linked profilamentous protein responsible for the monomeric state of the actin protein (Carlsson *et al.*, 1977). Profilin is among those proteins which are found in abundance inside the cell and with help of immunofluorescence and electron microscopy, it was found to be located on the edges of the cytoplasm and occasionally links with an internal shell of the plasma membrane (Hartwig *et al.*, 1989). The molecular mass of profilin varies between 12 – 15 KDa (Kilo Daltons), having 125-153 amino acids with pI (Isoelectric point) from 4.3 to 9.2 (Santos and Van Ree, 2011).

The count of profilin genes in various organisms is related to intricacy. Lower eukaryotes have one, two or three genes for the profilin. Three profilin genes were found in smaller eukaryotes like *Dictyostelium discoideum* (González-Velasco *et al.*, 2019) and *Caenorhabditis elegans* (Polet *et al.*, 2006). Mammals contain four profilin specific genes annotated as Pfn1 to Pfn4. Pfn1 gene translates into prevalent isoform profilin 1 (Schluter *et al.*, 1997). Pfn2 transcripts two splice forms out of which profilin II translates in neuron cells and profilin IIb expresses in kidney cells of mice (Di Nardo *et al.*, 2000). Translated product of Pfn3 and Pfn4 gene was found in kidney of rat and human testis respectively (Hu *et al.*, 2001; Obermann *et al.*, 2005). Plants contain the highest count of profilin genes, with a maximum number of up to ten. However, it may be noted that few of these genes may be pseudogenes, and apart

from these pseudogenes, the rest of the genes translates into two forms, based on sequence resemblance (Huang *et al.*, 1996) and biochemical features (Kovar *et al.*, 2000). In *Arabidopsis* plant, profilin 1, 2 and 3 isoforms were found to be expressed in every tissue of the plant whereas profilin 4 and 5 were found to be expressed majorly in developed pollen (Kandasamy *et al.*, 2002).

Numerous studies have been carried out previously in identification of food allergens, their classification as pan-allergens, identifying IgE cross-reactivity among food and aeroallergens and computationally assisted prediction of food allergens has all been discussed (Singh *et al.*, 2021). The first identification of profilin as an allergen was observed in the pollen of birch (*Betula verrucosa*) in 1991 and by immunological assays and homologous sequence to existing profilin, this protein was designated as **Bet v 2** (Valenta *et al.*, 1991). Most importantly, profilin was found to have allergic potential in peanuts (*Arachis hypogaea*), soybean (*Glycine max*), tomato (*Solanum lycopersicum*), apple (*Malus domestica*), banana (*Musa acuminata*), orange (*Citrus sinensis*), wheat (*Triticum aestivum*) (Kleber-Janke *et al.*, 1999; Rihs *et al.*, 1999; Willeroider *et al.*, 2003; Ma *et al.*, 2006; Reindl *et al.*, 2002; Lopez-Torrejón *et al.*, 2005; Rihs *et al.*, 1994). The prevalent nature and conserved structure of profilin in plants corresponds to IgE cross-reactivity in profilin among pollen, plant food and latex sources and thus regarded as pan-allergen (Santos and Van Ree, 2011). Profilin sensitisation has been determined in individuals experiencing food allergy. Profilin has been defined as minor allergen in individuals having an allergy to peanut (Kleber-Janke *et al.*, 1999), carrot (*Daucus carota*) (Ballmer-Weber *et al.*, 2001), celery (*Apium graveolens*) (Ballmer-Weber *et al.*, 2001) and pineapple (*Ananas comosus*) (Reindl *et al.*, 2002). In individuals facing allergy to melon (*Cucumis melo*) (López-Torrejón *et al.*, 2005a), orange (*Citrus sinensis*) (Lopez-Torrejón *et al.*, 2005) and soybean (Rihs *et al.*, 1999), the profilin sensitisation was 71%, 78-87% and 69%, respectively and thus regarded as a major allergen. Profilins are also associated with pollen food-allergy syndrome (leading to oropharyngeal pruritus and anaphylaxis), in which the individual sensitised by the profilin from ragweed (*Ambrosia artemisiifolia*), birch (*Betula verrucosa*) and mug wort (*Artemisia vulgaris*) develop cross-reactivity against the profilins from apple, pineapple, carrot, soybean, celery,

banana and peach (Carlson and Coop, 2019). Additionally, profilin in the atopic individuals caused oral epithelial remodelling leading to inflammatory conditions. Clinical study has proven the linkage of profilin with oral epithelial inflammation which subsequently got prolonged through profilin exposure, present in numerous vegetable and fruits (Rosace *et al.*, 2019). Recent studies have shown that profilin isoforms length ranges from 100 to 130 amino acids long with four helices and seven sheets in their secondary structure (Włodarczyk *et al.*, 2022). Study on profilin isoforms has shown sequence dissimilarity in profilin amino acid sequences but they possess parallel tertiary structure as observed in case of tomato, latex, and apple profilins (Włodarczyk *et al.*, 2022). From this we can conclude that despite the sequence difference in profilin allergens from various sources their structural homology contributes to their IgE cross-reactivity patterns and classification as pan allergens.

Allergic reactions mediated by sesame are on the rise globally and presently, its regulatory processes and responsible factors are not clearly distinguished. Sesame has been recognised as allergen by the 32 developed countries including European Union, Canada, Japan, Australia, and New Zealand (Gangur and Acharya, 2021). Recently, USA has also passed a bill (FASTER Act 2021) to label sesame as the major food allergen. India and China are the largest harvesters of sesame but still there are no evidence of reported allergic reactions in these countries which raises two hypotheses stating either there is no such incidence of sesame mediated food-allergy in those countries or certainly there is scarcity of standardised food-allergy tests and thus remains camouflaged (Gangur and Acharya, 2021). The protein allergens in sesame includes albumin, oleosin and vicilin whereas profilin has not been recognised as allergen by the International Union of Immunological Societies (IUIS) (Gangur and Acharya, 2021). Researchers has also pointed out that there may be presence of additional protein allergens in the sesame seeds which are not identified yet (Gangur and Acharya, 2021).

In the present study tertiary structure prediction and evaluation of the allergen profilins (apple, pineapple, wheat and soybean) will be carried out to analyse their biologically-active conformation and further with the advent of virtual screening

suitable pharmacophores will be screened out. Differential amino acid usage signatures of the profilin gene family (allergen and non-allergen) will be reported and further those variations will be linked to their physicochemical properties. The observed distinguished features will be employed as descriptors to develop machine learning based prediction models. The present study will be focused on identifying the uncharacterised sesame seed allergens in the reference proteome by virtue of the developed machine-learning based model. Finally, expression analysis study of the identified allergens will be carried out by RNASeq data analysis.

CHAPTER 2

Review of Literature

2.1 Historical glimpse of food allergy

Antique records pertaining to role of food in the pathogenesis and cure of any ailment is nonsensical, numinous, and preposterous (Cohen, 2008). The first document discussing the concern towards “food-hypersensitivity” was by Shen Nong (2735 BC) and Huang Di (2698 to 2598 BC) in “*Shi Jin-Jing*” (prohibition related to foods), which suggested caution to a female expecting child against non-vegetarian foods like meat, chicken and shrimp and causatives of skin lesion commonly regarded as urticaria and eczema nowadays (Cohen, 2008). In 460 to 377 BC the understanding of Hippocrates regarding atopic individuals suffering from cheese consumption stated that there is something different in the body components of these individuals. He also pointed out that if cheese consumption is not good for human, then it might have harmed all individuals. Presently, this understanding of his pointing out “difference in body components” in atopic individuals is defined as Immunoglobulin-Type-E (IgE) (Cohen, 2008). After this Pedacious Dioscorides (50 AD), Aretaeus the Cappadocian (120 to 180 AD) and Claudius Galen (130 to 200 AD) observed the inimical outcome by consumption of cheese and milk and finally in 1900 the antigen causing discomfort was identified in milk from cow who pasture on wheat bran, hay of peanut and tops of ragweed (Cohen, 2008). Herodotus (484 to 425 BC) has discussed about the Egypt people’s non preferential nature towards consumption of pulses and later observed in the Greek-Roman physicians warning to lentil, bean, and pea consumption (Cohen, 2008). The inception of food mediated hypersensitive reactions among individuals, J. B. van Helmont in 1662 reported a study describing the individual experiencing asthma attack after he consumed fish (Cohen, 2008). The ground-breaking discovery of Richet and Portier in 1902 unravelling about the anaphylaxis and later in 1906 Pirquet’s explanation to allergy as modified response paved the way to unveil mechanisms regulating food-mediated hypersensitive reactions (Cohen, 2008). With this the scientists from European countries initiated their research focusing on clinical studies of the patients and anaphylaxis in guinea pigs (Laroche *et al.*, 1930). In 1930,

Laroche with the co-authors published a study which discusses about food allergen mediated urticaria, asthma and gastro-intestinal complications and found different food sources like milk, eggs and wheat were the mediators of these hypersensitive reactions (Laroche *et al.*, 1930). The European scientists termed these conditions as alimentary anaphylaxis (Laroche *et al.*, 1930). Further, mollusc, fish and crustacean were identified as potential source of food-induced hypersensitive reactions (Laroche *et al.*, 1930). On the other hand, American scientists to a greater extent relied upon the skin tests for distinguishing the food-allergens (Schloss, 1912). The applicability of skin prick tests for the identification of food-allergen mediated hypersensitivity substantiated by Oscar Schloss in 1912 by systematically examining this response in an 8-year-old kid aroused after consuming egg, almond, and oat (Schloss, 1912). Schloss was also able to chemically purify the food extract used as a reagent to proceed with skin prick tests (Schloss, 1912). Further this process of chemical extraction of food reagent for skin prick food-allergen testing was advanced by the work of Coca, and, several physicians found it handy to carry out skin prick test using these materials and thus commercialisation of these test kits took place beyond the use of concerned physicians (Cohen, 2008). Albert Vandeer in 1933 delivered a speech to Society for Study of Asthma and Allied Conditions in which he depreciated the current practice of skin prick test as the standard protocol for food allergy testing by pointing out that it puts the patient in a bizarre and nonpractical prescribed food consumption with no significant results (Cohen, 2008). Further he raised the concern that usefulness of analytical test for food allergy identification needs further characteristic insight of the allergens, benchmarking of the chemicals used in tests and associated clinical investigations (Cohen, 2008). Taking these points into the consideration May presented double-blind placebo-controlled food challenge for diagnosis of food-allergy and later it became a benchmark analysis to assess food-mediated responses (May, 1976). Vaughan along with Frances Wilson in 1930 bring about a revolution in the associated domain by developing a categorical system of the plant extracted foods based on homologous features and further present with a model food- allergen plant which represents all plants falling in its clade (Vaughan, 1930). Additionally, researchers involved in study of plants were able to perceive the formation of new species from their progeny because of single mutational event

which correspond to presence of same allergen in discrete plants by the term cross-reactivity (Cohen, 2008). Another major contribution was from Matthew Walzer in 1931, providing a catalogue in which he classified and categorised food-allergen sources and discussed their implication in regulating hypersensitive reactions in suffering individuals (Cohen, 2008).

2.2 Impact of processing techniques on allergenicity of food allergens

Majority of the food mediated allergic hypersensitive reactions are associated with milk, egg, wheat, fish, soy, tree nut, peanut, and shellfish (Verhoeckx *et al.*, 2015). Various food processing techniques like thermal, homogenisation, hydrolysis, fermentation, and enzymatic treatments tend to predominate the allergenicity potential of antigenic food protein partially (Verhoeckx *et al.*, 2015). Recently, the application of fermentation to soybean found as the best processing method to obtain hypoallergenic soybean (Pi *et al.*, 2021). Fermentation and hydrolysis of milk has proven to be effective against limiting the allergenic potential of milk allergen proteins (Verhoeckx *et al.*, 2015). Heat-treated milk results in whey protein denaturation, while casein does not affect by heat treatment because of the absence of secondary, tertiary, and quaternary protein conformations. Homogenization of milk does not have any role in altering their allergenic potential. By sterilisation, around 25% of whey proteins left intact, while, remaining portion undergoes denaturation and triggered Maillard reaction resulting in significant loss of allergic nature of milk used in edible products made from milk (Porter, 1978; Michalski and Januel, 2006; Huffman and de Barros Ferreira, 2011; Bu *et al.*, 2013). Previously, several studies have observed that consumption of heat-treated eggs by the kids with egg allergy resulted in significant decrease (50 to 85%) in their allergic symptoms (Lemon-Mulé *et al.*, 2008; Turner *et al.*, 2013; Cortot *et al.*, 2012). The consumption of baked products from hazelnuts also reduces its allergenicity significantly in atopic individuals (Worm *et al.*, 2009; Hansen *et al.*, 2003). The incubation of peanuts in boiled water for 20 minutes has proved to be an effective processing measure to reduce IgE binding potential for allergenic food proteins from peanut as confirmed by immunoblotting. Autoclaving (2.56 atm for 30 min) of the roasted peanuts also resulted in decreased IgE binding efficiency for peanut allergenic proteins. Hydrolysis

effect on roasted peanut resulted in hydrolysis of peroxidase, digestive enzymes, and reduced level of allergens viz., Ara-h-1 and Ara h 3, while no effect observed on raw peanut (Yu *et al.*, 2011; Beyer *et al.*, 2001; Cabanillas *et al.*, 2012; Chung *et al.*, 2004).

2.3 IgE cross-reactivity of the food allergens

IgE cross-reactivity pertaining to allergic response refers to the generation of specific IgE against a food allergen protein leading to sensitisation and later this IgE triggers the same response against a protein generally belonging to the same family but from different organism because of inability to distinguish the original sensitiser protein causing allergic reaction (Chruszcz *et al.*, 2018). Profilins possess high sequence identity and conserved tertiary structure which contributes to frequent IgE cross-reactivity among pan allergens (Chruszcz *et al.*, 2018; Mari, 2001). Sensitisation to profilins from pollen also accounts for pollen-food allergy syndromes like mugwort-celery-spice syndrome and ragweed-melon allergic reactions (Asero and Amato, 2011; Ebner *et al.*, 1998). Profilin from pollens, plant foods and latex have been implicated with IgE cross-reactivity and regarded as pan-allergens (Santos and Van Ree, 2011). Various studies have confirmed the profilin sensitization in individuals experiencing food allergy (Lopez-Torrejon *et al.*, 2005; Rosace *et al.*, 2019). The homology of allergenic food protein from peanut has been demonstrated with allergen proteins from soy, legumes, and tree nuts causing IgE cross-reactivity reactions among atopic individuals (Popescu, 2015). The allergen Jun a 3 (Pathogenesis-related 5 protein), from mountain cedar contains the homologous protein sequence from pepper, cherry, kiwi, tomato, and apple (Popescu, 2015). The cross-reactivity among food and aeroallergens of animal, plant and fungal origin has clinical complications of respiratory allergy in patients sensitised with cross-reactive aero and food allergens resulting into oral allergy syndromes, which may extend into severe anaphylaxis (Popescu, 2015). Various cross-reactive food and aeroallergens groups are listed in Table 2.1 below (Popescu, 2015).

Table 2.1: Various cross-reactive food and aeroallergen groups responsible for IgE-cross-reactivity.

<i>Cross-Reactive Group</i>	<i>Allergen Name</i>	<i>Source Organism (Common Name)</i>	<i>Allergen Classification (Food/Aero)</i>
<i>Number 1</i>	Bet v 1	European White Birch	Aeroallergens
	Aln g 1	European Alder	
	Mal d 1	Apple	
	Pru p 1	Peach	Food allergens
	Api g 1	Celery	
	Gly m 4	Soybean	
	Bet v 2	European White Birch	
	Ole e 2	Olive	Aeroallergens
	Che a 2	Lambsquarters	
	Art v 4	Mugwort	
	Amb a 8	Short Ragweed	
Api g 4	Celery		
<i>Number 2</i>	Dau c 4	Carrot	Food allergens
	Pru p 4	Peach	
	Cuc m 2	Muskmelon	
	Mus xp 1	Banana	
	Sin a 4	Yellow Mustard	
	Pla a 3	London Plane Tree	
	Ole e 7	Olive	Aeroallergens
	Art v 3	Mugwort	
	Amb a 6	Short Ragweed	
	Api g 2	Celery	

Number 3	Pru p 3	Peach	Food allergens
	Cuc m LTP	Musk Melon	
	Mus a 3	Banana	
	Sin a 3	Yellow Mustard	
	Der p 10	European House Dust mite	
Number 4	Bla g 7	German Cockroach	Aeroallergens
	Pen m 1	Black Tiger Shrimp	Food allergens
	Myt e 1	Mussel	
	Fel d 2	Cat	
	Can f 3	Dog	Aeroallergens
	Equ c 3	Domestic Horse	Food allergens
	Bos d 6	Domestic Cattle	
Number 5	Sus s 6	Domestic Pig	Food allergens

Some of the crucial syndromes which arises because of cross-reactivity between food and aeroallergens from plant origin are listed in Table 2.2 below (Popescu, 2015).

Table 2.2: Various syndromes associated with cross-reactive food and aeroallergens.

<i>Name of Syndrome</i>	<i>Cross-reactive food and aeroallergens</i>
<i>Birch-apple syndrome</i>	Mal d 1(Apple) homolog to Bet v 1 (European White Birch)
<i>Cypress-peach syndrome</i>	Pru p 3 (Peach) (Non-specific lipid transfer protein)
<i>Celery- Mugwort- Spice Syndrome</i>	Art v 4 (Mugwort), Api g 5 (Celery) homologs to Art v 60 KDa
<i>Mugwort- Peach association</i>	Art v 4 (Mugwort) profilin, Art v 3 (lipid transfer protein)

2.4 Classification of allergen proteins in various protein families

A remarkable study was performed in 2009 which classified the allergen proteins into their respective protein families and motifs (Ivanciuc *et al.*, 2009). It was suggested that distinct similar region on allergen proteins having homologous structure may be of potential significance than overall sequence similarity (Ivanciuc *et al.*, 2009). To find those distinct similar regions on allergen proteins they categorised allergen proteins and their sub-domains present in Structural Database of Allergenic Proteins (SDAP) to their respective protein family using Pfam database (Ivanciuc *et al.*, 2009). Those SDAP allergenic proteins were categorised into 130 Pfams, out of which 31 Pfams contained at least four allergens (Ivanciuc *et al.*, 2009). The copious allergen families classified in Pfam database are represented in Figure 2.1.

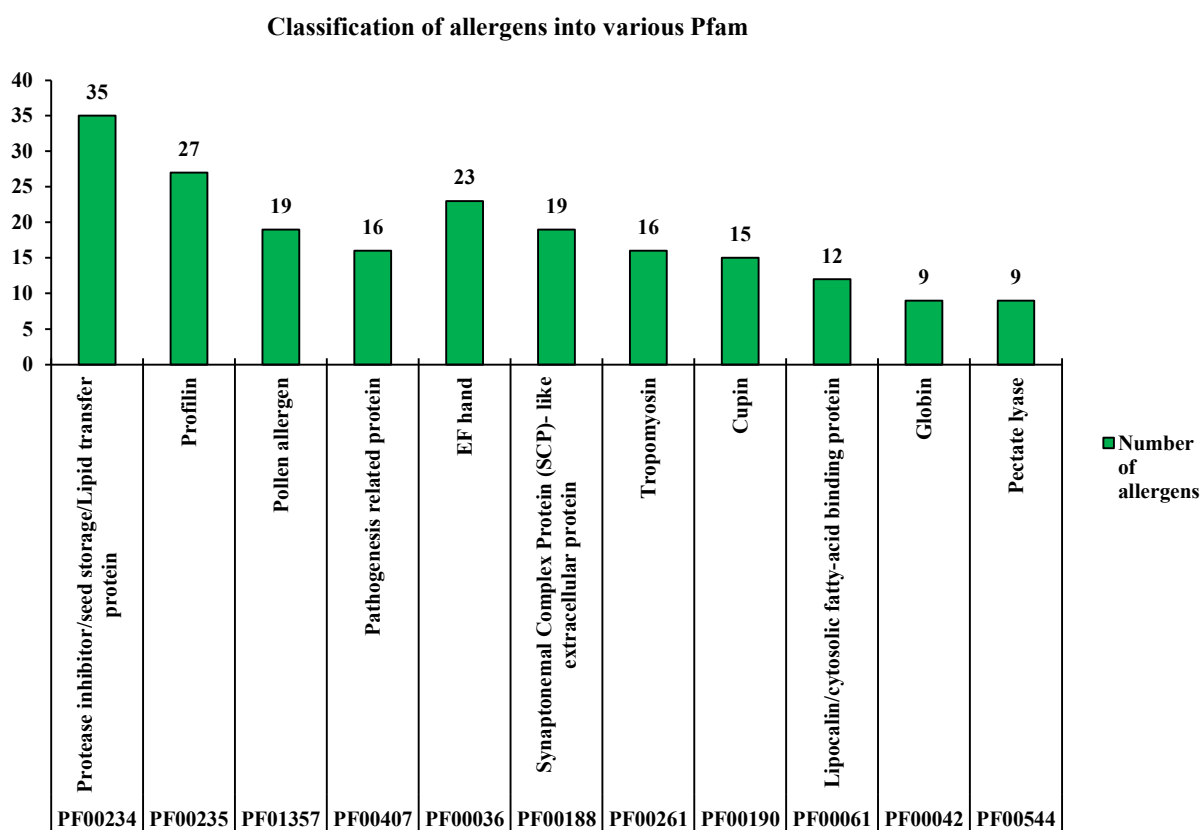


Figure 2.1 Classification of various allergen families by Pfam database. The horizontal axis represents various allergen families whereas vertical axis represents number of allergens in that family.

The Protease inhibitor/Lipid Transfer Protein/seed storage protein family having Pfam code PF00234 includes 34 allergens found in plants. Tertiary structure of three allergens namely Pru p 3, Hor v 1 and Zea m 14 has been classified under this family. Pathogenesis related/Bet v 1 protein family (PF00407) includes pollen, seed, and fruit tissue allergenic proteins. Birch pollen allergen (Bet v 1) has been classified as a major allergen which cross-reacts with allergenic proteins from alder pollen, hazel, chestnut, apple, pear, and stone fruit respectively. Cupin family (PF00190) represents beta-barrel domain families and EF hand family (PF00036) includes calcium attaching sequence motifs primarily observed in parvalbumin. **Profilin family (PF00235) consists of plant profilins having molecular weight of 12-15 kDa and are recognised as conserved proteins. The profilin allergen, Ara h 5 from peanut has been classified under this family (Ivanciuc *et al.*, 2009).**

2.5 Evolution of computationally assisted protein allergen prediction tools

Majority of the allergic reactions are triggered by presence of allergenic proteins in food, pollen, and various other substances in the environment. Considering the health risks posed by allergic responses, it is essential to evaluate their potential allergenicity. The application of omics engineering and processing techniques to alter or introduce a new protein in food or food products has emerged exponentially in previous years, which makes it essential to evaluate allergenicity of these altered or new proteins in order to ensure their non-allergenic nature. **Bioinformatics has played a central role in the endeavours associated with allergenicity assessment by development of various models and web-servers to assess the potential allergenicity** based on amino acid sequence and their associated structural information (Kadam *et al.*, 2016). In the upcoming part, some of the important allergenicity assessment tools developed by advent of computational biology has been reviewed.

2.5.1 Structural Database of Allergenic Proteins (SDAP)

SDAP, a specialised allergen database introduced in 2003, provides sequence, structural, and IgE binding region related annotation of the protein allergens. The database was aimed to ease the sequence homology search among the allergen proteins. The web-server can be accessed by the link <https://fermi.utmb.edu/>. On 12 Feb. 22, the

database contained annotation of 1658 allergenic proteins with 101 protein allergens having 3D structure information cross-referenced from Protein Data Bank (PDB). For the allergenic proteins the database also provides a comprehensive list of their classification in Protein Family (Pfam). The database allows the user to categorically access the allergens present in food, insects, animals, fungi, pollen, and mites (Ivanciuc *et al.*, 2003). User can perform the operation to evaluate their query protein for allergenicity by performing a similarity search against the SDAP database allergens (Ivanciuc *et al.*, 2003). The similarity search performed by the SDAP uses property distance function to provide user with search output (Ivanciuc *et al.*, 2003). Conclusively, the SDAP allows the user to access the allergen database and search for allergenicity of query protein by performing similarity search.

2.5.2 AIgPred

In 2006, a tool named as AIgPred was developed by (Saha and Raghava, 2006), following the in-silico approach for allergenicity assessment of query protein and finding the presence of any IgE binding region. They used a dataset of 578 allergens and 700 non-allergens for training and testing in order to develop a model based on support vector machine (SVM). Features like amino acid constitution and dipeptide constitution retained 85.02%, 84% of precision and specificity respectively (Saha and Raghava, 2006). In motif-based approach by using Multiple Em for Motif Elicitation (MEME)/Motif Alignment they achieved 93.94% and 33.34% sensitivity and specificity respectively (Saha and Raghava, 2006). In their third approach known IgE epitopes were used as a database to search for allergenic proteins which upon validation returned 98.14% specificity and 17.47% sensitivity. In their fourth part, simple BLAST search was performed in contrast to allergenic peptides for the prediction of the allergenic proteins. At last, with the combination of two or more than two methods hybrid method was developed and validated on a dataset of 323 allergens and 101725 non-allergens (Saha and Raghava, 2006). To our knowledge this was the first study which reported the application of machine learning classifiers for allergenicity assessment of any protein.

2.5.3 AllerTool

In 2007, (Zhang *et al.*, 2007) designed a web-based tool named as AllerTool for evaluating the allergenicity and cross-reactivity of the allergen proteins. The tool allows the user to graphically explore the allergen protein's cross-reactivity and allergenicity evaluation model based on **similarity search**. They also developed a support vector machine-based approach to find the allergenicity of query protein with 86% of sensitivity and specificity respectively (Zhang *et al.*, 2007). **The web link at <http://research.i2r.a-star.edu.sg/AllerTool/> on 17 Feb. 22, was found to be inaccessible implying the server has been obsolete.**

2.5.4 SVM based allergen prediction model

A machine-learning based approach, in 2008 was employed to find out the allergenic proteins by means of evolutionary relationship and claimed comparatively enhanced sensitivity and specificity. SVM was used for training and testing a dataset of 693 positive proteins and 1041 negative proteins retrieved from Swiss-Prot (<https://www.uniprot.org/>) and SDAP (<https://fermi.utmb.edu/>) (Kumar and Shelokar, 2008). Features like protein **peptide residues, dipeptide construction and pseudo amino acid construction** were used which resulted in an accuracy of 86.3%, 86.5% and 82.1% respectively. They also used **Position Specific Scoring Matrix (PSSM)**, which provided highest accuracy of 90.1%. For validation of model, 10-fold cross-validation method was used in which dataset was arbitrarily categorised into 10 subsets. The accuracy of the SVM model using PSSM was found superior as compared to existing allergen prediction models like AIgPred and WebAllergen. It was concluded that use of the evolutionary information could be of greater importance for developing highly sophisticated allergen prediction models (Kumar and Shelokar, 2008).

2.5.5 AllerTOP

A web server-based tool named as AllerTop was introduced in 2013 for the allergen prediction. This method presents the first **non-alignment-based approach** for the allergen assessment. They retrieved a sum of 2395 allergens and non-allergens were obtained from the same genus to constitute the positive and negative dataset respectively. They represented features of allergen protein sequences by z1, z2, z3 descriptors and changed to consistent vectors with ACC transformation. Five machine

learning-based approaches implemented in order to find the best-fit model for the prediction, out of which, k nearest neighbour (at k=3) gives the best model in terms of accuracy and was incorporated in the AllerTop web server. AllerTop on comparison with other existing prediction servers, performed better by giving 94% sensitivity (Dimitrov *et al.*, 2013). The updated version of AllerTop can be accessed by the link <https://www.ddg-pharmfac.net/AllerTOP/method.html>.

2.5.6 Allergenicity assessment by Chou's pseudo amino acid composition

(Mohabatkar *et al.*, 2013), in 2013 published a document, which predicts the allergenicity of query proteins by Chou's pseudo amino acid composition and machine learning approaches. Chou's pseudo amino acid composition methodology was developed to improve the allergen prediction efficiency of proteins at sub cellular location and membrane proteins respectively. They used Support Vector Machine for the prediction, which considers vector presentation of the sequences obtained from sequence characteristics. The dataset used to generate the model was obtained from AlgPred web-server. The dataset contained 460 positive protein entries and 560 negative protein entries. In Chou's simulated amino acid construction, they considered hydrophobic, hydrophilic, isoelectric point, pK1 and pK2 characteristics of the amino acids. Accuracy obtained on the dataset by them was 91.9% and Mathew's correlation coefficient value obtained was 0.82. They also compared these prediction results with other programs like AlgPred and found that their algorithm gives efficient accuracy than the one compared with (Mohabatkar *et al.*, 2013).

2.5.7 proAP

In 2013, (Wang *et al.*, 2013), comprehensively analysed computational based methods for allergen prediction. They combined these methods and developed a new tool named as proAP. Support vector machine, sequence and motif-based approaches for allergen prediction were analysed and they found that support vector machine-based approach provided top accuracy and specificity when tested on a dataset of 989 verified allergens and 244,538 non-allergens (Wang *et al.*, 2013). ProAP tool was able to predict the query allergen through world-wide-web search available at <http://gmobl.sjtu.cn/proAP/main.html>, but the link was inaccessible on 21 Feb. 22.

2.5.8 Artificial neural network-based allergen prediction model

Artificial neural network-based model was developed to evaluate the allergenic potential of the various proteins. Two distinct algorithms were coded consisting of three and four steps and their prediction potential was determined on 2427 positive and 2427 negative instances respectively. The positive protein sequences were retrieved from Central Science Laboratory allergen database, Food Allergen Research and Resource Program allergen database, SDAP and Allergome database, while, the non-allergen protein sequences of commonly used food such as bread wheat, tomato, potato, pepper, Asian and African rice were obtained from Swiss-Prot to constitute the negative dataset. In three-step algorithm, firstly characters of amino acids which includes size, hydrophobicity, relative abundance, beta strand and helix propensities were considered. Secondly, they converted strings into vectors of same length via auto and cross covariance. In last step, Artificial neural network was employed to develop the model. In terms of their performance, three-step algorithm was able to predict 82% of the allergens and non-allergens in contrast to four-step algorithm, which was able to predict 76% positive and negative protein entries. They also compared various web tools available for predicting allergenic potential and found that some tools identify allergens and some non-allergens with great accuracy, and finally concluded that utilization of multiple prediction tools is necessary for accurate allergenicity assessment of query proteins (Dimitrov *et al.*, 2014).

