



# Distinguishing allergens from non-allergenic homologues using Physical–Chemical Property (PCP) motifs<sup>☆</sup>

Wenzhe Lu<sup>a</sup>, Surendra S. Negi<sup>a</sup>, Catherine H. Schein<sup>a</sup>, Soheila J. Maleki<sup>b</sup>, Barry K. Hurlburt<sup>b</sup>, Werner Braun<sup>a,\*</sup>

<sup>a</sup> Sealy Center for Structural Biology and Molecular Biophysics, Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555-0304, United States

<sup>b</sup> Southern Regional Research Center, U.S. Department of Agriculture, Agriculture Research Service, 1100 Robert E. Lee Boulevard, New Orleans, LA 70124, United States

## ARTICLE INFO

### Keywords:

Physical-chemical property motifs  
Allergens  
Birch pollen allergen  
Profilin  
Pectate lyase  
human microbiome

## ABSTRACT

Quantitative guidelines to distinguish allergenic proteins from related, but non-allergenic ones are urgently needed for regulatory agencies, biotech companies and physicians. In a previous study, we found that allergenic proteins populate a relatively small number of protein families, as characterized by the Pfam database. However, these families also contain non-allergenic proteins, meaning that allergenic determinants must lie within more discrete regions of the sequence. Thus, new methods are needed to discriminate allergenic proteins within those families. Physical–Chemical Properties (PCP)-motifs specific for allergens within a Pfam class were determined for 17 highly populated protein domains. A novel scoring method based on PCP-motifs that characterize known allergenic proteins within these families was developed, and validated for those domains. The motif scores distinguished sequences of allergens from a large selection of 80,000 randomly selected non-allergenic sequences. The motif scores for the birch pollen allergen (Bet v 1) family, which also contains related fruit and nut allergens, correlated better than global sequence similarities with clinically observed cross-reactivities among those allergens. Further, we demonstrated that the average scores of allergen specific motifs for allergenic profilins are significantly different from the scores of non-allergenic profilins. Several of the selective motifs coincide with experimentally determined IgE epitopes of allergenic profilins. The motifs also discriminated allergenic pectate lyases, including Jun a 1 from mountain cedar pollen, from similar proteins in the human microbiome, which can be assumed to be non-allergens. The latter lacked key motifs characteristic of the known allergens, some of which correlate with known IgE binding sites.

## 1. Introduction

The prevalence of allergic diseases has risen dramatically in the last few decades, especially in industrialized nations (Sicherer and Sampson, 2014; Platts-Mills, 2015). A study from the American Academy of Allergy Asthma & Immunology indicates that allergic diseases with medium to severe symptoms, including asthma, rhinitis, and sinusitis affect millions of people in the US. Food proteins induce severe allergic reactions in about 6% of US children under age 3 and 3.5–4% of the overall US population (Gupta et al., 2011). Severe anaphylaxis from drug and food allergies cause hundreds of deaths every year (Pleis et al., 2010). Some reactions to peanuts and other foods can lead to anaphylactic shock and, in extreme cases, fatalities. One of the major objections to introducing genetically modified foods and other products in our everyday life is the concern for potential risks of allergenicity in

those foods and related products. Thus, any new protein in food must be assessed for the risk of allergenicity. For these reasons, we have developed bioinformatics-based methods to help regulators and companies to identify potential hazards and minimize the chance of allergenicity and toxicity (Delaney et al., 2008; Goodman et al., 2008; Ivanciuc et al., 2009a).

The US FDA has established a list of the eight most common allergenic foods (milk, eggs, fish, crustacean shellfish, tree nuts, peanuts, wheat and soybeans) which account for 90% of food allergic reactions. Foods containing these ingredients must be labeled accordingly (FDA, 2004). It is also well established that allergens may be cross-reactive with other allergens (Schein et al., 2007; Bonds et al., 2008; Maleki et al., 2011). For example, individuals allergic to the major birch pollen allergen Bet v 1 may develop a secondary allergy to the similar protein in peanut, Ara h 8 (Mittag et al., 2004). Similarly, a variety of plant

<sup>☆</sup> Software Availability: The PCPmer software can be downloaded from the website at <http://bose.utmb.edu/software.html>.

\* Corresponding author.

E-mail address: [webraun@utmb.edu](mailto:webraun@utmb.edu) (W. Braun).

profilins may cause cross reactions (Radauer et al., 2006). One of the goals of our research is to use computational methods, combined with experimental data, to discriminate potentially cross-reactive proteins from harmless homologues.

Criteria for the evaluation of allergenicity by comparing a query protein with sequences of known allergens have been proposed by the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) (FAO/WHO, 2001; FAO/WHO, 2003). However, these guidelines tend to over-predict potential cross-reactivity with allergens, producing a large number of false positives (Bjorklund et al., 2005; Goodman, 2006; Silvanovich et al., 2006). This can greatly increase the burden of experimental validation which is expensive, time consuming and difficult to interpret (Goodman and Leach, 2004). Modifying these guiding criteria, for example, by increasing the cutoff value of 35% sequence identity and 6 exact-match residues (Ladics et al., 2007; Goodman, 2008), will also decrease the sensitivity of the method. Therefore, new bioinformatics tools based on classification algorithms or sequence profiles/motifs have been developed (Radauer, 2017). Methods using sequence profiles rely on nonspecific motifs derived from a limited number of known IgE epitopes. Here, we describe a more comprehensive approach, based on the collection of allergenic sequences in our Structural Database of Allergenic Proteins (SDAP) (Ivanciuc et al., 2003). We generated allergen-specific motifs based on conserved physical and chemical properties (PCP-motifs) for 17 protein domains, classified in the Pfam database (Finn et al., 2016). These 17 domains account for 80% of the well characterized major allergens. Here, we show how computational scoring of query sequences using these PCP-motifs, gives a more specific evaluation of potential allergenicity for novel proteins. The method can also be used to suggest potential cross-reactivity between different allergen sources, especially in combination with methods based on 3D protein structures (Negi and Braun, 2017).

## 2. Methods

### 2.1. Datasets

Using our collection of allergenic sequences in the SDAP, we identified 17 protein domains, as classified in the Pfam database (Finn et al., 2016) that contained 10 or more related allergens (Table 1). There are many iso-allergens which share highly similar sequences and their presence may bias the motif extraction results. To overcome this problem the CD-HIT (Li and Godzik, 2006) program was used to select representative sequences from each cluster at 95% sequence identity cutoff. Finally, the ClustalW program was used to generate multiple sequence alignments (Larkin et al., 2007). Sequence motifs specific for each family were generated using the PCPmer program, as described previously (Ivanciuc et al., 2009b; Schein et al., 2005; Mathura et al., 2003). We identified typically 5 to 10 motifs for each allergen family.

In order to validate our computational prediction method, we created another large dataset of non-allergenic proteins from the UniProtKB database (Magrane and Consortium, 2011). For this purpose, we used the CD-HIT program to cluster and select representative sequences with a 40% sequence identity cutoff. Potential allergenic proteins from this dataset were removed if the sequence met any of the following criteria: 1) the description of the sequence contained any keyword related to allergen (allergen, allergy, lipid transfer protein, profilin, lipocalin, pectate lyase, tropomyosin, melittin, thaumatin, seed storage protein); 2) the sequence belonged to any of the Pfam containing allergenic sequences, such as PF00235 or PF00407. In addition, sequences shorter than 50 amino acids were also removed. Using these criteria, we obtained a final dataset of 84,939 non-allergen sequences.

We also collected all the protein sequences for some of the protein domains in the PFAM database, e.g. the profilins. Those protein domains contain in general allergens and non-allergens. We then used our allergen specific PCP motifs and calculated motif scores as the average

of scores in all allergenic proteins (A) and for all sequences in a family (G). An example is given for the profilin family in Table 3 to show the specificity of those motifs.

### 2.2. Scoring method

The scoring method makes use of the physical-chemical property (PCP) descriptors E1 to E5 of the 20 amino acids as determined previously (Venkatarajan and Braun, 2001). The algorithm implemented in PCPmer to generate sequence motifs for a protein family was previously published (Ivanciuc et al., 2009b; Schein et al., 2005; Mathura et al., 2003). Briefly, PCPmer identifies motifs as linear segments with highly conserved descriptors E1 to E5 within a family of aligned sequences. These sequence motifs are quantitatively described with the average values  $\langle V_i^p \rangle$  and standard deviations  $\sigma_i^p$  of the five descriptors E1 to E5 (denoted here with  $p = 1 \dots 5$ ) at each position  $i$  in the motifs derived from the multiple sequence alignment. These motif profiles can then be used to evaluate any query sequence if they contain similar profiles. Sequences with similar motif profiles are likely to share similar functions or properties (Schein et al., 2005; Mathura et al., 2003; Garcia et al., 2009).

Using this strategy, we developed a new scoring method for the recognition of potential allergenic query sequences. To determine how well a motif matches a sequence, the motif is aligned to every position  $k$  of the query sequence and the score values  $S(k)$  are calculated based on a Lorentzian scoring scheme. For each position  $i$  within a motif of length  $n$ , a partial score  $S_{k,i}^p$  is first calculated as

$$S_{k,i}^p = \left[ 1 + \left( \frac{V_{k+i}^p - \langle V_i^p \rangle}{W\sigma_i^p + \Phi} \right)^2 \right]^{-1} \quad (1)$$

Here,  $k$  denotes the position of the motif in the query sequence,  $i$  is the position index within the motif which varies from 0 to  $n-1$ ;  $V_{k+i}^p$  are the five quantitative descriptors (E1-E5) of the amino acid at the position  $k + i$  in the sequence;  $\langle V_i^p \rangle$  and  $\sigma_i^p$  is the average PCP values and corresponding standard deviations of the five descriptors ( $p = 1 \dots 5$ ) at the position  $i$  of the motif,  $W$  is the weight for standard deviation (by default set to 1.5) and the small positive shift  $\Phi$  (set to 0.001) was added to prevent overflow during calculation when the standard deviation is zero.

The score value for a motif of length  $n$  aligned at the position  $k$  in the sequence is then calculated as the average of the partial scores of each position  $i$  in the motif.

$$S(k) = \frac{1}{5n} \sum_{p=1..5} \sum_{i=0, \dots, n-1} S_{k,i}^p \quad (2)$$

The resulting score  $S$  of a motif against a query sequence is the maximum of the score value calculated for each position  $k$ . The scores  $S$  are values from the interval 0 to 1, with 1 indicating a perfect match.

Usually more than one motif will be generated from a multiple sequence alignment. A total score is calculated to evaluate if a query sequence matches a set of  $m$  motifs with the scores  $S_1 \dots S_m$

$$S_{tot} = \sum_{j=1}^m \frac{S_j | \langle S \rangle_{aln} - \langle S \rangle_{db} |}{(\sigma_{aln} + \sigma_{db})^2 + \varepsilon} \quad (3)$$

Where  $j$  indicates the motif number,  $\langle S \rangle_{aln}$  and  $\sigma_{aln}$  are the mean and standard deviation of scores of that motif  $j$  in the multiple sequence alignment that was used to generate motifs;  $\langle S \rangle_{db}$  and  $\sigma_{db}$  are the mean and standard deviation of scores of that motif in the query set of sequences. If there is only one query sequence, i.e.  $\sigma_{db} = 0$ , a small positive shift  $\varepsilon$  (by default, set to  $10^{-10}$ ) is added to prevent division by zero. To evaluate the potential allergenicity of proteins, motifs for only allergenic sequences obtained from the 17 protein families (Table 1) were used as criteria for the scoring. A query sequence was scored against all 17 sets of motifs using the method described above, resulting in 17

**Table 1**

Allergens in the top 17 highly-populated protein families used for identifying PCP motifs in this study.

| Pfam    | No. of allergens | No. of sequences | Domain                                                           | Related food sources             |
|---------|------------------|------------------|------------------------------------------------------------------|----------------------------------|
| PF00234 | 35               | 109              | Plant lipid transfer proteins                                    | peanut, soybean, tree nut, wheat |
| PF00235 | 27               | 38               | Profilin                                                         | peanut, soybean, wheat           |
| PF01357 | 24               | 47               | Grass pollen allergens                                           | soybean, lobster, wheat          |
| PF00036 | 23               | 64               | Calcium-binding proteins                                         | fish                             |
| PF00407 | 20               | 114              | PR-10 like proteins                                              | peanut, soybean, fruits          |
| PF00188 | 19               | 19               | SCP-like proteins                                                |                                  |
| PF00190 | 16               | 45               | Cupins                                                           | peanut, soybean, tree nut        |
| PF00261 | 16               | 15               | Tropomyosins                                                     | crustacean shellfish             |
| PF03330 | 15               | 35               | Rare lipoprotein A (RlpA)-like double-psi beta-barrel            | soybean, lobster, wheat          |
| PF00061 | 13               | 16               | Lipocalins                                                       | domestic cattle                  |
| PF01190 | 10               | 15               | Pollen proteins, Ole e 1 family                                  | tomato                           |
| PF00544 | 9                | 17               | Pectate lyases                                                   |                                  |
| PF01620 | 8                | 21               | Ribonucleases (pollen allergens)                                 | barley                           |
| PF02221 | 8                | 15               | MD-2-related lipid-recognition (ML) domain                       |                                  |
| PF00187 | 5                | 17               | Chitin recognition proteins                                      | wheat                            |
| PF00273 | 5                | 20               | Serum albumin family                                             | chicken, domestic cattle         |
| PF06757 | 2                | 30               | Insect allergen related repeat, nitrile-specifier detoxification |                                  |

scores  $S_{tot,l}$ , where  $l = 1 \dots 17$ . Since the scores of the query sequence are based on motifs from different protein families,  $\langle S \rangle_{aln}$  and  $\sigma_{aln}$  are different for different families, which means that  $S_{tot,l}$  of the query sequence are not on the same scale. In order to compare the scores of the query sequence, the set of non-allergenic sequences from UniProt were used as background and the total score  $S_{tot,l}^b$  for each protein family was calculated. Then the scores of the query sequences were converted to standard scores  $z_l$  (eq 4) according to the background score distribution of non-allergenic sequences  $S_{tot,l}^b$  for each family.

