Investigating Peanut Allergy Sensitization Through Predicted Protein Protein Interactions

Jacob Halle, Hajin Sim

Background

Peanut allergy is a significant health concern affecting individuals worldwide. For those sensitive to peanuts, exposure can lead to symptoms ranging from mild discomfort and itching, to severe anaphylactic shock[1]. Furthermore, the prevalence of peanut sensitivity has increased 3.5-fold in the past two decades[2] . Despite the severity and increasing prevalence of this allergen, the mechanism for allergic sensitization remains largely unknown.

Allergic reactions occur when B-lymphocytes mistake a foreign particle that is not inherently harmful for a dangerous pathogen. This leads to IgE antibody production and an adaptive immune response that will trigger on the next exposure to the allergen[3]. While this pathway is well understood, the question remains of why B-cells mistake non-harmful foreign substances for pathogens. Understanding the molecular interactions between peanut allergens (Ara h) and human immune system proteins is essential for elucidating the mechanisms underlying peanut allergy sensitization. In this study, we employed computational approaches to predict protein-protein interactions (PPIs) and perform analysis on these to investigate the interactions between peanut allergens and the immune system as well as the sensitization pathway.

Methods

Data Collection

For data collection, we retrieved human immune system protein data from UniProt, filtering for organisms classified as "human" and grouped by Gene Ontology terms related to immune system processes, resulting in a dataset comprising 37,250 proteins.[4] Peanut allergen proteins were obtained by filtering UniProt for peanut taxonomy and proteins annotated with allergens, supplemented with allergens from Allergen.org, resulting in a total of 30 protein sequences.[5]

Topsy Turvy

To obtain PPIs, we utilized the novel deep-learning algorithm Topsy Turvy. This algorithm takes a hybrid approach that combines insights from sequence-based bottom up approaches with structure information from global-based top down approaches.[6] The algorithm takes only protein sequences as inputs, which makes it an appropriate choice

for this large PPI screen, although it does utilize global and molecular aspects of protein interaction during training[6]. Another advantage of this model is that it performs very well across species, which made it particularly appealing for this project[6]. The development team behind Topsy-turvy provided pre-trained human models as well, which relieved a very large computational challenge.

A comprehensive dataset of 1,117,499 combinations was created from all possible pairings of the 37,250 human immune-related proteins and 30 peanut allergens. Due to computational demands, It was necessary to process this dataset using High-Performance Computing (HPC) infrastructure. The pretrained model 'topsy_turvy_v1.sav' was used for predictions. Output included positive interactions with scores above 0.5 and negative interactions. The team behind Topsy-Turvy considers any pairing with an interaction above 0.5 as a positive result, as the two proteins are more likely to interact than not. Contact maps representing residue interaction probabilities were generated and visualized using heatmaps.

MEME[7]

For molecular motif identification, we utilized MEME (Multiple Em for Motif Elicitation) analysis. High-confidence interactions (scores >0.7) predicted by Topsy Turvy were further analyzed. FASTA sequences of human immune proteins and peanut allergen proteins involved in these interactions were subjected to MEME analysis separately. MEME identified conserved motifs within human immune proteins and potential allergen epitopes in peanut proteins.

GO Term Enrichment

GO Term Enrichment analysis was performed on high-confidence pairs (>0.7). This was done using the 'enrichGO' function from the 'clusterProfile' package in Rstudio. The org.Hs.eg.db database was used to obtain GO terms related to biological processes. The high-confidence pairs were compared to all screened human proteins; not the entire proteome. A false discovery rate of 0.05 was used to correct for multiple hypothesis testing.

Reactome Pathway Analysis[8]

The 'Reactome' Pathway database was used to find overrepresented pathways involved in the predicted binding partners. High-confidence pairs were used (>0.7) and inferences from IntAct were used as well. This database was also used to browse pathways specifically related to interesting binding partners, such as CD209.

Results

The Topsy Turvy algorithm predicted 5,064 PPIs (>0.5 interaction probability) between human immune system proteins and peanut allergens. Contact maps derived from these predictions demonstrated residue interaction probabilities, highlighting potential sites of interaction.