2.5.9 Allerdicator

In 2014, (Dang and Lawrence, 2014) developed a tool named Allerdicator for the prediction of allergenic proteins by application of text classification. This method was based on sequence-mediated prediction of allergens and was able to predict large number of protein sequences with effective accuracy in quick response time. The tool directs the protein sequences into text document and applies support vector machine for the allergen prediction. On comparison with other existing prediction tools like AllerHunter, AlgPred and SORTALLER, this method surpasses accuracy of these tools and retrieves the results in quick time. They also found that Allerdicator examined approximately 540000 protein sequences from Uniprot (<https://www.uniprot.org/>) in around six minutes and identified less than 1% of sequences as allergenic (Dang and

Lawrence, 2014). However, the respective web link <http://allerdicator.vbi.vt.edu/> for the tool was inaccessible on 24 Feb. 22.

2.5.10 PREAL^w

In 2017, (Li and Wang, 2017), introduced a new algorithm named as PREAL^w which combines PREAL, FAO/WHO methodology for allergen prediction and motif-based method for allergenicity assessment. It considers weighted score for the allergenicity prediction of query sequences. This method was termed as integrative because of the combination of various prediction methods in order to embed the characteristic of various methods and overcome the limitation of individual predictions. The method was regarded as best suitable for the prediction of crops allergens. The dataset used for the prediction consists of 830 recognized allergens and non-allergens. The method developed gives the accuracy of 85.9% with area under curve value 0.87 which surpasses prediction accuracy of FAO/WHO and PREAL criteria for allergen prediction respectively. They also provided status of the allergens in crops (soybean, wheat, and maize). These crops had 3988 allergens out of which 846 were confirmed and 3142 allergens were predicted by the generated model. It was observed that soybean had 92 confirmed and 299 predicted, rice had 151 confirmed and 927 predicted, maize had 121 confirmed and 932 predicted and wheat contained 482 confirmed and 984 predicted allergens respectively (Li and Wang, 2017). The web-server for the developed model can be accessed at <http://lilab.life.sjtu.edu.cn:8080/prealw/index.php>.

2.5.11 Random Forest model for allergenicity assessment

In 2019, (Westerhout *et al*, 2019), designed a random forest model for the allergenicity assessment of query protein sequences. The dataset was constituted by 525745 instances out of which 1673 were considered allergen and rest non-allergen protein instances. The model was developed on 29 descriptors including physicochemical and biochemical parameters. Some of the features include GRAVY, aliphatic index, instability index, extinction coefficient, amino acids number, secondary structure proportion and positively and negatively charged amino acids. The developed model

was able to achieve more than 85% of sensitivity, specificity, and accuracy respectively (Westerhout *et al*, 2019).

2.5.12 Deep Learning and Ensemble Learning for allergenicity assessment

Recently, in 2021, (Wang *et al*, 2021) employed deep learning and ensemble learning techniques for the allergenicity assessment of the query proteins. The dataset used for model development consists of 583 positive and 600 negative protein instances which were non-redundant respectively. The dataset was further processed to obtain the **pseudo amino acid composition** vectors which were further subjected as input for deep learning and ensemble model development and lastly comparative analysis with other available models was carried out. Through five-fold cross-validation they observed that deep learning model produced comparatively highest area under curve (AUC) value of 0.95. They also pointed out the concern that deep learning model required prior training and this time was observed to be prolonged in order to perform evaluation of the query sequences (Wang *et al*, 2021).

2.5.13 AIgPred 2.0

(Sharma *et al*, 2021) in 2021, introduced an updated version of the AIgPred (section 2.5.2) for the allergenicity assessment. The model was developed on 10075 positive and negative instances respectively. The dataset was employed in the ratio of 80:20 for training and testing respectively and 5-fold cross validation was opted. The techniques employed for model development were homology-based assessment, presence of IgE epitopes, motif similarity, machine learning classifiers and hybrid approach combining all the techniques out of which the best model provided them with receiver operating characteristics curve (ROC) value of 0.98 (Sharma *et al*, 2021). **The web-server for the developed model can be accessed at <https://webs.iitd.edu.in/raghava/algpred2/>.**

Conclusively, the food allergy exists from the ancient time as discussed in the historical background of the section which present the different aspects of considering the impact of food allergy reactions and their understanding towards avoiding the consumption of allergenic food sources. Further insights into the food allergens revealed the antigenic part which is responsible for causing the associated reactions and thus the immunological bases were also revealed. Further the race to develop a standardised

food allergy test kit had numerous complications and till now the associated worldwide organisations rely on a weight of evidence manner to assess the allergenicity potential. We also reviewed some papers in which we have observed a significant reduction in allergenicity of the food sources undergoing various processing techniques and can conclude to a point that allergenic protein potential might have a significance to their conformation. The evolution of the allergenicity prediction tools by utilising numerous sequence and structural features helped to analyse the present state of the allergenicity prediction models.

Chapter 03

Research Gap

Presently, computational biology approaches have been extensively utilised for the allergenicity assessment of foreign proteins. Numerous web-servers and models have been developed recently which are based on homology search with known allergen, motif-based approach, identification of IgE epitopes, support vector machine, and pseudo amino-acid composition, and artificial neural networks for allergenicity assessment of proteins. Some studies have also utilised the physicochemical features like aliphatic index, GRAVY, molecular weight, polarity, amino acid composition and hydrophobicity as input features and resulted in improved efficiency of the model.

- After an in-depth analysis of these utilities, it was observed that systematic physicochemical space mining pertaining to food allergens have been partially explored.
- For the first-time Correspondence and RAAU (relative amino acid usage) analyses will be employed on food allergen dataset to dynamically explore their physicochemical space and further those findings (distinct protein allergen features) will be translated towards development of machine learning based protein allergenicity assessment model.

Chapter 4

Research Objectives

Objective 1: Computational analysis (sequence, structure, and function) of food crop allergen proteins.

Objective 2: Machine learning (SVM and RF) tool development for prediction of food allergens.

Objective 3: Application of the developed model/tool for genome-wide prediction of food allergen.

Objective 4: Expression validation of few food allergens by RNA-Seq data analysis.

Chapter 5

Materials and Methods

5.1 Computational analysis (sequence, structure, and function) of food crop allergen proteins

5.1.1 Multiple sequence alignment of human versus food allergen profilins

Protein sequences of human profilins having Uniprot accessions P07737 (Human profilin-1), P35080 (Human profilin-2), P60673 (Human profilin-3) and Q8NHR9 (Human profilin-4) were aligned with food allergen profilins Q84RR7 (Apple profilin), Q94JN2 (Pineapple profilin), B6EF35 (Wheat profilin) and O65810 (Soybean profilin) respectively by using the ClustalW web-server (Thompson *et al.*, 2003). Apart from selecting slow/accurate parameter, rest of the settings were kept at default to obtain the alignments.

5.1.2 Molecular modelling of the food allergen profilins

Protein sequences of the profilins from apple, pineapple, wheat and soybean were subjected to homology modelling (unavailability of experimentally elucidated 3D structure) by the SWISS Model webserver <https://swissmodel.expasy.org/> at default parameters to obtain their three-dimensional confirmations. This tool develops the homology model by recognition of the template structure for the target protein followed by target protein sequence alignment with the template structure. After this the model generation and evaluation are done by the SWISS MODEL server (Waterhouse *et al.*, 2018). Subsequently, the template pdb structures were retrieved from RCSB (Research Collaboratory for Structural Bioinformatics) PDB (Protein Data Bank) at <https://www.rcsb.org/> shown in Table 5.1.

5.1.3 Molecular dynamics simulation analysis

The molecular dynamic simulation studies help to evaluate the stability of the modelled confirmations (Kar *et al.*, 2021). The molecular dynamic simulation of the modelled structures was performed for 10 nano seconds using the Gromacs software version 5.1.1 (Kar *et al.*, 2021; Van Der Spoel *et al.*, 2005). The molecular dynamic

simulation steps were followed as referred in the (Kar *et al.*, 2021). The box center of 3.111, 3.111, 3.111 nm and box vector of dimension 6.223, 6.223, 6.223 were observed in cubic box type for the modelled pineapple profilin protein. OPLS-all atom force field (2001 amino acid dihedrals) was used to generate the topology of modelled structures (Jorgensen *et al.*, 1996). The simple point charge extended (spce) water model was considered to solvate the box and charged protein was observed having -7 electrons of total charge. The box was charge neutralised by adding 7 sodium ions and topology was processed. Fourier grid of dimension 40x40x40 and spacing of 0.156, 0.156, 0.156 was used for x, y and z axis respectively. NVT ensemble with constant volume and temperature were used to achieve the equilibration and the simulations were performed at 1.013 bar of pressure and 300 Kelvin temperature (Kar *et al.*, 2021). The molecular dynamics trajectory was analysed by the RMSD (Root Means Squared Deviation) values. Similarly, the simulations were carried out for the modelled profilins from apple, wheat and soybean respectively.

5.1.4 Virtual Screening of the modelled allergen profilins

Virtual screening of all the profilins listed in table 1 was carried out using the Pharmit web server available at <http://pharmit.csb.pitt.edu/index.php> (Sunseri and Koes, 2016). The web server allows the user to explore interactive chemical space and return significant hits using state-of-the-art algorithms. The significant hits were filtered based on energy minimised score and RMSD values. The pharmacophore search for the profilin protein was explored against the ZINC database (Sterling and Irwin, 2015). Pharmit also provides a hit screening option, which was employed in present study to retrieve more significant hits by considering Lipinski's Rule of Five (Lipinski, 2004) and Veber's rule for drug-likeness (Veber *et al.*, 2002). All the parameters of drug-likeness considered in the present study are detailed in Table 5.2.

5.1.5 Molecular Docking and Drug likeness

Top hits obtained by the Pharmit virtual screen were further subjected to docking analysis using fast and efficient molecular docking module AutoDock Vina available at <http://vina.scripps.edu/> (Trott and Olson, 2010). The docked conformations were

further processed by the PLIP (Protein-Ligand Interaction Profiler) webserver <https://projects.biotec.tu-dresden.de/plip-web/plip/index> to get non-covalent interaction map among the docked conformations. The physicochemical profiles of the pharmacophores for their bioavailability were revealed using SwissADME (Daina *et al.*, 2017) and admetSAR (Cheng *et al.*, 2012) web servers.

Table 5.1 List of profilins considered in the study along with their Uniprot accessions and templated PDB ID used for homology modelling.

<i>Profilin source organism</i>	<i>Allergome code</i>	<i>Template used</i>	<i>Uniprot ID</i>
<i>Apple</i>	Mal d 4	5NZB	Q84RR7
<i>Pineapple</i>	Ana c 1	5FDS	Q94JN2
<i>Wheat</i>	Tri a 12	5FEF	B6EF35
<i>Soybean</i>	Gly m 3	4ESP	O65810

Table 5.2 Hit screening parameters employed at Pharmit with reference to Lipinski's rule of five and Veber's rule.

<i>Screening parameter</i>	<i>Value range</i>
<i>Molecular weight</i>	150 to 500 Daltons
<i>No of rotatable bonds</i>	0 to 9
<i>LogP (measure of lipophilicity)</i>	0 to 5
<i>PSA (Polar Surface Area)</i>	0 to 140 Å ²
<i>Aromatics</i>	0 to 7
<i>No of H-bond acceptor</i>	0 to 10

No of H-bond donor

0 to 5

5.1.6 Data retrieval for differential amino acid usage analyses of profilin family

Profilin protein sequences were retrieved from the Uniprot knowledge base (<https://www.uniprot.org/>) with search keywords “profilin AND reviewed: yes”. Only those protein hits were included in the dataset whose protein names were designated as Profilin, Profilin 1, 2, 3 up to 12. A total of 408 hits were retrieved from keyword search out of which 200 hits constituted the final dataset. The instances from the query search including names other than profilin like Myb-related transcription factor..., suppressor of yeast deletion..., actin-cytoplasmic-2..., cell division control protein..., vasp, BN-1 and others were excluded from the dataset. The Allergome database, a repository of the allergen molecules, was referred to classify the profilin as allergens and non-allergens. The final dataset comprised of 164 allergen and 36 non-allergen profilins respectively. The index of the final dataset has been provided in Appendix 1.

5.1.7 Estimation of relative amino acid usage

Relative amino acid usage (RAAU) refers to number of times a particular amino acid occurs in a protein relative to the total number of amino acids in that protein. In other words, the amino acid usage reveals the frequency of each amino acid in a protein (Peden, 2000). RAAU of the dataset was calculated using the CodonW (Ver. 1.4.2) software available at http://www.molbiol.ox.ac.uk/cu_ (Peden, 2000).

$$\text{RAAU}(i) = \frac{\text{Number of 'i' amino acids}}{\text{Total number of amino acids}}$$

Where 'i' represents any of the amino acid

5.1.8 Correspondence analysis

Correspondence analysis (CoA), a multivariate statistical method, has been effective in addressing the variations in amino acid usage (Roy and Basak, 2021). It allows the user to explore the systematic synonyms codon/amino acid usage patterns of query genes/proteins (Peden, 2000). CoA represents major features of data variation by placing them along continuous axis according to the differential patterns observed, with each consecutive axis having a diminished effect (Greenacre, 1984). CodonW was used to generate CoA based on RAAU data of the concerned profilin sequences.

5.1.9 Calculation of physicochemical features

The web-servers GlobPlot 2 (<http://globplot.embl.de/>), Protein-Sol (<https://protein-sol.manchester.ac.uk/>) and PeptideCutter (https://web.expasy.org/peptide_cutter/) were used to calculate the disorder, solubility, and trypsin digestion properties of the profilin proteins respectively, with default parameters.

5.1.10 Statistical analysis

The SPSS v17.0 software was used to calculate Pearson's correlation analysis at significance levels of $P < 0.05$ and $P < 0.01$.

5.1.11 Sequence and structural analyses of allergen profilins

All the allergen profilins having experimentally resolved 3D structures were processed for in-depth sequence and structure analyses. The search query term “profilin” AND “allergen” at the PDB was used. The high-resolution PDB structures 5FEG (Hev b 8), 5EMO (Art v 4), 5NZB (Bet v 2), 6MBX (Cuc m 2), 7KYW (Phl p 12), 5FEF (Zea m 12), 4ESP (Ara h 5) and 5EM1 (Amb a 8) corresponding to the allergen profilins from *Hevea brasiliensis* (Hev b 8), *Artemisia vulgaris* (Art v 4), *Betula verrucosa* (Bet v 2), *Cucumis melo* (Cuc m 2), *Phleum pratense* (Phl p 12), *Zea mays* (Zea m 12), *Arachis hypogaea* (Ara h 5) and *Ambrosia artemisiifolia* (Amb a 8), respectively, were retrieved from PDB. Furthermore, the corresponding protein sequences were retrieved from UniProt and were subjected to multiple sequence alignment using the MEGAX software (Kumar et al. 2018). The z-scores of the alignment were computed employing the multi-Harmony web-server

<https://www.ibi.vu.nl/programs/shmrwww/> (Brandt *et al.*, 2010). The surface exposed residues were identified through the GETAREA web server (http://curie.utmb.edu/area_man.html) (Fraczkiewicz and Braun, 1998).

5.2 Machine learning (SVM and RF) tool development for prediction of food allergens

5.2.1 Dataset curation, descriptors selection and numerical indices calculation

The positive dataset for the study was constituted by including literature verified food allergen protein instances retrieved from Allergen online (<http://www.allergenonline.org/>), WHO/IUIS Allergen Nomenclature (<http://allergen.org/>) and SDAP (Structural Database of Allergenic Proteins) at (<https://fermi.utmb.edu/>), whereas the negative dataset was curated from Uniprot database having non-allergen protein instances verified from pathology and biotech section under protein annotation window (Not assigned any allergome ID or studies linked to allergenic properties stimulation). Further, all the instances which contains the keywords “*Profilin*”, “*Sesame*”, “*Sesamum indicum*” were excluded from the dataset to avoid biased results. The final dataset was constituted by 1200 non-redundant instances having equal ratio of positive and negative entries.

Differential amino acid usage analysis (profilin family) and in-depth literature studies revealed sequence and structure-based properties distinctive of allergen proteins. Further, these properties were mapped with AAIndex database at <https://www.genome.jp/aaindex/> to retrieve numerical indices of the respective descriptors. The final index of descriptors considered in the study are shown in Table 5.3. Lastly, the numerical indices of the entire manually curated dataset were generated by in-house developed script using MATLAB (Higham and Higham, 2016).

5.2.2 WEKA for machine learning model development

WEKA (Waikato Environment for Knowledge Analysis) version 3.8.4, allows the users to generate array of machine learning based models on the various datasets (Hall *et al.*, 2009). It includes algorithms of clustering, classification, regression and feature selection (Hall *et al.*, 2009). WEKA also allows to perform cross-validation of the

developed models along with data visualisation options (Hall *et al.*, 2009). The tool was employed in the present study to develop and cross-validate supervised machine learning models on the curated dataset. Initially the curated dataset was uploaded in the WEKA interface under the pre-process panel and under the classify tab the desired classifier was chosen. Further under the test option, ten-fold cross-validation was opted in which 90 percent of the data was used as training data and rest of the 10 percent was employed as testing data for all of the developed models.

Initially, ZeroR classifier was employed to calculate the baseline accuracy of the model with respect to the dataset (Hall *et al.*, 2009). The classifier acts as a model classifier to compare with other classifiers (Hall *et al.*, 2009). Further, Support Vector Machine (SVM) and Random Forest (RF) models were generated using functions.LibSVM and trees.RandomForest classifiers in the WEKA suite. The SVM classifier in case of linearly separable state provides an optimal hyperplane which best separates each class. The optimal hyperplane is achieved by maximising the margin, which is a distance from hyperplane boundary to the closest class value (Jackins *et al.*, 2021). On the other hand, Random Forest (RF) classification systems are ranked/ordered classification in the form of decision-tree. This classification system identifies the best descriptors for a dataset based on probability and returns a tree-based classification model for the dataset (Jackins *et al.*, 2021).

5.2.2a Evaluation of the developed classifiers

True Positive (TP) Rate: TP rate is the proportion of examples which were classified as class x, among all examples which truly have class x, which means how much part of the class was captured. It is equivalent to Recall (Hall *et al.*, 2009). Similarly, accuracy is evaluated by (Saha *et al.*, 2006).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

False Positive (FP) Rate: FP rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x (Hall *et al.*, 2009).

Precision: The Precision is proportion of examples which truly have class x among all those which were classified as class x (Hall *et al.*, 2009).

F-Measure: The F score is simply a combined measure for precision and recall (Hall *et al.*, 2009).

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Mathew's Correlation Coefficient (MCC) Value: The MCC is a measure of the classifier's classification potential. Its value ranges from -1 to 1, where -1 represent the misclassification and 1 depicts perfect classification. MCC value of 0.5 signifies a random classification by the classifier (Hall *et al.*, 2009).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Receiver Operating Characteristics (ROC) Area: The ROC Curve is a measure to evaluate the performance of a binary classifier by plotting a graph between TP rate and FP rate (Hall *et al.*, 2009).

Table 5.3 Descriptors considered in the study along with their AAIndex accessions.

<i>Descriptor</i>	<i>AAIndex/Web server</i>	<i>Reference</i>
<i>Hydrophobicity</i>	ARGP820101	Kopper <i>et al.</i> , 2005
<i>Flexibility</i>	BHAR880101	Breiteneder and Mills, 2005
<i>Helix propensity</i>	KANM800101	Breiteneder and Mills, 2005
<i>Strand propensity</i>	GEIM800105	Breiteneder and Mills, 2005
<i>Size</i>	AllergenFP	Dimitrov <i>et al.</i> , 2014a
<i>Relative abundance</i>	AllergenFP	Dimitrov <i>et al.</i> , 2014a
<i>D1 (B-values)</i>	VNM940102	Han <i>et al.</i> , 2009
<i>D2 (Transfer free</i>	BULH740101	Han <i>et al.</i> , 2009

<i>energy to surface)</i>		
<i>D3 (Knowledge-based membrane-propensity scale from 3D_Helix)</i>	PUNT030102	Han <i>et al.</i> , 2009
<i>D4 (Normalized frequency of beta-turn)</i>	CHOP780203	Han <i>et al.</i> , 2009
<i>D5 (Normalized frequency of C-terminal non beta region)</i>	CHOP780211	Han <i>et al.</i> , 2009

5.3 Application of the developed model/tool for genome-wide prediction of food allergen

To proceed with independent dataset testing of the developed model, the representative genome of *Sesamum indicum* cultivar: Zhongzhi No. 13, BioProject accession: PRJNA268358 was considered (Wang *et al.*, 2016). The proteome corresponding to the accession having 24,106 instances was retrieved and associated numerical indices were calculated for the dataset by in-house developed script. Finally, using WEKA the test dataset was evaluated by allocating the corresponding indices file in arff (Attribute Related File Format) file extension and predictions were analysed.

5.4 Expression validation of few food allergens by RNA-Seq data analysis

5.4.1 Transcriptomic data retrieval from NCBI-SRA

Transcriptomic data of indigenous *Sesamum indicum* having accession ID-SRR12153208 performed on Illumina NovaSeq 6000 platform was retrieved from NCBI-SRA toolkit. The run was having 18.7 Giga bases with total size of 5.6 GB. The run generated 6202914 raw reads with each read having length of 151 bases.

5.4.2 Evaluation by the FastQC

This tool was employed to check for the quality of the raw sequencing data which should be resolved before moving further to avoid any false positive results (Andrews, 2010). The module analyses the raw sequencing data by various measures like per sequence quality score which displays a graphical representation of sequence quality score in the data and displays warning if the average quality score value lowers than 27 representing 0.2 percent sequencing error rate (Andrews, 2010). Further, it allows to detect the presence of adapters in the raw sequencing data which should be filtered before further processing (Andrews, 2010).

5.4.3 Trimmomatic for the removal of adapter content from the transcriptomic data

The Trimmomatic tool in the present study was employed to resolve the presence of adapter content in the raw data in order to prevent from generating false positive results. Additionally, this tool also helps to remove the low-quality bases or any other type of unwanted contamination from the query data (Bolger *et al.*, 2014). The performance of the current protocol was carried out on high-performance computational system- Dell Precision Tower 3620 (RAM- 32GB) on Ubuntu Desktop 20.04 LTS operating system.

```

at org.usadellab.trimmomatic.TrimmomaticSE.process(TrimmomaticSE.java:197)
at org.usadellab.trimmomatic.TrimmomaticSE.run(TrimmomaticSE.java:321)
at org.usadellab.trimmomatic.Trimmomatic.main(Trimmomatic.java:85)
(base) precislontower@precislontower-Precision-Tower-3620:~/Documents/bhupender/Sesame_RNAseq$ java -jar /usr/share/java/trimmomatic.jar SE -threads 4 -phred33 -trinlog logfiletr
rln -summary summarystats -quiet /home/precislontower/Documents/bhupender/Sesame_RNAseq/sesame_sra_data.fastq sesame_sra_data_trimmomatic.fastq ILLUMINACLIP:/usr/share/trimmomat
ic/TruSeq3-SE-fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
Exception in thread "main" java.io.FileNotFoundException: /home/precislontower/Documents/bhupender/Sesame_RNAseq/sesame_sra_data.fastq (No such file or directory)
at java.base/java.io.FileInputStream.open0(Native Method)
at java.base/java.io.FileInputStream.open(FileInputStream.java:219)
at java.base/java.io.FileInputStream.<init>(FileInputStream.java:157)
at org.usadellab.trimmomatic.Fastq.FastqParser.parse(FastqParser.java:135)
at org.usadellab.trimmomatic.TrimmomaticSE.process(TrimmomaticSE.java:197)
at org.usadellab.trimmomatic.TrimmomaticSE.run(TrimmomaticSE.java:321)
at org.usadellab.trimmomatic.Trimmomatic.main(Trimmomatic.java:85)
(base) precislontower@precislontower-Precision-Tower-3620:~/Documents/bhupender/Sesame_RNAseq$ java -jar /usr/share/java/trimmomatic.jar SE -threads 4 -phred33 -trinlog logfiletr
rln -summary summarystats -quiet /home/precislontower/Documents/bhupender/Sesame_RNAseq/sesame_sra_data.fastq sesame_sra_data_trimmomatic.fastq ILLUMINACLIP:/usr/share/trimmomat
ic/TruSeq3-SE-fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
Exception in thread "Thread-0" java.lang.StringIndexOutOfBoundsException: String index out of range: 0
at java.base/java.lang.StringLatin1.charAt(StringLatin1.java:47)
at java.base/java.lang.String.charAt(String.java:693)
at org.usadellab.trimmomatic.Fastq.FastqParser.parseOne(FastqParser.java:65)
at org.usadellab.trimmomatic.Fastq.FastqParser.next(FastqParser.java:179)
at org.usadellab.trimmomatic.threading.ParserWorker.run(ParserWorker.java:42)
at java.base/java.lang.Thread.run(Thread.java:829)
(base) precislontower@precislontower-Precision-Tower-3620:~/Documents/bhupender/Sesame_RNAseq$ java -jar /usr/share/java/trimmomatic.jar SE -phred33 -trinlog logfiletrln -summar
y summarystats -quiet /home/precislontower/Documents/bhupender/Sesame_RNAseq/sesame_sra_data.fastq sesame_sra_data_trimmomatic.fastq ILLUMINACLIP:/usr/share/trimmomatic/TruSeq3-
SE-fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
Exception in thread "Thread-0" java.lang.StringIndexOutOfBoundsException: String index out of range: 0
at java.base/java.lang.StringLatin1.charAt(StringLatin1.java:47)
at java.base/java.lang.String.charAt(String.java:693)
at org.usadellab.trimmomatic.Fastq.FastqParser.parseOne(FastqParser.java:65)
at org.usadellab.trimmomatic.Fastq.FastqParser.next(FastqParser.java:179)
at org.usadellab.trimmomatic.threading.ParserWorker.run(ParserWorker.java:42)
at java.base/java.lang.Thread.run(Thread.java:829)
(base) precislontower@precislontower-Precision-Tower-3620:~/Documents/bhupender/Sesame_RNAseq$ java -jar /usr/share/java/trimmomatic.jar SE -phred33 /home/precislontower/Documen
ts/bhupender/Sesame_RNAseq/sesame_sra_data.fastq sesame_sra_data_trimmomatic.fastq ILLUMINACLIP:/usr/share/trimmomatic/TruSeq3-SE-fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4
:15 MINLEN:36
TrimmomaticSE: Started with arguments:
-phred33 /home/precislontower/Documents/bhupender/Sesame_RNAseq/sesame_sra_data.fastq sesame_sra_data_trimmomatic.fastq ILLUMINACLIP:/usr/share/trimmomatic/TruSeq3-SE-fa:2:30:1
0 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
Automatically using 4 threads
Using Long Clipping Sequences: 'ACATCGGAAGACGCTCGTGTAGCGGAAGAGTGA'
Using Long Clipping Sequence: 'ACATCGGAAGACGACACGCTCAACTCCAGTCA'
ILLUMINACLIP: Using 0 prefix pairs, 2 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences
Exception in thread "Thread-0" java.lang.StringIndexOutOfBoundsException: String index out of range: 0
at java.base/java.lang.StringLatin1.charAt(StringLatin1.java:47)
at java.base/java.lang.String.charAt(String.java:693)
at org.usadellab.trimmomatic.Fastq.FastqParser.parseOne(FastqParser.java:65)
at org.usadellab.trimmomatic.Fastq.FastqParser.next(FastqParser.java:179)
at org.usadellab.trimmomatic.threading.ParserWorker.run(ParserWorker.java:42)
at java.base/java.lang.Thread.run(Thread.java:829)
Input Reads: 1549899 Surviving: 1547943 (99.78%) Dropped: 192957 (1.22%)
TrimmomaticSE: Completed successfully
(base) precislontower@precislontower-Precision-Tower-3620:~/Documents/bhupender/Sesame_RNAseq$

```

Figure 5.1 Command line execution window of Trimmomatic tool.

5.4.4 Spliced Transcripts Alignment to a Reference (STAR) for transcriptome mapping and differential expression studies

This module was utilised in the present investigation for contig assignment/construction of the transcriptomic data. The contig assignment was carried out by mapping the transcripts to annotated features of *Sesamum indicum* cultivar: Zhongzhi No. 13, BioProject accession: PRJNA 268358. STAR algorithms are well known to provide intensive accuracy with mapping speed of 50x more than other alignment programs but it requires large storage space (Dobin *et al.*, 2013). The mapping process of the program includes looking for the seeds followed by cluster generation, joining the reads and finally scoring of the reads (Dobin *et al.*, 2013). For every read of the query transcriptomic data, the STAR algorithms look for highest sequence length that overlaps with the reference genome. Now, the unmatched part of the read is further searched to look for its highest sequence length that matches the reference genome and by this generation of numerous seeds takes place (Dobin *et al.*, 2013). In case of the unmatched read sequence, the algorithm extends the previous seeds and performs the scoring which in turn if gave poor alignment score then the associated read portion will be assigned as contamination or adapter content. After this the generated seeds are joined together (a) based on their tendency of being close to the corresponding seeds (b) the seeds which are mapped to single location in the reference genome (c) arrangement of reads which gives best mapping score based on insertion, deletion, mismatch and gaps in order to form the complete read.