$$z_l = \frac{S_{tot,l} - \langle S_{tot,l}^b \rangle}{\sigma_{tot,l}^b} \quad (4)$$

Finally, the highest  $z$  score (eq 5) of the query sequence was selected to indicate the extent of allergenicity.

$$z = \max(z_l) \quad (5)$$

### 2.3. Multidimensional scaling of experimental cross-reactivity data

The clinical cross-reactivities among the birch pollen allergen Bet v 1 and several fruit and nut allergens reported by Pfiffner et al (Pfiffner et al., 2012), are given by percentages  $P_{ab}$  within those patient sera that are positive against an allergen  $a$ , and are also positive against allergen  $b$ . These data may not be symmetric, and  $P_{ab}$  does not necessarily equal  $P_{ba}$  because the extent of binding of the allergen  $a$  to anti- $b$  IgE antibody is not necessarily the same as the binding of  $b$  to anti- $a$  IgE. We defined a symmetric distance between the pair of allergens  $a$  and  $b$  as the geometric mean,

$$D = -\log \sqrt{P_{ab} \cdot P_{ba}} \quad (6)$$

When there are higher percentages of patients positive to the pair of allergens, the distance is smaller. We further visualized the resulting multidimensional distance matrix on a 2D scale (MDS) to show a “map” of the allergens corresponding to experimentally determined cross-reactivity. See Section 3.2 for further details.

### 2.4. Binding efficiency test

To validate the significance of the motifs identified by PCPmer, we tested the binding of peptides containing these motifs to IgE from a peanut allergic patient. Peptide microarray assays are powerful tools to study the binding efficiency for many different peptides. Peptide sequences overlapping the motifs were synthesized and spotted onto glass microarray slides by JPT Peptide Technologies (Berlin, Germany). For the assay, slides were placed in the slide holders within the hybridization chamber of the HS 400 Pro (Tecan, Austria, GmbH,

Salzburg, Austria). The blocking solution, serum, and other solutions were injected into each slide chamber through the injection port. Slides were blocked for one hour at 4 °C, washed three times and incubated with centrifuged serum from a well-characterized patient highly allergic to peanut, with IgE to Ara h 2 and Ara h 5 peptides (Patient 4 from Maleki et al., 2011) at 4 °C overnight (14–16 hours). The slide was then washed. Next, mouse-anti-human IgE (murine IgG antibody; Southern Biotech, Birmingham, AL) was added and the slide was incubated for 1 hour at room temperature in the dark, then washed as before and incubated with labeled anti-mouse IgG Cy3 (green dye read at 532 nm) (Invitrogen, Grand Island, NY) for another 1 hour at room temperature in the dark. The slides were washed and dried by centrifugation prior to scanning with a GenePix-4000B scanner and GenePix-Pro7 software. The binding of patient's IgE to the peptides was measured by the fluorescence signal of Cy3. The resultant fluorescence intensities were organized in a spreadsheet for later statistical analysis.

## 3. Results

### 3.1. Allergenic-specific motifs for 17 protein families highly populated with allergenic proteins

Previously, we annotated most of the allergenic sequences in the SDAP database according to their protein domain classification in Pfam (Finn et al., 2016). Even though there are more than 1300 allergenic sequences in SDAP, they populate only a relatively small number of PFAM domains (Ivanciuc et al., 2009b). Here, we selected 17 protein families, each containing more than 10 allergenic sequences (Table 1). To avoid a bias by iso-allergens, we also removed sequence pairs with more than 95% sequence identity, and kept only one representative sequence for groups of highly related sequences. Multiple sequence alignments of each group were prepared with ClustalW. A set of PCP-motifs were generated from the alignment of allergenic sequences in each family with PCPmer (Mathura et al., 2003). The PCP motifs for those 17 protein allergens are given in Fig. S1 in the Supplementary material. These 17 protein families listed in Table 1 (FDA (2004)).

To determine the specificity of these allergen-specific motifs for distinguishing allergenic sequences from non-allergenic ones, we used the scoring method described in the Method section to compare query sequences to them. We generated scores for a large set of non-allergenic sequences from Uniprot and for all the allergenic sequences from SDAP (Fig. 1). The distribution of scores shows a clear separation between non-allergenic sequences and allergenic sequences from the 17 protein families, indicating the ability of the method to distinguish allergenic sequences from the background. Our motif database only contains motifs for the 17 families (Table 1), thus allergens not related to those

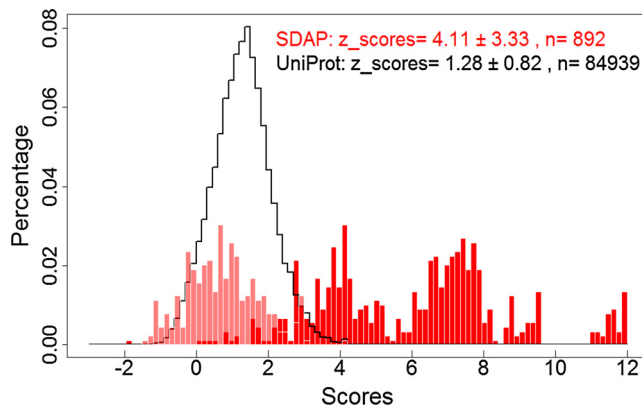


**Fig. 1.** Z-scores distinguish the allergenic members of the top 17 families from which the PCP motifs were extracted (dark red bars, Z-score =  $4.11 \pm 3.33$ ) from 84,939 non-allergenic sequences from Uniprot (black line,  $1.28 \pm 0.82$ ) and allergenic sequences from other allergen families (light red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

families cannot currently be identified. However, we hope to overcome this limitation in the future. As homologues of allergens from less populated Pfam are sequenced, they will be added to our database and used for further analysis.

To estimate an empirical cutoff score for the distinction between allergenic and non-allergenic proteins, we considered all the allergen protein sequences in the 17 protein families as true entities and all non-allergenic sequences in the UniProt database as false entities, and determined the true positive rates (recalls) and false positive rates at different score values (Fig. 2). As the cutoff score increases, the positive rates for both allergenic sequences and non-allergenic sequences decreases. The maximum difference of positive rate occurs at a score value of 2.7, where above this score, only 4.52% of the non-allergenic sequences are reported as false positives and 92.96% of the allergenic sequences in the top 17 protein families are detected.

### 3.2. Correlation between motif scores and clinically determined cross-reactivities in the birch pollen Bet v 1 family

Detection of cross-reactivity between allergens is of great interest because similar allergens can be found in diverse sources. One of the

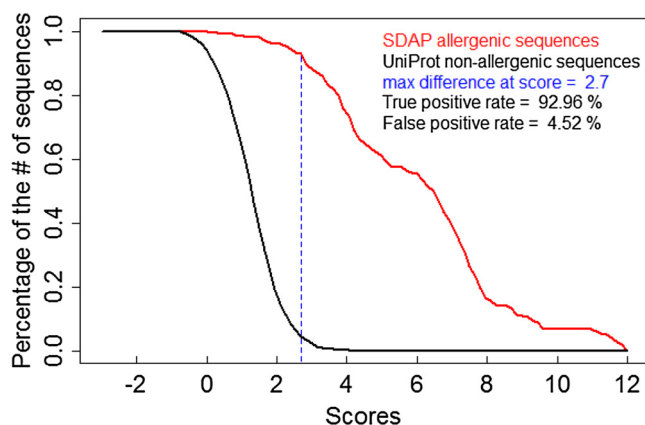


**Fig. 2.** PCP-motif based Z-scores differentiate allergens in the major families from non-allergenic sequences. Red: Z-scores for cumulative percentage of allergenic sequences in the top 17 protein families from SDAP; Black: non-allergenic sequences from Uniprot; Blue: maximum difference between the percentage of allergenic sequences and non-allergenic sequences at score = 2.7 (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

major advantages of our motif approach, compared to simple global sequence comparison, is the sensitivity of the scores to individual regions of the protein that are related to the allergenic effect. To demonstrate this effect we used the empirical quantitative data of the amount of clinical cross-reactivities in the Bet v 1 family as found in a large scale study of human sera (Piffner et al., 2012). Clinical cross-reactions can occur when an individual sensitized to one allergen is exposed to other allergens that contain allergens with similar sequence or 3D-structure (Aalberse and Serology, 2015; Aalberse, 2007; Bublin and Breiteneder, 2014; Bublin et al., 2013; Pichler et al., 2015; Gadermaier et al., 2011; Ferreira et al., 2004; Jenkins et al., 2005). Severe symptoms may develop in individuals who unknowingly consume foods containing cross-reactive allergens. For example, the birch pollen allergen, Bet v 1, is cross-reactive with several fruits and nut allergens, including Pru av 1 from cherry, Mal d 1 from apple, Gly m 4 from soybean, and Ara h 8 from peanut (Mittag et al., 2004; Mittag et al., 2005; Mittag et al., 2006; Egger et al., 2006; Bolhaar et al., 2005).

To determine the ability of our motifs to classify potentially cross-reactive allergens, we used a dataset (Fig. 5 (Piffner et al. (2012))) that assayed the binding of IgE in patient sera to chips coated with Bet v 1 or several different proteins in the same family. Their results revealed a hierarchy in the degree of cross-reactivity among allergens and identified six allergens, Bet v 1, Cor a 1.0401, Cor a 1.0101, Pru p 1, Aln g 1 and Mal d 1, which share a greater portion of patients who are positive to them than the other allergens. We converted the percentages of patients who reacted positively to an allergen, to a distance matrix representing the extent of cross-reactivity by using equation 6 (Table S1). A small distance between two allergens means that a high percentage of patients cross-reacted to that pair of allergens.

We next determined to what extent the allergen-specific motifs from the Bet v 1 family matched each individual allergen sequence, which was calculated as a z-score (Table 2). Fig. 3 “maps” the experimentally determined clinical cross-reactivities as points in a 2-dimensional plot, where highly cross-reacting allergens are located near each other in the map. Six highly cross-reactive allergens, allergens 1 to 6, cluster to the top left of the map, and those allergens also have the highest z-scores, colored in red. The less cross-reactive allergens are scattered throughout the rest of the map with low scores.