Table 1. Top ten predicted interactions

Human immune	Peanut allergen	Interaction Probability
Q14160	Q9SQI9	0.9385416508
Q14005	Q9SQI9	0.9231057763
P84022	Q9SQI9	0.8999634385
P0DOX8	Q9AXI1	0.8909175396
P27361	Q9SQI9	0.8822197914
L7RXH5	Q9SQI9	0.8822197914
Q1HBJ4	Q9SQI9	0.8779109716
P28482	Q9SQI9	0.8779109716
P84022	B0YIU5	0.8751564026
A0A286R9D9	A0A444XS96	0.8742918372

Among the 5,064 predicted PPIs, the table below lists the top five specific peanut allergen proteins with the highest frequencies of predicted interactions with human immune system proteins.

Table 2. Top 5 frequencies of peanut allergens with binding interaction probability >0.5

Peanut Protein	Number of Predicted Positive Interactions
Ara h 3	1258
Ara h 13	427
Ara h 4	312
Ara h 8	282

Ara h 15	245

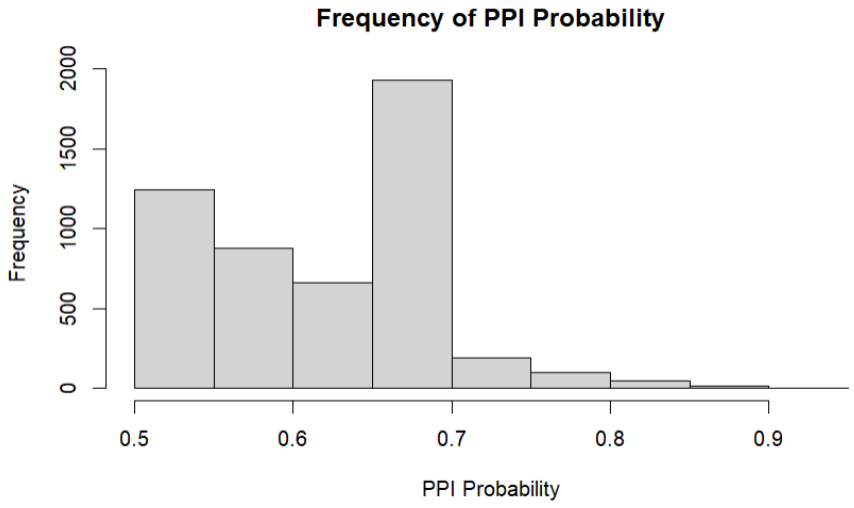
356 peanut-human combinations exhibited high positive interaction scores (>0.7) with human immune system proteins. Among these interactions, the 5 most frequently involved peanut proteins are shown in Table 3.

Table 3. Top 5 frequencies of peanut allergens with binding interaction probability >0.7

Peanut Protein	Number of High Positive Interaction Scores
Ara h 8 - isoform	48
Ara h 8	43
Ara h 3	36
Ara h 18	34
Ara h 5	30

The frequency PPI predictions generally decreases with increasing probability, as to be expected. However, there is a large spike in binding probabilities from 0.65-0.7.

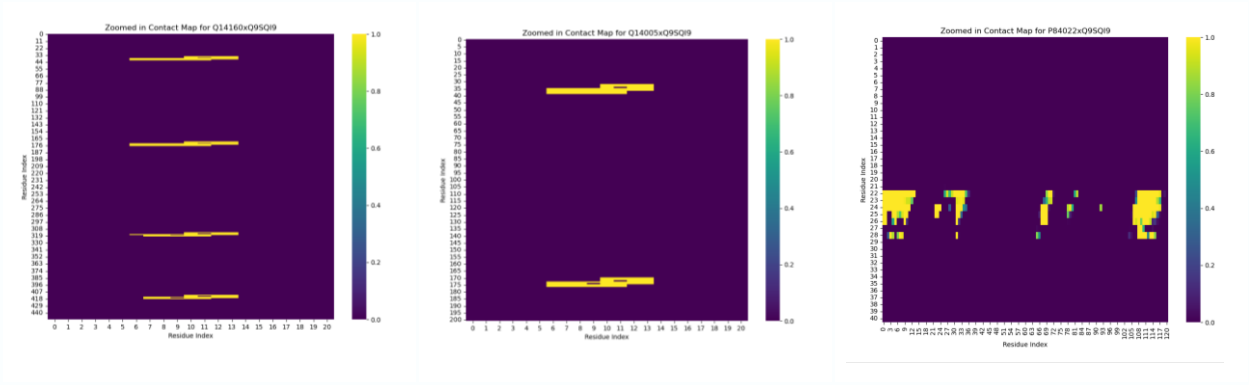
Figure 1. Histogram of PPI probability



For each interaction with probability over 0.5, a contact map was generated that shows the residues that are predicted to interact. The top 3 interactions are shown in Figure 2. The vertical axis represents residues for the human immune system protein, while the horizontal axis represents residues of the peanut allergen protein. In the leftmost

contact map in Figure 2, four potential sites on the human immune protein sequence are predicted to interact with one site on the allergen. By tracing back the indices of these residues, the specific sequence of sites likely involved in interaction could be identified.