By using -quantMode option in the STAR (version 2.7.9a) binary number of reads mapped to the reference genes of *Sesamum indicum* Zhongzhi No., 13 were analysed (Dobin *et al.*, 2013). A read count is considered only if it matches by one or more nucleotides to a single gene of reference genome (Dobin *et al.*, 2013). After this, normalised mapped reads value was calculated by using the following equation.

$$\text{Normalized mapped reads percent for n gene} = \frac{\text{Number of mapped reads for n gene}}{\text{Total number of mapped reads}} \times 100$$


```

EditPad Pro 8 - [D:\DOCTORATE\THESIS\OBJECTIVE MEDIATED WORK\Objective#4\STAR output files\Log.out]
File Edit Project Search Go Block Mark Fold Tools Macros Extra Convert Options View Help
Log.out
STAR version=2.7.3a
STAR compilation time,server.dir=<not set in Debian>
#### Command Line:
STAR --quantMode TranscriptomeSAM GeneCounts --genomeDir /home/precisiontower/STAR-2.7.9a/index/ --readFilesIn /home/precisiontower/STAR-2.7.9a/sesame_sra_data_trimmomatic.fastq --sjdbGTFfile /home/precisiontower/STAR-2.7.9a/S_indicum_v1.0_annotatedfeatures.gtf/ncbi_dataset/data/GCF_000512975.1/genomic.gtf
#### Initial USER parameters from Command Line:
#### All USER parameters from Command Line:
quantMode      TranscriptomeSAM  GeneCounts      ~RE-DEFINED
genomeDir      /home/precisiontower/STAR-2.7.9a/index/  ~RE-DEFINED
readFilesIn    /home/precisiontower/STAR-2.7.9a/sesame_sra_data_trimmomatic.fastq  ~RE-DEFINED
sjdbGTFfile    /home/precisiontower/STAR-2.7.9a/S_indicum_v1.0_annotatedfeatures.gtf/ncbi_dataset/data/GCF_000512975.1/genomic.gtf  ~RE-DEFINED
#### Finished reading parameters from all sources

#### Final user re-defined parameters-----:
genomeDir      /home/precisiontower/STAR-2.7.9a/index/
readFilesIn    /home/precisiontower/STAR-2.7.9a/sesame_sra_data_trimmomatic.fastq
sjdbGTFfile    /home/precisiontower/STAR-2.7.9a/S_indicum_v1.0_annotatedfeatures.gtf/ncbi_dataset/data/GCF_000512975.1/genomic.gtf
quantMode      TranscriptomeSAM  GeneCounts

#### Final effective command line:
STAR --genomeDir /home/precisiontower/STAR-2.7.9a/index/ --readFilesIn /home/precisiontower/STAR-2.7.9a/sesame_sra_data_trimmomatic.fastq --sjdbGTFfile /home/precisiontower/STAR-2.7.9a/S_indicum_v1.0_annotatedfeatures.gtf/ncbi_dataset/data/GCF_000512975.1/genomic.gtf --quantMode TranscriptomeSAM GeneCounts

Finished loading and checking parameters

```

Figure 5.2 STAR command line execution window.

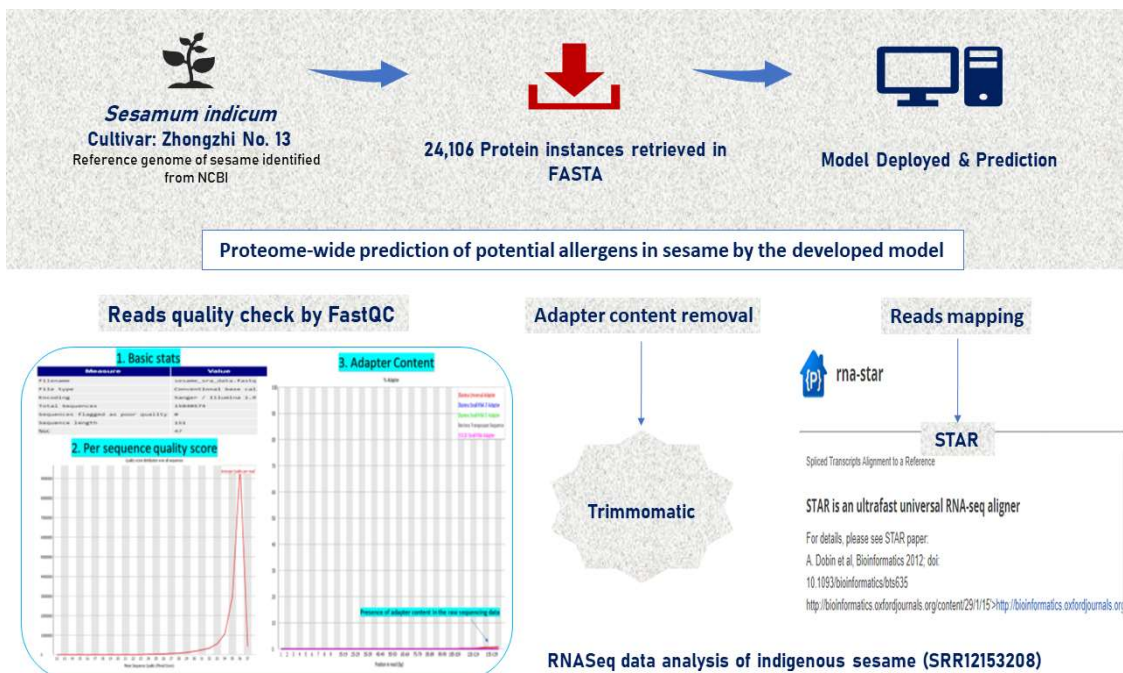


Figure 5.3 Schematic representation of proteome wide prediction and RNASeq data analysis for sesame.

Chapter 6

Results & Discussion

6.1 Computational analysis (sequence, structure, and function) of food crop allergen proteins

The omnipresence of profilin expression in eukaryotic organisms has led to allergenicity assessment triggered by profilin proteins in animal and human (Santos and Van Ree, 2011). During the first identification of profilin allergen Bet v 2, in white birch pollen (*Betula verrucosa*), the IgE auto-reactivity to the human profilin has been confirmed by immunoblot assay in birch pollen sensitised individuals (Valenta *et al.*, 1991). Although, the binding dynamics of IgE to human profilin were observed to be nominal but through basophil activation test the histamine release was found in only those patients which had higher IgE expression against allergen profilin from birch pollen. To further investigate on this account, we have performed sequence analysis of all human profilin isoforms against the allergen profilins from apple, pineapple, wheat, and soybean.

6.1.1 Sequence alignment profiles of human versus food allergen profilins

Human profilin sequences were aligned with food allergen profilins (apple, pineapple, wheat, and soybean) to find the degree of identity among them. The multiple sequence alignment profile of the human and food allergen profilins is shown in Figure 6.1 and subsequently the percent identity matrix of obtained alignment is shown in Table 6.1. Human profilin-1 showed 22%, 17%, 21%, 22% of identity with food profilins from apple, pineapple, wheat, and soybean respectively. Human profilin-2 isoform was found to be 27%, 24%, 23%, 22% identical to apple, pineapple, wheat, and soybean profilins respectively. Similarly, lower sequence identity of 12%, 19%, 23%, 12% of human profilin-3 was seen with apple, pineapple, wheat, and soybean profilins respectively. Human profilin-4 showed 24%, 24%, 21%, 24% of sequence identity against apple, pineapple, wheat, and soybean profilins respectively. Thus, percent identity matrix revealed a very low level (< 24%) of

identity among aligned human and food allergen profilins from apple, pineapple, wheat, and soybean.

This lower sequence identity of human profilins against profilin food allergens from apple, pineapple, wheat, and soybean corresponds to previous studies which have also demonstrated 34% sequence identity of human profilin against the birch pollen allergen profilin (Valenta *et al.*, 1991). Another major contributor to rarely observed animal profilin mediated allergic responses is the taxonomical relativeness of profilins from human and animal and also observed in our investigation, a low degree of similarity between human and plant profilins from apple, pineapple, wheat, and soybean which cancels out the phenomenon of IgE cross-reactivity (Santos and Van Ree, 2011). In addition to this, the plant profilins generate the sensitivity via the respiratory system and then due to high-level similarity with food profilins exhibits cross-reactivity, while animal profilins are consumed orally and gets digested before being presented to the associated digestive immune system (Santos and Van Ree, 2011).



Figure 6.1 Multiple sequence alignment profile of human versus food allergen profilins. The identical amino acids identified by the alignment are highlighted in yellow colour.

Table 6.1 Percent identity profile of the aligned human and food allergen profilins respectively.

	<i>Human profilin-1</i>	<i>Human profilin-2</i>	<i>Human profilin-3</i>	<i>Human profilin-4</i>
<i>Apple profilin</i>	22 %	27 %	12 %	24 %
<i>Pineapple profilin</i>	17 %	24 %	19 %	24 %
<i>Wheat profilin</i>	21 %	23 %	23 %	21 %
<i>Soybean profilin</i>	22 %	22 %	12 %	24 %

6.1.2 Structural evaluation of modelled food allergen profilins

Elucidation of 3-dimensional conformations of proteins are essential to explore their functional properties at macromolecular level, which in-turn opens a wide array of its utilisations in associated domains (Waterhouse *et al.*, 2018). Proteins and their associated interactions are essential components responsible for regulating various cellular-level processes and their comprehension allows us to regulate these systems (Waterhouse *et al.*, 2018). Although, electron microscopic based technologies have evolved the dynamics for structural elucidation, but there is a void between protein-protein association determination strategies (including yeast-2-hybridisation assay, phage display method and affinity chromatography) which provides high-rate output and number of newly experimentally resolved protein conformations. This certainly demands the intervention of computational strategies to perform protein modelling functionalities (Waterhouse *et al.*, 2018).

Profilin from apple, pineapple, wheat, and soybean has been associated with triggering IgE mediated responses in atopic individuals (Ma *et al.*, 2006; Reindl *et al.*,

2002; Rihs *et al.*, 1994; Rihs *et al.*, 1999). In the present study profilin allergens were explored computationally to develop homology models as their experimentally resolved structures were unavailable. The homology modelling of the profilin from apple, pineapple, wheat, and soybean opens the space for secondary-structure level comparisons. The best-modelled structures of these profilins are shown in Figure 6.2. The deduced 3D model shows that the profilins possess an equal number of helices and strands, contributing to their structural-level similarity. Profilin from apple, wheat and soybean were observed to have four helices while pineapple has three helices. All the modelled profilins also possess seven strands and 11 loops, as shown in Figure 6.2. Similar observations have been obtained in case of profilin structures from tomato and latex validating the correctness of deduced structure in the present analysis (Włodarczyk *et al.*, 2022).

The Ramachandran plot has been utilised extensively in protein informatics and is recognised as best tool to evaluate the quality of modelled proteins. The plot analyses the modelled conformations by projecting the phi (N-C α) and psi (C α -C) angles of amino acids in a 2-dimensional space. The plot considers Vander walls radius by taking atoms as spheres and for this, phi and psi angles possible rotations represents favoured regions respectively and the angles at which these atomic sphere clashes were referred to as disallowed regions in the plot (Ramachandran *et al.*, 1963).

The Ramachandran plot of the profilins revealed that 96.12%, 96.12%, 96.90% and 98.43% amino acids of the apple, pineapple, wheat, and soybean, respectively, lies under the favoured region as shown in Figure 6.3. These values validate the accuracy of steric conformations in the modelled profilin structures.

The MolProbity score, evaluating the modelled-protein quality, corresponds to the crystallographic resolution at which the modelled structure assumed to be obtained (Davis *et al.*, 2007). MolProbity score of 1.31, 1.16, 1.37 and 1.10 for the modelled profilins from apple, pineapple, wheat, and soybean respectively depicts the good model quality as the lower resolution value corresponds to better modelled structure.

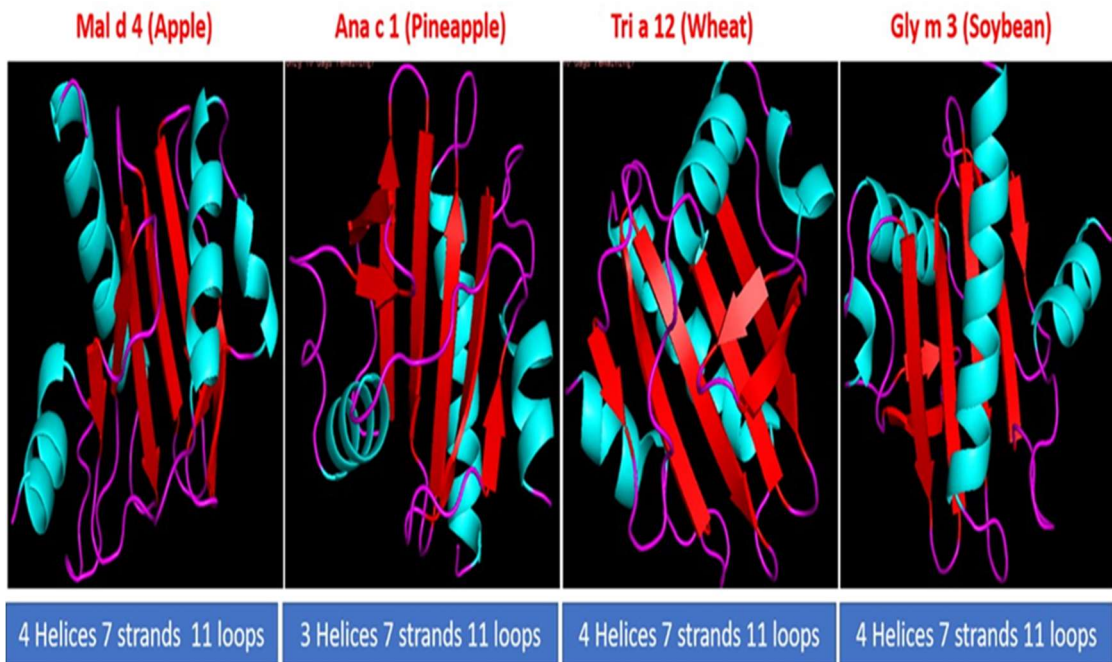


Figure 6.2 Homology modelled 3D structures of profilin from apple, pineapple wheat and soybean. Various secondary conformations of the model like helix, strand and loop are represented by blue, red, and purple colour respectively.

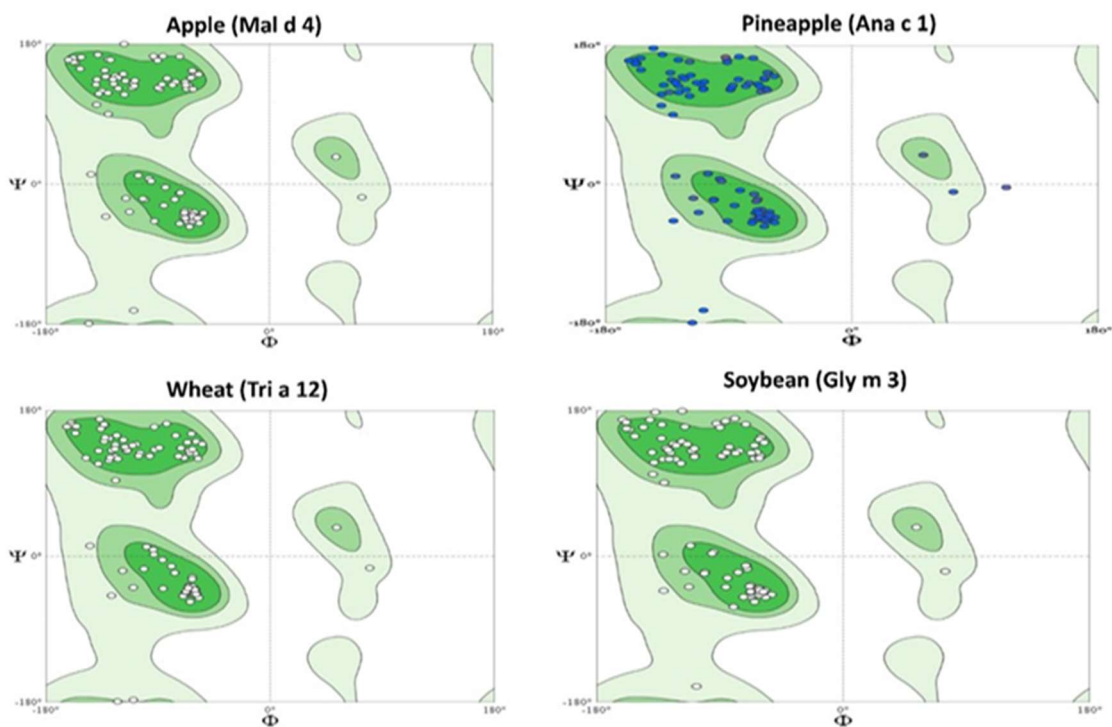


Figure 6.3 Ramachandran plot of the modelled profilins from apple, pineapple, wheat, and soybean.

6.1.3 Molecular dynamics simulation analysis of the modelled profilin structures

The modelled allergen profilins from apple, pineapple, wheat, and soybean were subjected to molecular dynamics simulation for 10 nano seconds and based on RMSD values their conformational stability was analysed. RMSD values are obtained based on comparing the atomic co-ordinates of the query molecule at any time of simulation with the coordinates of the reference structure (Van Der Spoel *et al.*, 2005). The RMSD plot of the 10 nano second simulation is shown in Figure 6.4. Initially, we saw a spike in the RMSD of all the modelled profilins which was around 2 Å for apple and pineapple, whereas it was around 3 Å for wheat and soybean. The fluctuation in the RMSD was observed approximately till 0.64 nano second for apple. The pineapple modelled structure was observed to fluctuate till 0.19 nano second. The fluctuation in wheat modelled profilin structure was observed until 0.60 nano second. Lastly, soybean profilin structure was observed to show fluctuation in their atomic coordinates till 0.17 nano second. After the observed fluctuations in the initial phases of the simulation their atomic coordinates were observed to attain equilibrium in all cases. Thereafter, stability was observed in all the modelled profilins which was around 1.2 Å for modelled apple profilin. Pineapple simulated modelled structure showed stability by showing fluctuations around 2.3 Å till complete simulation. Wheat profilin was subjected to simulation which showed fluctuations around 0.2 Å post equilibrium attainment. Soybean profilin was observed to fluctuate with RMSD value of 2.2 Å post equilibrium till the complete simulation cycle. Least RMSD value corresponding to better stability was seen in modelled profilin structure of wheat and apart from this all other profilins were not showing fluctuation more than 2.2 Å after the initial phase. From this, the stability of modelled profilins can be conferred and hence were opted as suitable receptor for the virtual screening studies.

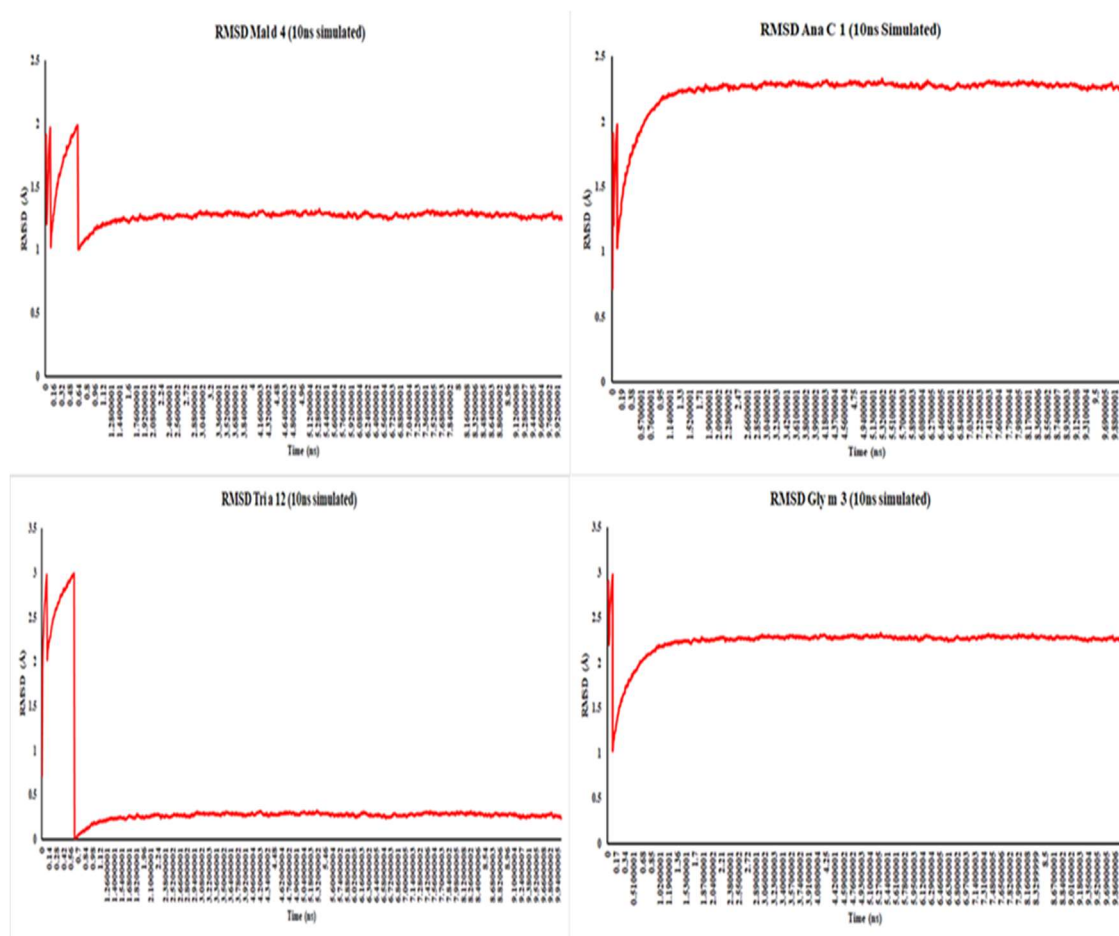


Figure 6.4 RMSD plot of the modelled allergen profilins from Mal d 4 (apple), Ana c 1 (pineapple), Tri a 12 (wheat) and Gly m 3 (soybean). The x-axis represents the time in nano seconds, whereas y-axis denotes RMSD values in Å.

6.1.4 Virtual screening of the modelled profilins from apple, pineapple, wheat, and soybean

The utilisation of structural dynamics of a drug-target molecule has become an important aspect to screen activity-specific pharmacophores in a time and resource saving manner (Lionta *et al.*, 2014). Application of tertiary structure information has advantage over the conventional drug discovery hierarchy since it utilises the conformational aspect of the target molecule causing any disease (Lionta *et al.*, 2014). The application of computational strategies during the process of drug design and development has completely changed the dynamics of pharmacophore identification by providing significant hits in lesser time and resource utilisation (Lionta *et al.*,

2014). Virtual screen allows the user to perform computationally assisted screening against the query receptor molecule to identify the best hits from libraries of purchasable pharmacophores (Lionta *et al.*, 2014).

Taking this into account the virtual screening against the modelled allergen profilin was performed by using a suitable web server to search for the suitable pharmacophores against the ZINC (purchasable) library. The Pharmit web server retrieved the best pharmacophore hits based on energy minimised score and RMSD values. The best hits retrieved for the apple, pineapple, wheat, and soybean profilins are shown in Table 6.2.

Table 6.2 Top ranked inhibitors screened by the Pharmit web server against ZINC (Purchasable) database for the allergen profilins.

<i>Profilin source organism</i>	<i>ZINC ID</i>	<i>Score</i>	<i>mRMSD</i>
<i>Malus domestica (Apple)</i>	ZINC000524729534	-9.99	0.778
<i>Ananas comosus (Pineapple)</i>	ZINC000000041632	-7.48	0.965
<i>Triticum aestivum (Wheat)</i>	ZINC000065529251	-7.92	0.733
<i>Glycine max (Soybean)</i>	ZINC000257349595	-7.85	1.455

Further the best identified pharmacophores against the considered allergen profilins were explored as per their annotation in the referenced database. The first

pharmacophore ZINC000524729534, against the modelled allergen profilin has chemical formula of $C_{27}H_{42}O_6$, having 6 rings, 33 heavy atoms with molecular weight of 462.627. The compound as per their characterisation in the dataset was categorised under the keyword anodyne and natural products. Next pharmacophore ZINC000000041632 screened against pineapple modelled allergen profilin was having chemical formula $C_{15}H_9NO_4$ with 3 rings, 20 heavy atoms and molecular weight of 267.24. The pharmacophore was categorised under the standard library of the database as observed by associated keywords against the compound. The pharmacophore ZINC000065529251 screened against the modelled wheat allergen profilin has molecular weight of 347.418 and chemical formula $C_{21}H_{21}N_3O_2$. The pharmacophore bears 4 rings, 26 heavy atoms and 6b hetero atoms. This compound was observed to be categorised under the anodyne library by the database and its analogs belong to the category of natural products. At last, ZINC000257349595 screened against the modelled soybean profilin was having a molecular weight of 352.471 and chemical formula $C_{20}H_{32}O_5$. The pharmacophore was having 4 rings and 25 heavy atoms. The pharmacophore was categorised under anodyne library by the database. Previously no data has been reported pertaining to the application of these pharmacophores and thus the present investigation through computationally was able to recognise their activity against the modelled allergen profilins (apple, pineapple, wheat, and soybean).

6.1.5 Molecular docking analysis of the selected profilins against screened pharmacophores

The observed activity of the screened pharmacophores allowed us to explore their molecular-level interactions with the modelled food allergen profilins from apple, pineapple, wheat, and soybean. The molecular docking programs allows the user to anticipate non-covalent interactions arising by virtue of the docked conformations between the receptor and its pharmacophore (Trott and Olson, 2010). In these algorithms the receptor molecules are deduced as fixed entity except for those regions which are capable of any rotation (Trott and Olson, 2010). The docked conformations of the receptor and ligand molecule is achieved by minimum energy space of the conformation, shape of the conformation and thermal factor (Trott and Olson, 2010).

Another major aspect of docking algorithms is to optimally attain the docked conformation accuracy while utilising the lesser computational time (Trott and Olson, 2010). AutoDock Vina is one such algorithm with above mentioned features, was utilised in the present investigation to reveal the molecular interactions among the receptor food allergen profilins (apple, pineapple, wheat, and soybean) and screened pharmacophores against them.

The visualisation of docked conformations of the profilins from apple, pineapple, wheat, and soybean against the screened pharmacophores were shown in Figure 6.5. Further, binding energy scores and the amino acids of the profilins taking part in the non-covalent interactions with pharmacophores are shown in Table 6.3. The docked conformation of the modelled apple allergen profilin and its screened pharmacophore were having -8.0 kcal/mol of the binding potential. Various non-covalent interactions were observed in the docked complex including hydrogen bond formation at Glu107 and Glu108 residue number of the receptor profilin protein. Hydrophobic interactions were observed at position Pro40 and Leu110 of the apple profilin receptor molecule. Binding potential of -7.5 kcal/mol was observed at docked conformation of pineapple profilin and its pharmacophore. In this docked conformation, hydrophobic interactions were observed at sites Lys43, Glu46, Ala49, Leu60 and Tyr66 of the pineapple profilin. Binding energy of -7.5 kcal/mol was observed with docked conformation of modelled wheat profilin and its screened pharmacophore. In the docked conformation, hydrophobic interactions were observed at Glu45, Glu46, Ala48, Lys52, His59, Leu60, Thr63 and Phe66 positions of the wheat profilin receptor respectively. Salt bridge was also observed in the docked conformation at the position Glu46 of the wheat profilin receptor molecule. The docked conformation of the soybean profilin with its screened pharmacophore exhibited -10.0 kcal/mol of the binding energy. Hydrogen bonds in the docked conformation was observed at Asp53, Gly58, Gly77 and Gly80 position of the soybean profilin receptor molecule. Hydrophobic interactions were observed in all the docked conformations. At the same time, hydrogen bonds were seen in the docked conformations of apple and soybean profilins, and lastly, a salt bridge was observed in the docked conformation of wheat profilin. Interestingly the pharmacophores were observed to be involved with non-

covalent interaction at GLU46 and LEU60 position of the profilins from pineapple and wheat respectively. Various non-covalent interactions observed in the docked conformations of profilins are responsible for their strong binding energy potential and thus qualifies for the bioavailability profiling of associated pharmacophores.

Mal d 4 Vs ZINC000524729534 Ana c 1 vs ZINC000000041632 Tri a 12 vs ZINC000065529251 Gly m 3 Vs ZINC000257349595

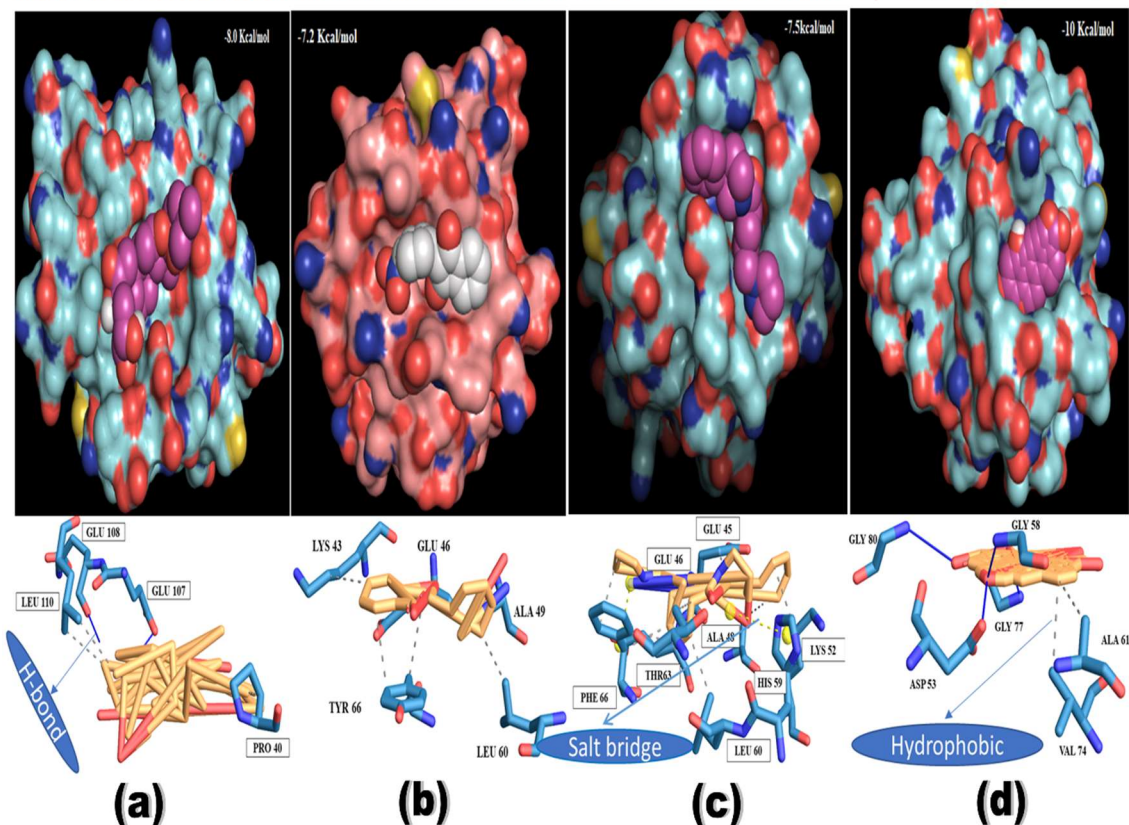


Figure 6.5 Docked conformations of the allergen profilins from apple, pineapple, wheat, and soybean along with their non-covalent interactions are referred to (a), (b), (c) and (d) respectively. Non-covalent interactions namely hydrogen bond, hydrophobic interactions and salt bridge are denoted by blue solid line, silver dotted line and yellow dotted lines respectively.