All of 10 of these allergens related to Bet v 1 have PCP-motif based Z-scores greater than 5, which clearly distinguishes them from non-allergenic proteins. In addition, the results imply that allergens most likely to cross react will have high Z-scores ( $> 6$ ), independent of their absolute sequence identity. The rationale for this observation is that those allergens are highly similar in those sequence regions that we identified by our analysis to be allergen-specific. To show that this similarity found by our z-score is not simply a consequence of a high global sequence similarity between those allergens, we compared our z-scores to the sequence identities between Bet v 1 as a representative of the 6 allergens in the most cross-reacting cluster to all other allergens (last column in Table 2). The scatter plot of the z-scores versus the sequence identities (Fig. 4) shows the high similarities of the z-scores

**Table 2**

Z-scores and sequence identities to Bet v1 for the ten cross-reacting allergens related to Bet v1.

| Allergen     | Uniprot identifier | Pfam identifier | z-scores | Seq ID to Bet v1 |
|--------------|--------------------|-----------------|----------|------------------|
| Bet v 1      | P45431             | BEV1B_BETPN     | 6.94     | 100              |
| Cor a 1.0401 | Q9SWR4             | Q9SWR4_CORAV    | 7.12     | 68.1             |
| Cor a 1.0101 | Q39454             | Q39454_CORAV    | 7.24     | 72.5             |
| Pru p 1      | Q216V8             | Q216V8_PRUPE    | 7.12     | 61.3             |
| Aln g 1      | P38948             | MPAG1_ALNGL     | 7.01     | 83.8             |
| Mal d 1      | P43211             | MAL11_MALDO     | 6.13     | 59.1             |
| Ara h 8      | B0YIU5             | B0YIU5_ARAHY    | 5.35     | 48.4             |
| Gly m 4      | P26987             | SAM22_SOYBN     | 5.72     | 50.0             |
| Api g 1      | P49372             | ALL1_APIGR      | 5.26     | 42.9             |
| Dau c 1      | O04298             | DAU1_DAUCA      | 5.09     | 40.3             |

**Table 3**

Allergen specific motifs in the profilin family and corresponding Ara h 5 peptides used for identifying their IgE reactivities. Second column: Sequences of Ara h 5, the profilin of peanut allergens, representing the allergen specific motifs in the profilin family. Column A: the average of the scores S, calculated by Eq. (2) for the sequences of allergenic profilins Column G: the average S scores for all sequences in the complete profilin family (PF00235). Column SA: the solvent accessibility of the peptides in the 3D structure of Ara h 5. Column 6 and 7: 15mer peptides of Ara h 5 overlapping with the motif regions and their corresponding IgE binding, measured as the average of 3 fluorescence intensity ( $F_{532}$ ) readings on the microarray for a highly peanut allergic individual, most reactive to peptides of Ara h 2 and Ara h 5. Peptides with IgE binding significantly higher than the average background level of 130,000 are highlighted.

| ID | Sequence in Ara h 5                  | A    | G    | SA% | Peptide sequence | IgE binding |
|----|--------------------------------------|------|------|-----|------------------|-------------|
| 1  | 2 WQTYVDNHLHC 12                     | 0.97 | 0.58 | 36  | MSWQTYVDNHLHCEI  | 833,465     |
| 2  | 26 GQDGGVWAQS 35                     | 0.97 | 0.63 | 28  | YVDNHLHCEIEGDHL  | 1,675,387   |
| 3  | 38 FPQFKPEE 45                       | 0.94 | 0.58 | 48  | LGQDGGVWAQSSHFP  | 304,733     |
| 4  | 56 PGSLAPTQ 63                       | 0.99 | 0.57 | 43  | SSHFPQFKPEETAI   | 88,464      |
| 5  | 71 YMVIGQE 77                        | 1.00 | 0.55 | 52  | GVWAQSSHFPQFKPE  | 1,780,267   |
| 6  | 81 IIPGKKPGGGVT 92                   | 0.98 | 0.62 | 21  | MNDFAEPPGSLAPTGL | 66,548      |
| 7  | 111 PGQCNM 116                       | 0.97 | 0.56 | 37  | KYMVIQGEPGAIIIPG | 64,903      |
| 8  | 119 ERLGDY 124                       | 0.99 | 0.62 | 36  | AIIIPGKKPGGGVTIE | 133,896     |
| *  | 103 GIYDKPMTPGQCNMIVERLGDY124        | 0.98 | 0.55 | 35  | TPGQCNMIVERLGDY  | 56,629      |
|    | * C-terminal motif, contains 7 and 8 |      |      |     | NMIVERLGDYLIDTG  | 104,354     |
|    |                                      |      |      |     | KTNQALIGIYDKPM   | 324,851     |
|    |                                      |      |      |     | LIIGIYDKPMTPGQC  | 151,028     |
|    |                                      |      |      |     | RLGDYLIDTGL      | 647,289     |

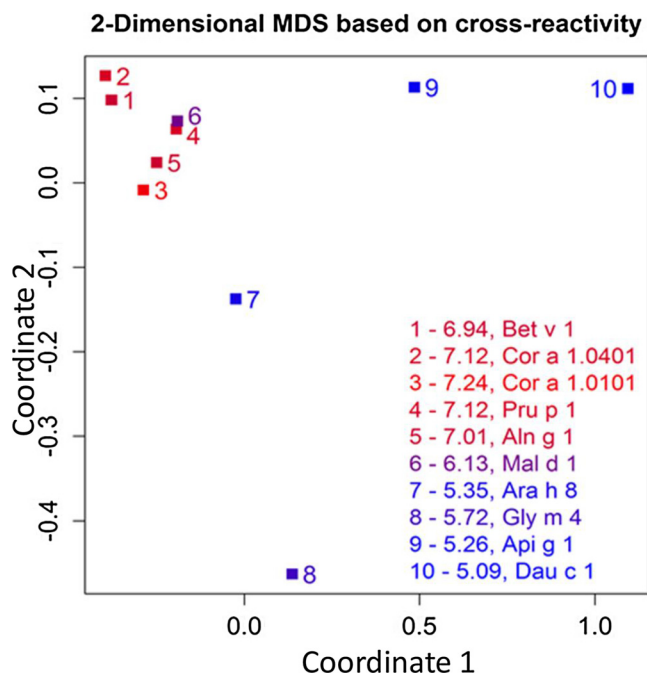


**Fig. 3.** PCP-motif based scores clearly distinguish cross-reacting allergens (red) from other Bet v 1 homologues (blue). 2-dimensional scaling of the distances (Table S1) of clinical cross-reactivities for the 10 known allergens in Table 2. Each 2D data point represents an allergen with the colors indicating the z-scores, where red points are high scoring (z-score > 6) and blue have low Z-scores. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

within the top 5 allergens. The values are all in the range of 7 except for a slightly lower value for Mal d 1, whereas the overall sequence identities varies for those allergens in a wide range from 60% to 100%. Thus, the z-scores correlate better with the experimental data on cross reactivity than could be predicted from overall sequence identity alone.

### 3.3. Comparison of allergen specific motif scores with IgE binding affinities to serum from a peanut allergic patient

In the next test for the biological relevance of our motifs in the profilin family, we wanted to know to what extent high scoring motifs can be used to predict IgE binding affinities of the corresponding

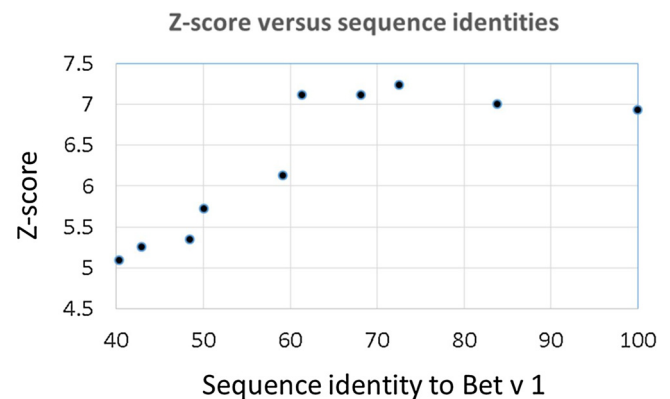


**Fig. 4.** Scatterplot of the z-scores based on the PCP motifs for the allergens related to Bet v 1 versus their sequence identities to Bet v 1. Bet v 1 was used as a representative of the highly cross-reactive allergens that cluster to the upper left corner in Fig. 3.

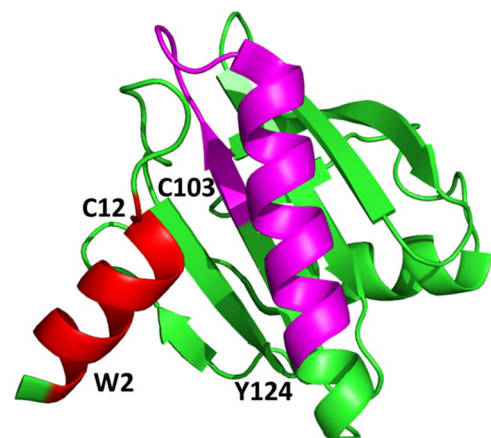


**Fig. 5.** The N- and C-terminal motifs of the profilin family (red and purple areas) map to one face of the 3D-crystal structure of the peanut allergen Ara h 5 (PDB entry: 4ESP) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

peptides, and if the motif scores distinguish allergenic from non-allergenic profilins. Profilins (Pfam ID PF00235) are a large group of diverse proteins, some of which are allergens from plants such as timothy grass,

mugwort, celery or bell pepper. They have been implicated in sensitization to both pollen and pollen-associated food allergens (Radauer et al., 2006; Cabanos et al., 2010). However, most profilins, such as those from mammals, are not considered allergenic (Valenta et al., 1991). At least 17 different peanut proteins have been characterized as allergens (Mueller et al., 2014; Schein et al., 2010; Nesbit et al., 2012; Chen et al., 2016), including the peanut profilin Ara h 5. We derived 8 allergen-specific motifs from the allergenic profilins in SDAP (Table 3, second column). The 8 motifs represent all allergenic profilins and are specified here by the corresponding peptide sequence in Ara h 5. These motifs are highly specific for the allergenic subfamily. To support this statement, we calculated for each of the 8 allergen-specific motifs the scores for the individual allergenic profilin sequences and list their average values over all allergenic sequences in Table 3, column A. We repeated this procedure for all non-allergenic profilin sequences and determined those averages (Table 3, column G). The average motif scores for those regions are significantly higher for the allergenic members than the average scores over all non-allergenic profilins.

Overlapping 15mer peptides from the Ara h 5 sequence containing these motifs were synthesized on a peptide microarray, as described in Methods. Table 3 shows the IgE binding specificity for peptides containing the motifs at different positions, measured using serum from a patient with severe peanut allergy, whose IgE recognizes peptides from Ara h 2 and Ara h 5 most strongly on the microarrays (Maleki et al., 2011). Six peptides, highlighted in Table 3, bind IgE with an intensity significantly above the background level. The positions of the most strongly recognized peptides in Ara h 5 cluster at the N and C termini of the protein (Wang et al., 2013) (Fig. 5) and correspond approximately to IgE epitopes identified for the birch pollen profilin Bet v 2 (Fedorov et al., 1997). These peptides represent conformationally constrained areas within the protein structure, which may be retained to some extent in the 15mer peptides used. Although the highest binding peptides (containing motifs 1 and 3) also had reasonable surface exposure (SA) in the Ara h 5 3D structure, there was no direct correlation between their overall surface accessibility SA and IgE binding. The degree of IgE binding also varied depending on the position of the motif within the 15mer peptide sequence.

### 3.4. Scores of sequences from the human microbiome

The microorganisms that make up the human microbiome outnumber the cells of the human body by an order of magnitude (Savage, 1977; Relman and Falkow, 2001) and play vital roles in maintaining health. To determine whether there were sequences related to known allergens, we scanned a database of protein sequences identified from the microbial communities that make up the human microbiome (Human Microbiome Project, 2012; Group et al., 2009; Gevers et al., 2012). The data are based on samples collected from the airways, skin, oral cavity, gastrointestinal tract and vagina of healthy humans. As these organisms exist in our body without causing immune response or allergic reaction, their proteins can be considered as non-allergenic.

We previously established several pollen pectate lyases as allergens (Midoro-Horiuti et al., 2003). These enzymes degrade plant cell walls (Marin-Rodriguez et al., 2002) and are found in plant pathogens and many plant tissues, including pollen and ripening fruits (Wing et al., 1990). Bacteria in the digestive tract of animals, who feed on plants and many in the normal human microbiome also produce and secrete pectate lyases (Hugouvieux-Cotte-Pattat et al., 2014), which may aid in digestion in the gut. BLAST searches using the pollen allergen Jun a 1 revealed a test set of presumably non-allergenic pectate lyases from the microbiome database that contained domains of the same protein family (PF00544) as our known allergens. We then compared the motif scores for sequences in the complete pectate lyase family to those in the subset of allergenic pectate lyases, including those of Amb a 1, Cup a 1, Jun a 1 and others (Midoro-Horiuti et al., 1999; Griffith et al., 1991; Aceituno et al., 2000) as catalogued in SDAP, and in the subset of Jun a

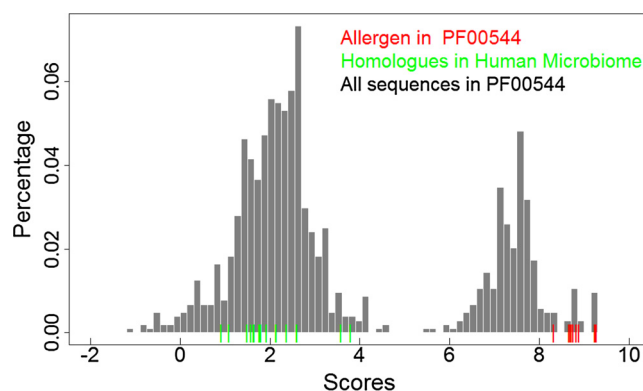


Fig. 6. Homologues of the allergenic pectate lyases in the human microbiome (green bars) have low PCP-motif based z-scores compared to the allergenic sequences in that family (red bars). As a comparison we also calculated the scores for all 1037 protein sequences in the complete pectate lyase family (PF00544) (grey bars). Red: z-scores for the 17 allergenic sequences from SDAP; Green: z-scores for the 18 homologues in the human microbiome (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

1 homologs in the microbiome.