Figure 2. Contact map of top 3 highest probability interactions

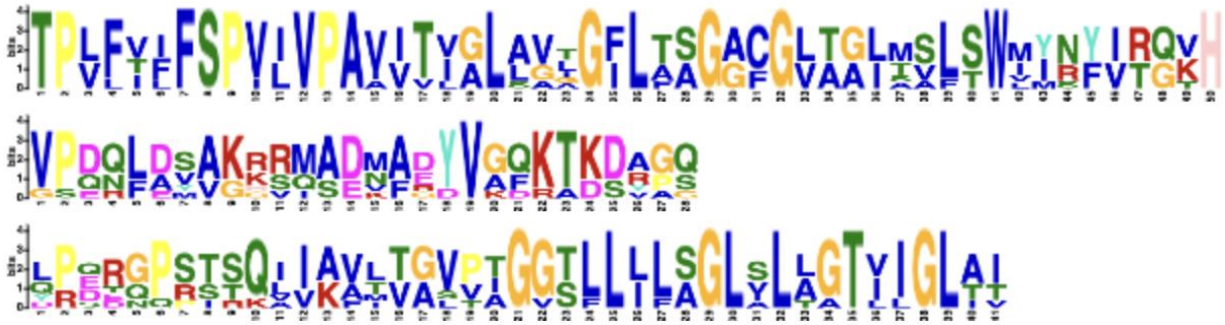


MEME analysis unveiled conserved motifs within human immune system proteins known to interact with Ara h allergens, providing insights into immune recognition mechanisms. Additionally, MEME identified potential allergen epitopes within peanut proteins that may trigger immune responses.