Table 6.3 Interaction map of the best docked conformations of profilins from apple, pineapple, wheat, and soybean. Various interactions are represented by coded superscripts such as residues involved in hydrogen bond formation are represented by H, hydrophobic interactions by h, salt bridges by S.

<i>Docked Conformation</i>	<i>Binding Energy (Kcal/mol)</i>	<i>Interacting Residues</i>
<i>Malus domestica (Apple) vs ZINC000524729534</i>	-8.0	Pro40 ^h , Glu107 ^H , Glu108 ^H and Leu110 ^h
<i>Ananas comosus (Pineapple) vs ZINC000000041632</i>	-7.2	Lys43 ^h , Glu46 ^h , Ala49 ^h , Leu60 ^h and Tyr66 ^h
<i>Triticum aestivum (Wheat) vs ZINC000065529251</i>	-7.5	Glu45 ^h , Glu46 ^S , Ala48 ^h , Lys52 ^h , His59 ^S , Leu60 ^h , Thr63 ^h and Phe66 ^h
<i>Glycine max (Soybean) vs ZINC000257349595</i>	-10.0	Asp53 ^H , Gly58 ^H , Ala61 ^h , Val74 ^h , Gly77 ^H and Gly80 ^H

6.1.6 Bioavailability analysis of the pharmacophores

The bioavailability values of the docked pharmacophores were calculated by the SwissADME and admetSAR web servers to account for their compatibility in the form of oral dosage. According to Veber and Lipinski rule of five (LRO5), the parameters like drug-likeness, the number of hydrogen bond donor, the number of hydrogen bond acceptor, topological polar surface area, molecular weight, human intestinal absorption, and Abbott bioavailability score were determined as shown in

Table 6.4. LRO5 allows the user to screen potential pharmacophores in accordance with their physicochemical descriptors to behave as orally bioavailable molecule (Lipinski *et al.*, 1997). This rule considers physicochemical properties and provides with a range of these descriptors which helps in screening suitable pharmacophores. Features like molecular weight, number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA) and partition coefficient are considered (Lipinski *et al.*, 1997). As per the rule, an ideal orally bioavailable pharmacophore should have molecular weight less than 500. Number of HBD and HBA should be less than five and ten respectively for orally bioavailable molecule. The partition coefficient of a molecule refers to its potential to enter the cell membrane (Ahuja and Scypinski, 2001). The partition coefficient is calculated by measuring the amount of the pharmacophore in equal ratio of organic and aqueous solvent after the state of equilibrium has been attained (Ahuja and Scypinski, 2001). The acceptable value of orally bioavailable pharmacophore with reference to LRO5 should be less than 5 (Ahuja and Scypinski, 2001). The higher partition coefficient of a pharmacophore signifies its unchallenging potential to enter the cell membrane (Ahuja and Scypinski, 2001). Veber had examined that pharmacophore with less than ten rotatable bonds and polar surface area of less than or equal to 140 \AA^2 , were observed as suitable candidates to behave as orally bioavailable pharmacophores (Veber *et al.*, 2002). Studies had shown that bioavailability of a pharmacophore is primarily driven by the human intestinal absorption process (Hou *et al.*, 2007). This suggests the need to evaluate the HIA potential of the candidate pharmacophores and thus in the present scenario the HIA indices of all the screened pharmacophores were evaluated. Another score to evaluate the bioavailability of the pharmacophore is Abott Bioavailability score (Martin, 2005). The score allows the user to predict that whether a test compound will have bioavailability of more than ten percent in rats (Martin, 2005). The study pointed out that charge of the pharmacophore at human physiological pH drive the bioavailability profile of the respective molecule (Martin, 2005). Taking these into account, all the above discussed parameters were evaluated for the screened pharmacophores and their evaluation is discussed below.

The pharmacophore ZINC000524729534 screened against the modelled allergen profilin from apple was having six and three HBA and HBD respectively. The molecular weight of the pharmacophore was 462.62 Daltons, with no rotatable bonds. The topological polar surface area of the compound was calculated to be 96.22 Å². Moreover, the compound was observed to HIA positive with index value of 0.81 and accepting the conditions of LRO5. The Abott bioavailability score of the pharmacophore was 0.55 suggesting the compound to be a good candidate to behave as orally bioavailable drug. Next, the pharmacophore ZINC000000041632, screened against the modelled pineapple profilin, had the molecular weight of 267.24 with topological polar surface area of 83.12 Å². Number of HBA and HBD in the pharmacophore were four and one respectively with presence of two rotatable bonds. The HIA index for the pharmacophore was evaluated as positive with index of 1.00 suggesting the compound to be highly suitable to behave efficiently in the form of oral dosage. The Abott bioavailability score of 0.56 also signify the pharmacophore to be good candidate to act as orally bioavailable drug and further these suggestions were consolidated by the point that it was also satisfying the parameters of LRO5. The pharmacophore ZINC000065529251, screened against the modelled allergen profilin from wheat had molecular weight of 347.41 Daltons and topological polar surface area of 58.36 Å². The number HBA and HBD in the pharmacophores were observed to be three and One respectively with presence of 5 rotatable bonds. The HIA index was observed to be positive for the compound with index of 1.00 signifying the compound to be highly preferred to act as orally bioavailable drug. The Abott bioavailability score for the respective pharmacophore was 0.55 and was satisfying all the parameters of LRO5 to perform as orally bioavailable drug. At last, the compound ZINC000257349595, screened against the modelled allergen profilin from soybean possessed molecular weight of 352.47 Daltons and topological polar surface area of 97.99 Å². The number of HBD and HBA in the pharmacophore were four and five respectively with presence of two rotatable bonds. The positive HIA index of the pharmacophore with value 0.98 suggests oral bioavailability potential of the compound. The Abott bioavailability score of 0.56b and fulfilling the conditions of LRO5 signifies the compound's suitability to behave as orally bioavailable drug. Conclusively, HIA values of the pharmacophores in the order of 0.81, 1.00, 1.00, 0.98

for apple, pineapple, wheat, and soybean depicts their potential to act as orally bioavailable drug and all the pharmacophores were also satisfying the conditions of Veber's rule and Lipinski's rule of five to serve as drug-like molecules. Abbott bioavailability scores for all the pharmacophores also depict them as strong drug-like molecules.

Table 6.4 Bioavailability analysis of the pharmacophores ZINC000524729534, ZINC000000041632, ZINC000065529251 and ZINC000257349595 respectively.

<i>Parameters</i>	<i>ZINC000524 729534</i>	<i>ZINC000000 041632</i>	<i>ZINC000006 5529251</i>	<i>ZINC00025 7349595</i>
<i>Number of H-bond donor (HBD)</i>	3	1	1	4
<i>Number of H-bond acceptor (HBA)</i>	6	4	3	5
<i>Number of rotatable bonds</i>	0	2	5	2
<i>Topological Polar Surface Area (Å²)</i>	96.22	83.12	58.36	97.99
<i>Molecular Weight (Daltons)</i>	462.62	267.24	347.41	352.47
<i>Human Intestinal Absorption</i>	0.81 (HIA+)	1.00 (HIA+)	1.00 (HIA+)	0.98 (HIA+)

<i>Drug likeness (Lipinski's Rule of five)</i>	YES	YES	YES	YES
<i>Abbott bioavailability score</i>	0.55	0.56	0.55	0.56

6.1.7 Differential amino acid usage analyses of profilin family

Amino acid usage analysis in the present investigation was carried out to explore whether there exists any difference in the amino acid utilisation patterns of the considered allergen and non-allergen profilin instances. The investigation will provide insight towards the relative amino usage patterns exhibited by the corresponding profilin allergen family whose differential behaviour, if observed, will further lead to explore the basis of this differential usage patterns in the associated family.

6.1.7.1 CoA based on the RAAU data of the profilin family

The present investigation is first of its kind to highlight differential patterns of amino acid usage patterns exhibited by the allergen and non-allergen profilins. Correspondence analysis (CoA) based on relative amino acid usage (RAAU) data was performed to explore the variations in the amino acid usage patterns among the profilin gene family. The principal axes of separation, Axes 1 and 2 of RAAU data, was used to generate the CoA. It was evident from our analysis (Figure 6.6) that the allergen and non-allergen profilins produced discrete clusters, signifying differential amino acid usage patterns. Moreover, the amino acids, namely methionine, proline, histidine, glutamine, glutamic acid, tryptophan, and glycine were more frequently represented among the allergen profilins as compared to the non-allergens ($P < 0.05$) (Table 6.5). The observed differential amino acid usage patterns among the allergen and non-allergen profilins form the foundation of our investigation to explore whether these patterns influence their physicochemical and structural properties.

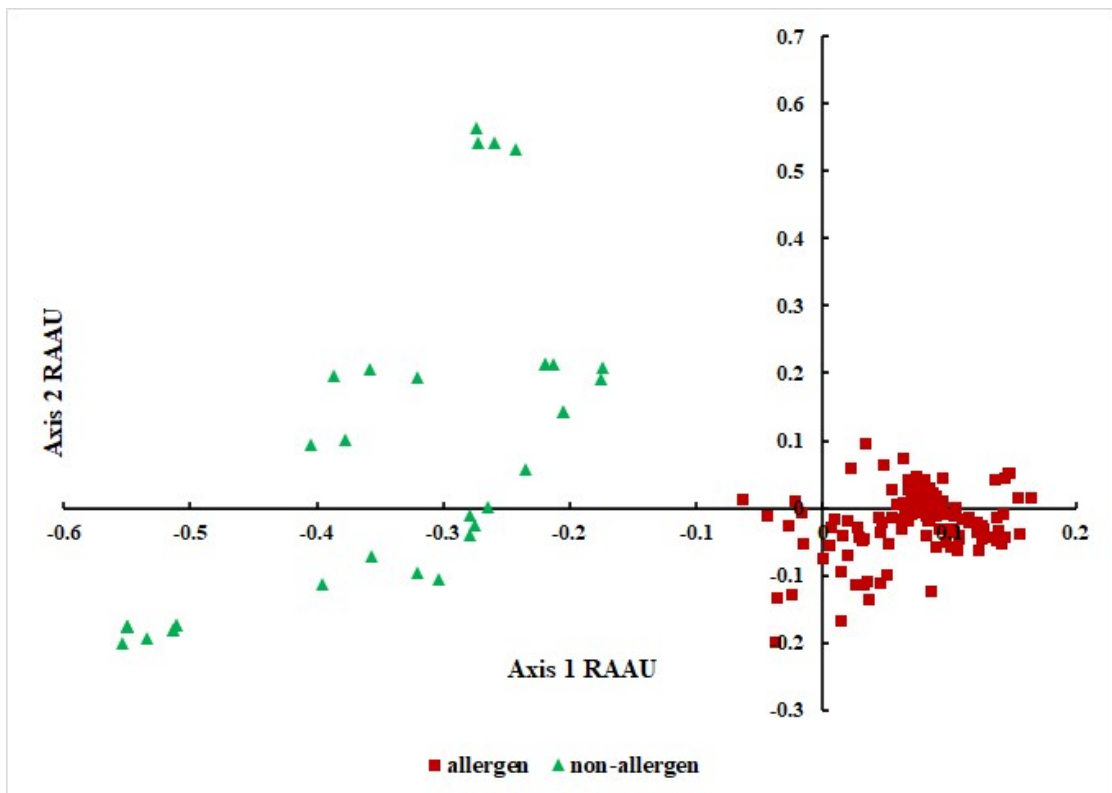


Figure 6.6 Correspondence analysis of the profilin gene family plotted against Axis 1 and Axis 2 of RAAU data. The allergen instances are represented as red squares, whereas non-allergens are denoted as green triangles.

Table 6.5 Over/under-represented amino acids based on normalised RAAU of the allergen and non-allergen profilins.

<i>Amino Acid</i>	<i>RAAU allergen</i>	<i>RAAU non-allergen</i>	<i>Significance</i>	<i>Amino Acid</i>	<i>RAAU allergen</i>	<i>RAAU non-allergen</i>	<i>Significance</i>
<i>Methionine</i>	5.99	3.58	P < 0.05	<i>Valine</i>	8.81	10.55	P < 0.05
<i>Proline</i>	6.53	4.11	P < 0.05	<i>Serine</i>	4.48	6.47	P < 0.05

<i>Histidine</i>	4.73	2.47	P < 0.05	<i>Threonine</i>	7.77	10.02	P < 0.05
<i>Glutamine</i>	7.17	2.91	P < 0.05	<i>Alanine</i>	9.78	12.86	P < 0.05
<i>Glutamic acid</i>	9.21	5.16	P < 0.05	<i>Asparagine</i>	2.60	6.52	P < 0.05
<i>Tryptophan</i>	2.00	1.44	P < 0.05	<i>Lysine</i>	6.67	8.13	P < 0.05
<i>Glycine</i>	18.76	11.75	P < 0.05	<i>Arginine</i>	2.52	7.02	P < 0.05

6.1.7.2 Physicochemical basis of differential amino acid usage patterns in profilin gene family

The differential amino acid usage patterns observed among the allergen and non-allergen profilins forms the foundation to investigate whether these distinct usages influence their physicochemical properties or not. Interestingly, features like disorder, solubility and trypsin digestion were found to be critically important among the allergen and non-allergen profilins. The allergen proteins possess distinguishable physicochemical features which make them to be recognised as foreign molecules by the host immune system and subsequently the individual gets sensitized and develop immune response against them (Xue et al. 2011).

6.1.7.2a Profilin disorder

Disorder in proteins refers to the state in which they do not have stable tertiary structures, yet are biologically active at physiological environment (Xue et al., 2011).

Previous studies have shown that allergen proteins are highly disordered and

contribute to be a part of the allergen representative peptides (ARPs) (Xue *et al.*, 2011). The protein disorder of the allergen and non-allergen profilins were found to be 19.13% and 5.32%, respectively, which revealed that allergen profilins were highly disordered. It was evident from our correlation analysis that protein disorder exhibited a strong positive correlation ($r = 0.680$) with Axis 1 of RAAU ($P < 0.01$) (Table 6.6). Therefore, presence of disordered amino acid residues in allergen profilins, as revealed from the present study, may contribute to the immune responses and subsequent hypersensitive reactions among atopic individuals.

6.1.7.2b Solubility

Solubility of a protein refers to its potential to get dissolved into the solution (Zayas 1997). Protein solubility is principally affected by various factors including the sequence of amino acids that constitute the protein (Zayas, 1997). The average solubility index of the allergen profilins (0.63 ± 0.04) was found to be relatively higher than that of the non-allergens (0.57 ± 0.07) ($P < 0.01$). A significant positive correlation ($r = 0.348$) ($P < 0.01$) (Table 6.6) of solubility with Axis 1 of RAAU data revealed a contribution of solubility index in producing the distinct clusters among the allergen and non-allergen profilins, based on RAAU data (Figure 6.6). Our observation is consistent with the fact that high solubility of allergens is implicated in activation of immune responses and triggering of allergic responses (Pekar and Untersmayr, 2018).

6.1.7.2c Trypsin Digestion of profilins

Trypsin is one of the several digestive enzymes involved in the gastrointestinal digestion of proteins found in food (Pekar and Untersmayr, 2018). The sensitisation of allergens in the gastrointestinal tract requires them to be resistant against various proteolytic enzymes like trypsin (Pekar and Untersmayr, 2018). This resistance stimulates the host to pose an immune response (Pekar and Untersmayr, 2018). It has been suggested that allergenic proteins possess higher potential to escape gastrointestinal digestion (Pekar and Untersmayr, 2018; Breiteneder and Mills, 2005). Our analysis revealed that allergen profilins were less prone to trypsin digestion (8.24 ± 0.93) in comparison to the non-allergen profilins (14.44 ± 2.59). Trypsin digestion was found to display a strong negative correlation ($r = -0.822$) with Axis 1 of RAAU

($P < 0.01$) (Table 6.6), which signified that this factor is an imperative determinant in producing the discrete clusters among the allergen and non-allergen profilins, based on RAAU data (Figure 6.6). Previous studies have reported a decrease in the IgE binding capacity of peanut (reduced by 100 times) and soybean (reduced by 10 times) upon their digestion with human digestive enzymes (pepsin, trypsin, chymotrypsin, and peptidase of intestine) (Burks *et al.*, 1992). Moreover, the celery food allergic patients with decreased gastrointestinal acid exudation, when provided with digestive-enzyme, processed celery and resulted in reduced IgE binding, cross-linking capability, and allergic responses (Untersmayr *et al.*, 2008). It is worth mentioning that altered gastrointestinal digestive enzyme activity contributes to increased IgE reactivity of allergenic food proteins (Pali-Schöll *et al.*, 2018). In addition, factors like improper food allergen protein digestion due to increased gastric pH or decreased digestion ability, presence of specific T-helper type 2 adjuvants, and altered microbiome of digestive system are key contributors of allergic responses to food (Pali-Schöll *et al.*, 2018).

Table 6.6 Pearson's correlation coefficient values of the physicochemical features of profilins with Axes 1 and 2 of RAAU data.

	<i>Amino Acid Disorder Usage</i>	<i>Solubility</i>	<i>Trypsin digestion</i>
<i>Axis 1 (RAAU)</i>	0.680**	0.348**	-0.822**
<i>Axis 2 (RAAU)</i>	-0.044	-0.135	0.302**

** indicates a significant correlation at 0.01 level (2-tailed)

6.1.8 Sequence analysis of the allergen profilins revealed conserved motifs

Sequence analysis of the allergen profilins was carried out by building their alignment profiles and subsequent screening of conserved motifs at 70% identity level. The sequence alignment has been shown in the Figure 6.7. The conserved motifs observed at 70% identity level revealed that the allergen profilins have identical motifs SWQ,

YVD, VWA, LAPTG, KYMVIQGE, VIRGKKG, KKT, GIY, PGQCNM and LGDYL which are represented by * in the Figure 6.7. The z-scores for the conserved motifs were observed as nan (not a number)/undefined, which denote completely conserved residues with a standard deviation value of zero. The empirical significance of the obtained alignment and conserved sites was estimated by z-score, which determines the deviation of the actual score from mean of the random scores generated by performing the randomizations of the actual alignment (Brandt *et al.*, 2010). The z-score for the conserved residues is marked by nan (not a number) signifying completely conserved sites (Appendix 2).

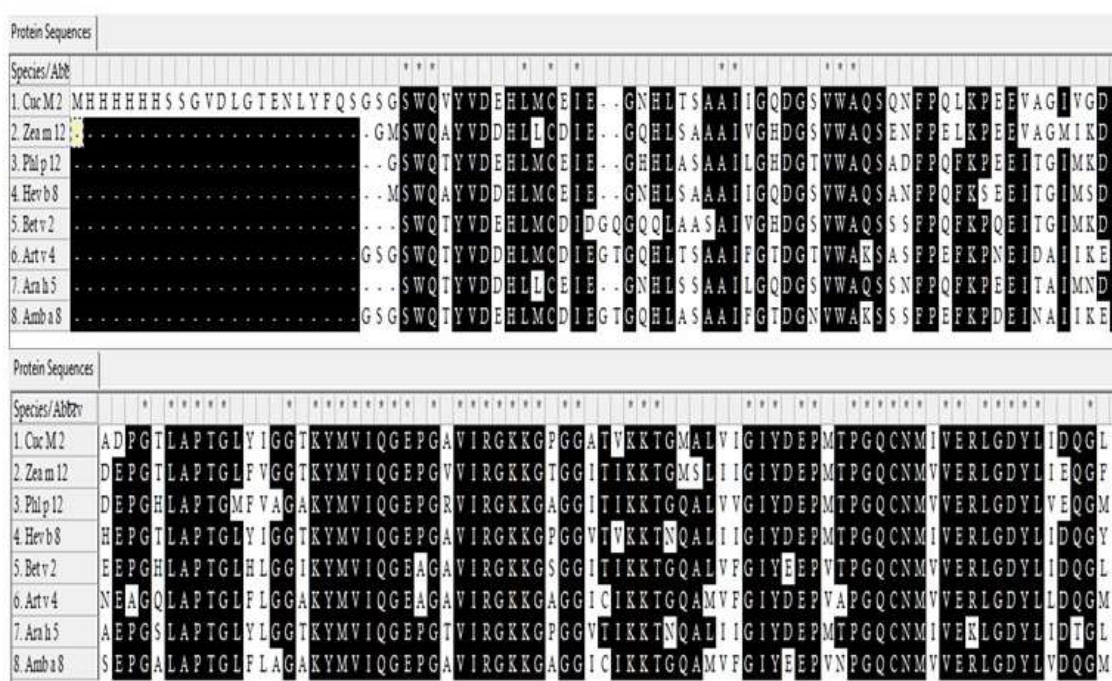


Figure 6.7 Multiple sequence alignment profile of the profilin allergens performed by MEGAX. The residues with background colour in black represent the conserved sites at 70% identity level among allergen profilins.

6.1.9 Structural analysis of the allergen profilins

The over-represented amino acids among the allergen profilins, in comparison to the non-allergens, are presented in Table 6.5. Majority of these over-represented amino acids have been observed to be surface exposed as shown in Table 6.7 and Figure 6.8. The amino acids namely Methionine (Met), Proline (Pro), Histidine (His), Glutamine

(Gln), Glutamic acid (Glu), Tryptophan (Trp), and Glycine (Gly) were observed to have propensity to remain surface exposed in the x-ray resolved structures of allergen profilins from *Hevea brasiliensis* (Hev b 8), *Artemisia vulgaris* (Art v 4), *Betula verrucosa* (Bet v 2), *Cucumis melo* (Cuc m 2), *Phleum pratense* (Phl p 12), *Zea mays* (Zea m 12), *Arachis hypogaea* (Ara h 5), and *Ambrosia artemisiifolia* (Amb a 8) respectively. The experimentally resolved structure of allergen profilin *Hevea brasiliensis* (Hev b 8) was observed to have propensity of amino acids (three letter code and its position) Met (73), Pro (57, 62, 79, 89, 109 and 112), His (55), Gln (4, 41, 76 and 99), Glu (14, 45, 78 and 108), and Gly (17, 58, 69, 88 and 130) to be surface exposed in the associated structure. The structure of *Artemisia vulgaris* (Art v 4) was having amino acids Met (75), Pro (46, 64, 111 and 114), Gln (78, 101 and 131), Glu (16, 43, 58, 80 and 110), and Gly (17, 19, 60, 90, 100, 115 and 132) exposed at the surface in their respective structure. The amino acids namely Met (119), Pro (46, 59, 64, 111 and 114), His (10), Gln (18, 20, 21, 43, 47, 101, 116 and 131), Glu (9, 57, 58, 80, 109 and 110), Trp (3), and Gly (17, 19, 79, 90, 100, 115 and 132) were observed to be surface exposed in the structure of *Betula verrucosa* (Bet v 2). In the structure of allergen profilin from *Cucumis melo* (Cuc m 2), the amino acids namely Met (73, 99 and 117), Pro (44, 57, 62, 79, 89, 109 and 112), His (10 and 19), Gln (37, 41 and 76), Glu (9, 14, 16, 45, 78 and 108), and Gly (17, 69, 88, 98, 113 and 130) were observed to be surfaced exposed respectively. The structural conformation of allergen profilin from *Phleum pratense* (Phl p 12), had amino acids Met (73), Pro (44, 57, 62, 79, 109 and 112), His (19), Gln (41, 76, 99 and 114), Glu (9, 14, 16, 45, 56, 108 and 128), and Gly (1, 17, 58, 69, 88, 98, 113 and 130) exposed in their resolved structure. The structure of allergen profilin from *Zea mays* (Zea m 12), had amino acids namely Met (73, 99 and 117), Pro (44, 57, 62, 79, 109 and 112), His (19), Gln (4 and 76), Glu (16, 37, 41, 45, 56, 78, 108 and 128), and Gly (17, 58, 69, 88, 98 and 130) surface exposed. The experimentally deduced crystal structure of allergen profilin from *Arachis hypogaea* (Ara h 5) was observed to have amino acids Met (73), Pro (44, 57, 62, 79, 89, 109 and 112), His (10 and 19), Gln (76 and 99), Glu (16, 45 and 108), Trp (3), and Gly (17, 58, 88, 113 and 130) exposed on their surface. The allergen profilin from *Ambrosia artemisiifolia* (Amb a 8) had amino acids namely Met (75), Pro (46, 64, 81, 111 and 114), His (21), Gln (78, 101 and 131), Glu (9, 16, 43,

58, 80, 109 and 110), and Gly (17, 19, 60, 71, 82, 92, 100, 115 and 132) exposed to their surface. Conclusively, the detailed structural analysis revealed that majority of the overrepresented amino acids in the allergen profilins were analysed to be surface exposed, which indicate their potential to interact with their surrounding cellular environment and thus place them in a position where they are more prone to be recognised by the cells of the host immune system.

Table 6.7 The surface exposed residues found in the allergen profilin in various organisms. These amino acids were over-represented by the allergen profilins on comparison to non-allergen profilins.

<i>Profilin</i>							
<i>Allergen</i>	<i>Methionine</i>	<i>Proline</i>	<i>Histidine</i>	<i>Glutamine</i>	<i>Glutamic</i>	<i>Tryptophan</i>	<i>Glycine</i>
<i>Source</i>	<i>(Position)</i>	<i>(Position)</i>	<i>(Position)</i>	<i>(Position)</i>	<i>Acid</i>	<i>(Position)</i>	<i>(Position)</i>
<i>Organism</i>					<i>(Position)</i>		
<i>Hevea brasiliensis</i> (<i>Hev b 8</i>)	Met (73)	Pro (57, 62, 79, 89, 109 and 112)	His (55)	Gln (4, 41, 76 and 99)	Glu (14, 45, 78 and 108)	-	Gly (17, 58, 69, 88 and 130)
<i>Artemisia vulgaris</i> (<i>Art v 4</i>)	Met (75)	Pro (46, 64, 111 and 114)	-	Gln (78, 101 and 131)	Glu (16, 43, 58, 80 and 110)	-	Gly (17, 19, 60, 90, 100, 115 and 132)
<i>Betula verrucosa</i> (<i>Bet v 2</i>)	Met (119)	Pro (46, 59, 64, 111 and 114)	His (10)	Gln (18, 20, 21, 43, 47, 101, 116 and 131)	Glu (9, 57, 58, 80, 109 and 110)	Trp (3)	Gly (17, 19, 79, 90, 100, 115 and 132)
<i>Cucumis melo</i> (<i>Cuc m 2</i>)	Met (73, 99 and 117)	Pro (44, 57, 62, 79, 89, 109 and 112)	His (10 and 19)	Gln (37, 41 and 76)	Glu (9, 14, 16, 45, 78 and 108)	-	Gly (17, 69, 88, 98, 113 and 130)
<i>Phleum</i>	Met (73)	Pro (44,	His (19)	Gln (41,	Glu (9,	-	Gly (1,

Figure 6.8 Surface exposed residues of allergen profilins depicted from their pdb structures by PyMol. The over-represented amino acids which were surface exposed are shown in red colour while rest of the protein is shown in grey colour as ribbons.

6.2 Machine learning (SVM and RF) tool development for prediction of food allergens

The exponential increase of the omics data by the advent of high-throughput approaches has opened the gates for implementation of data-mining based computational approaches. Data-mining based approaches has evolved the healthcare system by providing state-of-the-art data-driven solutions for disease diagnosis and thereby offering affordable treatment options (Jackins *et al.*, 2021). There has been a considerable progress in the area of machine learning assisted allergenicity assessment of query proteins. Various models and web servers had developed employing machine learning strategies to achieve optimum accuracy for the allergenicity assessment of query proteins (Saha and Raghava, 2006; Zhang *et al.*, 2007; Kumar and Shelokar, 2008; Dimitrov *et al.*, 2013; Wang *et al.*, 2013; Wang *et al.*, 2013a; Mohabatkar *et al.*, 2013; Dang and Lawrence, 2014; Dimitrov *et al.*, 2014; Wang *et al.*, 2021; Sharma *et al.*, 2021). After the extensive literature analysis of these developed models, we observed that selection as well as relevance of that descriptor influence the accurate allergenicity assignment of query proteins. For instance, we might have case where the model is giving-off very good accuracy but it might not be able to capture the entire allergenic classes. By taking these observations into account we performed simple but exhaustive approach to extract out features relevant to allergen proteins. The profilin allergen family was taken as the model class and subsequently its differential amino acid usage patterns were revealed as discussed in section 6.1.7. Further, the observed differential pattern was linked with those features which were associated with allergen proteins after their evaluation a difference in indices of those respective features were observed among allergen and non-allergen profilins. For the first we presented with an investigation where differential amino acid usage patterns findings were translated to extract significant features pertaining to allergen proteins. With this we further shifted towards utilisation of these features

in the form of descriptors to develop machine learning assisted allergenicity assessment model as discussed below.

6.2.1 Models developed employing WEKA classifiers

WEKA suite was used as a platform to generate various machine learning classifiers such as ZeroR, LibSVM and Random Forest respectively. The models were developed using manually curated dataset having equal ratio of binary classes followed by stratified 10- folds cross-validation.