We identified 18 homologues of the mountain cedar pollen allergen Jun a 1 sequence in the human microbiome database with an E-value below  $10^{-3}$  (Table S2). All sequences of the pectate lyase family (PF00544) were downloaded from the Pfam database as background. Scores were calculated with our allergen specific motifs for all sequences in the three sets: the allergenic pectate lyases, the homologs in the human microbiome and the complete pectate lyase family. The distribution of the score values in Fig. 6 indicates a clear separation between allergenic sequences in PF00544 (with scores in the range from 8 to 10, red), and their homologues in the human microbiome (with values between 1 and 4, green). The scores for all the sequences in PF00544 show a bimodal distribution. One high scoring group is clustered close to allergenic sequences, thus there might be other potential allergens in this family. These would be worthwhile testing in future studies, to refine our knowledge of potentially cross-reacting allergens in this large family. The majority of sequences are near the non-allergenic, microbiome sequences (black). The bimodal distribution probably reflects also the difference of plant and bacterial enzymes. Overall, our method gives better results for distinguishing allergenic pectate lyases than the current FAO/WHO guidelines, which correctly detect all allergenic pectate lyases, but indicated that 15 of our 18 non-allergenic homologues would also be allergenic (false positive rate of 83%). Our PCP-motif-based method can minimize false positive rates while recognizing all allergenic sequences.

## 4. Discussion

The most important challenge of our bioinformatics analysis is to distinguish allergenic proteins from non-allergens belonging to the same protein family that share a similar 3D structure. We show here, using several different examples, that Z-scores based on PCP motifs can distinguish known allergenic proteins from non-allergens. As Fig. 1 shows, a cumulative score value based on motifs from 17 different allergen families can distinguish allergens from non-allergens in a global analysis. The motifs also provide a tool to determine allergens likely to cross-react, based on analysis of available experimental evidence for Bet v 1 and related fruit and nut allergens (Fig. 3). More detailed analysis of individual allergen families shows how specific motifs may correlate with IgE epitopes (Table 3). They can be used to distinguish allergens from non-allergenic proteins, even when these have similar enzymatic motifs as the pectate lyases from the human microbiome (Fig. 6 and Supplementary material Table S2).

The method is particularly useful if several allergenic protein sequences have been experimentally verified within a family of homologous proteins. Whereas most bioinformatics tools group proteins based on overall sequence and structure, our motif methods dissect the protein families even further to define more subtle characteristics of the allergenic proteins. For allergens in each family, we typically find multiple motifs of short length (5–15 amino acids), which can serve as fingerprints to reduce the rate of false positives. Therefore, our method has higher sensitivity in distinguishing allergenic from non-allergenic sequences with high overall similarity.

Other motif-based methods have been proposed, based on the MEME software (Pfeiffer et al., 2012; Stadler and Stadler, 2003), or wavelet transforms and Hidden Markov Model (HMM) profiles in WebAllergen (Li et al., 2004; Riaz et al., 2005). However, the motifs generated by both methods are in general longer. For example, MEME uses peptides with a length of ~50 (MEME) and WebAllergen > 30. Such long peptides represent protein domains, not motifs by our definition, and are therefore less specific, as our results show. A second type of method is based on machine learning and classification algorithms. Applications such as Allerdicator (Dang and Lawrence, 2014), SORTALLER (Zhang et al., 2012), or AllerHunter (Muh et al., 2009) construct different feature vectors by decomposing sequences in short sections and evaluate the frequencies of amino acids and/or short peptides. The accuracy of these methods depends on how similar the query sequence is to the known allergens, which may reduce their selectivity (Schein et al., 2007).

The results from our motif scoring tests suggest IgE reactive segments within known allergens can be verified with further experiments. As the motifs also include areas conserved solely for protein function or structure, we cannot assume that there will be a one-to-one correspondence between all high-scoring motifs and their ability to bind to the IgE in serum from any individual patient. Further microarray analysis, using sera from many patients with high IgE to the same protein, will be needed to validate the IgE reactive segments and determine the predictive value of similarity to any of the motifs. It is also well known that most of the IgE epitopes are conformational in nature, therefore new refinement of methods will be needed to identify such epitopes. The PCP-motif approach as outlined here can be even more precise, if coupled with other recently developed methods, (Radauer, 2017; Mirsky et al., 2013; Garino et al., 2016), and 3D structural analysis (Negi and Braun, 2017; Dall'Antonia et al., 2011).

The most significant clinical use of the method will be in predicting likely cross-reactivities between known allergens (Bonds et al., 2008; Hamilton and Kleine-Tebbe, 2015). One advance here, shown in Fig. 3, is a way to directly convert patient-based data into quantitative maps. The results can be used to determine likely cross-reactive allergens, for use in counseling patients on which foods they may safely consume or those they should avoid.

## 5. Conclusion

We have developed a novel method to quantitatively estimate the potential allergenicity of query sequences using allergen-specific motifs for major allergens in highly populated protein families. The scoring method is able to distinguish non-allergenic and allergenic sequences, even those with significant overall similarity. The motifs overlap with experimentally known IgE epitopes for peptides from peanut and other allergens, as illustrated by literature examples and peptide microarray data presented here. The allergenicity scores for pectate lyase allergens and their homologous non-allergenic sequences from the human microbiome in PF00544 show a clear separation, consistent with the hypothesis that proteins of resident microbes are tolerated. Scores for individual allergen sequences against motifs derived for allergens related to Bet v 1 correlated well with experimental data on their clinically observed cross-reactivity. The motif list can be improved in the future to include members of less studied allergen families. The ability

of the method to predict likely IgE epitopes and cross reactivities can also be improved by combining the linear motifs with predicted conformational IgE epitopes, and by using methods to identify cross-reactive surface patches of allergens.

The PCPmer program is available online at <http://bose.utmb.edu/>.

## Conflict of interest

None declared.

## Acknowledgements

We thank Hsiaopo Cheng for expert technical assistance. This work was supported by grants from the National Institutes of Health (R21 AI1090909, to WB), the U.S. Environmental Protection Agency for STAR Research Assistance Agreements (No. RD834823, to WB and RE-83406601 to CHS), and the U.S. Department of Agriculture, Agricultural Research Service (to SJM, BKH and CHS). The article has not been reviewed by the EPA, and the views expressed in this document are solely those of the authors.

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: <https://doi.org/10.1016/j.molimm.2018.03.022>.

## References

- Aalberse, R.C., 2007. Assessment of allergen cross-reactivity. *Clinical Molecular Allergy* : CMA 5, 2.
- Aalberse, R.C., Serology, Shrimp, 2015. We need tests with more and less cross-reactivity. *J. Allergy Clin. Immunol. In Pract.* 3, 530–531.
- Aceituno, E., Del Pozo, V., Minguez, A., Arrieta, I., Cortegano, I., Cardaba, B., Gallardo, S., Rojo, M., Palomino, P., Lahoz, C., 2000. Molecular cloning of major allergen from *Cupressus arizonica* pollen: Cup a 1. *Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol.* 30, 1750–1758.
- Bjorklund, A.K., Soeria-Atmadja, D., Zorzet, A., Hammerling, U., Gustafsson, M.G., 2005. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics* 21, 39–50.
- Bolhaar, S.T., Zuidmeer, L., Ma, Y., Ferreira, F., Bruijnzeel-Koomen, C.A., Hoffmann-Sommergruber, K., van Ree, R., Knulst, A.C., 2005. A mutant of the major apple allergen, Mal d 1, demonstrating hypo-allergenicity in the target organ by double-blind placebo-controlled food challenge. *Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol.* 35, 1638–1644.
- Bonds, R.S., Midoro-Horiuti, T., Goldblum, R., 2008. A structural basis for food allergy: the role of cross-reactivity. *Curr. Opin. Allergy Clin. Immunol.* 8, 82–86.
- Bublin, M., Breiteneder, H., 2014. Cross-reactivity of peanut allergens. *Curr. Allergy Asthma Rep.* 14.
- Bublin, M., Kostadinova, M., Radauer, C., Hafner, C., Szepefalusi, Z., Varga, E.M., Maleki, S.J., Hoffmann-Sommergruber, K., Breiteneder, H., 2013. IgE cross-reactivity between the major peanut allergen Ara h 2 and the nonhomologous allergens Ara h 1 and Ara h 3. *J. Allergy Clin. Immunol.* 132, 118–124.
- Cabanas, C., Tandang-Silvas, M.R., Odijk, V., Brostedt, P., Tanaka, A., Utsumi, S., Maruyama, N., 2010. Expression, purification, cross-reactivity and homology modeling of peanut profilin. *Protein Express. Purif.* 73, 36–45.
- Chen, X., Negi, S.S., Liao, S., Gao, V., Braun, W., Dreskin, S.C., 2016. Conformational IgE epitopes of peanut allergens Ara h 2 and Ara h 6. *Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol.* 46, 1120–1128.
- Dall'Antonia, F., Gieras, A., Devanaboyina, S.C., Valenta, R., Keller, W., 2011. Prediction of IgE-binding epitopes by means of allergen surface comparison and correlation to cross-reactivity. *J. Allergy Clin. Immunol.* 128, 872–879 e878.
- Dang, H.X., Lawrence, C.B., 2014. Allerdicator: fast allergen prediction using text classification techniques. *Bioinformatics* 30 (8), 1120–1128. <http://dx.doi.org/10.1093/bioinformatics/btu004>.
- Delaney, B., Astwood, J.D., Cunney, H., Conn, R.E., Herouet-Guichene, C., Macintosh, S., Meyer, L.S., Privalle, L., Gao, Y., Mattsson, J., Levine, M., Safety, I.I.F.B.C.T.Fo.P., 2008. Evaluation of protein safety in the context of agricultural biotechnology. *Food Chemical Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc.* 46 (Suppl. 2), S71–S97.
- Egger, M., Mutschlechner, S., Wopfner, N., Gadermaier, G., Briza, P., Ferreira, F., 2006. Pollen-food syndromes associated with weed pollinosis: an update from the molecular point of view. *Allergy* 61, 461–476.
- FAO/WHO, 2001. Evaluation of Allergenicity of Genetically Modified Foods, Report of a Joint FAO/WHO Expert Consultation.
- FAO/WHO, 2003. Report of the Fourth Session of the Codex Ad Hoc Intergovernmental Task Force on Foods Derived from Biotechnology.
- FDA, 2004. Food Allergen Labeling and Consumer Protection Act of 2004 (Public Law 108-282, Title II).