Figure 3. Top 3 most statistically significant discovered motifs

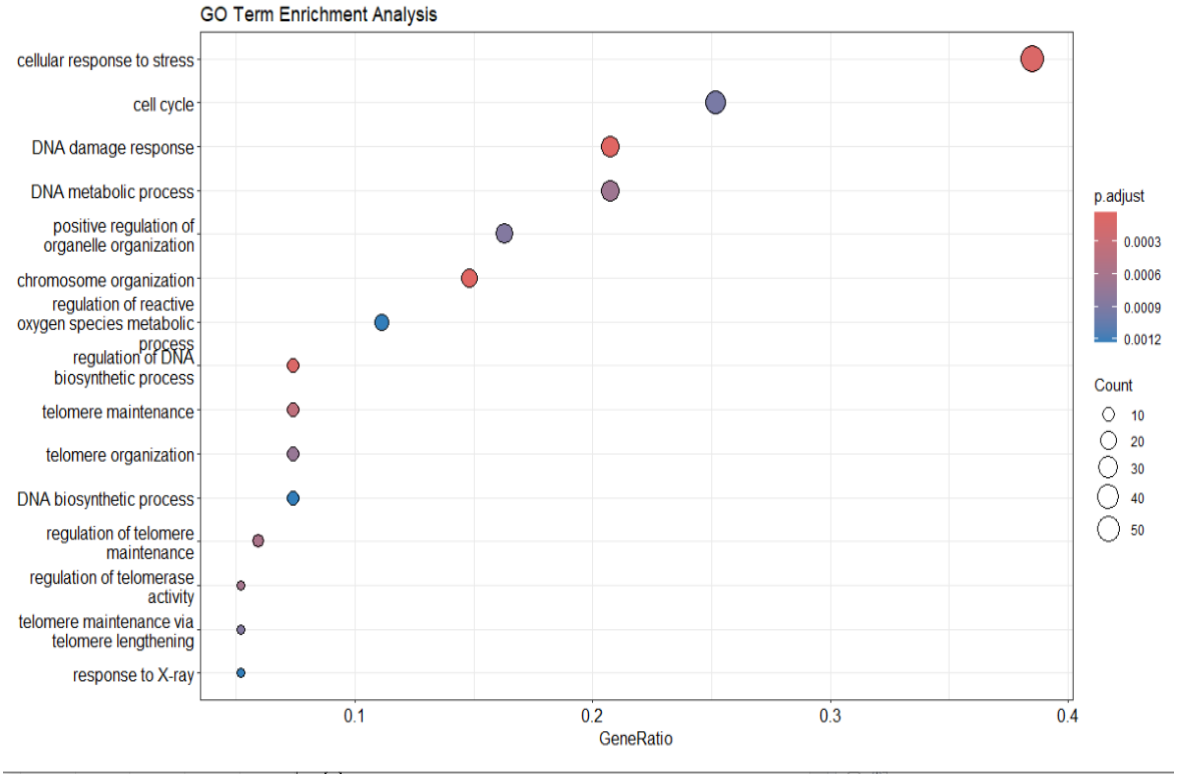
a. Human immune system proteins

b. Peanut allergen proteins



Gene Ontology Enrichment analysis provided functional insights to the high confidence binding pairs. This analysis revealed that the binding partners were most over represented in processes related to cellular response to stress, cell cycle, and DNA damage response. The background for this analysis was all screened proteins; not the entire transcriptome

Figure 4. GO Term Enrichment



The top ten overexpressed pathways using the Reactome database are shown below. The p-values must be taken with a grain of salt, as the reactome database compares inputted proteins against the entire transcriptome, and only a subset was screened against in this experiment.

Table 4. Top 10 Significant Pathways Associated with Human Binding Partners

Pathway identifier	Pathway name	Entities ratio	Entities pValue	Entities FDR

R-HSA-389948	PD-1 signaling	0.002119324	1.22E-06	0.002600155
R-HSA-380108	Chemokine receptors bind chemokines	0.003660651	1.69E-04	0.179658655
R-HSA-202430	Translocation of ZAP-70 to Immunological synapse	0.001926659	4.15E-04	0.294355795
R-HSA-933542	TRAF6 mediated NF-kB activation	0.001926659	8.05E-04	0.374633053
R-HSA-5218900	CASP8 activity is inhibited	7.71E-04	0.001077048	0.374633053
R-HSA-9670439	Signaling by phosphorylated juxtamembrane, extracellular and kinase domain KIT mutants	0.001798215	0.001232346	0.374633053
R-HSA-9669938	Signaling by KIT in disease	0.001798215	0.001232346	0.374633053
R-HSA-5603029	IkBA variant leads to EDA-ID	5.14E-04	0.00292464	0.6727235
R-HSA-450302	activated TAK1 mediates p38 MAPK activation	0.001733993	0.005701461	0.6727235
R-HSA-9652169	Signaling by MAP2K mutants	3.85E-04	0.006076858	0.6806007

Another interesting binding partner that was not involved in an over-represented pathway, but was significantly interacting with 10 separate peanut proteins was CD209. CD209 is a pattern recognition receptor that initiates a pathway that results in IgE antibodies.[9]

Discussion

In this paper, we present the utilization of the novel protein-protein interaction algorithm Topsy-Turvy. Over a million different combinations of proteins related to the human immune system and known peanut allergens were tested to determine their probability of interacting. Through this method, we identified over 5,000 potential novel interactions between peanut allergens and the human immune system. Of these, 356 were predicted with high confidence (>0.7). Interestingly, some allergens had many more predicted binding partners than others; specifically Arah3 (see table 1). This suggests that Arah3 may have a structure that more closely resembles typical immune protein binding partners than other peanut proteins. An alternative hypothesis to this is, since Arah3 has many moderate binding partners (~0.66), that Arah3's structure has many potential binding domains that are not highly specific. This preferential binding of Arah3

with confidence between 0.65-0.7 is the cause of the spike seen in the otherwise decreasing trend observed in Figure 1.

We used MEME to identify overrepresented motifs in both human proteins in high confidence pairs and peanuts. The motifs in human proteins represent likely binding sites for the peanut allergens. While most allergy responses are mediated through IgE antibodies, other secondary immune response triggers may play a role through these conserved motifs. The conserved motifs among peanut proteins did not align with any known epitopes, suggesting the potential discovery of novel epitopes.

The results of the GO term enrichment analysis reveal the function properties of the proteins that the peanut allergens preferentially bind to. The exact reasoning for this preferential binding is somewhat unclear, though we propose a hypothesis. Proteins related to the enriched GO terms could, on average, have more binding partners than those that are not enriched.