The output of the developed models using the manually curated dataset employing stratified 10-fold cross validation was shown in Figure 6.9 and Table 6.8 respectively. The confusion matrix in Figure 6.9 depicts the classification accuracy of the developed model by considering all the classes in the dataset. For the developed ZeroR classifier, the confusion matrix depicts those 600 instances were accurately classified as allergens whereas 600 were classified as non-allergens. Further, the confusion matrix pertaining to LibSVM classifier signified those 353 instances were accurately classified as allergens whereas 157 non-allergen instances were classified as allergens. Similarly, this model misclassified 247 instances of allergens as non-allergens whereas 443 instances were correctly assigned as non-allergen class. In case of Random Forest classifier, two classes ‘a’ and ‘b’ in the confusion matrix were representing allergens and non-allergens respectively. The column ‘a’ in the confusion matrix of RF classifier signified those 521 instances were classified accurately as ‘a’, whereas 62 instances have been misclassified as ‘a’. On the other hand, column ‘b’ of the matrix signified those 79 instances were wrongly classified as ‘b’ and 538 instances were correctly assigned to class ‘b’. From the confusion matrix it is evident that out of all the developed models the random forest model performed comparatively superior in terms of accurate assignment of allergen and non-allergen classes. After this those developed models were evaluated by the help of associated classifiers in order to select a best model for further analysis.

=== Confusion Matrix ===	=== Confusion Matrix ===	=== Confusion Matrix ===
<pre> a b <-- classified as 600 0 a = allergen 600 0 b = non allergen </pre>	<pre> a b <-- classified as 353 247 a = allergen 157 443 b = non allergen </pre>	<pre> a b <-- classified as 521 79 a = allergen 62 538 b = non allergen </pre>
<i>ZeroR</i>	<i>LibSVM</i>	<i>Random Forest</i>

Figure 6.9 Confusion matrix for the ZeroR, LibSVM and Random Forest classifiers.

After this the developed ZeroR, LibSVM and RF models were evaluated based on number of accurately classified instances in the dataset, accuracy, TP rate, FP rate, Precision, Recall, F-measure, MCC value, and ROC curve area respectively. The RF model was able to correctly classify 1059 instances in the dataset which was highest as compared to ZeroR and LibSVM as they classified 660 and 796 instances of the dataset correctly. The RF model produced highest accuracy of 0.88, whereas LibSVM and ZeroR model provided accuracy of 0.66 and 0.50 respectively. The TP rate of 0.88 was highest by the RF model and ZeroR and LibSVM classifiers provided TP rate of 0.40 and 0.66 respectively. FP rate of 0.11 was lowest in case of RF model as ZeroR and LibSVM model gave FP value of 0.50 and 0.33 respectively. Further, precision value of 0.88 was highest of the RF model as compared to LibSVM model with 0.66 respectively. The recall value of 0.88 was also highest for the RF model as ZeroR and LibSVM model produced recall of 0.49 and 0.64 respectively. The F-measure index of 0.88 was highest for RF model as LibSVM model gave F-measure of 0.64. The MCC value of 0.76 was also highest for RF model as LibSVM model gave 0.33 value for MCC. At last RF model with ROC area of 0.95 was highest as compared to ZeroR and LibSVM model which gave ROC area of 0.50 and 0.66 respectively. From the above discussion it is evident that RF model outperforms all the developed classifiers in terms of the observed parameters and thus was opted for further analysis.

Table 6.8 Evaluation of the developed classifiers based upon various parameters.

<i>Parameters</i>	<i>ZeroR</i>	<i>LibSVM</i>	<i>Random Forest</i>
<i>Correctly Classified Instances</i>	600	796	1059
<i>Accuracy</i>	0.50	0.66	0.88
<i>TP rate</i>	0.40	0.66	0.88
<i>FP rate</i>	0.50	0.33	0.11
<i>Precision</i>	-	0.66	0.88
<i>Recall</i>	0.49	0.64	0.88
<i>F-measure</i>	-	0.63	0.88
<i>MCC</i>	-	0.33	0.76
<i>ROC Area</i>	0.50	0.66	0.95

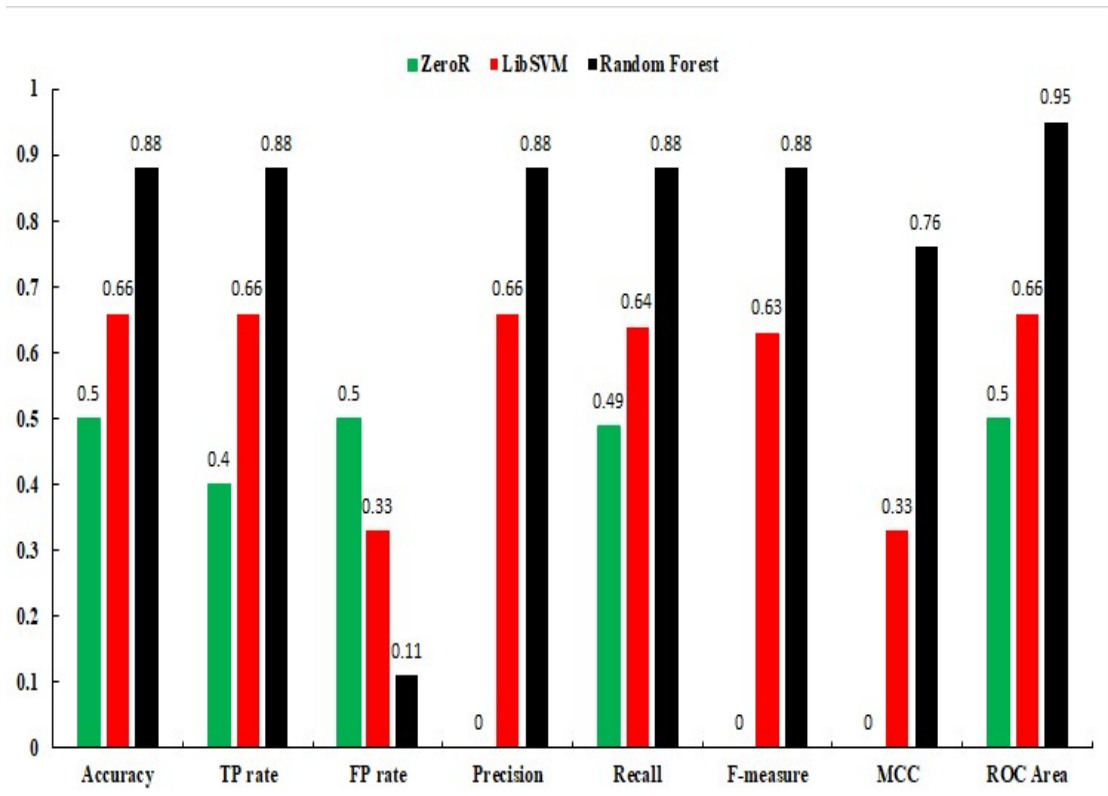


Figure 6.10 Clustered column based graphical evaluation of the ZeroR, LibSVM and RF classifiers.

The above representation clearly denotes that Random Forest model performed comparatively superior than others on the curated dataset by producing highest TP rate, precision, recall, F-measure, MCC value and ROC area respectively. The MCC value of 0.76 signified superior classification accuracy of the developed model.

6.2.1a ROC curve analysis of the developed models

The ROC curve is a two-dimensional representation between TP rate and FP rate (Centor, 1991). This technique has been extensively utilised for identifying the positive or negative samples in a test (Centor, 1991). The ROC curve for all the developed models has been shown in Figure 6.11. Area under ROC has been established as an effective measure which determines the probability of a model to correctly distinguish among the binary classes with higher area under ROC signifying more accurateness of the model in binary classification (0- misclassification, 0.5- random classification and 1- perfect classification) (Centor, 1991). Comparatively, area under ROC of the Random Forest model was observed to be highest with a value

of 0.95, signifying the perfection of the model in binary classification of the instances. Therefore, based on these parameters the random forest model was evaluated as most reliable model among all the developed classifiers.

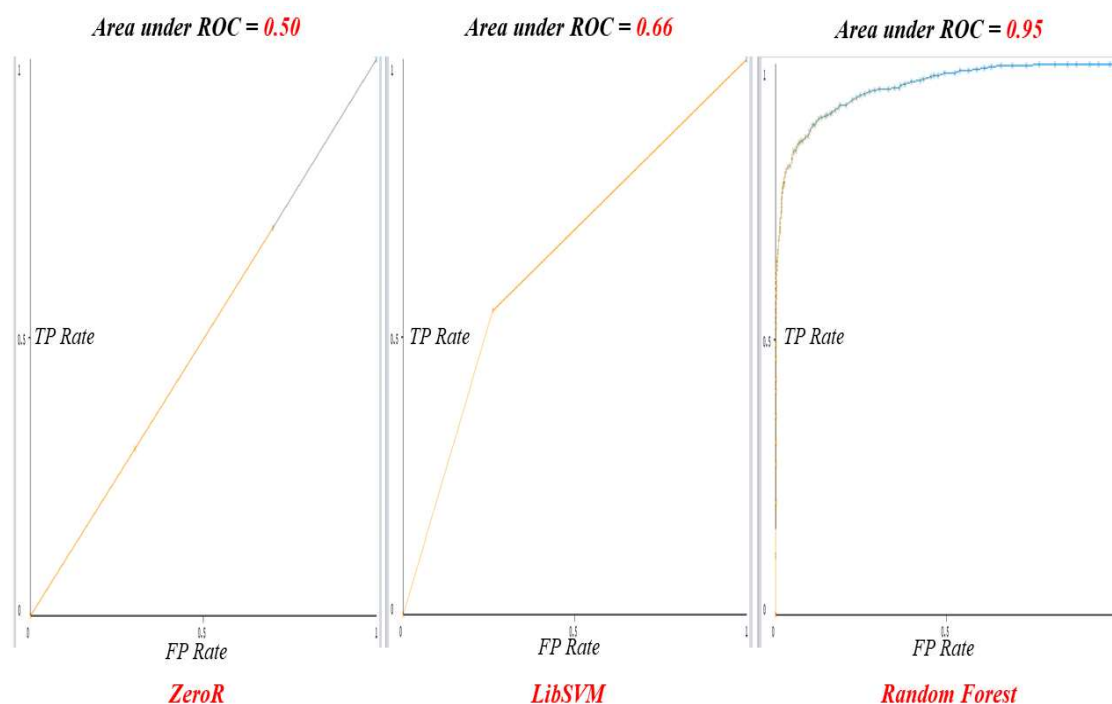


Figure 6.11 ROC curves for the ZeroR, LibSVM and Random Forest classifiers. The horizontal axis represents FP rate whereas vertical axis denotes TP rate.

6.2.2 Comparative evaluation of the developed model with existing classifiers

The developed model was evaluated comparatively with existing classifiers based on parameters such as sensitivity, specificity, accuracy, MCC and ROC curve respectively, as these have been extensively utilised for statistical evaluation of machine learning based classifiers in the associated disciplines (Saha and Raghava, 2006; Zhang *et al.*, 2007; Kumar and Shelokar, 2008; Dimitrov *et al.*, 2013; Wang *et al.*, 2013; Wang *et al.*, 2013a; Mohabatkar *et al.*, 2013; Dang and Lawrence, 2014; Dimitrov *et al.*, 2014; Wang *et al.*, 2021; Sharma *et al.*, 2021). The models developed by (Saha and Raghava, 2006; Wang *et al.*, 2013; Wang *et al.*, 2013a; Dimitrov *et al.*, 2014; Wang *et al.*, 2021; Sharma *et al.*, 2021) employing amino acid features-based model development. The first evaluator sensitivity corresponds to the magnitude of accurate positive class (allergen) prognostication by the classifier (Wang *et al.*, 2013;

Sharma *et al*, 2021). Sensitivity of the developed model (0.86) was observed as equivalent with existing models like AIGPred_AAC (0.88) (Saha and Raghava, 2006) and AIGPred2_AAC_SVM (0.87) (Sharma *et al*, 2021). Apart from this the sensitivity of the developed model was observed to be superior than considered models for evaluation. Specificity on the other hand signifies the proportion of accurately classified negative class instances (non-allergens). Specificity of the developed model (0.89) was highest from all the considered models except for PREAL (0.91) (Wang *et al.*, 2013a) and PREAL^w (0.94) (Wang *et al.*, 2013) algorithms. However, it may be noted that highest specificity of the PREAL^w model contributes to the application of similarity-based approaches which ultimately provides superior evaluator indices. Although these similarity-based methods provide excellent specificity but lacks in identification of allergens which possess poor sequence-similarity (Wang *et al.*, 2013). An instance was observed where “*alpha purothionin*” an allergenic protein in wheat was not classified as allergen by similarity-based method (FAO/WHO criteria), while other methods (motif based, descriptors based) were able to classify this instance as allergenic (Wang *et al.*, 2013). This certainly demanded an evolution in “*descriptor selection*” for computational based allergenicity assignment of query proteins. The next evaluator, accuracy is a measure of accurately classified positive and negative instances by the developed model (Wang *et al.*, 2013). Accuracy of the developed model (0.88) was found highest among all the existing classifiers signifying model’s ability to correctly assign both positive and negative instances. The MCC value is a measure of classification potential of the developed model whose indices ranges from -1 to 1, where -1 represents mis-classification, 0 represents random classification and 1 represents flawless classification potential of the model (Westerhout *et al*, 2019). The MCC value (0.76) of the developed model was highest amongst all the existing classifiers with second highest value of 0.70 by the AIGPred_AAC classifier. Finally, ROC area is a measure to evaluate prediction potential of the model by plotting a graph between TP rate and FP rate. The ROC area of 0.95 by the developed model was highest among all the considered classifiers signifying superior classification potential of the developed model. Thus, the above comparison of developed model outperforms the considered classifiers, signifying the

role of employed descriptors extracted by differential amino acid usage analysis in computationally assisted allergenicity assignment of query protein instances.

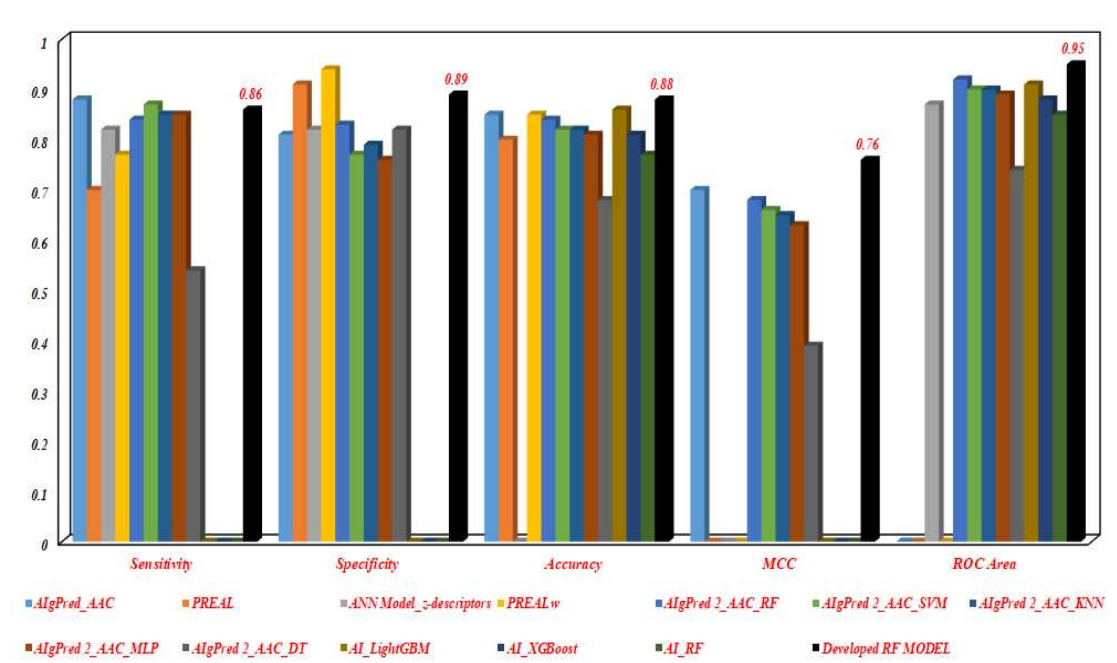


Figure 6.12 Comparative analysis of the developed RF model with existing classifiers. The horizontal axis represents the parameters namely sensitivity, specificity, accuracy, MCC value and ROC area respectively. The vertical axis represents the indices of the respective parameters. All the considered classifiers are coded by distinct colours respectively.

6.3 Application of the developed model/tool for genome-wide prediction of food allergen

Sesame mediated allergic reactions are on the rise globally and it has been regulated as allergen in 32 developed countries (Gangur and Acharya, 2021). Sesame induced allergic responses among atopic individuals presents array of responses (Gangur and Acharya, 2021). These responses include anaphylaxis and dermatitis induced by lipid allergens in sesame which are resulted from non-IgE mediated reactions (Gangur and Acharya, 2021). The protein allergens in sesame are responsible for IgE mediated responses resulting in anaphylaxis, diarrhoea, vomit, asthma, and respiratory system allergic responses. Apart from this, unidentified allergens in sesame seeds induces food-protein induced enterocolitis and eosinophilic esophagitis (Gangur and Acharya,

2021). On this account the present investigation was carried out to perform proteome wide allergenicity assessment of sesame to distinguish novel allergens.

To perform the genome wide prediction on developed Random Forest model, initially the reference genome of *Sesamum indicum* cultivar: Zhongzhi No. 13 was considered and subsequently the corresponding proteome having 24106 instances were retrieved as discussed under 5.3 subsection of the methodology.

The ARFF structured dataset of sesame (Figure 6.13) was further supplied to the developed model which classified 2381 instances as potential allergens. The log file of predictions carried out by RF model is shown in Figure 6.14. The pie chart for the prediction summary was shown in Figure 6.15.

Figure 6.13 was categorised into three part to comprehend the ARFF structured dataset of sesame. The first part named as dataset identifier represents the name of supplied dataset. Second part representing dataset attributes includes all the considered descriptors on which the RF model was developed. The last part numerical indices signified the calculated indices for each instance in the dataset whose values are separated by comma. At last, a question mark was placed at each instance which signified the unknown categorisation of the instance in the dataset.

```

@relation Food-allergens-dataset
@attribute Hydrophobicity numeric
@attribute Flexibility numeric
@attribute Helix-propensityaaindex numeric
@attribute Strand-propensityaaindex numeric
@attribute Size numeric
@attribute Relative-abundance numeric
@attribute D1 numeric
@attribute D2 numeric
@attribute D3 numeric
@attribute D4 numeric
@attribute D5 numeric
@attribute Type {allergen,'non allergen'}

@data
0.97563,0.43387,1.0273,0.97971,0.028804,-0.04191,1.3075,-0.91592,-0.0017959,0.99535,0.99902,?
0.96083,0.43816,1.0099,1.0208,0.03126,-0.041409,1.3078,-0.95912,-0.0074586,0.97807,0.98412,?
0.98172,0.43773,1.0179,1.0321,0.025253,-0.05046,1.3069,-0.98247,-0.0070402,0.96629,0.96595,?

```

Dataset Identifier

Dataset Attributes

Numerical indices of the dataset

Figure 6.13 ARFF structured dataset for the sesame proteome.

=== Predictions on test set ===

inst#	actual	predicted	error	prediction
1	1:?	2:non allergen		0.87
2	1:?	2:non allergen		0.89
3	1:?	2:non allergen		0.83
4	1:?	2:non allergen		0.92
5	1:?	2:non allergen		0.86
6	1:?	2:non allergen		0.78
7	1:?	2:non allergen		0.69
8	1:?	2:non allergen		0.77
9	1:?	2:non allergen		0.93
10	1:?	2:non allergen		0.64
11	1:?	1:allergen	0.75	
12	1:?	1:allergen	0.64	
13	1:?	2:non allergen		0.9
14	1:?	2:non allergen		0.76
15	1:?	2:non allergen		0.88
16	1:?	2:non allergen		1
17	1:?	1:allergen	0.81	
18	1:?	2:non allergen		0.8
19	1:?	2:non allergen		0.76
20	1:?	2:non allergen		0.8
21	1:?	2:non allergen		0.88
22	1:?	2:non allergen		0.94
23	1:?	2:non allergen		0.64
24	1:?	1:allergen	0.59	
25	1:?	1:allergen	0.9	
26	1:?	2:non allergen		0.55

Figure 6.14 RF model predictions log file on the sesame dataset.

Denotations in Figure 6.14

inst#: denotes the serial number of the instance in the sesame dataset.

actual: assigned class for the instance, ‘?’ was assigned initially as we want perform the predictions.

predicted: predicted class by the model (1- allergen, 2- non-allergen)

error: error in the predictions.

prediction: denotes the prediction probability of the assigned class.

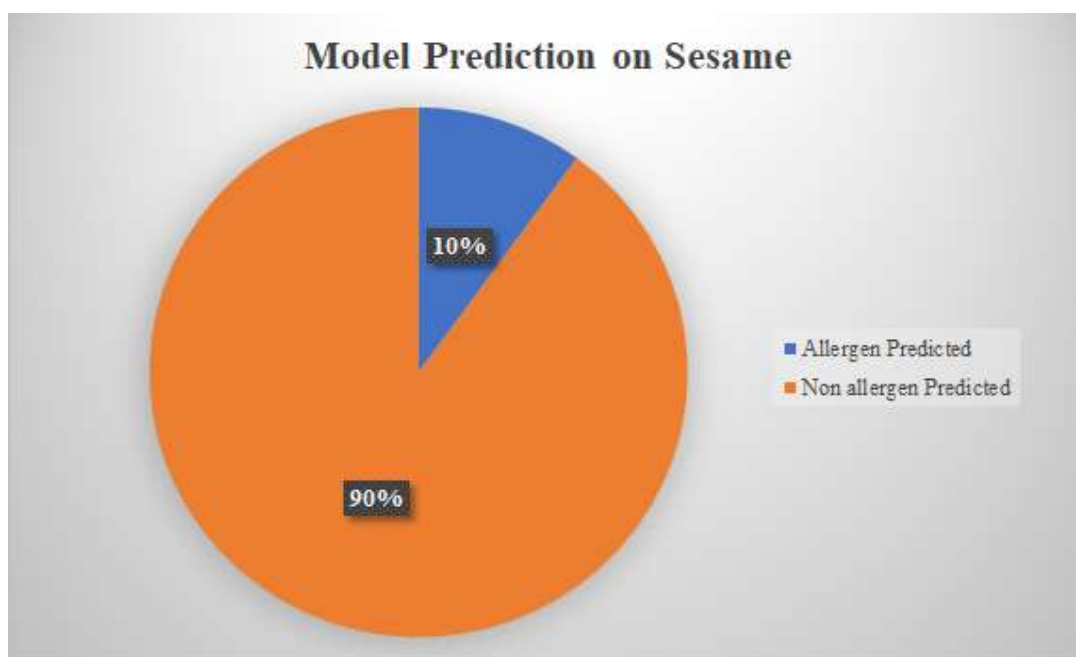


Figure 6.15 Pie chart for the overall prediction summary of sesame proteome. Different colours were used to distinguish the predicted allergen and non-allergen classes. The proportion of predicted allergens in the sesame proteome were represented by blue colour whereas non-allergens denoted by orange colour respectively.

Figure 6.15 clearly shows that around 10 percent of the entire sesame dataset was classified as potential allergen by the RF model. The developed model was able to correctly classify all known allergens in sesame namely oleosin H1 (Ses i 4), oleosin

L (Ses i 5), 11S globulin seed storage protein 2 (Ses i 6) and, 2S seed storage protein 1 (Ses i 2) in sesame with prediction probability values of 0.75 (inst#11), 0.81 (inst#17), 0.59 (inst#24) and, 0.90 (inst#25) respectively. The instances of profilin protein in the sesame proteome are automatically annotated by virtue of homology approach. The profilin instances namely inst#13622, inst#14113 and, inst#20609 were assigned as allergen with prediction probability of 0.69, 0.76 and, 0.77 respectively by the RF model. Eosinophilic esophagitis conditions among the atopic individuals are exponentially growing globally which results in conditions like dysphagia, puke and difficulty in growth (Spergel and Aceves, 2018). Food and aeroallergens are the inducers of this conditions allowing entry of eosinophil, basophil, mast cell and T-cells (Spergel and Aceves, 2018). Type-2 T-helper cells mediate these conditions whose dynamics were observed to be similar with conditions like asthma, allergic rhinitis, and atopic dermatitis (Spergel and Aceves, 2018). A case study carried out in 2016 reported the association of pan-allergens sensitisation among eosinophilic esophagitis patients (Hogan *et al.*, 2016). The study revealed that more than 85% of the subject patients (n = 66) were having the pan sensitisation to air-borne allergens and these patients were susceptible to be sensitised by exceptional foods (including nuts and legume) by virtue of the presence of pan-allergen profilin and PR-10 (Hogan *et al.*, 2016). With this we would like to suggest that unrecognised profilin allergen in sesame may be a contributor to these conditions however further investigations are required to consolidate these findings. The main objective behind this was to correlate the unidentified allergens mediated eosinophilic esophagitis conditions by sesame in associated patients with the computationally identified profilin pan allergen in sesame by the present investigation.

In a nut-shell, the RF model was able to not only classify the already identified allergens in sesame but also the profilin protein instances as allergenic with higher prediction probability. The observed classification potential of the RF model also corresponds to the relevancy of the descriptors employed. Additionally, among all the allergen predicted instances in sesame, 455 proteins were identified as uncharacterised. Uncharacterised proteins are those proteins whose structural,

functional, associational, and expressional analysis have not been evaluated (Hawkins and Kihara, 2007).

The prediction probability range for the allergen classified instances was analysed in terms of number of instances belonging to a particular range of the query proteome. After the evaluation it was observed that 1000 allergens were predicted in the prediction probability range of 0.50 to 0.59. There were 691 predicted allergen instances which had prediction probability range of 0.60 to 0.69. Further 361 instances were observed to be categorised under the prediction probability range of 0.70 to 0.79. There were 232 predicted allergen instances lied with in the prediction probability range of 0.80 to 0.89. At last, 97 allergen predicted instances had prediction probability range of 0,90 to 1.00. The graphical representation for the prediction probability range-based classification of allergen instances was shown in Figure 6.16.

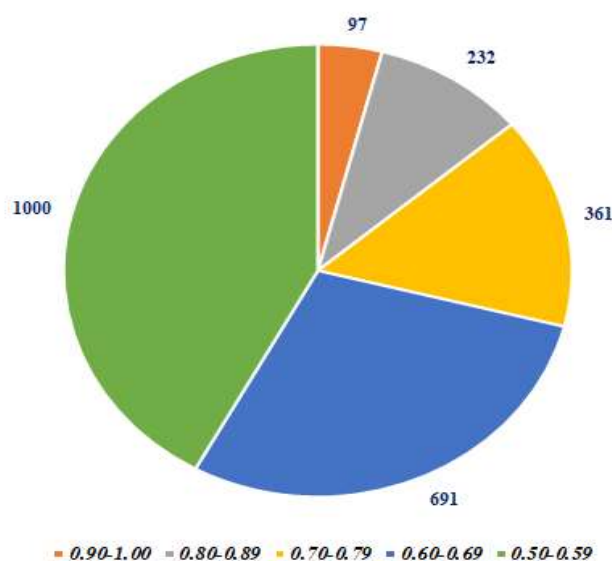


Figure 6.16 Number of instances classified as allergens based on prediction probability range for sesame proteome dataset. Green colour was used to denote the prediction probability range of 0.50 0.59, whereas, blue colour signified the prediction probability range of 0.60 to 0.69. Yellow colour denoted the number of allergens predicted instances under prediction probability range of 0.70 to 0.79, whereas grey colour denoted the prediction probability range of 0.80 to 0.89. Orange colour

signified the number of allergens predicted instances under the prediction probability range of 0.90 to 1.00.

6.4 Expression validation of few food allergens by RNA-Seq data analysis

The concept of mapping the transcriptional data to their respective proteome has been well established (Vogel and Marcotte, 2012; Chan *et al.*, 2015; Ogburn *et al.*, 2017; Karnaneedi *et al.*, 2020). A study conducted recently, had employed the transcriptomic data from the various shrimp species to measure the expression levels of their corresponding allergen proteins (Karnaneedi *et al.*, 2020). Taking the studies of (Chan *et al.*, 2015; Ogburn *et al.*, 2017; Karnaneedi *et al.*, 2020) into account we have performed the expression analysis of the transcript from sesame to distinguish new allergens and dwell into their corresponding expression profiles.

Transcriptomic data of indigenous *Sesamum indicum* having accession ID-SRR12153208 was retrieved from NCBI-SRA toolkit and subjected to expressional analysis.

6.4.1 Evaluation by the FastQC

The FastQC report of the raw transcriptomic data is shown in Figure 6.17. Part 1 of the figure represents basic statistics related to the sequenced data are presented in which total sequences were 15840574 with uniform length of 151 bases for each sequence and 47 % of GC content was observed. The part 2 of the figure representing per sequence quality score was evaluated by the Phred score (Ewing *et al.*, 1998). The score allows the user to evaluate the accuracy of the nucleotide calling by the sequencing machine (Ewing *et al.*, 1998). The Phred score generally ranges from 2 to 40 with higher indices signifying more accurate nucleotide calling (Ewing *et al.*, 1998). The average Phred quality score per read was observed to be 36, which corresponds to the range of best nucleotide call accuracy. In part 3 of the Figure 6.17 adapter content of the raw data was observed and it was found that there is presence of adapters in the data at positions 120-139 which needs to be resolved before further analysis with data.