- Fedorov, A.A., Ball, T., Mahoney, N.M., Valenta, R., Almo, S.C., 1997. The molecular basis for allergen cross-reactivity: crystal structure and IgE-epitope mapping of birch pollen profilin. *Structure* 5, 33–45.
- Ferreira, F., Hawranek, T., Gruber, P., Wopfner, N., Mari, A., 2004. Allergic cross-reactivity: from gene to the clinic. *Allergy* 59, 243–267.
- Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–285.
- Gadermaier, G., Hauser, M., Egger, M., Ferrara, R., Briza, P., Santos, K.S., Zennaro, D., Girbl, T., Zuidmeer-Jongejan, L., Mari, A., Ferreira, F., 2011. Sensitization prevalence, antibody cross-reactivity and immunogenic peptide profile of Api g 2, the non-specific lipid transfer protein 1 of celery. *PLoS One* 6, e24150.
- Garcia, T.I., Oberhauser, A.F., Braun, W., 2009. Mechanical stability and differentially conserved physical-chemical properties of titin Ig-domains. *Proteins* 75, 706–718.
- Garino, C., Coisson, J.D., Arlorio, M., 2016. In silico allergenicity prediction of several lipid transfer proteins. *Comput. Biol. Chem.* 60, 32–42.
- Gevers, D., Knight, R., Petrosino, J.F., Huang, K., McGuire, A.L., Birren, B.W., Nelson, K.E., White, O., Methe, B.A., Huttenhower, C., 2012. The human microbiome project: a community resource for the healthy human microbiome. *PLoS Biol.* 10.
- Goodman, R.E., 2006. Practical and predictive bioinformatics methods for the identification of potentially cross-reactive protein matches. *Mol. Nutr. Food Res.* 50, 655–660.
- Goodman, R.E., 2008. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. *Food Chemical Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc.* 46 (Suppl. 10), S24–S34.
- Goodman, R.E., Leach, J.N., 2004. Assessing the allergenicity of proteins introduced into genetically modified crops using specific human IgE assays. *J. AOAC Int.* 87, 1423–1432.
- Goodman, R.E., Vieths, S., Sampson, H.A., Hill, D., Ebisawa, M., Taylor, S.L., van Ree, R., 2008. Allergenicity assessment of genetically modified crops—what makes sense? *Nat. Biotechnol.* 26, 73–81.
- Griffith, J.J., Pollock, J., Klapper, D.G., Rogers, B.L., Nault, A.K., 1991. Sequence polymorphism of Amb a I and Amb a II, the major allergens in *Ambrosia artemisiifolia* (short ragweed). *Int. Arch. Allergy Appl. Immunol.* 96, 296–304.
- Group, N.H.W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., Baker, C.C., Di Francesco, V., Howcroft, T.K., Karp, R.W., Lunsford, R.D., Wellington, C.R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A.R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M.H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., Guyer, M., 2009. The NIH human microbiome project. *Genome Res.* 19, 2317–2323.
- Gupta, R.S., Springston, E.E., Warrier, M.R., Smith, B., Kumar, R., Pongracic, J., Holl, J.L., 2011. The prevalence, severity, and distribution of childhood food allergy in the United States. *Pediatrics* 128, e9–e17.
- Hamilton, R.G., Kleins-Tebbe, J., 2015. Molecular allergy diagnostics: analytical features that support clinical decisions. *Curr. Allergy Asthma Rep.* 15, 57.
- Hugouvieux-Cotte-Pattat, N., Condemine, G., Shevchik, V.E., 2014. Bacterial pectate lyases, structural and functional diversity. *Environ. Microbiol. Rep.* 6 (5), 427–440. <http://dx.doi.org/10.1111/1758-2229.12166>.
- Human Microbiome Project, 2012. A framework for human microbiome research. *Nature* 486, 215–221.
- Ivanciu, O., Schein, C.H., Braun, W., 2003. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.* 31, 359–362.
- Ivanciu, O., Schein, C.H., Garcia, T., Oezguen, N., Negi, S.S., Braun, W., 2009a. Structural analysis of linear and conformational epitopes of allergens. *Regul. Toxicol. Pharmacol.* 54, S11–S19.
- Ivanciu, O., Garcia, T., Torres, M., Schein, C.H., Braun, W., 2009b. Characteristic motifs for families of allergenic proteins. *Mol. Immunol.* 46, 559–568.
- Jenkins, J.A., Griffiths-Jones, S., Shewry, P.R., Breiteneder, H., Mills, E.N., 2005. Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: an in silico analysis. *J. Allergy Clin. Immunol. Pract.* 115, 163–170.
- Ladics, G.S., Bannon, G.A., Silvanovich, A., Cressman, R.F., 2007. Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens. *Mol. Nutr. Food Res.* 51, 985–998.
- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948.
- Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.
- Li, K.B., Issac, P., Krishnan, A., 2004. Predicting allergenic proteins using wavelet transform. *Bioinformatics* 20, 2572–2578.
- Magrane, M., Consortium, U., 2011. UniProt knowledgebase: a hub of integrated protein data. *Database J. Biol. Databases Curation* 2011 bar009.
- Maleki, S.J., Teuber, S.S., Cheng, H., Chen, D., Comstock, S.S., Ruan, S., Schein, C.H., 2011. Computationally predicted IgE epitopes of walnut allergens contribute to cross-reactivity with peanuts. *Allergy* 66, 1522–1529.
- Marin-Rodriguez, M.C., Orchard, J., Seymour, G.B., 2002. Pectate lyases, cell wall degradation and fruit softening. *J. Exp. Bot.* 53, 2115–2119.
- Mathura, V.S., Schein, C.H., Braun, W., 2003. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Bioinformatics* 19, 1381–1390.
- Midoro-Horiuti, T., Goldblum, R.M., Kurosky, A., Wood, T.G., Schein, C.H., Brooks, E.G., 1999. Molecular cloning of the mountain cedar (*Juniperus ashei*) pollen major allergen, Jun a 1. *J. Allergy Clin. Immunol. Pract.* 104, 613–617.
- Midoro-Horiuti, T., Mathura, V., Schein, C.H., Braun, W., Yu, S., Watanabe, M., Lee, J.C., Brooks, E.G., Goldblum, R.M., 2003. Major linear IgE epitopes of mountain cedar pollen allergen Jun a 1 map to the pectate lyase catalytic site. *Mol. Immunol.* 40, 555–562.
- Mirsky, H.P., Cressman Jr., R.F., Ladics, G.S., 2013. Comparative assessment of multiple criteria for the in silico prediction of cross-reactivity of proteins to known allergens. *Regul. Toxicol. Pharmacol.* 67, 232–239.
- Mittag, D., Akkerdaas, J., Ballmer-Weber, B.K., Vogel, L., Wensing, M., Becker, W.M., Koppelman, S.J., Knulst, A.C., Helbling, A., Hefle, S.L., Van Ree, R., Vieths, S., 2004. Ara h 8, a Bet v 1-homologous allergen from peanut, is a major allergen in patients with combined birch pollen and peanut allergy. *J. Allergy Clin. Immunol. Pract.* 114, 1410–1417.
- Mittag, D., Vieths, S., Vogel, L., Wagner-Loew, D., Starke, A., Hunziker, P., Becker, W.M., Ballmer-Weber, B.K., 2005. Birch pollen-related food allergy to legumes: identification and characterization of the Bet v 1 homologue in mungbean (*Vigna radiata*), Vig r 1. *Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol.* 35, 1049–1055.
- Mittag, D., Batori, V., Neudecker, P., Wiche, R., Friis, E.P., Ballmer-Weber, B.K., Vieths, S., Roggen, E.L., 2006. A novel approach for investigation of specific and cross-reactive IgE epitopes on Bet v 1 and homologous food allergens in individual patients. *Mol. Immunol.* 43, 268–278.
- Mueller, G.A., Maleki, S.J., Pedersen, L.C., 2014. The molecular basis of peanut allergy. *Curr. Allergy Asthma Rep.* 14.
- Muh, H.C., Tong, J.C., Tammi, M.T., 2009. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PLoS One* 4, e5861.
- Negi, S.S., Braun, W., 2017. Cross-React: a new structural bioinformatics method for predicting allergen cross-reactivity. *Bioinformatics* 33, 1014–1020.
- Nesbit, J.B., Hurlburt, B.K., Schein, C.H., Cheng, H., Wei, H., Maleki, S.J., 2012. Ara h 1 structure is retained after roasting and is important for enhanced binding to IgE. *Mol. Nutr. Food Res.* 56, 1739–1747.
- Pfiffner, P., Stadler, B.M., Rasi, C., Scala, E., Mari, A., 2012. Cross-reactions vs co-sensitization evaluated by in silico motifs and in vitro IgE microarray testing. *Allergy* 67, 210–216.
- Pichler, U., Hauser, M., Wolf, M., Bernardi, M.L., Gadermaier, G., Weiss, R., Ebner, C., Yokoi, H., Takai, T., Didierlaurent, A., Rafaianni, C., Briza, P., Mari, A., Behrendt, H., Wallner, M., Ferreira, F., 2015. Pectate lyase pollen allergens: sensitization profiles and cross-reactivity pattern. *PLoS One* 10, e0120038.
- Platts-Mills, T.A., 2015. The allergy epidemics: 1870–2010. *J. Allergy Clin. Immunol. Pract.* 136, 3–13.
- Pleis, J.R., Ward, B.W., Lucas, J.W., 2010. Summary health statistics for U.S. Adults: National Health Interview Survey, 2009. *Vital Health Stat.* 249 (10), 1–207 Data from the National Health Survey.
- Radauer, C., 2017. Navigating through the jungle of allergens: features and applications of allergen databases. *Int. Arch. Allergy Immunol.* 173, 1–11.
- Radauer, C., Willeroider, M., Fuchs, H., Hoffmann-Sommergruber, K., Thalhammer, J., Ferreira, F., Scheiner, O., Breiteneder, H., 2006. Cross-reactive and species-specific immunoglobulin E epitopes of plant profilins: an experimental and structure-based analysis. *Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol.* 36, 920–929.
- Relman, D.A., Falkow, S., 2001. The meaning and impact of the human genome sequence for microbiology. *Trends Microbiol.* 9, 206–208.
- Riaz, T., Hor, H.L., Krishnan, A., Tang, F., Li, K.B., 2005. WebAllergen: a web server for predicting allergenic proteins. *Bioinformatics* 21, 2570–2571.
- Savage, D.C., 1977. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* 31, 107–133.
- Schein, C.H., Ivanciu, O., Braun, W., 2005. Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens. *J. Agric. Food Chem.* 53, 8752–8759.
- Schein, C.H., Ivanciu, O., Braun, W., 2007. Bioinformatics approaches to classifying allergens and predicting cross-reactivity. *Immunol. Allergy Clin. North Am.* 27, 1–27.
- Schein, C.H., Ivanciu, O., Midoro-Horiuti, T., Goldblum, R.M., Braun, W., 2010. An allergen portrait gallery: representative structures and an overview of IgE binding surfaces. *Bioinform. Biol. Insights* 4, 113–125.
- Sicherer, S.H., Sampson, H.A., 2014. Food allergy: epidemiology, pathogenesis, diagnosis, and treatment. *J. Allergy Clin. Immunol.* 133, 291–307 quiz 308.
- Silvanovich, A., Nemeth, M.A., Song, P., Herman, R., Tagliani, L., Bannon, G.A., 2006. The value of short amino acid sequence matches for prediction of protein allergenicity. *Toxicol. Sci.* 90, 252–258.
- Stadler, M.B., Stadler, B.M., 2003. Allergenicity prediction by protein sequence. *FASEB J. Official Publ. Federation Am. Soc. Exp. Biol.* 17, 1141–1143.
- Valenta, R., Duchene, M., Pettenburger, K., Sillaber, C., Valent, P., Bettelheim, P., Breitenbach, M., Rumpold, H., Kraft, D., Scheiner, O., 1991. Identification of profilin as a novel pollen allergen; IgE autoreactivity in sensitized individuals. *Science* 253, 557–560.
- Venkatarajan, M.S., Braun, W., 2001. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J. Mol. Model.* 7, 445–453.
- Wang, Y., Fu, T.J., Howard, A., Kothary, M.H., McHugh, T.H., Zhang, Y., 2013. Crystal structure of peanut (*Arachis hypogaea*) allergen Ara h 5. *J. Agric. Food Chem.* 61, 1573–1578.
- Wing, R.A., Yamaguchi, J., Larabell, S.K., Ursin, V.M., McCormick, S., 1990. Molecular and genetic characterization of two pollen-expressed genes that have sequence similarity to pectate lyases of the plant pathogen *Erwinia*. *Plant Mol. Biol.* 14, 17–28.
- Zhang, L., Huang, Y., Zou, Z., He, Y., Chen, X., Tao, A., 2012. SORTALLER: predicting allergens using substantially optimized algorithm on allergen family featured peptides. *Bioinformatics* 28, 2178–2179.

## Supplementary Material:

**Table S1.** Distance matrix generated from the clinical observed cross-reactivities in the Bet v 1 family

|                     | Bet v 1 | Cor a 1.0401 | Cor a 1.0101 | Pru p 1 | Aln g 1 | Mal d 1 | Ara h 8 | Gly m 4 | Api g 1 | Dau c 1 |
|---------------------|---------|--------------|--------------|---------|---------|---------|---------|---------|---------|---------|
| <b>Bet v 1</b>      | 0       | 0.416        | 0.216        | 0.27    | 0.239   | 0.223   | 0.466   | 0.783   | 0.836   | 1.488   |
| <b>Cor a 1.0401</b> | 0.416   | 0            | 0.434        | 0.384   | 0.414   | 0.401   | 0.604   | 0.828   | 0.933   | 1.512   |
| <b>Cor a 1.0101</b> | 0.216   | 0.434        | 0            | 0.295   | 0.255   | 0.295   | 0.444   | 0.663   | 0.84    | 1.387   |
| <b>Pru p 1</b>      | 0.27    | 0.384        | 0.295        | 0       | 0.236   | 0.211   | 0.361   | 0.647   | 0.737   | 1.285   |
| <b>Aln g 1</b>      | 0.239   | 0.414        | 0.255        | 0.236   | 0       | 0.249   | 0.376   | 0.647   | 0.707   | 1.368   |
| <b>Mal d 1</b>      | 0.223   | 0.401        | 0.295        | 0.211   | 0.249   | 0       | 0.378   | 0.647   | 0.719   | 1.285   |
| <b>Ara h 8</b>      | 0.466   | 0.604        | 0.444        | 0.361   | 0.376   | 0.378   | 0       | 0.461   | 0.589   | 1.171   |
| <b>Gly m 4</b>      | 0.783   | 0.828        | 0.663        | 0.647   | 0.647   | 0.647   | 0.461   | 0       | 0.683   | 1.117   |
| <b>Api g 1</b>      | 0.836   | 0.933        | 0.84         | 0.737   | 0.707   | 0.719   | 0.589   | 0.683   | 0       | 0.606   |
| <b>Dau c 1</b>      | 1.488   | 1.512        | 1.387        | 1.285   | 1.368   | 1.285   | 1.171   | 1.117   | 0.606   | 0       |