Reactome pathway analysis reveals the most overrepresented pathways that human proteins in high confidence pairs were involved in. The p-values for these results must be taken cautiously, as there is no option in the Reactome database to specify the reference set. The most overrepresented pathway, albeit with unclear confidance, is PD-1 signaling. Interestingly, PD-1 signaling plays a large role in immune system regulation. Interference with this pathway can lead to an increased immune response[9]. This may be why peanut allergies tend to have severe reactions. The triggering of IgE mediated immune response with interruption to the PD-1 signaling pathway could lead to the enhanced response seen in those sensitive to peanuts.

Another interesting pathway is the predicted binding of ten separate peanut proteins to CD209. CD209 is a pattern recognition receptor and is expressed on the surface of dendritic cells.[9] This receptor is the beginning of a pathway that results in the presentation of antigens to T-cells.[9] We hypothesize the possible binding of peanut allergens to CD209 could be responsible for initial peanut sensitization.

Conclusion

In conclusion, this study underscores the utility of computational methods in elucidating protein-protein interactions relevant to peanut allergy. With these methods, we were able to identify over 5000 novel inter-species PPIs. We identified the common motifs that contribute to these interactions, and predict they could serve as novel antigens. GO term enrichment was used to predict the functional properties of the human proteins that preferentially bind to peanut allergens. We were then able to use our predicted

interactions to investigate the pathways that lead to peanut allergy sensitization, and the accompanying severity that affects sensitive individuals. These findings contribute to our understanding of the complex biological interactions surrounding allergic immune response, and lay a foundation for further experimental studies aimed at developing targeted therapies for peanut allergy management.

References

1. Patel R, Koterba AP. Peanut Allergy. [Updated 2023 Jul 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK538526/
2. Lange L, Klimek L, Beyer K, Blümchen K, Novak N, Hamelmann E, Bauer A, Merk H, Rabe U, Jung K, Schlenter W, Ring J, Chaker A, Wehrmann W, Becker S, Mülleneisen N, Nemat K, Czech W, Wrede H, Brehler R, Fuchs T, Jakob T, Ankermann T, Schmidt SM, Gerstlauer M, Zuberbier T, Spindler T, Vogelberg C. White paper on peanut allergy - part 1: Epidemiology, burden of disease, health economic aspects. Allergo J Int. 2021;30(8):261-269. doi: 10.1007/s40629-021-00189-z. Epub 2021 Sep 28. PMID: 34603938; PMCID: PMC8477625..
3. Knight J. The lymphatic system 4: Allergies, anaphylaxis and anaphylactic shock [Internet]. 2020 [cited 2024 May 14]. Available from: https://www.nursingtimes.net/clinical-archive/immunology/the-lymphatic-system-4-allergies-anaphylaxis-and-anaphylactic-shock-14-12-2020/
4. The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2023, *Nucleic Acids Research*, Volume 51, Issue D1, 6 January 2023, Pages D523–D531, https://doi.org/10.1093/nar/gkac1052
5. https://www.allergen.org/search.php?allergensource=peanut&searchsource=Search
6. Rohit Singh, Kapil Devkota, Samuel Sledzieski, Bonnie Berger, Lenore Cowen, Topsy-Turvy: integrating a global view into sequence-based PPI prediction, Bioinformatics, Volume 38, Issue Supplement_1, July 2022, Pages i264–i272, https://doi.org/10.1093/bioinformatics/btac258
7. Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
8. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, Haw R, Jassal B, Matthews L, May B, Petryszak R, Ragueneau E, Rothfels K, Sevilla C, Shamovsky V, Stephan R, Tiwari K, Varusai T, Weiser J, Wright A, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Research. 2024. doi: 10.1093/nar/gkad1025.

9. Gupta RK, Gupta GS. DC-SIGN Family of Receptors. Animal Lectins: Form, Function and Clinical Applications. 2012 Mar 20:773–98. doi: 10.1007/978-3-7091-1065-2_36. PMCID: PMC7122914
10. Arasanz H, Gato-Cañas M, Zuazo M, Ibañez-Vea M, Breckpot K, Kochan G, Escors D. PD1 signal transduction pathways in T cells. Oncotarget. 2017 Apr 19;8(31):51936-51945. doi: 10.18632/oncotarget.17232. PMID: 28881701; PMCID: PMC5584302.