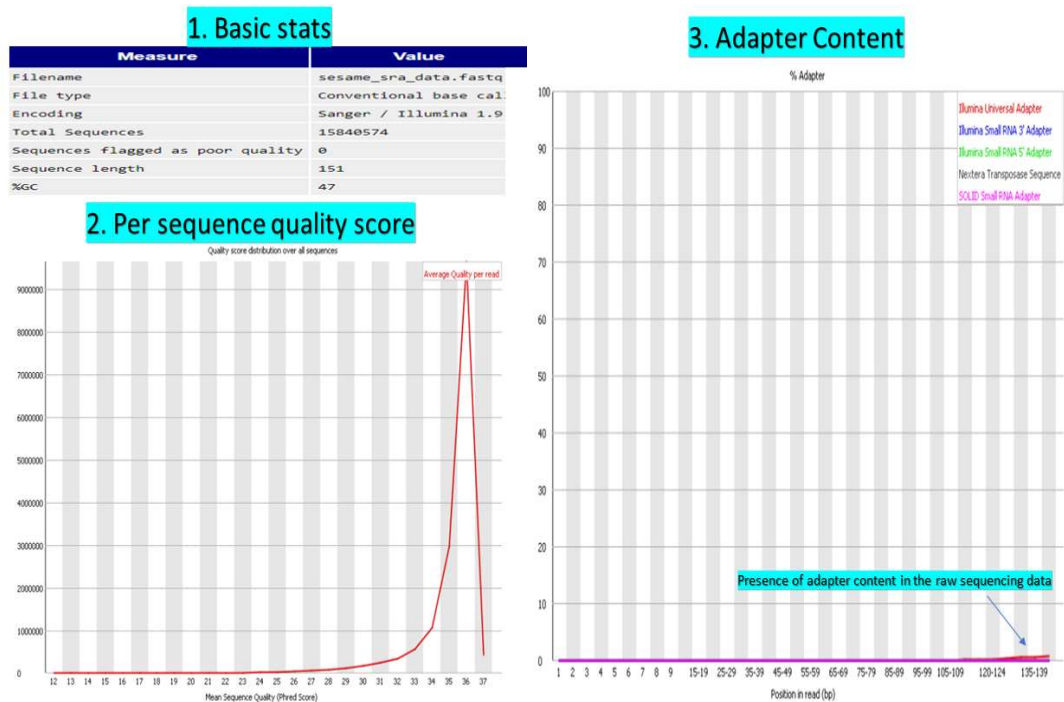


Figure 6.17 FastQC report of the raw sequencing data. The report corresponds to three parts including basis stats, per sequence quality score and adapter content respectively.

6.4.2 Trimmomatic analysis for the removal of adapter content from the transcriptomic data

The execution and completion window for the Trimmomatic program was shown in Figure 5.1 of the methodology section. The output of program showed that total 15840000 reads were provided as input out of which 1.22 % (192957) of the reads were filtered by the program and 15647043 reads constituted the processed dataset.

In order to validate the observed exclusions from the raw data FastQC was employed on the processed data whose graphical report have been shown in Figure 6.18. Part 3 of the Figure 6.18 indicated that adapter content of processed data showed a flat line which corresponds to removal of the adapter content and hence can be processed to further analysis. Further this leads to the formation of variable length of the sequencing reads ranging from 36-151 bases as shown in part 1 (basic stats) of the Figure 6.18.

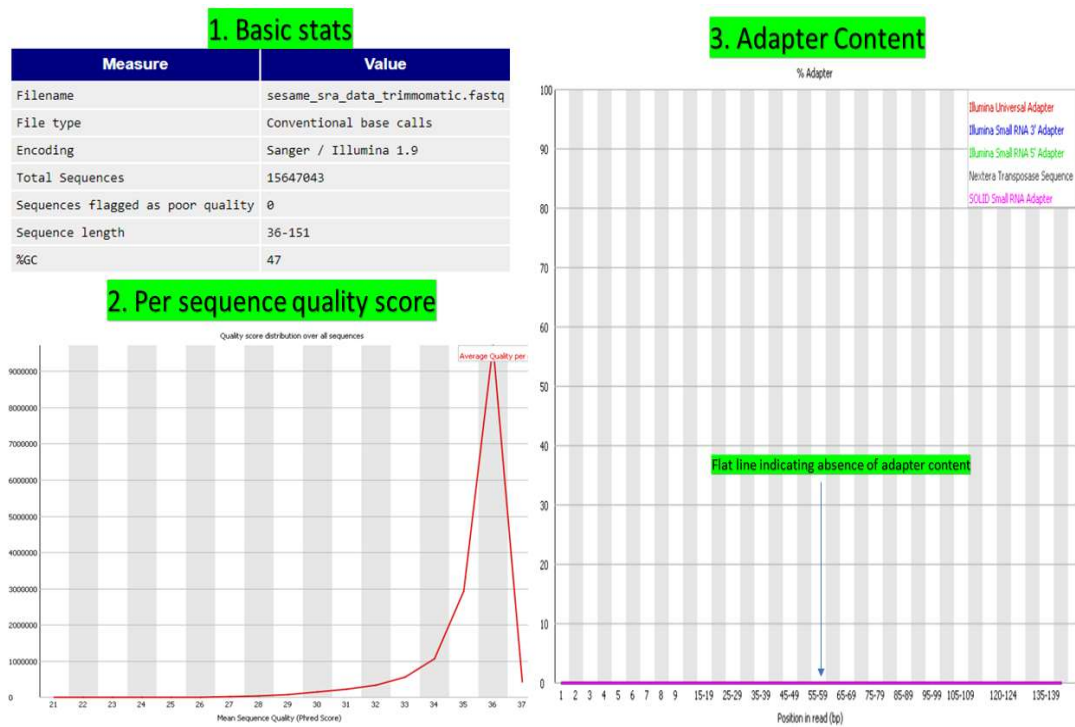


Figure 6.18 FastQC report of the processed sequencing data. The report corresponds to three parts including basis stats, per sequence quality score and adapter content respectively.

6.4.3 STAR analysis for transcriptome mapping and relative expression analysis

The STAR module was employed in the present investigation to map the generated contigs/scaffolds to the reference genome of sesame and calculate the number of mapped reads per gene. A screenshot of the output log file has been shown in Figure 6.19. The log file in Figure 6.19 clearly states that number of input reads provided for the mapping was 15647043 out of which 92.70% (14505083) reads were mapped uniquely (mapQ score of 255) to the reference genome with the average mapping length of 148.71 bases. After this the query reads which were mapped to multiple locus of the reference genome were analysed in which 2.44 % (381598) of the reads were found to be mapped with multiple locus and 0.02% (3307) of reads were mapped with too many loci. Finally, no reads were found to be mismatched with the reference genome but 4.81% (753323) were unmapped because of the too short read length. During this process, mapping speed of 40.38 million reads per hour was monitored on the computational system.

Started job on	Aug 21 14:43:23
Started mapping on	Aug 21 14:44:30
Finished on	Aug 21 15:07:45
Mapping speed, Million of reads per hour	40.38
Number of input reads	15647043
Average input read length	149
UNIQUE READS:	
Uniquely mapped reads number	14505083
Uniquely mapped reads %	92.70%
Average mapped length	148.71
Number of splices: Total	6501947
Number of splices: Annotated (sjdb)	6344588
Number of splices: GT/AG	6372711
Number of splices: GC/AG	91727
Number of splices: AT/AC	6158
Number of splices: Non-canonical	31351
Mismatch rate per base, %	0.35%
Deletion rate per base	0.02%
Deletion average length	2.46
Insertion rate per base	0.02%
Insertion average length	2.42
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	381598
% of reads mapped to multiple loci	2.44%
Number of reads mapped to too many loci	3307
% of reads mapped to too many loci	0.02%
UNMAPPED READS:	
Number of reads unmapped: too many mismatches	0
% of reads unmapped: too many mismatches	0.00%
Number of reads unmapped: too short	753323
% of reads unmapped: too short	4.81%
Number of reads unmapped: other	3732
% of reads unmapped: other	0.02%
CHIMERIC READS:	

Figure 6.19 Screenshot of the output log file of STAR module. The file corresponds to the mapping statistics of the query transcriptome to its reference genome.

The completion of the STAR workflow generated a file named ‘Reads per gene.out.tab’ which corresponds to the number of reads of the transcriptomic data aligned with reference genome. Figure 6.20 represents the mapped reads plot for the identified allergen instances from sesame. In order to retrieve more insightful information out of this, the locus corresponding to the protein names were mapped using the Uniprot database and thus proteins which were predicted as allergens were analysed. Figure 6.20 represents the number of mapped reads per gene to their respective protein and normalised number of mapped reads. It was observed that expression of already identified allergens in sesame was found in the descending order as 11s globulin seed storage protein 2 (Ses i 6) (Beyer *et al.*, 2007), 2S seed storage protein 1 (Ses i 2) (Beyer *et al.*, 2002), 2S albumin (Ses i 1) (Wolff *et al.*, 2003) and Oleosin H1 (Ses i 4) (Leduc *et al.*, 2006) with their mapped reads value of 448297, 304599, 237362 and 99141 respectively. Further, expression of the newly

predicted profilin food allergens in sesame by the developed RF model was found to be 3505, 99 and 6 reads of the query sesame transcriptomic data mapped with Profilin (loc105162034) (3505), Profilin (loc105257698) (99) and Profilin (loc105162033) (6) respectively. Additionally, other allergenic identified protein instances having relatively similar expression to those with highest expressing proteins were oleosin (25296) (22051), lipid transfer protein (32795), Major Facilitator Superfamily (MFS)-18 protein like (33706), Non-specific Lipid Transfer Protein (Ns-LTP)-3 like (21990), Uncharacterised ProteinLOC105168025 (21765), Glucan endo-1,3-beta-D-glucosidase (21005), extensin-2-like (17772), metallothionein-like protein-2 (12933), 36.4-kDa proline-rich protein (12030), and osmotin-like protein (8173). Most of the expressed proteins were also observed to be either annotated from homology (proteins with last word as “like”) or are uncharacterised proteins.

It may be noted that mapping the transcript reads to their corresponding protein segment does not completely represent the amount of protein present (Karnaneedi *et al.*, 2020). There are several other determinants like rate of protein translation and its regeneration, proficiency of the translational machinery and natural environmental conditions prior and throughout the translational process effects the protein expression level (Karnaneedi *et al.*, 2020).

A detailed discussion on prevalence, complications and regulation of sesame have been reviewed in the Introduction part section 1.5, which signifies the regulatory importance of sesame as food allergen. India and China have been distinguished as the top harvesters of sesame but there is no evidence of reported food allergy complications (Gangur and Acharya, 2021). This expression of unnoticed/unreported sesame food allergy in India may corresponds to concerns regarding unavailability of standard allergy tests leading to generation of biased diagnostic and epidemiological data (Bhattacharya *et al.*, 2018; Krishna *et al.*, 2020; Gangur and Acharya, 2021). Studies have been reported which confer the co-existence of food allergy and asthma leading to increased risk of life-threatening anaphylaxis (Wang and Liu, 2011; Foong *et al.*, 2107). In India, 37.9 million asthma patients have been diagnosed in 2018 but still no study have been carried out which evaluate its co-existence with food allergy (Krishna *et al.*, 2020). Thus, the abundance level of these distinguished major

allergens and newly identified allergens in sesame corresponds to their allergenic potential and signifies the importance of allergenicity assessment from food sources.

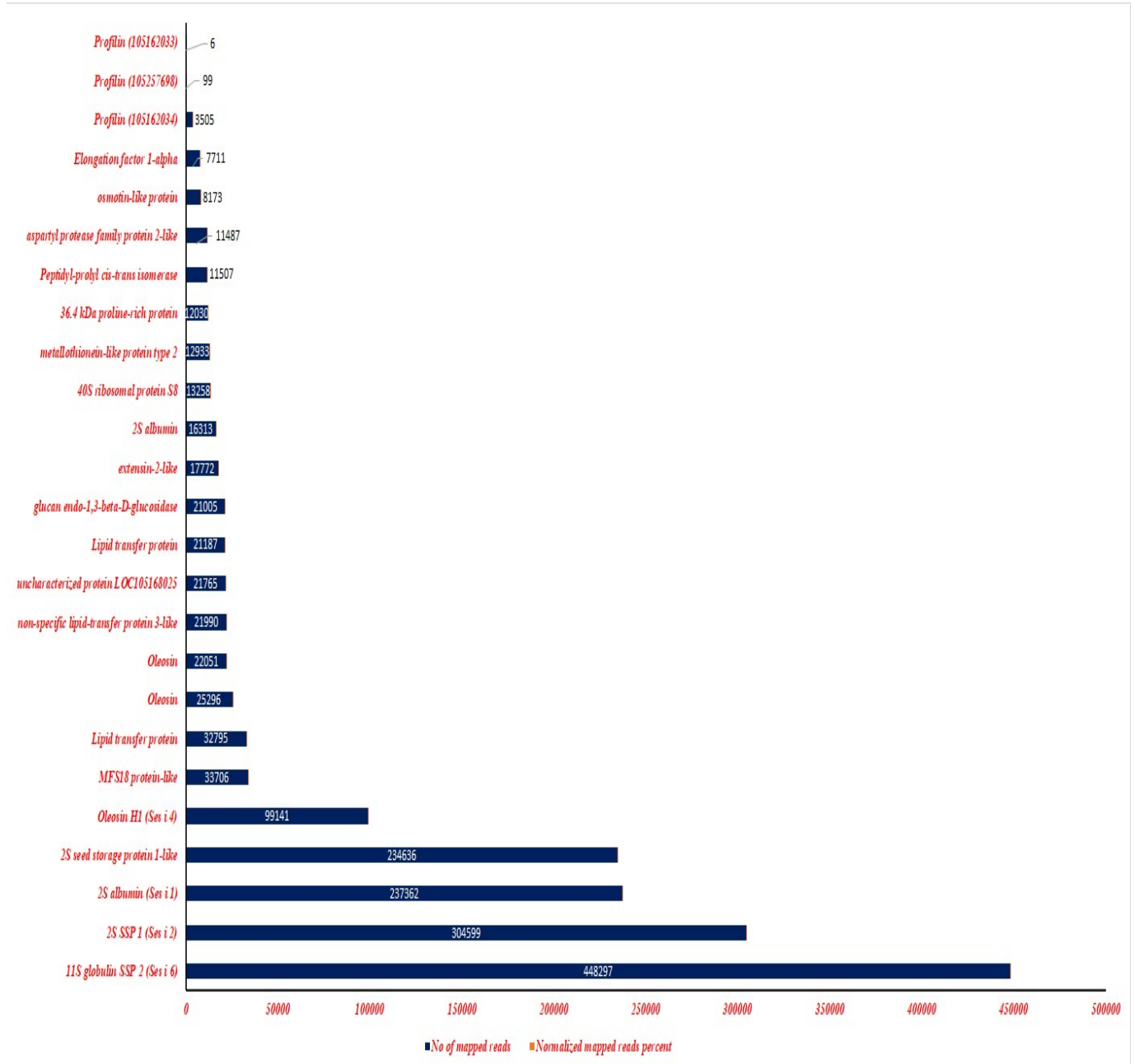


Figure 6.20 Number of mapped reads per gene for the sesame transcriptomic data. The horizontal axis represents the number of mapped reads. The vertical axis represents the identified allergen proteins by the developed model whose expressions were evaluated in the form of number of mapped reads.

Chapter 7

Summary & Conclusions

Food allergy is a predicament health concern affecting developed as well as developing nations. Primarily, proteins are accountable for inducing most food allergic reactions. At present more than 180 type of food sources have been identified to contain allergens and there are many to be explored. The FAO projection states that global population will be reaching the mark of 9.1 billion by 2050 and to feed this population an increase of 70 percent in food supply is required. Efforts have been carried out in the associated domain by worldwide researchers to produce genetically engineered crops with desired traits which certainly require expertise before its market viability. The identification of allergenic proteins in the food sources as well as assessing the allergenicity potential of foreign proteins is a vital checkpoint to assure food safety. The gold standard for food allergen identification includes serological and cytological assays but their applicability is limited by high cost and time-consuming process. As recommended by the expert committees, allergenicity assessment of the query protein is regulated in a “weight of evidence” manner which apart from standardised assays include computationally assisted allergenicity assessment. Numerous methods have been developed in the recent years involving application of similarity based, motif based, and epitope-based approaches. Further, utilisation of physicochemical space has also been observed in allergenicity assessment but there is a gap in exhaustive exploration of allergenic family to extract significant knowledge pertaining to their physicochemical nature and employing this information for the model development. Hence, the present study was objected towards allergen family based exhaustive feature extraction process and further employing this information in machine learning based model development.

In first phase of the study sequence, structure, and pharmacophore screening analyses of the food allergen profilins were carried out. Profilins from apple, pineapple, wheat, and soybean are responsible for triggering IgE mediated reactions in atopic individuals. Further, profilins from apple, pineapple, and soybean were also accounted for Pollen Food allergy syndrome. Multiple sequence alignment of the human

profilins and food allergen profilins revealed very low level of similarity, corresponding to the hypersensitive nature of the food allergen profilins in atopic individuals. Due to unavailability of experimentally resolved structure of these profilins, their homology modelled tertiary structure was deduced by which their structural conservation was revealed (4 helix, 7 strands and 11 loops) corresponding to their IgE cross-reactive potential and classification as pan-allergen. The pharmacophores identified against these profilins by virtue of virtual screen and molecular docking studies were showing efficient binding energy and non-covalent interactions and were in accordance with bioavailability profiles and thus can be considered as lead molecules for *in-vitro* studies.

After this, the study was focused towards unravelling the amino acid usage signatures of the profilin gene family. Multivariate statistical analysis revealed differential patterns of amino acid usage among the allergen and non-allergen profilins. It was evident from our analysis that the physicochemical features, including trypsin digestion, protein disorder and solubility of the profilins produced strong correlations with their differential patterns of amino acid usage. Interestingly, it was observed that relatively over-represented amino acids of the allergen profilins, on analysing their crystal structures were found to be surface exposed, suggesting that these over-represented amino acids might play a significant role in allergenicity assignment of profilins. Further, the dataset curation was carried out by retrieving equal instances of allergen and non-allergen protein instances from public-accessible databases namely Allergen Online, WHO/IUIS Allergen.org and SDAP. The final dataset consisted of non-redundant 1200 protein instances which was excluded by any profilin and sesame instances in either of the dataset classes. Machine learning based classifiers namely ZeroR, SVM and RF were developed employing the manually curated dataset out of which RF model was observed to be giving superior classification potential and thus opted for further analysis. The developed RF model outperforms the considered classifiers in terms of accuracy (0.88), MCC (0.76) and ROC area (0.95) indices signifying the role of employed descriptors extracted by differential amino acid usage analysis in computationally assisted allergenicity assignment of query protein instances.

Further, the developed RF model exhibiting best classification potential and was employed to perform the proteome wide prediction of allergenic instances in reference genome of sesame. The model was able to predict around 10 percent of the entire proteome as allergenic. The model classified the experimentally validated allergens like oleosin, 2s seed storage protein and 11s globulin seed storage proteins 2 in sesame as “*predicted allergens*” with superior prediction probability justifying its prediction accuracy. Apart from this, profilin protein instances in sesame were also classified as “*predicted allergens*” by the developed model with higher prediction probability. Thus, the study computationally validates the profilin protein in sesame as potential allergen and these finding can be taken up as preliminary basis to further consolidate these results. The RNASeq data analysis carried out on the transcriptomic data of sesame revealed the expression of 11s globulin seed storage protein 2, 2S seed storage protein 1, 2S albumin and Oleosin H1 allergenic proteins to be highest. Although profilin expression was observed to be comparatively less than the already distinguished allergens but their allergenic potential should not be overlooked due to their IgE cross-reactive nature.

In future, the model can be enriched by addition of more significant descriptors pertaining to food allergens by exploring the structural and sequence space to obtain optimum accuracy. The present study can be translated towards the development of public-accessible web server for potential allergenicity assessment of query proteins. The developed model can also be utilised by the food industries to screen out the potential instances of allergens from their dataset of genetically modified proteins which will eventually save time and resources being utilised for safety assessment of these proteins.

Bibliography

Ahuja, S. and Scypinski, S. eds., 2001. *Handbook of modern pharmaceutical analysis* (Vol. 3). Academic press.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.

Anvari, S., Miller, J., Yeh, C.Y. and Davis, C.M., 2019. IgE-mediated food allergy. *Clinical reviews in allergy & immunology*, 57(2), pp.244-260.

Asero, R., Mistrello, G. and Amato, S., 2011. The nature of melon allergy in ragweed-allergic subjects: A study of 1000 patients. In *Allergy and asthma proceedings* (Vol. 32, No. 1, p. 64). OceanSide Publications.

Ballmer-Weber, B.K., Wüthrich, B., Wangorsch, A., Fötisch, K., Altmann, F. and Vieths, S., 2001. Carrot allergy: double-blinded, placebo-controlled food challenge and identification of allergens. *Journal of allergy and clinical immunology*, 108(2), pp.301-307.

Beyer, K., Bardina, L., Grishina, G. and Sampson, H.A., 2002. Identification of sesame seed allergens by 2-dimensional proteomics and Edman sequencing: seed storage proteins as common food allergens. *Journal of Allergy and Clinical Immunology*, 110(1), pp.154-159.

Beyer, K., Grishina, G., Bardina, L. and Sampson, H.A., 2007. Identification of 2 new sesame seed allergens: Ses i 6 and Ses i 7. *Journal of allergy and clinical immunology*, 119(6), pp.1554-1556.

Beyer, K., Morrow, E., Li, X.M., Bardina, L., Bannon, G.A., Burks, A.W. and Sampson, H.A., 2001. Effects of cooking methods on peanut allergenicity. *Journal of Allergy and Clinical Immunology*, 107(6), pp.1077-1081.

Bhattacharya, K., Sircar, G., Dasgupta, A. and Bhattacharya, S.G., 2018. Spectrum of allergens and allergen biology in India. *International archives of allergy and immunology*, 177(3), pp.219-237.

- Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114-2120.
- Brandt, B.W., Feenstra, K.A. and Heringa, J., 2010. Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic acids research*, 38(suppl_2), pp.W35-W40.
- Breiteneder, H. and Mills, E.C., 2005. Molecular properties of food allergens. *Journal of Allergy and Clinical Immunology*, 115(1), pp.14-23.
- Bu, G., Luo, Y., Chen, F., Liu, K. and Zhu, T., 2013. Milk processing as a tool to reduce cow's milk allergenicity: a mini-review. *Dairy science & technology*, 93(3), pp.211-223.
- Burks, A.W., Tang, M., Sicherer, S., Muraro, A., Eigenmann, P.A., Ebisawa, M., Fiocchi, A., Chiang, W., Beyer, K., Wood, R. and Hourihane, J., 2012. ICON: food allergy. *Journal of Allergy and Clinical Immunology*, 129(4), pp.906-920.
- Burks, A.W., Williams, L.W., Thresher, W., Connaughton, C., Cockrell, G. and Helm, R.M., 1992. Allergenicity of peanut and soybean extracts altered by chemical or thermal denaturation in patients with atopic dermatitis and positive food challenges. *Journal of allergy and clinical immunology*, 90(6), pp.889-897.
- Cabanillas, B., Maleki, S.J., Rodríguez, J., Burbano, C., Muzquiz, M., Jiménez, M.A., Pedrosa, M.M., Cuadrado, C. and Crespo, J.F., 2012. Heat and pressure treatments effects on peanut allergenicity. *Food chemistry*, 132(1), pp.360-366.
- Carlson, G. and Coop, C., 2019. Pollen food allergy syndrome (PFAS): a review of current available literature. *Annals of Allergy, Asthma & Immunology*, 123(4), pp.359-365.
- Carlsson, L., Nyström, L.E., Sundkvist, I., Markey, F. and Lindberg, U., 1977. Actin polymerizability is influenced by profilin, a low molecular weight protein in non-muscle cells. *Journal of molecular biology*, 115(3), pp.465-483.
- Centor, R.M., 1991. Signal detectability: the use of ROC curves and their analyses. *Medical decision making*, 11(2), pp.102-106.

Chan, T.F., Ji, K.M., Yim, A.K.Y., Liu, X.Y., Zhou, J.W., Li, R.Q., Yang, K.Y., Li, J., Li, M., Law, P.T.W. and Wu, Y.L., 2015. The draft genome, transcriptome, and microbiome of *Dermatophagoides farinae* reveal a broad spectrum of dust mite allergens. *Journal of Allergy and Clinical Immunology*, 135(2), pp.539-548.

Cheng, F., Li, W., Zhou, Y., Shen, J., Wu, Z., Liu, G., Lee, P.W. and Tang, Y., 2012. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties.

Chruszcz, M., Kapingidza, A.B., Dolamore, C. and Kowal, K., 2018. A robust method for the estimation and visualization of IgE cross-reactivity likelihood between allergens belonging to the same protein family. *PLoS One*, 13(11), p.e0208276.

Chung, S.Y., Maleki, S.J. and Champagne, E.T., 2004. Allergenic properties of roasted peanut allergens may be reduced by peroxidase. *Journal of agricultural and food chemistry*, 52(14), pp.4541-4545.

Codex Alimentarius Commission, 2009. Foods derived from modern biotechnology. FAO/WHO, Rome, pp. 1-85.

Cohen, S.G., 2008. Food allergens: landmarks along a historic trail. *Journal of allergy and clinical immunology*, 121(6), pp.1521-1524.

Cortot, C.F., Sheehan, W.J., Permaul, P., Friedlander, J.L., Baxi, S.N., Gaffin, J.M., Dioun, A.F., Hoffman, E.B., Schneider, L.C. and Phipatanakul, W., 2012, May. Role of specific IgE and skin-prick testing in predicting food challenge results to baked egg. In *Allergy and asthma proceedings* (Vol. 33, No. 3, p. 275). OceanSide Publications.

Costa, J., Bavaro, S.L., Benede, S., Diaz-Perales, A., Bueno-Diaz, C., Gelencser, E., Klueber, J., Larre, C., Lozano-Ojalvo, D., Lupi, R. and Mafra, I., 2021. Are physicochemical properties shaping the allergenic potency of plant allergens?. *Clinical Reviews In Allergy & Immunology*.

Costa, J., Villa, C., Verhoeckx, K., Cirkovic-Velickovic, T., Schrama, D., Roncada, P., Rodrigues, P.M., Piras, C., Martín-Pedraza, L., Monaci, L. and

- Molina, E., 2022. Are physicochemical properties shaping the allergenic potency of animal allergens?. *Clinical reviews in allergy & immunology*, 62(1), pp.1-36.
- Daina, A., Michielin, O. and Zoete, V., 2017. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports*, 7(1), pp.1-13.
- Dang, H.X. and Lawrence, C.B., 2014. Allerdicator: fast allergen prediction using text classification techniques. *Bioinformatics*, 30(8), pp.1120-1128.
- Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall III, W.B., Snoeyink, J., Richardson, J.S. and Richardson, D.C., 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research*, 35(suppl_2), pp.W375-W383.
- Di Nardo, A., Gareus, R., Kwiatkowski, D. and Witke, W., 2000. Alternative splicing of the mouse profilin II gene generates functionally different profilin isoforms. *Journal of cell science*, 113(21), pp.3795-3803.
- Dimitrov, I., Flower, D.R. and Doytchinova, I., 2013. AllerTOP-a server for in silico prediction of allergens. In *BMC bioinformatics* 14(6), pp. 1-9.
- Dimitrov, I., Naneva, L., Bangov, I. and Doytchinova, I., 2014. Allergenicity prediction by artificial neural networks. *Journal of Chemometrics*, 28(4), pp.282-286.
- Dimitrov, I., Naneva, L., Doytchinova, I. and Bangov, I., 2014a. AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics*, 30(6), pp.846-851.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), pp.15-21.
- Ebner, C., Jensen-Jarolim, E., Leitner, A. and Breiteneder, H., 1998. Characterization of allergens in plant-derived spices: Apiaceae spices, pepper (Piperaceae), and paprika (bell peppers, Solanaceae). *Allergy*, 53, pp.52-54.

EFSA, 2010. Scientific Opinion on the assessment of allergenicity of GM plants and microorganisms and derived food and feed. *EFSA J.* 8, 1700. <https://doi.org/10.2903/j.efsa.2010.1700>

EFSA, 2017. Guidance on allergenicity assessment of genetically modified plants. *EFSA J.* 15, 4862. <https://doi.org/10.2903/j.efsa.2017.4862>

Ewing, B., Hillier, L., Wendl, M.C. and Green, P., 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome research*, 8(3), pp.175-185.

FAO, 2009. How to feed the world in 2050. https://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf

FAO, IFAD, UNICEF, WFP and WHO. 2021. The State of Food Security and Nutrition in the World 2021. Transforming food systems for food security, improved nutrition and affordable healthy diets for all. Rome, FAO. <https://doi.org/10.4060/cb4474en>

Florsheim, E.B., Sullivan, Z.A., Khoury-Hanold, W. and Medzhitov, R., 2021. Food allergy as a biological food quality control system. *Cell*.

Foong, R.X., du Toit, G. and Fox, A.T., 2017. Asthma, food allergy, and how they relate to each other. *Frontiers in pediatrics*, 5, p.89.

Fraczkiewicz, R. and Braun, W., 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of computational chemistry*, 19(3), pp.319-333.

Gangur, V. and Acharya, H.G., 2021. The global rise and the complexity of sesame allergy: prime time to regulate sesame in the United States of America?. *Allergies*, 1(1), pp.1-21.

González-Velasco, Ó., De Las Rivas, J. and Lacal, J., 2019. Proteomic and transcriptomic profiling identifies early developmentally regulated proteins in *Dictyostelium discoideum*. *Cells*, 8(10), p.1187.

Greenacre, M., 1984. Theory and applications of correspondence analysis. Academic Press, London.

Gupta, R.S., Warren, C.M., Smith, B.M., Jiang, J., Blumenstock, J.A., Davis, M.M., Schleimer, R.P. and Nadeau, K.C., 2019. Prevalence and severity of food allergies among US adults. *JAMA network open*, 2(1), pp.e185630-e185630.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), pp.10-18.

Han, P., Zhang, X. and Feng, Z.P., 2009. Predicting disordered regions in proteins using the profiles of amino acid indices. *Bmc Bioinformatics*, 10(1), pp.1-8.

Hansen, K.S., Ballmer-Weber, B.K., Lüttkopf, D., Skov, P.S., Wüthrich, B., Bindslev-Jensen, C., Vieths, S. and Poulsen, L.K., 2003. Roasted hazelnuts–allergenic activity evaluated by double-blind, placebo-controlled food challenge. *Allergy*, 58(2), pp.132-138.