**Table S2.** Homologues of pectate lyases in the Human Microbiome

| Seq_ID           | Protein name                   | Organism                                    | E-Value  |
|------------------|--------------------------------|---------------------------------------------|----------|
| HMPREF0004_0963  | conserved hypothetical protein | Achromobacter piechaudii ATCC 43553         | 2.00E-13 |
| HMPREF0005_04335 | hypothetical protein           | Achromobacter xylosoxidans C54              | 1.00E-12 |
| HMPREF9321_1726  | putative pectate lyase         | Veillonella atypica ACS+049+V+Sch6          | 7.00E-11 |
| HMPREF0908_0299  | pectate lyase                  | Selenomonas flueggei ATCC 43531             | 2.00E-10 |
| HMPREF1019_01313 | hypothetical protein           | Campylobacter sp. 10_1_50                   | 5.00E-10 |
| HMPREF9166_1475  | pectate lyase                  | Selenomonas sp. oral taxon 149 str. 67H29BP | 7.00E-10 |
| HMPREF1012_01654 | pel protein                    | Bacillus sp. BT1B_CT2                       | 1.00E-08 |
| HMPREF7545_0107  | pectate lyase                  | Selenomonas noxia ATCC 43541                | 2.00E-08 |
| HMPREF1012_03545 | pectate lyase                  | Bacillus sp. BT1B_CT2                       | 3.00E-06 |
| RUM_11790        | Pectate lyase                  | Ruminococcus champanellensis 18P13          | 7.00E-06 |
| HMPREF9412_2007  | pectate lyase                  | Paenibacillus sp. HGF5                      | 2.00E-05 |
| CUW_2559         | pectate lyase                  | Turicibacter sanguinis PC909                | 1.00E-04 |
| BACPEC_00859     | pectate lyase                  | Clostridiales                               | 2.00E-04 |

|                  |                                                    |                                 |          |
|------------------|----------------------------------------------------|---------------------------------|----------|
| HMPREF9402_0417  | pectate lyase                                      | Turicibacter sp.<br>HGF1        | 2.00E-04 |
| HMPREF6485_0196  | CHU large<br>protein<br>candidate<br>pectate lyase | Prevotella buccae<br>ATCC 33574 | 2.00E-03 |
| HMPREF0649_01089 | CHU large<br>protein<br>candidate<br>pectate lyase | Prevotella buccae<br>D17        | 2.00E-03 |
| HMPREF1146_2484  | pectate lyase                                      | Prevotella sp.<br>MSX73         | 2.00E-03 |
| PREVCOP_04892    | pectate lyase<br>related protein,<br>secreted      | Prevotella copri<br>DSM 18205   | 2.00E-03 |

**Fig. S1 PCP Motifs for the allergens in the 17 top protein families.** For each protein family of allergens the names of the families, their PFAM identifier, the reference sequence with gene bank identifier and allergen name is given. The reference sequence is used for the display of the locations of the motifs which are marked on the sequence. The significant amino acid residue positions within the motifs are given below with a graphical description of the standard deviations (dark blocks means conserved) and values for the five PCP descriptors.

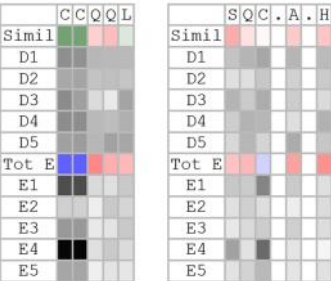
**(1) Allergen motifs for plant lipid transfer proteins**

**PFAM ID: PF00234**

**Reference sequence: GB\_170710\_151\_305**

**Allergen: Tri a gliadin wheat storage protein (Triticum aestivum)**

```
1  CRD VV LQQ H N I A H A S S Q V L Q Q S T Y Q L L Q Q L C C Q Q L L Q I P E Q S Q C Q A I H N V 50
51 A H A I I M H Q Q Q Q Q Q Q E Q K Q Q L Q Q Q Q Q Q Q Q L Q Q Q Q Q Q Q Q P S S Q V S F Q Q P 100
101 Q Q Q Y P S S Q V S F Q P S Q L N P Q A Q G S V Q P Q Q L P Q F A E I R N L A L Q T L P A M C N V Y 150
151 I P P H C
```



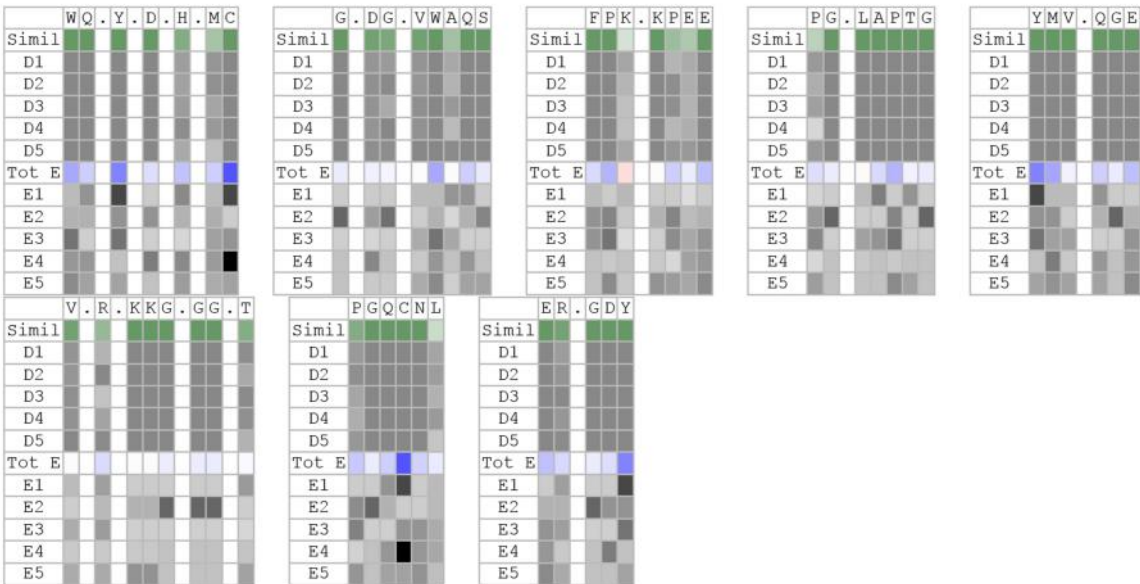
**(2) Allergen motifs for profilins**

**PFAM ID: PF00235**

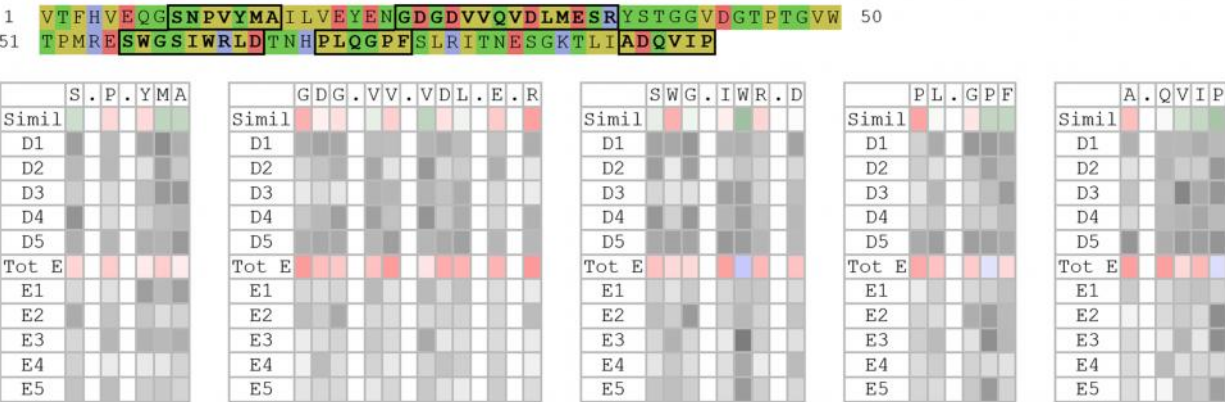
**Reference sequence: GB\_18652049\_2\_129**

**Allergen: Dau\_c\_4 profilin carrot allergen (Daucus carota)**

```
1  S W Q T Y V D D H L M C E V D G N P G Q Q L S A A A I I G H D G S V W A Q S S T F P K F K P E E I T 50
51 G I M K N F D E P G H L A P T G L Y L G G T K Y M V I Q G E P I A V I R G K K G S G G V T I K K T G 100
101 Q A L V F G V Y D E P V T P G Q C N L I V E R L G D Y L
```



(3) Allergen motifs for grass pollen allergens  
PFAM ID: PF01357  
Reference sequence: GB\_8118432\_220\_311  
Allergen: Ory\_s\_1 C-terminal domain rice pollen allergen (*Oryza sativa*)



(4) Allergen motifs for calcium binding proteins  
PFAM ID: PF00036  
Reference sequence: GB\_20796729\_42-72  
Allergen: Ran\_e\_1 alpha-parvalbumin (*Rana esculenta*)



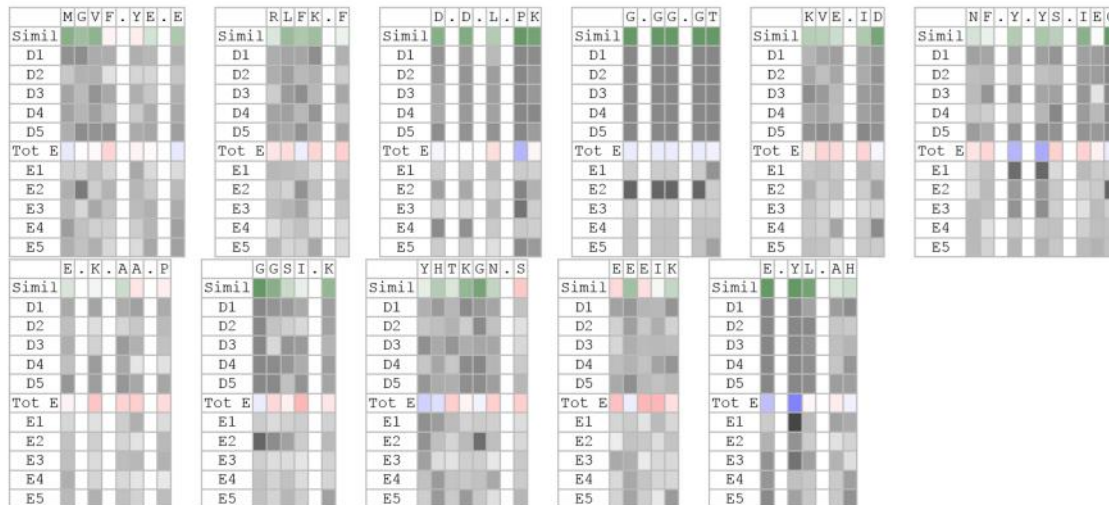
(5) Motifs for PR-10 related proteins

PFAM ID: PF00407

Reference sequence: GB\_5726304\_1\_156

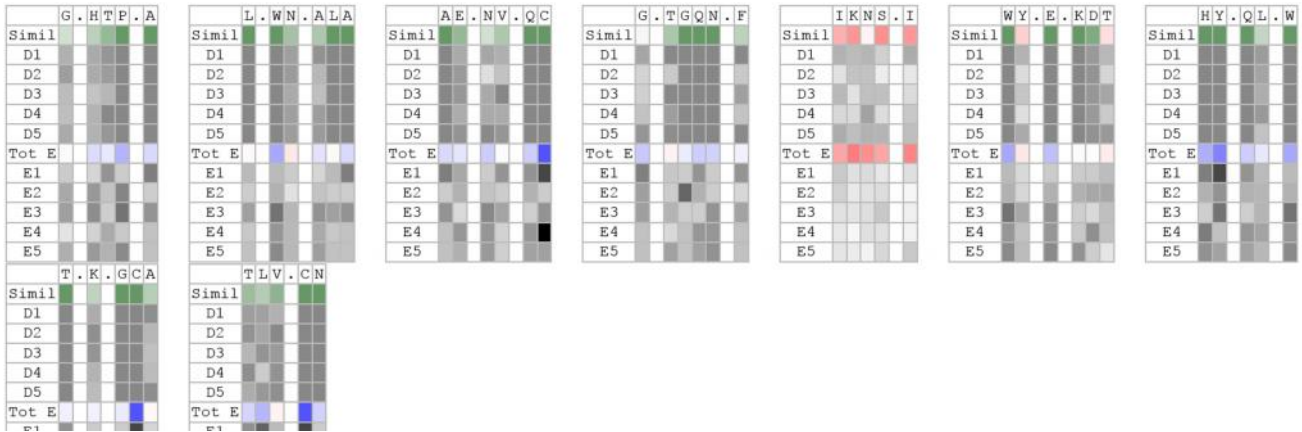
Allergen Cor\_a\_1.0401 major hazelnut allergen (*Corylus avellana*)

1 MGVFCYEDEATSVIPPA~~RLFKSF~~VL~~DADNLIPK~~VAPQHFTSAENLE~~GNGG~~ 50  
51 PGTIKKITFAEGNEFKYMKHKVEEIDHANFKYCYSIIEGGPLGHTLEKIS 100  
101 YEIKMAAAPHG~~GGSILK~~ITSKYHTKGNASIN~~EEEIK~~AGKEKAAGLPKAVE 150  
151 AYLLAH



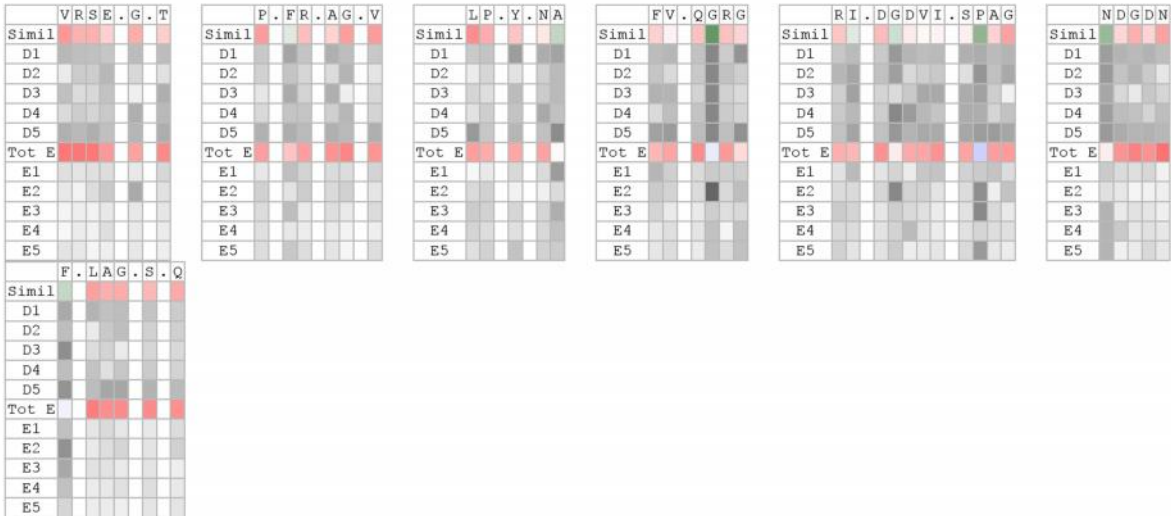
(6) Motifs for SCP-like extracellular proteins  
PFAM ID: PF00188  
Reference sequence: GB\_7638032\_73\_226  
Allergen Cte\_f\_2 salivary allergen 2 (Ctenocephalides felis felis)

1 LNAHNRRKRR LVAAGKGLLKDGVHTPIAAKMPNLTWNIALAKLAEYNVKQC 50  
51 EMKHDCAKTRHGHGTGQNLFFYGTTLSPIKNSTIAKMAVDGWYAESKDTRL 100  
101 EDIKKLTITIYPNGKPIGHYTQLIWNNTTKVGCAVSTYKKHSNGNMFNMTL 150  
151 VACN

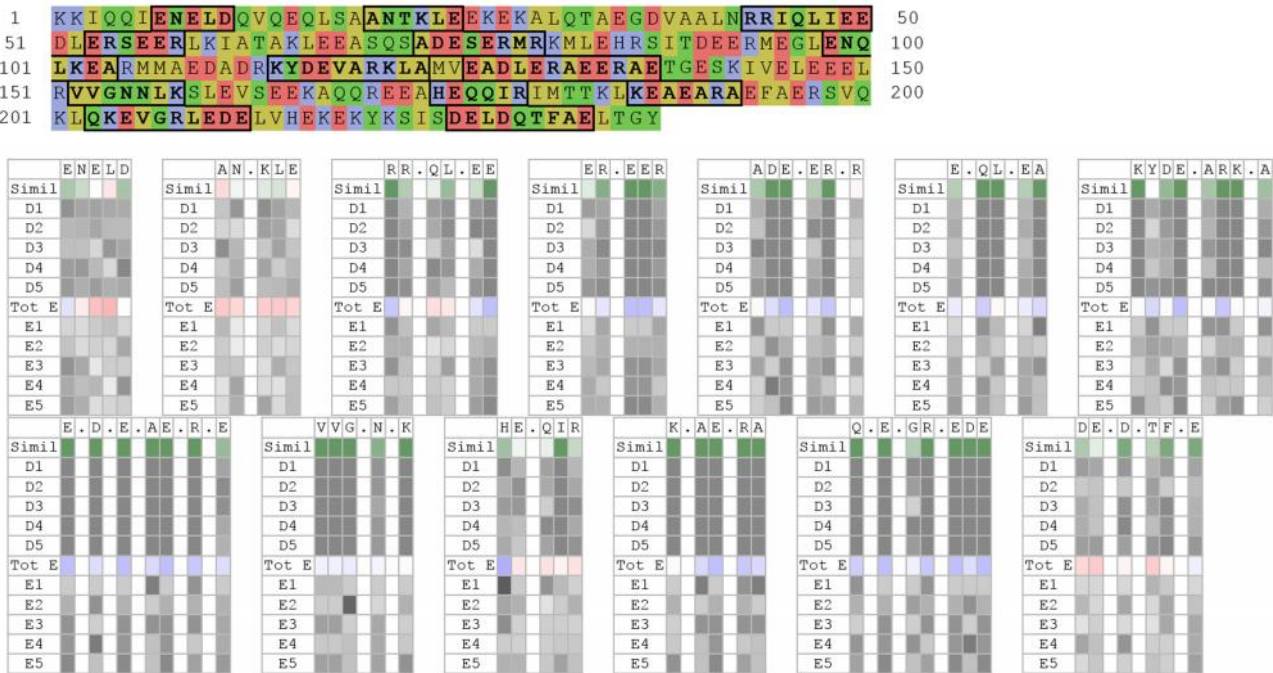


(7) Motifs for cupins  
PFAM ID: PF00190  
Reference sequence: GB\_2317670\_49\_275  
Allergen Fag\_e\_1 seed storage protein (Fagopyrum esculentum)

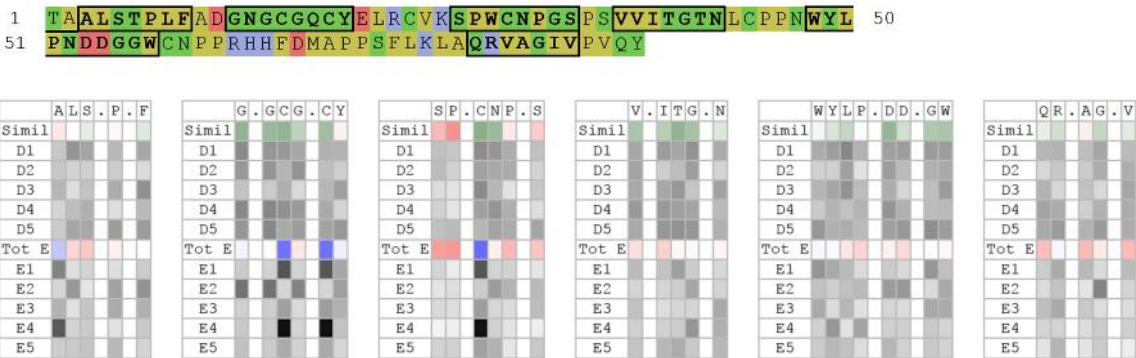
1 LTASEPSRRVRSEAGVTEIWDNDTPEFRCAGFVAVRVVVIQPGGLL LPSYS 50  
51 NAPIYITFVEQGRGVQGVVVP GCPETFQSESEFEYPQSQRDQRSRQSESEE 100  
101 SSRGDQRTROQSESEEFSSRGDQRTROQSESEEFSSRGDQRTROQSESEEFSSRGD 150  
151 QRTRQSESEEFSSRGDQHQKIFRIRDGDVIPSPAGVVQWTHNDGDN DLIS I 200  
201 TLYDANSFQNLQDGNVRNFFLAGQSKQ



(8) Motifs for tropomyosins  
PFAM ID: PF00261  
Reference sequence: SP\_O18416\_48\_284  
Allergen Der p 10 European house dust mite allergen (*Dermatophagoides pteronyssinus*)



(9) Motifs for RlpA-like double-psi beta barrel proteins  
PFAM ID: PF03330  
Reference sequence: GB\_6069656\_89\_175  
Allergen Ory\_s\_1 N-terminal domain rice pollen allergen (*Oryza sativa*)



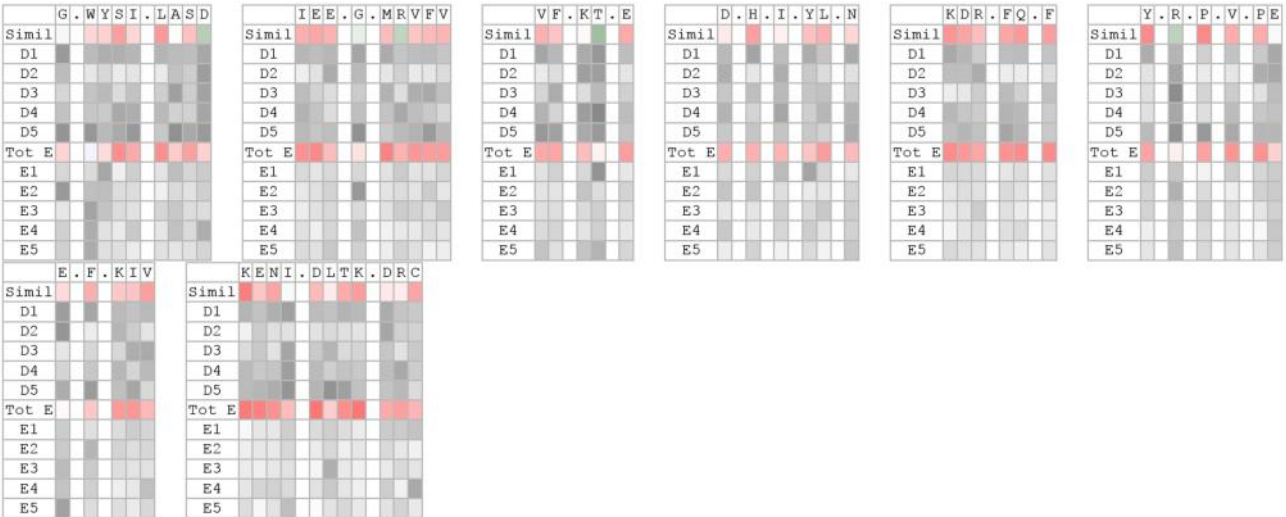
(10) Motifs for lipocalins

PFAM ID: PF00061

Reference sequence: SP\_Q95182\_32\_176

Allergen Equ\_c\_1 horse lipocalin (*Equus caballus*)

```
1  KISGEWYSIFLASDVKEKIEENGSMRVFVDVIRALDNSSLYAEYQTKVNG 50
51 ECTEFPMVFDKTEEDGVYSLNYDGYNVFRISEFENDEHIILYLVNFDKDR 100
101 PFQLFEFYAREPDVSPSEIKEEFVKIVQKRGIKKENIIDLTIDRC
```



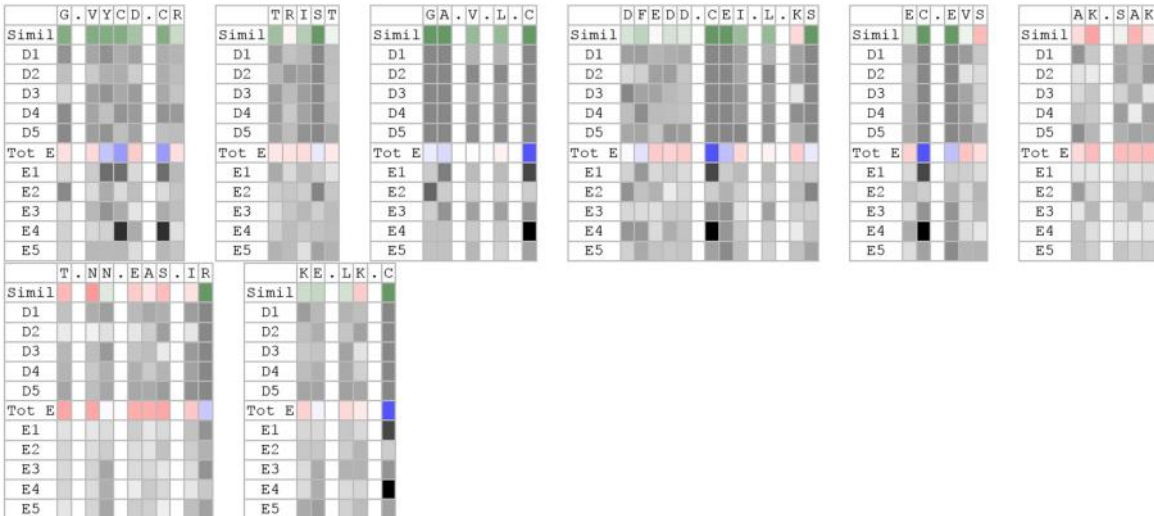
(11) Motifs for Pollen proteins (Ole e 1 family)

PFAM ID: PF01190

Reference sequence: GB\_22074346\_32\_155

Allergen Che\_a\_1 Ole e 1 like protein (*Chenopodium album*)

```
1  VQGMVYCDTCRIQFMTRISTIMEGATVKLECRNITAGTQTFKAEAVTDKV 50
51 GQYSIPVNGDFEDDICEIELVKS PNSECSSEVSHD VYAKQSAKVSLTSNNG 100
101 EASDIRSANALGFMRKEPLKECP E
```