Hartwig, J.H., Chambers, K.A., Hopcia, K.L. and Kwiatkowski, D.J., 1989. Association of profilin with filament-free regions of human leukocyte and platelet membranes and reversible membrane binding during platelet activation. *The Journal of cell biology*, 109(4), pp.1571-1579.

Hawkins, T. and Kihara, D., 2007. Function prediction of uncharacterized proteins. *Journal of bioinformatics and computational biology*, 5(01), pp.1-30.

Higham, D.J. and Higham, N.J., 2016. *MATLAB guide*. Society for Industrial and Applied Mathematics.

Hogan, M.B., Chawla, V., Scherr, R., Allenback, G., Wonnapharhown, A. and Wilson, N.W., 2016. Aeroallergen, Food and Panallergen Sensitization Patterns in Eosinophilic Esophagitis Patients. *Journal of Allergy and Clinical Immunology*, 137(2), p.AB232.

- Hou, T., Wang, J. and Li, Y., 2007. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *Journal of chemical information and modeling*, 47(6), pp.2408-2415.
- Hu, E., Chen, Z., Fredrickson, T. and Zhu, Y., 2001. Molecular cloning and characterization of profilin-3: a novel cytoskeleton-associated gene expressed in rat kidney and testes. *Nephron Experimental Nephrology*, 9(4), pp.265-274.
- Huang, S., McDowell, J.M., Weise, M.J. and Meagher, R.B., 1996. The Arabidopsis profilin gene family (Evidence for an ancient split between constitutive and pollen-specific profilin genes). *Plant physiology*, 111(1), pp.115-126.
- Huffman, L.M. and de Barros Ferreira, L., 2011. Whey-based ingredients. *Dairy ingredients for food processing*, 1, pp.179-198.
- Illi, S., von Mutius, E., Lau, S., Nickel, R., Grüber, C., Niggemann, B., Wahn, U. and Multicenter Allergy Study Group, 2004. The natural course of atopic dermatitis from birth to age 7 years and the association with asthma. *Journal of Allergy and Clinical Immunology*, 113(5), pp.925-931.
- Ivanciuc, O., Garcia, T., Torres, M., Schein, C.H. and Braun, W., 2009. Characteristic motifs for families of allergenic proteins. *Molecular immunology*, 46(4), pp.559-568.
- Ivanciuc, O., Schein, C.H. and Braun, W., 2003. SDAP: database and computational tools for allergenic proteins. *Nucleic acids research*, 31(1), pp.359-362.
- Jackins, V., Vimal, S., Kaliappan, M. and Lee, M.Y., 2021. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, 77(5), pp.5198-5219.
- Jorgensen, W.L., Maxwell, D.S. and Tirado-Rives, J., 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45), pp.11225-11236.

Kadam, K., Sawant, S., Jayaraman, V. and Kulkarni-Kale, U., 2016. Databases and Algorithms in Allergen Informatics. *Bioinformatics—Updated Features and Applications; IntechOpen: London, UK*, p.53.

Kandasamy, M.K., McKinney, E.C. and Meagher, R.B., 2002. Plant profilin isovariants are distinctly regulated in vegetative and reproductive tissues. *Cell motility and the cytoskeleton*, 52(1), pp.22-32.

Kar, P., Sharma, N.R., Singh, B., Sen, A. and Roy, A., 2021. Natural compounds from *Clerodendrum* spp. as possible therapeutic candidates against SARS-CoV-2: An in silico investigation. *Journal of Biomolecular Structure and Dynamics*, 39(13), pp.4774-4785.

Karnaneedi, S., Huerlimann, R., Johnston, E.B., Nugraha, R., Ruethers, T., Taki, A.C., Kamath, S.D., Wade, N.M., Jerry, D.R. and Lopata, A.L., 2020. Novel allergen discovery through comprehensive de novo transcriptomic analyses of five shrimp species. *International Journal of Molecular Sciences*, 22(1), p.32.

Kleber-Janke, T., Cramer, R., Appenzeller, U., Schlaak, M. and Becker, W.M., 1999. Selective cloning of peanut allergens, including profilin and 2S albumins, by phage display technology. *International archives of allergy and immunology*, 119(4), pp.265-274.

Kopper, R.A., Odum, N.J., Sen, M., Helm, R.M., Stanley, J.S. and Burks, A.W., 2005. Peanut protein allergens: the effect of roasting on solubility and allergenicity. *International archives of allergy and immunology*, 136(1), pp.16-22.

Kovar, D.R., Drøbak, B.K. and Staiger, C.J., 2000. Maize profilin isoforms are functionally distinct. *The Plant Cell*, 12(4), pp.583-598.

Krishna, M.T., Mahesh, P.A., Vedanthan, P.K., Mehta, V., Moitra, S. and Christopher, D.J., 2020. The burden of allergic diseases in the Indian subcontinent: barriers and challenges. *The Lancet. Global health*, 8(4), pp.e478-e479.

Kumar, K.K. and Shelokar, P.S., 2008. An SVM method using evolutionary information for the identification of allergenic proteins. *Bioinformation*, 2(6), p.253.

Ladics, G.S. and Selgrade, M.K., 2009. Identifying food proteins with allergenic potential: evolution of approaches to safety assessment and research to provide additional tools. *Regulatory Toxicology and Pharmacology*, 54(3), pp.S2-S6.

Laroche, G., Richet, C., ROWE, A.H., Mildred, P. and SAINT-GIRONS, F., 1930. L'Anaphylaxie Alimentaire. Alimentary Anaphylaxis. Gastro-intestinal Food Allergy. By G. Laroche, Charles Richet Fils and François Saint-Girons... Translated by Mildred P. Rowe and Albert H. Rowe. Preface by Albert H. Rowe. University of California Press.

Leduc, V., Moneret-Vautrin, D.A., Tzen, J.T.C., Morisset, M., Guerin, L. and Kanny, G., 2006. Identification of oleosins as major allergens in sesame seed allergic patients. *Allergy*, 61(3), pp.349-356.

Lemon-Mulé, H., Sampson, H.A., Sicherer, S.H., Shreffler, W.G., Noone, S. and Nowak-Wegrzyn, A., 2008. Immunologic changes in children with egg allergy ingesting extensively heated egg. *Journal of Allergy and Clinical Immunology*, 122(5), pp.977-983.

Li, J. and Wang, J., 2017. Improving allergen prediction in main crops using a weighted integrative method. *Interdisciplinary Sciences: Computational Life Sciences*, 9(4), pp.545-549.

Li, J., Ogorodova, L.M., Mahesh, P.A., Wang, M.H., Fedorova, O.S., Leung, T.F., Fernandez-Rivas, M., Mills, E.C., Potts, J., Kummeling, I. and Versteeg, S.A., 2020. Comparative study of food allergies in children from China, India, and Russia: the EuroPrevall-INCO surveys. *The Journal of Allergy and Clinical Immunology: In Practice*, 8(4), pp.1349-1358.

Lionta, E., Spyrou, G., K Vassilatis, D. and Cournia, Z., 2014. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry*, 14(16), pp.1923-1938.

Lipinski, C.A., 2004. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies*, 1(4), pp.337-341.

Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3), pp.3-25.

López-Torrejón, G., Crespo, J.F., Sánchez-Monge, R., Sánchez-Jiménez, M., Alvarez, J., Rodriguez, J. and Salcedo, G., 2005a. Allergenic reactivity of the melon profilin Cuc m 2 and its identification as major allergen. *Clinical & experimental allergy*, 35(8), pp.1065-1072.

Lopez-Torrejón, G., Ibanez, M.D., Ahrazem, O., Sánchez-Monge, R., Sastre, J., Lombardero, M., Barber, D. and Salcedo, G., 2005. Isolation, cloning and allergenic reactivity of natural profilin Cit s 2, a major orange allergen. *Allergy*, 60(11), pp.1424-1429.

Lyons, S.A., Clausen, M., Knulst, A.C., Ballmer-Weber, B.K., Fernandez-Rivas, M., Barreales, L., Bieli, C., Dubakiene, R., Fernandez-Perez, C., Jedrzejczak-Czechowicz, M. and Kowalski, M.L., 2020. Prevalence of food sensitization and food allergy in children across Europe. *The Journal of Allergy and Clinical Immunology: In Practice*, 8(8), pp.2736-2746.

Ma, Y., Zuidmeer, L., Bohle, B., Bolhaar, S.T.H., Gadermaier, G., Gonzalez-Mancebo, E., Fernandez-Rivas, M., Knulst, A.C., Himly, M., Asero, R. and Ebner, C., 2006. Characterization of recombinant Mal d 4 and its application for component-resolved diagnosis of apple allergy. *Clinical & Experimental Allergy*, 36(8), pp.1087-1096.

Mari, A., 2001. Multiple pollen sensitization: a molecular approach to the diagnosis. *International archives of allergy and immunology*, 125(1), pp.57-65.

Martin, Y.C., 2005. A bioavailability score. *Journal of medicinal chemistry*, 48(9), pp.3164-3170.

- May, C.D., 1976. Objective clinical and laboratory studies of immediate hypersensitivity reactions to foods in asthmatic children. *Journal of Allergy and Clinical immunology*, 58(4), pp.500-515.
- Michalski, M.C. and Januel, C., 2006. Does homogenization affect the human health properties of cow's milk?. *Trends in Food Science & Technology*, 17(8), pp.423-437.
- Mohabatkar, H., Mohammad Beigi, M., Abdolahi, K. and Mohsenzadeh, S., 2013. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Medicinal Chemistry*, 9(1), pp.133-137.
- Obermann, H., Raabe, I., Balvers, M., Brunswig, B., Schulze, W. and Kirchhoff, C., 2005. Novel testis-expressed profilin IV associated with acrosome biogenesis and spermatid elongation. *Molecular human reproduction*, 11(1), pp.53-64.
- Ogburn, R.N., Randall, T.A., Xu, Y., Roberts, J.H., Mebrahtu, B., Karnuta, J.M., Rider, S.D., Kissling, G.E., London, R.E., Pomés, A. and Arlian, L., 2017. Are dust mite allergens more abundant and/or more stable than other *Dermatophagoides pteronyssinus* proteins?. *Journal of Allergy and Clinical Immunology*, 139(3), pp.1030-1032.
- Pali-Schöll, I., Untersmayr, E., Klems, M. and Jensen-Jarolim, E., 2018. The effect of digestion and digestibility on allergenicity of food. *Nutrients*, 10(9), p.1129.
- Peden, J., 2000. Analysis of codon usage. Doctoral thesis, University of Nottingham, Nottingham.
- Pekar, J., Ret, D. and Untersmayr, E., 2018. Stability of allergens. *Molecular immunology*, 100, pp.14-20.
- Pi, X., Sun, Y., Fu, G., Wu, Z. and Cheng, J., 2021. Effect of processing on soybean allergens and their allergenicity. *Trends in Food Science & Technology*, 118, pp.316-327.

Polet, D., Lambrechts, A., Ono, K., Mah, A., Peelman, F., Vandekerckhove, J., Baillie, D.L., Ampe, C. and Ono, S., 2006. Caenorhabditis elegans expresses three functional profilins in a tissue-specific manner. *Cell motility and the cytoskeleton*, 63(1), pp.14-28.

Popescu, F.D., 2015. Cross-reactivity between aeroallergens and food allergens. *World journal of methodology*, 5(2), p.31.

Porter, J.W.G., 1978. The present nutritional status of milk proteins. *International Journal of Dairy Technology*, 31(4), pp.199-202.

Ramachandran, GN., Ramakrishnan, C. and Sasisekharan V. 1963. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7, pp.95-99.

Reindl, J., Rihs, H.P., Scheurer, S., Wangorsch, A., Haustein, D. and Vieths, S., 2002. IgE reactivity to profilin in pollen-sensitized subjects with adverse reactions to banana and pineapple. *International archives of allergy and immunology*, 128(2), pp.105-114.

Remington, B., Broekman, H.C.H., Blom, W.M., Capt, A., Crevel, R.W., Dimitrov, I., Faeste, C.K., Fernandez-Canton, R., Giavi, S., Houben, G.F. and Glenn, K.C., 2018. Approaches to assess IgE mediated allergy risks (sensitization and cross-reactivity) from new or modified dietary proteins. *Food and chemical toxicology*, 112, pp.97-107.

Rihs, H.P., Chen, Z., Ruëff, F., Petersen, A., Rozynek, P., Heimanna, H. and Baur, X., 1999. IgE binding of the recombinant allergen soybean profilin (rGly m 3) is mediated by conformational epitopes. *Journal of Allergy and Clinical Immunology*, 104(6), pp.1293-1301.

Rihs, H.P., Rozynek, P., May-Taube, K., Welticke, B. and Baur, X., 1994. Polymerase chain reaction based cDNA cloning of wheat profilin: a potential plant allergen. *International archives of allergy and immunology*, 105(2), pp.190-194.

- Rosace, D., Gomez-Casado, C., Fernandez, P., Perez-Gordo, M., del Carmen Dominguez, M., Vega, A., Belver, M.T., Ramos, T., Vega, F., Marco, G. and de Pedro, M., 2019. Profilin-mediated food-induced allergic reactions are associated with oral epithelial remodeling. *Journal of Allergy and Clinical Immunology*, 143(2), pp.681-690.
- Roy, A. and Basak, S., 2021. HIV long-term non-progressors share similar features with simian immunodeficiency virus infection of chimpanzees. *Journal of Biomolecular Structure and Dynamics*, 39(7), pp.2447-2454.
- Saha, S. and Raghava, G.P.S., 2006. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic acids research*, 34(suppl_2), pp.W202-W209.
- Sampath, V., Abrams, E.M., Adlou, B., Akdis, C., Akdis, M., Brough, H.A., Chan, S., Chatchatee, P., Chinthrajah, R.S., Cocco, R.R. and Deschildre, A., 2021. Food allergy across the globe. *Journal of Allergy and Clinical Immunology*, 148(6), pp.1347-1364.
- Santos, A. and Van Ree, R., 2011. Profilins: mimickers of allergy or relevant allergens?. *International archives of allergy and immunology*, 155(3), pp.191-204.
- Schloss, O.M., 1912. A CASE OF ALLEEGY TO COMMON FOODS. *American Journal of Diseases of Children*, 3(6), pp.341-362.
- Schluter, K., Jockusch, B.M. and Rothkegel, M., 1997. Profilins as regulators of actin dynamics. *Biochimica et Biophysica Acta-Molecular Cell Research*, 1359(2), pp.97-109.
- Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C. and Raghava, G.P., 2021. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Briefings in Bioinformatics*, 22(4), p.bbaa294.
- Singh, B., Karnwal, A., Tripathi, A. and Upadhyay, A.K., 2021. Food Allergens and Related Computational Biology Approaches: A Requisite for a Healthy Life. *Bioinformatics for agriculture: High-throughput approaches*, p.145.

Spergel, J. and Aceves, S.S., 2018. Allergic components of eosinophilic esophagitis. *Journal of Allergy and Clinical Immunology*, 142(1), pp.1-8.

Sterling, T. and Irwin, J.J., 2015. ZINC 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11), pp.2324-2337.

Sunseri, J. and Koes, D.R., 2016. Pharmit: interactive exploration of chemical space. *Nucleic acids research*, 44(W1), pp.W442-W448.

Tariq, S.M., Matthews, S.M., Hakim, E.A. and Arshad, S.H., 2000. Egg allergy in infancy predicts respiratory allergic disease by 4 years of age. *Pediatric Allergy and Immunology*, 11(3), pp.162-167.

Tedner, S.G., Asarnej, A., Thulin, H., Westman, M., Konradsen, J.R. and Nilsson, C., 2022. Food allergy and hypersensitivity reactions in children and adults—A review. *Journal of internal medicine*, 291(3), pp.283-302.

Thompson, J.D., Gibson, T.J. and Higgins, D.G., 2003. Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics*, (1), pp.2-3.

Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2), pp.455-461.

Turner, P.J., Mehr, S., Joshi, P., Tan, J., Wong, M., Kakakios, A. and Campbell, D.E., 2013. Safety of food challenges to extensively heated egg in egg-allergic children: a prospective cohort study. *Pediatric Allergy and Immunology*, 24(5), pp.450-455.

Untersmayr, E., Diesner, S.C., Brämswig, K.H., Knittelfelder, R., Bakos, N., Gundacker, C., Lukschal, A., Wallmann, J., Szalai, K., Pali-Schöll, I. and Boltz-Nitulescu, G., 2008. Characterization of intrinsic and extrinsic risk factors for celery allergy in immunosenescence. *Mechanisms of ageing and development*, 129(3), pp.120-128.

- Valenta, R., Duchene, M., Pettenburger, K., Sillaber, C., Valent, P., Bettelheim, P., Breitenbach, M., Rumpold, H., Kraft, D. and Scheiner, O., 1991. Identification of profilin as a novel pollen allergen; IgE autoreactivity in sensitized individuals. *Science*, 253(5019), pp.557-560.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. and Berendsen, H.J., 2005. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, 26(16), pp.1701-1718.
- Vaughan, W.T., 1930. Food allergens: I. A genetic classification, with results of group testing. *Journal of Allergy*, 1(5), pp.385-402.
- Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W. and Kopple, K.D., 2002. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12), pp.2615-2623.
- Verhoeckx, K.C., Vissers, Y.M., Baumert, J.L., Faludi, R., Feys, M., Flanagan, S., Herouet-Guicheney, C., Holzhauser, T., Shimojo, R., van der Bolt, N. and Wichers, H., 2015. Food processing and allergenicity. *Food and Chemical Toxicology*, 80, pp.223-240.
- Vogel, C. and Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics*, 13(4), pp.227-232.
- Wang, J. and Liu, A.H., 2011. Food allergies and asthma. *Current opinion in allergy and clinical immunology*, 11(3), p.249.
- Wang, J., Yu, Y., Zhao, Y., Zhang, D. and Li, J., 2013. Evaluation and integration of existing methods for computational prediction of allergens. *BMC bioinformatics*, 14(4), pp.1-9.
- Wang, J., Zhang, D. and Li, J., 2013a. PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection. *BMC systems biology*, 7(5), pp.1-9.

- Wang, L., Niu, D., Zhao, X., Wang, X., Hao, M. and Che, H., 2021. A Comparative Analysis of Novel Deep Learning and Ensemble Learning Models to Predict the Allergenicity of Food Proteins. *Foods*, 10(4), p.809.
- Wang, L., Xia, Q., Zhang, Y., Zhu, X., Zhu, X., Li, D., Ni, X., Gao, Y., Xiang, H., Wei, X. and Yu, J., 2016. Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. *BMC genomics*, 17(1), pp.1-13.
- Warrier, R. and Pande, H., 2016. Genetically engineered plants in the product development pipeline in India. *GM crops & food*, 7(1), pp.12-19.
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L. and Lepore, R., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1), pp.W296-W303.
- Westerhout, J., Krone, T., Snippe, A., Babé, L., McClain, S., Ladics, G.S., Houben, G.F. and Verhoeckx, K.C., 2019. Allergenicity prediction of novel and modified proteins: Not a mission impossible! Development of a Random Forest allergenicity prediction model. *Regulatory Toxicology and Pharmacology*, 107, p.104422.
- Willeroider, M., Fuchs, H., Ballmer-Weber, B.K., Focke, M., Susani, M., Thalhamer, J., Ferreira, F., Wüthrich, B., Scheiner, O., Breiteneder, H. and Hoffmann-Sommergruber, K., 2003. Cloning and molecular and immunological characterisation of two new food allergens, Cap a 2 and Lyc e 1, profilins from bell pepper (*Capsicum annuum*) and Tomato (*Lycopersicon esculentum*). *International archives of allergy and immunology*, 131(4), pp.245-255.
- Włodarczyk, K., Smolińska, B. and Majak, I., 2022. Tomato Allergy: The Characterization of the Selected Allergens and Antioxidants of Tomato (*Solanum lycopersicum*)—A Review. *Antioxidants*, 11(4), p.644.

Wolff, N., Cogan, U., Admon, A., Dalal, I., Katz, Y., Hodos, N., Karin, N. and Yannai, S., 2003. Allergy to sesame in humans is associated primarily with IgE antibody to a 14 kDa 2S albumin precursor. *Food and chemical toxicology*, 41(8), pp.1165-1174.

Worm, M., Hompes, S., Fiedler, E.M., Illner, A.K., Zuberbier, T. and Vieths, S., 2009. Impact of native, heat-processed and encapsulated hazelnuts on the allergic response in hazelnut-allergic patients. *Clinical & Experimental Allergy*, 39(1), pp.159-166.

Xue, B., Soeria-Atmadja, D., Gustafsson, M.G., Hammerling, U., Dunker, A.K. and Uversky, V.N., 2011. Abundance and functional roles of intrinsic disorder in allergenic proteins and allergen representative peptides. *Proteins: Structure, Function, and Bioinformatics*, 79(9), pp.2595-2606.

Yu, J., Ahmedna, M., Goktepe, I., Cheng, H. and Maleki, S., 2011. Enzymatic treatment of peanut kernels to reduce allergen levels. *Food chemistry*, 127(3), pp.1014-1022.

Zayas, J.F., 1997. Solubility of proteins. In *Functionality of proteins in food* (pp. 6-75). Springer, Berlin, Heidelberg.

Zhang, Z.H., Koh, J.L., Zhang, G.L., Choo, K.H., Tammi, M.T. and Tong, J.C., 2007. AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics*, 23(4), pp.504-

Index

A

Allergen: 1, 3, 5, 6, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 28, 29, 30, 32, 36, 37, 38, 41, 42, 43, 44, 45, 46, 49, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 67, 69, 71, 72, 73, 74, 77, 78, 79, 80, 81

Allergy: 1, 2, 3, 5, 6, 8, 9, 10, 11, 78, 80

Amino acid usage: 6, 23, 28, 29, 30, 51, 52, 53, 54, 61, 68, 81

AutoDock: 26, 45

B

BLAST: 16

Bioavailability: 27, 46, 48, 49, 50, 51, 81

Bioinformatics: 15, 25

C

Conserved: 5, 11, 15, 56, 57

Cross-reactivity: 5, 6, 11, 12, 13, 17, 37

Computational: 3, 5, 15, 19, 23, 24, 33, 39, 42, 43, 44, 45, 61, 67, 68, 72, 76, 80, 81, 82

Classifiers: 17, 22, 31, 62, 63, 64, 65, 66, 67, 68, 81

D

Descriptors: 7, 18, 21, 30, 31, 48, 61, 67, 68, 69, 72, 81, 82

Disorder: 29, 54, 81

Domain: 9, 14, 15, 38, 80

E

Epitope: 23, 80

Expression: 7, 34, 36, 72, 74, 76, 77, 79, 82

F

Food allergen: 1, 4, 5, 6, 9, 10, 11, 23, 25, 30, 32, 36, 37, 38, 44, 45, 55, 61, 69, 74, 78, 80, 81, 82

H

Hydrophobicity: 4, 18, 19, 23, 45, 46, 47

Hypersensitive: 8, 9, 10, 54, 81

I

Immunological: 1, 5

Identity: 11, 36, 37, 38, 56, 57

M

Model: 3, 4, 5, 7, 9, 15, 19, 20, 21, 22, 23, 24, 25, 26, 27, 30, 31, 32, 38, 39, 40, 41, 42, 43, 44, 45, 49, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 78, 79, 80, 81, 82

Machine learning: 7, 17, 18, 22, 23, 24, 30, 31, 61, 62, 67, 80

Molecular dynamics: 26, 41

N

Non allergen: 6, 15, 16, 18, 19, 20, 21, 28, 30, 51, 53, 54, 55, 57, 59, 61, 63, 67, 71, 81

O

Omics: 15, 61

P

Profilin: 4, 5, 6, 11, 13, 15, 25, 26, 27, 28, 29, 30, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 49, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 72, 78, 80, 81, 82

Protein: 3, 4, 6, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 36, 38, 39, 45, 54, 55, 60, 61, 67, 68, 69, 71, 72, 74, 77, 78, 79, 80, 81, 82

R

RNA-Seq: 24, 32, 74

RAAU: 23, 28, 29, 51, 52, 53, 54, 55, 56

S

Structure: 5, 6, 11, 14, 15, 16, 21, 24, 25, 26, 29, 30, 36, 39, 40, 41, 42, 54, 57, 58, 60, 69, 70, 80, 81

Sequence: 3, 4, 5, 6, 11, 15, 16, 18, 19, 20, 21, 24, 28, 29, 30, 33, 34, 36, 37, 54, 56, 57, 67, 74, 75, 76, 80, 82

T

Transcriptomic: 33, 34, 74, 75, 77, 78, 79, 82

W

WEKA: 30, 31, 32, 62

Appendices

Appendix 1 List of Uniprot Accessions with identifiers considered in the study for differential amino acid usage analysis.

<i>Given name</i>	<i>Uniprot Identifier</i>
<i>profilin_allergen_1</i>	>sp O82572 PROF1_RICCO Profilin-1 OS=Ricinus communis OX=3988 GN=PRO1 PE=2 SV=1
<i>profilin_allergen_2</i>	>sp P41372 PROF1_TOBAC Profilin-1 OS=Nicotiana tabacum OX=4097 GN=PRO1 PE=2 SV=1
<i>profilin_allergen_3</i>	>sp P49232 PROF1_WHEAT Profilin-1 OS=Triticum aestivum OX=4565 GN=PRO1 PE=2 SV=2
<i>profilin_allergen_4</i>	>sp P52184 PROF1_HORVU Profilin-1 OS=Hordeum vulgare OX=4513 GN=PRO1 PE=2 SV=1
<i>profilin_allergen_5</i>	>sp Q41344 PROF1_SOLLC Profilin-1 OS=Solanum lycopersicum OX=4081 GN=PRO1 PE=2 SV=1
<i>profilin_allergen_6</i>	>sp Q42449 PROF1_ARATH Profilin-1 OS=Arabidopsis thaliana OX=3702 GN=PRO1 PE=1 SV=1
<i>profilin_allergen_7</i>	>sp P25816 PROF1_BETPN Profilin-1 OS=Betula pendula OX=3505 GN=BETVII PE=1 SV=1
<i>profilin_allergen_8</i>	>sp A4KA39 PROF1_CORAV Profilin-1 OS=Corylus avellana OX=13451 PE=1 SV=1
<i>profilin_allergen_9</i>	>sp P35079 PROF1_PHLPR Profilin-1 OS=Phleum pratense OX=15957 GN=PRO1 PE=1 SV=1
<i>profilin_allergen_10</i>	>sp Q9XF40 PROF1_MALDO Profilin-1 OS=Malus domestica OX=3750 PE=1 SV=1
<i>profilin_allergen_11</i>	>sp O65809 PROF1_SOYBN Profilin-1 OS=Glycine max OX=3847 GN=PRO1 PE=1 SV=1
<i>profilin_allergen_12</i>	>sp Q64LH1 PROF1_AMBAR Profilin-1 OS=Ambrosia artemisiifolia OX=4212 GN=D106 PE=1 SV=1
<i>profilin_allergen_13</i>	>sp Q8H2C9 PROF1_ARTVU Profilin-1 OS=Artemisia vulgaris OX=4220 PE=1 SV=3
<i>profilin_allergen_14</i>	>sp O65812 PROF1_HEVBR Profilin-1 OS=Hevea brasiliensis OX=3981 PE=1 SV=1
<i>profilin_allergen_15</i>	>sp O24169 PROFA_OLEEU Profilin-1 OS=Olea europaea OX=4146 GN=PRO1 PE=1 SV=1
<i>profilin_allergen_16</i>	>sp P0DKD0 PROFD_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_17</i>	>sp P0DKD1 PROFG_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_18</i>	>sp P0DKD2 PROFH_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_19</i>	>sp P0DKD3 PROFI_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_20</i>	>sp P0DKD4 PROFJ_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_21</i>	>sp P0DKD5 PROFK_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_22</i>	>sp P0DKD6 PROFL_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_23</i>	>sp P0DKD7 PROFM_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_24</i>	>sp P0DKD8 PROFN_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_25</i>	>sp P0DKD9 PROFO_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1

<i>profilin_allergen_26</i>	>sp P0DKE1 PROFS_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_27</i>	>sp P0DKE8 PROAD_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_28</i>	>sp P0DKE9 PROAE_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_29</i>	>sp P0DKF0 PROAF_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_30</i>	>sp P0DKF2 PROAH_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_31</i>	>sp P0DKF3 PROAI_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_32</i>	>sp P0DKF5 PROAN_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_33</i>	>sp P0DKF6 PROAO_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_34</i>	>sp P0DKF7 PROAP_OLEEU Profilin-1 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_35</i>	>sp Q9XG85 PROF1_PARJU Profilin-1 OS=Parietaria judaica OX=33127 GN=PRO1 PE=1 SV=1
<i>profilin_allergen_36</i>	>sp P35081 PROF1_MAIZE Profilin-1 OS=Zea mays OX=4577 GN=PRO1 PE=1 SV=1
<i>profilin_allergen_37</i>	>sp Q9SQI9 PROF_ARAHY Profilin OS=Arachis hypogaea OX=3818 PE=1 SV=1
<i>profilin_allergen_38</i>	>sp Q9XF38 PROF_PYRCO Profilin OS=Pyrus communis OX=23211 PE=1 SV=1
<i>profilin_allergen_39</i>	>sp Q93YI9 PROF_CAPAN Profilin OS=Capsicum annuum OX=4072 PE=1 SV=1
<i>profilin_allergen_40</i>	>sp C6JWH0 PRF01_KALTU Profilin Sal k 4.0101 OS=Kali turgidum OX=151250 PE=1 SV=1
<i>profilin_allergen_41</i>	>sp P49233 PROF2_WHEAT Profilin-2 OS=Triticum aestivum OX=4565 GN=PRO2 PE=2 SV=2
<i>profilin_allergen_42</i>	>sp A4K9Z8 PROF2_BETPN Profilin-2 OS=Betula pendula OX=3505 PE=1 SV=1
<i>profilin_allergen_43</i>	>sp A4KA40 PROF2_CORAV Profilin-2 OS=Corylus avellana OX=13451 PE=1 SV=1
<i>profilin_allergen_44</i>	>sp O24650 PROF2_PHLPR Profilin-2 OS=Phleum pratense OX=15957 GN=PRO2 PE=1 SV=1
<i>profilin_allergen_45</i>	>sp Q9XF41 PROF2_MALDO Profilin-2 OS=Malus domestica OX=3750 PE=1 SV=1
<i>profilin_allergen_46</i>	>sp O65810 PROF2_SOYBN Profilin-2 OS=Glycine max OX=3847 GN=PRO2 PE=1 SV=1
<i>profilin_allergen_47</i>	>sp Q93YG7 PROF2_SOLLC Profilin-2 OS=Solanum lycopersicum OX=4081 PE=1 SV=1
<i>profilin_allergen_48</i>	>sp Q64LH2 PROF2_AMBAR Profilin-2 OS=Ambrosia artemisiifolia OX=4212 GN=A0418 PE=1 SV=1
<i>profilin_allergen_49</i>	>sp Q8H2C8 PROF2_ARTVU Profilin-2 OS=Artemisia vulgaris OX=4220 PE=1 SV=3
<i>profilin_allergen_50</i>	>sp Q9STB6 PROF2_HEVBR Profilin-2 OS=Hevea brasiliensis OX=3981 GN=PRO2 PE=1 SV=1
<i>profilin_allergen_51</i>	>sp Q9ST99 PROF2_TOBAC Profilin-2 OS=Nicotiana tabacum OX=4097 GN=PRO2 PE=1 SV=1
<i>profilin_allergen_52</i>	>sp A4GCR5 PROFE_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_53</i>	>sp A4GCR8 PROFQ_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_54</i>	>sp A4GD50 PROFT_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1

	SV=1
<i>profilin_allergen_55</i>	>sp A4GD56 PROAA_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_56</i>	>sp A4GDR3 PROAT_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_57</i>	>sp A4GDR8 PROAX_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_58</i>	>sp A4GDS6 PROBA_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_59</i>	>sp A4GDS7 PROBB_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_60</i>	>sp A4GDT0 PROBD_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_61</i>	>sp A4GDT4 PROBG_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_62</i>	>sp A4GDT9 PROBI_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_63</i>	>sp A4GDU0 PROBJ_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_64</i>	>sp A4GDU5 PROBM_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_65</i>	>sp A4GE39 PROBP_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_66</i>	>sp A4GE44 PROBR_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_67</i>	>sp A4GE47 PROBT_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_68</i>	>sp A4GE48 PROBU_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_69</i>	>sp A4GE53 PROBX_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_70</i>	>sp A4GE55 PROBZ_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_71</i>	>sp A4GFB7 PROCA_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_72</i>	>sp O24170 PROFB_OLEEU Profilin-2 OS=Olea europaea OX=4146 GN=PRO2 PE=1 SV=1
<i>profilin_allergen_73</i>	>sp P0DKE3 PROFV_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_74</i>	>sp P0DKE5 PROFX_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_75</i>	>sp P0DKE7 PROAC_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_76</i>	>sp P0DKF1 PROAG_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_77</i>	>sp P0DKF8 PROAQ_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_78</i>	>sp P0DKF9 PROAR_OLEEU Profilin-2 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_79</i>	>sp Q9T0M8 PROF2_PARJU Profilin-2 OS=Parietaria judaica OX=33127 GN=PRO2 PE=1 SV=1
<i>profilin_allergen_80</i>	>sp P35082 PROF2_MAIZE Profilin-2 OS=Zea mays OX=4577 GN=PRO2 PE=1 SV=2
<i>profilin_allergen_81</i>	>sp Q8GT39 PROF_PRUPE Profilin OS=Prunus persica OX=3760 PE=1 SV=1
<i>profilin_allergen_82</i>	>sp Q9XF37 PROF_APIGR Profilin OS=Apium graveolens OX=4045 PE=1 SV=1

<i>profilin_allergen_83</i>	>sp Q5FX67 PROF_CUCME Profilin OS=Cucumis melo OX=3656 PE=1 SV=1
<i>profilin_allergen_84</i>	>sp P49234 PROF3_WHEAT Profilin-3 OS=Triticum aestivum OX=4565 GN=PRO3 PE=2 SV=2
<i>profilin_allergen_85</i>	>sp A4KA44 PROF3_CORAV Profilin-3 OS=Corylus avellana OX=13451 PE=1 SV=1
<i>profilin_allergen_86</i>	>sp O24282 PROF3_PHLPR Profilin-3 OS=Phleum pratense OX=15957 GN=PRO3 PE=1 SV=1
<i>profilin_allergen_87</i>	>sp Q9XF42 PROF3_MALDO Profilin-3 OS=Malus domestica OX=3750 PE=1 SV=1
<i>profilin_allergen_88</i>	>sp Q64LH0 PROF3_AMBAR Profilin-3 OS=Ambrosia artemisiifolia OX=4212 GN=D03 PE=1 SV=1
<i>profilin_allergen_89</i>	>sp Q9M7N0 PROF3_HEVBR Profilin-3 OS=Hevea brasiliensis OX=3981 PE=1 SV=1
<i>profilin_allergen_90</i>	>sp A4GCR6 PROFF_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_91</i>	>sp A4GDQ6 PROAK_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_92</i>	>sp A4GDR4 PROAU_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_93</i>	>sp A4GDR9 PROAY_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_94</i>	>sp A4GDS9 PROBC_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_95</i>	>sp A4GDT1 PROBE_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_96</i>	>sp A4GDT5 PROBH_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_97</i>	>sp A4GDU2 PROBK_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_98</i>	>sp A4GDU6 PROBN_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_99</i>	>sp A4GE42 PROBQ_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_100</i>	>sp A4GE49 PROBV_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_101</i>	>sp A4GE54 PROBY_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_102</i>	>sp A4GFC0 PROCE_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_103</i>	>sp A4GFC3 PROCG_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_104</i>	>sp O24171 PROFC_OLEEU Profilin-3 OS=Olea europaea OX=4146 GN=PRO3 PE=1 SV=1
<i>profilin_allergen_105</i>	>sp P0DKE2 PROFU_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_106</i>	>sp P0DKE6 PROAB_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_107</i>	>sp P0DKF4 PROAM_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_108</i>	>sp P0DKG0 PROCB_OLEEU Profilin-3 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_109</i>	>sp P35083 PROF3_MAIZE Profilin-3 OS=Zea mays OX=4577 GN=PRO3 PE=1 SV=1
<i>profilin_allergen_110</i>	>sp P84177 PROF1_CITSI Profilin OS=Citrus sinensis OX=2711 PE=1 SV=2
<i>profilin_allergen_111</i>	>sp Q8L5D8 PROF_PHODC Profilin OS=Phoenix dactylifera OX=42345 PE=1 SV=1

<i>profilin_allergen_112</i>	>sp Q9XF39 PROF_PRUAV Profilin OS=Prunus avium OX=42229 PE=1 SV=1
<i>profilin_allergen_113</i>	>sp Q84V37 PROF_CHEAL Profilin OS=Chenopodium album OX=3559 PE=1 SV=1
<i>profilin_allergen_114</i>	>sp O81982 PROF_HELAN Profilin OS=Helianthus annuus OX=4232 PE=1 SV=1
<i>profilin_allergen_115</i>	>sp A4KA45 PROF4_CORAV Profilin-4 OS=Corylus avellana OX=13451 PE=1 SV=1
<i>profilin_allergen_116</i>	>sp Q9M7M9 PROF4_HEVBR Profilin-4 OS=Hevea brasiliensis OX=3981 PE=1 SV=1
<i>profilin_allergen_117</i>	>sp A4GCR7 PROFP_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_118</i>	>sp A4GD58 PROAJ_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_119</i>	>sp A4GDQ8 PROAL_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_120</i>	>sp A4GDR1 PROAS_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_121</i>	>sp A4GDR6 PROAV_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_122</i>	>sp A4GDS0 PROAZ_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_123</i>	>sp A4GDU3 PROBL_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_124</i>	>sp A4GE50 PROBW_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_125</i>	>sp A4GFB9 PROCD_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_126</i>	>sp A4GFC2 PROCF_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_127</i>	>sp A4GFC4 PROCH_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_128</i>	>sp A4KA49 PROCI_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_129</i>	>sp P0DKE4 PROFW_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_130</i>	>sp P0DKG1 PROCC_OLEEU Profilin-4 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_131</i>	>sp A4KA31 PROF4_PHLPR Profilin-4 OS=Phleum pratense OX=15957 PE=1 SV=1
<i>profilin_allergen_132</i>	>sp O22655 PROF4_MAIZE Profilin-4 OS=Zea mays OX=4577 GN=PRO4 PE=1 SV=1
<i>profilin_allergen_133</i>	>sp Q9FUB8 PROF_BRANA Profilin OS=Brassica napus OX=3708 PE=2 SV=1
<i>profilin_allergen_134</i>	>sp Q8GSL5 PROF_PRUDU Profilin OS=Prunus dulcis OX=3755 PE=1 SV=1
<i>profilin_allergen_135</i>	>sp Q8SAE6 PROF_DAUCA Profilin OS=Daucus carota OX=4039 PE=1 SV=1
<i>profilin_allergen_136</i>	>sp O49894 PROF_MERAN Profilin OS=Mercurialis annua OX=3986 PE=1 SV=1
<i>profilin_allergen_137</i>	>sp W8P570 PRF03_KALTU Profilin Sal k 4.0301 OS=Kali turgidum OX=151250 PE=1 SV=1
<i>profilin_allergen_138</i>	>sp A4KA41 PROF5_CORAV Profilin-5 OS=Corylus avellana OX=13451 PE=1 SV=1
<i>profilin_allergen_139</i>	>sp Q9M7M8 PROF5_HEVBR Profilin-5 OS=Hevea brasiliensis OX=3981 PE=1 SV=1
<i>profilin_allergen_140</i>	>sp A4GD54 PROFY_OLEEU Profilin-5 OS=Olea europaea OX=4146 PE=1 SV=1

<i>profilin_allergen_141</i>	>sp A4KA50 PRO CJ_OLEEU Profilin-5 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_142</i>	>sp A4KA32 PROF5_PHLPR Profilin-5 OS=Phleum pratense OX=15957 PE=1 SV=1
<i>profilin_allergen_143</i>	>sp Q9FR39 PROF5_MAIZE Profilin-5 OS=Zea mays OX=4577 GN=PRO5 PE=1 SV=1
<i>profilin_allergen_144</i>	>sp A4KA43 PROF6_CORAV Profilin-6 OS=Corylus avellana OX=13451 PE=1 SV=1
<i>profilin_allergen_145</i>	>sp Q9LEI8 PROF6_HEVBR Profilin-6 OS=Hevea brasiliensis OX=3981 PE=1 SV=1
<i>profilin_allergen_146</i>	>sp A4KA51 PROCK_OLEEU Profilin-6 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_147</i>	>sp A4KA33 PROF6_PHLPR Profilin-6 OS=Phleum pratense OX=15957 PE=1 SV=1
<i>profilin_allergen_148</i>	>sp A4KA55 PROF6_MAIZE Profilin-6 OS=Zea mays OX=4577 PE=1 SV=1
<i>profilin_allergen_149</i>	>sp A4KA52 PROCL_OLEEU Profilin-7 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_150</i>	>sp A4KA34 PROF7_PHLPR Profilin-7 OS=Phleum pratense OX=15957 PE=1 SV=1
<i>profilin_allergen_151</i>	>sp A4KA56 PROF7_MAIZE Profilin-7 OS=Zea mays OX=4577 PE=1 SV=1
<i>profilin_allergen_152</i>	>sp A4KA53 PROCM_OLEEU Profilin-8 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_153</i>	>sp A4KA36 PROF8_PHLPR Profilin-8 OS=Phleum pratense OX=15957 PE=1 SV=1
<i>profilin_allergen_154</i>	>sp A4KA57 PROF8_MAIZE Profilin-8 OS=Zea mays OX=4577 PE=1 SV=1
<i>profilin_allergen_155</i>	>sp A4KA54 PROCN_OLEEU Profilin-9 OS=Olea europaea OX=4146 PE=1 SV=1
<i>profilin_allergen_156</i>	>sp A4KA37 PROF9_PHLPR Profilin-9 OS=Phleum pratense OX=15957 PE=1 SV=1
<i>profilin_allergen_157</i>	>sp A4KA58 PROF9_MAIZE Profilin-9 OS=Zea mays OX=4577 PE=1 SV=1
<i>profilin_allergen_158</i>	>sp A4KA38 PRO10_PHLPR Profilin-10 OS=Phleum pratense OX=15957 PE=1 SV=1
<i>profilin_allergen_159</i>	>sp A4KA59 PRO10_MAIZE Profilin-10 OS=Zea mays OX=4577 PE=1 SV=1
<i>profilin_allergen_160</i>	>sp A4KA60 PRO11_MAIZE Profilin-11 OS=Zea mays OX=4577 PE=1 SV=1
<i>profilin_allergen_161</i>	>sp Q5VMJ3 PROFX_ORYSJ Profilin LP04 OS=Oryza sativa subsp. japonica OX=39947 GN=Os06g0152100 PE=2 SV=1
<i>profilin_allergen_162</i>	>sp P83647 PROFX_ORYSI Profilin LP04 OS=Oryza sativa subsp. indica OX=39946 GN=OsI_020954 PE=1 SV=2
<i>profilin_allergen_163</i>	>sp Q9FUD1 PROFA_ORYSJ Profilin-A OS=Oryza sativa subsp. japonica OX=39947 GN=Os10g0323600 PE=2 SV=1
<i>profilin_allergen_164</i>	>sp A4KA61 PRO12_MAIZE Profilin-12 OS=Zea mays OX=4577 PE=1 SV=1
<i>profilin_non_allergen_1</i>	>sp P33828 PROF_VAR67 Profilin OS=Variola virus (isolate Human/India/Ind3/1967) OX=587200 GN=A42R PE=3 SV=1
<i>profilin_non_allergen_2</i>	>sp P68695 PROF_VACCC Profilin OS=Vaccinia virus (strain Copenhagen) OX=10249 GN=A42R PE=3 SV=1
<i>profilin_non_allergen_3</i>	>sp Q6RZE1 PROF_RABPU Profilin OS=Rabbitpox virus (strain Utrecht) OX=45417 GN=RPXV150 PE=3 SV=1
<i>profilin_non_allergen_4</i>	>sp Q77TH1 PROF_VACCT Profilin OS=Vaccinia virus (strain Tian Tan) OX=10253 GN=TA53R PE=3 SV=1
<i>profilin_non_allergen_5</i>	>sp Q80DT4 PROF_CWPXG Profilin OS=Cowpox virus (strain GRI-90 / Grishak) OX=265871 GN=A44R PE=3 SV=1
<i>profilin_non_allergen_6</i>	>sp Q8V2L6 PROF_CAMPM Profilin OS=Camelpox virus (strain M-96) OX=203173 GN=CMLV161 PE=3 SV=1
<i>profilin_non_allergen_7</i>	>sp O57243 PROF_VACCA Profilin OS=Vaccinia virus (strain Ankara) OX=126794 GN=MVA154R PE=3 SV=1

<i>profilin_non_allergen_8</i>	>sp Q77DS0 PROF_CWPXB Profilin OS=Cowpox virus (strain Brighton Red) OX=265872 GN=CPXV179 PE=3 SV=1
<i>profilin_non_allergen_9</i>	>sp Q8JL78 PROF_ECTVM Profilin OS=Ectromelia virus (strain Moscow) OX=265874 GN=EVM141 PE=1 SV=1
<i>profilin_non_allergen_10</i>	>sp Q775N7 PROF_CAMPS Profilin OS=Camelpox virus (strain CMS) OX=203172 GN=CMP158R PE=3 SV=1
<i>profilin_non_allergen_11</i>	>sp Q8V4T7 PROF_MONPZ Profilin OS=Monkeypox virus (strain Zaire-96-I-16) OX=619591 GN=A42R PE=1 SV=1
<i>profilin_non_allergen_12</i>	>sp Q9XW16 PROF1_CAEL Profilin-1 OS=Caenorhabditis elegans OX=6239 GN=pfn-1 PE=2 SV=1
<i>profilin_non_allergen_13</i>	>sp P02584 PROF1_BOVIN Profilin-1 OS=Bos taurus OX=9913 GN=PFN1 PE=1 SV=2
<i>profilin_non_allergen_14</i>	>sp P62962 PROF1_MOUSE Profilin-1 OS=Mus musculus OX=10090 GN=Pfn1 PE=1 SV=2
<i>profilin_non_allergen_15</i>	>sp P62963 PROF1_RAT Profilin-1 OS=Rattus norvegicus OX=10116 GN=Pfn1 PE=1 SV=2
<i>profilin_non_allergen_16</i>	>sp P68696 PRO1A_ACACA Profilin-1A OS=Acanthamoeba castellanii OX=5755 PE=1 SV=2
<i>profilin_non_allergen_17</i>	>sp Q95VF7 PRO1B_ACACA Profilin-1B OS=Acanthamoeba castellanii OX=5755 PE=1 SV=3
<i>profilin_non_allergen_18</i>	>sp Q20025 PROF2_CAEL Profilin-2 OS=Caenorhabditis elegans OX=6239 GN=pfn-2 PE=2 SV=3
<i>profilin_non_allergen_19</i>	>sp P19984 PROF2_ACACA Profilin-2 OS=Acanthamoeba castellanii OX=5755 PE=1 SV=3
<i>profilin_non_allergen_20</i>	>sp P26200 PROF2_DICDI Profilin-2 OS=Dictyostelium discoideum OX=44689 GN=proB PE=1 SV=1
<i>profilin_non_allergen_21</i>	>sp P35080 PROF2_HUMAN Profilin-2 OS=Homo sapiens OX=9606 GN=PFN2 PE=1 SV=3
<i>profilin_non_allergen_22</i>	>sp Q09430 PROF2_BOVIN Profilin-2 OS=Bos taurus OX=9913 GN=PFN2 PE=1 SV=2
<i>profilin_non_allergen_23</i>	>sp Q4R4P8 PROF2_MACFA Profilin-2 OS=Macaca fascicularis OX=9541 GN=PFN2 PE=2 SV=1
<i>profilin_non_allergen_24</i>	>sp Q9JJV2 PROF2_MOUSE Profilin-2 OS=Mus musculus OX=10090 GN=Pfn2 PE=1 SV=3
<i>profilin_non_allergen_25</i>	>sp Q5R4E2 PROF2_PONAB Profilin-2 OS=Pongo abelii OX=9601 GN=PFN2 PE=2 SV=3
<i>profilin_non_allergen_26</i>	>sp Q9EPC6 PROF2_RAT Profilin-2 OS=Rattus norvegicus OX=10116 GN=Pfn2 PE=1 SV=3
<i>profilin_non_allergen_27</i>	>sp Q21193 PROF3_CAEL Profilin-3 OS=Caenorhabditis elegans OX=6239 GN=pfn-3 PE=2 SV=1
<i>profilin_non_allergen_28</i>	>sp Q32PB1 PROF3_BOVIN Profilin-3 OS=Bos taurus OX=9913 GN=PFN3 PE=2 SV=1
<i>profilin_non_allergen_29</i>	>sp M0RCP6 PROF3_RAT Profilin-3 OS=Rattus norvegicus OX=10116 GN=Pfn3 PE=2 SV=1
<i>profilin_non_allergen_30</i>	>sp P60673 PROF3_HUMAN Profilin-3 OS=Homo sapiens OX=9606 GN=PFN3 PE=2 SV=1
<i>profilin_non_allergen_31</i>	>sp Q8T8M2 PROF3_DICDI Profilin-3 OS=Dictyostelium discoideum OX=44689 GN=proC PE=3 SV=1
<i>profilin_non_allergen_32</i>	>sp Q9DAD6 PROF3_MOUSE Profilin-3 OS=Mus musculus OX=10090 GN=Pfn3 PE=1 SV=1
<i>profilin_non_allergen_33</i>	>sp Q2NKT1 PROF4_BOVIN Profilin-4 OS=Bos taurus OX=9913 GN=PFN4 PE=2 SV=1
<i>profilin_non_allergen_34</i>	>sp Q8NHR9 PROF4_HUMAN Profilin-4 OS=Homo sapiens OX=9606 GN=PFN4 PE=1 SV=1
<i>profilin_non_allergen_35</i>	>sp Q9D6I3 PROF4_MOUSE Profilin-4 OS=Mus musculus OX=10090 GN=Pfn4 PE=1 SV=1
<i>profilin_non_allergen_36</i>	>sp Q5IRJ7 PROF4_RAT Profilin-4 OS=Rattus norvegicus OX=10116

Appendix 2 The z-score table of the generated alignment for eight allergen profilins from *Hevea brasiliensis* (Hev b 8), *Artemisia vulgaris* (Art v 4), *Betula verrucosa* (Bet v 2), *Cucumis melo* (Cuc m 2), *Phleum pratense* (Phl p 12), *Zea mays* (Zea m 12), *Arachis hypogaea* (Ara h 5) and *Ambrosia artemisiifolia* (Amb a 8) respectively. The z-score of the conserved motifs SWQ, YVD, VWA, LAPTG, KYMVIQGE, VIRGKKG, KKT, GIY, PGQCNM and LGDYL are highlighted in yellow colour.

Alignment (Residue position)	Sequence Harmony Z-score	Multi-Relief Z-score	Consensus Strings			
81	-1.46	2.17	AS	EN	AH	D
93	-3.37	1.74	L	L	I	V
85	-0.73	0.83	AS	HQ	T	HT
52	-0.53	0.95	FL	FV	I	LV
54	-1.19	0.93	QT	HT	Q	H
63	-0.22	0.83	S	AS	AQ	AE
115	-0.16	0.71	AP	AS	P	AT
71	-1.22	1.02	DE	NQ	E	E
78	-1.53	0.79	KN	K	GS	K
107	-1.08	0.82	AT	A	A	RV
44	-0.95	0.54	NQ	Q	N	HQ
64	-0.73	0.84	NS	S	N	DN
92	-0.64	0.35	FY	FH	Y	F
129	-0.87	0.84	FI	F	I	IV
105	-3.00	2.00	P	A	P	P
120	-3.18	1.99	I	I	V	I
154	-2.59	2.92	D	D	D	E
157	1.09	-1.06	LM	LM	LY	FM
29	-0.70	0.70	T	T	AV	AT
42	-0.58	0.75	-T	QT	-	-
118	-0.47	0.59	IV	I	AV	I
74	1.44	-2.02	NT	DT	AT	AT
43	-1.04	1.15	-G	G	-	-
75	-1.18	1.12	A	AG	G	G

136	-1.15	1.15	MV	V	M	M
144	-1.18	0.85	IV	V	I	V
96	0.22	0.33	AT	AI	T	AT
47	1.20	-1.22	AS	AT	ST	AS
137	0.40	-0.72	NT	AT	T	T
57	0.88	-1.54	NS	ST	S	ST
153	0.83	-1.51	IV	IL	I	IV
25	1.29	-1.71	-G	-G	GM	GM
38	-0.25	0.07	DE	D	E	DE
24	1.49	-1.94	-S	-S	-S	-G
77	1.51	-1.84	IM	IM	MV	IM
36	0.44	0.13	LM	M	M	LM
61	0.35	-0.66	KQ	KQ	Q	Q
68	0.44	-0.70	F	F	FL	FL
73	0.31	-0.70	I	I	IV	IV
79	0.39	-0.66	DE	DE	D	D
94	0.45	0.06	AG	G	G	AG
119	0.31	-0.66	CT	CT	T	T
124	0.33	-0.05	GN	G	GN	G
125	0.40	-0.70	Q	Q	MQ	MQ
127	0.31	-0.66	LM	LM	L	L
133	0.35	0.47	DE	DE	D	D
1	-1.07	-0.35	-	-	-M	-
2	-1.10	-0.35	-	-	-H	-
3	-1.08	-0.35	-	-	-H	-
4	-1.07	-0.35	-	-	-H	-
5	-1.07	-0.35	-	-	-H	-
6	-1.10	-0.35	-	-	-H	-
7	-1.08	-0.35	-	-	-H	-
8	-1.10	-0.35	-	-	-S	-
9	-1.12	-0.35	-	-	-S	-
10	-1.10	-0.35	-	-	-G	-
11	-1.08	-0.35	-	-	-V	-

12	-1.09	-0.35	-	-	-D	-
13	-1.10	-0.35	-	-	-L	-
14	-1.09	-0.35	-	-	-G	-
15	-1.10	-0.35	-	-	-T	-
16	-1.10	-0.35	-	-	-E	-
17	-1.12	-0.35	-	-	-N	-
18	-1.07	-0.35	-	-	-L	-
19	-1.09	-0.35	-	-	-Y	-
20	-1.10	-0.35	-	-	-F	-
21	-1.07	-0.35	-	-	-Q	-
22	-1.12	-0.35	-	-	-S	-
40	-1.11	0.68	E	DE	E	E
45	-1.12	0.68	H	HQ	H	H
49	-1.10	0.68	A	AS	A	A
70	-1.10	0.54	P	P	PS	P
76	-0.65	0.69	I	I	I	IM
82	-1.07	-0.35	E	E	DE	E
83	-1.13	-0.84	P	AP	P	P
91	-0.67	0.42	L	L	L	LM
126	-0.66	0.69	A	A	A	AS
147	-1.04	0.26	KR	R	R	R
155	-1.09	0.26	QT	Q	Q	Q
23	0.75	-0.82	-G	-G	-G	-
48	0.77	-0.49	S	AS	AS	AS
67	0.87	-0.30	EQ	EQ	Q	EQ
128	0.96	-0.15	IV	V	IV	IV
26	nan	nan	S	S	S	S
27	nan	nan	W	W	W	W
28	nan	nan	Q	Q	Q	Q
30	nan	nan	Y	Y	Y	Y
31	nan	nan	V	V	V	V
32	nan	nan	D	D	D	D
33	1.37	-1.68	DE	DE	DE	DE

34	nan	nan	H	H	H	H
35	nan	nan	L	L	L	L
37	nan	nan	C	C	C	C
39	nan	nan	I	I	I	I
41	nan	nan	G	G	G	G
46	nan	nan	L	L	L	L
50	nan	nan	A	A	A	A
51	nan	nan	I	I	I	I
53	nan	nan	G	G	G	G
55	nan	nan	D	D	D	D
56	nan	nan	G	G	G	G
58	nan	nan	V	V	V	V
59	nan	nan	W	W	W	W
60	nan	nan	A	A	A	A
62	nan	nan	S	S	S	S
65	nan	nan	F	F	F	F
66	nan	nan	P	P	P	P
69	nan	nan	K	K	K	K
72	nan	nan	E	E	E	E
80	nan	nan	F	F	F	F
84	nan	nan	G	G	G	G
86	nan	nan	L	L	L	L
87	nan	nan	A	A	A	A
88	nan	nan	P	P	P	P
89	nan	nan	T	T	T	T
90	nan	nan	G	G	G	G
95	nan	nan	G	G	G	G
97	nan	nan	K	K	K	K
98	nan	nan	Y	Y	Y	Y
99	nan	nan	M	M	M	M
100	nan	nan	V	V	V	V
101	nan	nan	I	I	I	I
102	nan	nan	Q	Q	Q	Q

103	nan	nan	G	G	G	G
104	nan	nan	E	E	E	E
106	nan	nan	G	G	G	G
108	nan	nan	V	V	V	V
109	nan	nan	I	I	I	I
110	nan	nan	R	R	R	R
111	nan	nan	G	G	G	G
112	nan	nan	K	K	K	K
113	nan	nan	K	K	K	K
114	nan	nan	G	G	G	G
116	nan	nan	G	G	G	G
117	nan	nan	G	G	G	G
121	nan	nan	K	K	K	K
122	nan	nan	K	K	K	K
123	nan	nan	T	T	T	T
130	nan	nan	G	G	G	G
131	nan	nan	I	I	I	I
132	nan	nan	Y	Y	Y	Y
134	nan	nan	E	E	E	E
135	nan	nan	P	P	P	P
138	nan	nan	P	P	P	P
139	nan	nan	G	G	G	G
140	nan	nan	Q	Q	Q	Q
141	nan	nan	C	C	C	C
142	nan	nan	N	N	N	N
143	nan	nan	M	M	M	M
145	nan	nan	V	V	V	V
146	nan	nan	E	E	E	E
148	nan	nan	L	L	L	L
149	nan	nan	G	G	G	G
150	nan	nan	D	D	D	D
151	nan	nan	Y	Y	Y	Y
152	nan	nan	L	L	L	L

nan	nan	G	G	G	G
-----	-----	---	---	---	---

List of Publications

- **Singh, B.**, Ahanathapillai, V., Sharma, N.R., Jan, S., Roy, A. and Upadhyay, A.K., 2022. Structural insights into the amino acid usage variations in the profilin gene family. *Amino Acids*, 54(3), pp.411-419.
- **Singh, B.**, Jan, S., Kumar Upadhyay, A. and Raj Sharma, N., 2022. In silico investigation for drug-like pharmacophores against food allergen profilins. *Allergo Journal International*, pp.1-9.
- **Singh, B.**, Karnwal, A., Tripathi, A. and Upadhyay, A.K., 2021. Food Allergens and Related Computational Biology Approaches: A Requisite for a Healthy Life. In *Bioinformatics for agriculture: High-throughput approaches* (pp. 145-160). Springer, Singapore. (ISBN978-981-33-4791-5)

Paper presented in International Conferences

- Paper entitled “*Computational analysis of the food-allergen profilins reveals key compounds to combat its associated allergic responses*” presented in *International Conference on Fundamental and Applied Sciences* (ICFAS2021) organised by Faculty of Science and I.Q.A.C. from 24th March 2021 to 26th March 2021. Position secured- Second Prize.
- Paper entitled “*Protein Pratyurjak Pragukti: A random forest model for anticipation of potential allergens*” presented in *International Conference on Plant Physiology and Biotechnology* (ICPPB) held from 10-12 September 2021 organized by Department of Molecular Biology and Genetic Engineering, School of Bio-engineering and Biosciences, Lovely Professional University, Punjab.