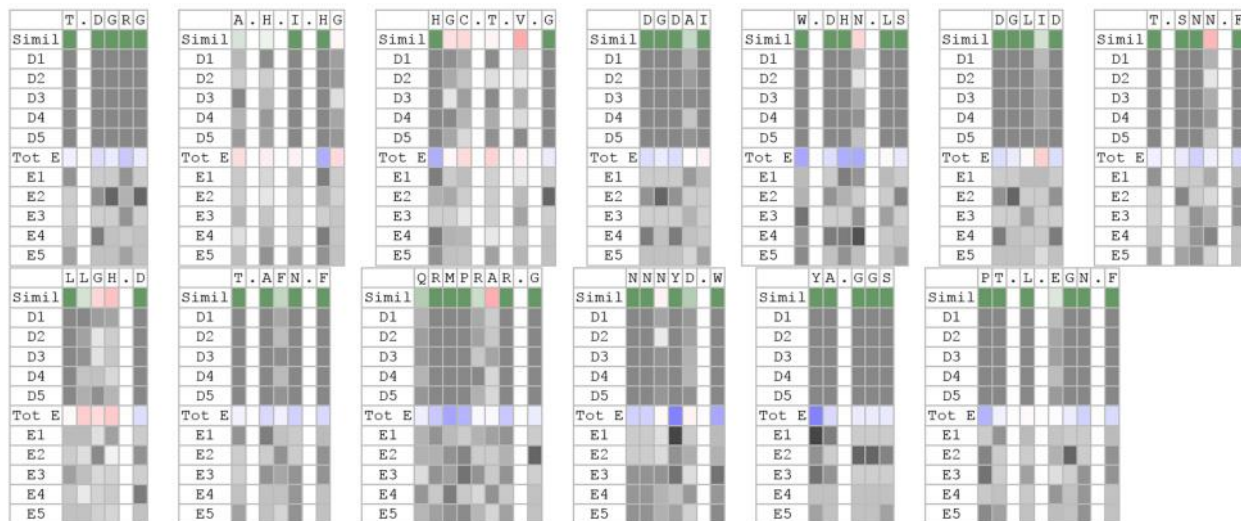
### (12) Motifs for pectate lyases

PFAM ID: PF00544

Reference sequence: GB\_19069497\_102\_289

Allergen Cup\_a\_1 cypress pollen allergen (*Cupressus arizonica*)

```
1  QMPLYVAGYK TIDGRG ADVHLGNGGPCLFMR TASHVILHGLHI HGCNTSV 50
51 LGDVLVSE SIGVEPVHAQ DGDATMNRNVTNA WIDHNSLS DCS DGLIDVTL 100
101 GSTGI TISNNHFFNHHKVM LLGHDDTY DDDISMKV TVAFNQF GPNAGQRM 150
151 PRARYGLVHVANNNDQWNIYAIGGSNPTILSEGNSE
```



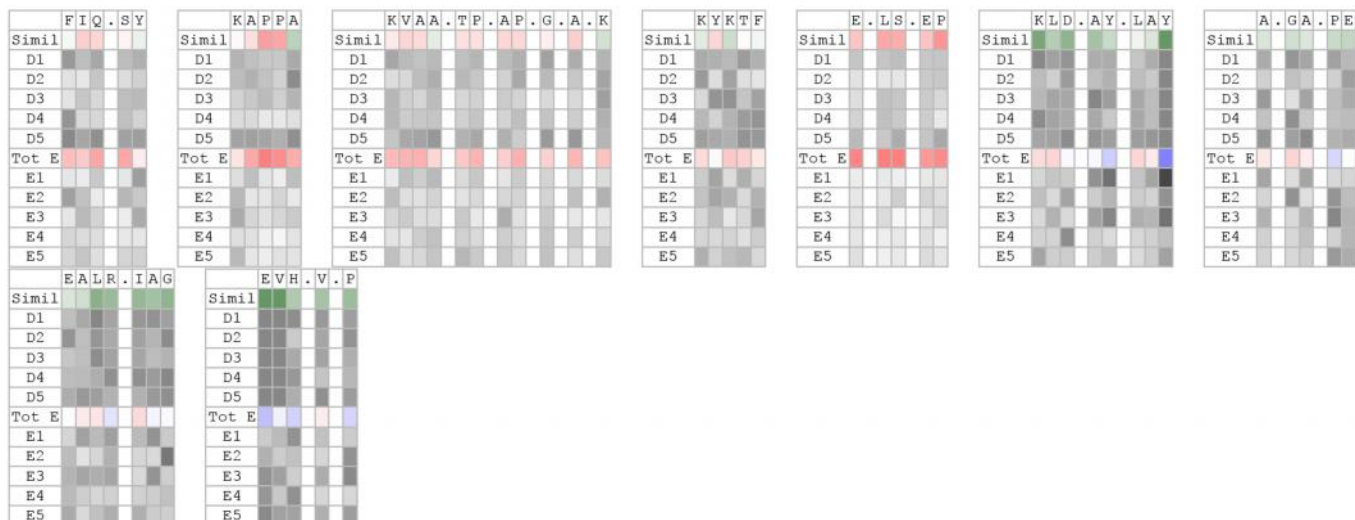
### (13) Motifs for Ribonucleases (pollen allergen)

PFAM ID: PF01620

Reference sequence: SP\_P22284\_42\_237

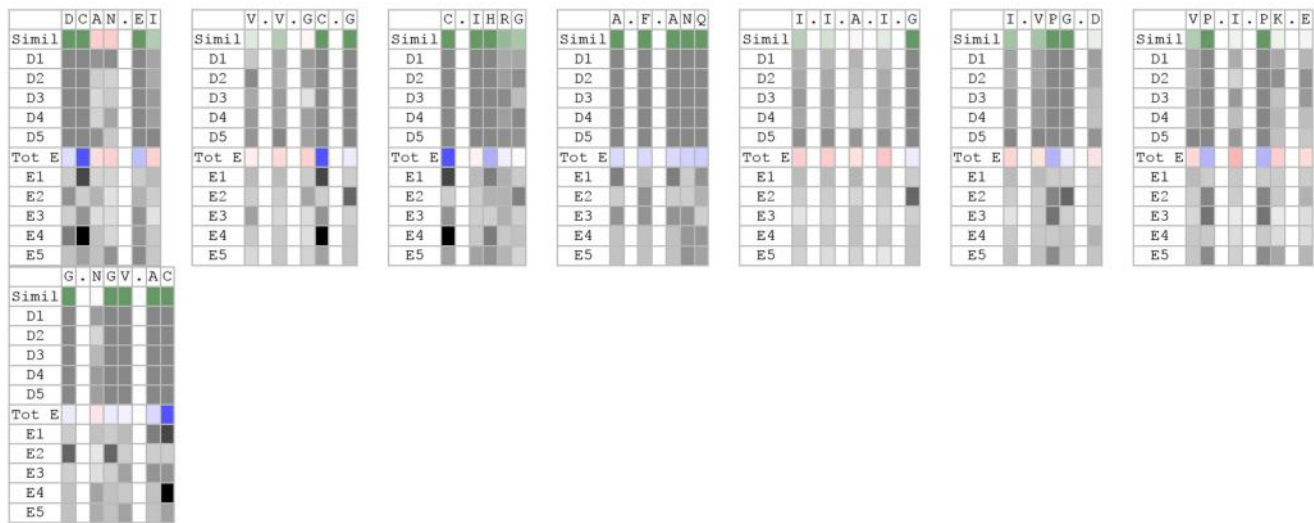
Allergen Poa p 9 grass pollen allergen (*Poa pratensis*)

```
1  PDAVGFAKKLDAFIQTSYLSTKAAEPKEKFDLFVLSL TEVLRFMAGAVKA 50
51 PPA SKFPAPAP KVAAYTPAAPAGAAPKATTDEQKLI EKINVGFKAAVAA 100
101 AAGVPAAS KYKTFVATFGAASNKAFAEALSTEPKGA AVASSKAVLTS KLD 150
151 AAYKLAYKS AEGATPEAKYDAYVATLS EALRIIAGTLE VHGVKPPAA
```



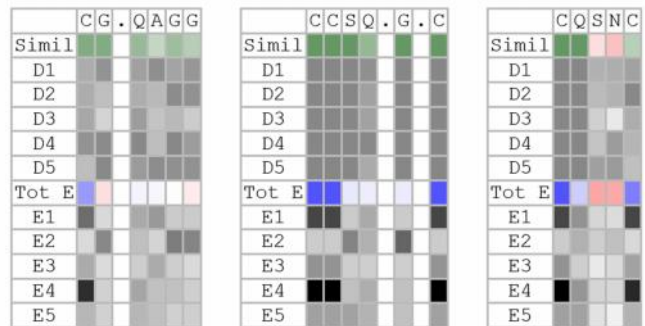
(14) Motifs for MD-2-related lipid-recognition domain  
PFAM ID: PF02221  
Reference sequence: SP\_Q9TZZ2\_17\_142  
Allergen Eur\_m\_2 MD-2 lipid recognition domain house dust mite (Euroglyphus maynei)

1 DQVDIKDCANHEIKKVMVPGCKGSEPCVIHRGTAFQLEAVFDANQNSNAA 50  
51 KIEIKATIDGVEIDVPGIDNNLCHFMKCPLVKGQEYDIKYTWNVPRIAPK 100  
101 SENVVVTVKLLGDNGLACAIATHAK



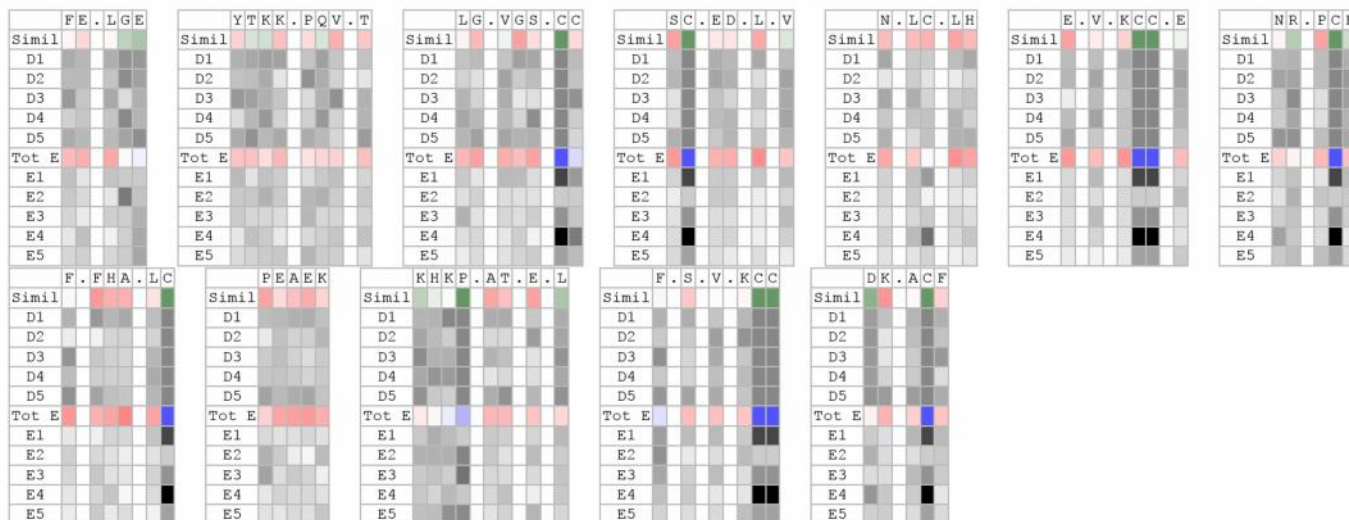
(15) Motifs for Chitin recognition protein  
PFAM ID: PF00187  
Reference sequence: SP\_P02877\_17\_58  
Allergen Hev\_b\_6.01 tropomyosin latex (Hevea brasiliensis)

1 AEQCGRQAGGKLCPNNLCCSQWGWCSTDEYCSPDHNCQSNC



**(16) Motifs for Serum albumin family**  
**PFAM ID: PF00273**  
**Reference sequence: SP\_P49064\_412\_592**  
**Allergen Fel d 2 cat saliva allergen (*Felis domesticus*)**

1 V K T N C E L F E K L G E Y G F Q N A L L V R Y T K K V P Q V S T P T L V E V S R S L G K V G S K C 50  
51 C T H P E A E R L S C A E D Y L S V L N R L C V L H E K T P V S E R V T K C C T E S L V N R R P C 100  
101 F S A L Q V D E T Y V P K E F S A E T F T F H A D L C T L P E A E K Q I K K Q S A L V E L L K H K P 150  
151 K A T E E Q L K T V M G D F G S F V D K C C A A E D K E A C F



**(17) Motifs for Insect allergen related repeat**  
**PFAM ID: PF06757**  
**Reference sequence: GB\_4572592\_332\_406**  
**Allergen Bla\_g\_1.0103 German cockroach allergen (*Blattella germanica*)**

1 I D D V L A I L P L D D L K A L F N E K L E T S P D F L A L Y N A I K S P E F Q S I V Q T L N A M P 50  
51 E Y Q N L L E K L R E K G V D V D K I I E L I R A

